РОССИЙСКАЯ АКАДЕМИЯ НАУК
ОТДЕЛЕНИЕ МАТЕМАТИЧЕСКИХ НАУК РАН
ВЫЧИСЛИТЕЛЬНЫЙ ЦЕНТР ИМ. А. А. ДОРОДНИЦЫНА РАН
при поддержке
РОССИЙСКОГО ФОНДА ФУНДАМЕНТАЛЬНЫХ ИССЛЕДОВАНИЙ
КОМПАНИИ FORECSYS

Математические методы распознавания образов

ММРО-14

г. Суздаль,
21–26 сентября 2009

Доклады XIV Всероссийской конференции

МАКС ПРЕСС
Москва, 2009

УДК 004.85+004.89+004.93+519.2+519.25+519.7
ББК 22.1:32.973.26-018.2
М34

М34 **Математические методы распознавания образов:** 14-я Всероссийская конференция. Владимирская обл., г. Суздаль, 21–26 сентября 2009 г.: Сборник докладов. — М.: МАКС Пресс, 2009. — 632 с.

ISBN 978-5-317-02947-0

В сборнике представлены доклады 14-й Всероссийской конференции «Математические методы распознавания образов», проводимой Вычислительным центром им. А. А. Дородницына Российской академии наук при финансовой и организационной поддержке РФФИ и компании Forecsys.

Конференция регулярно проводится один раз в два года, начиная с 1983 г., и является самым представительным российским научным форумом в области распознавания образов и анализа изображений, интеллектуального анализа данных, машинного обучения, обработки сигналов, математических методов прогнозирования.

УДК 004.85+004.89+004.93+519.2+519.25+519.7
ББК 22.1:32.973.26-018.2

ISBN 978-5-317-02947-0

Оргкомитет

Председатель: Журавлев Юрий Иванович, *академик РАН*
Зам. председателя: Матросов Виктор Леонидович, *академик РАН*
Ученый секретарь: Чехович Юрий Викторович, *к.ф.-м.н.*

Донской Владимир Иосифович, *д.ф.-м.н.*
Дедус Флоренц Федорович, *д.т.н.*
Немирко Анатолий Павлович, *д.ф.-м.н.*
Устинин Михаил Николаевич, *д.ф.-м.н.*
Инякин Андрей Сергеевич, *к.ф.-м.н.*
Песков Николай Владимирович, *к.ф.-м.н.*
Романов Михаил Юрьевич, *к.ф.-м.н.*

Программный комитет

Председатель: Рудаков Константин Владимирович, *чл.-корр. РАН*
Зам. председателя: Дюкова Елена Всеволодовна, *д.ф.-м.н.*
Ученый секретарь: Воронцов Константин Вячеславович, *к.ф.-м.н.*

Микаэлян Андрей Леонович, *академик РАН*
Жижченко Алексей Борисович, *чл.-корр. РАН*
Сойфер Виктор Александрович, *чл.-корр. РАН*
Местецкий Леонид Моисеевич, *д.т.н.*
Моттль Вадим Вячеславович, *д.ф.-м.н.*
Пытьев Юрий Петрович, *д.ф.-м.н.*
Рязанов Владимир Васильевич, *д.ф.-м.н.*
Рейер Иван Александрович, *к.т.н.*

Технический комитет

Председатель: Громов Андрей Николаевич

Гуз Иван Сергеевич
Ефимов Александр Николаевич
Ивахненко Андрей Александрович
Каневский Даниил Юрьевич
Лисица Андрей Валерьевич
Никитов Глеб Владимирович

Рецензенты

Рудаков Константин Владимирович, *чл.-корр. РАН*
Местецкий Леонид Моисеевич, *д.т.н.*
Моттль Вадим Вячеславович, *д.т.н.*
Мясников Владислав Валерьевич, *д.ф.-м.н.*
Немирко Анатолий Павлович, *д.ф.-м.н.*
Устинин Михаил Николаевич, *д.ф.-м.н.*
Федотов Николай Гаврилович, *д.ф.-м.н.*
Хачай Михаил Юрьевич, *д.ф.-м.н.*
Чернов Владимир Михаилович, *д.ф.-м.н.*
Воронцов Константин Вячеславович, *к.ф.-м.н.*
Гашников Михаил Валерьевич, *к.т.н.*
Глумов Николай Иванович, *к.т.н.*
Гуров Сергей Исаевич, *к.ф.-м.н.*
Дьяконов Александр Геннадиевич, *к.ф.-м.н.*
Инякин Андрей Сергеевич, *к.ф.-м.н.*
Майсурадзе Арчил Ивериевич, *к.ф.-м.н.*
Рейер Иван Александрович, *к.т.н.*
Стрижов Вадим Викторович, *к.ф.-м.н.*
Чехович Юрий Викторович, *к.ф.-м.н.*
Чичёва Марина Александровна, *к.т.н.*

Краткое оглавление

Фундаментальные основы распознавания и прогнозирования

Код раздела: TF (Theory and Fundamentals)

- Статистические основы обучения по прецедентам.
- Дискретно-логические основы обучения по прецедентам.
- Алгебраический подход к проблеме распознавания.
- Проблема обобщающей способности.
- Теория возможности и неопределённые нечёткие модели.
- Устойчивость обучения.
- Байесовский вывод.

Точные оценки вероятности переобучения для монотонных и унимодальных семейств алгоритмов*

Ботов П. В.
pbotov@forecsys.ru
Московский физико-технический институт

В рамках комбинаторного подхода получены точные оценки вероятности переобучения h-мерной монотонной сетки и приближённые оценки для h-мерной унимодальной сетки. Показано, что в случае единичной длины контрольной подвыборки вероятность переобучения линейно возрастает с ростом размерности h. Выявлена возможность аппроксимации унимодальных сеток монотонными удвоенной размерности. В экспериментах на данных из репозитория UCI исследована аппроксимируемость вероятности переобучения некоторых семейств алгоритмов её оценками для монотонных сеток подходящей размерности.

В рамках комбинаторного подхода в [1, 2] были получены точные верхние оценки вероятности переобучения для некоторых семейств алгоритмов простой структуры. В [2] такие оценки приводятся для монотонных и унимодальных цепочек алгоритмов, а также для единичной окрестности лучшего алгоритма. Монотонная цепочка является несколько идеализированной моделью однопараметрического связного семейства алгоритмов. Известно, что для линейного классификатора, разделяющего выборку без ошибок, непрерывное отклонение направляющего вектора разделяющей гиперплоскости от оптимального положения порождает монотонную цепочку. Однако при изменении условий этого модельного эксперимента — использовании нелинейной модели или неразделимой выборки — порождается цепочка не монотонная, хотя и близкая, в некотором смысле, к монотонной. Точное вычисление вероятности переобучения для таких случаев представляется гораздо более трудной задачей. Тем не менее, остаётся возможность аппроксимировать вероятность переобучения «реальных» семейств алгоритмов оценками, исходно полученными для «идеальных» семейств.

В данной работе рассматриваются многомерные обобщения монотонных и унимодальных цепочек — монотонные и унимодальные h-мерные сетки алгоритмов. Для них приводятся точные верхние оценки вероятности переобучения. Многомерные сетки являются сильно идеализированными моделями многопараметрических связных семейств алгоритмов. Например, для семейства линейных классификаторов в $\mathbb{R}^h$ монотонная h-мерная сетка порождается лишь при весьма специфическом расположении объектов выборки, которое едва ли может встретиться на практике. Тем не менее, оказывается, что вероятность переобучения «реальных» семейств неплохо аппроксимируется оценками, исходно полученными для h-мерной сетки, при подборе подходящего значения параметра размерности h.

Далее используются основные понятия и обозначения, введённые в [1, стр. 18 в этом сборнике]: генеральная выборка $\mathbb{X} = \{x_1, \ldots, x_L\}$, бинарная функция ошибки $I(a, x)$, число ошибок $n(a, X)$ алгоритма a на выборке $X \subseteq \mathbb{X}$, метод обучения $\mu\colon X \mapsto a$, функционал вероятности переобучения Q_ε, функция гипергеометрического распределения $H_L^{\ell,m}(s)$.

Монотонные семейства алгоритмов

Рассмотрим частный случай — монотонную цепочку, или же монотонную одномерную сетку.

Определение 1. *Множество алгоритмов* $A = \{a_0, a_1, \ldots, a_D\}$ *называется монотонной цепочкой алгоритмов, если* $I(a_d, x_i) \leqslant I(a_{d+1}, x_i)$ *для всех* $x_i \in \mathbb{X}$ *и* $n(a_d, \mathbb{X}) = m+d$ *при некотором* $m \geqslant 0$. *Алгоритм* a_0 *называется лучшим в цепочке.*

Пример 1. Представление монотонной цепочки в виде таблицы «объекты–алгоритмы»:

	a_0	a_1	a_2	a_3	a_4	$\cdots$
x_1	0	**1**	1	1	1	1
x_2	0	0	**1**	1	1	1
x_3	0	0	0	**1**	1	1
x_4	0	0	0	0	**1**	1
$\cdots$	0	0	0	0	0	**1**

Здесь d-й столбец соответствует вектору ошибок алгоритма a_d; i-я строка — объекту x_i; значение в (i, d)-й ячейке $I(a_d, x_i)$ равно 1, если алгоритм a_d ошибается на объекте x_i.

Определение 2. *Метод минимизации эмпирического риска* $\mu\colon [\mathbb{X}]^\ell \to A$ *называется пессимистичным, если в случаях, когда минимум* $n(a, X)$ *достигается на многих алгоритмах,* μ *выбирает алгоритм с бóльшим* $n(a, \mathbb{X})$. *Если же и таких алгоритмов несколько, то* μ *выбирает любой из них с равной вероятностью.*

Пессимистичный метод на практике нереализуем, но он даёт верхние оценки Q_ε, завышенность которых невелика и связана только с неоднозначным выбором минимума эмпирического риска.

*Работа поддержана РФФИ (проект №08-07-00422) и программой ОМН РАН «Алгебраические и комбинаторные методы математической кибернетики и информационные системы нового поколения».

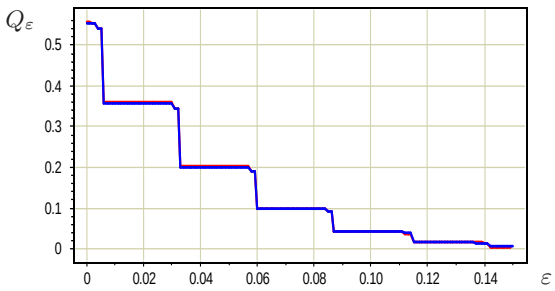


Рис. 1. Зависимости вероятности переобучения Q_ε и её эмпирической оценки $\hat{Q}_\varepsilon$ для монотонной цепочки. Параметры эксперимента: $L = 500$, $l = 460$, $m = 60$.

Теорема 1 (О монотонной цепочке [1]). *Пусть A — монотонная цепочка, метод обучения μ является пессимистичной минимизацией эмпирического риска и выполнено условие $k \leqslant D \leqslant L - m$. Тогда вероятность получить каждый из алгоритмов цепочки в результате обучения есть*

$$P_t = \mathsf{P}[\mu X = a_t] = C_{L-1-t}^{l-1} / C_L^l;$$

вероятность переобучения выражается в виде

$$Q_\varepsilon = \sum_{t=0}^{k} P_t H_{L-1-t}^{l-1,m}(s_t(\varepsilon)), \quad s_t(\varepsilon) = \frac{l}{L}(m+t-\varepsilon k).$$

Полученную оценку легко проверить эмпирически, сравнив с оценкой по методу Монте-Карло $\hat{Q}_\varepsilon$, вычисляемой как доля разбиений выборки, при которых величина переобученности превышает ε. Различие двух графиков слабо заметно, рис. 1.

Введём целочисленный вектор индексов $\boldsymbol{d} = (d_1, \ldots, d_h)$. Обозначим $|\boldsymbol{d}| = d_1 + \cdots + d_h$. На множестве векторов индексов введём покомпонентное отношение сравнения: $\boldsymbol{d} < \boldsymbol{d}'$, если $d_j \leqslant d'_j$, $j = 1, \ldots, h$, и хотя бы одно из неравенств строгое.

Определение 3. *Множество алгоритмов $A = \{a_{\boldsymbol{d}}\}_{|\boldsymbol{d}| \leqslant D}$, называется монотонной h-мерной сеткой алгоритмов, если выполнены два условия:*

1) *если $\boldsymbol{d} < \boldsymbol{d}'$, то для всех $x_i \in \mathbb{X}$ выполнено $I(a_{\boldsymbol{d}}, x_i) \leqslant I(a_{\boldsymbol{d}'}, x_i)$;*
2) *$n(a_{\boldsymbol{d}}, \mathbb{X}) = m + |\boldsymbol{d}|$ при некотором $m \geqslant 0$.*

Алгоритм $a_{\boldsymbol{0}}$ называется лучшим в сетке.

Множество алгоритмов с равным числом ошибок $t + m = n(a_{\boldsymbol{d}}, \mathbb{X})$ называются t-м слоем сетки.

Пример 2. Монотонная двумерная сетка для выборки из $L = 4$ объектов:

	$a_{0,0}$	$a_{1,0}$	$a_{2,0}$	$a_{0,1}$	$a_{1,1}$	$a_{2,1}$	$a_{0,2}$	$a_{1,2}$	$a_{2,2}$
x_1	0	**1**	1	0	**1**	1	0	**1**	1
x_2	0	0	**1**	0	0	**1**	0	0	**1**
x_3	0	0	0	**1**	**1**	**1**	1	1	1
x_4	0	0	0	0	0	0	**1**	**1**	**1**

Теорема 2 (О монотонной h-мерной сетке). *Пусть A — h-мерная монотонная сетка, метод μ является пессимистичной минимизацией эмпирического риска и выполнено условие $kh \leqslant Dh \leqslant L - m$. Тогда вероятность получить каждый из алгоритмов сетки в результате обучения есть*

$$P_{\boldsymbol{d}} = \mathsf{P}[\mu X = a_{\boldsymbol{d}}] = C_{L-h-|\boldsymbol{d}|}^{l-h} / C_L^l;$$

вероятность получить какой-либо из алгоритмов слоя $t + m$ есть

$$P_t = \sum_{|\boldsymbol{d}|=t} P_{\boldsymbol{d}} = C_{h+t-1}^{t} C_{L-h-t}^{l-h} / C_L^l;$$

вероятность переобучения выражается в виде

$$Q_\varepsilon = \sum_{t=0}^{k} P_t H_{L-h-t}^{l-h,m}(s_t(\varepsilon)), \quad s_t(\varepsilon) = \frac{l}{L}(m+t-\varepsilon k).$$

Теорема 1 получается из данной как частный случай при $h = 1$.

Монотонные сетки удобны тем, что позволяют просуммировать все события $[\mu X = a_{\boldsymbol{d}}]$ по одному слою с $|\boldsymbol{d}| = t + m$. В результате вид оценки оказывается не намного сложнее одномерного случая (монотонной цепочки). В унимодальных семействах, к которым мы сейчас переходим, такого уже не наблюдается, что затрудняет вычисления.

Унимодальные семейства

Определение 4. *Множество алгоритмов $A = \{a_{-D}, \ldots, a_{-1}, a_0, a_1, \ldots, a_D\}$ называется унимодальной цепочкой алгоритмов, если*

1) *если d, d' одного знака и $|d| < |d'|$, то для всех $x_i \in \mathbb{X}$ выполнено $I(a_d, x_i) \leqslant I(a_{d'}, x_i)$;*
2) *$n(a_d, \mathbb{X}) = m + |d|$ при некотором $m \geqslant 0$;*
3) *если $I(a_0, x_i) = 0$, то $I(a_d, x_i) + I(a_{d'}, x_i) \neq 2$ для всех $d > 0$, $d' < 0$, $x_i \in \mathbb{X}$.*

Алгоритм a_0 называется лучшим в цепочке.

Унимодальную цепочку можно представить как объединение двух монотонных цепочек с одинаковым лучшим алгоритмом a_0: правой ветви $\{a_0, a_1, \ldots, a_D\}$ и левой ветви $\{a_0, a_{-1}, \ldots, a_{-D}\}$. Третье условие в определении означает, что алгоритмы в левой и правой ветвях допускают ошибки на различных объектах, за исключением объектов, на которых ошибается лучший алгоритм.

Теорема 3 (Об унимодальной цепочке [2]). *Пусть A — унимодальная цепочка, метод μ является пессимистичной минимизацией эмпирического риска и выполнено условие $2k \leqslant 2D \leqslant L - m$. Тогда вероятность в результате обучения получить алгоритм a с $n(a, \mathbb{X}) = t + m$ есть*

$$P_t = \mathsf{P}\big[\mu X = a \colon n(a, \mathbb{X}) = t + m\big] = \\ = \Big(C_{L-2-2t}^{l-2} + 2\big(C_{L-t-1}^{l-1} - C_{L-2t-1}^{l-1}\big)\Big) / C_L^l; \quad (1)$$

вероятность переобучения выражается в виде

$$Q_\varepsilon = \frac{1}{C_L^l}\sum_{t=0}^{k}\Big(C_{L-2-2t}^{l-2}H_{L-2-2t}^{l-2,m}(s_t(\varepsilon)) + 2C_{L-t-1}^{l-1}H_{L-t-1}^{l-1,m}(s_t(\varepsilon)) - 2C_{L-2t-1}^{l-1}H_{L-2t-1}^{l-1,m}(s_t(\varepsilon))\Big). \quad (2)$$

Унимодальная h-мерная сетка является многомерным обобщением унимодальной цепочки.

Определение 5. *Множество алгоритмов* $A = \{a_{\boldsymbol{d}}\}_{|\boldsymbol{d}|\leqslant D}$ *называется* h*-мерной унимодальной сеткой алгоритмов, если*

1) *подмножество* A_j*, образуемое при изменении* j*-й компоненты вектора* $\boldsymbol{d}$ *при фиксированных остальных, является унимодальной цепочкой, при любом* j*;*
2) $n(a_{\boldsymbol{d}}, \mathbb{X}) = m + |\boldsymbol{d}|$ *при некотором* $m \geqslant 0$*.*

Алгоритм $a_{\mathbf{0}}$ *называется лучшим в сетке.*

Точную оценку вероятности переобучения пока удалось получить лишь в одномерном и двумерном случаях, причём она является довольно громоздкой. В произвольном h-мерном случае имеется только приближённый результат.

Теорема 4 (Унимодальная h-мерная сетка). *Пусть* A *— унимодальная* h*-мерная сетка, метод* μ *является пессимистичной минимизацией эмпирического риска и выполнено* $2hk \leqslant 2hD \leqslant L - m$*. Тогда вероятность переобучения выражается в виде* $Q_\varepsilon = \sum\limits_{t=0}^{k} Q_{\varepsilon,t}$*, где* $Q_{\varepsilon,t}$ *есть вклад алгоритмов* t*-го слоя, имеющих* $n(a, \mathbb{X}) = t + m,\ t = |\boldsymbol{d}|$*.*

$$Q_{\varepsilon,t} = \frac{1}{C_L^l}\left(\sum_{i=t}^{2t} B_i^t C_{L-2h-i}^{l-2h} H_{L-2h-i}^{l-2h,m}(s_t(\varepsilon))\right). \quad (3)$$

Здесь несколько первых членов ряда B_i^t*:*

$$\begin{aligned}
&B_0^0 = 1;\\
&B_1^1 = 2h, \quad B_2^1 = h;\\
&B_2^2 = 2h^2, \quad B_3^2 = 2h^2, \quad B_4^2 = \tfrac{1}{2}h(h+1);\\
&B_3^3 = \tfrac{2}{3}h(2h^2+1),\ B_4^3 = 2h^3,\ B_5^3 = h^2(h+1),\\
&B_6^3 = \tfrac{1}{6}h(h+1)(h+2);\\
&B_4^4 = \tfrac{2}{3}h^2(2h^2+1),\\
&B_5^4 = \tfrac{2}{3}h(2h^3 - 3h^2 + 10h - 6),\\
&B_6^4 = h^3(h+1),\\
&B_7^4 = \tfrac{1}{3}h^2(h+1)(h+2),\\
&B_8^4 = \tfrac{1}{24}h(h+1)(h+2)(h+3).
\end{aligned}$$

Выразить все коэффициенты B_i^t для произвольного t в короткой форме пока не удалось. Тем не менее, полученные результаты уже могут быть полезны: при небольших значениях длины контрольной выборки $k \leqslant 4$ этих коэффициентов достаточно для вычисления точного значения Q_ε. Численные расчёты также показали, что при «разумных» значениях параметров L, l, m, h, выписанные параметры B_i^t обеспечивают достаточно хорошее приближение точной оценки Q_ε.

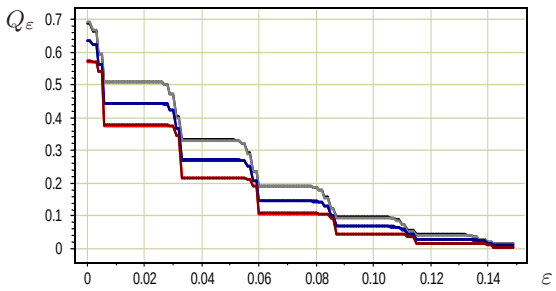


Рис. 2. Вероятность переобучения Q_ε унимодальных сеток размерностей 1, 3 и 5, и парные к ним графики для монотонных сеток соответствующих удвоенных размерностей 2, 6 и 10. В каждой паре графики почти сливаются. Параметры эксперимента: $L = 500,\ l = 460,\ m = 60,\ \varepsilon = 0{,}05$. Для унимодальных сеток использовались лишь слои с 0-го по 4-ый включительно.

Аппроксимация унимодальных сеток монотонными

Замечателен тот факт, что значение Q_ε для $2h$-мерной монотонной сетки в точности соответствует значению Q_ε для h-мерной унимодальной сетки, если положить $B_t^t = C_{2h+t-1}^t$ и $B_{t+i}^t = 0$, $i > 0$. В частности, в случае единичной длины контрольной выборки $k = |\bar{X}| = 1$ вероятность переобучения унимодальной сетки $Q_\varepsilon^{U,h}$ в точности равна вероятности переобучения монотонной сетки $Q_\varepsilon^{M,2h}$ удвоенной размерности:

$$Q_\varepsilon^{U,h} = Q_\varepsilon^{M,2h} = \frac{L-2h}{L}H_{L-2h}^{l-2h,m}(s_0(\varepsilon)) + \frac{2h}{L}H_{L-2h-1}^{l-2h,m}(s_1(\varepsilon)). \quad (4)$$

В общем случае оценки для унимодальных и монотонных сеток отличаются, но не сильно. Соответствующие графики для теоретических $Q_\varepsilon^{U,h}$ и $Q_\varepsilon^{M,2h}$ приведены на рис. 2. Из графиков также можно сделать вывод, что вклада слоёв с нулевого по четвёртый вполне достаточно для аппроксимации вероятности переобучения унимодальных сеток при данных значениях параметров L, l, m.

Зависимость вероятности переобучения от размерности сетки h

Большой интерес представляет зависимость вероятности переобучения Q_ε от размерности h. В случае единичной длины контроля, $k = 1$, имеем $H_L^{L-1,m}(z) = \frac{m}{L}\big[m - 1 \leqslant z\big] + \frac{L-m}{L}\big[m \leqslant z\big]$, что

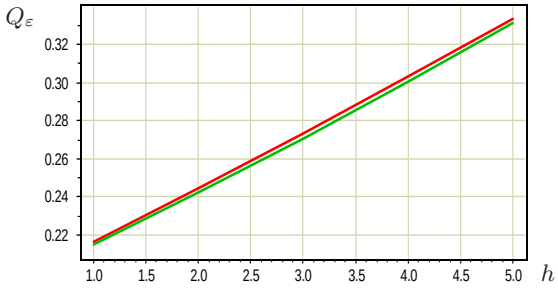


Рис. 3. Зависимость вероятности переобучения Q_ε от размерности h. Нижняя кривая — унимодальная сетка; верхняя — монотонная размерности $2h$. Параметры эксперимента: $L = 500,\ l = 460,\ m = 60,\ \varepsilon = 0{,}05$.

позволяет существенно упростить выражение (4):

$$Q_\varepsilon = \frac{m}{L}\left[\frac{L-m}{L-1} \geqslant \varepsilon\right] + \frac{h}{L}\left[\frac{L-m-1}{L-1} \geqslant \varepsilon\right] \simeq \frac{m+h}{L}.$$

Таким образом, зависимость при достаточно малых ε является линейной. В общем же случае $k > 1$ зависимость также близка к линейной в широком диапазоне параметров, рис. 3.

Сравнение монотонных сеток с реальными семействами

На основе платформы RapidMiner были получены экспериментальные зависимости Q_ε от ε для набора задач из репозитория UCI (sonar, wpbc, wdbc, breast-cancer-wisconsin, ripley) на методах классификации NaiveBayes, SVM, DecisionTree, NeuralNetwork. Экспериментальные кривые переобученности $\hat{Q}_\varepsilon(\varepsilon)$, полученные методом Монте-Карло, аппроксимировались кривыми переобученности монотонных сеток, при подборе параметра размерности h и фиксированных L, l и m. Оказалось, что чем сложнее семейство алгоритмов, или чем меньшую ошибку он даёт на обучении, тем большей размерности должно быть генерируемое им семейство алгоритмов, тем выше проходит кривая переобученности и выше размерность h аппроксимирующей монотонной сетки, см. рис. 4, 5 («более ступенчатые» кривые соответствуют монотонным сеткам, «более гладкие» — реальным семействам).

Примечателен тот факт, что на задаче breast-cancer-wisconsin метод NaiveBayes показал кривую переобученности, эквивалентную монотонной сетке размерности 1, рис. 5. Это можно интерпретировать как объективное существование в данной задаче одного информативного признака, который и был успешно выделен данным методом. DecisionTree на тех же данных не смог найти этот признак и показал бóльшую размерность, бóльшую переобученность и бóльшую ошибку на контроле.

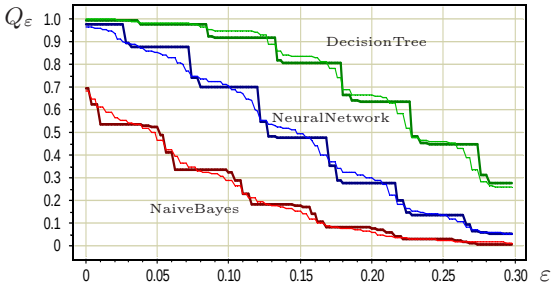


Рис. 4. Сравнение вероятности переобучения Q_ε алгоритмов NaiveBayes, NeuralNetwork, DecisionTree с аппроксимирующими монотонными сетками размерностей, соответственно, $h = 8,\ 28,\ 46$. Соответственно, ошибки на обучении/контроле: 0,28/0,32; 0,05/0,20; 0,02/0,26. Задача sonar, 20 признаков, $L = 208,\ l = 187$.

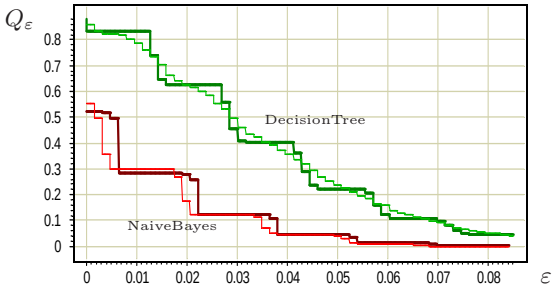


Рис. 5. Сравнение вероятностей переобучения Q_ε, полученных алгоритмами NaiveBayes и DecisionTree, с аппроксимирующими монотонными сетками размерностей $h = 1,\ 20$. Соответственно, ошибки на обучении/контроле: 0,04/0,04; 0,03/0,06. Задача breast-cancer-wisconsin, 9 признаков, $L = 700,\ l = 630$.

Выводы

Вероятность переобучения h-мерных унимодальных сеток достаточно точно аппроксимируется $2h$-мерными монотонными сетками. Причём в широком диапазоне параметров для численных оценок можно использовать лишь несколько первых слоёв сеток.

Вероятность переобучения реальных семейств алгоритмов достаточно точно аппроксимируется монотонными сетками подходящей размерности, что позволяет оценивать эффективную размерность задачи. Эффективная размерность характеризует не только саму задачу, но и используемый метод обучения, и определяется, главным образом, структурой локальной окрестности наилучшего алгоритма в семействе.

Литература

[1] *Воронцов К. В.* Комбинаторный подход к проблеме переобучения // Всеросс. конф. ММРО-14 — М.: МАКС Пресс, 2009 — С. 18–21 (в настоящем сборнике).

[2] *Воронцов К. В.* Точные оценки вероятности переобучения // Доклады РАН, 2009 (в печати).

Об унимодальности непрерывного расширения критерия Акаике*

Ветров Д. П., Кропотов Д. А., Пташко Н. О.
vetrovd@yandex.ru, dmitry.kropotov@gmail.com, ptashko@inbox.ru
Москва, ВМиК МГУ, Вычислительный Центр РАН

В работе рассматривается применение непрерывного расширения критерия Акаике (CAIC) к подбору параметров регуляризации в задаче обобщенной линейной регрессии. Значениями параметра регуляризации являются все симметричные неотрицательно определенные матрицы. Показывается, что на множестве всех таких матриц критерий CAIC является унимодальным. Получено явное условие вырожденности решающего правила (нулевого решения). Показано, что данное условие остается справедливым для семейства диагональных неотрицательно определенных матриц, а также для семейства матриц, пропорциональных единичной.

Для решения задач выбора моделей широко применяется информационный критерий Акаике [1], предлагающий подход, основанный на теории информации. Несмотря на то, что этот метод изначально был предложен для выбора из конечного числа моделей, он может быть расширен на случай бесконечного семейства моделей. Такое расширение позволяет заменить полный перебор на конечном множестве направленным поиском максимума непрерывного функционала, что существенно ускоряет процедуру выбора. В работах [2, 3] предложено подобное обобщение информационного критерия Акаике (CAIC) в применении для настройки параметров модели стационарной и нестационарной линейной регрессии. Одновременная настройка нескольких параметров модели с помощью критерия CAIC обычно проводится градиентным или покоординатным методами. В этой связи возникает вопрос о наличии у непрерывного критерия Акаике локальных максимумов. В данной работе рассматривается задача выбора модели с помощью CAIC в широком семействе всех симметричных неотрицательно определенных матриц. Доказано, что непрерывный критерий Акаике в этом семействе является унимодальной функцией. Выписано аналитическое решение для обобщенной линейной модели регрессии, в которой, согласно критерию Акаике, достигается наилучшая обобщающая способность. Также получено практически важное условие релевантности, допускающее непосредственную прямую проверку, которое позволяет отбросить заведомо неприемлемые модели. Рассматриваемое семейство матриц включает в себя важные подсемейства: семейство всех диагональных матриц и семейство матриц, пропорциональных единичной. Полученные здесь результаты для широкого семейства могут быть частично перенесены для данных подсемейств. В частности, показано, что условие релевантности остается справедливым и для рассматриваемых подсемейств.

*Работа выполнена при финансовой поддержке РФФИ, проекты №08-01-00405, №08-01-90016, №08-01-90427, №07-01-00211.

Непрерывное расширение критерия Акаике

Рассмотрим классическую задачу обобщенной линейной регрессии. Пусть $(X, \boldsymbol{t}) = \{(\boldsymbol{x}_i, t_i)\}_{i=1}^n$ — обучающая выборка, где $\boldsymbol{x}_i = (x_i^1, \ldots, x_i^d) \in \mathbb{R}^d$ — вектор наблюдаемых признаков объекта, $t_i \in \mathbb{R}$ — значение зависимой переменной. Зафиксируем некоторое множество базисных функций $\{\varphi_i(\boldsymbol{x})\}_{i=1}^m$, $\varphi_j : \mathbb{R}^d \to \mathbb{R}$. Требуется найти вектор весов $\boldsymbol{w} \in \mathbb{R}$ такой, что функция

$$y(\boldsymbol{x}) = \boldsymbol{w}^\intercal \boldsymbol{\varphi}(\boldsymbol{x}) = \sum_{j=1}^m w_j \varphi_j(\boldsymbol{x})$$

приближала бы значения переменной t на объектах обучающей выборки X. Пусть $\Phi = (\varphi_{ij})_{n\times m} = (\varphi_j(\boldsymbol{x}_i))_{n\times m}$ — матрица базисных функций, вычисленных для каждого объекта обучающей выборки. Классический подход к обучению линейной регрессии состоит в оптимизации регуляризованного правдоподобия

$$\boldsymbol{w}_{\text{MP}} = \arg\max_{\boldsymbol{w}} p(\boldsymbol{t} \mid X, \boldsymbol{w}) p(\boldsymbol{w} \mid \alpha), \qquad (1)$$

где

$$p(\boldsymbol{t} \mid X, \boldsymbol{w}) = \frac{1}{\sqrt{(2\pi)^n}\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\|\Phi\boldsymbol{w} - \boldsymbol{t}\|^2\right) \qquad (2)$$

— функция правдоподобия.

В качестве регуляризатора $p(\boldsymbol{w} \mid \alpha)$ часто рассматривается квадратичный функционал с некоторым коэффициентом регуляризации $\alpha \geqslant 0$:

$$p(\boldsymbol{w} \mid \alpha) = \left(\frac{\alpha}{2\pi}\right)^{m/2} \exp\left(-\frac{\alpha}{2}\sum_{j=1}^m w_j^2\right). \qquad (3)$$

В методе релевантных векторов RVM [4] для автоматического отбора релевантных базисных функций семейство регуляризаторов (3) предлагается расширить, и для каждого веса w_j ввести свой

коэффициент регуляризации α_j:

$$p(\boldsymbol{w}\,|\,\boldsymbol{\alpha}) = \prod_{j=1}^{m} \sqrt{\frac{\alpha_j}{2\pi}} \exp\left(-\frac{\alpha_j}{2} w_j^2\right) = \\ = \sqrt{\frac{\det(R)}{(2\pi)^m}} \exp\left(-\frac{1}{2}\boldsymbol{w}^\intercal R \boldsymbol{w}\right), \quad (4)$$

где $R = \operatorname{diag}(\alpha_1, \ldots, \alpha_m)$ — матрица регуляризации, $\alpha_j \geqslant 0$. Расширим используемое в RVM семейство регуляризаторов на случай всех (необязательно диагональных) симметричных неотрицательно определенных матриц $R = R^\intercal \succeq 0$. Такое семейство матриц позволяет не только находить релевантное подмножество базисных функций, но и одновременно выделять релевантные линейные комбинации исходных базисных функций.

Подбор матрицы регуляризации будем осуществлять, используя непрерывное расширение критерия Акаике, CAIC [2]:

$$R = \arg\max f(R);$$
$$f(R) = \log p(\boldsymbol{t}\,|\,X, \boldsymbol{w}_{\text{MP}}) - \operatorname{tr}\big(H(H{+}R)^{-1}\big). \quad (5)$$

Здесь $H = -\nabla\nabla \log p(\boldsymbol{t}\,|\,X, \boldsymbol{w}) = \sigma^{-2}\Phi^\intercal\Phi$.

Параметр σ также может быть найден путем максимизации CAIC, что приводит к следующему итеративному процессу его вычисления на основе текущего значения R:

$$(\sigma^2)^{\text{new}} = \frac{\|\boldsymbol{t} - \Phi\boldsymbol{w}_{\text{MP}}\|^2}{n - \operatorname{tr} H(H{+}R)^{-1}R(H{+}R)^{-1}}. \quad (6)$$

Решение задачи оптимизации

Обозначим через $\boldsymbol{w}_{\text{ML}}$ оценку максимального правдоподобия на выборке X. Можно показать, что при условии (2) и (4) максимизация CAIC (5) эквивалентна следующей оптимизационной задаче:

$$\begin{cases} -\frac{1}{2}\boldsymbol{w}_{\text{ML}}^\intercal H(H{+}R)^{-1}H(H{+}R)^{-1}H\boldsymbol{w}_{\text{ML}} + \\ + \boldsymbol{w}_{\text{ML}}^\intercal H(H{+}R)^{-1}H\boldsymbol{w}_{\text{ML}} - \\ \qquad - \operatorname{tr}(H(H{+}R)^{-1}) \to \max\limits_{R}; \\ R = R^\intercal \succeq 0. \end{cases} \quad (7)$$

Обозначим $\boldsymbol{v} = H^{\frac{1}{2}}\boldsymbol{w}_{\text{ML}}$,

$$A = H^{\frac{1}{2}}(H{+}R)^{-1}H^{\frac{1}{2}} = (I + H^{-\frac{1}{2}}RH^{-\frac{1}{2}})^{-1}.$$

Тогда задача (7) может быть переписана в следующем виде:

$$\begin{cases} -\frac{1}{2}\boldsymbol{v}^\intercal AA\boldsymbol{v} + \boldsymbol{v}^\intercal A\boldsymbol{v} - \operatorname{tr} A \to \max\limits_{A}; \\ A^{-1} - I \succeq 0; \\ A^\intercal = A. \end{cases} \quad (8)$$

Используя условие симметричности матрицы A, представим ее в виде $A = Q\Lambda Q^\intercal$, где $Q^\intercal = Q^{-1}$,

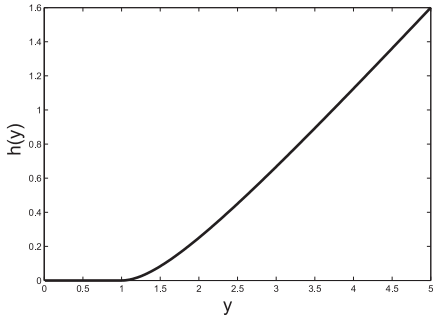


Рис. 1.

$\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_m)$ и $(\lambda_1, \ldots, \lambda_m)$ — набор собственных чисел матрицы A. Заметим, что такое разложение единственно. Обозначим $\boldsymbol{x} = Q^\intercal\boldsymbol{v}$. Тогда задача (8) переформулируется следующим образом:

$$\begin{cases} g(\boldsymbol{x}, \boldsymbol{\lambda}) = \sum\limits_{j=1}^{m}\big(-\frac{1}{2}x_j^2\lambda_j^2 + x_j^2\lambda_j - \lambda_j\big) \to \max\limits_{\boldsymbol{x},\boldsymbol{\lambda}}; \\ \sum\limits_{j=1}^{m} x_j^2 = \|\boldsymbol{v}\|^2; \\ 0 \leqslant \lambda_j \leqslant 1, \quad j = 1, \ldots, m. \end{cases} \quad (9)$$

Максимальные значения функции $g(\boldsymbol{x}, \boldsymbol{\lambda})$ при фиксированных $\boldsymbol{x}$ достигаются при следующих значениях $\boldsymbol{\lambda}$:

$$\lambda_j^*(\boldsymbol{x}) = \max\left(0, 1 - \frac{1}{x_j^2}\right). \quad (10)$$

Введем функцию $h(y)$ (см. рис. 1):

$$h(y) = \begin{cases} \frac{y}{2} + \frac{1}{2y} - 1, & y \geqslant 1; \\ 0, & 0 \leqslant y \leqslant 1. \end{cases} \quad (11)$$

Заметим, что функция $h(y)$ выпукла. Обозначим $y_j = x_j^2$. Тогда, подставив выражение (10) в систему (9), получим следующую задачу распределения ресурсов:

$$\begin{cases} \sum\limits_{j=1}^{m} h(y_j) \to \max\limits_{\boldsymbol{y}}; \\ \sum\limits_{j=1}^{m} y_j = \|\boldsymbol{v}\|^2. \end{cases} \quad (12)$$

В случае $\|\boldsymbol{v}\| \leqslant 1$ критерий $\sum\limits_{j=1}^{m} h(y_j)$ тождественно равен 0 (см. рис. 1), и компоненты y_j принимают любые значения от 0 до 1 при условии, что $\sum\limits_{j=1}^{m} y_j = \|\boldsymbol{v}\|^2$. Таким образом, все $x_j^2 < 1$ и, следовательно, все $\lambda_j^* = 0$, то есть $A = 0$, и матрица регуляризации $R^{-1} = 0$. В этом случае регуляризатор штрафует любую попытку настройки на данные, и решающее правило становится вырожденным с $\boldsymbol{w}_{\text{MP}} = \boldsymbol{0}$. На практике такой случай соответствует ситуации, когда выбранное семейство

базисных функций $\{\varphi_i(\boldsymbol{x})\}_{i=1}^m$ не позволяет восстановить зависимость скрытой переменной от признаков. Заметим, что данный результат остается справедливым в подсемействах моделей, отвечающих диагональным и скалярным (пропорциональным единичной) матрицам регуляризации, т. к. вырожденная матрица регуляризации $R^{-1} = 0$ входит в эти подсемейства. Процесс поиска наилучшей модели в этих подсемействах является итеративным, поэтому возможность отсечь заведомо неадекватное подсемейство по аналитически проверяемому условию

$$\boldsymbol{w}_{ML}^{\intercal} H \boldsymbol{w}_{ML} \leqslant 1 \tag{13}$$

позволяет значительно сократить время поиска наилучшей модели.

Если $\|\boldsymbol{v}\| > 1$, тогда, вследствие строгой выпуклости $h(y)$, максимум достигается в том случае, когда все ресурсы сосредоточены в одной из компонент y_i. Без ограничения общности выберем в качестве такой компоненты y_1, т. е. $y_1 = \|\boldsymbol{v}\|^2$, $y_2 = \cdots = y_m = 0$. Тогда первый собственный вектор матрицы A сонаправлен вектору $\boldsymbol{v}$. Остальные собственные векторы имеют нулевые собственные значения и могут быть выбраны произвольно при условии сохранения тождества $Q^{\intercal} = Q^{-1}$. Отсюда, используя определение матрицы A, получим выражение для матрицы R^{-1}:

$$\begin{aligned} R^{-1} &= H^{-\frac{1}{2}}(A^{-1} - I)^{-1} H^{-\frac{1}{2}} = \\ &= H^{-\frac{1}{2}} Q \operatorname{diag}\big(\|\boldsymbol{v}\|^2 - 1, 0, \ldots, 0\big) Q^{\intercal} H^{-\frac{1}{2}} = \\ &= \frac{\|\boldsymbol{v}\|^2 - 1}{\|\boldsymbol{v}\|^2} H^{-\frac{1}{2}} \boldsymbol{v}\boldsymbol{v}^{\intercal} H^{-\frac{1}{2}} = \\ &= \frac{\boldsymbol{w}_{\mathrm{ML}}^{\intercal} H \boldsymbol{w}_{\mathrm{ML}} - 1}{\boldsymbol{w}_{\mathrm{ML}}^{\intercal} H \boldsymbol{w}_{\mathrm{ML}}} \boldsymbol{w}_{\mathrm{ML}} \boldsymbol{w}_{\mathrm{ML}}^{\intercal}. \end{aligned}$$

Таким образом, конечное выражение для матрицы R не зависит от выбора компоненты y_i при решении задачи распределения ресурсов (12), следовательно, исходная задача (7) имеет единственную точку максимума. Итак доказана следующая

Теорема 1. *Функционал*

$$f(R) = \log p(\boldsymbol{t} \,|\, X, \boldsymbol{w}_{\mathrm{MP}}) - \operatorname{tr}\big(H(H+R)^{-1}\big)$$

является унимодальным на множестве всех матриц $R = R^{\intercal} \succeq 0$, *если* $\|H^{\frac{1}{2}} \boldsymbol{w}_{\mathrm{ML}}\| > 1$.

Выводы

Доказанная теорема позволяет использовать точку максимума R критерия CAIC для построения конечного решения задачи восстановления регрессии, гарантируя, что выбранная модель будет наилучшей среди всех допустимых моделей $R = R^{\intercal} \succeq 0$.

Кроме того, полученный результат важен для выбора модели в задаче линейной регрессии, рассмотренной в [5]. В указанной работе подбор коэффициентов регуляризации, связанных индивидуально с каждым весом, производится путем максимизации непрерывного критерия Акаике (CAIC). В этом случае критерий (5) оптимизируется на множестве всех диагональных матриц $R \succeq 0$. Унимодальность критерия на данном множестве остается открытой проблемой. Доказанная в данной работе теорема косвенно подтверждает предположение об унимодальности критерия и в этом случае. Полученное в работе условие релевантности (13) позволяет эффективно отсекать заведомо неадекватные наблюдаемым данным модели до начала итеративной настройки параметров модели. Аналитическое решение, максимизирующее непрерывный критерий Акаике, может быть использовано для построения решающих правил с лучшей обобщающей способностью.

Литература

[1] *Akaike H.* A new look at statistical model identification // IEEE Trans. Automatic Control. 1974. V. 25. P. 461–464.

[2] *Kropotov D. A., Vetrov D. P.* General solutions for information-based and bayesian approaches to model selection in linear regression and their equivalence // Pattern Recognition and Image Analysis. 2009. V. 3. P. 447–455.

[3] *Ezhova E., Mottl V., Krasotkina O.* Estimation of time-varying linear regression with unknown time-volatility via continuous generalization of the Akaike information criterion // World Academy of Sciences, Engineering and Technology. 2009. V. 51.

[4] *Tipping M. E.* The relevance vector machine // Advances Neural Information Processing Systems. 2000. V. 12. P. 652–658.

[5] *Kropotov D. A., Ptashko N. O., Vetrov D. P.* Relevant regressors selection by continuous AIC // Pattern Recognition and Image Analysis. 2009. V. 3. Pp. 456–464.

Метрики и меры опровержимости на формулах предикатной логики с вероятностями на измеримых классах моделей*

Викентьев А. А., Викентьев Р. А.
vikent@math.nsc.ru
Новосибирск, Институт математики СО РАН

Рассматриваются логические высказывания экспертов, представленные формулами языка первого порядка, формулами исчисления высказываний и формулами языка первого порядка с вероятностями. Предлагаются способы задания метрик (как мер близости) на таких высказываниях, определения мер информативности (опровержимости) и вероятности этих формул. Изучаются свойства введенных метрик и связанных с ними мер, приводятся примеры.

Введение

К настоящему времени достаточно хорошо развиты теория и методы построения решающих функций распознавания образов на основе анализа эмпирической информации, представленной в виде таблиц данных. Наряду с этим проявляется все больший интерес к анализу экспертной информации, заданной в виде вероятностных логических высказываний нескольких экспертов. В работах [1, 2, 4] поставлена задача введения меры информативности I на множестве классов эквивалентных формул (языка первого порядка), и сформулированы требования, которым должна удовлетворять эта функция, выражающая связь информативности компонент через нормированное расстояние ρ на множестве классов (с точностью до меры 0) эквивалентных формул. В настоящей работе предлагается общий подход к решению задачи с использованием измеримого класса метрических моделей и его расширений для вычисления (введения) расстояний, мер информативности и устанавливаются свойства, потверждающие требования. В дальнейшем вместо информативности будем использовать меру опровержимости, поскольку это больше отражает суть вводимой меры. Рассматривается задача введения расстояний на множестве вероятностных формул. Для предложенных расстояний справедливы специфические свойства метрики (как в [2]) и меры опровержимости. Наш подход отличается некоторой общностью (измеримые основные предикаты) и теоретической исследованием свойств предлагаемых метрик и решением задачи о введении информативности (как меры опровержимости). В решении поставленных задач играет важную роль измеримость сигнатурных предикатов в моделях и используется раздел матлогики — теория моделей, что позволяет изучать вопрос не только с точки зрения высказываний экспертов, но и общих «знаний» экспертов (гипотез), выраженных в виде аксиом.

*Работа выполнена при финансовой поддержке РФФИ, проект № 07-01-00331а, 08-07-00136а.

Свойства расстояний между формулами языка первого порядка

Пусть $\Omega = \{P_1^{m_1}, \ldots, P_t^{m_t}\}$ — фиксированная сигнатура, состоящая из конечного числа предикатных символов, которые выбираются для записи и изучения имеющихся связей между переменными в конкретной прикладной области. Случай исчисления высказываний описан в [2, 3]. Пусть задано некоторое исходное множество переменных $X = \{x_1, \ldots, x_p\}$. Обозначим через D_{x_j} конечное или измеримое множество значений переменной x_j. Пусть $A_n = \bigoplus_{j=1}^{p} D_{x_j}$ — множество с мерой n, которая равна сумма мер компонент. Далее, меру будем обозначать тем же символом, что и число элементов в конечном случае. В сигнатуру Ω для каждой переменной x_j включаем одноместный предикат P_{x_j}, выделяющий область значений переменной x_j в A_n. Пусть имеется число s экспертов и области возможных значений всех переменных. Модели (алгебраические системы в смысле А. И. Мальцева) задаются специалистами согласно утверждениям (аксиомам) и знаниям экспертов. С каждым экспертом связывается модель, согласно его знаниям интерпретируются сигнатурные предикаты. Каждый эксперт «задает» свою интерпретацию каждого предикатного символа $P_i^{m_i}$ сигнатуры Ω соответствующим отношением (с конечной мерой) на множестве A_n. В результате имеем множество конечномерных моделей $\{M_j\}_{j=1}^{s}$ (исходный класс моделей). «Высказывания» экспертов записываем в виде формул (возможно с кванторами) в многосортном языке первого порядка. Пусть F — система подмножеств множества $\bigcup_k A_n^k$, где $A_n^k = \underbrace{A_n \times \ldots \times A_n}_{k}$, образующая σ-алгебру. Нас интересуют такие подмножества S_j из F, для которых найдется формула ψ_j, отражающая «высказывание» эксперта, которая выделяет подмножество S_j (формульное подмножество). То есть S_j — это множество кортежей из $\bigcup_k A_n^k$, на которых выполняется формула ψ_j. Формула ψ_j либо отражает какое-то из высказанных «знаний» экспертов, либо

является их булевой комбинацией и навешиванием кванторов на некоторые переменные. В дальнейшем каждому рассматриваемому нами множеству S_j соответствует некоторая формула ψ_j. Пусть B — замыкание множества Ω относительно логических операций $\neg$, $\wedge$, $\vee$, $\to$ и кванторов $\forall$ и $\exists$ по переменным. Ясно, что рассматриваемое множество формул в B содержится.

Предполагается знакомство с основными понятиями из списка литературы. Далее предполагаем, что на множестве формул B задана вероятностная мера μ. Тем самым, вероятностная мера μ задана на элементах множества F. По согласованным «знаниям» экспертов мы имеем исходный класс моделей. Для полноты использования информации экспертов можно расширить измеримым образом исходный класс моделей. Учитывая одновременно новую информацию нескольких экспертов, можно уточнить каждую модель, каждую пару моделей (построить третью уточняющую их) и т. д. (тем самым получить еще не более счетное число новых моделей) и эти новые модели добавить в исходный класс. Обозначим произвольный такой класс моделей через $\mathrm{Mod}_n(\Omega)$. Введем расстояние на множестве «формул» экспертов с помощью моделей $\mathrm{Mod}_n(\Omega)$. Модели различаются носителями и интерпретациями сигнатурных предикатов, входящих в «знания» экспертов. Определим расстояние между формульными подмножествами (предикатами) в каждой модели $M_i \in \mathrm{Mod}_n(\Omega)$, как отнормированную меру их симметрической разности.

Расстоянием между формулами (формульными подмножествами) $P_k^{M_i}$ и $P_j^{M_i}$, определенными в модели M_i, назовем величину

$$\rho_{M_i}(P_k^{M_i}, P_j^{M_i}) = \mu(P_k^{M_i} \triangle P_j^{M_i}).$$

Расстояние между формулами, определенным множеством моделей $\mathrm{Mod}_n(\Omega)$, зададим как выборочное взвешенное среднее на множестве расстояний в измеримых моделях.

Расстоянием между формулами P_k и P_j, определенными на множестве $\mathrm{Mod}_n(\Omega)$, назовем величину

$$\rho_1(P_k, P_j) = \frac{\sum_{M_i \in \mathrm{Mod}_n(\Omega)} \rho_{M_i}(P_k^{M_i}, P_j^{M_i})}{|\mathrm{Mod}_n(\Omega)|}.$$

Расстоянием между предложениями φ и ψ (без свободных переменных) назовем величину

$$\rho_2(\varphi, \psi) = \frac{\left|\mathrm{Mod}((\neg\varphi\wedge\psi)\vee(\varphi\wedge\neg\psi))\right|}{|\mathrm{Mod}_n(\Omega)|}.$$

Ранее доказана теорема [4, 5], из которой следует, что предложенные расстояния действительно являются метриками. В теореме 1 обобщается этот результат и приведены некоторые свойства введенных расстояний. Далее вместо $\varphi(\bar{x})$, $\psi(\bar{x})$, $\chi(\bar{x})$ для краткости будем писать φ, ψ, χ.

Теорема 1. *Для любых формул φ, ψ, χ и любого измеримого расширения исходного класса конечномерных моделей для любого ρ_i $(i = 1, 2)$ выполняются следующие свойства.*

1) $0 \leqslant \rho_i(\varphi, \psi) \leqslant 1$;
2) $\rho_i(\varphi, \psi) = \rho_i(\psi, \varphi)$ *(симметричность);*
3) *Если* $\rho_i(\varphi, \psi) = \rho_i(\varphi_1, \psi_1)$ *и* $\rho_i(\varphi_1, \psi_1) = \rho_i(\varphi_2, \psi_2)$, *то* $\rho_i(\varphi, \psi) = \rho_i(\varphi_2, \psi_2)$;
4) $\rho_i(\varphi, \psi) \leqslant \rho_i(\varphi, \chi) + \rho_i(\chi, \psi)$;
5) $\varphi \equiv \psi \iff \rho_i(\varphi, \psi) = 0$;
6) $\varphi \equiv \neg\psi \iff \rho_i(\varphi, \psi) = 1$;
7) $\rho_i(\varphi, \psi) = 1 - \rho_i(\varphi, \neg\psi) = \rho_i(\neg\varphi, \neg\psi)$;
8) $\rho_i(\varphi, \psi) = \rho_i(\varphi \wedge \psi, \varphi \vee \psi)$;
9) $\rho_i(\varphi, \neg\varphi) = \rho_i(\varphi, \psi) + \rho_i(\psi, \neg\varphi)$.

Доказательство теоремы следует из определений, свойств меры в каждой модели, теоретико-модельных вычислений и аналогично доказательству из [4, 5].

Меры опровержимости и вероятности формул

С точки зрения важности информации, сообщенной экспертом, которому мы верим, естественно считать, что опровержимость непустой формулы тем выше, чем меньше число удовлетворяющих ей элементов (т. е. мера, определенная на подмножестве, задаваемом предикатной формулой). Поэтому введем меру опровержимости следующим образом.

Мерой опровержимости формулы $\varphi(\bar{x})$ назовем величину

$$I_i(\varphi(\bar{x})) = \rho_i(\varphi(\bar{x}), 1),$$

где 1 — тождественно истинный предикат (например, $\bar{x} = \bar{x}$). Для мер опровержимости соответствующих введенным выше расстояниям верна

Теорема 2. *Для любых формул исчисления предикатов φ и ψ от одних и тех же переменных и любого измеримого расширения исходного класса конечномерных моделей справедливы для любого ρ_i следующие свойства.*

1) $0 \leqslant I_i(\varphi) \leqslant 1$;
2) $I_i(1) = 0$;
3) $I_i(0) = 1$;
4) $I_i(\varphi) = 1 - I_i(\neg\varphi)$;
5) $I_i(\varphi) \leqslant I_i(\varphi \wedge \psi)$;
6) $I_i(\varphi) \geqslant I_i(\varphi \vee \psi)$;
7) $I_i(\varphi \wedge \psi) = \rho_i(\varphi, \psi) + I_i(\varphi \vee \psi)$;
8) *Если* $\varphi \equiv \psi$, *то* $I_i(\varphi) = I_i(\psi)$;
9) *Если* $\rho_i(\varphi, \psi) = 0$, *то* $I_i(\varphi \vee \psi) = I_i(\varphi \wedge \psi) = I_i(\varphi)$;
10) $I_i(\varphi \wedge \psi) = (I_i(\varphi) + I_i(\psi) + \rho_i(\varphi, \psi))/2$;
11) $I_i(\varphi \vee \psi) = (I_i(\varphi) + I_i(\psi) - \rho_i(\varphi, \psi))/2$.

Для доказательства теоремы используются введенные определения, свойства метрики из теоремы 1 и теоретико-модельных вычислений аналогично доказательствам [4].

На практике же эксперт обычно задает высказывание с его «вероятностью». А вопрос состоит

в изучении и согласовании таких высказываний, введении некоторой метрики на таких высказываниях. Первоочередной задачей, на наш взгляд, является определение вероятностей для формул с помощью теоретико-модельного подхода.

Вероятностью формулы $\varphi(\bar{x})$ назовем величину

$$P_i(\varphi(\bar{x})) = \rho_i(\varphi(\bar{x}), \text{false}).$$

Теорема 3. *Для любого измеримого расширения исходного класса конечномерных моделей для любых формул исчисления предикатов φ и ψ от одних и тех же переменных и для любого ρ_i $(i = 1, 2)$ справедливы следующие утверждения.*

1) $0 \leqslant P_i(\varphi) \leqslant 1$;
2) $P_i(1) = 1$;
3) $P_i(0) = 0$;
4) $P_i(\varphi) = 1 - P_i(\neg\varphi)$;
5) $P_i(\varphi) \geqslant P_i(\varphi \wedge \psi)$;
6) $P_i(\varphi) \leqslant P_i(\varphi \vee \psi)$;
7) $P_i(\varphi \wedge \psi) = P_i(\varphi \vee \psi) - \rho_i(\varphi, \psi)$;
8) *Если* $\varphi \equiv \psi$, *то* $P_i(\varphi) = P_i(\psi)$;
9) *Если* $\rho_i(\varphi, \psi) = 0$, *то* $P_i(\varphi \vee \psi) = P_i(\varphi \wedge \psi) = P_i(\varphi)$;
10) $P_i(\varphi \wedge \psi) = (P_i(\varphi) + P_i(\psi) - \rho_i(\varphi, \psi))/2$;
11) $P_i(\varphi \vee \psi) = (P_i(\varphi) + P_i(\psi) + \rho_i(\varphi, \psi))/2$.

Для доказательства теоремы 3 используются результаты теорем 1, 2 и то, что $P_i(\varphi(\bar{x})) = I_i(\neg\varphi(\bar{x}))$ и наличие аппроксимации произвольной реализации формулы в модели конечным числом измеримых множеств.

Напомним некоторые факты для логического исчисления высказываний. Рассмотрим «знания» экспертов, представленные формулами исчисления высказываний с вероятностями (вероятностные высказывания), т. е. высказывания вида «φ с вероятностью p_φ», где φ — формула исчисления высказываний. Используем сокращенную запись для таких высказываний: $B_i = \langle\varphi, p_\varphi\rangle$, $B_j = \langle\psi, p_\psi\rangle$. Пусть Σ — база знаний, состоящая из формул исчисления высказываний (в Σ содержатся все те формулы, с которыми будут работать эксперты), $S(\varphi)$ — носитель формулы φ, т. е. множество элементарных высказываний, используемых при написании формулы φ, и $S(\Sigma) = \bigcup\limits_{\varphi\in\Sigma} S(\varphi)$ — носитель совокупности знаний. Рассмотрим множество $P(S(\Sigma)) = 2^{S(\Sigma)}$ — множество всех подмножеств множества $S(\Sigma)$. Элементы множества $P(S(\Sigma))$ назовем моделями. Известно, что мощность множества $P(S(\Sigma))$ равна $2^{|S(\Sigma)|} = n$ (обозначим для простоты). Предположим, что эксперты говорят о вероятностях формул на множестве всех моделей, и каждое высказывание присутствует только с одной вероятностью (случай, когда у высказывания не одна вероятность будет рассмотрен ниже). Тогда интерпретируем вероятность, данную экспертом, следующим образом: $B = \langle\varphi, p_\varphi\rangle$ означает, что высказывание φ истинно на $n_\varphi = \lfloor n \cdot p_\varphi \rfloor$ моделях, где $n = 2^{|S(\Sigma)|}$ — число всех моделей. Пусть даны два вероятностных логических высказывания $B_i = \langle\varphi, p_\varphi\rangle$ и $B_j = \langle\psi, p_\psi\rangle$, и требуется вычислить расстояние $\rho(B_i, B_j)$ между такими высказываниями. Тогда, интерпретируя данные экспертами вероятности описанным выше способом, получаем, что высказывание φ истинно на $n_\varphi = \lfloor n \cdot \cdot p_\varphi \rfloor$ моделях, а высказывание ψ истинно на $n_\psi = \lfloor n \cdot p_\psi \rfloor$ моделях. Отметим, однако, что неизвестно на каких именно моделях каждое высказывание истинно, а также число моделей, на которых эти высказывания истинны одновременно. Рассмотрим следующую подзадачу. Пусть высказывание φ истинно на n_φ моделях, высказывание ψ истинно на n_ψ моделях и k — число моделей, на которых эти высказывания одновременно истинны. Требуется вычислить расстояние между высказываниями $B_i = \langle\varphi, p_\varphi\rangle$ и $B_j = \langle\psi, p_\psi\rangle$. Возникающие в дальнейшем расстояния обозначим через $\rho_k(B_i, B_j)$, где $k = t, t+1, \ldots, \min(n_\varphi, n_\psi)$, здесь и далее $t = \max(0, n_\varphi + n_\psi - n)$. Заметим, что значение k для каждой пары свое. Как и раньше [3,4,5], расстояние $\rho_k(B_i, B_j)$ определим как симметрическую разность, т. е.

$$\rho_k(B_i, B_j) = \frac{1}{n}(n_\varphi - k + n_\psi - k) = \frac{1}{n}(n_\varphi + n_\psi - 2k),$$

для каждого $k = t, t+1, \ldots, \min(n_\varphi, n_\psi)$. Для расстояний $\rho_k(B_i, B_j)$ справедливы все утверждения теоремы 1. (Здесь $B_i \equiv B_j \iff \varphi \equiv \psi$ и $p_\varphi = p_\psi$, это означает, что формулы φ и ψ истинны на одних и тех же моделях.)

Результаты для формул исчисления высказываний с вероятностями переносятся на формулы исчисления предикатов с вероятностями в моделях и произвольного измеримого класса измеримых моделей фиксированной сигнатуры. Остановимся на существенных моментах, отличающих этот случай от предыдущего. Рассмотрим «знания» экспертов, представленные формулами исчисления предикатов с вероятностями, т. е. высказывания вида: «$\varphi(\bar{x})$ с вероятностью p_φ», где $\varphi(\bar{x})$ — формула исчисления предикатов. Используем сокращенную запись для таких высказываний: $B_i = \langle\varphi(\bar{x}), p_\varphi\rangle$, $B_j = \langle\psi(\bar{x}), p_\psi\rangle$. Каждый эксперт «задает» свою интерпретацию каждого предикатного символа сигнатуры Ω соответствующим отношением с приписанной ему вероятностью, т. е. «знания» i-го эксперта могут быть записаны в виде: $\langle P_1^i(\bar{x}), p_1^i\rangle, \ldots, \langle P_t^i(\bar{x}), p_t^i\rangle$. $\{M_i\}_{i=1}^s$ — исходное множество моделей. Высказывания экспертов могут быть представлены любыми формулами исчисления предикатов данной сигнатуры. Для простоты зададим расстояние между предикатами P_l и P_j. Для начала определим расстояние между вероятностными интерпретациями $B_l^i = \langle P_l^i(\bar{x}), p_l^i\rangle$

и $B_j^i = \langle P_j^i(\bar{x}), p_j^i \rangle$ предикатов в каждой модели M_i. Можно предполагать, что расстояния вычисляются между предикатами одинаковой местности и от одних и тех же переменных. «Знание» $B_l^i = \langle P_l^i(x_1, \ldots, x_k), p_l^i \rangle$ означает, что предикат $P_l^i(x_1, \ldots, x_k)$ истинен на $n_{P_l^i} = \lfloor n \cdot p_l^i \rfloor$ кортежах (в смысле меры) длины k в модели M_i, где $n = \prod_{j=1}^{k} \left| D_{x_j} \right|$. Аналогично случаю исчисления высказываний пусть предикат $P_l^i(\bar{x})$ истинен на $n_{P_l^i}$ кортежах в модели M_i, предикат $P_j^i(\bar{x})$ истинен на $n_{P_j^i}$ кортежах в модели M_i и k^i — число (как мера) кортежей, на которых эти предикаты одновременно истинны, где $k^i = t, t+1, \ldots, \min(n_{P_l^i}, n_{P_j^i})$, $t = \max(0, n_{P_l^i} + n_{P_j^i} - n)$. Значение k^i для каждой пары предикатов свое. Тогда для каждого k^i расстояние $\rho_{k^i}(B_l^i, B_j^i)$ зададим формулой

$$\rho_{k^i}(B_l^i, B_j^i) = \frac{1}{n}(n_{P_l^i} + n_{P_j^i} - 2k^i).$$

Заметим, что для расстояний $\rho_{k^i}(B_l^i, B_j^i)$ справедливы утверждения теоремы 1. Применяя модельный подход [5, 4] и теорему 3, вычислим вероятности $P_{M_i}(P_l^i)$, $P_{M_i}(P_j^i)$ и расстояние $\rho_{M_i}(P_l^i, P_j^i)$ в модели M_i, затем вычислим вероятность $P_{M_i}(P_l^i \wedge P_j^i)$ и найдем меру (число в конечном случае) $k_0^i = \lfloor P_{M_i}(P_l^i \wedge P_j^i) \cdot n \rfloor$ — меру (число) кортежей, на которых предикаты одновременно истинны, вычисленное по мере в каждой модели. Далее поступаем аналогично случаю исчисления высказываний: в каждой модели M_i вычислим расстояния $\rho^i(B_l^i, B_j^i)$ с учетом вероятностей и в качестве расстояния $\rho(P_l, P_j)$ возьмем величину

$$\rho(P_l, P_j) = \frac{1}{s} \sum_{i=1}^{s} \rho^i(B_l^i, B_j^i).$$

либо (как выше в определении расстояния между формулами), используя взвешенную выпуклую сумму расстояний и измеримость класса рассматриваемых моделей. Для таких расстояний $\rho(P_l, P_j)$ справедлива теорема 1. Таким же образом вводится расстояние между двумя формулами языка первого порядка с вероятностями. Мера опровержимости формул с вероятностями (в случае исчисления высказываний и в случае языка первого порядка) вводится так же, как в пункте 2, т. е. $I(\varphi) = \rho(\varphi, 1)$, где 1 — тождественно истинная формула, а ρ — одно из введенных (по формулам выше) расстояний на вероятностных высказываниях. Для так введенных мер опровержимости справедливы свойства теоремы 2, доказанные нами для формул исчисления предикатов с использованием исходного измеримого класса моделей. Для хорошей кластеризации (таксономии) множества объектов и хорошего распознавания образов желательно, чтобы расстояние между своими представителями каждого кластера были малыми, а расстояния до представителей других кластеров по возможности большими. Это достигается заданием метрики в метрических моделях с учетом меры разнесенности изучаемых объектов: как реализаций формул в метрических моделях, так и изучаемых множеств моделей высказываний экспертов с упорядоченной частью элементарных знаний в случае исчисления высказываний каждым экспертом в отдельности.

Выводы

Исследование позволит решать вопросы, связанные с согласованием экспертных высказываний, построением адаптивных решающих функций, распознавания образов, а также при создании логических баз знаний, их кластеризации и разработки экспертных систем.

Авторы благодарят профессора Г. С. Лбова за постоянное внимание и интерес к этой работе.

Литература

[1] *Блощицын В. Я., Лбов Г. С.* О мерах информативности логических высказываний // Доклады Республиканской Школы-Семинара «Технология разработки экспертных систем». Кишинев, 1978. — С. 12–14.

[2] *Лбов Г. С., Старцева Н. Г.* Логические решающие функции и вопросы статистической устойчивости решений. — Новосибирск: Издательство Института математики 1999. — 212 с.

[3] *Vikent'ev A. A., Lbov G. S.* Setting the metric and informativeness on statements of experts // Pattern Recognition and Image Analysis. — 1997. — V. 7, № 2. — P. 175–189.

[4] *Vikentiev A. A., Koreneva L. N.* Measures of Proximity and Refutability of probabil. Logical Formulas on a small class of models // Pattern Recognition and Image Analysis. — 2004. — V. 4, № 3. — P. 452–462.

[5] *Vikent'ev A. A., Koreneva L. N.* Setting the metric and measures of informativity in predicate formulas corresponding to the statements of experts about hierarchical objects // Pattern Recognition and Image Analysis. — 2000. — V. 10, № 3.— P. 303–308.

[6] *Загоруйко Н. Г.* Прикладные методы анализа данных и знаний. — Новосибирск: Изд-во Института математики, 1999. — 270 с.

[7] *Загоруйко Н. Г., Бушуев М. В.* Меры расстояния в пространстве знаний. // Анализ данных в экспертных системах. Новосибирск, 1986. — Вып. 117: Вычислительные системы. С. 24–35.

[8] *Ершов Ю. Л., Палютин Е. А.* Математическая логика. — М.: Наука, 1991. — 320 с.

[9] *Gaifman H.* Concerning measures in the first order calculi. // Israel Math. — 1964. — V. 2, № 1. — P. 1–18.

Комбинаторный подход к проблеме переобучения*

Воронцов К. В.
vokov@forecsys.ru
Москва, Вычислительный Центр РАН

В рамках комбинаторного подхода получены точные оценки вероятности переобучения. Общая оценка формулируется в терминах порождающих и запрещающих множеств объектов, связанных с каждым алгоритмом семейства. Для некоторых семейств простой структуры порождающие и запрещающие множества выписываются в явном виде. Для общего случая предложен рекуррентный метод построения этих множеств. Упрощённый вариант рекуррентного метода приводит к верхней оценке вероятности переобучения, выраженной через профиль расслоения и связности семейства алгоритмов.

Получение точных оценок обобщающей способности является открытой проблемой в теории статистического обучения. Первые оценки VC-теории были сильно завышены [1] и в дальнейшем неоднократно уточнялись [3, 2, 5]. Однако наиболее интересные для практики случаи малых выборок и сложных семейств алгоритмов всё ещё остаются за границами применимости теории. Завышенные оценки лишь на качественном уровне описывают связь переобучения со сложностью семейства алгоритмов, и не всегда подходят для точных количественных предсказаний и управления процессом обучения. Остаётся открытым вопрос, не связано ли переобучение с какими-то более тонкими и пока не изученными явлениями.

Эксперименты [7, 8], показали, что вероятность переобучения зависит не только от сложности семейства (числа различных алгоритмов в нём), но и от степени их различности. Для получения точных оценок необходимо одновременно учитывать два эффекта: расслоение семейства по уровням ошибок и сходство алгоритмов в семействе. Пренебрежение одним из них сводит на нет все усилия, направленные на учёт второго. Методы обучения, «хорошо работающие» на практике, с необходимостью порождают расслоенные и связные семейства; иначе вероятность переобучения была бы близка к 1 уже при нескольких десятках алгоритмов в семействе.

Большинство стандартных сложностных оценок выводятся с помощью неравенства Буля (union bound), что и ведёт к их завышенности. В докладе развивается комбинаторный подход, не использующий неравенство Буля. Основная идея заключается в получении точных формул для эффективного вычисления функционалов полного скользящего контроля. Предполагается, что затем эти формулы будут использоваться для анализа причин переобучения и оптимизации гиперпараметров методов обучения. В докладе рассматривается общая техника получения таких формул для комбинаторного функционала вероятности переобучения.

*Работа поддержана РФФИ (проект №08-07-00422) и программой ОМН РАН «Алгебраические и комбинаторные методы математической кибернетики и информационные системы нового поколения».

Понятие вероятности переобучения

Пусть $\mathbb{X} = \{x_1, \ldots, x_L\}$ — конечное множество объектов, называемое *генеральной выборкой*; A — множество *алгоритмов*; $I\colon A \times \mathbb{X} \to \{0,1\}$ — бинарная функция ошибки. Если $I(a,x) = 1$, то говорят, что алгоритм a допускает ошибку на объекте x. *Вектором ошибок* алгоритма a называется L-мерный бинарный вектор $\big(I(a,x_i)\big)_{i=1}^{L}$.

Обозначим через $n(a,X)$ число ошибок алгоритма a на выборке $X \subseteq \mathbb{X}$. Частота ошибок или *эмпирический риск* алгоритма a на выборке X есть $\nu(a,X) = \frac{1}{|X|} n(a,X)$. Если $n(a,X) = 0$, то алгоритм $a \in A$ называется *корректным* на X.

Методом обучения называется отображение $\mu\colon X \mapsto a$, которое произвольной *обучающей выборке* $X \subset \mathbb{X}$ ставит в соответствие некоторый алгоритм $a \in A$. Метод обучения μ называется методом *минимизации эмпирического риска*, если

$$\mu X = \arg\min_{a \in A} n(a,X). \qquad (1)$$

Отклонением частот ошибок алгоритма a на двух выборках X и $\bar{X} = \mathbb{X}\backslash X$ называется разность частот $\delta(a,X) = \nu(a,\bar{X}) - \nu(a,X)$. *Переобученностью* метода μ на выборке X будем называть отклонение частот ошибок алгоритма $a = \mu X$:

$$\delta_\mu(X) = \nu\big(\mu X, \bar{X}\big) - \nu\big(\mu X, X\big).$$

Будем говорить, что метод μ *переобучен* при разбиении $X \sqcup \bar{X} = \mathbb{X}$, если $\delta_\mu(X) \geqslant \varepsilon$, где ε — положительный вещественный параметр.

Пусть все C_L^ℓ разбиений множества $\mathbb{X}$ на наблюдаемую обучающую выборку X длины ℓ и скрытую контрольную выборку $\bar{X}$ длины $k = L - \ell$ равновероятны. Это эквивалентно стандартному предположению о независимости наблюдений в генеральной выборке $\mathbb{X}$. Обозначим через $[\mathbb{X}]^\ell$ множество всех ℓ-элементных подмножеств выборки $\mathbb{X}$.

Задача заключается в получении точных верхних оценок *вероятности переобучения* для метода минимизации эмпирического риска μ:

$$Q_\varepsilon = \mathsf{P}\big[\delta_\mu(X) \geqslant \varepsilon\big] = \frac{1}{C_L^\ell} \sum_{X \in [\mathbb{X}]^\ell} \big[\delta_\mu(X) \geqslant \varepsilon\big]. \qquad (2)$$

Точная оценка для одного алгоритма

Пусть алгоритм a допускает m ошибок на генеральной выборке, $n(a,\mathbb{X}) = m$. Тогда вероятность допустить s ошибок на выборке X описывается гипергеометрической функцией вероятности:

$$\mathsf{P}\big[n(a,X){=}s\big] = h_L^{\ell,m}(s) = C_m^s C_{L-m}^{\ell-s}/C_L^\ell,$$

где $m \in \{0,\ldots,L\}$, аргумент s принимает целые значения от $s_0 = \max\{0, m-k\}$ до $s_1 = \min\{m,\ell\}$. При других целых m, s положим $C_m^s = h_L^{\ell,m}(s) = 0$.

Вероятность большого отклонения частот ошибок описывается гипергеометрической функцией распределения $H_L^{\ell,m}(z) = \sum\limits_{s=s_0}^{\lfloor z\rfloor} h_L^{\ell,m}(s)$:

$$\mathsf{P}\big[\delta(a,X) \geqslant \varepsilon\big] = H_L^{\ell,m}\left(\tfrac{\ell}{L}(m-\varepsilon k)\right), \qquad (3)$$

где значение $\left\lfloor \frac{\ell}{L}(m-\varepsilon k)\right\rfloor$ есть наибольшее число ошибок $n(a,X)$, при котором имеет место большое отклонение частот, $\delta(a,X) \geqslant \varepsilon$.

В пределе при $L,\ell,m \to \infty$ и $\frac{m}{L} \to p$ гипергеометрическое распределение переходит в биномиальное $h_L^{\ell,m}(s) \to C_\ell^s p^s (1-p)^s$, где p — вероятность ошибки. В теории статистического обучения при получения оценок обобщающей способности широко используется именно биномиальное распределение [3], а также завышенные верхние оценки «хвостов» биномиального распределения — неравенства Хёффдинга, Бенетта, Чернoффа и др. [4].

Гипергеометрическая оценка (3) является точной (не асимптотической, не завышенной) и не опирается на понятие «вероятности ошибки», не вполне корректное в задачах эмпирического предсказания с малым объёмом скрытой выборки.

Свойства расслоения и связности

В общем случае, когда имеется семейство алгоритмов A и метод обучения μ, справедлива оценка Вапника-Червоненкиса [1, 7]:

$$Q_\varepsilon \leqslant \sum_{m=\lceil \varepsilon k\rceil}^{L} \Delta_m H_L^{\ell,m}\left(\tfrac{\ell}{L}(m-\varepsilon k)\right) \leqslant \leqslant \Delta \max_m H_L^{\ell,m}\left(\tfrac{\ell}{L}(m-\varepsilon k)\right), \qquad (4)$$

где Δ — *коэффициент разнообразия* (shattering coefficient), равный числу различных векторов ошибок, порождаемых алгоритмами вида $a = \mu X$ по всевозможным обучающим выборкам X длины ℓ; Δ_m — коэффициент разнообразия m-го слоя — множества алгоритмов $a = \mu X$, допускающих ровно m ошибок на генеральной выборке, $n(a,\mathbb{X}) = m$.

Оценка (4) сильно завышена. Эксперименты на реальных задачах классификации [7] выявили два основных фактора завышенности — это пренебрежение свойствами расслоения и связности в семействах алгоритмов.

Свойство расслоения. В практических ситуациях множество алгоритмов *расслаивается* по уровням частоты ошибок $\nu(a,\mathbb{X})$. Основная масса алгоритмов концентрируется в области наихудшей частоты 50%, и лишь малая доля алгоритмов имеет низкий уровень ошибок. Это связано с универсальностью применяемых семейств алгоритмов. Для решения конкретной задачи с фиксированной функцией ошибки I и выборкой $\mathbb{X}$ подходит лишь малая доля алгоритмов из A. Подавляющее большинство алгоритмов «предназначены» для других задач, и в конкретной задаче практически не задействуются методом обучения μ. В то же время, понятие VC-размерности и другие распространённые меры сложности основаны на подсчёте числа всех алгоритмов в семействе (точнее, числа попарно различных векторов ошибок), без учёта вероятностей их получения. Эксперименты [7] показали, что пренебрежение эффектом расслоения может ухудшать оценку Q_ε в 10^2–10^5 раз.

Свойство связности. На практике часто применяются *связные семейства* алгоритмов, в которых для каждого алгоритма $a \in A$ найдутся другие алгоритмы $a' \in A$ такие, что векторы ошибок алгоритмов a и a' отличаются только на одном объекте [6]. Связные семейства порождаются методами классификации с непрерывной по параметрам разделяющей поверхностью. Это линейные классификаторы, машины опорных векторов с непрерывными ядрами, нейронные сети с непрерывными функциями активации, решающие деревья с пороговыми условиями ветвления, и многие другие. Чем больше в семействе схожих алгоритмов, тем сильнее завышено неравенство Буля, используемое при выводе VC-оценки (4). Эксперименты [7] показали, что пренебрежение эффектом сходства или связности может ухудшать оценку Q_ε в 10^3–10^4 раз.

Влияние расслоения и связности на вероятность переобучения было оценено в экспериментах на модельных данных [8]. Рассматривалась *цепочка алгоритмов* — последовательность векторов ошибок, в которой каждый последующий вектор отличается от предыдущего только на одном объекте. Цепочка является примером «одномерного» связного семейства. Вероятность переобучения Q_ε вычислялась методом Монте-Карло для цепочек и не-цепочек с расслоением и без расслоения. Оказалось, что если отсутствует хотя бы одно из двух свойств, то вероятность переобучения может оказаться близкой к единице уже при нескольких десятках алгоритмов в семействе. Поэтому численное точные оценки Q_ε с необходимостью должны учитывать оба свойства. До сих пор в теории статистического обучения не существовало подходов, способных дать точные оценки для цепочек с расслоением. Такой подход впервые предлагается в данной работе.

Порождающие и запрещающие подмножества объектов

Будем полагать, что все алгоритмы имеют попарно различные векторы ошибок. Тогда, очевидно, A — конечное множество.

Гипотеза 1. Для каждого алгоритма $a \in A$ можно указать индексное множество V_a, подмножества объектов $X_{av}, X'_{av} \subset \mathbb{X}$ и коэффициенты $c_{av} \in \mathbb{R}$ для каждого $v \in V_a$, такие, что при всех $X \in [\mathbb{X}]^\ell$

$$\big[\mu X{=}a\big] = \sum_{v \in V_a} c_{av} \big[X_{av} \subseteq X\big] \big[X'_{av} \subseteq \bar{X}\big]. \qquad (5)$$

Множества X_{av} будем называть *порождающими*, множества X'_{av} — *запрещающими*, множества $\mathbb{X}\backslash X_{av}\backslash X'_{av}$ — *нейтральными* для алгоритма a.

Теорема 1. *Для любых $\mathbb{X}$, A и μ существуют множества V_a, X_{av}, X'_{av}, $a \in A$, $v \in V_a$ удовлетворяющие* (5)*, причём можно полагать $c_{av} = 1$.*

Итак, гипотеза 1 верна всегда. Однако представление (5) в общем случае не единственно. Эффективно вычислимые оценки дают только такие представления, в которых множества $|V_a|$, $|X_{av}|$, $|X'_{av}|$ имеют небольшую мощность.

Введём для каждого алгоритма $a \in A$ и каждого индекса $v \in V_a$ обозначения:

$$\begin{aligned} L_{av} &= L - |X_{av}| - |X'_{av}|; \\ \ell_{av} &= \ell - |X_{av}|; \\ m_{av} &= n(a, \mathbb{X}\backslash X_{av}\backslash X'_{av}); \\ s_{av}(\varepsilon) &= \tfrac{\ell}{L}\big(n(a,\mathbb{X}) - \varepsilon k\big) - n(a, X_{av}). \end{aligned}$$

Теорема 2. *Если гипотеза 1 справедлива, то для всех $a \in A$ вероятность получить в результате обучения алгоритм a равна*

$$P(a) = \mathsf{P}\big[\mu X{=}a\big] = \sum_{v \in V_a} c_{av} P_{av}; \qquad (6)$$

$$P_{av} = \mathsf{P}\big[X_{av} \subseteq X\big]\big[X'_{av} \subseteq \bar{X}\big] = C^{\ell_{av}}_{L_{av}} / C^\ell_L; \qquad (7)$$

вероятность переобучения равна

$$Q_\varepsilon = \sum_{a \in A} \sum_{v \in V_a} c_{av} P_{av} H^{\ell_{av}, m_{av}}_{L_{av}} \left(s_{av}(\varepsilon)\right). \qquad (8)$$

Теорема 3. *Пусть гипотеза 1 справедлива, μ — метод минимизации эмпирического риска, для любой выборки $X \in [\mathbb{X}]^\ell$ множество A содержит корректный алгоритм a: $n(a, X) = 0$. Тогда вероятность переобучения принимает более простой вид:*

$$Q_\varepsilon = \sum_{a \in A} \big[n(a,\mathbb{X}) \geqslant \varepsilon k\big] P(a). \qquad (9)$$

Итак, для получения точных оценок Q_ε достаточно выписать систему порождающих и запрещающих множеств для каждого алгоритма $a \in A$.

Семейства простой структуры

Определение 1. *Множество алгоритмов $A = \{a_0, a_1, \ldots, a_D\}$ называется монотонной цепочкой алгоритмов, если $I(a_d, x_i) \leqslant I(a_{d+1}, x_i)$ для всех $x_i \in \mathbb{X}$ и $n(a_d, \mathbb{X}) = m{+}d$ при некотором $m \geqslant 0$. Алгоритм a_0 называется лучшим в цепочке.*

Монотонная цепочка алгоритмов — это простейшая модель однопараметрического связного семейства алгоритмов. Пусть для примера $\mathbb{X}$ — множество точек в $\mathbb{R}^n$; A — семейство линейных алгоритмов классификации $a(x, w) = \operatorname{sign}\langle x \cdot w\rangle$, $x \in \mathbb{R}^n$, с вектором весов $w \in \mathbb{R}^n$; функция потерь имеет вид $I(a, x) = \big[a(x, w) \neq y(x)\big]$, где $y(x)$ — истинная классификация объекта x, и выборка линейно разделима, т. е. существует $w^* \in \mathbb{R}^n$, при котором алгоритм $a(x, w^*)$ не допускает ошибок на $\mathbb{X}$. Тогда множество алгоритмов $\big\{a(x, w^* + t\delta) \colon t \geqslant 0\big\}$ образует монотонную цепочку для любого направляющего вектора $\delta \in \mathbb{R}^n$, за исключением некоторого конечного множества векторов. При этом $m = 0$.

Метод минимизации эмпирического риска μ называется *пессимистичным*, если в случаях, когда минимум $n(a, X)$ достигается на многих алгоритмах, μ выбирает алгоритм с бóльшим $n(a, \mathbb{X})$. Если же и таких алгоритмов несколько, то μ выбирает алгоритм с бóльшим порядковым номером. Пессимистичный метод на практике нереализуем, но он даёт верхние оценки Q_ε, завышенность которых невелика и связана только с неоднозначным выбором минимума эмпирического риска (1).

Теорема 4. *Пусть $A = \{a_0, \ldots, a_D\}$ — монотонная цепочка, μ — пессимистичный метод минимизации эмпирического риска, $k \leqslant D \leqslant L - m$. Тогда*

$$Q_\varepsilon = \sum_{d=0}^{k} P_d H^{\ell-1, m}_{L-d-1}\left(\tfrac{\ell}{L}(m + d - \varepsilon k)\right), \qquad (10)$$

где $P_d = C^{\ell-1}_{L-d-1} / C^\ell_L$ — вероятность получить алгоритм a_d методом μ.

Вывод этой оценки основан на явном построении порождающих и запрещающих множеств. Перенумеруем объекты так, чтобы каждый из алгоритмов a_d, $d = 1, \ldots, D$ допускал ошибку на объектах $x_1, \ldots, x_d$. Тогда справедлива гипотеза 1:

$$[\mu X{=}a_d] = [x_{d+1} \in X][x_1, \ldots, x_d \in \bar{X}].$$

Оценка (10) получается непосредственным применением теоремы 2, причём формулы (6) и (8) сильно упрощаются, т. к. $|V_a| = 1$ для всех $a \in A$.

Аналогичные точные оценки получены и для других связных семейств алгоритмов простой структуры: унимодальной цепочки, монотонных и унимодальных h-мерных сеток, окрестностей оптимального алгоритма.

Рекуррентные оценки

Перенумеруем алгоритмы $a_0, \ldots, a_D$ в порядке неубывания числа ошибок $n(a_d, \mathbb{X})$. Обозначим через μ_d пессимистичный метод, выбирающий алгоритмы только из подмножества $A_d = \{a_0, \ldots, a_d\}$. Рассмотрим переход от метода μ_{d-1} к методу μ_d при последовательном добавлении алгоритмов. Допустим, что для всех алгоритмов a_t, $t < d$, информация $\mathfrak{I}_t = \langle X_{tv}, X'_{tv}, c_{tv}\rangle_{v\in V_t}$ относительно метода μ_{d-1} уже известна. Найдём информацию $\mathfrak{I}_d$ и скорректируем информацию $\mathfrak{I}_t$, $t < d$, относительно метода μ_d. Необходимость коррекции связана с тем, что алгоритм a_d может «отбирать» разбиения у каждого из предыдущих алгоритмов.

Рассмотрим случай, когда в семействе A существует корректный алгоритм: $n(a_0, \mathbb{X}) = 0$.

Лемма 5. *Алгоритм a_d имеет только одно запрещающее множество $X'_d = \{x_i \in \mathbb{X}\colon I(a_d, x_i) = 1\}$:*

$$[\mu_d X{=}a_d] = [X'_d \subseteq \bar{X}].$$

Таким образом, $\mathfrak{I}_d = \langle \varnothing, X'_d, 1\rangle$.

Лемма 6. *Коррекция информации $\mathfrak{I}_t$, $t < d$ сводится к проверке трёх условий для каждого $v \in V_t$ такого, что $X_{tv} \cap X'_d = \varnothing$:*

а) если $X'_d \setminus X'_{tv} = \{x_i\}$ — одноэлементное множество, то x_i присоединяется к X_{tv};

б) если $|X'_d \setminus X'_{tv}| > 1$, то множество индексов V_t пополняется ещё одним элементом w и полагается $c_{tw} = -c_{tv}$, $X_{tw} = X_{tv}$, $X'_{tw} = X'_{tv} \cup X'_d$;

в) если $|X'_d \setminus X'_{tv}| = 0$, то из множества индексов V_t удаляется индекс v; соответственно, из $\mathfrak{I}_t$ удаляется вся тройка $\langle X_{tv}, X'_{tv}, c_{tv}\rangle$.

Леммы 5, 6 и теорема 3 позволяют рекуррентно вычислять вероятность переобучения Q_ε. На d-м шаге добавляется алгоритм a_d, вычисляется информация $\mathfrak{I}_d$; затем для всех $t < d$ корректируется информация $\mathfrak{I}_t$, вероятности P_{tv}, и обновляется текущая оценка Q_ε. После D-го шага она даёт точное значение вероятности переобучения.

Рекуррентная процедура может оказаться вычислительно неэффективной, если условие б) будет выполняться слишком часто. Каждый раз это приводит к добавлению ещё одного слагаемого в сумму (8). Оказывается, время вычисления можно сокращать, жертвуя точностью оценки.

Теорема 7. *Если при $c_{tv} = 1$ не выполнить условие б), то вычисленная оценка Q_ε не уменьшится.*

Рассмотрим *упрощённую рекуррентную процедуру*, в которой проверка б) не выполняется никогда. Тогда условие в) также никогда не будет выполняться. В результате каждому алгоритму a_d будет соответствовать только одна тройка $\langle X_d, X'_d, 1\rangle$, $|V_d| = 1$, и верхняя оценка вероятности переобучения будет выражаться через профиль расслоения и связности множества алгоритмов A.

Профиль расслоения и связности

Рассмотрим разбиение множества алгоритмов A на *слои* $A_m = \{a \in A\colon n(a, \mathbb{X}) = m\}$.

Связностью $q(a)$ алгоритма $a \in A$ будем называть число алгоритмов в следующем слое, допускающих ошибки на тех же объектах, что и a:

$$q(a) = \#\{a' \in A_{n(a,\mathbb{X})+1}\colon I(a,x) \leqslant I(a',x),\ x \in \mathbb{X}\}.$$

Теорема 8. *Если векторы ошибок всех алгоритмов из A попарно различны, $n(a_0, X) = 0$, и Δ_{mq} — число алгоритмов в m-м слое со связностью q, то*

$$Q_\varepsilon \leqslant \sum_{m=\lceil \varepsilon k\rceil}^{k} \sum_{q=0}^{L} \Delta_{mq} \frac{C^{\ell-q}_{L-m-q}}{C^{\ell}_{L}}. \qquad (11)$$

Предварительные эксперименты показали, что *профиль расслоения и связности* Δ_{mq} для многих семейств алгоритмов с высокой точностью является сепарабельным: $\Delta_{mq} \leqslant \Delta_m \lambda_q$, где Δ_m — коэффициент разнообразия m-го слоя, λ_q — доля алгоритмов m-го слоя, имеющих связность q. Вектор $(\Delta_q)_{q=0}^{L}$ логично называть *профилем расслоения*, а вектор $(\lambda_q)_{q=0}^{L}$ — *профилем связности* множества алгоритмов A. Профиль связности удовлетворяет условию нормировки $\sum_{q=0}^{L} \lambda_q = 1$.

В этих обозначениях слегка ухудшенная оценка (11) принимает следующий вид:

$$Q_\varepsilon \leqslant \sum_{m=\lceil \varepsilon k\rceil}^{k} \Delta_m \frac{C^{\ell}_{L-m}}{C^{\ell}_{L}} \sum_{q=0}^{L} \lambda_q \left(\frac{\ell}{L-m}\right)^q.$$

Первая часть этой оценки представляет собой в точности VC-оценку (4) для частного случая, когда $n(a_0, \mathbb{X}) = 0$. Вторая часть представляет собой «поправку на связность». Она быстро убывает с ростом q и делает оценку существенно более точной.

Литература

[1] *Вапник В. Н., Червоненкис А. Я.* Теория распознавания образов. — М.: Наука, 1974.

[2] *Herbrich R., Williamson R.* Algorithmic luckiness // *JMLR.* — 2002. — no. 3. — Pp. 175–212.

[3] *Langford J.* Quantitatively tight sample complexity bounds: Ph.D. thesis / Carnegie Mellon. — 2002.

[4] *Lugosi G.* On concentration-of-measure inequalities. — Machine Learning Summer School, Canberra. — 2003.

[5] *Philips P.* Data-dependent analysis of learning algorithms: Ph.D. thesis / ANU, Canberra. — 2005.

[6] *Sill J.* Monotonicity and connectedness in learning systems: Ph.D. thesis / CalTech. — 1998.

[7] *Vorontsov K. V.* Combinatorial probability and the tightness of generalization bounds // *Pattern Recognition and Image Analysis.* — 2008. — Vol. 18, no. 2. — Pp. 243–259.

[8] *Vorontsov K. V.* Splitting and similarity phenomena in the sets of classifiers and their effect on the probability of overfitting // *Pattern Recognition and Image Analysis.* — 2009. — Vol. 19, No. 3, Pp. 412—420.

Точечная оценка вероятности 0-события*

Гуров С. И.
sgur@cs.msu.ru
г. Москва, МГУ им. М. В. Ломоносова, факультет ВМиК

Предлагаются и обосновываются точечная оценка вероятности события, ни разу не наблюдавшегося в серии испытаний по схеме Бернулли (0-событие), для которого классические статистические методы дают на практике часто неприемлемую нулевую оценку. Даётся классификация 0-выборок по объёму.

> Гильденстерн. . . . время остановилось намертво, и поэтому выпавший в тот миг «орёл» повторяется в девяностый раз. . . *(Бросает монету. . .)*.
> *Том Стоппард.* Розенкранц и Гильденстерн мертвы.

Введение. Постановка задачи

Рассматривается оценивание неслучайной, но неизвестной вероятности p осуществления некоторого случайного события X в единичном испытании. При этом в $n > 0$ испытаниях по схеме Бернулли случайная величина числа успехов $m \in \{0, \ldots, n\}$ будет иметь биномиальное распределение $\binom{n}{m} p^m (1-p)^{n-m},\ p \in (0,1)$.

Точечная оценка $\hat{p}_{ml}$ максимального правдоподобия величины p даётся «классической формулой» (последнее равенство):

$$\hat{p}_{ml} = \arg\max_{p \in [0,1]} L(p \mid m, n) = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{m}{n}. \qquad (1)$$

Здесь $L(p \,|\, m, n) = p^m (1-p)^{n-m}$ — функция правдоподобия для биномиальной статистической модели, где $x = (x_1, \ldots, x_n),\ x_i \in \{0,1\},\ i = 1, \ldots, n$ — выборка, полученная в результате проведения n элементарных независимых экспериментов по наблюдению события X, содержащая m значений 1 и $n - m$ значений 0.

Данная оценка является является несмещенной, эффективной и состоятельной. Несмещенная функция оценки её дисперсии есть

$$\frac{m\,(n-m)}{n^3}. \qquad (2)$$

При $m = 0$ говорят, что X является *0-событием*. В этом случае формула (1) даёт нулевую точечную оценку вероятности наблюдения X, а формула (2) — нулевое оценочное значение её дисперсии. Всё это приводит к тому, что на практике оценка $\hat{p} = 0$ часто неприемлема. Такая ситуация может сложиться, например, при оценке вероятности ошибок корректного классифицирующего алгоритма.

В данной работе, являющийся развитием [4], предлагается и обосновывается ненулевая точечная оценка 0-события. Автору неизвестны публикации по данной проблеме.

Известные оценки

Частотный подход. В случае 0-события классические методы частотного подхода к решению задач математической статистики [1, 8] определяют нижнюю границу $p^-(n)$ доверительного интервала при коэффициенте доверия η как нулевую, а верхнюю $p^+(n)$ — как решение уравнения

$$I_x(1, n) = \eta.$$

Здесь $I_x(\cdot,\cdot)$ — отношение неполной B(бета)-функции к полной B-функции. Таким образом, имеем:

$$I_x(1, n) = n \int_0^x (1-t)^{n-1} dt = 1 - (1-x)^n = \eta,$$

откуда

$$p^+(n) = 1 - \sqrt[n]{1-\eta}.$$

Обычно полагают $\eta = 0.95$ или $\eta = 0.99$.

Использование $p^+(n)$ в качестве точечной оценки p, как правило, является неоправданным, дающим слишком завышенное значение вероятности: с близкой к 1 достоверностью будем иметь $p \leqslant p^+$. Однако, от точечной оценки не требуется, чтобы отклонение её значения от истинного было односторонним почти всегда.

Бейесовский подход. При использования бейесовского подхода встаёт вопрос о конкретизации априорного распределения. Будем рассматривать наиболее интересную ситуацию отсутствия результатов аналогичных экспериментов, проводимых ранее, то есть, когда использование того или иного метода восстановления априорного распределения на их основе (эмпирический бейесовский подход) невозможно. В этих случаях обычно прибегают к закону недостаточного основания Лапласа, который устанавливает, что если ничего не известно о параметре и он изменяется на конечном интервале, то в качестве априорного распределения принимают равномерное.

Априорное распределение будем, как обычно, выбирать из семейства сопряженных априорных

*Работа выполнена при финансовой поддержке РФФИ (проекты №07-01-00211-а, 08-01-00405-а) и компании Intel Corporation.

распределений относительно биномиальной статистической модели, которое составляют плотности Be_p-распределений (или распределений Бернулли)

$$\mathrm{Be}_p(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1}$$

с параметрами $a, b > 0$. Равномерное распределение $U(0,1)$ на интервале $(0,1)$ есть $\mathrm{Be}_p(1,1)$. Тогда плотность вероятности апостериорного распределения будет $\mathrm{Be}_p(1, n{+}1)$. Математическое ожидание полученного апостериорного распределения, как известно, есть

$$\mu = (n+1)\int\limits_0^1 p(1-p)^n dp \,=\, \frac{1}{n+2}, \qquad (3)$$

а медиана — $\mathrm{med} = 1 - 1/\sqrt[n]{2}$.

Бейесовскую точечную оценку определяемой величины обычно полагают равной математическому ожиданию (что отражает закон следования Лапласа) или медиане апостериорного распределения, как доставляющим минимумы среднеквадратических потерь и среднего отклонения соответственно. Таким образом, имеем две оценки:

$$\hat{p}_{B^U_\mu}(n) = \frac{1}{n+2} \quad \text{и} \quad \hat{p}_{B^U_{\mathrm{med}}}(n) = 1 - \sqrt[n]{0.5}.$$

В любом случае, ясно, что для не слишком малых n обе приведённые оценки являются завышенными, поскольку основаны на предположении о равномерном априорном распределении p, что мало согласуется с фактом 0-события.

Оценка $\hat{p}_0$

0-событие имеет место, когда в результате проведения n элементарных экспериментов по наблюдению события X получают *0-выборку* $x^0 = (0,\ldots,0)$ длины $n \geqslant 1$. Считаем, что любая другая информация о событии X отсутствует и не может быть дополнительно получена.

Далее для оценки вероятности p появления X в единичном эксперименте будет использоваться понятие коэффициента доверия $\eta \in (0,1)$. Пусть $\hat{p}$ — выбранная оценка вероятности p события X, а $\mathsf{P}(n,\hat{p})$ — вероятность некоторого события, связанного с наблюдённым 0-событием, и на основании которого делаются те или иные выводы относительно X. Будем считать значение $\mathsf{P} = \mathsf{P}(n,\hat{p})$ превосходящим выбранный коэффициент доверия:

$$\mathsf{P} \geqslant \eta. \qquad (4)$$

При этом будет иметь место непривычная зависимость $\mathsf{P}(n,\hat{p}) \to 1$ при $\hat{p} \to 0$, что связано с нулевой оценкой p по (1). Поэтому здесь коэффициент доверия выражает не степень достоверности некоторого события, а степень «уступки», на которую мы можем пойти для получения оценки, уклоняющейся от теоретически истинного, но неприемлемого для нас значения.

Оценка $\hat{p}_\eta$. При истинном значении оцениваемой вероятности p вероятность P осуществившегося 0-события есть $\mathsf{P} = (1-p)^n$. По (4) имеем:

$$p \leqslant \hat{p}_\eta = 1 - \sqrt[n]{\eta} \simeq \frac{\ln(1/\eta)}{n}.$$

Оценка $\hat{p}_r$. Мы будем говорить, что некоторое случайное событие X, наблюдаемое в единичном эксперименте по схеме Бернулли с вероятностью $p \in (0,1)$, определяет случайный процесс $\mathfrak{X}_p$ с дискретным временем, который и порождает выборку x^0 как реализацию этого процесса.

Идея получения оценки $\hat{p}_r(n)$ состоит в замене рассмотрения реализации x^0 процесса $\mathfrak{X}_p$ некоторой другой его реализацией x^1, которая содержит хотя бы одно значение 1.

Построим требуемую реализацию x^1. Рассмотрим процесс $\mathfrak{X}_q$, определяемый вероятностью q наблюдения события X в единичном эксперименте по схеме Бернулли, и x^1 — реализацию указанного процесса. Пусть объём выборки x^1 есть $N \geqslant 1$, причём $M \geqslant 1$ значений в выборке нулевые. Далее воспользуемся оценкой (1). Определим допустимые значения M и N из условия достоверности равенства $p = q$ не менее η.

Для решения поставленной задачи воспользуемся точным критерием Фишера сравнения вероятностей, лежащих в основе двух биномиальных распределений [8], сводящимся к анализу так называемых *таблиц* 2×2. В нашем случае имеем таблицу

0	n	n
M	$N-M$	N
M	$N-M+n$	$N+n$

(5)

Применение данного критерия вызвано тем, что использования общего критерия анализа 2×2 таблиц возможно лишь при достаточно больших значениях элементов таблицы, что в нашем случае заведомо не имеет места, поскольку одно из таких значений нулевое.

Вероятность $\mathsf{P} = \mathsf{P}(N, M; n)$ того, что таблица порождена одним значением вероятности, будет равна

$$\mathsf{P} = \frac{n!\,N!\,M!\,(N-M+n)!}{(N+n)!}\,\frac{1}{n!\,M!\,(N-M)!} = \\ = \frac{\binom{N}{M}}{\binom{N+n}{M}}. \qquad (6)$$

Известна асимптотика

$$\frac{\binom{n-s}{k}}{\binom{n}{k}} \sim \exp\Big\{-\frac{sk}{n} - \frac{s^2k + sk^2}{2n^2}\Big\},$$

(справедливая в нашем случае, доказательство опускаем), которая даёт

$$\mathsf{P} = \frac{\binom{(N+n)-n}{M}}{\binom{N+n}{M}} \sim \exp\Big\{-\frac{nM}{N+n}\Big(1+\frac{M+n}{N+n}\Big)\Big\}.$$

Тогда по (4) имеем:

$$-\frac{nM}{N+n}\Big(1+\frac{M+n}{N+n}\Big) \lesssim \ln\eta,$$

а, полагая по (1) $\hat{p}_r = M/N$ и считая $N \gg 1$, получим

$$n\hat{p}_r(1+\hat{p}_r) \gtrsim \ln\frac{1}{\eta}.$$

Отсюда, пренебрегая величиной $\hat{p}_r^2$, окончательно получим $\hat{p}_r \simeq \ln(1/\eta)/n = \hat{p}_\eta$.

Таким образом, обе построенные оценки практически совпадают. Данную оценку обозначим $\hat{p}_0$:

$$\hat{p}_0(n) = 1 - \sqrt[n]{\eta} \simeq \frac{\ln(1/\eta)}{n} \simeq \frac{1-\eta}{\eta n}; \qquad (7)$$

её и предлагается принимать как точечную оценку вероятности 0-события. Приведённые асимптотики (перечисленные в порядке понижения точности с завышением оценки) справедливы для практических значений η и не слишком малых n.

Несколько более грубые рассуждения, основанные на фиксации определённого значения N, приводят, как следствие $\mathsf{P} \to \max$, к $M = 1$. Тогда $\mathsf{P} = N/(N+n)$, по (4) имеем:

$$N = \Big]\frac{\eta n}{1-\eta}\Big[\qquad (8)$$

и $p \leqslant \hat{p} = M/N = (1-\eta)/(\eta n)$, что совпадает с (7).

Очевидно, для реальных значений η и $n{>}3$

$$\hat{p}_0(n) < \hat{p}_{B^U_{\mathrm{med}}}(n) < \hat{p}_{B^U_\mu}(n) < p^+(n).$$

Случай малой выборки

Предложенная оценка $\hat{p}_0$ интуитивно кажется слишком заниженной для малых значениях n. Построим оценку $\hat{p}(n)$ для этого случая.

При малых n факт 0-события не противоречит предположению о достаточно больших значениях вероятности p. Поэтому оправданным представляется следующий подход. Для некоторой N-элементной выборки с M единичными значениями найдём по (6) вероятность $\mathsf{P}(N, M; n)$ того, что таблица (5) порождена одним значением вероятности, осредним оценку M/N в соответствии с введённым вероятностным распределением на выборках, и данное среднее значение

$$\hat{p}_N(n) = \frac{\sum_{M=0}^{N} \frac{M}{N}\,\mathsf{P}(N, M; n)}{\sum_{M=0}^{N} \mathsf{P}(N, M; n)} \qquad (9)$$

будем принимать за искомую оценку для данного $N \geqslant 1$.

Элементарно показывается (см. 6), что

$$\sum_{M=0}^{N} \mathsf{P}(N, M; n) = \frac{N+n+1}{n+1},$$

и знаменатель (9) определён. Для числителя аналогично показывается, что

$$\sum_{M=0}^{N} \frac{M}{N}\,\mathsf{P}(N, M; n) = \frac{N+n+1}{(n+1)(n+2)}.$$

Отсюда

$$\hat{p}_N(n) = \frac{1}{n+2} = \hat{p}_{B_\mu}(n)$$

(и приятной неожиданностью оказывается независимость $\hat{p}_N(n)$ от N).

Полученный результат заставляет сделать вывод, что при малых значениях n обоснованной точечной оценкой вероятности 0-события является бейесовская оценка по математическому ожиданию при равномерном априорном распределении.

Интересно заметить, что формулу (3) можно проинтерпретировать как вычисление среднего значения вероятности p при распределении $\mathrm{Be}_p(1, n{+}1)$, выражающем, в рамках фидуциального фишеровского подхода [11], степень уверенности в равенстве текущего значения p действительному значению вероятности 0-события. Таким образом, (9) оказывается дискретным аналогом (3), что и объясняет совпадение оценок $\hat{p}_N(n)$ и $\hat{p}_{B^U_\mu}(n)$.

Когда какую оценку использовать?

Сразу укажем, что мы не рассматриваем случаи, когда ясны принципы (точнее, желательность более вероятного отклонения) выбора точечной оценки в данной предметной области исследования: здесь всё зависит от того, насколько важно то или иное редкое событие.

Если указанные принципы отсутствуют, то для ответа на поставленный в заголовке вопрос необходимо определиться, что понимать под «малой выборкой».

Разные авторы по разному определяют это понятие (обзор см. в [3]). Для случая 0-выборки требуется конкретные указания, при каких значениях n обосновано использовать оценку $\hat{p}_{B^U_\mu}(n)$, а при каких — $\hat{p}_0(n)$. Понятно, что абсолютно объективных критериев такого выбора существовать не может. Мы, однако, предложим указанное разбиение на основе некоторых, как представляется, разумных соображений.

Нижняя граница. Прежде всего, кажется ясным, что при совсем малых значениях n никаких статистических выводов делать вообще нельзя.

Заметим, что при $n < 4$ имеем $\hat{p}_{B_{\text{med}}}(n) < \hat{p}_{B^U_\mu}(n)$, и обратное отношение при $n \geqslant 4$. Указанное значение представляется естественной границей для отделения понятия «малая выборка» от случая недостаточности данных для любых статистических выводов. Таким образом считаем, что при $1 \leqslant n \leqslant 3$ можно только констатировать факт 0-события при данном числе испытаний.

Аналогичный вывод сделан в работе [5]: *Один из основных вопросов математической статистики: какова должна быть минимально необходимая информация для получения требуемой достоверности результата... Если подразумевать под условиями отсутствие каких-либо ограничений по точности конечного результата статистического анализа, то ответ на поставленный вопрос дал Р. Фишер [13, 12]. Минимальное число образцов не может быть меньше 4. В противном случае, неизбежно возникает систематическая ошибка (смещение). Наличие смещения — первый признак отсутствия достаточности статистики [9]. Ряд авторов подтверждал вывод Фишера.*

Также при проверке гипотезы о значении отношения ξ наблюдаемых абсолютных частот a и b на основе χ^2-критерия со статистической надёжностью 95% требуется [8]

$$\hat{\chi}^2 = \frac{(\xi a - b)^2}{\xi(a+b)} < \chi^2 = 3.841.$$

При определении равенства вероятностей, порождающих выборки как реализации случайных процессов, полагаем $\xi = 1$, что приводит к соотношению $|a-b| < 3.841$. Поскольку применение данного критерия предполагает $0 < a \leqslant b$, вместо 0-события рассматриваем противоположное ему *полное событие*, для которого $b = n$. Поэтому для того, чтобы с указанной надёжностью считать выборку с b значениями 1 другой реализацией того же случайного процесса, что и породивший 0-выборку той же длины, необходимо, чтобы значение $n - a$ не превосходило 3[1]. Таким образом, различие может быть статистически определено лишь при длине выборки $n \geqslant 4$.

Верхняя граница. Верхнюю границу для малой выборки кажется естественным установить равной N по (8) при $n = 1$. При этом (рассматриваем, напомним, 0-события) представляется более естественными значения, определяемые величинами $\eta = 0.95$ или близкими к указанной, то есть $n = 20 \div 30$. Меньшие значения n не дают возможности статистически достоверно определить совпадение вероятностей единичных событий, связанных с данными выборками. Предложенное значение n совпадает с общепринятой границей для понятия «малая выборка» в случае нормального распределения, где она принимается равной 30 [10].

[1]Интересно, что граничное значение 3 (так называемая «бонгардовская тройка») часто возникает в комбинаторных исследованиях на неслучайность событий [2, 7, 6].

Выводы

В результате предлагается следующая классификация 0-выборок по объёму с указанием точечной оценки $\hat{p}$ вероятности 0-события:

n	Тип 0-выборки, $\hat{p}$
$1, 2, 3$	никаких оценок дать нельзя
от 4 до 20–30	«малая» 0-выборка, $\hat{p} = \hat{p}_{B^U_\mu}(n)$
более 20–30	«большая» 0-выборка, $\hat{p} = \hat{p}_0$

Резкий скачок значения предложенной оценки при переходе от «малой» выборки к «большой» вызван экстремальностью самого исследуемого понятия: в малых выборках осуществление 0-события представляется вполне возможным даже при p, не обязательно близких к 0, в то время как в больших выборках оно с необходимостью означает либо крайне малую величину p, либо вообще невозможность события X.

Литература

[1] *Большев Л. Н., Смирнов Н. В.* Таблицы математической статистики. — М.: Наука, 1983. — 464 с.

[2] *Бонгард М. М.* Проблема узнавания. — М.: Наука, 1967. — 320 с.

[3] *Гуров С. И.* Оценка надёжности классифицирующих алгоритмов. — М.: Издательский отдел ф-та ВМиК МГУ, 2002. — 45 с.

[4] *Гуров С. И.* Оценка вероятности ни разу не наблюдённого события // — 2009. — Вып. 2 (в печати).

[5] *Гусев А. В., Лидский Э. А., Мироненко О. В.* Малые выборки при оценке работоспособности и надежности электронных компонентов. Часть I. // Chip news – Инженерная микроэлектроника. — 2002. — № 1. — С. 52–56.

[6] *Донской В. И., Башта А. И.* Дискретные модели принятия решений при неполной информации. — Симферополь: Таврия, 1992. — 166 с.

[7] *Закревский А. Д.* Логика распознвания. — М.: Едиториал УРСС, 2003. — 144 с.

[8] *Закс Л.* Статистическое оценивание. — М.: Статистика, 1976. — 560 с.

[9] *Леман Э.* Теория точечного оценивания. — М.: Наука, 1991. — 448 с.

[10] *Прохоров Ю. В.* Малая выборка // БСЭ. — М.: Сов. энциклопедия, 1969–1978.

[11] *Уилкс С.* Математическая статистика. — М.: Наука, 1967. — 632 с.

[12] *Фишер Р.* Статистические методы для исследователей. — М.: Гостехиздат, 1958. — 267 с.

[13] *Fisher R. A.* On the mathematical foundations of theoretical statistics // Phil. Trans. Roy. Soc., Ser. A. — 1921, V. 222. — С. 309-368.

Обобщение семейства алгоритмов вычисления оценок*

Докукин А. А.
dalex@ccas.ru
Москва, Вычислительный Центр РАН

В статье описывается обобщение семейства алгоритмов вычисления оценок, в формуле оценок которого учитываются, т.е. наказываются или поощряются, все возможные комбинации принадлежности аргумента к заданному классу и близости к нему. Изучается задача оптимизации высоты обобщенного АВО. Показывается её сводимость к ранее изученному частному случаю.

Классический АВО

Рассмотрим задачу распознавания в следующей стандартной постановке [1]. Имеются две выборки векторов из n-мерного признакового пространства: обучающая и контрольная. Для определённости предполагаем, что первая содержит m объектов: $S_1,\ldots,S_m$, вторая q: $S^1,\ldots,S^q$. Предполагается также, что множество допустимых объектов разбито на l классов K_j. Классификация объекта S задается информационным вектором $\alpha(S)=(\alpha_1(S),\ldots,\alpha_l(S))$ по системе предикатов «$S\in K_j$». Классификация объектов обучающей выборки известна. Требуется построить алгоритм, который, используя информацию об обучающей выборке $I_0=\big(S_1,\ldots,S_m,\alpha(S_1),\ldots,\alpha(S_m)\big)$, восстанавливал бы классификацию контрольной.

Таким образом, задача распознавания Z определяется совокупностью начальной информации и контрольных объектов $Z=Z(I_0,S^1,\ldots,S^q)$.

Семейство алгоритмов вычисления оценок (АВО) определяется следующим образом [2]:

1. Каждому признаку приписывается некоторый вес p_i, $i=1,\ldots,n$.

2. Выделяются некоторые подмножества множества признаков, которые называются опорными. Их совокупность обозначается Ω_A. Каждому опорному множеству $\omega\in\Omega_A$ приписывается вес

$$p(\omega)=\sum_{i\in\omega}p_i.$$

Каждому набору опорных множеств Ω_A приписывается вес $p(\Omega_A)=\sum\limits_{\omega\in\Omega_A}p(\omega)$.

3. Вводится функция близости двух объектов по опорному множеству $B_\omega(S,S')$. Если не оговорено особо, везде далее будет использоваться пороговая функция близости, т. е. два объекта $S=\{a_1,\ldots,a_n\}$ и $S'=\{b_1,\ldots,b_n\}$ будут считаться близкими, если для всех опорных признаков выполняются неравенства

$$\rho_i(a_i,b_i)\leqslant\varepsilon_i,\quad\forall i\in\omega,$$

где ρ_i — полуметрика или метрика, заданная на области значения i-го признака.

Числа ε_i, $i=1,\ldots,n$, называются порогами функции близости.

4. Свой вес $\gamma(S_j)$ приписывается каждому объекту S_j обучающей выборки $j=1,\ldots,m$.

Для каждого класса $\tilde K_j=K_j\cap\{S_1,\ldots,S_m\}$ и его дополнения $C\tilde K_j=\{S_1,\ldots,S_m\}\setminus\tilde K_j$ вводятся веса $\gamma(\tilde K_j)=\sum\limits_{S\in\tilde K_j}\gamma(S)$ и $\gamma(C\tilde K_j)=\sum\limits_{S\in C\tilde K_j}\gamma(S)$.

5. Оценка объекта за класс вычисляется по следующей формуле:

$$\Gamma_j(S^i)=x_1\frac{1}{Q_1}\Gamma_1^j(S^i)+x_0\frac{1}{Q_0}\Gamma_0^j(S^i);\qquad(1)$$

$$\Gamma_1^j(S^i)=\sum_{S\in\tilde K_j}\gamma(S)\sum_{\omega\in\Omega_A}p(\omega)B_\omega(S,S^i);\qquad(2)$$

$$\Gamma_0^j(S^i)=\sum_{S\in C\tilde K_j}\gamma(S)\sum_{\omega\in\Omega_A}p(\omega)\bar B_\omega(S,S^i).\qquad(3)$$

где $\bar B(S,S^i)=1-B(S,S^i)$; $x_1,x_0\in\{0,1\}$ — коэффициенты, отвечающие за включение компонент оценки; Q_0,Q_1 — коэффициенты нормировки, которые в общем случае зависят от j и могут использоваться для устранения разницы в мощности классов обучающей выборки. Однако далее будет рассматриваться модель $Q_1=Q_0=1$, которую принято считать классчической и использовать в теоретических построениях [2].

6. Алгоритм вычисления оценок определяется как суперпозиция $A=B\circ C$, где $B(I_0,S^1,\ldots,S^q)=\|\Gamma_{ij}\|_{q\times l}=\|\Gamma_j(S^i)\|_{q\times l}$, $C(\|\Gamma_{ij}\|_{q\times l})=\|\beta_{ij}\|_{q\times l}$, B — распознающий оператор, C — решающее правило, $\beta_{ij}\in\{0,1,\Delta\}$ — окончательный ответ о принадлежности объекта S^i классу K_j, соответственно, нет, да, не известно.

Заметим, что оценка принадлежности объекта классу в приведённом определении обладает простой интерпретацией: распознаваемый объект поощряется за близость к прецедентам своего класса и за удалённость от обучающих векторов чужих классов. Очевидно, что эти случаи отражают только два из четырех возможных типов отношения близости объектов и их принадлежности выделенному классу. Конструкция АВО может быть обобщена путем введения штрафов за близость объекта к прецедентам чужих классов и за удалённость от прецедентов своего.

*Работа выполнена при поддержке грантов РФФИ №08-01-00636-а, №08-07-00437-а, а также гранта Президента РФ, НШ–5294.2008.1.

Обобщение АВО

Запишем определение обобщённого АВО формально. Введём предикат принадлежности объекта классу — $P(S, K_j)$:

$$P(S, K_j) = \begin{cases} 1, & S \in K_j, \\ 0, & S \notin K_j. \end{cases} \tag{4}$$

Применительно к каждой тройке (S, S', K_j), $S' \in K_j$, пара $\big(P(S, K_j), B_\omega(S, S')\big)$ может принимать четыре различных значения. Сопоставим каждому из них соответственный весовой коэффициент $x_{00}, x_{01}, x_{10}, x_{11} \geqslant 0$:

		$B_\omega(S,S')$	
		0	1
$P(S,K_j)$	0	x_{00}	$-x_{01}$
	1	$-x_{10}$	x_{11}

Формулы (1)–(3) оценки объекта за класс из пункта 5 определения АВО можно переписать, добавив новые компоненты и воспользовавшись введёнными обозначениями:

$$\Gamma_j(S^i) = x_{11}\frac{1}{Q_{11}}\Gamma^j_{11}(S^i) - x_{10}\frac{1}{Q_{10}}\Gamma^j_{10}(S^i) - x_{01}\frac{1}{Q_{01}}\Gamma^j_{01}(S^i) + x_{00}\frac{1}{Q_{00}}\Gamma^j_{00}(S^i); \tag{5}$$

$$\Gamma^j_{11}(S^i) = \sum_{S\in\tilde{K}_j}\gamma(S)\sum_{\omega\in\Omega_A} p(\omega)B_\omega(S, S^i); \tag{6}$$

$$\Gamma^j_{10}(S^i) = \sum_{S\in\tilde{K}_j}\gamma(S)\sum_{\omega\in\Omega_A} p(\omega)\bar{B}_\omega(S, S^i); \tag{7}$$

$$\Gamma^j_{01}(S^i) = \sum_{S\in C\tilde{K}_j}\gamma(S)\sum_{\omega\in\Omega_A} p(\omega)B_\omega(S, S^i); \tag{8}$$

$$\Gamma^j_{00}(S^i) = \sum_{S\in C\tilde{K}_j}\gamma(S)\sum_{\omega\in\Omega_A} p(\omega)\bar{B}_\omega(S, S^i); \tag{9}$$

где $Q_{00}, Q_{01}, Q_{10}, Q_{11}$ — нормировочные коэффициенты, введенные для наглядности интерпретации оценок. Далее, аналогично классическому случаю, исследуется модель $Q_{00} = Q_{01} = Q_{10} = Q_{11} = 1$.

Воспользовавшись обозначением

$$x^\alpha = \begin{cases} x, & \alpha = 1; \\ \bar{x}, & \alpha = 0; \end{cases}$$

а также формулами (4), (6)–(9), получим более компактную запись для выражения (5).

Во-первых, запишем общую формулу для слагаемых (6)–(9):

$$\Gamma^j_{\alpha\beta}(S^i) = \sum_{S\in\{S_1,\ldots,S_m\}} P^\alpha(S, K_j)\gamma(S) \times \times \sum_{\omega\in\Omega_A} p(\omega)B^\beta_\omega(S, S^i).$$

Далее, просуммировав по всем возможным значениям $\alpha = 0, 1,\ \beta = 0, 1$, получим формулу обобщенного АВО в следующем виде:

$$\Gamma^j(S^i) = \sum_{\alpha=0,1,\beta=0,1}(-1)^{\alpha+\beta}x_{\alpha\beta} \times \times \sum_{S\in\{S_1,\ldots,S_m\}} P^\alpha(S, K_j)\gamma(S)\sum_{\omega\in\Omega_A} p(\omega)B^\beta_\omega(S, S^i).$$

Очевидно, положив $x_{11} = x_1$, $x_{10} = 0$, $x_{01} = 0$, $x_{00} = x_0$, $\alpha = \beta$, получим компактную формулу для АВО, приведенную в первом разделе:

$$\Gamma^j(S^i) = \sum_{\alpha=0,1} x_\alpha \sum_{S\in\{S_1,\ldots,S_m\}} P^\alpha(S, K_j)\gamma(S) \times \times \sum_{\omega\in\Omega_A} p(\omega)B^\alpha_\omega(S, S^i).$$

Таким образом, классический АВО действительно является частным случаем нового семейства, что автоматически приводит к сохранению основных теоретических результатов о корректности и устойчивости семейства. В частности, справедлива классическая теорема о существовании корректного полинома над семейством АВО [2] и её усиленный вариант [4].

Высота обобщённого АВО

Ранее был предложен подход к обучению АВО на основе максимизации высоты отдельных слагаемых распознающего полинома [5]. Под высотой оператора Ψ в данном случае понимается разность между максимумом оценки правильной пары (объект, класс), т. е. пары, в которой объект принадлежит классу, и минимумом — неправильной [3]:

$$h(\Psi, \Phi) = \min_{(S,j)\in M_1}\Psi_j(S) - \max_{(S',v)\in M_0}\Psi_v(S'), \tag{10}$$

где Φ — множество объектов, на котором вычисляется высота; $\Psi_j(S)$ — оценка объекта S за j-й класс;

$$M_1 = \{(S, j)\colon S \in \Phi,\ S \in K_j\};$$
$$M_0 = \{(S, j)\colon S \in \Phi,\ S \notin K_j\}.$$

На основе упомянутого подхода были разработаны различные алгоритмы обучения АВО [5, 6], при построении которых рассматривался случай $(x_{11}, x_{10}, x_{01}, x_{00}) = (1, 0, 0, 0)$. Покажем, что в задаче максимизации высоты АВО выбор коэффициентов не важен.

Обозначим $\Theta^j(S)$ произвольную линейную комбинацию вида

$$\Theta^j(S) = \sum_{u,v\in\{0,1\}}\vartheta_{uv}(-1)^{u+v}\Gamma^j_{uv}(S),$$

где $\vartheta_{uv} \geqslant 0,\ u, v \in \{0, 1\},\ \vartheta_{uv}$ не равны нулю одновременно. Кроме того, ϑ_{uv} не зависят от j.

Справедлива следующая теорема.

Теорема 1. *Пусть в исходной задаче распознавания классы не пересекаются. Тогда для произвольного объекта S справедливо равенство*

$$\Theta^j(S) = a\Gamma^j_{11}(S) + b\gamma(\tilde{K}_j) + c\gamma(C\tilde{K}_j) + d(S),$$

где $a, b, c \in \mathbb{R}$ — константы для данного ABO, $a > 0$, $d(S) \in \mathbb{R}$ не зависит от j.

Доказательство. Из формул оценок (6), (7) получаем равенство

$$\begin{aligned}\Gamma^j_{11}(S^i) + \Gamma^j_{10}(S^i) = \\ = \sum_{S\in\tilde{K}_j}\gamma(S)\sum_{\omega\in\Omega_A} p(\omega)\left(B_\omega(S,S^i) + \bar{B}_\omega(S,S^i)\right) = \\ = p(\Omega_A)\gamma(\tilde{K}_j). \quad (11)\end{aligned}$$

Далее, из (8), (9) следует, что

$$\Gamma^j_{01}(S^i) + \Gamma^j_{00}(S^i) = p(\Omega_A)\gamma(C\tilde{K}_j). \quad (12)$$

Эти формулы справедливы, поскольку порог ε фиксирован. Кроме того, поскольку классы не пересекаются, из (6), (8) получаем выражение:

$$\begin{aligned}\Gamma^j_{11}(S^i) + \Gamma^j_{01}(S^i) = \\ = \sum_{S\in\{S_1,\dots,S_m\}}\gamma(S)\sum_{\omega\in\Omega_A} p(\omega)B_\omega(S,S^i) = \\ = \sum_{i=1}^{l}\Gamma^i_{11}(S^i). \quad (13)\end{aligned}$$

Обозначим $Y(S^i) = \sum\limits_{i=1}^{l}\Gamma^i_{11}(S^i)$.

Из (11)–(13) получаем:

$$\begin{aligned}\Theta^j(S^i) = \Gamma^j_{11}(S^i)\sum_{u,v\in[0,1]}\vartheta_{uv} - \\ - p(\Omega_A)\gamma(\tilde{K}_j)\vartheta_{10} + \\ + p(\Omega_A)\gamma(C\tilde{K}_j)(\vartheta_{00} - \vartheta_{01}) - \\ - Y(S^i)\vartheta_{00}.\end{aligned}$$

Таким образом, $a = \vartheta_{11} + \vartheta_{10} + \vartheta_{01} + \vartheta_{00} > 0$ по условию. Коэффициенты b и c равны, соответственно, $b = -\vartheta_{10}p(\Omega_A)$, $c = (\vartheta_{00} - \vartheta_{01})p(\Omega_A)$, и зависят только от конфигурации опорных множеств. Наконец, $d(S) = -Y(S)\vartheta_{00}$ зависит от обучающей выборки в целом, а не от отдельных классов.

Теорема доказана.

Доказанная теорема имеет важное следствие. Для краткости рассмотрим случай, когда высота алгоритма вычисляется на множестве из одного объекта. Этот случай, кроме того, актуален при построении корректного полинома [2, 3, 5].

Пусть $\{S\}$ — множество из одного объекта, на котором требуется вычислить высоту, причём для определенности $S \in K_j$. Формула (10) высоты оператора Θ с помощью теоремы 1 приводится к виду

$$\begin{aligned}h(\Theta,\{S\}) = \Theta^j(S) - \max_{i\neq j}\Theta^i(S) = \\ = a\Gamma^j_{11}(S) + b\gamma(\tilde{K}_j) + c\gamma(C\tilde{K}_j) - \\ - \max_{i\neq j}\left(a\Gamma^i_{11}(S) + b\gamma(\tilde{K}_j) + c\gamma(C\tilde{K}_j)\right). \quad (14)\end{aligned}$$

Таким образом, все изученные алгоритмы максимизации высоты [5, 6] для набора коэффициентов (1, 0, 0, 0) в том же виде пригодны для оптимизации произвольной их линейной комбинации Θ. При этом сложность вычисления высоты в заданной точке практически не меняется — требуется заранее один раз рассчитать веса всех классов и их дополнений (в общем случае также требуется вычислить значение $Y(S)$ для всех контрольных объектов). Более того, при равном количестве объектов в классах формула (14) ещё более упрощается

$$h(\Theta,\{S\}) = a\left(\Gamma^j_{11}(S) - \max_{i\neq j}\Gamma^i_{11}(S)\right),$$

что позволяет вычислять высоту всех возможных комбинаций аналитически, получив оптимальные пороги для случая $(1, 0, 0, 0)$.

Литература

[1] *Журавлёв Ю. И.* Корректные алгебры над множеством некорректных (эвристических) алгоритмов I // Кибернетика. — 1977. — № 4. — С. 14–21.

[2] *Журавлёв Ю. И.* Корректные алгебры над множеством некорректных (эвристических) алгоритмов II // Кибернетика. — 1977. — № 6. — С. 21–27.

[3] *Журавлёв Ю. И., Исаев И. В.* Построение алгоритмов распознавания, корректных для заданной контрольной выборки // Ж. вычисл. матем. и матем. физ. — 1979. — Т. 19, № 3. — С. 729–738.

[4] *Докукин А. А.* О построении в алгебраическом замыкании одного алгоритма распознавания // Ж. вычисл. матем. и матем. физ. — 2001. — Т. 41, № 12. — С. 1873–1877.

[5] *Докукин А. А.* Об одном методе построения оптимального алгоритма вычисления оценок // Ж. вычисл. матем. и матем. физ. — 2006. — Т. 46, № 4. — С. 754–760.

[6] *Докукин А. А.* О построении выборок для тестирования приближённых методов оптимизации алгоритмов вычисления оценок // Ж. вычисл. матем. и матем. физ. — 2006. — Т. 46, № 5. — С. 978–983.

Разрешимость и регулярность алгоритмов нечёткой разметки точечных конфигураций

Дорофеев Н. Ю.
cmc.nick@gmail.com
Москва, МГУ им. М. В. Ломоносова

В работе рассматриваются вопросы разрешимости и регулярности задач разметки точечных конфигураций, а также обосновывается переход от исходной дискретной задачи к непрерывной.

Рассматривается задача построения обучаемых алгоритмов классификации точек в плоских конфигурациях [1]. Конечной плоской конфигурацией называют вектор $\bar{A}^d = \left(S^1, \ldots, S^d\right) = \left(\left(t^1, v^1\right), \ldots, \left(t^d, v^d\right)\right)$. Обычно считается, что либо $t^1 < \ldots < t^d$, либо $t^1 \leqslant \ldots \leqslant t^d$, причем при $t^i = t^{i+1}$ выполнено $v^i = v^{i+1}$, $i = 1, \ldots, d-1$. Множество всех d-точечных плоских конфигураций обозначают K^d, при этом $K = \bigcup\limits_{d=1}^{\infty} K^d$ есть множество всех конфигураций.

Требуется каждой точке конфигурации сопоставить элемент из некоторого множества, называемого словарём разметки. В работе [2] были рассмотрены вопросы разрешимости и регулярности задач выделения трендов. Указанные задачи были сведены к задаче классификации точек в плоских точечных конфигурациях. Там же были даны определения и получены критерии локальной разрешимости и локальной регулярности этих задач.

Следует отметить, что полученные критерии опирались на понятия сдвиг-эквивалентности конфигураций и сдвиг-эквивалентности окрестностей, в которых конфигурации и окрестности считались эквивалентными, если они совпадали с точностью до сдвига. Неразрешимыми в этом случае считались задачи, в которых эквивалентным конфигурациям и окрестностям соответствовали различные разметки.

В настоящей работе изучаются вопросы локальной разрешимости и локальной регулярности задач разметки точечных конфигураций в условиях модифицированного отношения эквивалентности окрестностей. Отметим, что в случае указанного выше отношения сдвиг-эквивалентности достаточно минимального изменения любой точки окрестности, например, как следствие шумов в данных, и отношение сдвиг-эквивалентности теряется.

Разумной представляется идея потребовать, чтобы «похожим» окрестностям сопоставлялись похожие метки. Для этого будем предполагать, что чем дальше находится точка от опорной, тем меньшее влияние она может оказать на разметку опорной точки. Таким образом, изменение точек на периферии окрестности должно минимально сказываться на разметке опорной точки. И наоборот: даже несильное изменение точек окрестности, близких к опорной, должно существенно влиять на разметку опорной точки.

Окрестности, разметки, метрики

Определение 1. *Подконфигурацией $P_{\bar{S}^d}$ конфигурации $\bar{S}^d$ называется любое подмножество точек из $\bar{S}^d$: $P_{\bar{S}^d} \subseteq \bar{S}^d$. Подконфигурация $\tilde{P}_{\bar{S}^d}$ конфигурации $\bar{S}^d$ называется связанной, если $\tilde{P}_{\bar{S}^d} = \left(S^{i_1}, S^{i_1+1}, \ldots, S^{i_2} \subseteq \bar{S}^d\right)$, где $1 \leqslant i_1 \leqslant i_2 \leqslant d$.*

Определение 2. *Нечёткой окрестностью точки S^i называется тройка $O_{\bar{S}^d}(S^i) = \left\{S^i, \tilde{P}_{\bar{S}^d}, g(t)\right\}$, $1 \leqslant i_1 \leqslant i \leqslant i_2 \leqslant d$, содержащая саму точку S^i, некоторую связную подконфигурацию $\tilde{P}_{\bar{S}^d}$ конфигурации $\bar{S}^d$, включающую эту точку, а также унимодальную функцию принадлежности $g(t)$, определяющую степень принадлежности точек подконфигурации к окрестности.*

Точку S^i будем далее называть *опорной точкой* окрестности $O_{\bar{S}^d}\left(S^i\right)$. В дальнейшем термин окрестность будет употребляться именно в смысле последнего определения.

Если окрестность содержит одну лишь опорную точку, будем говорить о *тривиальной окрестности*. Заметим, что в работе [2] в силу специфики задачи дано иное определение тривиальности.

Определение 3. *На конфигурации задана система окрестностей $O_{\bar{S}^d}$, если каждой точке поставлена в соответствие некоторая её окрестность. Система окрестностей Ω задана на K, если для каждой конфигурации из K задана система окрестностей. Пусть далее на K задана нетривиальная система окрестностей Ω, т. е. система, не содержащая окрестностей, которые были бы тривиальными.*

В случаях, когда нас будут интересовать элементы пространства Ω и опорная точка нам будет не столь важна, окрестность будет обозначаться одной буквой, например, $O \in \Omega$.

Зафиксируем *словарь разметки* — конечное множество меток $M = \{\mu^1, \ldots, \mu^d\}$, $m \geqslant 1$. Множество $M_\Delta = M \cup \Delta$, $\Delta \notin M$, где Δ — специальная метка, интерпретируемая как «не размечено», будем называть *расширенным множеством меток* или *расширенным словарем разметки*. При фиксированном множестве меток M и, соответственно,

расширенном множестве меток M_Δ *разметкой* длины d называется любая последовательность $\bar{\mu}^d = \{\mu^1, \ldots, \mu^d\}$ длины $d \geqslant 1$, если $\mu^i \in M$, или *частичной разметкой* длины d, если $\mu^i \in M_\Delta$.

Введём понятие *пространства меток*. Элементами этого пространства будут являться выпуклые комбинации меток:

$$M^* = \left\{ \mu = \sum_{i=1}^{m} a_i \mu_i \;\middle|\; a_i \geqslant 0,\ \sum_{i=1}^{m} a_i = 1 \right\},$$

где $\mu_i \in M$, $i = 1, \ldots, m$.

По аналогии с расширенным словарём разметки можно задать *расширенное пространство меток*: $M^*_\Delta = M^* \cup \Delta$.

Определение 4. *Алгоритмом разметки окрестностей A будем называть отображение $\Omega \to M^*_\Delta$, которое ставит в соответствие всякой нетривиальной окрестности некоторую метку из M^*_Δ.*

С помощью алгоритма разметки окрестностей можно определить действие *алгоритма разметки конфигураций* $\bar{A}$ как последовательное применение алгоритма A к окрестности каждой точки конфигурации:

$$\bar{A}\left(\bar{S}^d\right) = \left(A(O_{\bar{S}^d}(S^1)), \ldots, A(O_{\bar{S}^d}(S^d))\right)$$

Следующим шагом является введение оценок сходства на пространстве нечётких окрестностей и пространстве меток. Это позволит более гибко, чем с помощью отношения сдвиг-эквивалентности, оценивать близость окрестностей.

Определение 5. *Метрика ρ на пространстве Ω называется внутренней если для любых двух точек O_1 и O_2 и $\varepsilon > 0$ найдётся их ε-середина, то есть точка O_ε такая, что*

$$\rho\left(O_1, O_\varepsilon\right) < \tfrac{1}{2}\rho\left(O_1, O_2\right) + \varepsilon \quad и$$
$$\rho\left(O_2, O_\varepsilon\right) < \tfrac{1}{2}\rho\left(O_1, O_2\right) + \varepsilon.$$

Будем считать, что на множестве окрестностей задана некоторая внутренняя метрика $\rho\left(O_1, O_2\right)$: $\Omega \times \Omega \to [0, 1]$, то есть $\rho(a, b) + \rho(b, c) \leqslant \rho(a, c)$, $\rho(a, b) = \rho(b, a)$, $\rho(a, b) = 0 \Leftrightarrow a \cong b$, где $\cong$ обозначает сдвиг-эквивалентность.

В пространстве меток M^*_Δ зададим метрику $l\left(\mu^1, \mu^2\right)$. Дополнительно потребуем, чтобы расстояние от всякой метки до Δ равнялось 0.

Пусть $\alpha > 0$ — параметр задачи, устанавливающий соответствие между близостью окрестностей и близостью меток. Этот параметр указывает, насколько должны быть близки метки, в зависимости от близости окрестностей. При больших α точкам, окрестности которых значительно близки, можно будет сопоставить достаточно различные метки. При малых α наоборот потребуется даже мало похожим окрестностям ставить близкие метки. В предельном случае при $\alpha \to \infty$ мы, фактически, вернёмся к случаю сдвиг-эквивалентности: сдвиг-эквивалентным окрестностям будут ставиться одинаковые метки, в то время как на разметку остальных окрестностей никакие ограничения, связанные с локальностью, не накладываются.

Аксиомы разметки

Из содержательных соображений следует, что не все разметки каждой конкретной конфигурации являются «разумными» (подходящими). Для описания требований к подходящим разметкам, вводятся системы аксиом (правил) разметки. Отметим, что из аксиом вытекают ограничения на семейства алгоритмов разметки.

Определение 6. *Аксиомами (или правилами) разметки называется набор эффективно вычислимых предикатов $\Pi = \{\pi_1, \ldots, \pi_k\}$:*

$$\pi_i \colon \Omega \times M^* \to \{0, 1\}.$$

Тот же символ Π будет использоваться и для обозначения конъюнкции предикатов π_i:

$$\Pi = \bigcap_{i=1}^{k} \pi_i, \quad \Pi \colon \Omega \times M^* \to \{0, 1\}.$$

Для удобства записи в дальнейшем определим действие аксиом на пространстве конфигураций K:

$$\Pi\left(\bar{S}^d, \bar{\mu}^d\right) = \bigcap_{i=1}^{d} \Pi\left(S^i, \mu^i\right).$$

Определение 7. *Систему аксиом разметки $\Pi = \{\pi_i\}$ будем называть α-Ω-локальной, если $\forall O_1 \in \Omega,\ \forall \mu_1 \in M^*\colon\ \Pi(O_1, \mu_1) = 1\ \ \exists \varepsilon \in (0, \frac{1}{\alpha})\colon$ $\forall O_2 \in \Omega\colon\ \ \rho(O_1, O_2) \leqslant \varepsilon\ \ \exists \mu_2 \in M^*\colon$*

$$\Pi\left(O_2, \mu_2\right) = 1 \ \ и \ \ l\left(\mu_1, \mu_2\right) \leqslant \alpha\varepsilon.$$

Далее будем считать, что зафиксирована некоторая α-Ω-локальная система аксиом $\Pi = \{\pi_i\}$.

Определение 8. *Метка $\mu \in M^*_\Delta$ называется допустимой для окрестности $O \in \Omega$, если выполнено равенство $\Pi\left(O, \mu\right) = 1$.*

Частичная разметка $\bar{\mu}^d$ конфигурации $\bar{S}^d$ называется допустимой, если выполнено равенство $\Pi\left(\bar{S}^d, \bar{\mu}^d\right) = 1$.

Определение 9. *Будем называть подходящим алгоритм разметки A, для которого верно $\Pi\left(O, \gamma\right) = 1\ \ \forall O \in \Omega$, где $\gamma = A(O)$.*

Разрешимость

Пусть задан набор прецедентов:

$$H = \left\{ \left(\bar{S}_i^{d_i}, \bar{\mu}_i^{d_i}\right) \;\middle|\; \bar{S}_i^{d_i} \in K^{d_i},\ \bar{\mu}_i^{d_i} \in M^{d_i},\ i = 1, \ldots, q \right\}.$$

Определение 10. *Множеством окрестностей O_H набора прецедентов H в смысле системы окрестностей Ω называется множество окрестностей всех точек всех конфигураций набора H.*

Множеством частично размеченных окрестностей O_H^μ набора прецедентов называется множество:

$$O_H^\mu = \left\{(O,\mu) \;\middle|\; O \in O_H,\ \mu \in M_\Delta\right\}.$$

Задача, в которой близким в смысле метрики ρ окрестностям присваиваются далёкие в смысле метрики l метки, противоречит здравому смыслу, поэтому логичным будет исключить возможность возникновения подобных ситуаций. Эта идея формализуется в понятии противоречивости.

Нас будут интересовать локальные алгоритмы, то есть алгоритмы, использующие информацию о некоторой окрестности входных данных. Классическое определение локального алгоритма вычисления информации дано в [4]. Определение, данное ниже, учитывает специфику исследуемой задачи.

Определение 11. *Алгоритм разметки A будем называть α-Ω-H-локальным тогда и только тогда, когда для любых окрестностей O_1 и O_2 выполнено условие:*

$$\rho(O_1,O_2) = r \;\Rightarrow\; l(\mu_1,\mu_2) \leqslant \alpha r$$

и для всех $O' \in \Omega$ верно

$$\rho(O,O') \geqslant \tfrac{1}{\alpha}\ \forall O' \in \Omega \;\Leftrightarrow\; A(O') = \Delta,$$

где $\mu_1 = A(O_1)$ и $\mu_2 = A(O_2)$.

Нетрудно заметить, аналогию между первым требованием к локальному алгоритму и понятием модуля непрерывности функции. Более того, если рассматривать метку как функцию от окрестности, требование локальности алгоритма окажется аналогом условия Липшица с константой α.

Определение 12. *Задача Z называется α-Ω-локально разрешимой тогда и только тогда, когда для неё существует подходящий α-Ω-H-локальный алгоритм.*

Теорема 1. *Задача Z α-Ω-локально разрешима тогда и только тогда, когда существует непрерывная функция $f\colon \Omega \to M_\Delta^*$ такая, что $\Pi(O, f(O))$ для всякой окрестности из Ω.*

Определение 13. *Набор прецедентов H будем называть α-Ω-локально противоречивым, если для всех i выполнено $\Pi\left(\bar{S}_i^{d_i}, \bar{\mu}_i^{d_i}\right) = 1$, но существуют $O_1 \in O_H$, $O_2 \in O_H$ и $\varepsilon > 0$ такие, что $\rho(O_1,O_2) \leqslant \varepsilon$, но $l(\mu_1,\mu_2) > \alpha\varepsilon$.*

Заметим, что противоречивый набор в смысле [2] будет и α-Ω-локально противоречивым. В то же время существуют α-Ω-локально противоречивые наборы, не являющиеся противоречивыми в смысле [2].

Теорема 2. *Задача Z α-Ω-локально разрешима тогда и только тогда, когда набор прецедентов H не является α-Ω-локально противоречивым.*

Регулярность

Определение 14. *Задача разметки конечных плоских конфигураций Z называется α-Ω-локально регулярной тогда и только тогда, когда Z α-Ω-локально разрешима для любых допустимых частичных разметок всех конфигураций из H.*

α-Ω-локально разрешимая задача Z является α-Ω-локально регулярной тогда и только тогда, когда для произвольных $O_1, O_2 \in O_H$ из правил разметки вытекает выполнение условия:

$$\rho(O_1,O_2) = r \;\Rightarrow\; \rho(\mu_1,\mu_2) \leqslant \alpha r,$$

где μ_1 и μ_2 — произвольные допустимые метки для O_1 и O_2 соответственно.

Выводы

Основной целью работы было исследование поведения классов разрешимых и регулярных задач разметки точечных конфигураций при переходе от сравнения окрестностей с помощью отношения сдвиг-эквивалентности к более гибкому отношению. Основным принципом была выбрана идея о том, что близким окрестностям должны соответствовать близкие метки.

Для проверки этого условия были введены пространство меток и метрика на этом пространстве, а также метрика на пространстве окрестностей. В результате этого соответствующим образом изменились основные понятия, в том числе и важнейшее понятие противоречивого набора. Был определён класс α-Ω-H-локальных алгоритмов. Были сформулированы и доказаны разрешимости и регулярности задач разметки конечных точечных конфигураций. В результате классы разрешимых и регулярных задач сузились. При этом всякая разрешимая задача в полученных определениях является разрешимой и относительно отношения сдвиг эквивалентности. И наоборот, задача неразрешимая относительно отношения сдвиг эквивалентности будет неразрешимой и во введённых терминах. Таким образом, полученный результат является своего рода обобщением уже имеющихся в данной области работ.

Кроме того, было доказано, что в полученной задаче разрешимость эквивалентна существованию

непрерывной функции из пространства окрестностей в пространство меток. Фактически был осуществлён переход от дискретной задачи разметки конфигураций к непрерывной.

Литература

[1] *Чехович Ю. В.* Об обучаемых алгоритмах выделения трендов // Искусственный интеллект (научно-теоретический журнал НАН Украины). — 2002. — № 2. — С. 298–305.

[2] *Рудаков К. В., Чехович Ю. В.* Алгебраический подход к проблеме синтеза обучаемых алгоритмов выделения трендов // Доклады Академии наук. — 2003. — Т. 388, № 1. — С. 33–36.

[3] *Рудаков К. В.* Об алгебраической теории универсальных и локальных ограничений для задач классификации // Распознавание, классификация, прогноз. — М.: Наука, 1989. — С. 176–201.

[4] *Журавлёв Ю. И.* Избранные научные труды. — М.: Магистр, 1998. — 420 с.

Алгебраические замыкания обобщённой модели алгоритмов распознавания, основанных на вычислении оценок*

Дьяконов А. Г.
djakonov@mail.ru
МГУ им. М. В. Ломоносова

Представлены последние достижения в алгебраическом подходе к решению задач распознавания, связанные с полным описанием алгебраических замыканий классической модели алгоритмов вычисления оценок (ABO), а также её обобщения. Обобщённая модель вычисления оценок ориентирована на задачи с произвольным способом задания объектов, в том числе «непризнаковым». Описание замыканий позволило окончательно решить многие проблемы: оценки степени корректного замыкания, геометрических критериев корректности, обоснования понятия корректность и т. д.

Введение

В начале 1970-х годов Ю. И. Журавлёвым была описана модель алгоритмов вычисления оценок (ABO) для решения задач распознавания [1]. Каждый алгоритм модели являлся суперпозицией распознающего оператора и решающего правила. Распознающий оператор строил матрицу оценок принадлежности контрольных объектов (которые алгоритм должен классифицировать) классам. Решающее правило на основе этой матрицы классифицировало контрольные объекты. В ABO были отражены многие эвристические методы, применявшиеся при решении прикладных задач, и модель стала весьма универсальным языком описания алгоритмов распознавания.

В конце 1970-х годов Ю. И. Журавлёвым был предложен алгебраический подход к решению задач распознавания [2]: корректный алгоритм (который не делает ошибок на контрольный выборке) было предложено искать в виде алгебраического выражения над некорректными (эвристическими) алгоритмами. Над распознающими операторами были введены операции сложения, умножения на константу и умножения как операции над матрицами оценок, которые они порождают (умножение проводилось поэлементно). При фиксированном решающем правиле эти операции индуцировали операции над алгоритмами распознавания. Множество всех полиномов степени не выше k над алгоритмами некоторой модели было названо алгебраическим замыканием k-й степени этой модели (при $k = 1$ — линейным замыканием). Было доказано, что существует и в явном виде выписывается корректный алгоритм-полином над некорректными ABO [2]. Основное достижение алгебраического подхода — строгое доказательство «чисто алгебраическими» методами того, что в теории распознавания возможно построение «хороших» алгоритмов на базе «плохих». Недостатки одного конкретного алгоритма устраняются достоинствами остальных. Эту идею используют многие современные и практически эффективные конструкции: комитеты, бустинг (boosting), усреднение по ансамблю, модульное обучение (modular learning), области компетентности, нелинейные монотонные корректирующие операции. Позже в рамках алгебраического подхода К. В. Рудакову удалось создать язык для описания и исследований задач преобразования информации — теорию локальных и универсальных ограничений [3].

Исследования модели ABO, в основном, были сконцентрированы на оптимизации алгоритмов и алгебраических выражений над ними [4, 5, 6]. Строились корректные алгоритмы из алгебраических замыканий, но не изучалось множество всех алгоритмов, т.е. упускался важный объект исследования: алгебраические замыкания. До настоящего времени они не были даже достаточно детально описаны, чтобы анализировать, есть ли в них «хорошие» алгоритмы, как их синтезировать, и т. д. Технику анализа алгебраических конструкций в рамках классической теории удалось построить В. Л. Матросову для решения фундаментальной задачи о теоретическом обосновании надёжности алгоритмических построений в алгебраическом подходе [7, 8]. С её помощью удалось решить несколько важных проблем, связанных с корректностью алгебраических замыканий (возможностью получить операторами замыкания произвольную матрицу оценок в рассматриваемой задаче распознавания). Эти результаты, в определённом смысле, поставили гораздо больше проблем, причём более сложных, постановка и решение которых описаны ниже.

Решённые проблемы

За последние годы получены следующие результаты в рамках алгебраического похода к распознаванию.

1. Предложена удобная техника для описания и исследования алгебраических замыканий конечных степеней — теория систем эквивалентностей. Каждой задаче распознавания сопостовляется система эквивалентностей, в терминах которой просто описываются алгебраические замыкания моде-

*Работа выполнена при финансовой поддержке РФФИ, проекты № 08-07-00305-а и № 08-01-00636-а.

ли АВО. Переходу к алгебраическому замыканию фиксированной степени соответствуют специальные преобразования эквивалентностей.

2. Получены новые критерии корректности алгебраического замыкания конечной степени и критерии разрешимости задач алгоритмами из этого замыкания. Критерии позволяют описывать разделяющие поверхности алгоритмов из алгебраических замыканий в пространстве контрольных объектов и простые семейства корректных полиномов.

3. Получена неулучшаемая в общем случае оценка степени корректного алгебраического замыкания модели АВО. Для важных частных случаев получены пониженные оценки.

4. Исследованы пополнения линейного замыкания полиномов конечной степени над АВО операциями нормировки и деления. Нормировка рассмотрена при различных способах её определения: по сумме, максимуму, отрезку. Получены формулы для размерности соответствующих замыканий и критерии корректности. Показана представимость замыканий в стандартной форме (в алгебраических выражениях нет вложения новых операций).

5. Исследовано понятие корректности модели относительно семейства решающих правил. Получены общие критерии реализации классификации с помощью алгоритма из линейного замыкания произвольной модели с решающим правилом, на которое накладываются требования частичной монотонности. Рассмотрены специальные требования: построчная монотонность, постолбцовая монотонность, реализация классификаций из заданного множества. Показано, что в некоторых случаях следует использовать «неклассическое» определение корректности.

Ниже формально описаны некоторые из полученных результатов.

Задача распознавания

Множество допустимых объектов M содержит объединение множеств K_j, $j = 1, \ldots, l$, называемых *классами*. Каждому объекту $S \in M$ соответствует бинарный *вектор классификации* $\tilde{\alpha}(S) = (\alpha_j(S))_{j=1}^{l}$, где $\alpha_j(S)$ — значение предиката «$S \in K_j$», $j = 1, \ldots, l$. Для каждой пары $(S^t, S_i) \in M \times M$ можно вычислить значения функций

$$\rho_\Omega(S^t, S_i) \in \mathrm{E}_Z,\ \Omega \in \Omega_Z,$$

где Ω_Z — конечное множество параметров, E_Z — частично упорядоченное множество. Функция ρ_Ω играет ту же роль, что и «расстояние» в классической постановке [4]. Задача распознавания состоит в том, чтобы построить алгоритм A, который по набору $\tilde{S}^m = \{S^t\}_{t=1}^m$ *эталонных* объектов с известными векторами $\{\tilde{\alpha}(S^t)\}_{t=1}^m$ для набора $\tilde{S}_q = \{S_t\}_{t=1}^q$ *контрольных* объектов строит их векторы классификаций — *классифицирует* (распознаёт).

Обобщённая модель АВО

Алгоритм обобщённой модели вычисления оценок является суперпозицией распознающего оператора B и решающего правила C: $A = B \circ C$. Для объектов из $\tilde{S}_q$ оператор B получает матрицу $\Gamma[B] = \|\Gamma_{ij}[B]\|_{q\times l}$, ij-й элемент которой — *оценка принадлежности объекта* S_i *к классу* K_j:

$$\Gamma_{ij}[B] = \sum_{a,b=0,0}^{1,1} x_{abj} \sum_{\Omega\in\Omega_A} \sum_{S^t\in\tilde{K}_j^a} w^t w(\Omega) \mathrm{B}_\Omega^{\tilde{e},b}(S^t, S_i),$$

где $w^t \in \mathbb{Q}^+ = \{x \in \mathbb{Q} \mid x \geqslant 0\}$ при $t \in \{1, \ldots, m\}$ — *вес* t*-го объекта*, $w(\Omega) \in \mathbb{Q}^+$ при $\Omega \in \Omega_A$ — *вес учёта* Ω*-й близости*, $\tilde{e} \in \mathrm{E}_Z$, $\mathrm{B}_\Omega^{\tilde{e},b}(S^t, S_i)$ — *функция близости* такая, что

$$\mathrm{B}_\Omega^{\tilde{e},1}(S^t, S_i) = 1 - \mathrm{B}_\Omega^{\tilde{e},0}(S^t, S_i) = \begin{cases} 1, & \rho_\Omega(S^t, S_i) \leqslant \tilde{e}, \\ 0, & \rho_\Omega(S^t, S_i) \nleqslant \tilde{e}, \end{cases}$$

$$\tilde{K}_j^a = \begin{cases} \tilde{S}^m \cap K_j, & a = 1, \\ \tilde{S}^m \backslash K_j, & a = 0. \end{cases}$$

В этой работе для простоты считаем, что $\Omega_A = \Omega_Z$ и $x_{ab} = x_{abj} \in \{0, (-1)^{a+b}\}$ для всех $j = 1, \ldots, l$, $(a, b) \in \{0, 1\}^2$ (*модель без нормировок*).

Простейшее решающее правило — *пороговое правило* C_{c_1,c_2}: $C_{c_1,c_2}(\|\Gamma_{ij}[B]\|_{q\times l}) = \|\alpha_{ij}\|_{q\times l}$,

$$\alpha_{ij} = \begin{cases} 1, & \Gamma_{ij} \geqslant c_2, \\ \Delta, & c_1 \leqslant \Gamma_{ij} < c_2, \\ 0, & \Gamma_{ij} < c_1, \end{cases}$$

$(i, j) \in QL = \{1, \ldots, q\} \times \{1, \ldots, l\}$, $c_1 \in \mathbb{Q}$, $c_2 \in \mathbb{Q}$, $c_1 \leqslant c_2$. Символ Δ обозначает отказ от классификации.

Алгебра над алгоритмами

Следующие операции над распознающими операторами индуцируют соответствующие операции над алгоритмами при фиксированном решающем правиле: $\Gamma[B_1 + B_2] = \Gamma[B_1] + \Gamma[B_2]$,

$$\Gamma[cB] = c\Gamma[B], \quad \Gamma[B_1 \cdot B_2] = \Gamma[B_1] \circ \Gamma[B_2].$$

Символ $\circ$ используем для обозначения адамарова (поэлементного) умножения. Для множества B^* всех операторов обобщённой модели АВО вводим понятия: *линейного замыкания*

$$\mathbf{L}(B^*) = \{c_1 B_1 + \ldots + c_r B_r \mid r \in \mathbb{N}, \\ c_1, \ldots, c_r \in \mathbb{Q},\ B_1, \ldots, B_r \in B^*\},$$

алгебраического замыкания k*-й степени*

$$\mathbf{U}^k(B^*) = \mathbf{L}\big(\{B_1 \cdot \ldots \cdot B_s \mid \\ B_1, \ldots, B_s \in B^*, 1 \leqslant s \leqslant k\}\big),$$

алгебраического замыкания $\mathbf{U}(B^*) = \bigcup\limits_{k=1}^{\infty} \mathbf{U}^k(B^*)$.

Операторы разметки

Определение 1. *Оператор* $D_{\Omega,t,a,\tilde{e}}$:

$$\Gamma_{ij}[D_{\Omega,t,a,\tilde{e}}] = I[\alpha_j(S^t) = a] \cdot I[\rho_\Omega(S^t, S_i) = \tilde{e}],$$

$(i,j) \in QL$, *где* $a \in \{0,1\}$, $\tilde{e} \in \mathrm{E}_Z$; $I[\pi] = 1$, *если выполняется условие* π, $I[\pi] = 0$ — *в противном случае, называется оператором разметки.*

Теорема 1. *Справедливо равенство* $\mathbf{U}^k(B^*) = \mathbf{U}^k(D^*)$, *где* $D^* = \{D_{\Omega,t,a,\tilde{e}}\}_{\Omega,t,a,\tilde{e}}$ — *множество операторов разметки.*

Пусть $\sim^Q_{\Omega,t}$ и $\sim^L_{\Omega,t}$ — эквивалентности соответственно на множествах $\{1,\ldots,q\}$ и $\{1,\ldots,l\}$ такие, что

$$\begin{aligned} i \sim^Q_{\Omega,t} j \quad &\Leftrightarrow \quad \rho_\Omega(S^t, S_i) = \rho_\Omega(S^t, S_j), \\ i \sim^L_{\Omega,t} j \quad &\Leftrightarrow \quad \alpha_i(S^t) = \alpha_j(S^t), \end{aligned}$$

$\mathbf{\Theta}(\sim)$ — множество характеристических векторов классов эквивалентности $\sim$. Пусть $\Gamma^{*,k}$ — множество матриц оценок операторов из $\mathbf{U}^k(B^*)$. Тогда множество $\Gamma^{*,1}$ является линейным замыканием множества

$$\bigcup_{t=1}^{m} \bigcup_{\Omega\in\Omega_A} \Big(\big\{\tilde{\theta}\tilde{\alpha}^\intercal \mid \tilde{\theta} \in \mathbf{\Theta}(\sim^Q_{\Omega,t}),\ \tilde{\alpha} \in \mathbf{\Theta}(\sim^L_{\Omega,t})\big\}\Big).$$

Таким образом, матрицы операторов разметки являются характеристическими векторами (ql-мерными, записанными в матричной форме) классов эквивалентностей $\sim_{\Omega,t}$:

$$(i_1, j_1) \sim_{\Omega,t} (i_2, j_2) \Leftrightarrow (i_1 \sim^Q_{\Omega,t} i_2)\ \&\ (j_1 \sim^L_{\Omega,t} j_2).$$

Аналогичное «удобное» представление можно получить и для множества $\Gamma^{*,k}$ (это линейное замыкание множества характеристических векторов классов эквивалентностей некоторой системы эквивалентностей).

Определение 2. *Задача распознавания называется регулярной, если выполняются условия регулярности:*

1) $\big|\{\tilde{K}^1_1, \ldots, \tilde{K}^1_l\}\big| = l$;
2) $\big|\{(\rho_\Omega(S^t, S_i))_{\Omega\in\Omega_A,\, t\in\{1,\ldots,m\}}\}_{i=1}^{q}\big| = q$;
3) $\tilde{S}^m \cap \tilde{S}_q = \varnothing$.

Определение 3. *Модель распознающих операторов называется корректной (относительно задачи распознавания), если для любой матрицы из* $\mathbb{Q}^{q\times l}$ *найдётся оператор модели, который её порождает.*

С помощью техники операторов разметки и систем эквивалентностей классический результат алгебраического подхода о корректности алгебраического замыкания в классе регулярных задач [2] получается в виде критерия.

Теорема 2. *Модель* $\mathbf{U}(B^*)$ *корректна тогда и только тогда, когда выполнены первое и второе условия регулярности. При*

$$k \geqslant \max\big[\lfloor\log_2 q\rfloor + \lfloor\log_2 l\rfloor,\ 1\big]$$

справедливо равенство $\mathbf{U}(B^*) = \mathbf{U}^k(B^*)$, *где* $\lfloor x\rfloor$ — *наибольшее целое число, не превосходящее числа* x. *При* $1 \leqslant k < \lfloor\log_2 q\rfloor + \lfloor\log_2 l\rfloor$ *существует (регулярная) задача распознавания с* q *контрольными объектами и* l *классами, в которой* $\mathbf{U}(B^*) \neq \mathbf{U}^k(B^*)$.

Критерии корректности

Пусть оператор разметки $B_{\Omega,t,(i,j)}$ такой, что его матрица оценок равна $\tilde{\theta}\tilde{\alpha}^\intercal$: $\theta_i = 1$, $\alpha_j = 1$, $\tilde{\theta} = (\theta_1, \ldots, \theta_q)^\intercal \in \mathbf{\Theta}(\sim^Q_{\Omega,t})$, $\tilde{\alpha} = (\alpha_1, \ldots, \alpha_l)^\intercal \in \in \mathbf{\Theta}(\sim^L_{\Omega,t})$. Пусть

$$B_{(i,j)} = \sum_{t=1}^{m} \sum_{\Omega\in\Omega_A} B_{\Omega,t,(i,j)}.$$

Считаем, что $F(B) = a_k B^k + \cdots + a_1 B + a_0 B_E$ при $F(x) = a_k x^k + \cdots + a_1 x + a_0$, где $\Gamma[B_E] = \|1\|_{q\times l}$ (очевидно, что $B_E \in \mathbf{L}(B^*)$). Следующая теорема обобщает классический результат Ю. И. Журавлёва [2] о реализации с помощью алгоритмов вида

$$\sum_{(a,b)\in QL} c_{(a,b)} B^k_{(a,b)}$$

произвольной классификации в регулярной задаче.

Теорема 3. *Справедливо равенство*

$$\mathbf{L}\big(\{F_k(B_{(i,j)})\}_{(i,j)\in QL}\big) = \mathbf{U}^k(B^*),$$

где $F_k(x) = a_k f_k(x) + \cdots + a_1 f_1(x) + a_0$, $a_k > 0$, $a_{k-1}, \ldots, a_0 \geqslant 0$, $f_r(x) = x(x-1)\cdot\ldots\cdot(x-r+1)$.

Зафиксируем на множестве QL лексикографический порядок и в соответствии с ним выпишем матрицу

$$H_k = \|h^k_{(i,j),(a,b)}\|_{ql\times ql},$$

в которой $\big((i,j),(a,b)\big)$-й элемент равен числу $F_k\big(\Gamma_{ij}[B_{(a,b)}]\big)$.

Теорема 4. *Модель* $\mathbf{U}^k(B^*)$ *корректна тогда и только тогда, когда* $\det(H_k) \neq 0$.

Теорема 5. *Классификация* $\|\alpha_{ij}\|_{q\times l} \in \{0,1\}^{q\times l}$ *реализуется алгоритмом из замыкания* $\mathbf{U}^k(B^*)$ *с пороговым решающим правилом тогда и только тогда, когда совместна система неравенств относительно переменных* $x_{11}, \ldots, x_{ql}$:

$$\begin{cases} h^k_{(i,j),(1,1)} x_{11} + \cdots + h^k_{(i,j),(q,l)} x_{ql} > 0, & (i,j) \in I_1, \\ h^k_{(i,j),(1,1)} x_{11} + \cdots + h^k_{(i,j),(q,l)} x_{ql} < 0, & (i,j) \in I_0, \end{cases}$$

где $I_\alpha = \{(i,j) \in QL \mid \alpha_{ij} = \alpha\}$, $\alpha \in \{0,1\}$.

Конус CUT_n описывает всевозможные l_1-метрики на n-точечных множествах пространств $\mathbb{R}^r$, $r \in \mathbb{N}$ [9].

Теорема 6. *Пусть в регулярной задаче распознавания $ql > 1$, тогда функция $\rho^k\colon QL \times QL \to [0,\,1]$,*

$$\rho^k\big((i,j),(a,b)\big) = \\ = 1 - \left(\frac{\big|\{(\Omega,t) \mid (i,j) \sim_{\Omega,t} (a,b)\}\big|}{m|\Omega_A|}\right)^k,$$

является метрикой из конуса CUT_{ql}, для которой

$$\mathbf{U}^k(B^*) = \mathbf{L}\big(\{P_{(a,b)}\}_{(a,b)\in QL}\big)$$

при $\Gamma_{ij}[P_{(a,b)}] = \rho^k\big((i,j),(a,b)\big)$, $(i,j) \in QL$, и для которой определитель $ql \times ql$-матрицы метрики отличен от нуля тогда и только тогда, когда замыкание $\mathbf{U}^k(B^*)$ корректно.

В алгебраическом подходе к распознаванию традиционно исследовались модели при одном фиксированном (достаточно простом) решающем правиле. Возникает идея, что критерии корректности можно ослабить, если выбирать решающее правило из «достаточно богатого семейства». Как показывают результаты, описанные ниже, критерии корректности ослабить не удаётся при накладывании на множество решающих правил «естественного» ограничения построчной монотонности: если $C\big(\|\Gamma_{ij}\|_{q\times l}\big) = \|\alpha_{ij}\|_{q\times l}$, то для любого $i \in \{1,\ldots,q\}$ и любой пары $(j_1,j_2) \in \{1,\ldots,l\}^2$ справедливо

$$\Gamma_{ij_1} \leqslant \Gamma_{ij_2} \Rightarrow \alpha_{ij_1} \leqslant \alpha_{ij_2}.$$

Определение 4. *Модель R^* распознающих операторов называется корректной относительно множества решающих правил C^*, если*

$$\forall \mathrm{A} \in \{0,1\}^{q\times l}\ \ \exists B \in R^*\ \ \exists C \in C^*\colon\ C\big(\Gamma[B]\big) = \mathrm{A}.$$

Теорема 7. *Модель $\mathbf{U}^k(B^*)$ корректна относительно семейства всех построчно монотонных решающих правил тогда и только тогда, когда она корректна или $l = 1$.*

При введении некоторых дополнительных ограничений, например постолбцовой монотонности, критерии корректности совпадают.

Пополнение другими операциями

При пополнении линейного замыкания новой операцией часто достаточно рассматривать алгебраические выражения, в которых нет вложений новой операции (т. н. замыкание в стандартной форме). Например, если рассмотреть операцию взятия обратной по Адамару матрицы

$$\mathrm{D}\big(\|h_{ij}\|_{q\times l}\big) = \big\|h_{ij}^{-1}\big\|_{q\times l},$$

определяемую на матрицах без нулевых элементов, и операцию над операторами, индуцированную ею, то корректный алгоритм представляется в виде

$$\sum_j \sum_i c_j \mathrm{D}(B_1 + c_i B_2) \circ C,$$

где $\{B_1, B_2\} \subseteq \mathbf{L}(B^*)$. Просто выписываются формулы для размерности и критерии корректности пополненных замыканий при пополнении делением или нормировкой, при различных способах её определения: по максимуму, по сумме и по отрезку. Например, при нормировке по сумме каждый элемент матрицы делится на сумму элементов строки, его содержащей.

Литература

[1] *Журавлев Ю. И., Никифоров В. В.* Алгоритмы распознавания, основанные на вычислении оценок // Кибернетика. — 1971. — № 3. — С. 1–11.

[2] *Журавлев Ю. И.* Корректные алгоритмы над множествами некорректных (эвристических) алгоритмов. II // Кибернетика. — 1977. — № 6. — С. 21–27.

[3] *Рудаков К. В.* Об алгебраической теории универсальных и локальных ограничений для задач классификации // Распознавание, классификация, прогноз. Математические методы и их применение. М.: Наука, 1989. — Вып. 1. — С. 176–201.

[4] *Журавлев Ю. И.* Об алгебраическом подходе к решению задач распознавания или классификации // Пробл. кибернетики. — 1978. — Вып. 33. — С. 5–68.

[5] *Журавлёв Ю. И, Рязанов В. В., Сенько О. В.* «РАСПОЗНАВАНИЕ». Математические методы. Программная система. Практические применения. — М.: Фазис, 2006. — 176 с.

[6] *Воронцов К. В.* О проблемно-ориентированной оптимизации базисов задач распознавания // Ж. вычисл. матем. и матем. физ. — 1998. — Т.38, №5. — С. 870–880.

[7] *Матросов В. Л.* Синтез оптимальных алгоритмов в алгебраических замыканиях моделей алгоритмов распознавания // Распознавание, классификация, прогноз. Математические методы и их применение. М.: Наука, 1989. — Вып. 1. — С. 149–176.

[8] *Матросов В. Л.* О критериях полноты модели алгоритмов вычисления оценок и её алгебраических замыканий // Докл. АН СССР. — 1981. — Т.258, №4. — С. 791–796.

[9] *Деза М. М., Лоран М.* Геометрия разрезов и метрик. — М.: МЦНМО, 2001. — 736 с.

Критерии корректности алгебраического замыкания модели ABO в задачах с порядковыми признаками*

Иофина Г. В.
giofina@mail.ru
Московский физико-технический институт

Определены условия регулярности задачи распознавания с порядковыми признаками, а также получены условия корректности алгебраического замыкания модели ABO. Рассмотрен случай, когда в матрицах расстояний на признаках элементы не возрастают по строкам и не убывают по столбцам.

В работе рассматривается задача распознавания Z в стандартной постановке [1]. Даны эталонные (обучающие) объекты $\tilde{S}^m = \{S^i\}_{i=1}^m$, каждый из которых принадлежит одному или нескольким классам $K_1, \ldots, K_l$. Для контрольных объектов $\tilde{S}_q = \{S_i\}_{i=1}^q$ надо определить принадлежность к этим классам. Каждый объект S задан своим признаковым описанием $\boldsymbol{a}(S) = \big(a_1(S), \ldots, a_n(S)\big)$, где $a_j(S) \in \tilde{N}_j = \{0, \ldots, N_j - 1\}$.

Для решения задачи будет использоваться алгоритм вычисления оценок (ABO), введённый в 1970-х годах Ю. И. Журавлёвым для обобщения основных подходов к решению задач распознавания образов [1, 4].

Алгоритм вычисления оценок $A = B \circ C$ представляется в виде суперпозиции распознающего оператора B и решающего правила C.

Оператор B по описаниям контрольных объектов и обучающей информации вычисляет матрицу оценок $\Gamma[B] = \big(\Gamma_{ij}[B]\big)_{q\times l}$, где

$$\Gamma_{ij}[B] = x_1 \sum_{\Omega\in\Omega_A} \sum_{S^t\in\tilde{S}^m\cap K_j} w^t w(\Omega) B_\Omega^{\tilde{\varepsilon}}(S^t, S_i) + \\ + x_0 \sum_{\Omega\in\Omega_A} \sum_{S^t\in\tilde{S}^m\setminus K_j} w^t w(\Omega)\big(1 - B_\Omega^{\tilde{\varepsilon}}(S^t, S_i)\big),$$

где $x_0, x_1 \in \{0, 1\}$, Ω_A — множество подмножеств $\{1, \ldots, n\}$, называемое системой опорных множеств, $w^t \in \mathbb{Q}_+$ — вес t-го объекта, $t = 1, \ldots, m$, $w(\Omega) \in \mathbb{Q}_+$ — вес опорного множества $\Omega \in \Omega_A$. Бинарная функция $B_\Omega^{\tilde{\varepsilon}}(S^t, S_i)$ задаётся параметрами $\tilde{e} = (\varepsilon_1, \ldots, \varepsilon_n)$ и обращается в единицу тогда и только тогда, когда выполнены условия $\rho_i(a_i, b_i) \leqslant \varepsilon_i$ для всех $i \in \Omega$. В работе [2] показано, что при изучении алгебраических и линейных замыканий можно, без ограничения общности, рассматривать функций близости такого вида.

Решающее правило C по матрице оценок классифицирует объекты, т. е. получает матрицу $(\alpha_{ij})_{q\times l}$, в которой $\alpha_{ij} = 1$, если правило относит объект S_i к j-му классу, $\alpha_{ij} = 0$, если правило не относит объект S_i к j-му классу. В работе рассматриваются пороговые правила $C\left((\gamma_{ij})_{q\times l}\right) = (\alpha_{ij})_{q\times l}$, где $\alpha_{ij} = 1$, если $\gamma_{ij} \geqslant c$, и $\alpha_{ij} = 0$, если $\gamma_{ij} < c$.

*Работа выполнена при финансовой поддержке РФФИ, проекты №08-01-00636 и №08-01-00405.

Эквивалентность метрик

Будем считать, что функции расстояний на признаках $\rho_j\colon \tilde{N}_j \times \tilde{N}_j \to \tilde{M}_j = \{0, \ldots, M_j - 1\}$, $j = 1, \ldots, n$ удовлетворяют всем условиям метрики.

Определение 1. *Метрической характеристикой признака j алгоритма ABO в дальнейшем будем называть пару $\{\rho_j, \varepsilon_j\}$.*

Определение 2. *Будем говорить, что метрические характеристики признаков s и r эквивалентны относительно задачи распознавания Z (и обозначать $\{\rho_s, \varepsilon_s\} \overset{Z}{\sim} \{\rho_r, \varepsilon_r\}$), если при их использовании в ABO алгоритмы A_{ρ_s,ε_s} и A_{ρ_r,ε_r} дают одинаковые результаты для всех объектов контрольной выборки, то есть для всех S_i, $i = 1, \ldots, q$ выполняются равенства $A_{\rho_s,\varepsilon_s}(S_i) = A_{\rho_r,\varepsilon_r}(S_i)$.*

Если $\varepsilon_s = \varepsilon_r = \tilde{\varepsilon}$, то будем говорить об эквивалентности метрик и обозначать эквивалентность метрик, как $\rho_s \overset{Z,\tilde{\varepsilon}}{\sim} \rho_r$.

Справедлива следующая

Теорема 1. *Пусть для решения задачи распознавания с порядковыми признаками Z при использовании ABO с фиксированными параметрами метрические характеристики $\{\rho_j, \varepsilon_j\}$, $j = 1, \ldots, n$ также фиксированы. Тогда справедливы следующие утверждения:*

1. *Если $\varepsilon_j = 0$ или $\varepsilon_j > M_j - 1$, то все метрики ρ_j эквивалентны относительно задачи распознавания Z.*
2. *Если $0 < \varepsilon_j < M_j - 1$, то существует метрика $\rho_j^*\colon \tilde{N}_j \times \tilde{N}_j \to \{0, 1, 2\}$ такая, что метрические характеристики $\{\rho_j, \varepsilon_j\}$ и $\{\rho_j^*, 1\}$ эквивалентны относительно задачи Z, $\{\rho_j, \varepsilon_j\} \overset{Z}{\sim} \{\rho_j^*, 1\}$.*

Первое утверждение следует из того, что при $\varepsilon_j = 0$ условие $\rho_j(\boldsymbol{a}, \boldsymbol{b}) \leqslant \varepsilon_j$ выполняется тогда и только тогда, когда $a_j = b_j$, что не зависит от выбираемой метрики. При $\varepsilon_j > M_j - 1$ условие $\rho_j(\boldsymbol{a}, \boldsymbol{b}) \geqslant \varepsilon_j$ выполняется для всех метрик.

Для доказательства утверждения 2 можно рассмотреть преобразование

$$\rho_j^*(\boldsymbol{a}, \boldsymbol{b}) = \begin{cases} 1, & \text{если } \rho_j(\boldsymbol{a}, \boldsymbol{b}) \leqslant \varepsilon_j; \\ 2, & \text{в противном случае.} \end{cases}$$

Если положить $\varepsilon_j^* = 1$, то неравенства $\rho_j(\boldsymbol{a}, \boldsymbol{b}) \leqslant \leqslant \varepsilon_j$ и $\rho_j^*(\boldsymbol{a}, \boldsymbol{b}) \leqslant \varepsilon_j^*$ будут выполняться одновременно, и, следовательно, одинаково влиять на результат работы алгоритма. Теорема доказана.

Теорема 1, без ограничения общности, позволяет для распознавания использовать только метрики, принимающие значения на множестве $\{0, 1, 2\}$.

Заметим, что метрику $\rho_j^*\colon \tilde{N}_j \times \tilde{N}_j \to \{0, 1, 2\}$ можно представить, как матрицу расстояний размера $N_j \times N_j$ с нулевой диагональю и недиагональными элементами из множества $\{1, 2\}$ (элементы таких матриц всегда удовлетворяют всем условиям метрики).

Очевидно, что $\rho_j^* \overset{Z,1}{\sim} \rho_i^*$ тогда и только тогда, когда $\rho_j^* = \rho_i^*$. Поэтому число неэквивалентных метрик в АВО при решении задач распознавания равно $2^{N_j(N_j-1)/2}$.

Корректность алгебраического замыкания

Алгебраический подход к решению задач распознавания, предложенный Ю. И. Журавлёвым, заключается во введении следующих операций над матрицами оценок [3]:

1) сложение:

$$\big(\Gamma_{ij}[B_1 + B_2]\big)_{q\times l} = \big(\Gamma_{ij}[B_1]\big)_{q\times l} + \big(\Gamma_{ij}[B_2]\big)_{q\times l};$$

2) умножение на число:

$$\big(\Gamma_{ij}[cB]\big)_{q\times l} = c\big(\Gamma_{ij}[B]\big)_{q\times l};$$

3) умножение (адамарово):

$$\big(\Gamma_{ij}[B_1B_2]\big)_{q\times l} = \big(\Gamma_{ij}[B_1]\big)_{q\times l} \circ \big(\Gamma_{ij}[B_2]\big)_{q\times l} = \big(\Gamma_{ij}[B_1] \cdot \Gamma_{ij}[B_2]\big)_{q\times l}.$$

Определение 3. *Линейным замыканием $L(B^*)$ множества B^* распознающих операторов АВО называется множество всех линейных комбинаций из B^*:*

$$L(B^*) = \big\{c_1B_1 + \cdots + c_rB_r \;\big|\; r \in \mathbb{N}, c_1, \ldots, c_r \in \mathbb{Q}, B_1, \ldots, B_r \in B^*\big\}.$$

Алгебраическим замыканием k-ой степени $U^k(B^*)$ называется множество полиномов от операторов B^* степени не выше k:

$$U^k(B^*) = L\big(\{B_1 \cdots B_s | B_1, \ldots, B_s \in B^*, 1 \leqslant s \leqslant k\}\big).$$

Алгебраическое замыкание — множество всех полиномов $U(B^*) = \bigcup\limits_{k=1}^{\infty} U^k(B^*)$.

Множество распознающих алгоритмов будем называть *моделью* распознающих алгоритмов.

Определение 4 ([2, 4, 5]). *Задача распознавания называется регулярной, если выполнены следующие условия:*

1) $\big|\{\tilde{K}_1, \ldots, \tilde{K}_l\}\big| = l$, *где* $\tilde{K}_j = \tilde{S}^m \cap K_j$ *(множества эталонов каждого из классов попарно различны);*
2) *в матрице* $\boldsymbol{R}$ *размера* $q \times mn$,
$$\boldsymbol{R} = \big(\big(\rho_1(S^j, S_i), \ldots, \rho_n(S^j, S_i)\big)\big)_{i=1\,j=1}^{q\quad m},$$
все q строк попарно различны (в контрольной выборке нет ни одной пары объектов, неразличимых относительно эталонов);
3) $\tilde{S}^m \cap \tilde{S}_q = \varnothing$ *(обучающая и контрольная выборки не пересекаются).*

Определение 5 ([2, 4, 5]). *Модель R^* распознающих алгоритмов называется корректной, если*

$$\forall \Gamma \in \mathbb{Q}^{q\times l} \;\; \exists B \in R^*\colon \; \Gamma[B] = \Gamma.$$

В работах [2, 4, 6] были найдены критерии корректности линейного и алгебраического замыканий моделей АВО для фиксированной метрики. В частности, была доказана следующая

Теорема 2. *Модель $U(B^*)$ корректна тогда и только тогда, когда выполнены первое и второе условия регулярности.*

В дальнейшем будем считать, что первое условие регулярности выполнено. Найдем метрики, при которых выполнено второе условие регулярности для рассматриваемой задачи, тем самым получим критерий корректности алгебраического замыкания АВО для задач с порядковыми признаками.

Задача состоит в нахождении условий на ρ_j, при которых все строки матрицы $\boldsymbol{R}$ различны.

Переставим столбцы в матрице так, чтобы первые m столбцов соответствовали сравнению контрольных объектов со всеми обучающими объектами по первому признаку. Следующие m столбцов соответствовали сравнению по второму признаку, и т. д. Рассмотрим m столбцов, соответствующих j-му признаку.

Если в контрольной выборке есть два объекта S_r и S_p таких, что $a_j(S_r) = a_j(S_p)$, то, очевидно, что в подматрице расстояний подстроки r и p совпадут. Поэтому максимальное число различных подстрок равно N_j. Ясно, что при некоторых ρ_j определённые строки могут совпадать, и тогда число различных строк будет меньше N_j.

Определение 6. *Будем говорить, что объекты разделимы (полностью разделимы) по j-му признаку, если число различных строк в подматрице матрицы $\boldsymbol{R}$, соответствующей j-му признаку, равно числу значений j-го признака на объектах контрольной выборки.*

Признак j будем называть полностью делящимся признаком. Если условие не выполнено, то признак j будем называть неполностью делящимся признаком.

1. Найдем условие, при котором объекты разделимы по j-му признаку.

1.1. Пусть $a_j(S_r) = a_j(S^p)$, тогда $\rho_j(S_r, S^p) = 0$ для всех метрик ρ_j. Аналогично, для всех метрик ρ_j: если $a_j(S_r) \neq a_j(S^p)$, то $\rho_j(a_j(S_r), a_j(S^p)) \neq 0$. Таким образом, автоматически выделяется группа одинаковых строк, характеризующихся положением нулей.

Если всю обучающую выборку можно разбить на группы, характеризующиеся положениями нулей, то для получения такого разбиения можно использовать любую метрику, и дальнейшее деление по выбранному признаку невозможно. Очевидно, что в данном случае число непересекающихся классов равно числу различных значений j-го признака на объектах контрольной выборки.

1.2. Пусть после выделения описанных выше групп осталось некоторое число строк, принадлежащих одной группе. Они характеризуются элементами сравнения несовпадающих значений j-го признака на объектах обучающей и контрольной выборок.

Пусть число разных значений j-го признака на объектах из обучающей выборки равно s_j, а из контрольной — t_j. Ни одно значение из этих s_j на объектах из обучающей выборки не равно ни одному значению из t_j на объектах контрольной выборки (иначе можно было бы разделить объекты по получившимся нулям в п. 1.1). Число возможных заданий метрик на этих значениях признака равно числу определений матриц размера $s_j \times t_j$, т. е. $\binom{2^{s_j}}{t_j} = \frac{2^{s_j}!}{t_j!(2^{s_j}-t_j)!}$.

Заметим, что матрицу расстояний можно выбрать тогда и только тогда, когда $\binom{2^{s_j}}{t_j} \geqslant 1$, что при $s_j, t_j \geqslant 1$ эквивалентно $2^{s_j} \geqslant t_j$.

Таким образом, справедлива

Теорема 3. *Объекты обучающей выборки разделимы по j-му признаку тогда и только тогда, когда $2^{s_j} \geqslant t_j$.*

2. Рассмотрим случай с n признаками.

Утверждение 4. *Пусть все объекты в контрольной выборке различны. Тогда, если объекты обучающей выборки разделимы по каждому признаку, то все строки в матрице расстояний $\boldsymbol{R}$ различны.*

Проведём доказательство от противного. Допустим, что в задаче распознавания с различными контрольными объектами при разделимости объектов контрольной выборки по каждому признаку в матрице расстояний $\boldsymbol{R}$ существуют две одинаковые строки.

Пусть есть разделение по каждому признаку. Тогда каждой j-й подматрице матрицы $\boldsymbol{R}$ можно однозначно сопоставить значение j-го признака на объекте из контрольной выборки. Следовательно, каждой строке в матрице расстояний можно однозначно сопоставить объект из контрольной выборки. Поэтому, если строки одинаковы, то и объекты в контрольной выборке были одинаковыми, что противоречит условию различности объектов. Теорема доказана.

Из теоремы 3 следует

Утверждение 5. *Пусть все объекты в контрольной выборке различны. Тогда, если объекты обучающей выборки разделимы по всем признакам кроме одного (j-го), то в матрице расстояний могут быть совпадающие строки.*

Пусть, без ограничения общности, в рассматриваемой задаче нет разделения по признакам $j = 1, \ldots, r$, и есть разделимость по $j = r+1, \ldots, n$. По теореме 3 это означает, что $2^{s_j} < t_j$ для $j = 1, \ldots, r$ и $2^{s_j} \geqslant t_j$ для $j = r+1, \ldots, n$.

Определим произвольным образом матрицы расстояний для полностью делящихся признаков. Сгруппируем объекты по одинаковым наборам значений признаков $j = r+1, \ldots, n$. Для решения задачи надо определить функции расстояний для признаков $j = 1, \ldots, r$ так, чтобы в каждой группе были различные строки.

Очевидно, в каждой группе можно провести деление на подгруппы, характеризующиеся положением нулей. Теперь в каждой подгруппе надо доопределить значения расстояний на соответствующих значениях признаков на объектах, т. е. значения попарных расстояний в матрицах размеров $s_1 \times t_1, \ldots, s_r \times t_r$ так, чтобы в каждой подгруппе были различные векторы. Каждую матрицу можно определить $\prod_{i=1}^{r} 2^{s_i t_i}$ способами. Выбор согласованных во всех группах метрик можно осуществить их перебором.

Если существует разбиение объектов на группы такое, что в каждой подгруппе оказываются различные векторы, то для него второе условие регулярности выполняется, и задача распознавания с выбранными метриками является регулярной. Иначе задача регулярной не является. Так как регулярность задачи в данном случае равносильна условию корректности модели относительно данной задачи, то получаем следующую теорему.

Теорема 6. *Пусть в задаче осуществлено деление на подгруппы по полностью делящимся признакам. Тогда задача регулярна (а модель $U(B^*)$ корректна) тогда и только тогда, когда для неполностью делящихся признаков можно определить значения матриц попарных расстояний так, чтобы в каждой подгруппе, характеризующейся положением нулей, были различные векторы.*

Заметим, что задача проверки задачи Z на регулярность (а модель $U(B^*)$ на корректность) име-

ют сложность $O(n) + O\left(\prod_{i=1}^{r} 2^{s_i t_i}\right)$. Действительно, проверка на разделимость по признаку и деление на группы имеет сложность $O(n)$. Если есть r неполностью делящихся признаков, то дополнительная проверка будет осуществляться со сложностью $O\left(\prod_{i=1}^{r} 2^{s_i t_i}\right)$.

Пример 1. Пусть $N_1 = N_2 = 5$, $S^1 = (2,1)$, $S^2 = (1,2)$; $S_1 = (3,3)$, $S_2 = (4,4)$, $S_3 = (3,4)$, $S_4 = (4,3)$. Очевидно, что объекты разделимы по каждому признаку ($s_1 = s_2 = t_1 = t_2 = 2$, $2^{s_j} \geqslant t_j$, $j = 1,2$). Поэтому рассматриваемая задача регулярна (а модель $U(B^*)$ корректна).

Метрики с выполненным условием порядка

Рассмотрим задачу с порядковыми признаками. Пусть на множествах значений признаков задано отношение порядка: $0 < 1 < \cdots < N_j - 1$, и функции расстояний на признаках $\rho_j(x,y)$ удовлетворяют дополнительному *условию порядка*: если $x \geqslant y$, то $\rho_j(x,z) \geqslant \rho_j(y,z)$ для всех $z = 0,\ldots,N_j - 1$ таких, что $z \leqslant y$.

Таким образом, если рассматривать метрику $\rho(s,d)$ как матрицу попарных расстояний $C = (c_{sd})_{s,d=1}^{N_j}$, $c_{sd} \in \{0,1,2\}$, то эта матрица симметричная, с нулевой диагональю. Элементы матрицы, находящиеся выше главной диагонали, не возрастают по строкам (сверху вниз) и не убывают по столбцам (слева направо). Найдем условия корректности алгебраического замыкания модели ABO для данного случая.

Все заключения, полученные для общего случая задач с порядковыми признаками, верны с той разницей, что условие разделимости по j-му признаку принимает следующий вид.

Теорема 7. *Объекты обучающей выборки разделимы по j-му признаку тогда и только тогда, когда $s_j \geqslant t_j + 1$.*

Действительно, задача отличается только тем, что множество матриц, среди которых осуществляется поиск, является подмножеством матриц в общем случае. Характеристикой данного подмножества является неубывание значений по столбцам и невозрастание по строкам, то есть $c_{ij_1} \leqslant c_{ij_2}$ для всех $j_1 \leqslant j_2$, и $c_{i_1 j} \leqslant c_{i_2 j}$ для всех $i_1 \geqslant i_2$.

Для определения метрик на множествах значений j-го признака на объектах обучающей и контрольной выборок (определение элементов матрицы размера $s_j \times t_j$) необходимо для каждой строки i определить номер столбца d, в котором встречается первая двойка, т. е. такой, что $c_{id_1} = 1$ для всех $d_1 < d$, и $c_{id_2} = 2$ для всех $d_2 \geqslant d$. Это можно сделать $\binom{s_j}{t_j}$ способами. Всего имеется $s_j + 1$ мест, на которых может первый раз встретиться двойка. Поэтому выбор возможен при $\binom{s_j+1}{t_j} \geqslant 1$, а при $s_j,\ t_j \geqslant 1$ это эквивалентно $s_j + 1 \geqslant t_j$.

Перебор метрик в задаче с порядковыми признаками уменьшается, т. к. рассматриваются только функции расстояний, удовлетворяющие условию порядка (в теореме 6 добавится ограничение на выбираемые метрики).

Это сокращает сложность перебора до величины $O\left(\prod_{i=1}^{r} \binom{s_i+t_i}{t_i}\right)$, определяемой числом монотонных слов длины t_j в алфавите из s_j+1 символов [7]. Таким образом, сложность проверки задачи на регулярность составит $O(n) + O\left(\prod_{i=1}^{r} \binom{s_i+t_i}{t_i}\right)$.

Пример 2. Пусть $N_1 = 7$, $S^1 = 1$, $S^2 = 2$; $S_1 = 3$, $S_2 = 4$, $S_3 = 5$, $S_4 = 6$. Здесь $s_1 = 2$, $t_1 = 4$, поэтому условие $s_1 + 1 \geqslant t_1$ не выполнено, и задача распознавания с порядковыми признаками не является корректной. Заметим, что аналогичная задача на дискретных признаках — корректа.

Выводы

Итак, в работе были получены условия корректности алгебраического замыкания ABO для задачи распознавания с порядковыми признаками при фиксированной метрике. Были найдены условия, когда задача является регулярной, а, следовательно, алгебраическое замыкание ABO корректным. Также был рассмотрен более общий случай, когда признаки принимали значения из ограниченного дискретного множества.

Полученные результаты могут быть использованы в дальнейшем исследовании алгебраических замыканий ABO, главным образом, в задачах распознавания с порядковыми признаками.

Литература

[1] *Журавлёв Ю. И.* Об алгебраическом подходе к решению задач распознавания или классификации // Пробл. киберн. — 1978. — Вып. 33. — С. 5–68.

[2] *Дьяконов А. Г.* Алгебра над алгоритмами вычисления оценок: Учебное пособие. — М.: Издат. отд. ф-та ВМиК МГУ, 2006. — 70 с.

[3] *Журавлёв Ю. И.* Корректные алгоритмы над множествами некорректных (эвристических) алгоритмов. Часть I. // Кибернетика. — 1977. — № 4. — С. 5–17.

[4] *Журавлёв Ю. И.* Корректные алгоритмы над множествами некорректных (эвристических) алгоритмов. Часть II. // Кибернетика. — 1977. — № 6. — С. 21–27.

[5] *Дьяконов А. Г.* Критерии корректности алгебраических замыканий модели алгоритмов вычисления оценок // ДАН. — 2008. — Т. 420, № 6. — С. 732–735.

[6] *Рудаков К. В.* Алгебраическая теория универсальных и локальных ограничений для алгоритмов распознавания: Дис. докт. физ.-матем. наук. — М.: ВЦ РАН, 1992. — 146 с.

[7] *Журавлёв Ю. И., Флёров Ю. А.* Дискретный анализ. Часть I: Учебное пособие. — М.: МФТИ, 1999. — 136 с.

Критерии k-сингулярности систем точек в алгебраическом подходе к распознаванию*

Карпович П. А., Дьяконов А. Г.
pkarpovich@mail.ru, djakonov@mail.ru
Москва, ф-т ВМК МГУ им. М. В. Ломоносова

Получены критерии корректности модели алгоритмов вычисления оценок в терминах матриц попарных расстояний. Представлена теория k-сингулярности систем точек, её приложения в алгебраическом подходе к распознаванию, обзор последних результатов и дальнейшие направления исследований.

Введение

В [1] был предложен алгебраический подход к решению задач распознавания, основная идея которого — представление корректного алгоритма в виде алгебраического выражения над некорректными (эвристическими) алгоритмами. Для этого каждый распознающий алгоритм представляется в виде суперпозиции распознающего оператора и решающего правила. Распознающий оператор получает $q\times l$-матрицу оценок, q — число контрольных объектов, l — классов, ij-й элемент которой — оценка принадлежности i-го контрольного объекта к j-му классу. Решающее правило по этой матрице оценок осуществляет классификацию. Операции над алгоритмами вводятся как операции над матрицами оценок, полученными распознающими операторами (решающее правило предполагается фиксированным).

В [2] впервые получены критерии корректности алгебраических замыканий и разрешимости задач алгоритмами этих замыканий. Под корректностью модели понимается возможность получить произвольную матрицу оценок алгоритмами модели (а следовательно, произвольную классификацию). В [2] исследовались полиномиальные замыкания — множества всех полиномов степени не выше k над алгоритмами модели. К сожалению, полученные критерии не допускают простой геометрической интерпретации в терминах исходной задачи.

В [3] представлены метрические критерии корректности и разрешимости. Далее опишем их для одной модели алгоритмов распознавания, а также перечислим основные направления исследований, связанные с их применением и последние результаты в этой области.

Модель АВО

Модель алгоритмов вычисления оценок (АВО) была введена в [4] как средство описания наиболее часто применявшихся в 1970-е годы распознающих алгоритмов. Благодаря своей универсальности и простоте реализации модель стала одной из наиболее хорошо исследованных и часто применяемых на практике (см. [5, 6]). Опишем результаты метрической теории корректности для одной подмодели модели АВО.

Рассмотрим задачу распознавания с двумя непересекающимися классами в стандартной признаковой постановке [5]. Множество

$$M = K_1 \cup K_2,\ K_1 \cap K_2 = \varnothing,$$

состоит из объектов, заданных признаковыми описаниями

$$S = \big(f_1(S), \ldots, f_n(S)\big) \in M_1 \times \ldots \times M_n,$$

где M_i — метрическое пространство с метрикой ρ_i. Задача распознавания состоит в том, чтобы построить алгоритм A, который по набору $\tilde{S}^m = \{S^t\}_{t=1}^m$ эталонных объектов, для которых известна классификация (принадлежность классам K_1, K_2), получает классификацию контрольных объектов $\tilde{S}_q = \{S_t\}_{t=1}^q$. Для просторы рассмотрим случай $M_1 = \ldots = M_n = \mathbb{R},\ \rho_i(x,y) = |x-y|,\ \ i = 1, \ldots, n$.

Распознающий оператор алгоритма модели вычисления оценок с одноэлементными опорными множествами (см. [1, 5]) по описаниям контрольных объектов $\tilde{S}_q$ получает матрицу оценок

$$\Gamma[B] = ||\Gamma_{ij}[B]||_{q\times l},\ l = 2,$$

$$\Gamma_{ij}[B] = \sum_{a,b=0,0}^{1,1} x_{ab} \sum_{h=1}^{n} \sum_{S^t \in \tilde{K}_j^a} w^t w_h \mathrm{B}_h^{\varepsilon_h, b}(S^t, S_i),$$

где $w^t \in \mathbb{Q}^+ = \{x \in \ Q \mid x \geqslant 0\}$ при $t = 1, \ldots, m$ (вес t-го объекта), $w_h \in \mathbb{Q}^+$ при $\Omega \in \Omega_A$ (вес h-го признака),

$$\mathrm{B}_h^{\varepsilon_h, b}(S^t, S_i) = \begin{cases} b, & \rho_h\big(f_h(S^t), f_h(S_i)\big) \leqslant \varepsilon_h, \\ 1-b, & \rho_h\big(f_h(S^t), f_h(S_i)\big) > \varepsilon_h, \end{cases}$$

$$\tilde{K}_j^a = \begin{cases} \tilde{S}^m \cap K_j, & a = 1, \\ \tilde{S}^m \backslash K_j, & a = 0, \end{cases}$$

$(a,b) \in \{0,1\}^2$, $x_{ab} \in \{0, (-1)^{a+b}\}$.

Для задачи с двумя непересекающимися классами обычно применяется *решающее правило по максимуму*: если $\Gamma_{i1} > \Gamma_{i2}$, то объект S_i относят к первому классу, если $\Gamma_{i1} < \Gamma_{i2}$, то ко второму (при равенстве оценок отказываются от классификации).

*Работа выполнена при финансовой поддержке РФФИ, проект № 08-07-00305-а.

Операции над распознающими операторами вводят как операции над их матрицами оценок:

$$\Gamma[B_1 + B_2] = \Gamma[B_1] + \Gamma[B_2],$$

$$\Gamma[cB] = c\Gamma[B], \quad \Gamma[B_1 \cdot B_2] = \Gamma[B_1] \circ \Gamma[B_2],$$

умножение $\circ$ поэлементное. При фиксированном решающем правиле эти операции индуцируют алгебру над алгоритмами. Для множества B^* всех операторов рассматриваемой модели АВО вводим понятия: *линейного замыкания*

$$\mathbf{L}(B^*) = \{c_1 B_1 + \ldots + c_r B_r \mid r \in \mathbb{N}, \\ c_1, \ldots, c_r \in \mathbb{Q},\ B_1, \ldots, B_r \in B^*\},$$

алгебраического замыкания k-й степени

$$\mathbf{U}^k(B^*) = \mathbf{L}(\{B_1 \cdot \ldots \cdot B_s \mid B_1, \ldots, B_s \in B^*, \\ 1 \leqslant s \leqslant k\}),$$

алгебраического замыкания $\mathbf{U}(B^*) = \bigcup\limits_{k=1}^{\infty} \mathbf{U}^k(B^*)$.

Для матрицы H через $\mathbf{U}^k(H)$ обозначаем множество всех значений полиномов степени не выше k над столбцами этой матрицы (умножение поэлементное).

Задача с резнесёнными эталонами

Определение 1. *Задача распознавания (в рассматриваемой постановке) называется задачей с разнесёнными эталонами, если для любого признака $r \in \{1, \ldots, n\}$ найдутся эталонные объекты S^{t_1}, S^{t_2}, для которых $f_r(S^{t_1}) \neq f_r(S^{t_2})$.*

Требование «разнесённости эталонов» достаточно естественное: если есть признак, принимающий одно значение на всех (эталонных) объектах, то обычно его исключают по причине низкой информативности.

Теорема 1. *В задаче с двумя непересекающимися классами и разнесёнными эталонами алгебраическое замыкание $\mathbf{U}(B^*)$ корректно тогда и только тогда, когда $|\tilde{S}_q| = q$.*

Теорема 2. *В задаче с двумя непересекающимися классами и разнесёнными эталонами множество матриц оценок операторов замыкания $\mathbf{U}^k(B^*)$ есть*

$$\{[h_1\, h_2] \mid \{h_1, h_2\} \subseteq \mathbf{U}^k(H)\},$$

где H — матрица попарных l_1-расстояний системы контрольных объектов $\tilde{S}_q$.

Следствие 1. *В рассматриваемой задаче добавление новых эталонных объектов не оказывает влияние на корректность алгебраических замыканий исследуемой модели.*

В общем случае этот результат не имеет места, более того, эталонные объекты можно добавить так, что линейное замыкание модели будет корректным.

Следствие 2. *В рассматриваемой задаче алгебраическое замыкание k-й степени модели АВО с системой всех одноэлементных опорных множеств корректно тогда и только тогда, когда размерность пространства $\mathbf{U}^k(H)$ равна q.*

В следующем разделе подробно исследуем этот критерий корректности. Нас интересует, в первую очередь, геометрическая интерпретация — каким конфигурациям контрольных объектов соответствует корректность модели $\mathbf{U}^k(B^*)$.

Теорема 3. *Пусть $|\tilde{S}_q| = q$ в задаче с двумя непересекающимися классами и разнесенными эталонами. Алгебраическое замыкание k-й степени модели АВО с системой всех одноэлементных опорных множеств корректно при $k \geqslant \min\{n, \lfloor \log_2 q \rfloor\}$. Оценка степени является неулучшаемой.*

k-сингулярность

Исследуем задачу о размерности пространства значений полиномов ограниченной степени над столбцами матрицы. Отметим, что задача имеет приложения не только в теории распознавания (см. теорему 2), но и в теории интерполяции: выяснение возможности представления функции, заданной лишь на конечном множестве, в виде суммы функций из определенного класса, часто сводится к анализу матрицы попарных расстояний точек этого множества. Например, точное представление функции в классе радиальных базисных функций (RBF) [7] или жёстких функций (riddle functions) [8] возможно тогда и только тогда, когда такая матрица попарных l_p-расстояний невырождена. Из классической серии работ И. Шенберга (см. полный перечень ссылок в [8]) следует, что матрица невырождена для конечной системы $\{\tilde{s}_i\}_{i=1}^q$ попарно различных точек пространства $\mathbb{R}^m$ и метрики l_p, $p \in (1, 2]$, $q > 1$. Особый интерес представляет случай $p = 1$, для которого критерий вырожденности был получен только в 1993 году [8]. Представим решение более общей задачи [9]: критерии неполноты размерности пространства значений полиномов ограниченной степени от столбцов матрицы попарных l_1-расстояний (операция умножения поэлементная), а также новые критерии вырожденности такой матрицы.

Пусть задана система попарно различных точек $S = \{\tilde{s}_i\}_{i=1}^q$ пространства $\mathbb{R}^m$, $|S| = q \geqslant 2$. Пусть

$$\{(\tilde{s}_i)_t \mid i \in \{1, \ldots, q\}\} = \{a_{t0}, \ldots, a_{tp(t)}\} = A_t,$$

$a_{t0} < \ldots < a_{tp(t)}$ при $t \in \{1, \ldots, m\}$, $(\tilde{s}_i)_t$ — t-я координата точки $\tilde{s}_i$.

Определение 2. *Система точек S называется k-сингулярной, если размерность пространства $\mathbf{U}^k[S] = \mathbf{U}^k(P_S)$ меньше q, где P_S — матрица попарных l_1-расстояний системы точек S.*

Заметим, что в определении можно матрицу P_S заменить на матрицу попарных расстояний Хэмминга. Без ограничения общности считаем, что $p(t) \geqslant 1$ для всех $t \in \{1, \ldots, m\}$ (t-ю координату можно удалить при $p(t) = 1$, не изменив $\mathbf{U}^k(P_S)$). Ниже представим новые критерии k-сингулярности [9, 10] (обзор некоторых известных критериев 1-сингулярности можно найти в [8]).

Теорема 4. *Пусть z — биективное отображение множества $\{1, \ldots, C_m^k\}$ на множество сочетаний без повторений объема k (из множества $\{1, \ldots, m\}$). Система точек $\{\tilde{s}_i\}_{i=1}^q$ является k-сингулярной тогда и только тогда, когда является 1-сингулярной система точек $\{\tilde{d}_i\}_{i=1}^q$ пространства $\mathbb{R}^{C_m^k}$ такая, что для всех $t \in \{1, \ldots, C_m^k\}$ справедливо*

$$(\tilde{d}_i)_t = (\tilde{d}_j)_t \quad \Leftrightarrow \quad \forall r \in z(t) \quad (\tilde{s}_i)_r = (\tilde{s}_j)_r.$$

Теорема 5. *Система точек S не является k-сингулярной при $k \leqslant m$ тогда и только тогда, когда любая функция $f(x_1, \ldots, x_m)$ на точках системы S может быть представлена в виде конечной суммы функций, каждая из которых зависит от k переменных. Любая функция $f(x_1, \ldots, x_m)$ на точках множества S может быть представлена в виде конечной суммы функций, каждая из которых зависит от $\lfloor \log_2 |S| \rfloor$ переменных.*

Пусть G — минимальная группа (с операцией суперпозиция), содержащая преобразования $g_{t,x,y}$: $\mathbb{R}^m \to \mathbb{R}^m$ при всех $t \in \{1, \ldots, m\}$, $x \in \mathbb{R}$, $y \in \mathbb{R}$ такая, что

$$g_{t,x,y}(s_1, \ldots, s_m) = \\ = \begin{cases} (s_1, \ldots, s_m), & s_t \notin \{x, y\}, \\ (s_1, \ldots, s_{t-1}, y, s_{t+1}, \ldots, s_m), & s_t = x, \\ (s_1, \ldots, s_{t-1}, x, s_{t+1}, \ldots, s_m), & s_t = y. \end{cases}$$

Теорема 6. *Для любого преобразования $g \in G$ справедливо равенство $\mathbf{U}^k[S] = \mathbf{U}^k[g(S)]$, где $g(S) = \{g(\tilde{s}_i)\}_{i=1}^q$.*

Следствие 3. *Система точек S k-сингулярна тогда и только тогда, когда k-сингулярна система точек $g(S)$. Достаточно ограничиться рассмотрением систем точек на целочисленной решетке, поскольку*

$$\mathbf{U}^k\big[\{(a_{1,b(i,1)}, \ldots, a_{m,b(i,m)})\}_{i=1}^q\big] = \\ = \mathbf{U}^k[\{(b(i,1), \ldots, b(i,m))\}_{i=1}^q].$$

Теорема 7. *Система точек $S = \{\tilde{s}_i\}_{i=1}^q$ является 1-сингулярной тогда и только тогда, когда существует такое подмножество $X \subseteq \{1, \ldots, q\}$, что для любого преобразования $g \in G$ система точек $\{g(\tilde{s}_i)\}_{i \in X}$ не отделима от системы точек $\{g(\tilde{s}_i)\}_{i \in \{1,\ldots,q\} \setminus X}$ гиперплоскостью.*

Неформально условие теоремы 7 можно переформулировать следующим образом: система точек не является 1-сингулярной тогда только тогда, когда при любом разбиении ее на две непересекающиеся подсистемы они разделимы с помощью «суперпозиции» некоторого преобразования $g \in G$ и гиперплоскости. В условии теоремы отделимость гиперплоскостью можно заменить отделимостью с помощью гиперплоскости, проходящей через ноль, или любой фиксированной с уравнением $a_1 x_1 + \ldots + a_m x_m + a_0 = 0$, $0 \notin \{a_1, \ldots, a_m\}$.

Критерий k-сингулярности также можно сформулировать в виде условия существования линейной зависимости «антипотенциальных» функций $\rho^k(\tilde{s}, \tilde{s}_i)$:

Теорема 8. *Система точек $S = \{\tilde{s}_i\}_{i=1}^q$ пространства $\mathbb{R}^m$ является k-сингулярной тогда и только тогда, когда существует ненулевой вектор $(c_1, \ldots, c_q)$, для которого при всех $\tilde{s} \in \mathbb{R}^m$ справедливо равенство $\sum\limits_{i=1}^q c_i \rho^k(\tilde{s}, \tilde{s}_i) = 0$, где ρ — метрика Хэмминга или l_1-метрика.*

Теорема перестает быть верной при изометричном вложении системы точек в другое пространство. Для $k = 1$ теорема сформулирована в [3], в общем случае — в [10].

Геометрия k-сингулярных систем

Пусть $N(f) = \{x \mid f(x) \neq 0\}$ — *носитель* функции $f\colon X \to \mathbb{R}$, а $f(Y) = \{f(x) \mid x \in Y\}$.

Пусть функция $\Pi\colon \mathbb{R}^m \to \mathbb{R}$ такая, что $N(\Pi) \subseteq X = \bigtimes\limits_{t=1}^m \{a_t, b_t\}$, $r = \big|\{t \in \{1, \ldots, m\} \mid a_t \neq b_t\}\big|$. Если $\Pi(X) = \{1\}$ или $\Pi(X) = \{-1\}$, то функция Π называется *константным параллелепипедом* (к. п.) размерности r, а если в каждой точке $(c_1, \ldots, c_m)$ множества X она равна $(-1)^c$, где $c = \big|\{t \in \{1, \ldots, m\} \mid c_t = a_t\}\big|$, то она называется *размеченным параллелепипедом* (р. п.) размерности r. Элементы множества X называются *вершинами* р. п. (к. п.). Параллелепипед называется *последовательным относительно множества A*, если $N(\Pi) = \bigtimes\limits_{t=1}^m Y^t$ и для всех $t \in \{1, \ldots, m\}$ найдётся $r \in \{1, \ldots, p(t)\}$ такой, что $Y^t \subseteq \{a_{t,r-1}, a_{tr}\}$.

Теорема 9. *Система точек S k-сингулярна тогда и только тогда, когда найдётся конечная сумма Σ функций из множества Π^k, для которой $\varnothing \neq N[\Sigma] \subseteq S$. Утверждение справедливо для множеств Π^k следующих функций:*

1) *к. п. размерности больше k с одной общей вершиной;*

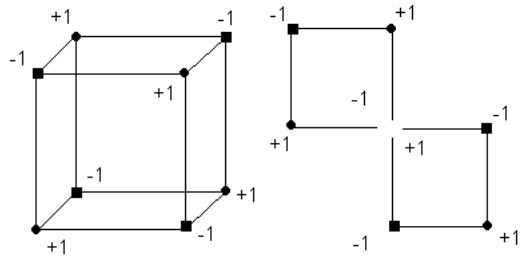


Рис. 1. Размеченный параллелепипед (р. п.) размерности 3 (слева) и пример 1-сингулярной системы в $\mathbb{R}^2$ (носитель суммы двух р. п. размерности 2).

2) р. п. размерности больше k с одной общей вершиной;
3) р. п. размерности $(k+1)$ (см. рис. 10);
4) последовательных относительно A р. п. размерности $(k+1)$;
5) последовательных относительно A к. п. размерности $(k+1)$.

Пункт *3)* теоремы обобщает результат [8].

Дальнейшие исследования

В этой работе приложения результатов о k-сингулярных системах к распознаванию описаны лишь для одной подмодели модели АВО и задачи распознавания с двумя непересекающимися классами. Аналогичные приложения есть и в общем случае (см., например, [3]): некорректность алгебраического замыкания k-й степени модели АВО эквивалентна k-сингулярности системы ql точек, заданной матрицей попарных l_1-расстояний. При представлении k-сингулярной системы точек с помощью носителя суммы р. п. в явном виде получается ненулевой вектор пространства, ортогонального к пространству $\mathbf{U}^k[S]$, а применительно к задаче распознавания — классификация, не реализуемая алгоритмом из $\mathbf{U}^k(B^*)$.

Теорема 10. *В задаче распознавания с двумя непересекающимися классами и разнесёнными эталонами для любого оператора B из алгебраического замыкания k-й степени модели АВО с системой всех одноэлементных опорных множеств существуют функции $g_{r_1,\ldots,r_k}$, $1 \leqslant r_1 < \ldots < r_k \leqslant n$, от k переменных, для которых*

$$\Gamma_{i1}[B] = \sum_{1\leqslant r_1<\ldots<r_k\leqslant n} g_{r_1,\ldots,r_k}\big(f_{r_1}(S_i),\ldots,f_{r_k}(S_i)\big).$$

Последний результат позволяет сводить задачу поиска оптимальных алгоритмов в алгебраических замыканиях к задаче регрессии. Алгоритмы линейного замыкания являются «суперпозициями» линейной регрессии $g_1(f_1(S)) + \ldots + g_n(f_n(S))$ и решающего правила, а также, в некотором смысле, суперпозициями элементов группы G и линейного классификатора. Это определяет вид разделяющих поверхностей в пространстве $\tilde{S}_q$ для модели $\mathbf{L}(B^*)$.

Дальнейшие исследования планируется посвятить разработке методов поиска оптимальных алгоритмов в алгебраических замыканиях, а также методов представления k-сингулярной системы в виде объединения подсистем, которые уже не являются k-сингулярными. Это позволит разбивать признаковое пространство на области компетентности алгоритмов [11]. Здесь представляют интерес и теоретические результаты, например оценки на число подсистем в разбиении 1-сингулярной системы. Верхняя оценка для минимального числа таких подсистем получается с помощью применения жадного алгоритма разделения на подсистемы, поскольку в системе $\{\tilde{s}_i\}_{i=1}^q$ из $\mathbb{R}^m$ всегда можно найти подсистему из $\lceil \sqrt[m]{q} \rceil$ точек, не являющуюся 1-сингулярной.

Литература

[1] *Журавлёв Ю. И.* Корректные алгоритмы над множествами некорректных (эвристических) алгоритмов. II // Кибернетика. — 1977. — № 6. — С. 21–27.

[2] *Матросов В. Л.* О критериях полноты модели алгоритмов вычисления оценок и её алгебраических замыканий // Докл. АН СССР. — 1981. — Т. 258, № 4. — С. 791–796.

[3] *Дьяконов А. Г.* Критерии корректности алгебраических замыканий модели алгоритмов вычисления оценок // Докл. РАН. — 2008. — Т. 420, № 6. — С. 732–735.

[4] *Журавлёв Ю. И., Никифоров В. В.* Алгоритмы распознавания, основанные на вычислении оценок // Кибернетика. — 1971. — № 3. — С. 1–11.

[5] *Журавлёв Ю. И.* Об алгебраическом подходе к решению задач распознавания или классификации // Пробл. кибернетики. — 1978. — Вып. 33. — С. 5–68.

[6] *Журавлёв Ю. И, Рязанов В. В., Сенько О. В.* «РАСПОЗНАВАНИЕ». Математические методы. Программная система. Практические применения. — М.: Фазис, 2006. — 176 с.

[7] *Baxter B. J. C.* Conditionally positive functions and p-norm distance matrices // Constr. Approx. — 1991. — № 7. — P. 427–440.

[8] *Reid L. Sun X.* Distance matrices and ridge function interpolation // Constr. Approx. — 1993. — V. 45. — P. 1313–1323.

[9] *Дьяконов А. Г.* Критерии вырожденности матрицы попарных l_1-расстояний и их обобщения // Докл. РАН. — 2009. — Т. 425, № 1. — С. 11–14.

[10] *Карпович П. А.* Критерий k-сингулярности системы точек и оптимальное разбиение на подсистемы // Материалы XVI Международной конференции студентов, аспирантов и молодых учёных: секция ВМК, М.: Изд. отд. ф-та ВМК МГУ, МАКС Пресс, 2009. — С. 33.

[11] *Растригин Л. А., Эренштейн Р. Х.* Коллективные правила распознавания. — М.: Энергия, 1981. — 244 с.

Структуры сходства в семействах алгоритмов классификации и оценки обобщающей способности*

Кочедыков Д. А.
dkochedykov@forecsys.ru
Москва, Вычислительный Центр РАН

Работа выполнена в рамках комбинаторного подхода к статистической теории обучения. Предложены два типа верхних оценок вероятности переобучения, по-разному учитывающих структуру сходства между алгоритмами в семействе.

Проблема обобщающей способности является одной из центральных в теории статистического обучения [1–6]. Обычно при решении задач обучения по прецедентам фиксируется некоторое *семейство алгоритмов* (функций классификации, регрессии, прогнозирования, и т. п.) и *метод обучения*, который по наблюдаемой обучающей выборке выбирает некоторый алгоритм из данного семейства. Классическим примером является метод *минимизации эмпирического риска* [6], который выбирает алгоритм, допускающий наименьшее число ошибок на обучающей выборке. Под *обобщающей способностью* метода обычно понимают его способность выбрать такой алгоритм, который будет допускать мало ошибок и на новых контрольных данных, неизвестных (скрытых) в момент обучения.

Алгоритм, у которого частота ошибок на контрольной выборке существенно выше (более чем на заданную величину ε) частоты ошибок на обучающей выборке, называется *переобученным*. Если метод обучения выбрал такой алгоритм, то говорят, что метод *переобучился*. Получение достаточно точных (не сильно завышенных) оценок вероятности переобучения остаётся в настоящее время открытой проблемой.

В классических работах Вапника-Червоненкиса [6] и многих более поздних подходах (см. обзоры [1, 2]) вероятность переобучения оценивается сверху исходя из *принципа равномерной сходимости* — как вероятность того, что хотя бы один из алгоритмов в семействе переобучен. Последняя, в свою очередь, оценивается сверху при помощи неравенства Буля (union bound) как сумма вероятностей переобучения по всем алгоритмам семейства. Применение неравенства Буля ведёт к сильно завышенным оценкам. Оно является точным только когда события «алгоритм переобучен» для всех алгоритмов в семействе независимы. Однако на практике чаще используются семейства, содержащие огромное количество схожих алгоритмов, для которых эти события существенно зависимы.

В данной работе вводится естественная метрика на множестве алгоритмов и рассматриваются порождаемые ею структуры сходства, позволяющие в явном виде учитывать эффекты *расслоения* и *сходства* в семействах алгоритмов [3, 4, 5] для получения более точных верхних оценок вероятности переобучения.

Понятие вероятности переобучения

Далее будем использовать понятия и обозначения, введенные в [3, стр. 18 в данном сборнике]:

X — обучающая выборка объектов;
$\bar{X}$ — контрольная выборка объектов;
$\mathbb{X} = X \sqcup \bar{X}$ — генеральная выборка объектов;
A — множество алгоритмов;
$I(a, x)$ — индикатор ошибки $a \in A$ на $x \in \mathbb{X}$;
$n(a, S)$ — число ошибок $a \in A$ на выборке $S \subseteq \mathbb{X}$;
$\nu(a, S)$ — частота ошибок $a \in A$ на $S \subseteq \mathbb{X}$;
$[\mathbb{X}]^l$ — множество всех выборок $X \subset \mathbb{X}$ длины l;
$\mu\colon [\mathbb{X}]^l \to A$ — метод обучения.

Будем говорить, что алгоритм $a \in A$ *переобучен* при разбиении $X \sqcup \bar{X} = \mathbb{X}$, если $\nu(a, \bar{X}) - \nu(a, X) \geqslant \varepsilon$.

Легко проверить, что условие переобученности равносильно $n(a, X) \leqslant \left\lfloor \frac{l}{L}(m - \varepsilon(L - l)) \right\rfloor \equiv s_m(\varepsilon)$, где $m = n(a, \mathbb{X})$ — число ошибок алгоритма a на генеральной выборке. Величину $s_m = s_m(\varepsilon)$ будем называть *порогом переобучения*.

Заметим, что s_m монотонно зависит от произвольно задаваемого параметра ε. В дальнейшем нам будет удобнее, наоборот, задавать значение s_m, и по нему определять ε.

Заметим также, что для порога переобучения справедливо $|s_m - s_{m-1}| \leqslant 1$ для всех $m = 1, \ldots, L$.

Пусть $\varphi(X)\colon [\mathbb{X}]^l \to \{\text{ложь}, \text{истина}\}$ — произвольный предикат на множестве выборок $[\mathbb{X}]^l$. Обозначим через $[\varphi(X)]$ индикатор, равный 1, если предикат $\varphi(X)$ истинен, и 0 иначе. Предполагая, что все разбиения $\mathbb{X} = X \sqcup \bar{X}$ равновероятны [3], будем определять вероятность события $\varphi(X)$ как

$$\mathsf{P}\left[\varphi(X)\right] = \frac{1}{C_L^\ell} \sum_{X \in [\mathbb{X}]^l} \left[\varphi(X)\right].$$

Введём предикат $U_a(X)$, обозначающий событие «алгоритм a переобучен на выборке X»:

$$U_a(X) \equiv \left[n(a, X) \leqslant s_{n(a,\mathbb{X})}\right].$$

*Работа поддержана РФФИ (проект №08-07-00422) и программой ОМН РАН «Алгебраические и комбинаторные методы математической кибернетики и информационные системы нового поколения».

Вероятность переобучения метода μ определяется как $Q_\mu = \mathsf{P}\left[U_{\mu(X)}\right]$. Функционал равномерной сходимости [6] определяется как вероятность того, что в семействе A существует переобученный алгоритм:

$$Q_A = \mathsf{P}\Big[\bigvee_{a\in A} U_a(X)\Big].$$

Он даёт верхнюю оценку вероятности переобучения: $Q_\mu \leqslant Q_A$ для любого метода обучения μ.

Заметим, что значение Q_A мало́, когда большинство алгоритмов в A переобучены приблизительно на одних и тех же выборках $X \in [\mathbb{X}]^l$, то есть за счет сходства алгоритмов. Получение достаточно точных оценок для Q_A невозможно без учёта эффекта сходства [5]. Получение точных оценок для Q_μ требует дополнительного учета свойств метода обучения μ и здесь не рассматривается.

Целью данной работы является выявления структурных характеристик семейства A, влияющих на вероятность переобучения, и получение достаточно точных верхних оценок для Q_A.

Оценка для одного алгоритма и оценка Вапника-Червоненкиса

Рассмотрим один алгоритм a с числом ошибок на генеральной выборке $n(a,\mathbb{X}) = m$. Вероятность того, что на обучающей выборке $X \in [\mathbb{X}]^l$ алгоритм будет иметь заданное число ошибок s определяется гипергеометрическим распределением:

$$\mathsf{P}\left[n(a,X) = s\right] = C_m^s C_{L-m}^{l-s}/C_L^l \equiv h_m(s);$$
$$\mathsf{P}\left[n(a,X) \leqslant s\right] = \sum_{t=0}^{s} h_m(t) \equiv H_m(s).$$

Выберем функцию порога переобучения в виде η-квантили распределения $H_m(s)$:

$$s_m = s_m(\eta) = \max\left\{s\colon H_m(s) \leqslant \eta\right\}. \qquad (1)$$

Тогда вероятность переобучения U_a есть

$$\mathsf{P}\left[U_a\right] = \mathsf{P}\left[n(a,X) \leqslant s_m(\eta)\right] \leqslant \eta.$$

Можно полагать $\mathsf{P}\left[U_a\right] \approx \eta$ с точностью до дискретности множества значений $H_m(s)$. Соответственно, параметр η в дальнейших оценках имеет смысл оценки $\mathsf{P}\left[U_a\right]$ для одного алгоритма a.

Оценка вероятности переобучения для конечного семейства, $|A| < \infty$, по Вапнику-Червоненкису [6] получается с помощью неравенства Буля как сумма оценок по всем алгоритмам в A:

$$Q_A = \mathsf{P}\Big[\bigvee_{a\in A} U_a\Big] \leqslant \sum_{a\in A} \mathsf{P}\left[U_a\right] \leqslant |A|\cdot\eta.$$

Связность и расслоение семейства

Определим расстояние между алгоритмами как хеммингово расстояние между их векторами ошибок: $\rho(a_1,a_2) = \operatorname{card}\left\{x \in \mathbb{X}\colon I(a_1,x) \neq I(a_2,x)\right\}$. Алгоритмы, расстояние между которыми равно 1, будем называть *смежными*.

Графом смежности семейства A назовем граф со множеством вершин A и множеством ребер $\{(a_1,a_2) \in A\times A\colon \rho(a_1,a_2) = 1\}$, соответствующих парам смежных алгоритмов.

Слоем семейства A назовем множество

$$A_m = \left\{a \in A\colon n(a,\mathbb{X}) = m\right\}, \quad m = 0,\ldots,L.$$

Набор чисел $D_m = |A_m|$, $m = 0,\ldots,L$ представляет *профиль расслоения* семейства A [3, 4, 7].

В [8] показано, что если $\mathbb{X}$ является независимой выборкой из некоторого распределения вероятностей, то при достаточно слабых требованиях к семейству алгоритмов и генерирующему распределению граф смежности получающегося семейства A оказывается связным с вероятностью 1. Будем называть множество A со связным графом смежности *связным* семейством.

Неравенства Бонферрони-Галамбоса

Вероятность Q_A может быть разложена с помощью формул включения-исключения:

$$Q_A = \mathsf{P}\Big[\bigvee_{a\in A} U_a\Big] = \sum_{j=1}^{|A|}\sum_{\substack{J\subset A\\|J|=j}} (-1)^{j-1}\,\mathsf{P}\Big[\prod_{a\in J} U_a\Big].$$

В силу знакопеременности данного ряда подсумма $\sum_{j=1}^{j_0}$ до четного/нечетного j_0 дает, соответственно, нижнюю/верхнюю оценку Q_A. Подсумма $\sum_{j=1}^{1}$ есть неравенство Буля. Верхние и нижние оценки, основанные на выборе определенных систем подмножеств J называются неравенствами *типа Бонферрони*. Большое число таких неравенств и методов их получения описано в [10].

В [9] показано, что если для произвольного множества событий $\{U_a,\ a \in A\}$ выбрать такое подмножество их пар $T = \{(a,a') \in A\times A\}$, чтобы граф (A,T) был деревом, то будет верна верхняя оценка

$$\mathsf{P}\Big[\bigvee_{a\in A} U_a\Big] \leqslant \sum_{a\in A}\mathsf{P}\left[U_a\right] - \sum_{\{a,a'\}\in T}\mathsf{P}\left[U_aU_{a'}\right]. \qquad (2)$$

Если семейство алгоритмов A — связное, то мы всегда можем выделить в его графе смежности дерево (A,T). Тогда в (2) имеем вероятности $\mathsf{P}\left[U_aU_{a'}\right]$ одновременного переобучения пар таких алгоритмов, что их векторы ошибок отличаются на одном объекте. Можно показать, что для таких пар $\mathsf{P}\left[U_aU_{a'}\right]$ равно либо $\mathsf{P}\left[U_a\right]$ либо $\mathsf{P}\left[U_{a'}\right]$, в зависимости от номера слоя.

Лемма 1. *Если алгоритмы* $a_m \in A_m$ *и* $a_{m+1} \in A_{m+1}$ *таковы, что* $\rho(a_m,a_{m+1}) = 1$, *то*

$$\left[U_mU_{m+1}\right] = \begin{cases}\left[U_m\right], & \text{если } s_m < s_{m+1};\\ \left[U_{m+1}\right], & \text{если } s_m = s_{m+1}.\end{cases}$$

Этот факт, в совокупности с тем, что число слагаемых во второй сумме (2) лишь на единицу меньше числа слагаемых в первой сумме, дает надежду на то, что оценка (2) для связных семейств может оказаться достаточно точной. Более пристальное рассмотрение показывает, что в суммах (2) взаимно сокращаются слагаемые, соответствующие алгоритмам лишь части слоев. Алгоритмы остальных слоев оставляют в оценке вклады размера $\big|\mathsf{P}\,[U_a] - \mathsf{P}\,[U_{a'}]\big|$, $\rho(a,a') = 1$. Эта величина оказывается достаточно велика в сравнении с вероятностью переобучения отдельного алгоритма, что дает в итоге оценку, лишь немного лучшую, чем оценка Вапника-Червоненкиса.

Теорема 2. *Для любых* $\mathbb{X}$, A *и* $\eta \in (0,1)$ *справедлива оценка*

$$Q_A \leqslant \eta + \sum_m D_m h_m(s_m(\eta)) \leqslant \\ \leqslant \eta + |A| \max_m h_m(s_m(\eta))\,.$$

Отметим, что данная оценка отличается от оценок типа Вапника-Червоненкиса [6, 4] только заменой левого «хвоста» гипергеометрического распределения $H_m(s_m)$ на значение гипергеометрической вероятности в точке $h_m(s_m)$.

Учет степени связности

Другой подход заключается в том, чтобы разложить вероятность Q_A по *цепному правилу*. Если задан произвольный порядок на множестве A, то

$$Q_A = \mathsf{P}\Big[\bigvee_{d=1}^{|A|} U_d\Big] = \sum_{d=1}^{|A|} \mathsf{P}\left[U_d \bar{U}_{d-1} \ldots \bar{U}_1\right]. \qquad (3)$$

В частном случае, если воспользоваться верхней оценкой $\mathsf{P}\left[U_d \bar{U}_{d-1} \ldots \bar{U}_1\right] \leqslant \mathsf{P}\,[U_d]$, то цепное правило переходит в неравенство Буля.

Упорядочим алгоритмы в A в порядке убывания номеров слоев: $A_L, A_{L-1}, \ldots, A_0$, причём порядок алгоритмов внутри одного слоя несущественен. Рассмотрим произвольный алгоритм $a \in A_m$ такой, что что у него есть некоторое число q смежных алгоритмов в слое A_{m+1}: $b_1, \ldots, b_q$. Тогда можно заметить, что условие U_a и условия $\bar{U}_{b_1}, \ldots, \bar{U}_{b_q}$ в определённом смысле противоречат друг другу, и, как следствие, вероятность $\mathsf{P}\left[U_a \bar{U}_{b_1} \ldots \bar{U}_{b_q}\right]$ должна быть достаточно мала.

Более того, оказывается, что для такой «верхней единичной окрестности» произвольного алгоритма a в графе смежности можно выписать точное значение вероятности $\mathsf{P}\left[U_a \bar{U}_{b_1} \ldots \bar{U}_{b_q}\right]$.

Лемма 3. *Пусть* $a \in A_m$, $b_d \in A_{m+1}$, $\rho(a, b_d) = 1$ *для всех* $d = 1, \ldots, q$, *и пусть* s_m *— произвольная функция порога переобучения. Тогда*

$$\mathsf{P}\left[U_a \bar{U}_{b_1} \ldots \bar{U}_{b_q}\right] = \begin{cases} 0, & s_{m+1} = s_m + 1; \\ \dfrac{C_m^{s_m} C_{L-m-q}^{l-s_m-q}}{C_L^l}, & s_{m+1} = s_m. \end{cases}$$

Пусть D_{mq} — число алгоритмов в m-ом слое $A_m \subseteq A$, имеющих ровно q смежных алгоритмов в $m{+}1$-ом слое. Будем называть D_{mq} *профилем расслоения и связности* [5].

Пусть N_0 — число алгоритмов, не имеющих смежных в последующих слоях.

Теорема 4. *Для любых* $\mathbb{X}$, A *и* $\eta \in (0,1)$ *справедлива оценка*

$$Q_A \leqslant \eta N_0 + \sum_{m,q} D_{mq} \left(\frac{l - s_m(\eta)}{L - m}\right)^q h_m(s_m(\eta))\,, \quad (4)$$

где сумма берётся только по тем слоям m, *для которых* $s_m(\eta) = s_{m-1}(\eta)$.

Сравнивая с оценкой теоремы 2, можно заметить, что учёт степени связности q даёт уменьшение оценки приблизительно на фактор α^q, где $\alpha = \frac{l - s_m}{L - m} < 1$, то есть, оценка уменьшается экспоненциально с увеличением степени связности q. Это вполне соответствует интуитивному представлению, что семейство с большей степенью связности должно иметь меньшую вероятность Q_A.

Эксперимент с семейством линейных классификаторов

Рассмотрим модельную задачу классификации с двумя классами — гауссовскими p-мерными распределениями с единичными ковариационными матрицами и расстоянием между центрами, равным 3. Число объектов каждого класса в $\mathbb{X}$ одинаково и равно $L/2$. При достаточно большом L такие классы линейно не разделимы.

Рассмотрим семейство линейных классификаторов $a(x) = \operatorname{sign}(w^{\intercal} x - b)$, $x, w \in \mathbb{R}^p$, $b \in \mathbb{R}$. Индикатор ошибки $I(a,x)$ определим как несовпадение ответа $a(x)$ с истинным классом объекта x. Тогда множество A есть множество различных векторов ошибок $\big(I(a,x_i)\big)_{i=1}^L$, порождаемых при всевозможных значениях параметров w, b.

Известна оценка $|A| \leqslant 2(C_L^0 + \cdots + C_L^p)$, см. [6]. Поскольку мощность A огромна, для оценки профиля D_{mq} используем случайную равномерную подвыборку алгоритмов $\tilde{A} \subset A$. Тогда несмещенная оценка профиля расслоения-связности

$$\hat{D}_{mq} = \tfrac{|A|}{|\tilde{A}|} \sum_{a \in \tilde{A}} \big[r^+(a,\mathbb{X}) = q\big]\big[n(a,\mathbb{X}) = m\big],$$

где $r^{\pm}(a,\mathbb{X}) = \#\big\{a' \in A_{m\pm1} \colon \rho(a,a') = 1\big\}$ — степень связности алгоритма a с последующим / предшествующим слоем семейства A.

На рис. 1 приведена оценка профилей расслоения $\hat{D}_m = \sum_q \hat{D}_{mq}$ и связности $\hat{D}_q = \sum_m \hat{D}_{mq}$. Заметим, что бóльшая часть алгоритмов семейства сконцентрирована возле частоты ошибок 0,5 и связности, немного превышающей размерность p.

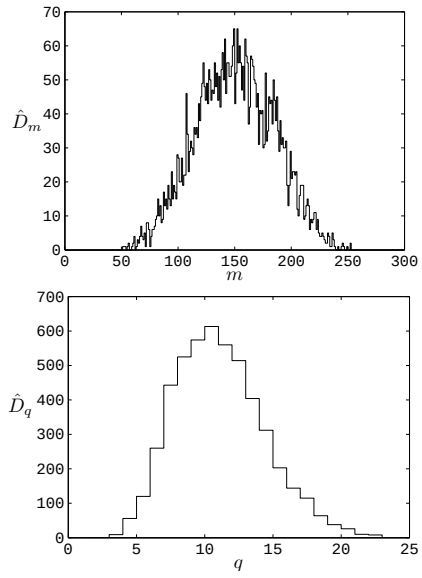


Рис. 1. Профили расслоения $\hat{D}_m$ и связности $\hat{D}_q$ семейства линейных классификаторов при $p = 8$, $|\mathbb{X}| = 300$, $|\tilde{A}| = 5 \cdot 10^3$.

Пусть $\bar{q}^{\pm} = \frac{1}{|\tilde{A}|} \sum_{a \in \tilde{A}} r^{\pm}(a, \mathbb{X})$ — оценка средней степени связности семейства. На рис. 2 приведены зависимости $\bar{q}^{+}$ и $\bar{q}^{-}$ от размерности пространства p. Они практически совпадают и близки к линейным, причём средняя связность растет немного быстрее, чем размерность.

Наконец, на рис. 3 показано отношение оценки Вапника-Червоненкиса $Q_{VC} = |\tilde{A}| \cdot \eta$ к оценке по профилю расслоения-связности (4), в зависимости от размерности пространства p. Заметим, что в (4) первое слагаемое оценивается нулём, так как в выборке $\tilde{A}$ нет алгоритмов, не имеющих связей со следующим слоем в A.

Относительная величина оценки (4) падает с ростом размерности p экспоненциально, немного быстрее, чем 2^{-p}. Уже при небольшой размерности (7–8 признаков) полученная оценка на 3 порядка лучше, чем оценка Вапника-Червоненкиса.

Заключение

В работе получены две верхние оценки вероятности переобучения, показывающие, что сходство алгоритмов внутри семейства приводит к улучшению обобщающей способности. Обе оценки представляют возможности для дальнейшего улучшения. Первая — путём учёта слагаемых более высоких порядков в неравенствах типа Бонферрони. Вторая — путём учёта большего числа условий с каждом члене цепного разложения. Пока остаются открытыми вопросы, какие ещё структурные характеристики графа связности семейства алгоритмов влияют на обобщающую способность, и какие графы связности имеют семейства алгоритмов, используемые на практике.

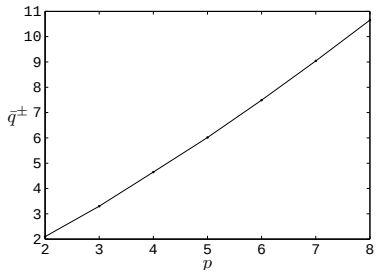


Рис. 2. Зависимость средней связности $\bar{q}$ семейства линейных классификаторов от размерности p.

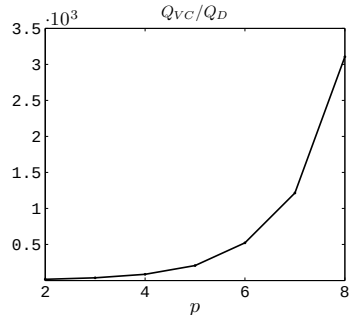


Рис. 3. Относительная завышенность оценки Вапника-Червоненкиса для линейных классификаторов, в зависимости от размерности p.

Литература

[1] *Boucheron S., Bousquet O., Lugosi G.* Theory of classification: A survey of some recent advances. — *ESAIM: Probability and Statistics.* — 2005. — Vol. 9 — Pp. 323–375.

[2] *Philips P.* Data-dependent analysis of learning algorithms: Ph.D. thesis / ANU, Canberra. — 2005.

[3] *Воронцов К. В.* Комбинаторный подход к проблеме переобучения // ММРО-14 — М.: МАКС Пресс, 2009 — С. 18–21 (в настоящем сборнике).

[4] *Vorontsov K. V.* Combinatorial probability and the tightness of generalization bounds // *Pattern Recognition and Image Analysis.* — 2008. — Vol. 18, no. 2. — Pp. 243–259.

[5] *Vorontsov K. V.* Splitting and similarity phenomena in the sets of classifiers and their effect on the probability of overfitting // *Pattern Recognition and Image Analysis.* — 2009. — Vol. 19, no. 3, Pp. 412–420.

[6] *Вапник В. Н., Червоненкис А. Я.* Теория распознавания образов. — М.: Наука, 1974.

[7] *Кочедыков Д. А.* Комбинаторные оценки обобщающей способности методов обучения по прецедентам с расслоением по наблюдаемой частоте ошибок // Труды 51-й научн. конф. МФТИ. — 2009.

[8] *Sill J.* Monotonicity and connectedness in learning systems: Ph.D. thesis / CalTech. — 1998.

[9] *Hunter D.* An upper bound for the probability of a union. — *J. Appl.Probab..* — 1976. — Vol. 13. — Pp. 597-603.

[10] *Galambos J., Simonelli I.* Bonferroni-type Inequalities with Applications. — Springer, 1996.

Исследование распределений расстояний точек евклидова пространства при случайных аффинных преобразованиях

Лясникова С. М., Жаркиx А. А.
LyasnikovaSM@yandex.ru
Мурманск, Мурманский Государственный Технический Университет

В работе исследуется зависимость положения конечного множества точек плоскости от действия некоторого аффинного преобразования. Исходное множество точек, как и множество точек, полученное после преобразования, представляется в виде точек из $\mathbb{R}^{2N}$. Изменение положения оценивается как расстояние в метрике l_2 (евклидовой метрике) между исходным и преобразованным множеством точек. Для некоторых частных случаев получены плотности распределения вероятностей и формулы для вычисления начальных моментов этих расстояний.

Конечные множества точек используются для описания различных объектов в физике, технике и социальных науках. Как частный случай можно представить изображение на экране компьютера, на телевизионном экране, на экране цифрового фотоаппарата и т. д. Во многих технических системах осуществляется измерение характеристик объекта, содержащего конечное число точек, в условиях каких-то случайных воздействий. К таким случайным воздействиям можно отнести случайные повороты, случайные отражения, случайные переносы и др. Такие воздействия могут существовать и в совокупности. В работе исследуется зависимость положения конечного множества точек плоскости от действия некоторого аффинного преобразования. Изменение положения оценивается как расстояние в евклидовой метрике между исходным и преобразованным множеством точек. В данной работе рассматриваются задачи вычисления плотности вероятности и начальных моментов указанных расстояний при случайных поворотах или отражениях. Рассматривается также случай, когда матрица преобразования имеет одинаково распределенные случайные компоненты. В перспективе предполагается решение аналогичных задач для случайных аффинных преобразований произвольного вида.

Множество N точек на плоскости как вектор пространства $\mathbb{R}^{2N}$

Во многих практических задачах наборы точек плоскости рассматриваются не обособлено, а как единый геометрический объект. Удобно рассматривать этот объект как элемент евклидова пространства $\mathbb{R}^{2N}$. Мы рассматриваем случай, когда координаты любой точки набора представляют собой две последовательно идущие координаты точки из $\mathbb{R}^{2N}$. При этом предполагается, что мы сохраняем соответствие между точками исходного объекта и преобразованного. Это позволит однозначно рассмотреть расстояния между исходным геометрическим объектом из $\mathbb{R}^{2N}$ и полученным в результате аффинного преобразования.

Аффинные преобразования на плоскости

Преобразование плоскости называется аффинным, если оно взаимно однозначно, и образом любой прямой является прямая. Аффинное преобразование на плоскости задаётся как линейное преобразование: $\boldsymbol{x'} = A\boldsymbol{x} + \boldsymbol{b}$, где $\boldsymbol{x} \in \mathbb{R}^2$ — исходная точка, $\boldsymbol{x'} \in \mathbb{R}^2$ — точка в преобразованной системе координат, $\boldsymbol{b} \in \mathbb{R}^2$ — вектор переноса, A — некоторая 2×2-матрица с определителем, отличным от нуля.

Случайные аффинные преобразования

В практических задачах случайными могут быть как элементы матрицы A, так и элементы вектора $\boldsymbol{b}$. В общем случае это шесть вещественных чисел, которые могут удовлетворять различным законам распределения. Очевидно, что в общем случае сложно получить вероятностные характеристики распределения расстояний между преобразованным и исходным геометрическими объектами. В данной работе мы рассматриваем следующие частные случаи аффинных преобразований.

1. Множество случайных поворотов относительно точки, задаваемой фиксированным вектором $\boldsymbol{b}$, $A = \begin{pmatrix} \cos\varphi & \sin\varphi \\ -\sin\varphi & \cos\varphi \end{pmatrix}$. Угол φ равномерно распределен в интервале $[0; 2\pi)$.
2. Множество случайных отражений относительно прямой, повернутой на угол φ, проходящей через фиксированную точку $\boldsymbol{b}$, $A = \begin{pmatrix} \cos 2\varphi & \sin 2\varphi \\ \sin 2\varphi & -\cos 2\varphi \end{pmatrix}$. Угол φ равномерно распределен в интервале $[0; 2\pi)$.
3. Вектор $\boldsymbol{b} = 0$, а элементы матрицы A являются независимыми случайными величинами с произвольными законами распределения.

Распределения расстояний при случайных поворотах и отражениях

Пусть имеется точка в $\mathbb{R}^{2N}$ $A = A_i(x_i, y_i)$, $i = 1, \ldots, N$. Координаты этой точки подвергаются одному из двух преобразований. Преобразование 1 — это поворот каждой проекции A на плоскости на один и тот же случайный угол φ отно-

сительно фиксированной точки (x_0, y_0). Преобразование 2 — это отражение тех же самых проекций на плоскости относительно оси, повернутой на случайный угол φ относительно точки (x_0, y_0).

Теорема 1. *В результате одного из двух преобразований получается новая точка в $\mathbb{R}^{2N}$. Случайный угол φ равномерно распределен в интервале $[0, 2\pi)$. Тогда плотность вероятности расстояний между исходной и полученной точкой определяется выражением:*

$$P_D(d) = \begin{cases} \dfrac{2X}{\pi\sqrt{Y^2 - Z^2}}, & \text{для } p_1 \leqslant d \leqslant p_2; \\ 0, & \text{в других случаях.} \end{cases}$$

В случае преобразования 1:

$$X = 1,\ Y = d_{\max},\ Z = d; \quad p_1 = 0,\ p_2 = Y,$$

где $d_{\max} = 2 \cdot \sum\limits_{i=1}^{N} \sqrt{(x_i - x_0)^2 + (y_i - y_0)^2}$ *— максимальное расстояние между исходным и преобразованным множествами точек.*

В случае преобразования 2:

$$X = d, \quad Y = E, \quad Z = d^2 - \frac{A + C}{2};$$

$$p_1 = \sqrt{\frac{A + C}{2} - E}, \quad p_2 = \sqrt{\frac{A + C}{2} + E},$$

$$E = \sqrt{B^2 + \left(\frac{C - A}{2}\right)^2},$$

$$A = 4\sum_{i=1}^{N}(x_i - x_0)^2, \quad C = 4\sum_{i=1}^{N}(y_i - y_0)^2,$$

$$B = 4\sum_{i=1}^{N}(x_i - x_0)(y_i - y_0).$$

Теорема 2. *Пусть выполняются условия Теоремы 1, тогда начальные моменты расстояния d:*

$$\bar{d^k} = T\sum_{r=0}^{m} \frac{\Gamma\left(\frac{k}{2} + 1\right)}{\Gamma\left(\frac{k}{2} - 2r + 1\right)(2r)!} \cdot \frac{C_{2r}^{\,r}\, H^{2r}}{U^{2r}\, 2^{2r}},$$

где горизонтальная черта над переменной d означает статистическое усреднение, $\Gamma(\cdot)$ — гамма-функция, r — счетчик.

$$m = \begin{cases} [\frac{n}{2}], & k = 2n; \\ \infty, & k = 2n{+}1. \end{cases}$$

Здесь n — любое натуральное число.
В случае преобразования 1:

$$T = \frac{d_{\max}^k}{2^{\frac{k}{2}}}, \quad U = H = 1.$$

В случае преобразования 2:

$$T = U = \left(\frac{A + C}{2}\right)^{\frac{k}{2}}, \quad H = \sqrt{B^2 + \left(\frac{C - A}{2}\right)^2}.$$

Теорема 3. *Пусть задана точка из $\mathbb{R}^{2N}$, множество координат этой точки разбивается на две произвольные части. Все разбиения равновероятны. Все части подвергаются одному из двух преобразований: преобразованию 1 или преобразованию 2. При преобразовании 1, одна часть подвергается повороту относительно фиксированной точки на случайный угол φ_1, а другая — на случайный угол φ_2. При преобразовании 2 одна часть подвергается отражению относительно прямой повернутой относительно фиксированной точки на случайный угол φ_1, а другая — на случайный угол φ_2. Случайные углы φ_1 и φ_2 равномерно распределены в интервале $[0, 2\pi)$. Тогда плотность вероятности вычисляется в виде свертки плотностей, определенных в Теореме 1:*

$$P_D(d) = \frac{d}{\pi}\int\limits_{-\infty}^{\infty} J_0(L_1)\, J_0(L_2)\, e^{-iwv}\, dw,$$

где $J_0(*)$ — функция Бесселя.

В случае преобразования 1:

$$L_1 = \frac{d_{\max 1}^2}{2}, \quad L_2 = \frac{d_{\max 2}^2}{2},$$

$$v = \frac{d_{\max 1}^2 \cdot \cos\varphi_1}{2} + \frac{d_{\max 2}^2 \cdot \cos\varphi_2}{2},$$

$$d_{\max 1} = 2 \cdot \sum_{i=1}^{N_1} \sqrt{(x_{i1} - x_{01})^2 + (y_{i1} - y_{01})^2},$$

$$d_{\max 2} = 2 \cdot \sum_{i=1}^{N_2} \sqrt{(x_{i2} - x_{02})^2 + (y_{i2} - y_{02})^2},$$

$$N_1 + N_2 = N.$$

То есть $d_{\max 1}$, $d_{\max 2}$ — это некоторые функции от координат 1-ой и 2-ой части соответственно.

В случае преобразования 2:

$$L_1 = \sqrt{B_1^2 + \left(\frac{C_1 - A_1}{2}\right)^2},\ L_2 = \sqrt{B_2^2 + \left(\frac{C_2 - A_2}{2}\right)^2},$$

$$v = \sqrt{B_1^2 + \left(\frac{C_1 - A_1}{2}\right)^2}\sin(2\varphi_1 + \beta_1) +$$

$$+ \sqrt{B_2^2 + \left(\frac{C_2 - A_2}{2}\right)^2}\sin(2\varphi_2 + \beta_2),$$

$$\beta_1 = \text{const}, \quad \beta_2 = \text{const}\,.$$

Теорема 4. *Пусть выполняются условия Теоремы 3, тогда начальные моменты расстояния d:*

$$\bar{d^k} = S^{\frac{k}{2}}\sum_{r=0}^{m}\left(\frac{1}{S^{2r}} \cdot \frac{\Gamma\left(\frac{k}{2} + 1\right)}{\Gamma\left(\frac{k}{2} - 2r + 1\right)(2r)!} \cdot \frac{(2n)!}{2^{2n}} \times \right.$$

$$\left. \times \sum_{r=0}^{n} \frac{1}{(r!)^2((n - r)!)^2}\, L_1^{2r}\, L_2^{2n-2r}\right),$$

где обозначения аналогичны обозначениям Теоремы 2.

В случае преобразования 1:

$$S = \frac{d^2_{\max 1} + d^2_{\max 2}}{2}.$$

В случае преобразования 2:

$$S = \frac{A_1 + A_2 + C_1 + C_2}{2}.$$

Преобразования с произвольным распределением компонент матрицы преобразования

Пусть имеется точка в $\mathbb{R}^{2N}$. Координаты этой точки подвергаются преобразованию, которое описывается матрицей $A = \left(\begin{smallmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{smallmatrix}\right)$. Предполагается, что элементы матрицы являются независимыми случайными величинами с произвольными законами распределений. Задача вычисления плотности вероятностей распределения расстояний является достаточно сложной и сводится к вычислению кратных несобственных интегралов, зависящих от параметра. В данной работе мы решили более простую задачу вычисления начальных моментов произвольного порядка. Квадрат расстояния между точками из $\mathbb{R}^{2N}$ при преобразовании, описываемом матрицей A, имеет вид:

$$d^2 = \sum_{i=1}^{N}\Big((F^2 + a_{21}^2)x_i^2 + \\ + 2x_i y_i(Fa_{12} + Pa_{21}) + \\ + (P^2 + a_{12}^2)y_i\Big),$$

где $F = a_{11} - 1,\ P = a_{22} - 1,\ \left(\begin{smallmatrix} x_i \\ y_i \end{smallmatrix}\right)$ — вектор исходных координат точки из $\mathbb{R}^{2N}$.

Пусть

$$T = F^2 + a_{21}^2,\ B = Fa_{12} + Pa_{21},\ C = P^2 + a_{12};$$
$$X = \sum_{i=1}^{N} x_i^2 = \text{const}, \quad Y = \sum_{i=1}^{N} x_i y_i = \text{const},$$
$$Z = \sum_{i=1}^{N} y_i^2 = \text{const}\,.$$

Тогда квадрат произвольного расстояния будет иметь вид $d^2 = TX + 2BY + CZ$. При вычислении начальных моментов любого порядка используются формулы полиномиального и биномиального разложения. Тогда k-й начальный момент есть

$$\bar{d}^k = \sum_{s_i} \frac{(k/2)!}{s_1!\, s_2!\, s_3!} \cdot X^{s_1}(2Y)^{s_2} Z^{s_3}\, \overline{T^{s_1} B^{s_2} C^{s_3}},$$

где горизонтальная черта над переменной d означает статистическое усреднение, суммирование производится по всем $s_1, s_2, s_3 \geqslant 0$ таким, что $s_1 + s_2 + s_3 = k/2$.

$$\overline{T^{s_1} B^{s_2} C^{s_3}} = \sum_{l_1=0}^{s_1}\sum_{l_2=0}^{s_2}\sum_{l_3=0}^{s_3} C_{s_1}^{l_1} C_{s_2}^{l_2} C_{s_3}^{l_3} \times \\ \times f^{(2l_1+l_2)} m_{21}^{(s_1-l_1+s_2-l_2)} p^{(s_2-l_2+2l_3)} m_{12}^{(l_2+s_3-l_3)},$$

где $f^{(j)}$, $p^{(j)}$, $m_{12}^{(j)}$, $m_{21}^{(j)}$ — j-е начальные моменты случайных величин $a_{11}-1$, $a_{22}-1$, a_{12}, a_{21}, соответственно.

Выводы

Авторы предполагают, что аналогичные задачи могут быть поставлены и решены для различных классов гильбертовых пространств и случайных линейных операторов, действующих в таких пространствах. Задача может найти применение в радиотехнических приложениях, в задачах медицинской диагностике, криминалистике и др. В приложениях может возникнуть задача сравнения исходного изображения с зашумленным, где зашумленное изображение — это исходное изображение, подвергнутое различным аффинным преобразованиям или совокупности аффинных преобразований. Рассмотренные задачи и последующая разработка общей теории таких преобразований позволит определить, является ли зашумленное изображение исходным изображением с шумом или это совершенно другое изображение.

Литература

[1] *Румшиский Л. З.* Элементы теории вероятностей: учебн. пособие для вузов — М.: Наука, 1976. — 240 с.

[2] *Кудрявцев Л. Д.* Краткий курс математического анализа, Т.2 — Висагинас: Alfa, 1998. — 384 с

[3] *Лясникова С. М.* Вероятностные характеристики расстояний между точками евклидова пространства, отличающимися случайными поворотами или отражениями // Докл. XV Межд. конф. студентов, аспирантов и молодых ученых «Ломоносов», М.: Издательство МГУ, 2008. — С. 29–30.

Непрерывное обобщение информационного критерия Акаике для оценивания нестационарной регрессионной модели временного ряда с неизвестной степенью изменчивости коэффициентов*

Моттль В. В., Красоткина О. В., Ежова Е. О.
vmottl@yandex.ru, ko189177@yandex.ru, lena-ezhova@rambler.ru
Москва, ВЦ РАН; Тула, ТулГУ; Долгопрудный, МФТИ

Применение информационного критерия Акаике (AIC) для выбора класса модели из упорядоченного множества вложенных классов моделей ограничено предположением, что классы определяются возрастающей размерностью вектора параметров. Мы распространили принцип максимума информации по Кульбаку, лежащий в основе классического информационного критерия Акаике, на более широкий класс моделей, в котором размерность вектора параметров фиксирована, но свобода выбора его значений ограничена системой непрерывно вложенных семейств априорных плотностей распределения. Мы проиллюстрировали применение обобщенного критерия Акаике на задаче оценивания нестационарной линейной регрессионной модели временного ряда с неизвестной степенью изменчивости коэффициентов.

Широко используемый в современном анализе данных информационный критерий Акаике (AIC) [1] является простым и эффективным способом выбора наиболее адекватного класса модели из упорядоченного дискретного множества вложенных классов моделей.

В классической постановке критерия обычно рассматривается выборка $\mathbf{y} = (y_j,\ j = 1,\ldots,N)$ независимых случайных величин с неизвестной плотностью распределения $\varphi^*(y)$, принадлежащей некоторому параметрическому семейству $\varphi(y\,|\,\mathbf{c})$, $\mathbf{c} \in \mathbb{R}^m$. Часто размерность вектора параметров m оказывается очень большой и существенно превосходит размер обучающей выборки N, что делает бессмысленным применение для оценивания вектора параметров $\mathbf{c}$ принципа максимального правдоподобия

$$\hat{\mathbf{c}}(\mathbf{y}) = \arg\max_{\mathbf{c}} \ln \Phi(\mathbf{y}\,|\,\mathbf{c}),$$
$$\ln \Phi(\mathbf{y}\,|\,\mathbf{c}) = \sum_{j=1}^{N} \ln \varphi(y_j\,|\,\mathbf{c}). \qquad (1)$$

Если же предположить, что элементы вектора $\mathbf{c}$ обладают естественной упорядоченностью по степени значимости, и при этом $c_i = 0,\ n < i \leqslant m$:

$$\mathbf{c} = (\mathbf{c}_n, \mathbf{c}_{m-n}),\ \mathbf{c}_n \in \mathbb{R}^n,\ \mathbf{c}_{m-n} = \mathbf{0} \in \mathbb{R}^{m-n}, \quad (2)$$

то это позволит нам рассмотреть параметрическое семейство $\Phi(\mathbf{y}\,|\,\mathbf{c})$ как последовательность вложенных классов моделей $\Phi\big(\mathbf{y}\,|\,\mathbf{c}{=}(\mathbf{c}_n,\mathbf{0})\big)$ размерности $n = n_{\min},\ldots,n_{\max}$.

Критерий AIC в классической постановке является способом оценивания подходящей размерности вектора параметров, как меры сложности модели $\hat{n} = \arg\max_n \big(\ln \Phi\big(\mathbf{y}\,|\,\mathbf{c}_n(\mathbf{y}),\mathbf{0}\big) - n\big)$. Однако это формула получена в предположении, что гессиан $\nabla^2_{\mathbf{c}_n\mathbf{c}_n} \ln \Phi(\mathbf{y}\,|\,\mathbf{c}_n,\mathbf{0})$ в точке максимального правдоподобия имеет полный ранг, а значит и оценка $\hat{\mathbf{c}}_n(\mathbf{y})$ единственная. В более общем случае заменим штраф n на ранг матрицы

$$\hat{n} = \arg\max_n \Big\{ \ln \Phi\big(\mathbf{y}\,|\,\hat{\mathbf{c}}_n(\mathbf{y}),\mathbf{0}\big) - \\ - \operatorname{rank}\big[\nabla^2_{\mathbf{c}_n\mathbf{c}_n} \ln \Phi(\mathbf{y}\,|\,\hat{\mathbf{c}}_n(\mathbf{y}),\mathbf{0})\big]\Big\}. \quad (3)$$

В основе классического AIC лежит принцип максимизации информации по Кульбаку между моделью плотности распределения и настоящей гипотетической плотностью распределения.

$$n^* = \arg\max_n \int \big[\ln \Phi(\mathbf{y}\,|\,n,\mathbf{c}^*_n)\big]\,\Phi^*(\mathbf{y})d\mathbf{y} \qquad (4)$$

есть желаемая размерность в предположении, что $\Phi^*(\mathbf{y}) = \Phi(\mathbf{y}\,|\,\mathbf{c}^*_{n^*})$ с некоторым значением $(\mathbf{c}^*_{n^*}\mathbf{0})$, вырезанным из неизвестного $\mathbf{c}^* = (c^*_1,\ldots,c^*_m)$

Одним из первых применений AIC было моделирование нестационарного сигнала на дискретной временной оси, разделенной на неизвестное количество n интервальных блоков, и проверка локальной стационарности модели авторегрессии с фиксированным порядком k на каждом их них [2].

Со времен публикации первой статьи Акаике было предложено много модификаций этого критерия [3, 4, 5, 6]. Среди них байесовский информационный критерий BIC [3] нашел более широкое применение. Однако все они были нацелены на выбор размерности вектора параметров для случая известной упорядоченности его элементов по степени значимости.

В данной работе предлагается совершенно новое обобщение критерия Акаике, которое было вызвано необходимостью анализа нестационарного сигнала $(\mathbf{y},\mathbf{x}) = \big((y_t,\mathbf{x}_t),\ t = 1,\ldots,N\big)$, регрессионная модель которого

$$y_t = \mathbf{c}_t^{\intercal}\mathbf{x}_t + \eta_t, \quad \mathbf{c}_t,\ \mathbf{x}_t \in \mathbb{R}^k,$$
$$\eta_t \sim \mathcal{N}(\eta_t\,|\,0,\delta), \quad \mathsf{E}(\eta_t,\eta_s) = 0, \qquad (5)$$

меняется на интервале наблюдения. Очевидно, что при этом размерность вектора параметров в семействе условных плотностей распределения $\Phi(\mathbf{y}\,|\,\mathbf{x},\mathbf{c})$

*Работа выполнена при финансовой поддержке РФФИ, проекты №№ 09-07-00394, 08-01-00695, 08-01-12023, 08-07-90700.

оказывается фиксированной $\mathbf{c}{=}(\mathbf{c}_1,\ldots,\mathbf{c}_N)\in\mathbb{R}^{kN}$ и в k раз превосходит количество наблюдений. Вместо этого предполагается, что искомая последовательность коэффициентов представляет собой случайный марковский процесс

$$\mathbf{c}_t=\mathbf{c}_{t-1}+\xi_t,\ \xi_t\sim\mathcal{N}\left(\xi\,|\,\mathbf{0},\lambda\delta\mathbf{I}\right),\ \mathsf{E}\left(\xi_t\xi_s^{\mathsf{T}}\right)=\mathbf{0},\quad(6)$$

который начинается с неизвестного первого значения $\mathbf{c}_1\sim\mathcal{N}(\mathbf{c}_1\,|\,\mathbf{0},\rho\mathbf{I})$, $\rho\to\infty$. Параметр дисперсии шума λ является структурным параметром априорной модели и отвечает за степень временной нестационарности коэффициентов регрессии.

Это типичный пример задачи, в которой плавное изменение параметра λ определяет систему непрерывно вложенных априорных плотностей распределения $\Psi(\mathbf{c}\,|\,\lambda)$ вектора параметров модели, начиная от «однородного» распределения в $\mathbb{R}^k$ при $\lambda=0$ до «однородного» распределения в $\mathbb{R}^{kN}$ при $\lambda\to\infty$. Такая ситуация фактически представляет собой введение вместо дискретной последовательности целочисленных размерностей понятия «размытой размерности» вектора параметров $\mathbf{c}$, непрерывно меняющейся от k до kN при увеличении параметра λ. Естественно, что классический критерий AIC оказывается неприменимым для выбора наиболее подходящего для данного сигнала $(\mathbf{y},\mathbf{x})$ значения параметра $0<\lambda<\infty$.

В этой статье мы рассматриваем параметрическую модель плотности распределения неизвестной генеральной совокупности $F^*(\mathbf{y})$ как смесь условной плотности из заданного семейства $\Phi(\mathbf{y}\,|\,\mathbf{c})$, $\mathbf{c}\in\mathbb{R}^m$ и априорной плотности распределения вектора параметров $\Psi(\mathbf{c}\,|\,\lambda)$:

$$F(\mathbf{y}\,|\,\lambda)=\int\Phi(\mathbf{y}\,|\,\mathbf{c})\Psi(\mathbf{c}\,|\,\lambda)\,d\mathbf{c},\quad\mathbf{c}\in\mathbb{R}^m.\qquad(7)$$

Значение структурного параметра модели λ, оцененное по наблюдаемой выборке $\mathbf{y}$, обеспечивает оптимальную степень сокращения слишком большой размерности вектора параметров $\mathbf{c}$. Как только значение λ выбрано, результат анализа представляет собой байесовскую оценку вектора параметров $\mathbf{c}$

$$\hat{\mathbf{c}}_\lambda(\mathbf{y})=\arg\max_{\mathbf{c}}\big(\ln\Phi(\mathbf{y}\,|\,\mathbf{c})+\ln\Psi(\mathbf{c}\,|\,\lambda)\big).\qquad(8)$$

Мы будем эксплуатировать ту же идею, что и в (4), т. е. будем с помощью варьирования параметра λ пытаться обеспечить наилучшее приближение модельного распределения $F(\mathbf{y}\,|\,\lambda)$ (7) и неизвестного распределения генеральной совокупности $F^*(\mathbf{y})$. В частности, когда структурный параметр модели принимает целые значения $0\leqslant\lambda\leqslant m$, классический критерий Акаике получается из обобщенной непрерывной версии с помощью выбора соответствующей априорной плотности распределения $\Psi(\mathbf{c}\,|\,\lambda)$.

Мы проиллюстрируем применение обобщенного критерия Акаике на задаче оценивания нестационарной линейной регрессионной модели временного ряда с неизвестной степенью изменчивости коэффициентов.

Основные предположения о семействах параметрических плотностей

Предположения. Мы ограничимся здесь рассмотрением случая параметрических семейств плотностей распределения $\varphi(y\,|\,\mathbf{c})$, для которых логарифмическая функция правдоподобия $\ln\Phi(\mathbf{y}\,|\,\mathbf{c})$ асимптотически квадратична в окрестности оценки максимального правдоподобия $\mathbf{c}$, то есть для достаточно большого размера N выборки $\mathbf{y}=(y_j,\ j=1,\ldots,N)$ можно считать, что

$$\ln\Phi(\mathbf{y}|\mathbf{c})=\ln\Phi\big(\mathbf{y}|\hat{\mathbf{c}}(\mathbf{y})\big)+\tfrac{1}{2}\big(\mathbf{c}-\hat{\mathbf{c}}(\mathbf{y})\big)^{\mathsf{T}}\mathbf{A}\big(\mathbf{c}-\hat{\mathbf{c}}(\mathbf{y})\big),$$
$$\nabla_{\mathbf{c}}\log\Phi(\mathbf{y}\,|\,\mathbf{c})=\mathbf{A}\big(\mathbf{c}-\hat{\mathbf{c}}(\mathbf{y})\big).\qquad(9)$$

Причем гессиан $\mathbf{A}=\nabla^2_{\mathbf{cc}}\ln\Phi(\mathbf{y}\,|\,\mathbf{c})$, называемый информационной матрицей Фишера, не зависит от точки $\mathbf{c}$, в которой определен.

Рассмотрим теперь семейство плотностей априорного распределения скрытой переменной $\Psi(\mathbf{c}\,|\,\lambda)$. Будем полагать, что каждая из этих плотностей является нормальной, возможно вырожденной, с нулевым математическим ожиданием и ковариационной матрицей, определяемой значением структурного параметра λ. Это приводит к тому, что логарифмическая функция правдоподобия $\ln\Psi(\mathbf{c}\,|\,\lambda)$ есть квадратичная функция, достигающая своего максимального значения в нуле $\nabla_{\mathbf{c}}\ln\Psi(\mathbf{0}\,|\,\lambda)=\mathbf{0}$ и определяемая своим Гессианом $\mathbf{D}_\lambda=\nabla^2_{\mathbf{cc}}\ln\Psi(\mathbf{c}\,|\,\lambda)$, так что

$$\begin{aligned}\ln\Psi(\mathbf{c}\,|\,\lambda)&=\mathrm{const}_\lambda+\tfrac{1}{2}\mathbf{c}^{\mathsf{T}}\mathbf{D}_\lambda\mathbf{c},\\ \nabla_{\mathbf{c}}\ln\Psi(\mathbf{c}\,|\,\lambda)&=\mathbf{D}_\lambda\mathbf{c}.\end{aligned}\qquad(10)$$

Что касается неизвестной плотности распределения выходной переменной $F^*(\mathbf{y})$, то мы будем предполагать, что оно согласуется с семейством плотностей $\Phi(\mathbf{y}\,|\,\mathbf{c})$ в том смысле, что существует неизвестная плотность $\Psi^*(\mathbf{c})$, которая допускает представление

$$F^*(\mathbf{y})=\int\Phi(\mathbf{y}\,|\,\mathbf{c})\Psi^*(\mathbf{c})\,d\mathbf{c}.\qquad(11)$$

Принцип максимума информации по Кульбаку о распределении наблюдаемой переменной

Принцип максимизации по Кульбаку схожести модельного распределения $F(\mathbf{y}\,|\,\lambda)$ и распределения генеральной совокупности $F^*(\mathbf{y})$ приводит к такому критерию:

$$\lambda^*=\arg\max_\lambda\int\big[\ln F(\mathbf{y}\,|\,\lambda)\big]F^*(\mathbf{y})\,d\mathbf{y}.\qquad(12)$$

Однако непосредственная реализация критерия (12) невозможна хотя бы потому, что истинное распределение $F^*(\mathbf{y})$ неизвестно. Максимизация функции правдоподобия по одной доступной реализации $\ln F(\mathbf{y}\,|\,\lambda)$, как несмещенной оценки критерия, также бессмысленна, так как при этом будут предпочтительны значения структурного параметра, приводящие к слишком большим размерностям $\mathbf{c}\in\mathbb{R}^m$.

Для того, чтобы преодолеть «проклятие единственной выборки», мы применим идею компромисса, обосновывающего классический информационный критерий Акаике [1], а именно, вообразим существование другой независимой выборки $\tilde{\mathbf{y}}$. Пусть по ней получена произвольная байесовская оценка $\hat{\mathbf{c}}_\lambda(\tilde{\mathbf{y}})$ (8). Заменим $\ln F(\mathbf{y}\,|\,\lambda)$ в (12) на математическое ожидание $\ln\Phi(\mathbf{y}\,|\,\hat{\mathbf{c}}_\lambda(\tilde{\mathbf{y}}))$:

$$\hat{\lambda}=\arg\max_\lambda\int\Big\{\int\Big\{\int\big[\ln\Phi(\mathbf{y}\,|\,\hat{\mathbf{c}}_\lambda(\tilde{\mathbf{y}}))\big]\Phi(\tilde{\mathbf{y}}\,|\,\mathbf{c})\,d\tilde{\mathbf{y}}\Big\}\times\\ \times\,\Phi(\mathbf{y}\,|\,\mathbf{c})\,d\mathbf{y}\Big\}\Psi^*(\mathbf{c})\,d\mathbf{c}. \quad (13)$$

Теорема 1. *При предположениях* (9) *и* (10),

$$\int\Big\{\int\Big\{\int\big[\ln\Phi(\mathbf{y}\,|\,\hat{\mathbf{c}}_\lambda(\tilde{\mathbf{y}}))\big]\Phi(\tilde{\mathbf{y}}\,|\,\mathbf{c})\,d\tilde{\mathbf{y}}\Big\}\times\\ \times\,\Phi(\mathbf{y}\,|\,\mathbf{c})\,d\mathbf{y}\Big\}\Psi^*(\mathbf{c})\,d\mathbf{c}=\int J(\lambda\,|\,\mathbf{y})F^*(\mathbf{y})\,d\mathbf{y},$$

$$J(\lambda\,|\,\mathbf{y})=\ln\Phi\big(\mathbf{y}\,|\,\hat{\mathbf{c}}_\lambda(\mathbf{y})\big)-\mathrm{tr}\big(\mathbf{A}(\mathbf{A}+\mathbf{D}_\lambda)^{-1}\big). \quad (14)$$

Эта теорема указывает на построение непрерывного аналога классического AIC. Хотя распределение $\Psi^*(\mathbf{c})$ в (9) по-прежнему неизвестно, а значит непосредственно применить критерий (13) невозможно, но выражение (14) дает легко вычисляемую функцию $J(\lambda\,|\,\mathbf{y})$, которая является несмещенной оценкой полного критерия. Эту функцию можно также максимизировать по искомому значению структурного параметра:

$$\hat{\lambda}(\mathbf{y})=\arg\max_\lambda J(\lambda\,|\,\mathbf{y})=\\ =\arg\max_\lambda\Big(\ln\Phi\big(\mathbf{y}|\hat{\mathbf{c}}_\lambda(\mathbf{y})\big)-\mathrm{tr}\big(\mathbf{A}(\mathbf{A}+\mathbf{D}_\lambda)^{-1}\big)\Big). \quad (15)$$

Это и есть обобщенный информационный критерий Акаике (3). Сравнение критериев (15) и (3) позволяет интерпретировать штрафной член $\mathrm{tr}\big(\mathbf{A}(\mathbf{A}+\mathbf{D}_\lambda)^{-1}\big)$, как условную «размытую размерность» параметра $\mathbf{c}$, выбор которого ограничено распределением $\ln\Psi(\mathbf{c}\,|\,\lambda)$.

Частный случай: классический информационный критерий Акаике

Пусть структурный параметр принимает целые положительные числа $0\leqslant\lambda\leqslant m$ и урезает вектор параметров с упорядоченными элементами $\mathbf{c}=(\mathbf{c}_\lambda,\mathbf{c}_{m-\lambda})\in\mathbb{R}^m$, так же как и в (2) с $n=\lambda$, то есть $\mathbf{c}_\lambda\in\mathbb{R}^\lambda$, $\mathbf{c}_{m-\lambda}\in\mathbb{R}^{m-\lambda}$. Никакой априорной информации о векторе $\mathbf{c}$ нет, то есть $\Psi(\mathbf{c}_\lambda\,|\,\lambda)=\prod_{i=1}^{\lambda}\psi_i(c_i)$, $\psi_i(c_i)=\mathcal{N}(c_i\,|\,0,\sigma^2)$, $\sigma^2\to\infty$, $\Psi(\mathbf{c}_\lambda\,|\,y\lambda)\cong\mathrm{const}=0$, $\ln\Psi(\mathbf{c}_\lambda\,|\,\lambda)\cong\cong\mathrm{const}\ll 0$. Так как только первая часть вектора параметров входит в условную плотность $\Phi(\mathbf{y}\,|\,\mathbf{c}_\lambda,\mathbf{c}_{m-\lambda})$, то Гессиан $\mathbf{A}_\lambda=\nabla^2_{\mathbf{c}_\lambda\mathbf{c}_\lambda}\ln\Phi(\mathbf{y}\,|\,\mathbf{c}_\lambda,\mathbf{0})$ есть матрица размера $\lambda\times\lambda$.

При принятых предположениях (15) приводит к критерию (3):

$$\max_{\mathbf{c}_\lambda}\ln\Phi(\mathbf{y}\,|\,\mathbf{c}_\lambda,\mathbf{0})-\mathrm{rank}\,(\mathbf{A}_\lambda)\to\max_\lambda.$$

Применение критерия Акаике в задаче оценивания нестационарной регрессии: модельные эксперименты

В задаче оценки нестационарной регрессии (5)–(6) байесовская оценка скрытой последовательности коэффициентов регрессии $\mathbf{c}=(\mathbf{c}_1,\ldots,\mathbf{c}_N)\in\in\mathbb{R}^{kN}$ зависит только от отношения λ предполагаемых дисперсий шума в уравнениях наблюдения (5) и состояния (6), в то же время, ее статистические свойства существенно определяются дисперсией шума в модели наблюдения. Байесовская оценка вектора параметров $\mathbf{c}$ может быть получена минимизацией критерия Flexible Least Squares

$$\hat{\mathbf{c}}_\lambda(\mathbf{y})=\arg\min\Big(\sum_{t=1}^{N}\big(y_t-\mathbf{x}_t^\intercal\mathbf{c}_t\big)^2+\\ +\frac{1}{\lambda}\sum_{t=2}^{N}(\mathbf{c}_t-\mathbf{c}_{t-1})^\intercal(\mathbf{c}_t-\mathbf{c}_{t-1})\Big) \quad (16)$$

с помощью фильтра-интерполятора Калмана-Бьюси [8].

Представим модель в явной форме.

Будем полагать, что $\mathbf{y}=(y_1,\ldots,y_N)^\intercal\in\mathbb{R}^N$ и $\mathbf{c}=(\mathbf{c}_1^\intercal,\ldots,\mathbf{c}_N^\intercal)^\intercal\in\mathbb{R}^{kN}$ есть вектор-столбцы, $\mathbf{X}=(\mathbf{X}_{ts},\ t=1,\ldots,N)$ есть блочная матрица размера $kN\times N$ с блоками $\mathbf{X}_{ts}=(\mathbf{x}_t,t\neq s)$ размера $k\times 1$, $\mathbf{D}_\lambda=-\frac{1}{\lambda}\mathbf{B}$, где $\mathbf{B}$ есть квадратная блочно-трехдиагональная матрица размера $kN\times kN$ с диагональю $\big(\mathbf{I},2\mathbf{I},\ldots,2\mathbf{I},\mathbf{I}\big)$ и не диагоналями $\big(-\mathbf{I},\ldots,-\mathbf{I}\big)$, где $\mathbf{I}$ есть единичная матрица размера $k\times k$. Положим также дисперсию наблюдаемого шума равной единице $\delta=1$, тогда модель (5) будет давать функцию максимального правдоподобия

$$\ln\Phi(\mathbf{y}|\mathbf{c},\mathbf{X})=\ln\mathcal{N}(\mathbf{y}|\mathbf{X}^\intercal\mathbf{c},\mathbf{I})=\mathrm{const}+\tfrac{1}{2}\mathbf{c}^\intercal\mathbf{A}\mathbf{c},$$

гессиан которой $\mathbf{A}_N=-\mathbf{X}\mathbf{X}^\intercal$ размера $kN\times kN$ всегда вырожден и, если регрессоры $(x_{it},\ t=1,\ldots,N)$ линейно независимы, имеет максимальный ранг

$\operatorname{rank}\mathbf{A} = N$. Скрытая марковская модель коэффициентов регрессии (6) выражается семейством априорных плотностей распределения

$$\ln\Psi(\mathbf{y}\,|\,\lambda) = \ln\mathcal{N}(\mathbf{c}\,|\,\mathbf{0}, \lambda\mathbf{B}^{-1}) = \mathrm{const}_\lambda - \tfrac{1}{2}\lambda\mathbf{c}^\intercal\mathbf{B}\mathbf{c}.$$

Мы проанализировали 200 независимых реализаций случайного процесса (5) длинной $N = 50$, заданного линейной комбинацией трех регрессоров $(x_{it},\ t = 1,\ldots,N),\ i = 1,\ldots,k,\ k = 3$, представляющих собой случайный белый шум с нулевым средним, с коэффициентами регрессии, заданными синусоидальными последовательностями $c^*_{it} = 4\sin\big((2\pi/N)t + (2\pi/3)(i-1)\big)$, смещенными друг относительно друга по фазе 10%. Дисперсия шума в модели наблюдения составляла 10%, $\delta = 0.1\frac{1}{N}\sum\limits_{i=1}^{N}(\mathbf{x}_t^\intercal\mathbf{c}_t)^2$.

«Эффективная размерность» последовательности коэффициентов регрессии $(\mathbf{c}_1,\ldots,\mathbf{c}_N)$ равна числу регрессоров в случае нулевой дисперсии $\lambda \to 0$ и достигает длины временных рядов при $\lambda \to \infty$. Если последовательность коэффициентов регрессии фиксирована, то случайна лишь последовательность наблюдений, а значит и логарифм условной плотности распределения в байесовской оценке есть также случайная функция. Опыт показывает, что реализация случайной функции J критерия имеет единственную точку максимума. На Рис.1 представлен типичный график полученного критерия. Строгая унимодальность функции J критерия позволяет применять для нахождения точки максимума любой одномерный оптимизационный метод.

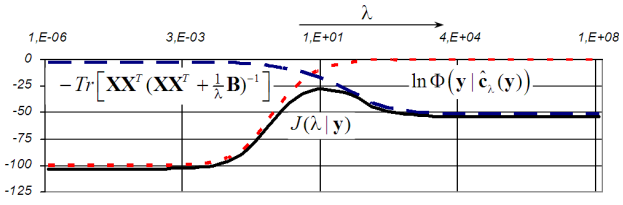


Рис. 1. Графики одной из реализаций случайной функции $J(\lambda\,|\,\mathbf{y})$ и ее компонент.

Для каждого из 200 смоделированных временных рядов были вычислены 2 значения параметра дисперсии $\hat\lambda$, во-первых, по принципу полученного обобщенного критерия Акаике (15), во-вторых, традиционным методом скользящего контроля [8]. Затем мы применили каждое из полученных значений к оставшимся 199 временным рядам, как к контрольному множеству, и сравнили истинную последовательность коэффициентов регрессии $(\mathbf{c}^*_1,\ldots,\mathbf{c}^*_N)$ с полученной оценкой $(\hat{\mathbf{c}}_{1,\hat\lambda},\ldots,\hat{\mathbf{c}}_{N,\hat\lambda})$ по критерию

$$\varepsilon_{\hat\lambda} = \sum_{t=1}^{N}(\hat{\mathbf{c}}_{t,\hat\lambda} - \mathbf{c}^*_t)^\intercal(\hat{\mathbf{c}}_{t,\hat\lambda} - \mathbf{c}^*_t)\Big/\sum_{t=1}^{N}(\mathbf{c}^*_t)^\intercal\mathbf{c}^*_t.$$

Мы получили следующие результаты:

критерий	$\varepsilon_{\hat\lambda}$	
	марковская модель	синусоидальные коэффициенты
обобщенный AIC	0,016	0,021
скольз. контроль	0,046	0,015

Выводы

Можно сделать вывод, что непрерывная версия критерия Акаике гораздо более чувствительна к выбору априорной модели процесса изменения коэффициентов регрессии, нежели критерий скользящего контроля. Поэтому, если априорная модель адекватна фактическому характеру изменения коэффициентов регрессии, то критерий Акаике предпочтителен.

Литература

[1] *Akaike H.* A new look at the statistical model identification // IEEE Trans. on Automatic Control, Vol. IC-19, No. 6, 1974, Pp. 716–723.

[2] *Kitagawa G., Akaike H.* A procedure for the modeling of no-stationary time series // Ann. Inst. Statist. Math., Vol. 30, Part B, 1987, Pp. 351–363.

[3] *Scharz G.* Estimating the dimtnsion of the model // The Annals of Statistics, Vol. 6, No. 2, 1978, Pp. 461–464.

[4] *Bozdogan H.* Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions // Psychometrica, Vol. 52, No. 3, 1987.

[5] *Spiegelhalter D., Best N., Carlin B. Van der Linde A.* Bayesian mesures of model complexity and fit // Journal of the Royal Statistical Society. Series B (Statistical Methodology), Vol. 64, No. 4, 2002, Pp. 583–639.

[6] *Rodrigues C. C.* The ABC of model selection: AIC, BIC and new CIC // AIP Conference Proceedings, Vol. 803, 2005, Pp. 80–87.

[7] *Markov M., Krasotkina O., Mottl V., Muchnik I.* Time-varying regression model with unknown time-volatility for nonstationary signal analyses // Proc. of the 8th IASTED Internation Conference on Signal and Image Processing. Honolulu, Hawaii, USA, August 14–16, 2006.

[8] *Markov M., Muchnik I., Mottl V., Krasotkina O.* Dynamic analysis of hedge funds // Proc. of the 3rd IASTED Internation Conference on Financal Engineering and Applications. Cambridge, Massachusetts, USA, October 9–11, 2006.

[9] *Bishop C. M.* Pattern Recognition and Machine Learning. Springer, 2006.

О точности интервальных оценок вероятности ошибочной классификации, основанных на эмпирическом риске*

Неделько В. М.
nedelko@math.nsc.ru
Новосибирск, Институт математики СО РАН

Работа посвящена проблеме оценивания риска в задаче классификации. Рассмотрен ряд примеров, которые позволяют проанализировать поведение сложностных оценок. Приведены асимптотически точные оценки риска для гистограммного классификатора при фиксированном отношении объёма выборки к сложности класса решающих функций.

Проблема построения достаточно точных оценок вероятности ошибочной классификации до настоящего времени остается открытой. Известно, что оценки Вапника-Червоненкиса [1] не могут быть существенно улучшены, если использовать только ёмкостную характеристику класса решающих функций. Актуальным является выявление характеристик множества классификаторов и метода обучения, которые бы позволили получать более точные оценки риска.

Постановка задачи

Пусть X — пространство значений переменных, используемых для прогноза, а $Y = \{0, 1\}$ — пространство значений прогнозируемых переменных, и пусть C — множество всех вероятностных мер на заданной σ-алгебре подмножеств множества $D = X \times Y$. При каждом $c \in C$ имеем вероятностное пространство $\langle D, B, \mathsf{P}_c \rangle$, где B — σ-алгебра, $\mathsf{P}_c[D]$ — вероятностная мера.

Решающей функцией называется соответствие $\lambda\colon X \to Y$.

Качество принятого решения оценивается заданной функцией потерь $\mathcal{L}\colon Y^2 \to [0, \infty)$. Под риском будем понимать средние потери:

$$R(c, \lambda) = \int_D \mathcal{L}(y, \lambda(x))\, d\,\mathsf{P}_c[D].$$

При $\mathcal{L}(y, y') = \left\{ \begin{smallmatrix} 0, & y = y' \\ 1, & y \neq y' \end{smallmatrix} \right.$ риск есть вероятность ошибочной классификации.

Пусть $\nu = \{(x^i, y^i) \in D \,|\, i = 1, \ldots, N\}$ — случайная независимая выборка из распределения $\mathsf{P}_c[D]$. Эмпирический риск определяется как средние потери на выборке:

$$\widetilde{R}(\nu, \lambda) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(y^i, \lambda(x^i)).$$

Пусть $Q\colon \{\nu\} \to \Lambda$ — алгоритм (метод) построения решающих функций, $\lambda_{Q,\nu}$ — функция, построенная по выборке ν методом Q, Λ — заданный класс решающих функций.

Заметим, что значение риска зависит от c — распределения, которое неизвестно. Поэтому возникает задача оценивания риска по выборке. Доверительный интервал для R будем задавать [2] в виде $[0, \hat{R}(\nu)]$. При этом должно выполняться условие:

$$\forall c \quad \mathsf{P}_c\big(R \leqslant \hat{R}(\nu)\big) \geqslant \eta,$$

где η — заданная доверительная вероятность. Здесь мы ограничиваемся односторонними оценками, поскольку на практике для риска важны именно оценки сверху. Таким образом, в данном случае построение доверительного интервала эквивалентно выбору функции $\hat{R}(\nu)$, которую будем называть оценочной функцией или просто оценкой (риска).

При построении оценок риска первая проблема, которую нужно решить — это сравнение качества различных оценок.

Можно положить, что задан функционал качества $K\big(F_{c,\hat{R}}(\cdot)\big)$, где $F_{c,\hat{R}}(\cdot)$ — функция распределения оценки $\hat{R}(\nu)$. Выбор данного функционала, так же как и выбор функции потерь, определяется практическими соображениями. Простейшим вариантом такого функционала является математическое ожидание.

При фиксированном распределении c функционал K позволяет сравнивать качество оценок риска и находить оптимальную оценку. Однако на практике c неизвестно, а оценки, оптимальной при всех распределениях, может не существовать. В этом случае естественным является поиск множества Парето-недоминируемых оценок.

Оценки для конечного множества классификаторов

Пусть p — вероятность «успеха» в схеме Бернулли. Для фиксированного классификатора в роли p будет выступать вероятность ошибочной классификации.

Обозначим через $\xi = \frac{N_e}{N}$ случайную величину, представляющую собой долю ошибочно классифицированных объектов обучающей выборки, $B(\gamma, N, p) = \mathsf{P}(\xi \leqslant \gamma) = \sum\limits_{0 \leqslant i \leqslant N\gamma} C_N^i p^i (1-p)^{N-i}$ — кумулятивное биномиальное распределение.

*Работа выполнена при финансовой поддержке РФФИ, проекты №07-01-00331-а и №08-01-00944-а.

Имеем: $\mathsf{P}\,(\xi \leqslant \gamma) = B(\gamma, N, p)$ — вероятность получить долю ошибочно классифицированных объектов (эмпирический риск) меньше γ.

Если приравнять данную вероятность заданному уровню значимости α, то получим уравнение, связывающее p и γ. Выразив p как функцию γ, получим границу доверительного интервала для вероятности ошибочной классификации.

Пусть $\hat{p}(\gamma)$ — функция, задаваемая уравнением $B(\gamma, N, \hat{p}(\gamma)) = \alpha$. Очевидно, что для любого p выполняется $\mathsf{P}(p > \hat{p}(\xi)) \leqslant \alpha$.

Рассмотрим конечное множество классификаторов Λ, $|\Lambda| = L$. Для каждого $\lambda \in \Lambda$ определена вероятность ошибочной классификации $p(\lambda)$.

Обозначим $A(\lambda)$ — событие $p(\lambda) > \hat{p}(\gamma)$.

Имеем:

$$\mathsf{P}\Big(p\big(\lambda(\nu)\big) > \hat{p}(\gamma)\Big) \leqslant \mathsf{P}\bigg(\sum_{\lambda\in\Lambda} A(\lambda)\bigg) \leqslant \sum_{\lambda\in\Lambda} \mathsf{P}\big(A(\lambda)\big) \leqslant \alpha L. \quad (1)$$

Таким образом, при доверительной вероятности $\eta = 1 - \alpha L$ функция $\hat{p}(\gamma)$ является доверительным интервалом для вероятности ошибочной классификации при любом методе $\lambda(\nu)$ выбора решающей функции из Λ. Получили один из вариантов оценок Вапника-Червоненкиса.

В оценке присутствуют три неравенства. На первый взгляд может показаться, что первое из них несущественно влияет на точность оценки для метода минимизации эмпирического риска по всему классу Λ, поскольку при минимизации эмпирического риска, как правило, максимизируется разность между риском и эмпирическим риском. Однако классификатор, минимизирующий эмпирический риск, не обязательно оказывается среди нарушающих доверительный интервал. Фактически это есть эффект «расслоения» [3]. Помимо «расслоения», погрешность оценки этим неравенством может быть вызвана тем, что метод классификации не всегда минимизирует эмпирический риск (использует более сложный критерий качества или приближенный метод оптимизации).

Второе неравенство (известное также как «union bound») обращается в равенство только в случае, когда вероятность произведения любой пары событий равна нулю. В случае независимых событий, вероятность которых отлична от 0 и 1, неравенство становится строгим. Однако его вклад в погрешность оценок в этом случае является несущественным. Так, например, если A_j — независимы (в совокупности) и $\mathsf{P}(A_j) = p$, то

$$\mathsf{P}\bigg(\sum_{j=1}^{L} A_j\bigg) = 1 - (1-p)^L \xrightarrow[L\to\infty]{pL=\text{const}} 1 - e^{-pL}.$$

При $1 - e^{-pL} = 0{,}2$ получим $pL \approx 0{,}22$.

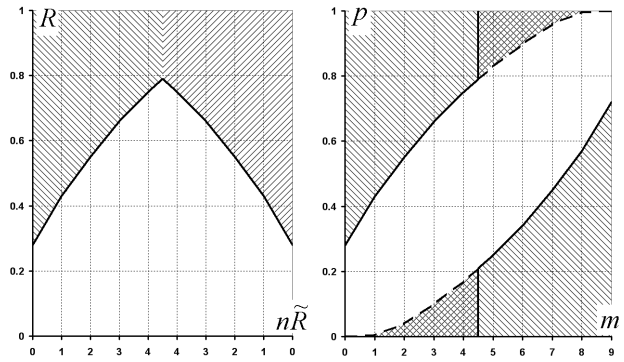


Рис. 1. Критические области.

Случай независимых событий представляет собой пример, когда неравенство union bound не вносит существенной погрешности в оценку риска, и именно на независимых событиях строится пример, показывающий, что сложностная оценка является асимптотически точной (рассмотрен далее). Вместе с тем, в большинстве практических ситуаций события $A(\lambda)$ зависимы, причем зависимость такова, что погрешность union bound становится весьма значительной. Это эффект «сходства» классификаторов [3].

Заметим, что эффекты «расслоения» и «сходства» связаны между собой, в частности, для независимых $A(\lambda)$ эффект «расслоения» также не имеет места (далее иллюстрируется на примере).

Последнее неравенство в оценке (1) есть следствие дискретности биномиального распределения и не дает существенного вклада в погрешность.

Простейшие примеры

Пусть $L = 1$. Оценка риска превращается в классический односторонний доверительный интервал для вероятности успеха в схеме Бернулли. Следующий пример: $L = 2$, причем λ_2 является инверсией λ_1, т. е. $\lambda_2(x) = 1 - \lambda_1(x)$. Можно для определенности считать, что первый классификатор всегда приписывает класс 0, второй — класс 1. Это соответствует методу классификации объектов по оценкам безусловных вероятностей классов (то есть в соответствии с выборочными частотами классов, без использования признакового пространства X). Вероятностная модель задается единственным параметром $p = \mathsf{P}(y = 0)$, и пусть m — количество объектов с $y = 0$ в выборке.

Оценка по формуле (1) совпадает с оценкой, полученной из двустороннего доверительного интервала для вероятности успеха в схеме Бернулли. Однако такой интервал неоптимален, поскольку включает в критическое множество лишние области (см. рис. 1). За счет «расслоения» часть критической области двустороннего доверительного интервала не входит в критическую область оценки риска. Это происходит, когда выход за грани-

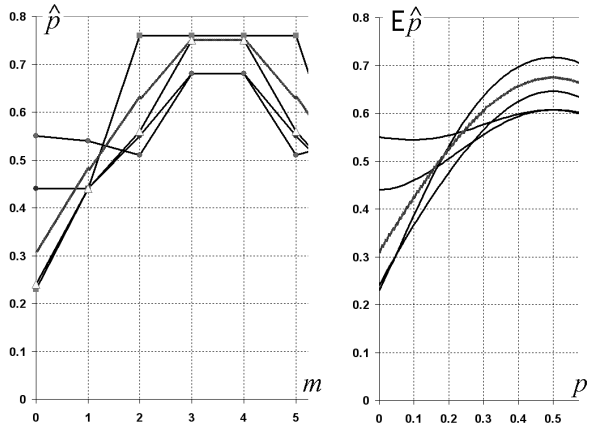


Рис. 2. Оценки, оптимальные по Парето.

цу двустороннего интервала происходит на классификаторе, который имеет не наилучший эмпирический риск. На рисунке эта область показана двойной штриховкой (критические множества — одинарной).

Заметим, что в рассматриваемом примере интервала, оптимального при всех распределениях, не существует. Если при построении одностороннего интервала свобода выбора отсутствует, то для двустороннего интервала можно варьировать вероятности выхода за левую и правую границы. Существуют различные доверительные интервалы, оптимальные по Парето. Некоторые из них (а также доминируемая сложностная оценка) приведены на рис. 2.

Асимптотические оценки

Будем рассматривать асимптотическое поведение оценок (1), когда абсолютный объем выборки стремится к бесконечности, но при этом относительно сложности остается малым [4], т. е. $N \to \infty$, $\varkappa = \frac{N}{\ln L} = \text{const}$.

Хотя рассматривается асимптотический случай, полученные выводы будут на качественном уровне применимы и к случаю достаточно малых выборок $N \approx 50$. Рассмотрим простейший вариант сложностной оценки [1]: $\hat{p}(\gamma) = \gamma + \varepsilon$, где

$$\varepsilon = \sqrt{\frac{\ln L - \ln(1-\eta)}{2N}} = \sqrt{\frac{1}{2\varkappa}\left(1 + \frac{\varkappa}{N}\ln(1-\eta)\right)}.$$

При $\varkappa \ll N$, что, как правило, выполняется на практике, оценка определяется главным образом значением $\varkappa$, в то время как N и η несущественны. Например, при $\varkappa = 5$ и $N = \infty$, имеем $\varepsilon = 0{,}316$, при $N = 50$ и $\eta = 0{,}5$ получим $\varepsilon = 0{,}327$, при $\eta = 0{,}9$ получим $\varepsilon = 0{,}351$.

Таким образом, при типичных значениях параметров оценка доверительного интервала слабо зависит от доверительной вероятности (если не выбирать значения близкие к 1), а объем выборки влияет, главным образом, опосредованно через $\varkappa$. Аналогичные рассуждения можно провести и для более тонких сложностных оценок. Это оправдывает рассмотрение предложенной асимптотики, позволяющей ограничиться наиболее существенным параметром — $\varkappa$.

Чтобы построить доверительный интервал риска, пользуясь (1), нужно найти $\hat{p}(\gamma)$, удовлетворяющую уравнению

$$LB(\gamma, N, \hat{p}(\gamma)) = 1 - \eta. \tag{2}$$

Справедлива известная [5] оценка

$$\frac{1}{\sqrt{2\pi N\gamma(1-\gamma)}}\exp\Big(-NH(\gamma,p) - \frac{1}{12N\gamma(1-\gamma)}\Big) \leqslant B(\gamma,N,p) \leqslant \exp\big(-NH(\gamma,p)\big),$$

где $H(\gamma,p) = \gamma\ln\frac{\gamma}{p} + (1-\gamma)\ln\frac{1-\gamma}{1-p}$ — дивергенция Кульбака-Лейблера.

Умножая неравенства на L, получаем, что решения уравнения (2) при $N \to \infty$, $\varkappa = \frac{N}{\ln L}$ сходятся к решению уравнения

$$H(\gamma, \hat{p}(\gamma)) = 1/\varkappa. \tag{3}$$

Решение $\hat{p}_\varkappa(\gamma)$ уравнения (3) оказывается неулучшаемой (без использования дополнительной информации) асимптотической оценкой риска на основе эмпирического риска.

Для доказательства этого факта рассмотрим пример, приведенный в [6]. Пусть дан набор переменных $X_1, \ldots, X_L$ и множество классификаторов $\lambda_1, \ldots, \lambda_L$, причем λ_j приписывает объекту класс, номер которого равен значению j-й переменной, то есть $f_{\lambda_j}(x) = x_j$. Распределение в D задается следующим образом:

$$\mathsf{P}(x,y) = \mathsf{P}(x\,|\,y)\,\mathsf{P}(y); \quad \mathsf{P}(x\,|\,y) = \sum_{j=1}^{L}\mathsf{P}(x_j\,|\,y);$$
$$\mathsf{P}(y=0) = p_0; \quad \mathsf{P}(x_j \neq y\,|\,y) = p,$$

где p_0 и p — параметры распределения, причем выбор p_0 не имеет значения. По построению, риск для любого классификатора λ_j равен p. Вместо (1) для алгоритма $\lambda(\nu)$, минимизирующего эмпирический риск, в рассмотренной модели можем выписать точную вероятность выхода из доверительного интервала:

$$\mathsf{P}\big(p(\lambda(\nu)) > \hat{p}(\gamma)\big) = 1 - (1 - B(\gamma,N,p))^L = 1 - \eta.$$

Аппроксимируя степень экспонентой, получаем: $LB(\gamma, N, \hat{p}(\gamma)) \approx -\ln\eta$. От (2) данное уравнение отличается только правой частью. Но, так как асимптотическое решение не зависит от правой части, получаем, что для полученного уравнения решение

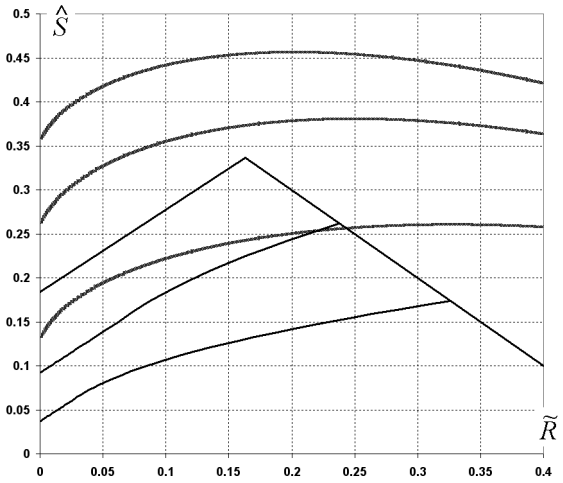


Рис. 3. Оценки для гистограммного классификатора.

есть $\hat{p}_{\varkappa}(\gamma)$. Если взять любую $\hat{p}(\gamma)$, строго меньшую $\hat{p}_{\varkappa}(\gamma)$ на некотором, ненулевой длины, интервале значений аргумента, то найдутся такие параметры рассмотренной модели, что вероятность выхода из доверительного интервала $\hat{p}(\gamma)$ будет больше любого $1-\eta<1$. Это доказывает, что оценка $\hat{p}_{\varkappa}(\gamma)$ является асимптотически неулучшаемой.

То, что без дополнительной информации нельзя добиться существенного улучшения оценок Вапника-Червоненкиса, общеизвестно. В данном разделе показано, что в рассмотренной асимптотике они неулучшаемы вовсе.

Точные асимптотические оценки для гистограммного классификатора

Пусть $X=\{1,\ldots,k\}$ — единственная дискретная переменная, и метод классификации минимизирует эмпирический риск в каждой точке.

Вероятностная мера c задается набором вероятностей $\zeta_j=\mathsf{P}(x{=}j),\ p_j=\mathsf{P}(y{=}0\,|\,x{=}j)$.

Данный пример имеет практическую ценность — он соответствует использованию histogram classifier, когда решение принимается по гистограммам частот [7].

В асимптотике доверительный интервал совпадает с $\widehat{S}+\widetilde{R}$, где $\widehat{S}=\sup\limits_{c\in C}\mathsf{E}R-\mathsf{E}\widetilde{R}$ — функция максимального (по всем распределениям) смещения эмпирического риска [4].

В рассмотренном случае существенную роль играет эффект «сходства» классификаторов. Их множество продуцирует всего $\log_2 L=k$ «ячеек», в то время как максимально возможное число «ячеек» (имеющее место в примере, иллюстрирующем достижимость сложностных оценок) составляет 2^L. Под «ячейками» здесь понимаются области порождаемого классификаторами разбиения признакового пространства. Число таких областей представляется перспективной для дальнейшего исследования характеристикой сложности семейства классификаторов, которая может использоваться наряду с ёмкостью.

На рис. 3 приведены доверительные интервалы для риска. Три гладкие кривые соответствуют асимптотическим сложностным оценкам $\hat{p}_{\varkappa}(\gamma)-\gamma$ при $\theta=1,2,5$, где $\theta=\frac{N}{k}$, $\varkappa=\frac{\theta}{(1-e^{-\theta})\ln 2}$. Кривые с изломами отражают точные значения максимального смещения эмпирического риска при таких же θ. Рисунок показывает различие точных универсальных оценок и точных оценок, учитывающих все особенности задачи.

Выводы

В работе рассмотрены достаточно простые примеры, которые позволяют проанализировать поведение сложностных оценок, а также наглядно проиллюстрировать эффекты «расслоения» и «сходства» классификаторов.

Несмотря на использование асимптотического приближения и рассмотрение самого простого случая гистограммного классификатора, полученные результаты имеют практическую ценность, поскольку исследованные примеры отражают существенные свойства оценок риска, проявляющиеся в реальных задачах.

Литература

[1] *Вапник В. Н., Червоненкис А. Я.* Теория распознавания образов. — Москва: Наука, 1974. — 415 с.

[2] *Неделько В. М.* Об интервальном оценивании риска для решающей функции // Таврический вестник информатики и математики. Изд-во НАН Украины. — 2008. — № 2. — С. 97–103.

[3] *Vorontsov K. V.* Combinatorial probability and the tightness of generalization bounds // Pattern Recognition and Image Analysis. — 2008. — Vol. 18, No. 2. — Pp. 243–259.

[4] *Nedelko V. M.* Estimating a Quality of Decision Function by Empirical Risk // LNAI 2734. Machine Learning and Data Mining in Pattern Recognition. Third Intern. Conference, MLDM 2003. Proceedings. Leipzig: Springer-Verlag. — 2003. — Pp. 182–187.

[5] *Боровков А. А.* Теория вероятностей. — Москва: Наука, 1986. — 432 с.

[6] *Langford J.* Quantitatively tight sample complexity bounds. Carnegie Mellon Thesis. — 2002. — `http://citeseer.ist.psu.edu/langford02quantitatively.html`. — 130 p.

[7] *Braga–Neto U. and Dougherty E. R.* Exact performance of error estimators for discrete classifiers // Pattern Recognition, Elsevier Ltd. — 2005. — Vol. 38, No. 11. — Pp. 1799–1814.

[8] *Неделько В. М.* Об эффективности эмпирических функционалов качества решающей функции // Всеросс. конф. ММРО-13. — М.: МАКС Пресс, 2007. — С. 47–49.

Возможность как альтернативная вероятности модель случайности: событийно-частотная интерпретация и эмпирическое построение*

Пытьев Ю. П.
yuri.pytyev@gmail.com
Москва, МГУ им. М. В. Ломоносова

Вероятностные модели стохастических объектов не могут быть построены эмпирически, если в процессе получения данных их вероятностные характеристики непредсказуемо эволюционируют. В докладе рассмотрены возможностные модели подобных объектов, охарактеризован класс стохастических объектов, возможностные модели которых могут быть построены эмпирически, а вероятностные — нет, и приведены адаптивные алгоритмы эмпирического построения их возможностных моделей, позволяющие получать с гарантированной вероятностью правильные модели на основе почти наверное конечного числа данных.

Что можно сказать о предопределенности[1] исходов стохастического эксперимента (СЭ), если его моделью является вероятностное пространство $(\Omega, \mathcal{P}(\Omega), \mathrm{Pr})$, в котором $\Omega = \{\omega_1, \omega_2, \ldots\}$? В частности, что можно сказать о *возможностях* исходов СЭ в этом случае, — об их *шансах*? Ясно лишь, что при любом определении возможности p_i исхода $\omega_i \in \Omega$ как значения меры (возможности $\mathrm{P}(\cdot)\colon \mathcal{P}(\Omega) \to [0,1]$), *при каждом испытании оценивающей, обусловленный свойствами* СЭ, *шанс его исхода ω_i в сравнении с шансами всех других его элементарных исходов*, естественно считать, что $\mathrm{P}(\{\omega_i\}) \stackrel{\mathrm{def}}{=} \mathrm{p}_i \geqslant \mathrm{p}_j \stackrel{\mathrm{def}}{=} \mathrm{P}(\{\omega_i\})$, если $\mathrm{Pr}(\{\omega_i\}) \stackrel{\mathrm{def}}{=} \mathrm{pr}_i \geqslant \mathrm{pr}_j \stackrel{\mathrm{def}}{=} \mathrm{Pr}(\{\omega_j\})$.

В данном случае принципиально то, что для такого заключения *не требуются значения* $\mathrm{pr}_1, \mathrm{pr}_2, \ldots$, *достаточно лишь знать, как они упорядочены.* Более того, такое заключение останется в силе, если *вероятности* $\mathrm{pr}_1, \mathrm{pr}_2, \ldots$ *произвольно изменяются от испытания к испытанию, оставаясь лишь одинаково упорядоченными*, например, согласно условию

$$1 \geqslant \mathrm{pr}_1 \geqslant \mathrm{pr}_2 \geqslant \ldots > 0, \quad \mathrm{pr}_1 + \mathrm{pr}_2 + \ldots = 1. \quad (1)$$

Рассмотрим СЭ, моделью которого является *класс* $\mathcal{P}r \stackrel{\mathrm{def}}{=} \{(\Omega, \mathcal{P}(\Omega), \mathrm{Pr}),\ \mathrm{Pr} \in \mathbb{P}r\}$ *дискретных вероятностных пространств, где* $\mathbb{P}r$ *— класс вероятностей* $\mathrm{Pr}(\cdot)\colon \mathcal{P}(\Omega) \to [0,1]$, удовлетворяющих условию (1). Знания одной лишь упорядоченности (1) вероятностей $\mathrm{pr}_1, \mathrm{pr}_2, \ldots$, конечно, недостаточно, чтобы охарактеризовать СЭ в терминах формализма теории вероятностей. Класс $\mathcal{P}r$ является «существенно недоопределенной» стохастической моделью СЭ, а если вероятности в (1) *произвольно изменяются от испытания к испытанию, то наблюдения за исходами* $\mathrm{СЭ} \times \ldots \times \mathrm{СЭ} = (\mathrm{СЭ})^n$ *не позволят ее «доопределить», как бы велико ни было* n.

В возможностной модели СЭ, модель которого определена как класс $\mathcal{P}r$, возможности $\mathrm{p}_i = \mathrm{P}(\{\omega_i\})$, $i = 1, 2, \ldots$, априори должны быть подчинены условию

$$1 = \mathrm{p}_1 \geqslant \mathrm{p}_2 \geqslant \ldots \geqslant 0, \quad (2)$$

согласованному с условием (1) ($\mathrm{p}_1 = 1$ — условие нормировки), а *каждая конкретная упорядоченность* в (2), *в которой встречаются только равенства и строгие неравенства, должна определить класс взаимно эквивалентных возможностей* $\mathrm{P}(\cdot)\colon \mathcal{P}(\Omega) \to [0,1]$ *и соответствующий класс эквивалентных пространств с возможностью* $(\Omega, \mathcal{P}(\Omega), \mathrm{P})$.

Обозначим $\mathbb{P}$ — класс возможностей, удовлетворяющих условию (2), и $\mathcal{P} \stackrel{\mathrm{def}}{=} \{(\Omega, \mathcal{P}(\Omega), \mathrm{P}),\ \mathrm{P} \in \mathbb{P}\}$ — соответствующий класс пространств с возможностью — *возможностную модель* СЭ.

Представим $\mathbb{P}$ в виде разбиения на классы *взаимно эквивалентных возможностей*, каждый из которых определит единственную с точностью до эквивалентности возможностную модель. Заметим, что *всякую конкретную упорядоченность* в (2) можно задать двоичным числом $e = 0{,}e_1e_2\ldots \in (0,1)$, в котором $e_i = 1$, если $\mathrm{p}_i > \mathrm{p}_{i+1}$, и $e_i = 0$, если $\mathrm{p}_i = \mathrm{p}_{i+1}$, $i = 1, 2, \ldots$ Обозначим $\mathbb{P}_{(e)}$ — *класс возможностей, упорядоченность значений* $\mathrm{P}(\{\omega_i\})$, $i = 1, 2, \ldots$, *которых определена значением* $e \in (0,1)$. Тогда $\mathbb{P}_{(e)} \bigcap \mathbb{P}_{(e')} = \varnothing$, если $e \neq e'$, и

$$\mathbb{P} = \bigcup_{e \in (0,1)} \mathbb{P}_{(e)}. \quad (3)$$

Шкала значений возможности.

Определим *шкалу* $\mathcal{L}$ *значений возможности* как интервал $[0,1]$ с естественной упорядоченностью $\leqslant$ и двумя бинарными операциями — сложением $\dotplus\colon [0,1] \times [0,1] \to [0,1]$ и умножением

*Работа выполнена при финансовой поддержке РФФИ, проект № 08-07-00133-а

[1] Напомним, что $\mathrm{Pr}(A)$ — прогнозируемое значение *частоты* события A в серии взаимно независимых испытаний, но не мера предопределенности или *возможности* A при каждом испытании.

$\bullet\colon [0,1]\times[0,1]\to[0,1]$, т. е. определим четвёрку $\mathcal{L}=([0,1],\leqslant,+,\bullet)$, и *группу* Γ *изотонных, сохраняющих каждую конкретную упорядоченность в* (2) *автоморфизмов* $\mathcal{L}$, порожденную группой строго монотонных непрерывных функций $\gamma(\cdot\,\cdot)\colon [0,1]\to[0,1]$, $\gamma(0)=0$, $\gamma(1)=1$, с групповой операцией $\circ$, определённой как $\gamma'\circ\gamma(a)\stackrel{\text{def}}{=}\gamma'(\gamma(a))$, $a\in[0,1]$. Поскольку Γ — группа *автоморфизмов* $\mathcal{L}$, то $\forall\ a,\ b\in[0,1]$ и $\gamma(\cdot)\in\Gamma$

$$\begin{aligned} &a*b\Leftrightarrow\gamma(a)*\gamma(b),\quad \gamma(a+b)=\gamma(a)+\gamma(b),\\ &\gamma(a\bullet b)=\gamma(a)\bullet\gamma(b),\quad \gamma(0)=0,\quad \gamma(1)=1,\end{aligned}\tag{4}$$

где $*$ означает либо $<$, либо $>$, либо $=$.

Теорема 1 ([1]). *Если*

1) *операции* $+$ *и* $\bullet$, *как отображения из* $[0,1]\times[0,1]$ *в* $[0,1]$, *непрерывны;*
2) *для любых* $a,b\in[0,\,1]$

$$\begin{aligned} &a\bullet b=b\bullet a,\quad 0\bullet a=0,\quad 1\bullet a=a,\\ &a+b=b+a,\quad 0+a=a,\quad 1+a=1;\end{aligned}\tag{5}$$

3) *для всех* $\gamma(\cdot)\in\Gamma$ *выполнены условия*[2] (4);

то $a+b=\max(a,b)$, $a\bullet b=\min(a,b)$, $a,b\in[0,1]$.

В шкале $\mathcal{L}=\big([0,1],\leqslant,+,\bullet\big)$ значений возможности[3] операции «$+$» и «$\bullet$» коммутативны, ассоциативны, и взаимно дистрибутивны.

Операция «$+$» определяет возможность любого события $A\in\mathcal{P}(\Omega)$,

$$\mathrm{P}(A)\stackrel{\text{def}}{=}\mathop{+}_{i:\omega_i\in A}\mathrm{p}_i\stackrel{\text{def}}{=}\sup_{i:\omega_i\in A}\mathrm{P}(\{\omega_i\}),\tag{6}$$

подобно тому, как для вероятности

$$\Pr(A)=\sum_{i:\omega_i\in A}\mathrm{pr}_i=\sum_{i:\omega_i\in A}\Pr(\{\omega_i\}),\quad A\in\mathcal{P}(\Omega).$$

Возможность, максимально согласованная с вероятностью

В [1] показано, что каждому классу $\mathbb{P}_{(e)}$ в (3) согласно условиям

$$\begin{aligned} &e_i=1\Leftrightarrow\mathrm{p}_i>\mathrm{p}_{i+1}\Leftrightarrow\\ &\qquad\Leftrightarrow\mathrm{pr}_1+\cdots+\mathrm{pr}_{i-1}+2\mathrm{pr}_i\stackrel{\text{def}}{=}f_i>1,\\ &e_i=0\Leftrightarrow\mathrm{p}_i=\mathrm{p}_{i+1}\Leftrightarrow f_i\leqslant 1,\ \ i=1,2,\ldots,\end{aligned}\tag{7}$$

где $\Leftrightarrow$ означает «если и только если», *взаимно однозначно сопоставлен класс вероятностей* $\mathbb{P}\mathrm{r}_{(e)}$, $e\in(0,1)$, *а разбиению* (3) *класса* $\mathbb{P}$ — *разбиение*

$$\mathbb{P}\mathrm{r}=\bigcup_{e\in(0,1)}\mathbb{P}\mathrm{r}_{(e)},\mathbb{P}\mathrm{r}_{(e)}\cap\mathbb{P}\mathrm{r}_{(e')}=\varnothing,\quad e\neq e',\ e,e'\in(0,1),\tag{8}$$

класса $\mathbb{P}\mathrm{r}$ вероятностей, распределенных согласно условиям (1), и при этом *для любых* $\mathrm{P}\in\mathbb{P}_{(e)}$ и $\mathrm{Pr}\in\mathbb{P}\mathrm{r}_{(e)}$ *можно указать монотонно неубывающую непрерывную* на $(0,1]$ *функцию* $\widetilde{\gamma}_e(\cdot)\colon[0,1]\to[0,1]$ из класса $\widetilde{\Gamma}(\mathrm{Pr})$, *такую, что для любого* $A\in\mathcal{P}(\Omega)$

$$\mathrm{P}(A)=\sup_{i:\,\omega_i\in A}\mathrm{p}_i=\widetilde{\gamma}_e\Big(\sum_{i:\,\omega_i\in A}\mathrm{pr}_i\Big)=\\ =\widetilde{\gamma}_e(\mathrm{Pr}(A)),\ e\in(0,1).\tag{9}$$

Класс $\widetilde{\Gamma}(\mathrm{Pr})$ всех таких функций $\widetilde{\gamma}_e(\cdot)$ определяется вероятностью $\mathrm{Pr}\in\mathbb{P}\mathrm{r}_{(e)}$, $e\in(0,1)$. *Любая возможность* $\mathrm{P}\in\mathbb{P}_{(e)}$ *называется максимально согласованной с любой вероятностью* $\mathrm{Pr}\in\mathbb{P}\mathrm{r}_{(e)}$, факт *максимальной* согласованности P с Pr выражает символ $\mathrm{Pr}\approx\!>\mathrm{P}$, означающий, что среди неравенств $\widetilde{\gamma}_e(\mathrm{pr}_1)\geqslant\widetilde{\gamma}_e(\mathrm{pr}_2)\geqslant\ldots$, *максимальное* число строгих неравенств, $e\in(0,1)$, [1]. Если $\mathrm{Pr}\approx\!>\mathrm{P}$, то каждое событие $A\in\mathcal{P}(\Omega)$ в $(\Omega,\mathcal{P}(\Omega),\mathrm{Pr})$ можно интерпретировать как событие в $(\Omega,\mathcal{P}(\Omega),\mathrm{P})$, а его возможность будет определена его вероятностью равенством (9), в этом смысле *возможность* P *называется* Pr*-измеримой.*

Далее ограничимся случаем *регулярных вероятностей*, для которых в (7) либо $f_i>1$, либо $f_i<1$ $i=1,2,\ldots$

Событийно-частотная интерпретация и эмпирическое построение. Вероятность не изменяется в процессе наблюдений

Согласно (9) и З. Б. Ч.[4], если $\mathrm{Pr}\approx\!>\mathrm{P}$, то для любых $A,B\in\mathcal{P}(\Omega)$, таких, что $\Pr(A)>\Pr(B)$, $\forall\,\tilde{\gamma}_e(\cdot)\in\tilde{\Gamma}(\mathrm{Pr})\ \ \forall\varepsilon>0\ \ \exists N=N(\varepsilon,A,B,\tilde{\gamma}_e(\cdot))$, $\forall\,n>N$

$$\begin{aligned}&\tilde{\gamma}_e(\mathrm{Pr}(A))-\tilde{\gamma}_e(\mathrm{Pr}(B))-\varepsilon\stackrel{\text{п. н.}}{<}\\ &\qquad\stackrel{\text{п. н.}}{<}\tilde{\gamma}_e(\nu^{(n)}(A))-\tilde{\gamma}_e(\nu^{(n)}(B))\stackrel{\text{п. н.}}{<}\\ &\qquad\qquad\stackrel{\text{п. н.}}{<}\tilde{\gamma}_e(\mathrm{Pr}(A))-\tilde{\gamma}_e(\mathrm{Pr}(B))+\varepsilon.\end{aligned}$$

---

[2] Согласно (4) и (5) 0 и 1 суть *нейтральные элементы* $\mathcal{L}$.

[3] Шкала $\mathcal{L}=([0,1],\leqslant,+,\bullet)$ — полная дистрибутивная решетка, в которой решеточные операции суть $a\vee b\stackrel{\text{def}}{=}\stackrel{\text{def}}{=}a+b$, $a\wedge b\stackrel{\text{def}}{=}a\bullet b$ [2].

[4] Если $\nu^{(n)}(A)$ — частота события $A\in\mathcal{A}$ в серии n взаимно независимых испытаний, модель которых $(\Omega,\mathcal{A},\mathrm{Pr})\times\ldots\times(\Omega,\mathcal{A},\mathrm{Pr})=(\Omega,\mathcal{A},\mathrm{Pr})^n$, то для всех $\varepsilon>0$ и $A\in\mathcal{A}$

$$\lim_{N\to\infty}\mathrm{Pr}^\infty\Big(\sup_{n\geqslant N}|\nu^{(n)}(A)-\mathrm{Pr}(A)|>\varepsilon\Big)=0,\tag{10}$$

т. е. $\nu^{(n)}(A)$ с увеличением n приближается и остается близкой к $\mathrm{Pr}(A)$, ибо согласно (10) $|\nu^{(n)}(A)-\mathrm{Pr}(A)|>\varepsilon$ лишь для Pr^∞-почти наверное (п. н.) конечного числа n испытаний (усиленный З. Б. Ч., $\nu^{(n)}(A)\xrightarrow[n\to\infty]{\text{п. н.}}\mathrm{Pr}(A)$, Pr^∞ — вероятность, определенная на борелевских множествах бесконечных последовательностей испытаний).

Поэтому, если

$$\mathrm{P}(A) = \tilde{\gamma}_e(\mathrm{Pr}(A)) > \tilde{\gamma}_e(\mathrm{Pr}(B)) = \mathrm{P}(B),$$

то, выбрав $\varepsilon \in \big(0, \tilde{\gamma}_e(\mathrm{Pr}(A)) - \tilde{\gamma}_e(\mathrm{Pr}(B))\big)$, найдем, что для всех $n > N$

$$\mathrm{P}(A) > \mathrm{P}(B) \Rightarrow$$
$$\Rightarrow \tilde{\gamma}_e(\nu^{(n)}(A)) \overset{\text{п. н.}}{>} \tilde{\gamma}_e(\nu^{(n)}(B)) \Rightarrow$$
$$\Rightarrow \nu^{(n)}(A) \overset{\text{п. н.}}{>} \nu^{(n)}(B).$$

Поэтому *в достаточно длинной последовательности взаимно независимых испытаний упорядоченность возможностей любых событий п. н. точно прогнозирует такую же упорядоченность их частот.* Такова эмпирическая интерпретация возможности; такая же интерпретация сохраняется и при изменяющейся вероятности [1].

Взаимно однозначное соответствие между $\mathbb{P}_{(e)}$ и $\mathbb{P}\mathrm{r}_{(e)}$, $e \in (0,1)$, в (3) и в (8) решает *проблему эмпирического построения возможностной модели* [1, 3].

Речь идет о задаче эмпирического восстановления класса $\mathbb{P}\mathrm{r}_{(e)} \subset \mathbb{P}\mathrm{r}$ вероятностей, содержащего вероятность Pr, определяющую модель испытаний $(\Omega, \mathcal{P}(\Omega), \mathrm{Pr})^n$, поскольку *класс* $\mathbb{P}\mathrm{r}_{(e)}$ *определяет класс* $\mathbb{P}_{(e)}$ *взаимно эквивалентных возможностей* P *и, следовательно, — класс* $(\Omega, \mathcal{P}(\Omega), \mathrm{P})$, $\mathrm{P} \in \mathbb{P}_{(e)}$ *взаимно эквивалентных возможностных моделей каждого испытания.*

Согласно условиям (7), задача эмпирического восстановления возможности сводится к задаче теории статистических решений, в которой на основе значений частот $\nu_i^{(n)}$, $i = 1, \ldots, s$, элементарных событий $\{\omega_i\}$, $i = 1, \ldots, s$, наблюденных в последовательности n взаимно независимых испытаний, модель которых определена как $(\Omega, \mathcal{P}(\Omega), \mathrm{Pr})^n$, $n = 1, 2, \ldots$, для всех $i = 1, \ldots, s$ требуется принять одну из гипотез $f_i > 1$ или $f_i < 1$, $s = 1, 2, \ldots$.

Рассмотрим следующий алгоритм принятия решений: для всех $i = 1, \ldots, s$ и каждого $n = 1, 2, \ldots$

$$\begin{aligned} &\bullet \text{ если } \hat{f}_i^{(n)} > 1 + \delta^{(n,s)}, \text{ то } \square_1\text{: считать } f_i > 1; \\ &\bullet \text{ если } \hat{f}_i^{(n)} < 1 - \delta^{(n,s)}, \text{ то } \square_2\text{: считать } f_i < 1; \\ &\bullet \text{ если } |\hat{f}_i^{(n)} - 1| \leqslant \delta^{(n,s)}, \text{ то } \circlearrowleft\text{: продолжить} \\ &\qquad\qquad\qquad\qquad\qquad\qquad \text{испытания,} \end{aligned} \tag{11}$$

где $\hat{f}_i^{(n)} \triangleq \nu_1^{(n)} + \ldots + \nu_{i-1}^{(n)} + 2\nu_i^{(n)}$, $i = 1, 2, \ldots$

В алгоритме (11) при каждом $n = 1, 2, \ldots$ проверяются условия решений $\square_1$, $\square_2$ и $\circlearrowleft$ для всех $i = 1, \ldots, s$; если при этом приняты только решения $\square_1$ или $\square_2$, то алгоритм завершен, если же хотя бы для одного i принято решение $\circlearrowleft$, то после каждого дополнительного испытания для всех $i = 1, \ldots, s$ проверяются условия решений $\square_1$, $\square_2$ и $\circlearrowleft$, ранее принятые решения корректируются, и так до тех пор, пока алгоритм не будет завершен.

Если значение $\alpha^{(s)}$ оценивает сверху вероятности ошибочных решений $\square_1$ и $\square_2$, то в (11)

$$\delta^{(n,s)} = \left(\frac{2}{n} \ln \frac{1}{\alpha^{(s)}}\right)^{\frac{1}{2}}, \quad n, s = 1, 2, \ldots, \tag{12}$$

и, как нетрудно убедиться, событие $\{1 - \delta^{(n,s)} \leqslant \leqslant \hat{f}_i^{(n)} \leqslant 1 + \delta^{(n,s)}\}$, приводящее в алгоритме (11) к решению $\circlearrowleft$ о продолжении испытаний, при любом условии $f_i > 1$ или $f_i < 1$, $i = 1, \ldots, s$, для каждого $s = 1, 2, \ldots$ может выполняться лишь для п. н. конечного числа n испытаний[5].

Следовательно, алгоритм (11) проверки любого конечного числа $2s$ гипотез $f_i > 1$ или $f_i < 1$, $i = 1, 2, \ldots$, будет завершен на основе данных п. н. конечного числа испытаний.

Теорема 2 ([3]). *Для любого $s = 1, 2, \ldots$ алгоритм (11) на основе п. н. конечного числа испытаний восстанавливает упорядоченность возможностей элементарных событий* $\mathrm{p}_1, \ldots, \mathrm{p}_s$, *совпадающую с истинной их упорядоченностью с вероятностью, большей* $1 - s\,\alpha^{(s)} = 1 - \alpha$, *если* $\alpha^{(s)} = \alpha/s$, *где α — априорная оценка вероятности ошибочного упорядочения возможностей s элементарных событий.*

Алгоритм эмпирического восстановления возможности. Вероятность изменяется от испытания к испытанию

Пусть модель n взаимно независимых испытаний определена как вероятностное пространство

$$(\Omega^{(n)}, \mathcal{P}(\Omega^{(n)}), \mathrm{Pr}^{(n)}_{1,\ldots,n}) = \\ = (\Omega, \mathcal{P}(\Omega), \mathrm{Pr}_1) \times \ldots \times (\Omega, \mathcal{P}(\Omega), \mathrm{Pr}_n), \tag{13}$$

в котором $\mathrm{Pr}^{(n)}_{1,\ldots,n} = \mathrm{Pr}_1 \times \ldots \times \mathrm{Pr}_n$, для каждого $j = 1, 2, \ldots$ $\mathrm{Pr}_j \in \mathbb{P}\mathrm{r}_{(e)}$ и, следовательно, *выполнено условие* $\mathbb{P}\mathrm{r}_{(e)}$*-измеримости восстанавливаемой возможности* P, *согласно которому для каждого* $i = 1, 2, \ldots$

$$\text{либо } e_i = 1 \Leftrightarrow \mathrm{p}_i > \mathrm{p}_{i+1} \Leftrightarrow \mathrm{F}_i^{(j)} > 1, \tag{14}$$

$$\text{либо } e_i = 0 \Leftrightarrow \mathrm{p}_i = \mathrm{p}_{i+1} \Leftrightarrow \mathrm{F}_i^{(j)} < 1, \tag{15}$$

где $\mathrm{F}_i^{(j)} \overset{def}{=} \mathrm{Pr}_j(\{\omega_1\}) + \ldots + \mathrm{Pr}_j(\{\omega_{i-1}\}) + 2\mathrm{Pr}_j(\{\omega_i\})$ для всех $j = 1, 2, \ldots$. Значение $e = 0{,}e_1e_2\ldots$, определяющее упорядоченность $\mathrm{p}_i = \mathrm{P}(\{\omega_i\})$, $i = 1, 2, \ldots$, разумеется, не известно и должно быть определено на основе результатов испытаний.

---

[5] Этот факт и равенство (12) следуют из леммы Хёфдинга [4], согласно которой, если случайные величины $\xi_1, \ldots, \xi_n$ взаимно независимы и $\mathrm{Pr}(a_k \leqslant \xi_k \leqslant b_k) = 1$, $k = 1, \ldots, n$, то $(\forall \varepsilon > 0)$ $\mathrm{Pr}\Big(\sum\limits_{k=1}^{n} \xi_k - \mathbf{E} \sum\limits_{k=1}^{n} \xi_k > n\varepsilon\Big) \leqslant \exp\Big(\frac{-2n^2\varepsilon^2}{\sum_{k=1}^{n}(b_k - a_k)^2}\Big)$.

В рассматриваемом случае проверяемые гипотезы определяются следующими условиями:

$$\text{либо } e_i = 1 \Leftrightarrow \mathrm{p}_i > \mathrm{p}_{i+1} \Leftrightarrow f_i^{(n)} > 1, \qquad (16)$$

$$\text{либо } e_i = 0 \Leftrightarrow \mathrm{p}_i = \mathrm{p}_{i+1} \Leftrightarrow f_i^{(n)} < 1, \qquad (17)$$

где $f_i^{(n)} \triangleq \mathrm{pr}_1^{(n)} + \ldots + \mathrm{pr}_{i-1}^{(n)} + 2\mathrm{pr}_i^{(n)}$,

$\mathrm{pr}_i^{(n)} \triangleq \frac{1}{n}\sum\limits_{j=1}^{n} \mathrm{Pr}_j(\{\omega_i\})$, $i = 1, \ldots, s$,

согласно которым восстанавливаемая возможность P максимально согласована с каждой *вероятностью* $\mathrm{Pr}^{(n)} \triangleq \frac{1}{n}\sum\limits_{j=1}^{n} \mathrm{Pr}_j$, $n = 1, 2, \ldots$, ибо $\big(\mathrm{Pr}_j \in \mathcal{P}r_{(e)},\ j{=}1, 2, ...\big) \to \big(\mathrm{Pr}^{(n)} \in \mathbb{P}r_{(e)},\ n{=}1, 2, ...\big)$.

Так как модель испытаний такова, что условия (14), (15) $\mathbb{P}\mathrm{r}_{(e)}$-измеримости восстанавливаемой возможности выполнены для каждого $i = 1, 2, \ldots$ и всех $j = 1, 2, \ldots$, то гипотезы (16), (17) суть следствия условий (14), (15), но, в отличие от последних, могут быть проверены эмпирически, поскольку при $n \to \infty$ $\hat{f}_i^{(n)} - f_i^{(n)} \xrightarrow{\text{п. н.}} 0$, где

$$\hat{f}_i^{(n)} \triangleq \nu_1^{(n)} + \ldots + \nu_{i-1}^{(n)} + 2\nu_i^{(n)},$$

$\nu_i^{(n)}$ — частота элементарного события $\{\omega_i\}$ в последовательности n испытаний[6], $i = 1, 2, \ldots$

Рассмотрим следующий алгоритм принятия решений в задачах проверки гипотез (16), (17): для всех $i = 1, \ldots, s$ и каждого $n = 1, 2, \ldots$

- если $\hat{f}_i^{(n)} > 1{+}\delta^{(n,s)}$, то $\square_1$: считать $f_i^{(n)}{>}1$;
- если $\hat{f}_i^{(n)} < 1{-}\delta^{(n,s)}$, то $\square_2$: считать $f_i^{(n)}{<}1$;
- если $|\hat{f}_i^{(n)}{-}1| \leqslant \delta^{(n,s)}$, то $\circlearrowleft$: продолжить испытания,

(18)

Значения $\delta^{(n,s)}$, $n, s = 1, 2, \ldots$ для (18) определим, как для (11), задав соответственно верхние границы $\alpha^{(s)}$, $s = 1, 2, \ldots$, вероятностей ошибочных решений $\square_1$ и $\square_2$ и получив равенства (12).

Наконец, если наблюденное значение

$$\hat{f}_i^{(n)} \in \left[1 - \left(\frac{2}{n}\ln\frac{1}{\alpha^{(s)}}\right)^{\frac{1}{2}},\ 1 + \left(\frac{2}{n}\ln\frac{1}{\alpha^{(s)}}\right)^{\frac{1}{2}}\right] \triangleq \triangleq \mathrm{I}^{(n,s)}, \qquad (19)$$

то, согласно (18), принимается решение $\circlearrowleft$, и испытания должны быть продолжены, причем *при изменяющейся вероятности событие* $\{\hat{f}_i^{(n)} \in \mathrm{I}^{(n,s)}\}$, *вообще говоря, может выполняться для бесконечно многих* $n = 1, 2, \ldots$

Требования к модели испытаний (13), гарантирующие, что произойдет п. н. конечное число событий (19), и алгоритм (18) будет завершен на основе данных п. н. конечного числа испытаний, сформулированы в следующей лемме.

Лемма 3. *Пусть при всех достаточно больших* n *и всех* $i = 1, \ldots, s$ $|f_i^{(n)}{-}1| \geqslant \delta^{(n,s)}(1{+}\varepsilon_{n,s})$, *где* $\delta^{(n,s)}$ *определены в (12),* $\varepsilon_{n,s} > 0$ *и удовлетворяют условиям* $\sum\limits_{n=1}^{\infty}(\alpha^{(s)})^{\varepsilon_{n,s}^2} < \infty$, *в которых* $\alpha^{(s)} = \alpha/s$, $s = 1, 2, \ldots$ *Тогда для каждого* $s = 1, 2, \ldots$ *происходит п. н. конечное число событий (19).*

Если модель испытаний удовлетворяет условиям, сформулированным в лемме 3 (условия леммы 3, очевидно, выполнены, если среди вероятностей $\mathrm{Pr}_1, \mathrm{Pr}_2, \ldots$ конечное число различных), то теорема 2 верна и в случае вероятности, изменяющейся от испытания к испытанию, если существует $e = 0{,}e_1e_2\ldots$, для которого условия (14) и (15) выполнены для всех $j = 1, 2, \ldots$

Выводы

В то время, как при эмпирическом оценивании вероятности, контролирующей результаты испытаний, необходимо, чтобы последняя была *зафиксирована условиями испытаний, при восстановлении возможности условия испытаний должны фиксировать один из классов* $\mathbb{P}\mathrm{r}_{(e)}$ в (8), *в пределах которого вероятность, контролирующая результаты наблюдений, может произвольно изменяться от наблюдения к наблюдению.* При известном *условии регулярности последней класс* $\mathbb{P}\mathrm{r}_{(e)}$ *восстанавливается безошибочно на основе п. н. конечного числа испытаний* [1].

Этот факт существенно расширяет класс стохастических объектов, математическая модель которых может быть построена эмпирически, расширяет за счет тех из них, для которых эмпирически может быть построена возможностная модель, а вероятностная — нет.

Литература

[1] *Пытьев Ю. П.* Возможность как альтернатива вероятности. Математические и эмпирические основы, применение. — М: Физматлит, 2007. — 464 с.

[2] *Биркгофф Г.* Теория решеток. — М: Наука, 1984. — 566 с.

[3] *Пытьев Ю. П.* Математические методы и алгоритмы эмпирического восстановления стохастических и нечетких моделей. // Интеллектуальные системы. — 2008. — Т. 11, № 1–4. — С. 227–329.

[4] *Hoeffding W.* Probability of sums of bounded random variables. // J. Amer. Statist. Assoc. — 2963. — v. 58, № 301. — Pp. 213–226.

[6] Поскольку для любых $A \in \mathcal{P}(\Omega)$ и $\varepsilon > 0$

$$\lim_{N\to\infty} \mathrm{Pr}^{\infty}\left(\sup_{n\geqslant N}\left|\nu^{(n)}(A) - \frac{1}{n}\sum_{j=1}^{n}\mathrm{Pr}_j(A)\right| > \varepsilon\right) = 0,$$

то есть $\nu^{(n)}(A) - \frac{1}{n}\sum\limits_{j=1}^{n}\mathrm{Pr}_j(A) \xrightarrow{\text{п. н.}} 0$ при $n \to \infty$.

Эмпирическое восстановление неопределенной нечеткой модели*

Фаломкина О. В., Пытьев Ю. П.
olesya.falomkina@gmail.com
Москва, МГУ им. М. В. Ломоносова

В докладе рассмотрены математические методы и адаптивные алгоритмы эмпирического построения неопределенных нечетких (НН) моделей стохастических объектов, в том числе таких, для которых эмпирическое построение вероятностных моделей принципиально невозможно.

Теория вероятностей, как математическая основа модели случайности, широко используется в прикладных и теоретических исследованиях благодаря двум ее фундаментальным аспектам: математическому, поскольку теория вероятностей есть специальный раздел фундаментальной теории меры и интеграла, и эмпирическому, основанному на простых, математически обоснованных, процедурах, позволяющих на основе данных событийно-частотных наблюдений получать сколь угодно точные аппроксимации вероятности, математического ожидания и т. п., с одной стороны, а с другой — как предсказывать (прогнозировать), так и эмпирически проверять предсказанные значения математически аккуратно определенных характеристик случайных процессов и явлений.

Тем не менее, теоретико-вероятностные методы де-факто оказались неэффективными при моделировании сложных физических, технических, социальных и экономических объектов, субъективных суждений и т. д.

Причины неэффективности вероятностных методов обусловлены принципиальными трудностями, возникающими при эмпирическом построении и верификации моделей названных объектов. Во-первых, последние зачастую не имеют хорошо определенной стохастической компоненты, а в тех случаях, когда ее удается выделить, возникают серьезные проблемы с построением и проверкой адекватности ее теоретико-вероятностной модели. Дело прежде всего в том, что в процессе построения модели объект и его окружение эволюционируют, их вероятностные характеристики, как правило, изменяются, и их оценки оказываются неадекватными. Во-вторых, даже если стохастическая природа объекта и его «стационарность» ясны, эмпирическое построение с приемлемой точностью его вероятностной модели может оказаться нереализуемым из-за необходимого объема наблюдений. В-третьих, если достаточно точная модель будет построена, она может оказаться настолько сложной, что проблемным окажется ее использование на практике[1].

Многие из перечисленных трудностей, возникающих при эмпирическом построении модели стохастического объекта, могут быть преодолены, если эмпирически восстанавливать его теоретико-возможностную (нечеткую) модель, в которой возможность (как и вероятность) является мерой, определяемой свойствами объекта, значение которой при любом наблюдении оценивает шанс (предопределенность) проявления каждого из этих свойств по сравнению с шансами проявления остальных свойств; при этом численные значения возможности не важны, имеет смысл их упорядоченность [1]. Методы и адаптивные алгоритмы эмпирического построения теоретико-возможностных моделей стохастических объектов подробно рассмотрены в [2], см. также доклад [4].

В докладе рассмотрены математические методы и адаптивные алгоритмы эмпирического восстановления неопределенных нечетких (НН) моделей стохастических объектов, в том числе таких, вероятностные модели которых принципиально не могут быть построены эмпирически. В НН моделях [3] нечеткость, неточность формулировок, относящаяся к содержанию информации, охарактеризована в терминах значений *мер возможности* и (или) *необходимости*, а их достоверность, истинность которых не может быть абсолютной в силу принципиальной неполноты знаний, охарактеризована в терминах значений *мер правдоподобия* и (или) *доверия*. В докладе предложены методы эмпирического восстановления названных мер, основанные на данных событийно-частотных наблюдений.

Неопределенные нечеткие элементы

Обозначим $(Y, \mathcal{P}(Y), \mathrm{P})$ — пространство с возможностью [4], в котором Y — множество элементарных событий, $\mathcal{P}(Y)$ — класс всех подмножеств Y, называемых событиями, $\mathrm{P}\colon \mathcal{P}(Y) \to [0,\,1]$ — мера возможности (возможность).

Возможность $\mathrm{P}(\cdot)$ определяется ее значениями $f^\eta(y) \triangleq \mathrm{P}(\{y\}) = \mathrm{P}(\eta = y),\ y \in Y$, на одноточечных подмножествах $\{y\} \subset Y$, а именно, $\mathrm{P}(A) \triangleq \triangleq \mathrm{P}(\eta \in A) = \sup\limits_{y \in A} f^\eta(y),\ A \in \mathcal{P}(Y)$.

Функция $f^\eta\colon Y \to [0,1]$ называется распределением возможностей значений канонически-

*Работа выполнена при финансовой поддержке РФФИ, проект № 08-07-00133-а

[1] Разумеется, речь не идет о стохастических объектах, динамические модели которых могут быть априори охарактеризованы математически, а эмпирически лишь уточнены и верифицированы.

го для $(Y, \mathcal{P}(Y), \mathrm{P})$ нечеткого элемента $\eta\colon Y \to \to (Y, \mathcal{P}(Y), \mathrm{P}(\cdot))$.

Обозначим аналогично $(\mathcal{U}, \mathcal{P}(\mathcal{U}), \mathrm{P}l(\cdot))$ — *пространство с правдоподобием*, в котором $\mathcal{U}$ — множество элементарных высказываний, $\mathcal{P}(\mathcal{U})$ — класс всех подмножеств $\mathcal{U}$ (высказываний), $\mathrm{P}l(\cdot)\colon \mathcal{P}(\mathcal{U}) \to [0, 1]$ — мера правдоподобия. Аналогично возможности $\mathrm{P}(\cdot)$, правдоподобие $\mathrm{P}l(\cdot)$ определяется распределением правдоподобий $g^{\tilde{u}}(\cdot)\colon \mathcal{U} \to [0, 1]$ значений канонического неопределенного элемента $\tilde{u}\colon (\mathcal{U}, \mathcal{P}(\mathcal{U}), \mathrm{P}l(\cdot)) \to \mathcal{U}$.

Определение 1 ([3]). *Неопределенным нечетким (НН) элементом, принимающим значения в X, называется образ $\tilde{\xi} \triangleq q(\eta, \tilde{u})$ (упорядоченной) пары $(\eta, \tilde{u})$ — нечеткого η и неопределенного $\tilde{u}$ элементов при отображении $q\colon Y \times \mathcal{U} \to X$.*

Пространства с возможностью $(Y, \mathcal{P}(Y), \mathrm{P})$ и с правдоподобием $(\mathcal{U}, \mathcal{P}(\mathcal{U}), \mathrm{P}l(\cdot))$ назовем НН моделью. В докладе рассмотрен адаптивный алгоритм эмпирического восстановления $(Y, \mathcal{P}(Y), \mathrm{P})$ и $(\mathcal{U}, \mathcal{P}(\mathcal{U}), \mathrm{P}l(\cdot))$.

Алгоритм эмпирического восстановления НН модели

Обозначим СЭ — стохастический эксперимент, вероятностная модель которого определена в [4], где $\Omega = \{\omega_1, \ldots, \omega_s\}$. В обозначениях, принятых в [4], алгоритм эмпирического восстановления НН модели СЭ формулируется следующим образом: для всех $i = 1, \ldots, s$ и каждого $n = 1, 2, \ldots$

$$\begin{array}{l} \bullet \text{ если } \hat{f}_i^{(n)} > 1, \text{ то } \square_1\text{: считать } f_i > 1; \\ \bullet \text{ если } \hat{f}_i^{(n)} < 1, \text{ то } \square_2\text{: считать } f_i < 1, \end{array} \tag{1}$$

где

$$f_i \triangleq \mathrm{pr}_1 + \ldots + \mathrm{pr}_{i-1} + 2\mathrm{pr}_i, \quad i = 1, 2, \ldots, \tag{2}$$

$$\hat{f}_i^{(n)} \triangleq \nu_1^{(n)} + \ldots + \nu_{i-1}^{(n)} + 2\nu_i^{(n)}, \quad i, n = 1, 2, \ldots \tag{3}$$

При каждом $n = 1, 2, \ldots$ определим для каждого $i = 1, \ldots, s$ случайную величину $\hat{\delta}^{(n,i)} = |\hat{f}_i - f_i|$, при которой алгоритм (11) из [4]: для всех $i = 1, \ldots, s$ и каждого $n = 1, 2, \ldots$

$$\begin{array}{l} \bullet \text{ если } \hat{f}_i^{(n)} > 1 + \delta^{(n,s)}, \text{ то } \square_1\text{: считать } f_i > 1; \\ \bullet \text{ если } \hat{f}_i^{(n)} < 1 - \delta^{(n,s)}, \text{ то } \square_2\text{: считать } f_i < 1; \\ \bullet \text{ если } |\hat{f}_i^{(n)} - 1| \leqslant \delta^{(n,s)}, \text{ то } \circlearrowleft\text{: продолжить испытания}, \end{array} \tag{4}$$

принял бы одно из решений $\square_1$ или $\square_2$, и зададим «порог» $\delta^{(n)}$, $n = 1, 2, \ldots$, определяющий «правило остановки» алгоритма и вероятность ошибочного восстановления модели. «Правило остановки» имеет вид $\min\limits_{1 \leqslant i \leqslant s} \hat{\delta}^{(n,i)} \geqslant \delta^{(n)}$. Минимальное $n = N_0$, при котором выполняется данное правило, определяет количество наблюдений, при котором вероятность ошибочного восстановления модели не больше заданного $\alpha = \exp\bigl(-\frac{1}{2} n (\delta^{(n)})^2\bigr)$, см (12) в [4].

Согласно [4], чем больше $\hat{\delta}^{(n,i)}$, тем более «правдоподобно» принятое алгоритмом (1) решение $\square_1$ или $\square_2$. Обозначим правдоподобие принятого решения $\square_1$ или $\square_2$ $\mathrm{P}l_i^{(n)}$, $i = 1, \ldots, s$, $n = 1, 2, \ldots$ Согласно свойствам меры правдоподобия [3], правдоподобие совокупности принятых алгоритмом (1) решений $\square_1$ или $\square_2$ при всех $i = 1, \ldots, s$ и любом $n = 1, 2, \ldots$ $\mathrm{P}l^{(n)} \leqslant \min\limits_{1 \leqslant i \leqslant s} \mathrm{P}l_i^{(n)}$.

Заметим, что $\mathrm{P}l^{(n)}$ не выражается через значения $\mathrm{P}l_i^{(n)}$, поскольку события «при $i = 1, \ldots, s$ приняты решения $\square_1$ или $\square_2$» не независимы.

Дуальная $\mathrm{P}l(\cdot)$ мера доверия $\mathrm{Bel}^{\tilde{u}}(\cdot)$ истинности утверждения, согласно которому $\tilde{u} \in A$, определяется равенством [3]

$$\mathrm{Bel}^{\tilde{u}}(A) \equiv \mathrm{Bel}(\tilde{u} \in A) = \vartheta\bigl(\mathrm{P}l^{\tilde{u}}(\mathcal{U} \backslash A)\bigr), \; A \in \mathcal{P}(A), \tag{5}$$

где функция $\vartheta(\cdot)\colon [0, 1] \to [0, 1]$ непрерывна, строго монотонно убывает, $\vartheta(0) = 1$, $\vartheta(1) = 0$.

Между правдоподобием и доверием имеют место следующие связи: $\forall A \in \mathcal{P}(\mathcal{U})$

$$\begin{gathered} \max(\mathrm{P}l^{\tilde{u}}(A), \mathrm{P}l^{\tilde{u}}(\mathcal{U} \setminus A)) = 1 \Leftrightarrow \\ \Leftrightarrow \min(\mathrm{Bel}^{\tilde{u}}(A), \mathrm{Bel}^{\tilde{u}}(\mathcal{U} \setminus A)) = 0, \\ \mathrm{P}l^{\tilde{u}}(A) < 1 \;\Rightarrow\; \mathrm{P}l^{\tilde{u}}(\mathcal{U} \setminus A) = 1 \;\Leftrightarrow\; \mathrm{Bel}^{\tilde{u}}(A) = 0, \\ \mathrm{Bel}^{\tilde{u}}(A) > 0 \;\Leftrightarrow\; \mathrm{P}l^{\tilde{u}}(\mathcal{U} \setminus A) < 1 \;\Rightarrow\; \mathrm{P}l^{\tilde{u}}(A) = 1 \end{gathered}$$

и, как следствие, $\mathrm{Bel}^{\tilde{u}}(A) \leqslant \mathrm{P}l^{\tilde{u}}(A)$.

Согласно [4], чем больше $\hat{\delta}^{(n,i)}$, тем больше следует доверять принятому алгоритмом (1) решению $\square_1$ или $\square_2$, $i = 1, 2, \ldots, s$, $n = 1, 2, \ldots$ Обозначим доверие к принятому решению $\square_1$ или $\square_2$ $\mathrm{Bel}_i^{(n)}$, $i = 1, 2, \ldots, s$, $n = 1, 2, \ldots$, $\mathrm{Bel}^{(n)}$ — доверие к совокупности таких решений, принятых при всех $i = 1, \ldots, s$ и фиксированном n. Как известно [3], $\mathrm{Bel}^{(n)} = \min\limits_{1 \leqslant i \leqslant s} \mathrm{Bel}_i^{(n)}$, поэтому значение $\mathrm{Bel}^{(n)}$ определено значением $\hat{\delta}^{(n)} = \min\limits_{1 \leqslant i \leqslant s} \delta^{(n,i)}$.

Тем самым определена мера доверия к возможностной модели СЭ, определенной алгоритмом (1), т. е. восстановлена его НН модель.

Литература

[1] *Пытьев Ю. П.* Возможность как альтернатива вероятности. Математические и эмпирические основы, применение. — М.: Физматлит, 2007. — 464 с.

[2] *Пытьев Ю. П.* Математические методы и алгоритмы эмпирического восстановления стохастических и нечетких моделей. // Интеллектуальные системы. — 2008. — Т. 11, № 1–4. — С. 227–329.

[3] *Пытьев Ю. П.* Неопределенные нечеткие модели и их применения // Интеллектуальные системы. — 2004. — № 8, Вып. 1–4. — С. 147–310.

[4] *Пытьев Ю. П.* Возможность как альтернативная вероятности модель случайности: событийно-частотная интерпретация и эмпирическое построение // ММРО-14. — 2009. — С. 60–63.

Точные оценки вероятности переобучения для симметричных семейств алгоритмов*

Фрей А. И.
frey@forecsys.ru
Московский Физико-технический институт

В комбинаторном подходе к проблеме переобучения основной задачей является получение вычислительно эффективных формул для вероятности переобучения и вероятности получить каждый из имеющихся алгоритмов в результате обучения. Предлагается подход, который позволяет проще выводить такие формулы в тех случаях, когда множество алгоритмов наделено некоторой группой симметрий. Приводятся примеры подобных ситуаций. Даётся определение рандомизированного метода обучения, для которого доказывается общая оценка вероятности переобучения.

Введение

При обучении алгоритмов классификации и прогнозирования по конечным выборкам часто возникает проблема переобучения, когда качество алгоритма, построенного по наблюдаемой обучающей выборке, оказывается значительно хуже на скрытой контрольной выборке.

В работах [1, 2, 4] рассматривался метод *минимизации эмпирического риска*. Он заключается в том, что из заданного множества (семейства) алгоритмов выбирается алгоритм, допускающий наименьшее число ошибок на обучающей выборке.

В следующей таблице показан пример, когда минимизация эмпирического риска приводит к переобучению. Столбцы таблицы соответствуют алгоритмам, строки — объектам генеральной выборки, единица в $[i,d]$-й ячейке таблицы означает, что алгоритм a_d допускает ошибку на объекте x_i. Первые три объекта составляют обучающую выборку, оставшиеся три — контрольную.

	a_1	a_2	…	a_d	…	a_D
x_1	0	1	…	**0**	…	1
x_2	1	1	…	**0**	…	0
x_3	0	0	…	**0**	…	0
x_4	1	1	…	**1**	…	1
x_5	1	0	…	**1**	…	0
x_6	0	0	…	**1**	…	0

В данном примере переобучение могло быть следствием «неудачного» разбиения генеральной выборки на обучение и контроль. Поэтому вводится функционал *вероятности переобучения*, равный доле разбиений выборки, при которых возникает переобучение [4, 3]. Этот функционал инвариантен относительно выбора разбиения и характеризует качество данного метода обучения на данной генеральной выборке.

Для некоторых семейств простой структуры (монотонных и унимодальных цепочек и h-мерных сеток) в [3, 5] найдены точные выражения вероятности переобучения.

В данной работе вводится понятия группы симметрий множества алгоритмов и рандомизированного метода обучения, позволяющие обобщить эти оценки для тех случаев, когда семейство алгоритмов наделено определённой симметрией.

Группа симметрий множества алгоритмов

Пусть задана генеральная выборка $\mathbb{X} = \left(x_i\right)_{i=1}^{L}$, состоящая из L объектов. Произвольный бинарный вектор $a \equiv \left(a(x_i)\right)_{i=1}^{L}$ длины L будем называть *алгоритмом*, и в случае $a(x_i) = 1$ говорить, что алгоритм a допускает ошибку на объекте x_i.

Обозначим через $\mathbb{A} = \{0,1\}^L$ множество всех возможных алгоритмов, через $\mathfrak{A} = 2^{\mathbb{A}}$ — множество всех возможных подмножеств (семейств) алгоритмов. Заметим, что $|\mathbb{A}| = 2^L$, $|\mathfrak{A}| = 2^{2^L}$.

Для большей наглядности дальнейших определений проведем следующую аналогию: пусть алгоритмы соответствуют точкам плоскости, множества алгоритмов — плоским фигурам. Зафиксируем некоторую группу преобразований плоскости (например, группу всевозможных движений). Тогда группа симметрии произвольной фигуры определяется как подгруппа, не изменяющая фигуру как множество точек плоскости.

В большинстве задач обучения по прецедентам порядок объектов в выборке не имеет значения. Поэтому в качестве исходной группы преобразований мы возьмем симметрическую группу S_L, элементы которой очевидным образом действуют как перестановки объектов и на генеральную выборку $\mathbb{X}$, и на произвольный алгоритм $a \in \mathbb{A}$, и (поэлементно) на произвольное множество алгоритмов $A \in \mathfrak{A}$.

Определение 1. *Группой симметрий $S(A)$ множества алгоритмов $A \in \mathfrak{A}$ будем называть его стационарную подгруппу:*

$$S(A) = \{\pi \in S_L : \pi(A) = A\}.$$

Каждый элемент группы симметрий $\pi \in S(A)$ переставляет алгоритмы a внутри множества A.

*Работа поддержана РФФИ (проект №08-07-00422) и программой ОМН РАН «Алгебраические и комбинаторные методы математической кибернетики и информационные системы нового поколения».

Значит, для любого $a \in A$ и любого $\pi \in S(A)$ выполнено $\pi(a) \in A$. Поэтому для группы $S(A)$, в отличии от всей группы S_L, естественным образом определено действие на множестве A.

Орбитой элемента m множества M, на котором действует группа G, называется подмножество $Gm = \{gm\colon g \in G\} \subset M$. Орбиты двух элементов m_1 и m_2 либо не пересекаются, либо совпадают. Это позволяет говорить о разбиении множества M на непересекающиеся орбиты: $M = Gm_1 \sqcup \ldots \sqcup Gm_k$.

Определение 2. *Орбиты действия группы симметрий $S(A)$ на множестве алгоритмов A будем называть классами идентичных алгоритмов.*

Совокупность всех орбит множества алгоритмов A обозначим через $\Omega(A)$. Представителя орбиты $\omega \in \Omega(A)$ будем обозначать через $a_\omega \in A$. Различных представителей одной и той же орбиты будем называть *идентичными алгоритмами*.

Согласно данному выше определению *алгоритм* $a \equiv \big(a(x_i)\big)_{i=1}^L$ является вектором, следовательно, зависит от нумерации объектов выборки. Однако ни группа симметрий $S(A)$, ни разбиение на классы идентичных алгоритмов $\Omega(A)$, уже не зависят от этой нумерации.

Теорема 1. *Для любого множества алгоритмов $A \in \mathfrak{A}$ и любой перестановки $\pi \in S_L$ группы $S(A)$ и $S(\pi(A))$ сопряжены: $S(\pi(A)) = \pi \circ S(A) \circ \pi^{-1}$.*

Сопряжение устанавливает изоморфизм групп $S(A)$ и $S(\pi(A))$. Остается лишь проверить, что действие изоморфных групп на множествах A и $\pi(A)$ действительно приведет к «одинаковому» разбиению на орбиты.

В следующей таблице приведен пример унимодальной цепочки [5]. Алгоритм a_0 является первым (и наилучшим) в цепочке; a_1, a_2, a_3 составляют левую ветвь; a_4, a_5, a_6 — правую.

	a_0	a_1	a_2	a_3	a_4	a_5	a_6
x_1	0	1	1	1	0	0	0
x_2	0	0	1	1	0	0	0
x_3	0	0	0	1	0	0	0
x_4	0	0	0	0	1	1	1
x_5	0	0	0	0	0	1	1
x_6	0	0	0	0	0	0	1

Перенумерацией объектов выборки ($x_1 \leftrightarrow x_4$, $x_2 \leftrightarrow x_5$, $x_3 \leftrightarrow x_6$) можно поменять левую и правую ветвь местами. Поэтому группой симметрии данного семейства является группа перестановок из двух элементов S_2. Идентичные алгоритмы в унимодальной цепочке — это пары алгоритмов с равным числом ошибок на полной выборке.

Естественно потребовать, чтобы идентичные алгоритмы имели равные шансы реализоваться при обучении и давали равный вклад в вероятность переобучения. В следующем параграфе мы обсудим связанные с этим требованием ограничения и предложим рандомизированный метод обучения, при котором это действительно так.

Рандомизированный метод обучения

При минимизации эмпирического риска может возникать неоднозначность — несколько алгоритмов могут иметь одинаковое число ошибок на обучающей выборке. В [4] для устранения неоднозначности и получения точных верхних оценок вероятности переобучения использовалась *пессимистичная* минимизация эмпирического риска — предполагалось, что в случае неоднозначности выбирается алгоритм с наибольшим числом ошибок на генеральной выборке $\mathbb{X}$. Это не устраняет неоднозначность окончательно. Возможны ситуации, когда несколько алгоритмов имеют наименьшее число ошибок на обучающей выборке X и одинаковое число ошибок на генеральной выборке $\mathbb{X}$. В таких случаях на множестве алгоритмов вводился линейный порядок, и среди неразличимых алгоритмов выбирался алгоритм с бóльшим номером. Введение приоритетности алгоритмов является искусственным приёмом, не имеющим адекватных аналогов среди известных методов обучения.

Ниже вводится рандомизированный метод обучения, лишенный этого недостатка.

Обычно метод обучения — это отображение, которое произвольной выборке X ставит в соответствие определённый алгоритм $a \in A$. Рандомизированный метод произвольной выборке X ставит в соответствие функцию распределения весов на множестве A. Эта функция нормирована так, что её можно интерпретировать как вероятность получить данный алгоритм в результате обучения.

Обозначим через $[\mathbb{X}]^\ell$ множество всех ℓ-элементных подмножеств $\mathbb{X}$. Каждый $X \in [\mathbb{X}]^\ell$ фиксирует разбиение генеральной выборки $\mathbb{X}$ на обучающую выборку X и контрольную выборку $\bar{X} = \mathbb{X}\backslash X$. Через $n(a, X) = \sum\limits_{x \in X} a(x)$ обозначим число ошибок алгоритма $a \in \mathbb{A}$ на множестве $X \subset \mathbb{X}$. Под действием $\pi(X)$ понимается поэлементное действие отображения $\pi\colon \mathbb{X} \to \mathbb{X}$ на каждый объект обучающей выборки: $\pi(X) = \{\pi(x)\colon x \in X\}$.

Определение 3. *Рандомизированным методом обучения будем называть отображение вида*

$$\mu : \mathfrak{A} \times [\mathbb{X}]^\ell \times \mathbb{A} \to [0, 1],$$

удовлетворяющее при любых $A \in \mathfrak{A}$, $X \in [\mathbb{X}]^\ell$, $a, b \in A$ и $\pi \in S_L$ следующим условиям:

1) $\sum\limits_{a \in A} \mu(A, X, a) = 1$;
2) $n(a, X) = n(b, X) \;\to\; \mu(A, X, a) = \mu(A, X, b)$;
3) $\mu(A, X, a) = \mu\big(\pi(A), \pi(X), \pi(a)\big)$.

Первое условие означает «вероятностную» нормировку весов алгоритмов. Кроме того, оно обеспе-

чивает нулевую «вероятность» алгоритмам, не принадлежащих множеству A.

Второе условие означает, что при любом разбиении $\mathbb{X} = X \sqcup \bar{X}$ вероятность получить алгоритм в результате обучения зависит только от количества ошибок алгоритма на обучении.

Третье условие означает, что метод обучения не учитывает порядок объектов в выборке.

Частотой ошибок алгоритма a на выборке $X \subset \mathbb{X}$ называется величина $\nu(a, X) = \frac{1}{|X|} n(a, X)$. Введем обозначение для разности частот ошибок алгоритма на контрольной и обучающей выборке: $\delta(a, X) = \nu(a, \bar{X}) - \nu(a, X)$.

Вероятностью получить алгоритм $a \in A$ в результате обучения назовем величину

$$P(a, A) = \frac{1}{C_L^\ell} \sum_{X \in [\mathbb{X}]^\ell} \mu(A, X, a).$$

Для произвольного $\varepsilon \in [0, 1]$ определим *вклад* алгоритма $a \in A$ в вероятность переобучения:

$$Q_\varepsilon(a, A) = \frac{1}{C_L^\ell} \sum_{X \in [\mathbb{X}]^\ell} \mu(A, X, a) \big[\delta(a, X) \geqslant \varepsilon\big].$$

и саму *вероятность переобучения*:

$$Q_\varepsilon(A) = \sum_{a \in A} Q_\varepsilon(a, A).$$

Следующая теорема утверждает, что функционалы $Q(A)$ и $P(A, a)$ не зависят от нумерации объектов выборки.

Теорема 2. *Вероятность получить алгоритм a в результате обучения, а также вероятность переобучения сохраняются при одновременном применении произвольной перестановки $\pi \in S_L$ к множеству A и алгоритму a:*

$$Q(\pi(A)) = Q(A),$$
$$P(\pi(A), \pi(a)) = P(A, a).$$

Теперь можно сформулировать в виде двух теорем основной результат данного раздела:

Теорема 3. *Вероятность получить идентичные алгоритмы в результате обучения одинакова:*

$$\forall \pi \in S(A) \text{ выполнено } P(\pi(a), A) = P(a, A).$$

Напомним, что $\Omega(A)$ — множество классов идентичных алгоритмов, $a_\omega \in A$ — произвольный представитель класса $\omega \in \Omega(A)$.

Теорема 4. *Идентичные алгоритмы дают равный вклад в вероятность переобучения:*

$$Q_\varepsilon(A) = \sum_{\omega \in \Omega(A)} |\,\omega|\, Q_\varepsilon(A, a_\omega).$$

Вероятность переобучения для семейств простой структуры

Определим метод минимизации эмпирического риска, удовлетворяющий определению 3. Выделим множество алгоритмов, допускающих минимальное число ошибок на обучающей выборке:

$$E_{A,X} = \operatorname*{Argmin}_{a \in A} n(a, X).$$

В результате обучения методом минимизации эмпирического риска все алгоритмы вне этого множества получают нулевой вес:

$$\mu(A, X, a) = \begin{cases} \frac{1}{|E_{A,X}|}, & a \in E_{A,X}; \\ 0, & a \notin E_{A,X}. \end{cases}$$

Монотонной цепочкой называется последовательность алгоритмов, в которой каждый следующий алгоритм допускает ошибки на тех же объектах, что предыдущий, и ещё на каком-то одном объекте.

Связкой из p монотонных цепочек называется множество алгоритмов, полученное объединением p штук монотонных цепочек равной длины («ветвей»), с общим первым алгоритмом, при условии, что множества объектов, на которых ошибаются алгоритмы ветвей, не пересекаются.

Связка из двух ветвей называется *унимодальной цепочкой*. Заметим, что монотонные и унимодальные цепочки можно рассматривать как модели однопараметрических семейств алгоритмов классификации с непрерывной по параметру разделяющей поверхностью [4, 3].

Нетрудно установить, что группа симметрии связки из p монотонных цепочек является симметрической группой S_p, действующей на ветви связки всевозможными перестановками. Таким образом, классы идентичных алгоритмов — это подмножества алгоритмов с одинаковым числом ошибок на полной выборке, называемые *слоями* [4].

В следующей теореме мы получим явную формулу вероятности переобучения для связки из p монотонных цепочек. При этом нам понадобится комбинаторный коэффициент $R^h_{D,p}(S, F)$, который зависит от параметров S и F, от числа монотонных цепочек p и от их длины D, а также от h — минимального значения параметра S. Коэффициент $R^h_{D,p}(S, F)$ равен числу способов представить число S в виде суммы p неотрицательных слагаемых, $S = t_1 + \ldots + t_p$, каждое из которых не превосходит D. При этом ровно F слагаемых не должно равняться D, а на первое слагаемое накладывается дополнительное ограничение $t_1 \geqslant h$.

Теорема 5. *Рассмотрим связку из p монотонных цепочек, в которой лучший алгоритм допускает m ошибок на полной выборке, длина каждой ветви*

без учета лучшего алгоритма — D. Тогда при обучении рандомизированным методом вероятность переобучения может быть записана в виде:

$$Q_\varepsilon(A) = \sum_{h=0}^{D} \sum_{S=h}^{pD} \sum_{F=0}^{p} \frac{|\omega_h| R^h_{D,p}(S,F)}{1+S} \frac{C^{\ell'}_{L'}}{C^{\ell}_{L}} H^{\ell',m}_{L'}(s(\varepsilon)),$$

где $L' = L-S-F$, $\ell' = \ell-F$, $s(\varepsilon) = \left\lfloor \frac{\ell}{L}(m+h-\varepsilon k) \right\rfloor$; $|\omega_h| = 1$ *при* $h = 0$ *и* $|\omega_h| = p$ *при* $h \geqslant 1$; $H^{\ell',m}_{L'}(s)$ — *функция гипергеометрического распределения [4].*

Связка из p монотонных цепочек является обобщением трёх частных случаев, рассмотренных в [3]: монотонной цепочки ($p = 1$), унимодальной цепочки ($p = 2$) и единичной окрестности лучшего алгоритма ($D = 1$). Вычисляя конкретные выражения комбинаторного коэффициента $R^h_{D,p}(S,F)$ для этих трех случаев, получим три следствия.

Следствие 1. *Для монотонной цепочки длины $D+1$ вероятность переобучения равна*

$$Q_\varepsilon = \sum_{h=0}^{D} \sum_{S=h}^{D} \frac{1}{1+S} \frac{C^{\ell'}_{L'}}{C^{\ell}_{L}} H^{\ell',m}_{L'}(s(\varepsilon)),$$

где $L' = L - S - [S{\neq}D]$, $\ell' = \ell - [S{\neq}D]$.

Следствие 2. *Для унимодальной цепочки с ветвями длины D вероятность переобучения равна*

$$Q_\varepsilon = \sum_{h=0}^{D} \sum_{t_1=h}^{D} \sum_{t_2=0}^{D} \frac{|\omega_h|}{1+S} \frac{C^{\ell'}_{L'}}{C^{\ell}_{L}} H^{\ell',m}_{L'}(s(\varepsilon)),$$

где $S = t_1 + t_2$, $F = [t_1{\neq}D] + [t_2{\neq}D]$, *остальные обозначения те же, что в теореме 5.*

Следствие 3. *Для единичной окрестности из $p+1$ алгоритма вероятность переобучения равна*

$$Q_\varepsilon(A) = \sum_{h=0}^{1} \sum_{S=h}^{p} \frac{|\omega_h| C^{S-h}_{p-h}}{1+S} \frac{C^{\ell'}_{L'}}{C^{\ell}_{L}} H^{\ell',m}_{L'}(s(\varepsilon)),$$

где $L' = L-p$, $\ell' = \ell+S-p$.

Численный эксперимент

На рис. 1 и 2 представлены результаты численных экспериментов. Из четырех кривых на каждом графике верхняя (жирная) соответствует пессимистической минимизации эмпирического риска [3, 4], нижняя — оптимистической. Две почти сливающиеся кривые между ними соответствуют рандомизированной минимизации эмпирического риска. Одна из них вычислена по доказанным формулам, вторая построена методом Монте-Карло по 10^5 случайных разбиений, при равновероятном выборе лучшего алгоритма в случаях неопределённости. Совпадение этих двух кривых подтверждает справедливость развитой выше теории.

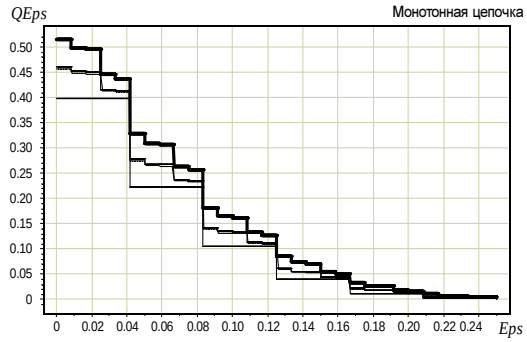


Рис. 1. Зависимость Q_ε от ε для монотонной цепочки при $L = 100$, $\ell = 60$, $D = 40$, $m = 20$.

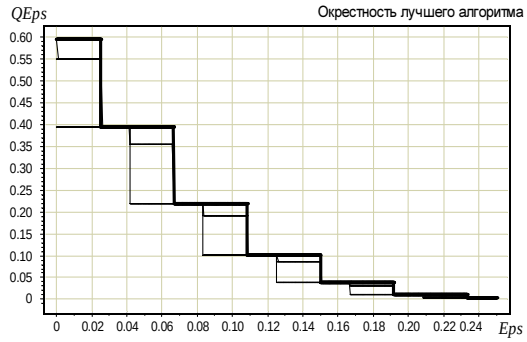


Рис. 2. Зависимость Q_ε от ε для единичной окрестности при $L = 100$, $\ell = 60$, $p = 10$, $m = 20$.

Выводы

Свойство симметрии семейств алгоритмов позволяет упрощать получение вычислительно эффективных формул вероятности переобучения. В частности, удалось вывести оценки для монотонной и унимодальной цепочек, а также для единичной окрестности лучшего алгоритма как следствия одной общей теоремы, в то время как ранее аналогичные оценки доказывались независимо и при неестественном предположении об априорной упорядоченности алгоритмов в семействе.

Литература

[1] *Вапник В. Н., Червоненкис А. Я.* Теория распознавания образов. — М.: Наука, 1974.

[2] *Vapnik V.* Statistical Learning Theory. — New York: Wiley, 1998.

[3] *Воронцов К. В.* Точные оценки вероятности переобучения // Доклады РАН, 2009 (в печати).

[4] *Воронцов К. В.* Комбинаторный подход к проблеме переобучения // Всеросс. конф. ММРО-14 — М.: МАКС Пресс, 2009 — С. 18–21 (в настоящем сборнике).

[5] *Ботов П. В.* Точные оценки вероятности переобучения для монотонных и унимодальных семейств алгоритмов // Всеросс. конф. ММРО-14 — М.: МАКС Пресс, 2009 — С. 7–10 (в настоящем сборнике).

[6] *Винберг Э. Б.* Курс алгебры // М.: Факториал Пресс, 2001. — 544 с.

О равновесии и неравновесии*

Хачай М. Ю., Мазуров Вл. Д., Шарф В. С.
vmazurov@imm.uran.ru
Екатеринбург, Институт математики и механики УрО РАН

В работе исследуется подход к описанию неравновесных ситуаций, в частности, исторических и экономических, с точки зрения циклов максимальных по включению совместных подсистем подходящих систем ограничений, неравенств или уравнений. Отмечается взаимосвязь условий существования простых циклов в графах максимальных совместных подсистем и комитетных решений таких систем. Отдельно исследуется структура графа максимальных совместных подсистем равномерно распределённой (по Гейлу) системы линейных неравенств.

Введение

Фундаментальное понятие равновесия пришло в экономику (как и в другие науки) из механики. Сейчас кризис, т. е. равновесие отсутствует. Но равновесие ведь всё равно наступит, а если не равновесие, то обход равновесия, обобщённое равновесие и так далее. Один из видов обобщённых равновесий — циклы. Один из видов циклов — циклы совместных подсистем (понятие введено нами).

Главная причина циклов — отсутствие «рога изобилия» (есть такая аксиома в математической экономике — она означает, что коэффициент полезного действия меньше единицы). Можно сказать, что все всегда живут в стеснённых обстоятельствах. Распоряжаться слишком ограниченными ресурсами — значит решать несовместные системы уравнений и неравенств. Решать — в том смысле, что находить решения максимальных совместных подсистем (м.с.п.) подходящих систем ограничений и потом циклически их реализовывать.

Равновесные циклы максимальных совместных подсистем в модели Вальраса можно использовать, когда система нелинейных уравнений в этой модели несовместна. В более общей модели вместо уравнений используем неравенства. Тогда также нужно использовать м.с.п. системы ограничений. В модели Леонтьева рассматриваются графы продуктивных матриц и матриц косвенных затрат. Более общий подход — использование графов м.с.п.

Когда некоторая система, работающая в определённом режиме, накапливает энтропию, она должна перейти к другому режиму, чтобы предотвратить своё саморазрушение. В случае изолированной системы состояние равновесия отвечает максимуму её энтропии. Энтропия — мера хаоса. Однако равновесие подвижно. Мы можем использовать для моделирования этой ситуации формулировку задачи выполнения ограничений. Как правило, эта задача несовместна, и тогда мы используем максимальные совместные подсистемы. От состояния, отвечающего данной м.с.п., мы переходим к соседней м.с.п.. Мы опираемся на доказанный ранее факт, что граф м.с.п. системы линейных неравенств связен.

Противоречивость системы условий объясняет цикличность многих процессов, в том числе и социальных. Циклическое равновесие — обобщение классического статического равновесия.

Противоречивость соотношений математической модели, даже относящейся к одному моменту, ведёт от статичности к процессуальности. К динамике. В частности, несовместность ведёт к цикличности. Логика разрешения противоречий динамизирует объекты и ситуации. Противоречия порождают развитие. Есть закономерности образования циклов при эволюции систем. Внешне циклы могут описываться, например, дифференциальными уравнениями. Но при этом не вскрываются фундаментальные причины циклического поведения сложных систем. Одна из причин циклов — противоречивые условия, при которых система может находиться в одном из классов состояний, отвечающих непротиворечивым подсистемам условий. Есть динамика циклического пробегания по максимальным совместным подсистемам системы условий. Другая причина — неустойчивость. При казалось бы тех же условиях система ведёт себя иначе.

Такого сорта явления известны в экономике, в химии, в биологии. При этом в биологии динамика, обусловленная противоречивостью, использует обратную связь: попав в очередную м.с.п., система определяет тот момент, когда надо переходить к другой м.с.п.. Циклы решений тупиковых (максимальных по включению) совместных подсистем наблюдаются также при неоднозначной интерпретации противоречивых данных [1]. В частности, при неоднозначной интерпретации противоречивых изображений.

Явление максимальных совместных подсистем возникает ввиду противоречивого наложения друг на друга биоритмов человека и ритмов среды.

При этом противоположности, фигурирующие в противоречивой модели, могут находиться в различных отношениях.

В качестве примера можно рассмотреть динамику региона в условиях циклического развития экономических процессов. Динамические ряды со-

*Поддержано междисциплинарной программой УрО РАН «Историческая динамика России».

стояний характеризуются более или менее длительными циклами на фоне структурных межотраслевых взаимосвязей. Многие циклы, например, суточные циклы организма, вызваны несовместимостью всех целей, если их свести воедино: например, надо и действовать, и сохранять энергию, предаваться и созерцанию, и размышлениям.

Циклическое равновесие, обобщение классического равновесия, тесно связано с понятием комитетного решения, являющегося, в свою очередь, дискретным обобщением понятия решения системы ограничений на случай несовместности последней. Фактически, комитетным решением является конечный набор элементов пространства, над которым задана система, такой, что каждому ограничению удовлетворяет большинство элементов набора. При этом каждый отдельный элемент не обязан разрешать систему в целом.

Графы максимальных совместных подсистем и комитетные решения

Пусть заданы множество X и набор его непустых подмножеств $D_1, D_2, \dots, D_m$. Рассмотрим систему включений:

$$x \in D_j, \quad j \in \mathbb{N}_m = \{1, \dots, m\}. \tag{1}$$

Система (1) называется несовместной, если $\bigcap\limits_{j=1}^{m} D_j = \varnothing$. Ряд утверждений, приведенных ниже, справедлив для произвольной системы (1), однако большая часть результатов будет сформулирована для её частного случая — системы неравенств:

$$f_j(x) > 0, \quad j \in \mathbb{N}_m, \tag{2}$$

где X — вещественное линейное пространство,

$$f_1, \dots, f_m \in F \subset \{X \to \mathbb{R}\},$$

а F — заданный класс функций (линейных, аффинных и т. п.). Систему

$$x \in D_j, \quad j \in L, \tag{1$_L$}$$

для произвольного непустого $L \subseteq \mathbb{N}_m$ будем называть подсистемой системы (1) с индексным множеством (индексом) L; множество её решений обозначим через $D(L) = \bigcap\limits_{j \in L} D_j$. Подсистему с индексным множеством L, являющимся собственным подмножеством $\mathbb{N}_m$, договоримся также называть собственной.

Определение 1. *Подсистема* (**??**$_L$) *называется максимальной совместной подсистемой (м.с.п.) системы* (1), *если выполнены условия*

1) $D(L) \neq \varnothing$;
2) $D(L \cup \{j\}) = \varnothing$ *для каждого* $j \in \mathbb{N}_m \setminus L$.

Видно, что система (1) либо совместна, либо имеет собственные м.с.п. Перейдём к определениям комитетных конструкций.

Определение 2. *Комитетным решением (комитетом) системы* (1) *называется конечная последовательность* $Q = (x^1, \dots, x^q)$, $x^i \in X$, *такая, что для каждого* $j \in \mathbb{N}_m$

$$\big|\{i \colon x^i \in D_j\}\big| > q/2.$$

Нетрудно убедиться, что при поиске комитетных решений системы (1) достаточно ограничиться рассмотрением комитетов, составленных из решений её максимальных совместных подсистем.

Структуру множества м.с.п. и условия существования комитетных решений несовместной системы ограничений удобно формулировать в терминах графов (гиперграфов) её максимальных совместных или минимальных несовместных подсистем. Понятие графа м.с.п. впервые было введено в работе [2] для системы строгих однородных линейных неравенств. Свойства этого графа подробно изучены в работах [3, 4], в работе [3] некоторые из них были обобщены на случай более общей системы включений.

Определение 3. *Графом максимальных совместных подсистем системы ограничений* (1) *называется конечный граф* $G = (V, E)$, *множество вершин которого совпадает с множеством* $J_1, \dots, J_T$ *индексов максимальных совместных подсистем системы, и* $\{J_i, J_j\} \in E$ *тогда и только тогда, когда* $J_i \cup J_j = \mathbb{N}_m$.

Пусть далее $G = (V, E)$ — произвольный граф. Степенью его вершины v называется число рёбер, инцидентных v, т. е. число $\big|\{e \in E : v \in e\}\big|$. Чередующаяся последовательность:

$$v_1, \{v_1, v_2\}, v_2, \{v_2, v_3\}, \dots, \{v_{l-1}, v_l\}, v_l, \tag{3}$$

в которой $v_j \in V$, $\{v_j, v_{j+1}\} \in E$, называется (v_1, v_l)-маршрутом. Часто маршрут задается последовательностью входящих в него вершин. Маршрут называется цепью, если все его ребра различны, и простой цепью, если все его вершины, кроме, может быть, крайних, различны. Маршрут называется циклическим, если $v_1 = v_l$. Циклическая цепь называется циклом, а простая — простым циклом. Число рёбер маршрута называется его длиной. Граф G называется связным, если для любых вершин $v_i \neq v_j$ в нём существует (v_i, v_j)-маршрут.

Связность графа м.с.п. несовместной системы включений тесно связана с наличием у неё комитетных решений. Справедливо, например, следующее непосредственно проверяемое

Утверждение 1. *Пусть* $J_1, \dots, J_{2s-1}, J_1$ — *цикл в графе м.с.п. системы* (1) *и* $x^i \in D(J_i)$. *Тогда последовательность* $(x^1, \dots, x^{2s-1})$ — *комитетное решение системы* (1).

Пусть далее X — топологическое пространство, в котором заданы упорядоченные пары множеств $(A_1, A'_1), \ldots, (A_m, A'_m)$. Определим множества $D_1, \ldots, D_m \subset X \times \{0, 1\}$ следующим образом:

$$D_j = \left\{ \begin{bmatrix} x \\ 1 \end{bmatrix} \,\middle|\, x \in A_j \right\} \cup \left\{ \begin{bmatrix} x \\ 0 \end{bmatrix} \,\middle|\, x \in A'_j \right\},$$

и рассмотрим систему включений:

$$y = \begin{bmatrix} x \\ x' \end{bmatrix} \in D_j, \quad j \in \mathbb{N}_m. \tag{4}$$

Видно, что произвольная подсистема с непустым индексом $L \subseteq \mathbb{N}_m$ системы (4) совместна (т. е. $D(L) = \bigcap\limits_{j \in L} D_j \neq \varnothing$) тогда и только тогда, когда $\left(\bigcap\limits_{j \in L} A_j \right) \cup \left(\bigcap\limits_{j \in L} A'_j \right) \neq \varnothing$. Справедлива [4]

Теорема 2. *Пусть множества A_j, A'_j открыты в X, $A_j \cap A'_j = \varnothing$ и $F_j = X \setminus (A_j \cup A'_j)$ нигде не плотно в X для всех $j \in \mathbb{N}_m$. Если множество*

$$X \setminus \left(\bigcup_{i \neq j} F_i \cap F_j \right)$$

связно, то граф м.с.п. системы (4) *также связен.*

Следствием теоремы 2 является теорема о связности графа м.с.п. системы линейных однородных неравенств:

$$(a_j, x) > 0, \quad j \in \mathbb{N}_m, \tag{5}$$

в которой $a_j, x \in \mathbb{R}^n$, $\|a_j\| = 1$ и $a_j \pm a_i \neq 0$ для любых $i, j \in \mathbb{N}_m$.

Убедимся в этом, проведя рассуждения согласно [3]. Сопоставим системе (5) подходящую систему (4), для чего положим

$$A_j = \{x \,|\, (a_j, x) > 0\}, \; A'_j = \{x \,|\, (a_j, x) < 0\}.$$

Нетрудно убедиться, что произвольное непустое подмножество $L \subseteq \mathbb{N}_m$ является индексом совместной подсистемы системы (5) тогда и только тогда, когда $D(L) \neq \varnothing$, поэтому множества (и графы) м.с.п. систем (4) и (5) совпадают. Поскольку F_j — гиперплоскости в $\mathbb{R}^n$, то F_j нигде не плотны в X, и определяемое в теореме множество F таково, что $X \setminus F$ связно. Следовательно, по теореме 2, граф м.с.п. построенной системы (4) связен, значит, связен и граф м.с.п. системы (5). Приведем еще несколько результатов, полученных в [3].

Теорема 3. *Пусть для $k \in \mathbb{N}_{n-1}$ каждая подсистема из $(k+1)$ неравенства системы* (5) *совместна. Тогда степень каждой вершины её графа м.с.п. не меньше $k+1$.*

Теорема 4. *Граф изоморфен графу м.с.п. подходящей системы* (5) *на плоскости тогда и только тогда, когда он является циклом нечётной длины q, где $1 \leqslant q \leqslant m$.*

В $\mathbb{R}^n$ близкий результат формулируется так:

Теорема 5. *Всякое ребро графа м.с.п. системы* (5) *принадлежит простому циклу длины, не большей m.*

Теорема 6. *Граф м.с.п. системы* (5) *содержит простой цикл нечетной длины, не превосходящей m.*

Теоремы 2–3 позволяют находить м.с.п. системы (5), строя маршруты в графе её м.с.п. Например, известно [2], что если J_1 — индекс м.с.п. несовместной системы (5), то найдется м.с.п. той же системы с индексом J_2 таким, что $J_2 \supset (\mathbb{N}_m \setminus J_1)$.

Теорема 6 позволяет с другой стороны взглянуть на вопрос существования комитета для системы линейных однородных наравенств. Согласно утверждению 1, если индексы $J_1, J_2, \ldots, J_{2k-1}$ образуют цикл в графе м.с.п. произвольной системы включений (1) (в частности, системы (5)), то указанная система разрешима комитетом, состоящим из решений соответствующих м.с.п., взятых для каждой по одному. Согласно утверждению теоремы, в случае системы однородных линейных неравенств условия существования комитетного решения и цикла нечётной длины (в графе её м.с.п.) эквивалентны.

Равномерно распределённые системы неравенств

Остановимся на рассмотрении одного частного случая системы линейных однородных неравенств

$$(a_j, x) > 0, \quad j \in \mathbb{N}_m, \tag{6}$$

интересного, с одной стороны тем, что такие системы в некотором смысле «наиболее противоречивы», а, с другой стороны, ввиду регулярности строения их графов м.с.п., задачи исследования таких систем, например, поиска комитетных решений с наименьшим числом элементов, имеют красивые и простые алгоритмы решения. Эти системы неравенств принято называть *равномерно распределенными по Гейлу.*

Сопоставим произвольному вектору $x \in \mathbb{R}^n$ множества

$$J_>(x) = \{j \in \mathbb{N}_m \;:\; (a_j, x) > 0\};$$
$$J_<(x) = \{j \in \mathbb{N}_m \;:\; (a_j, x) < 0\};$$
$$J_=(x) = \{j \in \mathbb{N}_m \;:\; (a_j, x) = 0\}.$$

Определение 4 ([5]). *Система линейных неравенств* (6) *при $m = 2k + n - 1$ равномерно распределена, если для каждого $0 \neq x \in \mathbb{R}^n$ выполнено условие $|J_>(x)| \geqslant k$.*

Теорема 7 ([5]). *Для произвольных натуральных $n > 1$ и k существует равномерно распределённая система* (6) *с $m = 2k + n - 1$.*

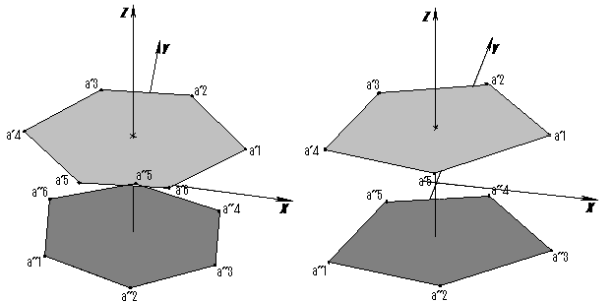


Рис. 1. Примеры равномерно распределённых систем неравенств

Приведем бесконечную серию примеров равномерно распределённых систем неравенств в $\mathbb{R}^3$. По определению, каждая такая система состоит из $2k+2$ неравенств (для некоторого натурального k). Исключив из рассмотрения тривиальный случай $k=1$, остановимся на системах неравенств

$$\begin{cases}(a'_i,x)>0, & |a'_i|=1, \quad i\in\mathbb{N}_{k+1};\\ (a''_j,x)>0, & |a''_j|=1, \quad j\in\mathbb{N}_{k+1},\end{cases}$$

векторы $a'_1,\dots,a'_{k+1},a''_1,\dots,a''_{k+1}$ левых частей которых имеют координаты

$$\begin{aligned} a'_j &= \frac{\sqrt{2}}{2}\begin{bmatrix}\cos\left(\frac{2\pi}{k+1}(j-1)\right)\\ \sin\left(\frac{2\pi}{k+1}(j-1)\right)\\ 1\end{bmatrix}, \quad j\in\mathbb{N}_{k+1};\\ a''_j &= \frac{\sqrt{2}}{2}\begin{bmatrix}\cos\left(\frac{\pi}{k+1}(2j+k)\right)\\ \sin\left(\frac{\pi}{k+1}(2j+k)\right)\\ -1\end{bmatrix}, \quad j\in\mathbb{N}_{k+1}.\end{aligned} \tag{7}$$

Таким образом, концы векторов a'_j и a''_j распределены на единичной сфере, являясь вершинами правильных m-угольников, вписанных в окружности

$$\left\{x\in\mathbb{R}^3\colon x_1^2+x_2^2=2,\ x_3=1\right\}$$

и

$$\left\{x\in\mathbb{R}^3\colon x_1^2+x_2^2=2,\ x_3=-1\right\},$$

и развёрнутых друг относительно друга на угол $\pi(m+1)/m$. На рис. 1 изображены множества векторов (7) при $k=5$ и $k=4$, соответственно.

Справедлив следующий критерий равномерной распределённости конкретной системы неравенств

Теорема 8 ([6]). *Система неравенств* (6) *равномерно распределена по Гейлу тогда и только тогда, когда одновременно выполнены условия:*

1) *каждые n векторов из $a_1,\dots,a_m$ линейно независимы;*
2) *если L — индекс м.с.п. системы* (6)*, то $|L| = k+n-1$.*

Теорема 8 имеет важные следствия. Равномерно распределённые системы в n–мерном пространстве наиболее близки по своим свойствам к хорошо изученным системам однородных неравенств, заданным на плоскости.

Заключение

Одной из причин циклов в экономике, биологии и некоторых других областях, является нехватка ресурсов. В этом случае возможны обходы равновесия, основанные на циклах во множестве максимальных по включению совместных подсистем соответствующих систем ограничений и комитетных конструкциях. В данной статье предлагается и обосновывается подобный подход.

Литература

[1] *Мазуров Вл. Д.* Неоднозначная интерпретация противоречивых данных // Труды ИММ УрО РАН. — 1984.

[2] *Тягунов Л. И.* О выделении последовательности максимальных совместных подсистем несовместной системы линейных неравенств // Математические методы планирования и управления в больших системах. — Свердловск: УНЦ АН СССР, 1973. — С. 152–162. — Деп. ВИНИТИ, № 7467–73.

[3] *Гайнанов Д. Н.* О графах максимальных совместных подсистем несовместных систем линейных неравенств. — Москва, 1981. — 46 с. — Деп. ВИНИТИ, № 229–81.

[4] *Гайнанов Д. Н., Новокшенов В. А., Тягунов Л. И.* О графах, порождаемых несовместными системами линейных неравенств // Мат. заметки. — 1983. — Т. 33, вып. 2. — С. 293–300.

[5] *Gale D.* Neighboring vertices on a convex polyhedron // Linear inequalities and related systems, edited by H.W.Kuhn and A.W.Tucker. — Princeton, 1956. — Pp. 255–263.

[6] *Khachay M. Yu.* On Approximate Algorithm of a Minimal Committee of a Linear Inequalities System // Pattern Recognition and Image Analysis. — 2003. — Vol. 13, № 3. — Pp. 459–464.

Об одном конструктивном подходе к построению обобщенных алгебраических ΣΠ-нейронов в одном абстрактном классе алгебр*

Шибзухов З. М.
szport@gmail.com
НИИ прикладной математики и автоматизации КБНЦ РАН, г.Нальчик

Рассматривается один класс обобщенных алгебраических ΣΠ-нейронов, построенных над алгебрами из некоторых достаточно широких абстрактных классов. Приводится математическое обоснование прямой конструктивной алгебраической процедуры для обучения корректных алгебраических ΣΠ-нейронов. Доказывается корректность рассмотренных классов алгебраических ΣΠ-нейронов относительно определенных классов обучающих последовательностей примеров.

В работе рассматривается задача обучения с учителем одного абстрактного класса алгебраических ΣΠ-нейронов. Впервые искусственный нейрон такого типа введен в [1] и назван ΣΠ-элементом (sigma-pi unit). Название ΣΠ-нейрона (sigma-pi neuron), по-видимому, впервые было введено в [4]. Показано, что модель ΣΠ-нейрона адекватно отражает некоторые процессы обработки информации в головном мозге [2]. Нейронные сети, состоящие из ΣΠ-нейронов традиционно называют ΣΠ-нейронными сетями (sigma-pi neural networks) [3] или кратко ΣΠ-сетями (sigma-pi networks). Краткое введение в ΣΠ-нейронные сети можно найти в [5]. Алгебраические ΣΠ-нейроны были предложены в [8].

В теоретической нейроинформатике известно, что любое непрерывное преобразование, определенное на компактном множестве, можно аппроксимировать при помощи искусственной нейронной сети, построенной на базе элементов, реализующих взвешенное суммирование и непрерывную скалярную функцию [6]. Из него следует, что любое непрерывное преобразование, определенное на компактном множестве, аппроксимируется при помощи сети из элементов, реализующих операции суммирования, умножения и элементов, реализующих произвольно выбранные нелинейные непрерывные скалярные функции. Алгебраические ΣΠ-нейроны представляют простейшие сети такого типа.

Большой интерес представляют *конструктивные алгебраические процедуры* построения *корректных распознающих алгоритмов* [7] вообще, и искусственных нейронных сетей в частности, которые позволяют одновременно формировать структуру алгоритма и настраивать его параметры, не прибегая при этом к решению сложных оптимизационных задач для достижения корректности его функционирования.

В [8, 9, 10] описаны некоторые *корректные классы* ΣΠ-нейронов, математически обоснован конструктивный метод прямого построения множеств корректных алгебраических ΣΠ-нейронов, имеющих в определенном смысле минимальную сложность, по заданной стандартной обучающей информации. Метод построения корректных алгебраических ΣΠ-нейронов был разработан в предположении, что входная информация кодируется в области целостности – коммутативном, ассоциативном кольце с единицей, без делителей нуля.

В настоящей работе рассматривается абстрактный класс алгебраических ΣΠ-нейронов для обработки информации, кодируемой в алгебраических структурах более общего вида, чем области целостности, в которых определены операции *псевдо-сложения* и *псевдо-умножения* (термины псевдо-сложения и псевдо-умножения введены в [11], но использованы для обозначения более узкого типа операций). Определяемые в работе функции псевдо-суммирования и псевдо-умножения обобщают традиционные операции суммирования и умножения. На их основе и определяется новая модель алгебраического ΣΠ-нейрона.

В настоящей работе описаны некоторые корректные классы алгебраических ΣΠ-нейронов, математически обосновывается прямой конструктивный алгебраический метод обучения с учителем корректных алгебраических ΣΠ-нейронов по стандартной обучающей информации.

Алгебры $\mathfrak{A}\langle +, \cdot \rangle$

Рассмотрим класс алгебр $\mathfrak{A} = \mathfrak{A}\langle +, \cdot \rangle$ на множестве $\boldsymbol{X} \supseteq \{0, 1\}$. Операция «+» удовлетворяет условию

$$0 + x = x + 0 = x,$$

операция «·» удовлетворяет условию

$$x_1 \cdot x_2 = 0 \;\Leftrightarrow\; x_1 = 0 \vee x_2 = 0.$$

Свойства коммутативности и ассоциативности для операций «+» и «·», вообще говоря, не требуются.

Пример 1. Пусть $\boldsymbol{X} = (-a, a) \subset \mathbb{R}$, и задано взаимно однозначное отображение $\mu\colon (-a, a) \to \mathbb{R}$ такое, что $\mu(0) = 0$. Операции «$\overset{\mu}{+}$» и «$\overset{\mu}{\cdot}$» определя-

*Работа выполнена при поддержке ОМН РАН по программе «Алгебраические и комбинаторные методы математической кибернетики и информационные системы нового поколения» и гранта РФФИ №09-01-00166-а.

ются следующим образом:

$$x \overset{\mu}{+} y = \mu^{-1}\big(\mu(x) + \mu(y)\big);$$
$$x \overset{\mu}{\cdot} y = \mu^{-1}\big(\mu(x) \cdot \mu(y)\big).$$

Эти операции задают на $(-a, a)$ структуру поля.

Пример некоммутативных арифметических операций на множестве бинарных деревьев можно найти в [12, 13].

Алгебраические псевдосуммирующие функции

Пусть $\{+\} \subset \boldsymbol{S}$ — множество функций $\Sigma\colon \boldsymbol{X}^2 \to \boldsymbol{X}$, таких, что

$$\Sigma(s, 0) = \Sigma(0, s) = s$$

и $\boldsymbol{H}$ — множество функций $\eta\colon \boldsymbol{X} \to \boldsymbol{X}$, таких что

$$\eta(s) = 0 \Leftrightarrow s = 0.$$

Пример 2. Пусть $\boldsymbol{X} = \mathbb{R}$. Возьмём множество операций сложения по степенному закону $\boldsymbol{S} = \{\Sigma_\alpha(x, y)\colon \alpha \in \mathbb{R}\}$,

$$\Sigma_\alpha(x, y) = x^{(\alpha)} + y^{(\alpha)},$$

где $x^{(\alpha)} = \operatorname{sign} x \cdot |x|^\alpha$. В частности, $\Sigma_1(x, y) = x + y$. Возьмём $\boldsymbol{H} = \{\sigma_w(x)\colon w \in \mathbb{R}\}$ — множество «сигмоидальных» функций вида

$$\sigma_w(s) = \frac{e^{wx} - e^{-wx}}{e^{wx} + e^{-wx}}.$$

Определим $\mathfrak{S}$ — класс всех алгебраических псевдосуммирующих функций, которые строятся по следующим правилам:

1) $\forall c \in \boldsymbol{X}\colon c \in \mathfrak{S}$;
2) $\forall \Sigma \in \boldsymbol{S}\ \forall \{\mathrm{sf}_1, \mathrm{sf}_2\} \subset \mathfrak{S}\colon \Sigma(\mathrm{sf}_1, \mathrm{sf}_2) \in \mathfrak{S}$;
3) $\forall \eta \in \boldsymbol{H}\ \forall \mathrm{sf} \in \mathfrak{S} : \eta(\mathrm{sf}) \in \mathfrak{S}$.

Пусть $\mathfrak{F}$ — некоторый класс функций. Определим $\mathfrak{S}[\mathfrak{F}]$ — класс всех *алгебраических* $\Sigma\Phi$*-функций*, который состоит из композиций

$$\Sigma\Phi = \mathrm{sf}(\Phi_1, \ldots, \Phi_m),$$

где $\mathrm{sf} \in \mathfrak{S}$, $\{\Phi_1, \ldots, \Phi_m\} \subset \mathfrak{F}$. То есть каждая $\Sigma\Phi$-функция представляет собой «псевдосумму» некоторого набора базисных функций.

Введем теперь решающие функции, которые преобразуют оценки, вычисляемые при помощи $\Sigma\Phi$-функций, в значения из некоторого целевого множества $\boldsymbol{Y}$.

Пусть $\mathfrak{R}$ — некоторый класс *решающих функций* вида $\rho\colon \boldsymbol{X} \to \boldsymbol{Y}$.

Определение 1. *Функция $\rho(s)$ — допустимая относительно $\mathfrak{S}$, если для любых $a \in \boldsymbol{X}$, $b \in \boldsymbol{X}\colon b \neq 0$ и $y \in \boldsymbol{Y}$ найдется функция $\mathrm{sf}(s_1, s_2) \in \mathfrak{S}$ такая, что $\rho\big(\mathrm{sf}(a, b)\big) = y$.*

Определение 2. *Класс $\mathfrak{R}$ — допустимый, если любая функция $\rho \in \mathfrak{R}$ — допустимая,*

Определим $\mathfrak{R} \circ \mathfrak{S}[\mathfrak{F}]$ — класс *алгебраических* $\Sigma\Phi$*-элементов*, который состоит из композиций вида $\rho\big(\Sigma\Phi\big)$, где $\Sigma\Phi \in \mathfrak{S}[\mathfrak{F}]$, ρ — допустимая решающая функция.

Треугольно упорядоченные последовательности функций

Пусть $\mathfrak{F}^n$ — некоторый класс функций $\boldsymbol{X}^r \to \boldsymbol{X}$, где $1 \leqslant r \leqslant n$. Рассмотрим последовательности функций $\{\Phi_k(\boldsymbol{x} \,|\, \boldsymbol{i}_k)\} \subset \mathfrak{F}^n$, где $\boldsymbol{x} = (x_1, \ldots, x_n)$; выражение $\boldsymbol{x} \,|\, \boldsymbol{i}$ обозначает набор аргументов $(x_{i_1}, \ldots, x_{i_r})$, где $\boldsymbol{i} = (i_1, \ldots, i_r)$; $\{\boldsymbol{i}_k\}$ — последовательность мультииндексов, $\boldsymbol{i}_k \subseteq \{1, \ldots, n\}$.

Пусть задана последовательность $\{\boldsymbol{x}_k\} \subset \boldsymbol{X}^n$.

Определение 3. *Последовательность функций $\{\Phi_k\}$ — треугольно упорядоченная на $\{\boldsymbol{x}_k\}$, если*

1) $\forall j < k\ \ \Phi_k(\boldsymbol{x}_j \,|\, \boldsymbol{i}_k) = 0$;
2) $\forall k\ \ \Phi_k(\boldsymbol{x}_k \,|\, \boldsymbol{i}_k) \neq 0$.

Другими словами, $\{\Phi_k\}$ — треугольно упорядоченная на $\{\boldsymbol{x}_k\}$, если матрица $b_{kj} = \big\{\Phi_k(\boldsymbol{x}_j)\big\}$ – треугольная с ненулевыми элементами на главной диагонали.

Пример 3. Пусть $\{\boldsymbol{x}_k\} \subseteq \{0, 1\}^n$ упорядочена по неубыванию величины скалярного произведения $c_1x_1 + \cdots + c_nx_n$, где $c_1 > 0, \ldots, c_n > 0$. Тогда последовательность произведений $\{\mathrm{p}_k(\boldsymbol{x} \,|\, \boldsymbol{i}_k)\}$, где

$$\mathrm{p}_k(\boldsymbol{x} \,|\, \boldsymbol{i}_k) = \prod_{i \in \boldsymbol{i}_k} x_i,$$

$\boldsymbol{i}_k = \{i : x_{ki} \neq 0\}$, является треугольно упорядоченной на $\{\boldsymbol{x}_k\}$.

Пример 4. Пусть $\boldsymbol{Q} = \{0, \ldots, Q-1\}$, $\{\boldsymbol{x}_k\} \subseteq \boldsymbol{Q}^n$ упорядочена по неубыванию величины скалярного произведения $c_1x_1 + \cdots + c_nx_n$, где $c_1 > 0, \ldots, c_n > 0$. Тогда последовательность произведений $\{\varphi_k(\boldsymbol{x})\}$, где

$$\mathrm{pf}_k(\boldsymbol{x}) = \prod_{i=1}^{n} \varphi(x_i, x_{ki}),$$

$\varphi(x, a) = 0 \Leftrightarrow x < a$, является треугольно упорядоченной на $\{\boldsymbol{x}_k\}$.

Ниже будет показано, что использование треугольно упорядоченных последовательностей функций в качестве базисных функций упрощает процедуру обучения, так что настройка весов осуществляется за один проход обучающей последовательности примеров.

Обозначим через $\mathfrak{F}\{\boldsymbol{x}_k\}$ класс всех последовательностей $\{\Phi_k\} \subset \mathfrak{F}^n$, треугольно упорядоченных на $\{\boldsymbol{x}_k\}$.

Пусть $\mathfrak{X}^n$ — некоторый класс конечных множеств $\{\boldsymbol{x}\} \subseteq \boldsymbol{X}^n$.

Определение 4. *$\mathfrak{F}^n$ — допустимый для $\mathfrak{X}^n$, если для любого конечного множества $\{\boldsymbol{x}\} \in \mathfrak{X}^n$ можно, пронумеровав его элементы, построить последовательность $\{\boldsymbol{x}_k\}$ и построить последовательность функций $\{\Phi_k\} \in \mathfrak{F}\{\boldsymbol{x}_k\}$.*

Функции, входящие в базисную треугольно упорядоченную последовательность часто содержат «несущественные» переменные, которые можно исключить без вреда: получающиеся в результате последовательности, содержа функции, которые зависят от меньшего числа переменных, в то же время являются треугольно упорядоченными. Выбор «несущественной» переменной для исключения, как правило, неоднозначен. Поэтому в результате можно построить множество треугольно-упорядоченных последовательностей, которые также можно использовать в качестве базисных. Изложим теперь это более формально.

Рассмотрим последовательность вложений $\mathfrak{F}^1 \subset \ldots \subset \mathfrak{F}^m \subset \ldots \subset \mathfrak{F}^n$, где $\mathfrak{F}^m$ — подкласс функций вида $\boldsymbol{X}^r \to \boldsymbol{X}$, $1 \leqslant r \leqslant m$.

Пусть для каждой функции $\Phi(\boldsymbol{x} \,|\, \boldsymbol{i}) \in \mathfrak{F}^m$ и каждого $i \in \boldsymbol{i}$ определена операция исключения i-го аргумента: $\partial^i \colon \mathfrak{F}^m \to \mathfrak{F}^{m-1}$.

Предположим, что она удовлетворяет следующему условию:

$$\Phi(\boldsymbol{x} \,|\, \boldsymbol{i}) \neq 0 \Rightarrow (\partial^i \Phi)(\boldsymbol{x} \,|\, \boldsymbol{i}') \neq 0,$$

где $\boldsymbol{i}' = \boldsymbol{i} \setminus \{i\}$.

Пусть $\{\Phi_k(\boldsymbol{x} \,|\, \boldsymbol{i}_k)\} \in \mathfrak{F}\{\boldsymbol{x}_k\}$. Рассмотрим преобразования последовательности $\{\Phi_k(\boldsymbol{x} \,|\, \boldsymbol{i}_k)\} \to \{\Phi'_k(\boldsymbol{x} \,|\, \boldsymbol{i}'_k)\}$, где $\Phi'_k(\boldsymbol{x} \,|\, \boldsymbol{i}'_k)$ получается из $\Phi_k(\boldsymbol{x} \,|\, \boldsymbol{i}_k)$ в результате последовательного исключения некоторых аргументов, а $\boldsymbol{i}'_k$ получается из $\boldsymbol{i}_k$ исключением индексов, исключаемых из $\Phi_k(\boldsymbol{x} \,|\, \boldsymbol{i}_k)$ аргументов. Данное преобразование будем называть *допустимым*, если $\{\Phi'_k(\boldsymbol{x} \,|\, \boldsymbol{i}'_k)\} \in \mathfrak{F}\{\boldsymbol{x}_k\}$. Тогда при помощи допустимых преобразований можно построить последовательности $\{\Phi_k(\boldsymbol{x} \,|\, \boldsymbol{i}_k)\}$, которые содержат только существенные для треугольной упорядоченности наборы аргументов. Такие последовательности функций условимся называть *существенными*. Таким образом, если задана некоторая последовательность $\{\Phi_k(\boldsymbol{x} \,|\, \boldsymbol{i}_k)\} \in \mathfrak{F}\{\boldsymbol{x}_k\}$, то применяя допустимые преобразования можно построить различные наборы существенных последовательностей функций, треугольно упорядоченных на $\{\boldsymbol{x}_k\}$.

Рекуррентный метод построения корректного ΣΦ-нейрона

Построение ΣΦ-элемента осуществляется по конечной обучающей последовательности $\{\boldsymbol{x}_k\} \in \mathfrak{X}^n$ и соответствующей ей последовательности $\{\mathrm{y}_k\} \subseteq \boldsymbol{Y}$, где $k = 1, \ldots, N$.

Определение 5. *ΣΦ-элемент* sfe *— корректный на $\{\boldsymbol{x}_k\}$ и $\{\mathrm{y}_k\}$, если $\forall k$:* $\mathrm{sfe}(\boldsymbol{x}_k) = \mathrm{y}_k$.

Построение начинается с некоторого начального ΣΦ-элемента $\mathrm{sfe}_0 = \rho(\Sigma\Phi_0)$, где $\Sigma\Phi_0$ — произвольная ΣΦ-функция (например, константа). В результате необходимо построить ΣΦ-элемент sfe, *корректный на $\{\boldsymbol{x}_k\}$ и $\{\mathrm{y}_k\}$*.

Перед началом или в процессе обучения строится последовательность функций $\{\Phi_k(\boldsymbol{x})\} \in \mathfrak{F}\{\boldsymbol{x}_k\}$ и последовательность ΣΦ-элементов $\{\mathrm{sfe}_k(\boldsymbol{x})\}$, где $\mathrm{sfe}_k = \rho\big(\Sigma\Phi_k\big)$, а

$$\Sigma\Phi_k = \begin{cases} \Sigma\Phi_{k-1}, & \text{если } \mathrm{sfe}_k(\boldsymbol{x}_k) = \mathrm{y}_k, \\ \mathrm{sf}_k\big(\Sigma\Phi_{k-1}, \Phi_k\big), & \text{если } \mathrm{sfe}_k(\boldsymbol{x}_k) \neq \mathrm{y}_k. \end{cases}$$

Функция $\mathrm{sf}_k(s_1, s_2) \in \mathfrak{S}$. Она подбирается так, чтобы выполнялось равенство:

$$\mathrm{y}_k = \rho\Big(\mathrm{sf}_k\big(\Sigma\Phi_{k-1}(\boldsymbol{x}_k),\, \Phi_k(\boldsymbol{x}_k)\big)\Big).$$

По определению, если ρ — допустимая, то $\mathrm{sf}_k(s_1, s_2)$ существует.

Применяя допустимые преобразования, можно построить множества существенных последовательностей функций $\{\Phi_k(\boldsymbol{x} \,|\, \boldsymbol{i}_k)\} \in \mathfrak{F}\{\boldsymbol{x}_k\}$. Таким образом можно построить множество ΣΦ-элементов $\{\mathrm{sfe}(\boldsymbol{x})\}$, корректных на $\{\boldsymbol{x}_k\}$ и $\{\mathrm{y}_k\}$. Это позволяет в ряде случаев применить в дальнейшем процедуры взвешенного голосования или линейные (выпуклые) свертки для повышения надежности функционирования группы корректных ΣΦ-нейронов.

Верна следующая

Лемма 1. $\forall k\ \forall j \leqslant k \colon \mathrm{sfe}_k(\boldsymbol{x}_j) = \mathrm{y}_j$.

Введем понятие корректного класса ΣΦ-элементов.

Определение 6. *Класс $\mathfrak{R} \circ \mathfrak{S}[\mathfrak{F}^n]$ — корректный на $\mathfrak{X}^n$, если для любой пары последовательностей $\{\boldsymbol{x}_k\} \in \mathfrak{X}^n$ и $\{\mathrm{y}_k\} \subseteq \boldsymbol{Y}$ существует* ΣΦ*-элемент* $\mathrm{sfe} = \rho\big(\Sigma\Phi\big)$ *такой, что* $\forall k \colon \mathrm{y}_k = \mathrm{sfe}(\boldsymbol{x}_k)$.

Из определения и леммы вытекает

Теорема 2. *Если $\mathfrak{R}$ — допустимый и $\mathfrak{F}^n$ — допустимый для $\mathfrak{X}^n$, то $\mathfrak{R} \circ \mathfrak{S}[\mathfrak{F}^n]$ — корректный на $\mathfrak{X}^n$.*

Рассмотрим некоторые классы *алгебраических псевдомультиплицирующих* функций, из которых при некоторых достаточно общих предположениях можно эффективно строить треугольно упорядоченные последовательности функций.

Алгебраические псевдомультиплицирующие функции

Пусть $\{\cdot\} \subset \boldsymbol{P}$ — множество функций $\Pi \colon \boldsymbol{X}^2 \to \boldsymbol{X}$, таких, что

$$\Pi(x_1, x_2) = 0 \Leftrightarrow x_1 = 0 \vee x_2 = 0,$$

и пусть $\boldsymbol{F}$ — множество функций $\chi\colon \boldsymbol{X} \to \boldsymbol{X}$ таких, что $\chi(0) = 0$.

Пример 5. При $\boldsymbol{X} = \mathbb{R}$ можно положить $\boldsymbol{P} = \{x_1^{(\alpha)} x_2^{(\beta)}\colon \alpha, \beta \in \mathbb{R}\}$ (при $\alpha = \beta = 1$ имеем обычное умножение), $\boldsymbol{F} = \{\operatorname{sign} x \cdot \ln(1 + |x|)\}$.

Определим $\mathfrak{P}$ — класс всех алгебраических псевдомультиплицирующих функций, которые строятся по следующим правилам:

1) $\forall c \in \boldsymbol{X}\colon c \in \mathfrak{P}$;
2) $\forall \Pi \in \boldsymbol{P}\, \forall \{\mathrm{pf}_1, \mathrm{pf}_2\} \subset \mathfrak{P}\colon \Pi(\mathrm{pf}_1, \mathrm{pf}_2) \in \mathfrak{P}$;
3) $\forall \chi \in \boldsymbol{F}\, \forall \mathrm{pf} \in \mathfrak{P}\colon \chi(\mathrm{pf}) \in \mathfrak{P}$.

Пусть $\mathfrak{F}$ — некоторый класс функций. Определим $\mathfrak{P}[\mathfrak{F}]$ — класс *алгебраических* ПФ*-функций*, который состоит из функций

$$\Pi\Phi = \Pi(\Phi_1, \ldots, \Phi_m),$$

где $\Pi \in \mathfrak{P}$, $\{\Phi_1, \ldots, \Phi_m\} \subset \mathfrak{F}$.

Алгебраический ΣΠ*-нейрон* реализует композицию $\mathrm{spn} = \rho(\mathrm{spf})$, где $\mathrm{spf} \in \mathfrak{S}\big[\mathfrak{P}\big]$. *Алгебраический* ΣΠΦ*-нейрон* реализует композицию $\mathrm{spn} = \rho(\mathrm{spf})$, где $\mathrm{spf} \in \mathfrak{S}\big[\mathfrak{P}[\mathfrak{F}]\big]$.

Во многих случаях обучающую последовательность входных векторов состоит из *разреженных* векторов. В этом случае нетрудно построить треугольно упорядоченную последовательность псевдомультиплицирующих функций.

Упорядоченные последовательности разреженных векторов

Пусть задана последовательность $\{\boldsymbol{x}_k\} \subset \boldsymbol{X}^n$ и последовательность мультииндексов $\{\boldsymbol{i}_k\}$, где $\boldsymbol{i}_k \subseteq \{1, \ldots, n\}$.

Определение 7. $\{\boldsymbol{x}_k\}$ *упорядочена по нулям относительно* $\{\boldsymbol{i}_k\}$, *если*

1) $\forall i \in \boldsymbol{i}_k\colon x_{ki} \neq 0$;
2) $\forall j < k\, \exists i\colon x_{ki} \neq 0 \wedge x_{ji} = 0$.

Пример 6. Если последовательность $\{\boldsymbol{x}_k\} \subseteq \{0, 1\}^n$ не содержит одинаковых элементов и упорядочена в порядке неубывания $x_1 + \cdots + x_n$, то она является упорядоченной по нулям относительно $\{\boldsymbol{i}_k^0\}$, где $\boldsymbol{i}_k^0 = \{i : k_{ki} \neq 0\}$.

Пусть задана последовательность функций $\{\mathrm{pf}_k(\boldsymbol{x} \,|\, \boldsymbol{i}_k)\}$, где $\mathrm{pf}_k \in \mathfrak{P}$. Если $\{\boldsymbol{x}_k\}$ упорядочена по нулям относительно $\{\boldsymbol{i}_k\}$, то $\{\mathrm{pf}_k(\boldsymbol{x} \,|\, \boldsymbol{i}_k)\} \in \mathfrak{P}\{\boldsymbol{x}_k\}$.

Не всякое множество векторов можно упорядочить по нулям. Однако, во многих случаях можно построить преобразование, которое исходное множество переводит в множество разреженных векторов, которое можно упорядочить по нулям относительно некоторой последовательности мультииндексов.

Это преобразование можно построить, используя метод покрытий. Пусть задано некоторое конечное множество $\{\boldsymbol{u}_k\} \subset \boldsymbol{X}^n$ и некоторое конечное покрытие $\{\boldsymbol{u}_k\} \subseteq \bigcup\limits_{p=1}^{L} U_p$, *разделяющее* векторы из $\{\boldsymbol{u}_k\}$, т. е.

$$\forall j \neq k\ \exists p\colon \boldsymbol{u}_j \in U_p \wedge \boldsymbol{u}_k \notin U_p.$$

Пусть для каждого p определена функция $\varphi_p(\boldsymbol{u})$ такая, что $\varphi_p(\boldsymbol{u}) = 0 \Leftrightarrow \boldsymbol{u} \notin U_p$. Пусть $\boldsymbol{i}_k = \{i\colon \boldsymbol{u}_k \in U_i\}$. Определим преобразование $\boldsymbol{\varphi}(\boldsymbol{u}) = \big(\varphi_1(\boldsymbol{u}), \ldots, \varphi_L(\boldsymbol{u})\big)$. Тогда $\{\boldsymbol{x}_k\}$, где $\boldsymbol{x}_k = \boldsymbol{\varphi}(\boldsymbol{u}_k)$, можно упорядочить по нулям относительно $\{\boldsymbol{i}_k\}$.

Упорядоченные последовательности векторов и треугольно упорядоченные последовательности функций

Пусть на $\boldsymbol{X}$ задано отношение линейного порядка «$\geqslant$».

Определение 8. *Для любой пары векторов* $\boldsymbol{x}'$ *и* $\boldsymbol{x}''$ *из* $\boldsymbol{X}^n$ *и мультииндекса* $\boldsymbol{i} \subseteq \{1, \ldots, n\}$ *определим отношения «*$\not\geqslant_{\boldsymbol{i}}$*» и «*$\geqslant_{\boldsymbol{i}}$*»:*

1) $\boldsymbol{x}' \not\geqslant_{\boldsymbol{i}} \boldsymbol{x}''$, *если* $\exists i \in \boldsymbol{i}\colon\ x'_i \not\geqslant x''_i$;
2) $\boldsymbol{x}' \geqslant_{\boldsymbol{i}} \boldsymbol{x}''$, *если* $\forall i \in \boldsymbol{i}\colon\ x'_i \geqslant x''_i$.

Определение 9. *Последовательность векторов* $\{\boldsymbol{x}_k\}$ *называется* $\not\geqslant$*-упорядоченной относительно* $\{\boldsymbol{i}_k\}$, *если* $\forall j < k$ *выполнено* $\boldsymbol{x}_j \not\geqslant_{\boldsymbol{i}_k} \boldsymbol{x}_k$.

Если $\forall k\colon \boldsymbol{i}_k = \{1, \ldots, n\}$, то будем говорить, что последовательность $\{\boldsymbol{x}_k\}$ — $\not\geqslant$-упорядоченная.

Пример 7. Если $\boldsymbol{X} = \mathbb{R}$ и $\{\boldsymbol{x}_k\}$ упорядочена по возрастанию величины $\sum c_i x_i$, где $c_i > 0$, то $\{\boldsymbol{x}_k\}$ — $\not\geqslant$-упорядоченная.

Введем понятие $\not\geqslant$*-упорядоченной пары последовательностей* относительно последовательности мультииндексов $\{\boldsymbol{i}_k\}$.

Определение 10. $\{\boldsymbol{x}_k\}$ *и* $\{\boldsymbol{a}_k\}$ *образуют* $\not\geqslant$*-упорядоченную пару относительно* $\{\boldsymbol{i}_k\}$, *если*

1) $\forall j < k\colon \boldsymbol{x}_j \not\geqslant_{\boldsymbol{i}_k} \boldsymbol{a}_k$;
2) $\forall k\colon \boldsymbol{x}_k \geqslant_{\boldsymbol{i}_k} \boldsymbol{a}_k$.

Пусть $\mathfrak{F}_{\not\geqslant}$ — класс функций $\varphi\colon \boldsymbol{X}^2 \to \boldsymbol{X}$, таких что $\varphi(x, a) = 0 \Leftrightarrow x \not\geqslant a$.

Рассмотрим класс $\mathfrak{P}[\mathfrak{F}_{\not\geqslant}]$, состоящий из функций вида

$$\Phi(\boldsymbol{x} \,|\, \boldsymbol{i};\ \boldsymbol{a} \,|\, \boldsymbol{i}) = \Pi\big(\varphi(x_{i_1}, \mathrm{a}_{i_1}), \ldots, \varphi(x_{i_r}, \mathrm{a}_{i_r})\big),$$

где $\Pi(x_{i_1}, \ldots, x_{i_r}) \in \mathfrak{P}$, $\boldsymbol{i} = (i_1, \ldots, i_r)$.

Если $\{\boldsymbol{x}_k\}$ и $\{\boldsymbol{a}_k\}$ образуют $\not\geqslant$-упорядоченную пару относительно $\{\boldsymbol{i}_k\}$, то $\{\Phi(\boldsymbol{x} \,|\, \boldsymbol{i};\ \boldsymbol{a}_k \,|\, \boldsymbol{i}_k)\}$ треугольно упорядочена на $\{\boldsymbol{x}_k\}$. Заметим, что если $\{\boldsymbol{x}_k\}$ — $\not\geqslant$-упорядоченная относительно $\{\boldsymbol{i}_k\}$, то в качестве $\{\boldsymbol{a}_k\}$ можно взять $\{\boldsymbol{x}_k\}$. Таким образом, для построения треугольно упорядоченной

последовательности функций достаточно треугольно упорядочить $\{\boldsymbol{x}_k\}$.

Иллюстративный пример

Для иллюстрации приведем простой пример применения ΣΠ-нейронов в задаче классификации. Пусть задана обучающая последовательность логических описаний объектов $\{\boldsymbol{x}_k\}$, где $\boldsymbol{x}_k \in \{0,1\}^n$, и соответствующая ей последовательность значений признака принадлежности объекта некоторому классу $\{y_k\}$, где $y_k \in \{0,1\}$.

Предполагаем, что обучающая информация непротиворечивая: для любой пары $\boldsymbol{x_{k'}}$ и $\boldsymbol{x_{k''}}$ из $\boldsymbol{x_{k'}} \neq \boldsymbol{x_{k''}}$ вытекает, что $y_{k'} \neq y_{k''}$.

Перед обучением упорядочим последовательности $\{\boldsymbol{x}_k\}$ и $\{y_k\}$ по неубыванию $|\boldsymbol{x}| = x_1 + \cdots + x_n$.

В результате последовательность произведений $\mathrm{p}_k(\boldsymbol{x} \,|\, \boldsymbol{i}_k^0) = \prod\limits_{i \in \boldsymbol{i}_k^0} x_i$, где $\boldsymbol{i}_k^0 = \{i\colon x_{ki} \neq 0\}$, является треугольно упорядоченной на $\{\boldsymbol{x}_k\}$.

Используя процедуру исключения «несущественных» переменных из функций $\mathrm{p}_k(\boldsymbol{x} \,|\, \boldsymbol{i}_k^0)$, получим для каждого k набор функций $P_k = \{\mathrm{p}_k(\boldsymbol{x} \,|\, \boldsymbol{i})\}$, где $\boldsymbol{i} \subseteq \boldsymbol{i}_k^0$, каждая из которых содержит только «существенные» аргументы.

Некоторым образом строим набор из m последовательностей функций $\{\mathrm{p}_k(\boldsymbol{x} \,|\, \boldsymbol{i})\}$, таких, что $\mathrm{p}_k \in P_k$. Затем применяем процедуру построения, описанную выше. С ее помощью строим набор корректных ΣΠ-нейронов $\{\mathrm{spn}_j(\boldsymbol{x})\}$.

Искомая характеристическая функция класса представляется в виде:

$$y = H\big(-m/2 + \sum_j \mathrm{spn}_j(\boldsymbol{x})\big),$$

где

$$H(s) = \begin{cases} 1, & \text{если} \quad s \geqslant 0 \\ 0, & \text{иначе.} \end{cases}$$

Заключение

Итак, мы определили абстрактный класс алгебраических ΣΠ-нейронов, который является обобщением моделей ΣΠ-нейрона, изучавшихся ранее. Показано, что в этой модели при помощи несложной конструктивной процедуры можно строить множества алгебраических ΣΠ-нейронов, которые являются корректными по построению. Теоретические значение данной работы состоит в том, что найден конструктивный метод быстрой генерации множеств алгебраических ΣΠ-нейронов, корректно функционирующих на обучающем множестве примеров. Важное достоинство данного метода состоит в том, что с его помощью можно осуществить построение каждого ΣΠ-нейрона всего за один проход обучающей последовательности примеров. Это, в частности, позволяет использовать его в процедурах комбинаторного характера для поиска корректных ΣΠ-нейронов, оптимальных по отношению ко внешним критериям качества. Экспериментальное исследование возможностей предложенной модели, конструктивного метода обучения, применения его на практике – предмет следующих исследований.

Литература

[1] *Rumelhart D. E., Hinton D. E., McClelland, J. L.* A general framework for parallel distributed processing // Parallel distributed processing: explorations in the microstructure of cognition – Eds: Rumelhart D. E., McClelland J. L. — 1986. MIT Press, Vol. 1, Pp. 45–76.

[2] *Mel B. W.* Sigma-pi column: A model of associative learning in celebral neocortex. — California Institute of Technology. CNS memo № 6. Technical report, Pasadena, California 91125, 1990.

[3] *Giles C. L., Maxwell T.* Learning, invariance and generalization in high-order neural networks // Applied Optics. 1987. — Vol. 26, № 23. — Pp. 4972–4978.

[4] *Mel B. W.* The sigma-pi model neuron: roles of the dendritic tree in associative learning // Soc. NeuroScience Abstr. 1990. — Vol. 16. — Pp. 205.4.

[5] *Шибзухов З.М.* ΣΠ-нейронные сети: Введение // XI Всероссийская научно-техническая конференция «Нейроинформатика–2009»: Лекции по нейроинформатике. — 2009, М: МИФИ. — С. 66–88.

[6] *Горбань А. Н.* Нейроинформатика. / Под. ред. Е. А. Новикова. — Новосибирск: Наука, 1998. — С. 18–46.

[7] *Журавлев Ю. И.* Об алгебраическом подходе к решению задач распознавания и классификации // Избранные научные труды. — М.: Магистр, 1998. — С. 229–323.

[8] *Шибзухов З. М.* Рекуррентный метод конструктивного обучения алгебраических ΣΠ-нейронов и ΣΠ-нейромодулей // Доклады РАН. — 2003. — Т.388, № 2. — С. 174–176.

[9] *Шибзухов З. М.* Рекуррентный метод конструктивного обучения некоторых сетей алгебраических ΣΠ-нейронов и ΣΠ--нейромодулей // Журнал вычислительной математики и математической физики. — 2003. — Т. 43, № 8. — С. 1298–1310.

[10] *Шибзухов З. М.* Конструктивные методы обучения ΣΠ-нейронных сетей. — М.: Наука, 2006. — 159 с.

[11] *Pap E.* g-Calculus // Univ. u N. Sadu Zb. Rad. Prirod-Mat. — 1993. — Vol. 23, № 1. — Pp. 145–156.

[12] *Loday J.-L.* Arithmetree // Algebra. — 2002. — Vol. 258, № 1. — Pp. 275–309.

[13] *Bruno A., Yasaki D.* The aruthmetic of trees. 25.08.2008. `http://arxiv.org/pdf/0809.4448`.

Методы и модели распознавания и прогнозирования

Код раздела: MM (Methods and Models)

- Статистические модели классификации и регрессии.
- Дискретные (логические) модели распознавания.
- Модели классификации на основе сходства и разделимости.
- Нейросетевые модели.
- Методы отбора и преобразования признаков.
- Методы построения алгоритмических композиций.
- Теория и методы прогнозирования временных рядов.
- Нестационарная регрессия.
- Обучение без учителя, кластеризация.
- Методы согласования экспертных оценок.

Оценки обобщающей способности бустинга с вероятностными входами*

Баринова О. В., Ветров Д. П.
obarinova@graphics.cs.msu.ru, vetrovd@yandex.ru
Москва, Московский Государственный Университет имени М. В. Ломоносова

В данной работе предлагается новая верхняя оценка ошибки классификации для композиций простых классификаторов, основанная на сведении бинарной задачи классификации с перекрывающимися распределениями классов к задаче классификации с неперекрывающимися классами. Предложенная верхняя оценка ошибки классификации использует оценки вероятностей принадлежности классам для объектов из обучающей выборки. В случае, когда оценки вероятностей точны, предложенная верхняя оценка ниже, чем известная оценка Шапира и др. [1]. В работе также предлагается новый метод обучения распознаванию образов на основе бустинга, который минимизирует предложенную верхнюю оценку. Результаты экспериментов на реальных данных показывают, что ошибка классификации для композиций, построенных предложенным методом, в среднем ниже, чем для композиций, построенных стандартным методом бустинга.

Введение

Многие исследователи отмечают, что методы бустинга оказываются крайне чувствительными к шуму в обучающей выборке [3, 4]. При этом наблюдается эффект переобучения, проявляющийся в росте тестовой ошибки с увеличением числа итераций бустинга. В данной работе рассматривается задача повышения обобщающей способности бустинга для случая бинарных задач классификации с перекрывающимися распределениями классов.

В статье предложен метод сведения бинарной задачи классификации с перекрывающимися распределениями классов к эквивалентной задаче классификации с неперекрывающимися распределениями классов, где каждому вектору признаков однозначно соответствует метка класса.

Предлагается новая верхняя оценка ошибки классификации для композиций простых классификаторов, основанная на предложенной редукции, которая не зависит от числа слабых классификаторов в композиции. Предложенная верхняя оценка имеет схожий вид с оценкой из теории отступов [1], однако в ней используются оценки условных вероятностей принадлежности классам для объектов из обучающей выборки. Если известные оценки условных вероятностей точны, то предложенная верхняя оценка ниже, чем известная оценка из теории отступов.

Далее, в данной работе предложен новый метод обучения распознаванию образов на основе бустинга, и доказано, что этот метод минимизирует предложенную оценку по аналогии с тем, как AdaBoost минимизирует верхнюю оценку из теории отступов. Для оценивания условных вероятностей принадлежностей классам для объектов из обучающей выборки в работе используется алгоритм, предложенный в [2] .

*Работа выполнена при финансовой поддержке РФФИ, проект № 08-01-00405.

Редукция задачи с перекрывающимися классами

Рассмотрим задачу бинарной классификации. Пусть множество $X = \mathbb{R}^k$ — множество векторов признаков, а множество ответов состоит из двух элементов $Y = \{-1, +1\}$. Пусть дана выборка объектов $S = \{(x_i, y_i)\}_{i=1}^m$, взятых из некоторого неизвестного распределения $P(x, y)$ на $X \times Y$. Оптимальным решением задачи классификации будем называть классификатор $f^*\colon X \to Y$, минимизирующий вероятность ошибочной классификации:

$$f^* \in \operatorname{Arg\,min} \mathsf{P}_{x,y}\big[y \neq f(x)\big].$$

Определение 1. *Будем называть задачу бинарной классификации задачей с перекрывающимися классами, если существует множество ненулевой меры, для которого обе метки классов имеют положительную вероятность:*

$$\mathsf{P}\big[p(+1 \mid x)\, p(-1 \mid x) > 0\big] > 0.$$

Если это условие не выполняется, будем говорить, что классы не перекрываются. Задачи с неперекрывающимися классами однозначно задаются распределением $P(x)$ и целевой функцией $y = c(x)$, поскольку метка класса однозначно задается вектором признаков объекта.

Для задач с перекрывающимися классами введем обозначения:

$$p_{\max}(x) = \max_y p(y \mid x), \quad p_{\min}(x) = \min_y p(y \mid x).$$

Основная идея данной работы заключается в том, чтобы свести задачу с перекрывающимися классами к задаче с неперекрывающимися классами.

Определение 2. *Будем называть две задачи бинарной классификации эквивалентными, если множества их оптимальных решений совпадают.*

Здесь и далее доказательства опущены из-за ограничений по объему.

Теорема 1. *Для любой бинарной задачи классификации с перекрывающимися классами, заданной распределением $P(x, y)$, существует эквивалентная ей задача классификации с неперекрывающимися классами, заданная распределением $\tilde{P}(x)$ и целевой функцией $\tilde{c}(x)\colon X \to Y$, где*

$$\tilde{P}(x) = \big(p_{\max}(x) - p_{\min}(x)\big)P(x)/(1-B);$$
$$\tilde{c}(x) = \arg\max_y p(y \mid x).$$

Причем для любого произвольного классификатора f верно следующее соотношение

$$\mathsf{P}[yf(x) \leqslant 0] = B + (1-2B)\,\tilde{\mathsf{P}}\big[\tilde{c}(x)f(x) \leqslant 0\big],$$

где B — вероятность ошибки байесовского классификатора $\tilde{c}(x)$.

Распределение $\tilde{P}(x)$ будем называть *модифицированным распределением.*

Теорема 1 дает способ свести задачу с перекрывающимися классами к задаче с неперекрывающимися классами. Этот результат будет использован далее для вывода новых верхних оценок ошибки классификации для композиций простых классификаторов.

Верхние оценки ошибки классификации

Предположим, что задан конечный набор слабых классификаторов H. Через C обозначим множество всех линейных комбинаций простых классификаторов из H, где коэффициенты линейной комбинации положительны и их сумма равна 1.

Предположим, что S — выборка из m объектов, выбранных независимо из некоторого распределения над $X \times Y$. Обозначим

$$\mathsf{P}_S\,\psi(x,y) = \frac{1}{m}\sum_{i=1}^{m} I_{\psi(x,y)};$$
$$\mathsf{E}_S\psi(x,y) = \frac{1}{m}\sum_{i=1}^{m} \psi(x,y);$$
$$\mathsf{E}_y\psi(x,y) = \frac{1}{|Y|}\sum_{y\in Y} p(y \mid x)\,\psi(x,y).$$

Отступом объекта (x, y) относительно классификатора $f(x)$ называется величина $yf(x)$.

Теория отступов, разработанная Шапиром и др. [1] на данный момент остается наиболее теоретически строгим объяснением хорошей обобщающей способности бустинга. Ниже мы приведем основные выводы теории отступов.

Теорема 2 (Шапир и др. [1]). *Пусть $P(x, y)$ — распределение на $X \times Y$. Пусть S — выборка из m объектов, выбранных независимо из распределения P. Предположим, что задан конечный набор слабых классификаторов H, принимающих значения из множества $\{-1, +1\}$. Пусть также задана константа $\delta > 0$. Тогда с вероятностью не менее δ все функции $f \in C$ удовлетворяют следующему неравенству для всех $\theta > 0$:*

$$\mathsf{P}[yf(s) \leqslant 0] \leqslant \mathsf{P}_S[yf(x) \leqslant \theta] + \varepsilon\big(m, |H|, \theta, \delta\big),$$

причем

$$\varepsilon\big(m, h, \theta, \delta\big) = O\Big(\sqrt{\tfrac{1}{m\theta^2}\log m \log h + \tfrac{1}{m}\log\tfrac{1}{\delta}}\,\Big).$$

Данная теорема гласит, что ошибка классификации любой композиции простых классификаторов, представляющей собой линейную комбинацию слабых классификаторов из H, где сумма коэффициентов в линейной комбинации равна 1, ограничена сверху суммой значения эмпирической функции распределения отступов на обучающей выборке и дополнительного члена, который зависит от числа объектов в обучающей выборке, числа простых классификаторов в семействе, и двух заданных констант.

Главное свойство дополнительного члена из данной теоремы состоит в том, что он не зависит от числа простых классификаторов в линейной комбинации и их коэффициентов. Если для вычисления этой верхней оценки задана фиксированная обучающая выборка, то данный дополнительный член представляет собой константу.

Таким образом, за счет выбора простых классификаторов в композиции и подбора их коэффициентов можно минимизировать первое слагаемое. В [1] было показано, что алгоритм AdaBoost минимизирует его при увеличении числа итераций.

Теорема 3 (Шапир и др. [1]). *Пусть метод обучения AdaBoost построил простые классификаторы с частотами ошибок на обучающей выборке $\varepsilon_1, \ldots, \varepsilon_T$. Тогда для любого θ, имеем*

$$\mathsf{P}_S[yf(x) \leqslant \theta] \leqslant 2^T \prod_{t=1}^{T} \sqrt{\varepsilon_t^{1-\theta}(1-\varepsilon_t)^{1+\theta}}.$$

В данной работе, следуя логике теории отступов, вводится новая оценка обобщающей способности для композиций простых классификаторов и предлагается метод обучения на основе бустинга, минимизирующий предлагаемую верхнюю оценку с увеличением числа слабых классификаторов в линейной комбинации.

Используя Теорему 1 и первую теорему Шапира, мы можем ограничить сверху ошибку классификации любой линейной комбинации простых классификаторов с положительными коэффициентами, сумма которых равна 1, используя выборку $\tilde{S}$ из модифицированного распределения $\tilde{P}$.

Теорема 4. *Пусть P — распределение на $X \times Y$. Предположим, что задан конечный набор слабых классификаторов H, принимающих значения*

из множества $\{-1,+1\}$. *Пусть эквивалентная задача с неперекрывающимися классами задана распределением* $\tilde{P}(x)$ *и целевой функцией* $\tilde{c}(x)$. *Пусть* $\tilde{S}$ *— выборка из* m *объектов, выбранных независимо из распределения* $\tilde{P}(x)$. *Пусть также задана константа* $\delta > 0$. *Тогда с вероятностью не менее* $1-\delta$ *для выбранной случайным образом выборки* $\tilde{S}$, *все функции* $f \in C$ *удовлетворяют следующему неравенству для всех* $\theta > 0$:

$$\mathsf{P}[yf(x) \leqslant 0] \leqslant B + (1-2B)\,\mathsf{P}_{\tilde{S}}\big[\tilde{c}(x)f(x) \leqslant 0\big] + \\ + (1-2B)\,\varepsilon(m,|H|,\theta,\delta),$$

где функция $\varepsilon(m,h,\theta,\delta)$ *та же, что в теореме 2.*

Полученная верхняя оценка имеет вид, схожий с верхней оценкой из теории отступов, однако дополнительный член в новой оценке в $(1-2B)^{-1}$ раз меньше, чем аналогичный дополнительный член из Теоремы Шапира. Также можно показать, что чем сильнее перекрываются классы (чем больше байесовская ошибка), тем больший выигрыш точности дает новая оценка.

Проблема заключается в том, что на практике вместо выборки $\tilde{S}$ из модифицированного распределения $\tilde{P}$ имеется выборка S из исходного распределения P. Далее речь пойдет о том, как, имея выборку из исходного распределения P и зная истинные значения условных вероятностей $p(+1\,|\,x_i)$ для объектов из обучающей выборки, аппроксимировать ошибку классификации на модифицированном распределении $\tilde{P}$. Следующее утверждение позволяет связать ошибку классификации на модифицированном распределении $\tilde{P}$ с эмпирическим распределением отступов на выборке S из исходного распределения P.

Теорема 5. *Пусть* P *— распределение на* $X \times Y$. *Предположим, что задан конечный набор слабых классификаторов* H, *принимающих значения из множества* $\{-1,+1\}$. *Пусть также задана константа* $\delta > 0$. *Пусть* S *— выборка из* m *объектов, выбранных независимо из распределения* P, *и для каждого объекта известны условные вероятности принадлежности классам. Тогда с вероятностью не менее* $1-\delta$ *для выбранной случайным образом выборки* S *все функции* f *из* C *удовлетворяют следующему неравенству для всех* θ:

$$\mathsf{P}\big[yf(x) \leqslant 0\big] \leqslant B + (1-2\hat{B}) \times \\ \times\, \mathsf{E}_S\left(\frac{p_{\max}(x) - p_{\min}(x)}{1-2\hat{B}} I_{[c(x)f(x)\leqslant\theta]}\right) + \\ + \tilde{\varepsilon}\big(m,|H|,\theta,\delta\big),$$

где B *— ошибка байесовского классификатора на распределении* P; $\hat{B} = \frac{1}{m}\sum\limits_{i=1}^{m} p_{\min}(x_i)$;

$$\tilde{\varepsilon}\big(m,|H|,\theta,\delta\big) = \varepsilon\big(m,|H|,\theta,\delta\big) - 2B\sqrt{\tfrac{1}{m}\ln|H|}.$$

Рассмотрим второе слагаемое из верхней оценки в Теореме 5. Оно представляет собой взвешенную сумму индикаторов события, причем сумма всех весов равна 1, поскольку

$$\frac{1}{m}\sum_{i=1}^{m}\big(p_{\max}(x_i) - p_{\min}(x_i)\big) = 1 - 2\hat{B}.$$

Таким образом, для вычисления второго слагаемого требуется лишь знание условных вероятностей принадлежности классам на обучающей выборке. Последнее условие редко выполняется на практике, однако условные вероятности можно оценить. Вопрос влияния погрешностей оценок условных вероятностей рассматривается в следующем утверждении.

Теорема 6. *Пусть* P *— распределение над* $X \times Y$. *Предположим, что задан конечный набор слабых классификаторов* H, *принимающих значения* $\{-1,+1\}$. *Пусть также заданы константы* $\delta > 0$, $\theta > 0$. *Пусть известны оценки условных вероятностей для объектов из обучающей выборки, удовлетворяющие условию* $|\hat{p}_{\max}(x_i) - p_{\max}(x_i)| < \alpha$, $i = 1,\ldots,m$. *Тогда верна следующая верхняя оценка ошибки классификации по обучающей выборке:*

$$\mathsf{P}[yf(x) > 0] \leqslant B + \\ + (1-2\tilde{B})\,\mathsf{E}_S\left(\frac{\hat{p}_{\max} - \hat{p}_{\min}}{1-2\tilde{B}} I_{[\hat{\mathsf{E}}_y f(x) < \theta + \beta(\alpha)]}\right) + \\ + \frac{2\alpha}{1-2\tilde{B}} + \tilde{\varepsilon}\big(m,|H|,\theta-\beta(\alpha),\delta\big),$$

где $\tilde{\varepsilon}\big(m,|H|,\theta-\beta(\alpha),\delta\big)$ *берется из Теоремы 5,*

$$\tilde{B} = \frac{1}{m}\sum_{i=1}^{m}\hat{p}_{\min}(x_i),$$

а $\beta(\alpha)$ *непрерывно зависит от* α, *причем* $\beta(0) = 0$.

Теоремы 5 и 6 предоставляют новые верхние оценки ошибки классификации для линейных комбинаций простых классификаторов с положительными коэффициентами. Вычислить эти оценки напрямую не представляется возможным, поскольку в них фигурирует значение ошибки байесовского классификатора на распределении P. Тем не менее, можно построить эффективный метод на основе бустинга для минимизации этих оценок. Такой алгоритм описывается в следующем разделе.

Бустинг с вероятностными входами

Следуя логике теории отступов, был разработан итеративный алгоритм, на вход которому подается обучающая выборка и значения оценок условных вероятностей принадлежности классам для объектов из обучения. Данный алгоритм минимизирует верхнюю оценку ошибки классификации из Теоремы 5 (в случае, когда известны точные значения

Алгоритм 1. Бустинг с вероятностными входами.

1: оценить условные вероятности
$\hat{p}(+1 \mid x_i),\ i = 1, \ldots, m;$
2: инициализировать веса
$D_1(i) := \dfrac{\hat{p}_{\max}(x_i) - \hat{p}_{\min}(x_i)}{1-2\tilde{B}},\ i = 1, \ldots, m;$
3: **для** $t = 1, \ldots, T$
4: настроить слабый классификатор h_t:
$\varepsilon_t = \sum\limits_{i=1}^{m} D_t(i)\, \hat{p}\big(y \neq h_t(x_i) \mid x_i\big) \to \min\limits_{h_t};$
5: вычислить вес слабого классификатора:
$\alpha_t := \dfrac{1}{2} \ln \dfrac{1-\varepsilon_t}{\varepsilon_t};$
6: пересчитать веса объектов из обучения:
$D'(i) := \frac{1}{Z} D_t(i) \mathsf{E}_y \exp(-y\alpha_t h_t(x));$
$D_{t+1}(i) := D'(i) \Big/ \sum\limits_{j=1}^{m} D'(j);\ i = 1, \ldots, m;$
7: **вернуть** итоговый классификатор:
$f(x) = \operatorname{sign} \sum\limits_{i=1}^{T} \alpha_t h_t(x);$

вероятностей) или Теоремы 6 (если используются оценки для условных вероятностей).

Следующая теорема показывает, что предложенный метод обучения минимизирует верхнюю оценку ошибки классификации из Теоремы 6.

Теорема 7. *Пусть предложенный алгоритм сгенерировал классификаторы с взвешенными ожидаемыми ошибками на обучающей выборке* $\varepsilon_1, \ldots, \varepsilon_T$. *Тогда для любого* $\theta > 0$ *выполняется неравенство:*

$$\mathsf{E}_S \left(\frac{\hat{p}_{\max}(x_i) - \hat{p}_{\min}(x)}{1 - 2\tilde{B}} I_{[\mathsf{E}_y(yf(x)) \leqslant \theta]} \right) \leqslant \leqslant 2^T \prod_{t=1}^{T} \sqrt{\varepsilon_t^{1-\theta}(1-\varepsilon_t)^{1+\theta}}.$$

Поскольку предложенный метод обучения использует оценки условных вероятностей для объектов из обучающей выборки, перед применением процедуры обучения требуется оценить условные вероятности. В данной работе для этого используется метод, предложенный в [2].

Эксперименты

Был проведен ряд экспериментов на задачах бинарной классификации из репозитория UCI, в которых предложенный метод сравнивался со стандартным методом AdaBoost, в качестве простых классификаторов были использованы элементарные пороговые правила. Данные случайным образом разделялись на две равные части: на одной из них настраивался классификатор, другая использовалась в качестве тестовой выборки для вычисления ошибки классификации, эта процедура повторялась 5 раз для каждого набора данных. Результаты представлены в таблице 1.

Таблица 1. Ошибки после 1000 итераций.

Задача	**AdaBoost**	**Предл. метод**
Austra	$16{,}0 \pm 0{,}1$	$13{,}1 \pm 0{,}1$
Australian	$17{,}4 \pm 0{,}1$	$13{,}0 \pm 0{,}1$
Breast	$5{,}4 \pm 0{,}0$	$4{,}4 \pm 0{,}1$
Checker09	$29{,}5 \pm 0{,}4$	$27{,}0 \pm 0{,}2$
German	$25{,}7 \pm 0{,}3$	$25{,}0 \pm 0{,}1$
Madelon	$41{,}5 \pm 0{,}3$	$38{,}0 \pm 0{,}1$
Magic04	$17{,}9 \pm 0{,}1$	$16{,}4 \pm 0{,}1$
Page	$3{,}3 \pm 0{,}1$	$2{,}9 \pm 0{,}1$
Pima	$26{,}2 \pm 0{,}1$	$24{,}7 \pm 0{,}2$
Vote	$4{,}6 \pm 0{,}2$	$4{,}3 \pm 0{,}2$

Выводы

В данной работе была получена новая верхняя оценка ошибки классификации для композиций простых классификаторов, основанная на сведении задачи классификации с перекрывающимися классами к задаче с неперекрывающимися классами. Предложенная верхняя оценка не зависит от числа простых классификаторов в композиции и имеет схожий вид с верхней оценкой из теории отступов. Для ее вычисления и оптимизации требуется знание оценок условных вероятностей принадлежности классам для объектов из обучающей выборки. В случае, когда оценки вероятностей точны, данная верхняя оценка является более точной, чем оценка из теории отступов.

В работе был предложен метод обучения на основе бустинга, который гарантированно минимизирует полученную оценку. В экспериментах использовался алгоритм из [2] для получения оценок условных вероятностей. Результаты экспериментов на реальных задачах из репозитория UCI демонстрируют, что предложенный метод в среднем работает лучше, чем стандартный AdaBoost.

Благодарности: Авторы хотели бы поблагодарить Д. А. Кропотова за помощь в обсуждении некоторых аспектов работы.

Литература

[1] *Schapire R., Freund Y., Bartlett P., Wee Sun Lee.* Boosting the margin: A new explanation for the effectiveness of voting methods // Proc. of the 14th International Conference on Machine Learning — 1997.

[2] *Vezhnevets A., Barinova O.* Avoiding boosting overfitting by removing confusing samples // Proc. of European Conference on Machine Learning — 2007.

[3] *Grove A. J., Schuurmans D.* Boosting in the limit: maximizing the margin of learned ensembles // Proc. of 15th National Conf. on Artifical Intelligence — 1998.

[4] *Dietterich T. G.* An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization // Machine Learning — Vol. 40, No. 2. — 1999.

Построение ансамбля логических моделей в кластерном анализе*

Бериков В. Б.
berikov@math.nsc.ru
Новосибирск, Институт математики СО РАН

В работе предлагается метод кластерного анализа, основанный на ансамбле древообразных логических моделей (решающих деревьев). При построении ансамбля учитываются расстояния между логическими высказываниями, описывающими кластеры; кроме того, используется процедура адаптивного планирования испытаний. Рассматриваются некоторые свойства байесовской модели классификации, которые используются при обосновании информационно-вероятностного критерия качества группировки. Приводятся результаты экспериментальных исследований, демонстрирующие преимущество предложенного метода.

Одной из актуальных проблем в кластерном анализе является обработка информации, описываемой разнотипными (количественными или качественными) характеристиками. В случае разнотипного пространства, возникает методологическая проблема определения в нем метрики. Другой актуальной проблемой является повышение устойчивости группировочных решений. В большинстве алгоритмов кластерного анализа результаты могут сильно меняться в зависимости от выбора начальных условий, порядка объектов, параметров настройки и т. п. Известно, что устойчивость решений может быть повышена путём применения ансамблей алгоритмов (см, например, [1]). При этом используются результаты («знания»), полученные различными алгоритмами, или одним алгоритмом, но с разными параметрами настройки, по различным подсистемам переменных и т. д. После построения ансамбля проводится нахождение итогового коллективного решения.

Одним из перспективных подходов к решению задач кластерного анализа является подход, основанный на логических моделях. Такого рода модели широко используются для решения задач распознавания и прогнозирования. Это объясняется хорошей интерпретируемостью результатов, имеющих вид логических закономерностей, высокой прогнозирующей способностью в условиях неопределённости, возможностью обрабатывать разнотипные переменные, выделять наиболее важные факторы. Разработке алгоритмов построения логических моделей кластерного анализа были посвящены например, работы [2, 3]. В работе [4] был описан метод построения логической функции в задаче кластерного анализа, основанный на рекурсивном алгоритме построения дерева решений. Этот алгоритм позволяет путём увеличения глубины перебора находить более сложные логические закономерности, описывающие структуру кластеров. В работе [5] был предложен метод кластерного анализа, основанный на коллективе деревьев решений, формируемых по случайно отобранным подсистемам переменных. Как показало статистическое моделирование, использование ансамбля позволяет значительно улучшить качество классификации, по сравнению с «несогласованными» деревьями. В данной работе предлагается метод, который при согласовании базовых решений дополнительно учитывает расстояния между логическими высказываниями, описывающими кластеры; а также применяет адаптивное планирование испытаний при построении ансамбля. Кроме того, предлагается новый информационно-вероятностный критерий качества группировки, основанный на байесовской модели классификации по конечному множеству событий.

Основные понятия

В стандартной постановке задачи кластерного анализа требуется сформировать сравнительно небольшое число групп (кластеров, классов) объектов, которые были бы как можно более схожими между собой внутри каждой группы, и как можно более отличающимися в разных группах.

Будем использовать вероятностный подход — предположим, что существует латентная переменная, имеющая недерминированный характер, определяющая принадлежность объектов классам. Требуется с минимальной вероятностью ошибки классифицировать объекты. Поскольку целевая переменная непосредственно ненаблюдаема, качество классификации можно оценить лишь косвенно (в рамках некоторой модели и т. д.).

Пусть имеется выборка объектов исследования $s = \{o^{(1)}, \ldots, o^{(N)}\}$, которая сформирована в результате отбора некоторых представителей генеральной совокупности. Требуется сформировать $K \geqslant 2$ классов; их число может быть как выбрано заранее, так и не задано (в последнем случае оптимальное количество групп должно быть определено автоматически). Каждый объект генеральной совокупности описывается с помощью набора переменных $X = (X_1, \ldots, X_n)$. Набор может включать переменные разных типов (количественные и качественные, под которыми будем понимать номинальные и булевы; а также порядковые). Пусть D_j обозначает множество значений переменной X_j.

*Работа выполнена при финансовой поддержке РФФИ, проекты №08-07-00136а, №07-01-00331а.

Обозначим через $x_j^{(i)} = X_j(o^{(i)})$ значение j-й переменной для i-го объекта.

Под группировочным решением будем понимать разбиение выборки $G = \{C^{(1)}, \ldots, C^{(K)}\}$, где $C^{(k)} = \{o^{(i_1)}, \ldots, o^{(i_{N_k})}\}$, N_k — число объектов, входящих в k-й кластер, $k = 1, \ldots, K$. Группировочной решающей функцией назовём отображение $f\colon s \to \{1, \ldots, K\}$.

Под древообразной логической моделью группирования данных будем понимать дерево, в котором внутренней вершине соответствует некоторая переменная X_j, а ветвям, выходящим из данной вершины, соответствует определенное высказывание вида $X_j(o) \in E_j^{(q)}$, где o — некоторый объект, $q = 1, \ldots, r$, $r \geqslant 2$ — число ветвей, выходящих из данной вершины, причём набор $E_j^{(1)}, \ldots, E_j^{(r)}$ есть разбиение множества значений D_j. Каждому m-ому листу (концевой вершине) дерева соответствует группа объектов выборки, удовлетворяющих цепочке высказываний, проверяемых по пути из корневой вершины в этот лист. Данной цепочке можно сопоставить логическое утверждение вида

$$\text{«Если } X_{j_1}(o) \in E_{j_1}^{(i_1)} \text{ И } \ldots \text{ И } X_{j_{q_m}}(o) \in E_{j_{q_m}}^{(i_{q_m})},$$
$$\text{то объект } o \text{ относится к } m\text{-й группе»},$$

где q_m — длина данной цепочки, $m = 1, \ldots, M$. Описанное дерево будем называть группировочным деревом решений. Этому дереву соответствует разбиение пространства переменных на M попарно непересекающихся подобластей $E^{(1)}, \ldots, E^{(M)}$, так что каждому m-му листу сопоставляется подобласть $E^{(m)}$. Разбиению пространства переменных, в свою очередь, соответствует разбиение выборки на подмножества $C^{(1)}, \ldots, C^{(M)}$.

Рассмотрим произвольную группу объектов $C^{(m)}$. Описанием этой группы назовём следующий набор высказываний: $X_1(o) \in T_1^{(m)}, \ldots, X_n(o) \in T_n^{(m)}$, где $T_j^{(m)}$ — отрезок $\left[\min\limits_{o \in C^{(m)}} X_j(o), \max\limits_{o \in C^{(m)}} X_j(o)\right]$ в случае количественной или порядковой переменной X_j, либо множество принимаемых значений $\{X_j(o) |\, o \in C^{(m)}\}$ в случае качественной переменной. Подобласть пространства переменных $T = T_1 \times \ldots \times T_n$, соответствующую описанию группы, назовём *таксоном*. Относительной мощностью (объёмом) j-й проекции таксона T назовём величину $\delta_j = \frac{|T_j|}{|D_j|}$, где через $|T_j|$ обозначена длина интервала (в случае количественной или порядковой переменной X_j) либо мощность (число значений) соответствующего подмножества в случае качественной переменной X_j, $j = 1, \ldots, n$. Под объёмом таксона будем понимать величину $\prod\limits_{j=1}^{n} \delta_j$.

В работе [4] был описан алгоритм, который осуществляет направленный поиск дерева, оптимального по заданному критерию качества. Предложены два варианта критерия. В первом варианте требуется найти дерево, для которого суммарный объём минимален, для заданного числа кластеров. Во втором варианте необходимо достичь компромисса между суммарным объёмов таксонов и их числом. Для построения дерева используется рекурсивный R-метод.

Ансамбль логических моделей

Будем формировать различные варианты базовых группировочных решений $G^{(l)}$, $l = 1, \ldots, L$, причём для каждого варианта определяется своя подсистема переменных, в пространстве которых проводится группировка с помощью R-метода. Выбор подсистемы будем проводить случайным образом. Требуется построить согласованное группировочное решение. При этом взаимно подтверждающиеся (устойчивые) закономерности, выявленные при построении отдельных решений, должны «усиливаться», а неустойчивые — «ослабляться».

Для выбора наилучшей согласующей функции могут быть использованы различные принципы. В ряде работ используется принцип, основанный на нахождении согласованной матрицы подобия (или различия) объектов. Обозначим через $S^{(l)}$ бинарную матрицу $S^{(l)} = \left(S_{iq}^{(l)}\right)_{N \times N}$, которая вводится для l-й группировки следующим образом:

$$S_{iq}^{(l)} = \begin{cases} 0, & o^{(i)} \text{ и } o^{(q)} \text{ лежат в одном кластере;} \\ 1, & \text{иначе;} \end{cases}$$

для всех i, $q = 1, \ldots, N$, $l = 1, \ldots, L$. После построения L группировочных решений можно сформировать согласованную матрицу различий $S = (S_{iq})_{i,j=1}^{N}$, где $S_{iq} = \frac{1}{L} \sum\limits_{l=1}^{L} S_{iq}^{(l)}$. Величина S_{iq} равна частоте классификации объектов $o^{(i)}$ и $o^{(q)}$ в разные группы в наборе группировок. Близкое к нулю значение величины означает, что данные объекты имеют большой шанс попадания в одну и ту же группу. Близкое к единице значение этой величины говорит о том, что шанс оказаться в одной группе у объектов незначителен.

В предлагаемом алгоритме используется аналогичный принцип, однако вместо бинарной матрицы $S^{(l)}$ в которой отражаются события типа «вхождение» (либо «невхождение») пары объектов в одну и ту же группу, предлагается использовать более информативную матрицу расстояний между кластерами, к которым отнесены объекты. Для вычисления расстояний между кластерами в разнотипном пространстве переменных будем использовать введённое в работе [4] расстояние между логическими высказываниями экспертов.

Пусть $T^{(s)}$, $T^{(q)}$ — два различных таксона. Рассматривалось два способа определения расстояния между ними. В первом варианте расстоя-

ние $\rho_E(T^{(s)},T^{(q)}) = \Big(\sum_{j=1}^n \rho_j^2\big(T^{(s)},T^{(q)}\big)\Big)^{1/2}$, где $\rho_j(T^{(s)},T^{(q)})$ — расстояние между j-ми проекциями областей. Во втором варианте $\rho_{\max}(T^{(s)},T^{(q)}) = \max\limits_j \rho_j(T^{(s)},T^{(q)})$. Если X_j — номинальная переменная, то расстояние между проекциями определяется как взвешенная мера симметрической разности: $\rho_j(T_j^{(s)},T_j^{(q)}) = |T_j^{(s)} \Delta\, T_j^{(q)}| \;/\; |D_j|$. Если X_j — количественная или порядковая переменная, $T_j^{(s)} = \big[a^{(s)},b^{(s)}\big]$, $T_j^{(q)} = \big[a^{(q)},b^{(q)}\big]$, то $\rho_j\big(T_j^{(s)},T_j^{(q)}\big) = \big(|a^{(s)}-a^{(q)}| + |b^{(s)}-b^{(q)}|\big) \;/\; 2|D_j|$.

После вычисления согласованной матрицы подобия для нахождения итогового варианта группировки будем применять иерархический метод построения дендрограммы, который в качестве входной информации использует попарные расстояния между объектами (расстояния между группами будем вычислять по принципу «средней связи»).

Адаптивное планирование при формировании ансамбля

В описанном алгоритме выбор подсистемы переменных осуществляется случайно. Этот выбор («испытание») можно организовать так, чтобы лучшие по критерию качества группировки варианты имели бы бóльшую вероятность отбора. В основе предлагаемой модификации лежит идея адаптивного случайного поиска наиболее информативной подсистемы переменных [6].

Первоначально все переменные имеют одинаковую вероятность отбора. Если после проведения очередного испытания критерий качества уменьшается, то переменные, входящие в подсистему, «наказываются» (т. е. вероятность их отбора уменьшается пропорционально критерию). В случае, когда критерий качества улучшается, то соответствующая подсистема переменных «поощряется» (т. е. вероятность их отбора увеличивается). Величина поощрения (наказания) зависит от числа проведённых экспериментов: чем больше проведено испытаний, тем больше соответствующий параметр.

Заметим, что существуют адаптивные методы построения группировочного ансамбля, основанные на генетическом алгоритме оптимизации. Однако при использовании такого алгоритма возникают проблемы, связанные со спецификой задачи кластерного анализа — с трудностью интерпретации операторов рекомбинации и кроссовера.

Использование байесовской модели классификации

В работе [4] и др. была предложена байесовская модель распознавания по конечному множеству событий, которая может использоваться для построения оптимальных по сложности логических функций. В данном параграфе рассматриваются некоторые свойства модели, связанные с информационно-вероятностным критерием качества логической функции в задаче кластерного анализа.

Пусть задано некоторое группировочное дерево решений, которому соответствует разбиение пространства переменных на M подобластей. Тогда число классов $K = M$. Обозначим $\theta = (p_1^{(1)},\ldots,p_M^{(K)})$, где $p_m^{(k)}$ есть вероятность принадлежности к m-й подобласти случайно выбранного объекта k-го класса. Пусть $\boldsymbol{\Theta}$ — множество возможных значений θ. Рассмотрим байесовскую модель распознавания по конечному множеству событий, которым соответствуют подобласти разбиения. В этой модели предполагается, что на $\boldsymbol{\Theta}$ определена случайная величина Θ с априорным распределением $\theta \sim \mathrm{Dir}(\boldsymbol{d})$, где $\boldsymbol{d} = \big(d_1^{(1)},\ldots,d_M^{(K)}\big)$ — вектор параметров распределения Дирихле. В работе [4] предложен способ задания параметров с помощью экспертной оценки степени «пересечения» между образами. При отсутствии априорных знаний можно использовать равномерное априорное распределение $d_m^{(k)} \equiv 1$.

Зададим энтропию вектора θ как $H(\theta) = -\sum\limits_{k,m} p_m^{(k)} \ln p_m^{(k)}$. Рассмотрим математическое ожидание $\mathsf{E}H(\Theta)$ данной величины, где усреднение проводится по множеству $\boldsymbol{\Theta}$.

Теорема 1. *Выполняется следующее равенство:*

$$\mathsf{E}H(\Theta) = \Psi(D+1) - \frac{1}{D}\sum_{k,m} d_m^{(k)}\Psi\big(d_m^{(k)}+1\big),$$

где $\Psi(z) = \frac{d}{dz}\ln\Gamma(z)$ *— дигамма функция,* $D = \sum\limits_{k,m} d_m^{(k)}$.

Пусть выборка преобразована в вектор предполагаемых частот попадания объектов каждого из латентных классов в подобласти $\boldsymbol{s} = \{n_1^{(1)},\ldots,n_M^{(K)}\}$, где $n_m^{(k)} = 0$ при $k \neq m$, и $n_m^{(k)} > 0$ при $k = m$, $k,m = 1,\ldots,M$. По свойству распределения Дирихле, апостериорное распределение $\Theta \mid \boldsymbol{s} \sim \mathrm{Dir}(\boldsymbol{d}+\boldsymbol{s})$. Из теоремы следует, что ожидаемая энтропия при условии известной выборки

$$\mathsf{E}H(\Theta \mid \boldsymbol{s}) = \Psi(D+N+1) - \\ - \frac{1}{D+N}\sum_{k,m}\big(d_m^{(k)}+n_m^{(k)}\big)\Psi\big(d_m^{(k)}+n_m^{(k)}+1\big).$$

Определение 1. *Назовём ожидаемым количеством информации, содержащейся в классифицированной выборке, величину*

$$I_{\boldsymbol{s}} = \mathsf{E}H(\Theta) - \mathsf{E}H(\Theta \mid \boldsymbol{s}).$$

Введённое понятие может использоваться как критерий информативности группировочной решающей функции. Оптимальному варианту разбиения соответствует максимальное значение критерия.

Экспериментальные исследования

Для определения качества разработанного метода применялась процедура, состоящая в многократном генерировании случайных выборок в соответствии с заданным распределением для каждого класса; построении согласованного группировочного решения для каждой выборки; нахождении усредненного по всем выборкам показателя качества. Качество группировки определяется как частота правильной классификации P_{cor}. Вычислялся также индекс Ранда IR, представляющий собой относительное число пар объектов, у которых либо одинаковые, либо разные номера классов в полученной и истинной группировке (значение индекса, близкое к 1, говорит о хорошей согласованности группировок). Этот индекс более удобен при числе классов $K \geqslant 3$. Для построения базовых деревьев использовался рекурсивный алгоритм.

Пример 1. Распределение для каждого из $K = 3$ классов является многомерным нормальным с ковариационной матрицей $\Sigma = \sigma_{ij}$, где $\sigma_{ij} = 3$. Число переменных $n = 100$; объём выборки для каждого класса равен 25. Вектора математических ожиданий переменных для каждого класса выбирались случайно из множества $\{1, \ldots, 10\}$. Каждое дерево строилось в подпространстве размерности 2. Усреднение проводилось по 100 случайным выборкам, являющихся реализациями смеси указанных распределений (с равными весами). На рис. 1 приведены полученные усредненные значения индекса Ранда при различном размере ансамбля.

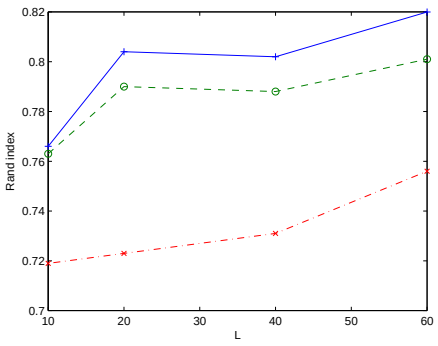


Рис. 1. Результаты работы различных вариантов алгоритма: «+» — вариант, использующий расстояние $\rho_{\max}$; «○» — использующий расстояние ρ_E; «×» — основанный на бинарной согласованной матрице различий.

Пример 2. В отличие от предыдущего примера, число классов $K = 2$; из общего числа переменных 50 являются количественными, а 50 — булевыми. Вектор математического ожидания для каждого класса выбирается случайно из множества, соответствующего вершинам единичного гиперкуба; ковариационная матрица является диагональной: $\Sigma = \sigma I$, где σ принимает значения из множества $\{0.01, 0.04, 0.1, 0.5, 1\}$. Для булевых переменных значения, полученные с помощью датчика случайных чисел, округляются до ближайшего целого из множества $\{0, 1\}$. Объём выборки для первого и второго класса равен 25. Число деревьев в ансамбле 10; каждое дерево строится в подпространстве размерности 2. В таблице 1 приведены значения усредненных показателей качества. Для сравнения, указаны аналогичные усредненные показатели для одиночных деревьев.

Таблица 1. Показатели качества алгоритмов.

Показатели	σ				
	0.01	0.04	0.1	0.5	1
P_{cor} (анс.)	1	1	0.999	0.85	0.75
IR (анс.)	1	1	0.998	0.75	0.62
P_{cor} (дер.)	1	0.87	0.85	0.65	0.61
IR (дер.)	1	0.98	0.81	0.57	0.53

Выводы

Экспериментальные исследования показывают, что применение ансамбля логических моделей позволяет заметно улучшить качество классификации. Это можно объяснить тем, что в коллективе «усиливаются» устойчивые закономерности, выявленных при построении отдельных решений, и «ослабляются» неустойчивые.

В дальнейшем планируется провести теоретическое исследование ансамблевых методов кластерного анализа, в том числе с использованием байесовских логико-вероятностных моделей.

Литература

[1] *Бирюков А. С., Рязанов В. В., Шмаков А. С.* Решение задач кластерного анализа коллективами алгоритмов // Журнал вычисл. матем. и матем. физики. — 2008. — Т. 48, № 1. — С. 176–192.

[2] *Michalski R., Stepp R., Diday E.* Automated construction of classifications: conceptual clustering versus numerical taxonomy // IEEE Trans. Pattern Anal. Mach. Intell. — 1983. — Vol. 5. — Pp. 396–409.

[3] *Лбов Г. С., Пестунова Т. М.* Группировка объектов в пространстве разнотипных переменных // В сб. «Анализ нечисловой информации в социологических исследованиях». М.: Наука, 1985. — С. 141–149.

[4] *Лбов Г. С., Бериков В. Б.* Устойчивость решающих функций в задачах распознавания образов и анализа разнотипной информации. — Новосибирск: Издательство Института математики, 2005. — 218 с.

[5] *Бериков В. Б.* Кластерный анализ с использованием случайного леса решений в пространстве переменных большой размерности // Вычислительные технологии. — 2008. — Т. 13. Часть 1. — С. 294–301.

[6] *Лбов Г. С.* Методы обработки разнотипных экспериментальных данных. — Новосибирск: Наука, 1981. — 160 с.

Сходство и компактность*

Борисова И. А., Дюбанов В. В., Загоруйко Н. Г., Кутненко О. А.
zag@math.nsc.ru
Институт математики им. С. Л. Соболева СО РАН, Новосибирский государственный университет

В статье вводится понятие функции конкурентного сходства (FRiS-функции), с помощью которой можно оценивать сходство между объектами и образами, получать количественные меры компактности образов и информативности признакового пространства. Описывается опыт использования предлагаемых мер сходства и компактности для решения задачи из области молекулярной биологии.

Мера сходства

Сходство $S(a,b)$ двух объектов a и b обычно оценивается величиной, которая зависит от расстояния $R(a,b)$ между этими объектами. Предполагается, что свойства расстояний — симметричность, рефлексивность, неравенство треугольника — проецируются и на меру сходства. Однако при распознавании образов нас интересует мера сходства с другими свойствами. Будем рассматривать сходство контрольного объекта z с объектами a и b, которые являются представителями (ближайшими объектами или эталонами) образов A и B, так что слова «сходство с объектом» будут означать то же, что и слова «сходство с образом». Для принятия решения о принадлежности контрольного объекта z к образу A недостаточно знать расстояние $R(z,a)$. Нужно знать также расстояние $R(z,b)$ и определить, что расстояние $R(z,a)$ является наименьшим из них. Следовательно, нужно иметь не абсолютную, а относительную меру сходства, величина которой зависит от расстояний до представителей конкурирующих образов. Если оценивается сходство между тремя объектами — a, b и c, то при оценке похожести объекта a на объект b должны учитываться расстояния $R(a,b)$ и $R(a,c)$, а при оценке похожести объекта b на объект a должны учитываться расстояния $R(b,a)$ и $R(b,c)$. Следовательно, относительная мера сходства $\bar{S}$ не обладает свойством симметричности: $\bar{S}(a,b) \neq \bar{S}(b,a)$. Не выполняется для этой меры и неравенство треугольника: сумма сходств между вершинами треугольника $\bar{S}(a,b)+\bar{S}(a,c)$ может быть как меньше, так и больше сходства $\bar{S}(b,c)$. Так что сходство, в отличие от расстояния, не образует *метрического* пространства. Относительная мера сходства, учитывающая конкурентную ситуацию, образует пространство, которое мы называем *конкурентным*.

Некоторые известные алгоритмы распознавания используют относительную меру сходства. Например, в методе k ближайших соседей (kNN) новый объект z распознается как объект образа A, если расстояние $R(z,A)$ до k ближайших объектов этого образа не только мало, но меньше, чем расстояние $R(z,B)$ до k ближайших объектов конкурирующего образа B. Оценка сходства в этом алгоритме делается в шкале порядка.

Более сложная мера сходства используется в алгоритме RELIEF [1]. Чтобы определить сходство объекта z с образом A в конкуренции с образом B используется величина

$$W_{A/B} = \frac{R(z,B) - R(z,A)}{R_{\max} - R_{\min}},$$

где $R_{\max}$ и $R_{\min}$ — максимальное и минимальное расстояния между всеми парами объектов. Сформулируем следующие требования, которым должна удовлетворять мера $F_{a/b}(z)$ сходства объекта z с объектом a в конкуренции с объектом b.

1. Мера сходства должна зависеть не от характера распределения всего множества объектов, а от особенностей распределения объектов в окрестности объекта z.

2. Если оценивается мера сходства объекта z с объектом a, и ближайшим соседом z является объект b, $b \neq a$, то при совпадении объектов z и a мера $F_{a/b}(z)$ должна иметь максимальное значение, равное $+1$, а при совпадении z с b — максимальное отрицательное значение, равное -1. Во всех остальных случаях мера конкурентного сходства принимает значения от -1 до $+1$.

3. При одинаковых расстояниях $R(z,a)$ и $R(z,b)$ объект z в равной степени будет похожим на объекты a и b,и меры сходства $F_{a/b}(z)$ и $F_{b/a}(z)$ должны быть равны 0.

Предлагаемая нами функция конкурентного сходства FRiS (Function of Rival Similarity) удовлетворяет всем этим требованиям [2]:

$$F_{a/b}(z) = \frac{R(z,b) - R(z,a)}{R(z,b) + R(z,a)}.$$

Выбор эталонов

Для выбора эталонных образцов (столпов), на основании сходства с которыми будет оцениваться конкурентное сходство контрольных объектов с образами, нами предлагается алгоритм FRiS-Stolp. Этот алгоритм выбирает эталоны следующим способом. Проверяется вариант, при котором первый

*Работа выполнена при финансовой поддержке РФФИ, проект № 08-01-00040, Международного фонда «Научный потенциал» и гранта АВЦП Рособразования, проект № 2.1.1/3235.

случайно выбранный объект $a_i,\ i = 1, \ldots, M_A$ образа A является единственным его столпом, а в качестве столпов образа B используются все его M_B объектов.

1. Для каждого объекта $a_j,\ j \neq i$, образа A находим расстояние r_{ji} до столпа a_i и расстояние r_{jb} до ближайшего к нему объекта b образа B. По этим расстояниям вычисляем значение функции сходства:

$$F_{ji/b} = \frac{r_{jb} - r_{ji}}{r_{jb} + r_{ji}}.$$

Чем больше эта величина, тем лучше объект a_i защищает объект a_j от включения его в состав образа B. Добавим полученную величину к счетчику C_i^1.

2. Повторив шаг 1 для всех $(M_A - 1)$ объектов $a_j,\ j \neq i$, получим в счетчике C_i^1 сумму оценок сходства объектов образа A с объектом a_i. Разделив эту сумму на $(M_A - 1)$, получим оценку F_i «обороноспособности» объекта a_i:

$$F_i^1 = \frac{C_i^1}{(M_A - 1)}.$$

3. Проделав шаги 1 и 2 для всех M_A объектов, мы получим оценки «обороноспособности» каждого из них. Теперь нужно проверить объект a_i на толерантность к объектам образа B. Для этого оценим сходство с a_i всех объектов $b_q,\ q = 1, \ldots, M_B$, образа B в предположении, что роль столпа этого образа будет играть объект b_s, который является ближайшим соседом объекта b_q.

4. Вычислим величину $F_{qs/i} = \frac{r_{qi} - r_{qs}}{r_{qi} + r_{qs}}$ сходства объекта b_q со своим столпом b_s в конкуренции со столпом a_i и добавим эту величину в счетчик C_i^2. Если эта величина положительна, то это повышает шансы объекта a_i стать столпом образа A. И наоборот. Повторив шаг 4 для всех объектов образа B, мы получим оценку F_i^2 толерантности объекта a_i по отношению к объектам образа B:

$$F_i^2 = \frac{C_i^2}{M_B}.$$

5. Если v — стоимость ошибки первого рода, а w — стоимость ошибки второго рода, то общую оценку F_i эффективности объекта a_i в качестве столпа образа A примем равной

$$F_i = \frac{vF_i^1 + wF_i^2}{v + w}.$$

Чем больше величина F_i^1, тем меньше будет ошибок первого рода (пропуск цели). Чем больше величина F_i^2, тем меньше будет ошибок второго рода (ложная тревога). Так что, их совместный учет должен отражать соотношение цен этих ошибок.

6. Повторяя шаги 4 и 5, мы получим такие оценки для для всех M_A объектов образа A. В качестве первого столпа образа A выбираем тот объект a_i, которому соответствует наибольшая величина F_i.

7. Затем выполним шаги 1–6 для объектов b_s образа B, $s = 1, \ldots, M_B$. Выбираем объект b_s, который получил наибольшую величину F_s, и объявляем его первым столпом образа B.

8. Теперь образы представлены не всеми объектами, а только своими столпами. В новых условиях выбор столпов может дать другой результат. Для проверки этого повторим шаги 1–7 с той разницей, что в качестве столпов конкурирующих образов будем использовать их столпы, выбранные на предыдущем этапе. Опыт показывает, что одной такой проверки оказывается достаточно.

9. Найдем объекты, сходство которых со своими столпами превышают заданный порог F^*, например, $F^* = 0$. Эти объекты образуют первые кластеры соответсвующих образов A и B.

10. Если не все M объектов вошли в эти кластеры, то для остальных объектов повторим шаги 1–9. При этом в качестве столпов конкурирующих образов будем использовать все их столпы, выбранные на предыдущих этапах. Шаг 10 повторяем до шага, после которого все объекты обучающей выборки оказываются включенными в свои кластеры. В итоге образы A и B будут представлены k_A и k_B столпами, соответственно.

Если количество образов $K > 2$, то задача сводится к предыдущей следующим способом. При выборе столпов последовательно для каждого образа (A) объекты всех остальных образов объединяются в один конкурирующий образ (B).

При нормальных распределениях в первую очередь будут выбраны столпы, расположенные в точках математического ожидания. Если распределения полимодальны и образы линейно неразделимы, столпы будут стоять в центрах мод. Количество столпов зависит от компактности образов.

Процесс распознавания с опорой на столпы состоит в оценке функций конкурентного сходства контрольного объекта z с двумя самыми близкими столпами разных образов. Решение принимается в пользу того образа, на столп которого контрольный объект похож больше всего.

Оценка компактности и цензурирование выборки

Практически все алгоритмы распознавания основаны на использовании гипотезы компактности. При определении компактности часто оперируют такими нечеткими понятиями, как «достаточно малое количество граничных точек», «не слишком вычурная граница» и т. д. Хотелось бы получить количественную меру компактности, значение ко-

торой было бы прямо связано с ожидаемой надежностью распознавания.

Одна из мер такого рода предложена в [3] и состоит в вычислении профиля компактности. Для каждого из M объектов a_i обучающей выборки все остальные $(M-1)$ объектов упорядочиваются по их расстоянию до a_i. При движении вдоль этих упорядоченных списков от первой позиции $j=1$ до последней $j=M-1$ в каждой порядковой позиции определяется количество объектов m_j, которые не принадлежат тому образу, которому принадлежит объект a_i. Величины $V_j = m_j/M,\ j=1,\ldots,M-1$, и формируют профиль компактности. Чем компактнее образы, тем для большего числа первых порядковых номеров профиля $j=1,\ldots,M-1$ выполнено $V_j = 0$. Переход от профиля к количественной оценке компактности может делаться разными способами. В работе [3] описывается связь между профилем компактности и функционалом полного скользящего контроля, который является естественной количественной оценкой и компактности, и обобщающей способности для метода kNN.

Для получения количественной оценки компактности можно использовать описанную выше FRiS-функцию. Будем оценивать компактность образа A в задаче распознавания K образов. При выборе столпов образа A мы получили оценки F_i для всех M_A его объектов. Компактностью $\mathcal{F}_\mathcal{A}$ образа A будем считать среднее значение этих величин:

$$\mathcal{F}_\mathcal{A} = \frac{1}{M_A}\sum_{i=1}^{M_A} F_i.$$

Общая оценка $\mathcal{F}$ компактности K образов в данном признаковом пространстве, а следовательно, и информативности этого пространства, может быть получена путем арифметического или геометрического усреднения. Для минимизации ошибок всех образов в среднем следует использовать арифметическое усреднение:

$$\mathcal{F}' = \frac{1}{K}\sum_{j=1}^{K} \mathcal{F}_j.$$

Если нужно, чтобы компактность самого некомпактного образа была максимально возможной, тогда нужно использовать среднегеометрическую величину:

$$\mathcal{F} = \sqrt[K]{\prod_{j=1}^{K} \mathcal{F}_j}.$$

Наши эксперименты показывают, что критерий $\mathcal{F}$ обычно дает лучший результат по сравнению с критерием $\mathcal{F}'$. Описанная мера компактности тем больше, чем выше плотность объектов внутри образов, и чем дальше образы отстоят друг от друга. Она используется в качестве меры информативности признакового пространства в алгоритме FRiS-GRAD [2].

Найденные в процессе выбора столпов оценки F_i позволяют наметить пути решения проблемы «цензурирования» выборки. Оценка F_i у объекта, находящегося в центре локального сгустка своих объектов, будет больше, чем у периферийных объектов. Для объектов, оказавшихся в окружении чужих объектов, величина F_i может иметь отрицательное значение. Такой объект будет приводить к увеличению числа столпов и ухудшать качество распознавания. По этой причине эти объекты целесообразно исключить из дальнейшего рассмотрения. Процесс цензурирования состоит из последовательного исключения объектов и пересчета компактности оставшихся объектов. Сначала исключается объект, обладающий наименьшим значением величины F_i. После пересчета компактности обнаружится, что она увеличилась. Одновременно выявляется другой объект с минимальным значением F_i, который является кандидатом на очередное исключение. Если этот процесс не останавливать, то максимальная компактность, равная 1, будет достигнута, когда останутся только объекты-столпы. Цензурирование должно остановиться на шаге, при котором достигает максимума критерий Q, отражающий два противоречивых желания: добиться максимального значения компактности при минимальном сокращении количества объектов обучающей выборки:

$$Q = f(\mathcal{F}, N_c/N_0),$$

где N_c/N_0 — доля выборки, сохранившейся после цензурирования. В настоящее время исследуются некоторые варианты этого критерия.

Ниже приводится пример использования описанных алгоритмов при решении одной из реальных задач.

Диагностика рака простаты по масс-спектрам белков

Анализируются данные о масс-спектре белковых форм, полученные с помощью спектрометра типа SELDI-MS-TOF [4]. Количество признаков (спектральных полос) — 15153. Представлены 4 класса пациентов с разным уровнем индекса PSA, характеризующего степень развития рака простаты: 63 здоровых пациента класса B имеют PSA < 1 ng/mL, 26 пациентов класса C имеют PSA $= 4 \div 10$ ng/mL, 43 пациента класса D имеют PSA > 10 ng/mL и 190 пациентов класса A имеют PSA > 4 ng/mL. Малое количество пациентов не позволяет разделить выборку на обучающую и контрольную. По этой причине будем обучаться на двух классах, а на контроль предъявлять объекты третьего класса.

О качестве обучения и распознавания будем судить, исходя из следующих соображений. Если упорядочить классы пациентов по степени проявле-

Таблица 1. Результаты экспериментов.

Обучение	Контроль	B	C	D
$[B\_D]$	A_{190} (>4)	3		187
$[B\_D]$	C_{26} $(4 \div 10)$	0		26
$[B\_C]$	A_{190} (>4)	1	189	
$[B\_C]$	D_{43} (>10)	3	40	
$[C\_D]$	A_{190} (>4)		168	22
$[C\_D]$	B_{63} (<1)		49	14
$[B\_C\_D]$	A_{190} (>4)	19	137	34

ния симптомов рака от самого здорового до самого больного, то класс B должен находиться в начале списка, за ним должен следовать класс C, и затем — класс D. Пациенты класса A должны оказаться среди пациентов классов C и D. Если построить правила для распознавания класса здоровых пациентов B от любого класса больных (например, класса C), то пациенты других классов больных (A и D) должны быть больше похожими на класс C, чем на B. Перебирая разные составы конкурирующих классов и фиксируя выбираемые при этом информативные характеристики, можно выделить подмножество характеристик, по которым классы будут отличаться друг от друга.

На первом этапе были сформированы две группы классов: первую группу представлял класс здоровых пациентов B, а во вторую группу были включены три класса больных пациентов — классы A, C и D. С помощью алгоритма FRiS-GRAD [2] в режиме Cross-Validation (10 этапов по 10% выборки на контроль) из 15153 признаков в состав 10 решающих правил было включено 24 признака. По этим правилам правильно распознано 275 объектов из 322 (85,4%). Надежность распознавания здоровых пациентов была равна 43 из 63 (68,3%), а больных — 232 из 259 (89,6%).

На следующем этапе делалась попытка из 24 найденных признаков выбрать информативные подсистемы для распознавания всех классов друг от друга. Результаты решения некоторых из этих задач представлены в таблице 1, из которой видно, что класс здоровых хорошо отличается от всех классов больных пациентов. Два класса больных (классы C и D) различаются хуже.

Кроме этой задачи описанными методами были успешно решены и другие задачи из области медицины. В задаче распознавания типов лейкемии по экспрессии генов, которая решалась многими другими авторами, получены результаты, превышающие результаты этих авторов [5]. В области физики успешно решена задача распознавания классов мелкодисперсных веществ по рентгеновским спектрам [6]. Использование FRiS-функции в задачах кластеризации и таксономии позволяет строить линейно неразделимые таксоны с автоматическим выбором наилучшего числа таксонов [2]. Достаточно успешным оказалось применение FRiS-функции в алгоритме прогнозирования. В международном конкурсе Data Mining CUP 2009 [7] по решению задачи прогнозирования спроса на разные книги в разных магазинах, участвовало 52 команды из Германии, США, Великобритании, Китая и других стран. Лучший результат получил оценку 17260, худший — 1938612. Наш результат 18353 оказался на четвертом месте. Общее свойство этих задач состояло в том, что количество признаков N на порядки превышало количество объектов M.

Выводы

Рассмотрение относительной меры сходства, учитывающей конкурентную обстановку, позволяет строить эффективные алгоритмы решения всех основных задач Data Mining. Функция конкурентного сходства дает возможность вычислять количественную оценку компактности образов и информативности признакового пространства и строить легко интерпретируемые решающие правила. Метод инвариантен к количеству образов, характеру их распределений и обусловленности обучающей выборки (соотношению между M и N). Трудоемкость метода позволяет использовать его для достаточно сложных реальных задач.

Литература

[1] *Kira K., Rendell L.* The Feature Selection Problem: Traditional Methods and a New Algorithm // Proc. 10th Nat'l Conf. Artificial Intelligence (AAAI-92). — 1992. — Pp. 129–134.

[2] *Zagoruiko N. G., Borisova I. A., Dyubanov V. V., Kutnenko O. A.* Methods of Recognition Based on the Function of Rival Similarity // Pattern Recognition and Image Analysis. — 2008. — V. 18. — Pp. 1–6.

[3] *Воронцов К. В., Колосков А. О.* Профили компактности и выделение опорных объектов в метрических алгоритмах классификации // Искусственный Интеллект. — 2006. — С. 30—33.

[4] *Ziener C., Foster P. S., Divall E. J., Hooker C. J., Langley A. J., Neely D.* Time-of-Flight corroboration on «conventional» ultra high intensity measurement // Central Laser Facility Annual Report. Chilton, UK. — 2001/2002.

[5] *Zagoruiko N. G., Borisova I. A., Dyubanov V. V., Kutnenko O. A.* Attributr selection through decision rules construction (algorithm FRiS-GRAD) // Proc. of 9th Intern Conf. Pattern Recognition and Image Analysis: New Information Technologies, Nizhni Novgorod, — 2008. — V. 2. — Pp. 335–338.

[6] *Богданов А. Б., Борисова И. А., Дюбанов В. В., Загоруйко Н. Г., Кутненко О. А., Кучкин А. В., Мещеряков М. А., Миловзоров Н. Г.* Интеллектуальный анализ спектральных данных // Автометрия. — 2009. № 1. — С. 92–101.

[7] http://www.prudsys.de/Service/Downloads/bin/DMC2009_Ergebnisliste.pdf

Оптимальные байесовские стратегии анализа релевантности для объектов с заданной структурой*

Виноградов А. П., Лаптин Ю. П.
vngrccas@mail.ru, laptin_yu_p@mail.ru
Москва, Вычислительный Центр им. А. А. Дородницына РАН;
Киев, Институт кибернетики им. В. М. Глушкова

Специальная задача типа CBR (content based retrieval) исследуется в предположении, что у объектов, собранных в большом хранилище данных, имеется стандартизованная структура с простыми подобъектами на нижнем уровне, которая считается заранее известной. Показано, что модель байесовской сети в этом случае является адекватной, и на данной основе разработан эффективный метод анализа релевантности объектов и предложен быстрый алгоритм построения оптимальных стратегий вычисления соответствующих оценок на узлах структуры.

Проблема отбора информации по содержанию из больших хранилищ или баз данных в настоящее время является актуальной для многих приложений. Здесь накоплен большой арсенал методов, включая методы анализа, распознавания и принятия решений, однако проблема всё ещё трудна для существующих методов и средств, и в её рамках пока не создано универсальных подходов, обеспечивающих достаточную эффективность на большом классе типов данных. Как показывает практика, наибольшего успеха здесь достигают методы, использующие те или иные полезные особенности записей. Исследования на данном пути проводятся во всём мире, и в качестве компромиссного промежуточного решения начинает формироваться комплексный подход, в котором для получения оценок релевантности производится выбор специальных процедур из большого списка — в соответствии с доступной априорной информацией о типе данных. В этом контексте создание различных, в том числе высоко специализированных, подходов является оправданным.

Одна из подобных особенностей состоит в том, что структура объектов может быть фиксированной и достаточно точно известной заранее. Примеры использования информации такого рода можно указать во многих областях, где типичной является задача поиска ограниченного набора структурированных объектов с наиболее подходящим наполнением среди многих других аналогичных. Например, это могут быть подборки подходящих по смыслу документов стандартного вида [2, 6, 8], сходных случаев в больших медицинских коллекциях [4, 5], наилучших кандидатур по типовым объективным данным, наилучших архитектурных решений для конкретных условий, технических спецификаций для взаимозаменяемых деталей, и т. д.

Во многих работах, в том числе и в указанных выше, известный вид структуры объекта задействуется лишь в утилитарных целях, т. е., для удобства вычислений, сокращения времен доступа, и т. п., но не используется непосредственно при анализе содержания. В то же время, в повторяющихся структурах сохраняются неизменными определенные логические связи между подструктурами, и это обстоятельство можно попытаться использовать.

Введение

Ниже исследуются некоторые вопросы, возникающие при поиске наборов подходящих по содержанию объектов с заданной сложной структурой в большой выборке данных. В работе показано (насколько известно авторам, впервые), что эта особенность представления может быть использована непосредственно при анализе релевантности содержания объекта тому или иному запросу, а именно, для организации быстрого поиска неудовлетворительных оценок релевантности сверху вниз — от оценки по объекту в целом до висячих вершин. На интуитивном уровне это означает, что предлагаемый метод поиска выдает заключение либо о релевантности содержания объекта в целом, либо о детализированной причине, почему содержание данного объекта нерелевантно запросу.

На основе указанных соображений создан новый метод анализа релевантности объектов и разработан быстрый алгоритм построения оптимальных стратегий оценивания, который базируется на использовании некоторой специальной модели байесовской сети [1, 3, 7, 9, 10] и имеет сложность $|V|^3$.

Далее предполагается, что нижние структурные уровни объектов занимают простые подобъекты с ограниченным числом атрибутов, и в ходе накопления выборки на пространстве атрибутов непрерывно обновляются эмпирические априорные распределения, определяющие для простых подобъектов степени их релевантности по отношению к тому или иному типу запроса. Данное предположение является ключевым и позволяет использовать преимущества модели байесовской сети при построении оптимальной стратегии оценки релевантности узлов структуры и объекта в целом.

*Работа выполнена при финансовой поддержке РФФИ, проект № 08-01-90427.

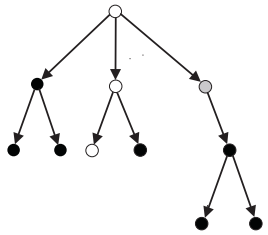


Рис. 1. Диаграмма результатов тестирования. Черные узлы релевантны ($\xi_i = 0$), белые узлы нерелевантны ($\xi_i = 1$), серый узел релевантен, но его собственное содержимое при тестировании обеспечивает пороговое значение степени релевантности ($\xi_i = 1$).

Дерево подструктур

Будем предполагать, что структуры объектов состоят из подструктур, и каждая подструктура в свою очередь может состоять из подструктур более низкого уровня. Отношения вложенности подструктур будем представлять ориентированным графом — деревом (V, E), где V — множество вершин-подструктур, E — множество дуг. Если $(i, j) \in E$, то подструктура j входит в подструктуру i. Висячим вершинам соответствуют подструктуры, состоящие из элементарных объектов, т. е. каждая такая подструктура состоит из одного объекта. Дерево (V, E) будем называть деревом подструктур, множество висячих вершин обозначим $\bar{V}$.

Если задан некоторый порог релевантности Δ, то структура (как и каждая ее подструктура) может быть в релевантном или в нерелевантном состоянии, в зависимости от того, превышен порог или нет. Поставим в соответствие вершинам $i \in V$ величины $\zeta_i \in \{0, 1\}$ так, что $\zeta_i = 0$, если подструктура i находится в релевантном состоянии, иначе $\zeta_i = 1$. Если точно установлено в результате некоторого исследования релевантности (далее используется термин «тестирование»), что подструктура $i \in V$ находится в релевантном состоянии, то и все подструктуры более низкого уровня, из которых она состоит, автоматически также находятся в релевантном состоянии. Если подструктура находится в нерелевантном состоянии, то автоматически хотя бы одна из подструктур более низкого уровня также находится в нерелевантном состоянии.

Предполагается, что для анализа релевантности и сравнения с порогом каждой подструктуре $i \in V$ соответствует некоторая заранее заданная процедура тестирования. Результатом тестирования является случайная величина $\xi_i = \zeta_i \vee \vartheta_i$, где $\vartheta_i \in \{0, 1\}$ — ошибка тестирования. Предполагается, что ϑ_i, $i \in V$, суть независимые случайные величины, для которых заданы вероятности $q_i^0 = \mathsf{P}\{\vartheta_i = 0\}$. Если $q_i^0 = 1$, то тестирование производится без ошибок. Для висячих вершин дерева подструктур оценки релевантности являются априорными, и следовательно, тестирование производится без ошибок. В процессе анализа последовательно выполняются процедуры тестирования вершин дерева подструктур (V, E). Цель тестирования — определить степень релевантности вершины дерева, т. е. объекта в целом, рис. 1.

Стратегии тестирования

Последовательность анализа релевантности подструктур, возникающую в результате применения некоторого правила выбора (в данном случае — алгоритма A, который описывается ниже) очередной вершины дерева (V, E) для тестирования на следующем шаге, будем называть стратегией тестирования R. Пусть $(i, j) \in E$. Стратегия реализуется, таким образом, как многошаговый процесс, на каждом шаге которого производится тестирование одной вершины в (V, E). Мы рассматриваем т. н. последовательные стратегии, в которых тестирование вершины j может выполняться только после тестирования вершины i, то есть если подструктура j входит в подструктуру i, то тестирование подструктуры i должно предшествовать тестированию подструктуры j (если оказалось, что подструктура i релевантна, $\xi_i = 0$, то отпадает необходимость тестировать входящие в нее подструктуры).

Множество вершин на текущем шаге, из которого выбирается вершина для тестирования, будем называть активным множеством. Пусть каждой вершине i дерева (V, E) поставлено в соответствие некоторое число (приоритет) γ_i, $i \in V$. Последовательную стратегию будем называть приоритетной, если на каждом шаге из активного множества выбирается вершина с наименьшим приоритетом.

При тестировании вершины $i \in V$ производятся затраты c_i. Стоимость анализа релевантностей есть сумма затрат на выполненные тестирования вершин. Заметим, что при этом стоимость анализа вершины может быть ниже суммарной стоимости тестирования подчинённых вершин (например, если среди атрибутов вершины есть поля, содержащие какие-либо предварительные оценки степени ее релевантности, которым можно доверять).

Пусть $\bar{\xi} = \{\bar{\xi}\}_{i \in V}$ — некоторая реализация случайных величин $\xi = \{\xi_i\}_{i \in V}$. Обозначим $U_{\bar{\xi}}(R)$ стоимость анализа релевантностей на реализации $\bar{\xi}$ при использовании стратегии R. Стоимостью стратегии R назовем математическое ожидание $U(R) = \mathsf{E}U_{\bar{\xi}}(R) = \sum U_{\bar{\xi}}(R) P_{\bar{\xi}}$,где суммирование производится по всем двоичным реализациям $\bar{\xi}$ на структуре дерева (V, E), вероятности в узлах которой определяются зависимостями в байесовской сети.

Анализ релевантности проводится при следующем дополнительном условии: нерелевантных висячих вершин достаточно обнаружить не более одной. То есть, процесс анализа либо завершается

сразу после обнаружения нерелевантной висячей вершины, либо продолжается, пока активное множество не станет пустым. Во втором случае устанавливается, что структура в целом релевантна.

Алгоритм построения оптимальной байесовской стратегии

Таким образом, если задан некоторый порог Δ и построены соответствующие оценки релевантности для простых объектов, то на этой основе можно рассчитать все производные распределения вероятностей для байесовской сети, фигурирующие в (1), и выбрать с их помощью оптимальную стратегию вычисления релевантности любого из объектов выборки. В результате получаем на выборке ранжирующую оценку степени релевантности объектов, которая всякий раз вычисляется с использованием этой оптимальной стратегии.

Для данной задачи разработан алгоритм А построения оптимальной (с минимальной стоимостью) приоритетной стратегии.

Алгоритм имеет рекуррентную форму. Определим поддерево V_i потомков узла i, и подмножество S_i прямых потомков этого узла. Пусть оптимальные приоритеты (оптимальные приоритетные стратегии) построены для поддеревьев $V_j, j \in S_i$. Оптимальные приоритеты для поддерева V_i вычисляются на основе значений оптимальных приоритетов для поддеревьев $V_j, j \in S_i$. Соотношения для рекуррентного пересчета являются результатом анализа условий оптимальности. Соответствующие вычисления включают объединение упорядоченных наборов вершин для поддеревьев $V_j, j \in S_i$ и переупорядочения результирующего набора со сложностью $|V|^2$, откуда следует оценка для сложности алгоритма в целом $|V|^3$.

Выводы

В работе изучались некоторые возможности, связанные с наличием фиксированной структуры представления у объектов в большом хранилище данных, которые могут быть использованы для организации эффективных на данной выборке процедур анализа релевантости объектов по содержанию. Показано, что для структур объектов в виде древовидного графа адекватной моделью логических связей между подобъектами служит байесовская сеть специального вида с бинарными переменными. Установлено, что при использовании данной модели всегда имеется простой алгоритм со сложностью $|V|^3$, который позволяет эффективно строить оптимальную байесовскую стратегию анализа релевантности объектов по их содержанию, представленному каждый раз в узлах структуры в распределенном виде. Основной результат состоит в создании эффективного приема задействования устойчивых структурных связей для организации оптимального по затратам процесса анализа релевантности объектов на большой выборке.

Предложенный подход может оказаться полезным при выработке рациональных эвристик для многих задач, где используются байесовские оценки аналогичного вида. В частности, представленная в работе схема анализа релевантности может применяться как прогрессивная, если на очередном шаге в качестве порогового используется экстремальное на выборке значение степени релевантности Δ^*, обнаруженное на предыдущих шагах. Наличие быстрого алгоритма позволяет в этом случае эффективно строить новую оптимальную стратегию, соответствующую порогу Δ^*.

Подход может быть адаптирован и на другие задачи поиска, оценивания и тестирования, где также применимы модели байесовских сетей с бинарными переменными.

Литература

[1] *Agosta J. M., Gardos T.* Bayesian network “smart” diagnostics // Intel Technology Journal — 2004. — Vol. 8, № 4. — С. 361–372.

[2] *Ahlgren P., Colliander C.* Document-document similarity approaches and science mapping: experimental comparison of five approaches // Journal of Informetrics — 2009. — Vol. 3, № 1. — С. 49–63.

[3] *Ahlswede R., Wegener I.* Search Problems. New York: John Wiley & Sons, 1987.

[4] *Del Fiol G., Haug P. J.* Classification models for the prediction of clinicians’ information needs // Journal of Biomedical Informatics — 2009. — Vol. 42, № 1. — С. 380–392.

[5] *Hliaoutakis A., Zervanou K., Petrakis E. G. M.* The AMTEx approach in the medical document indexing and retrieval application // Data & Knowledge Engineering — 2009. — Vol. 68, № 3. — С. 380–392.

[6] *Lin S.-S.* A document classification and retrieval system for R&D in semiconductor industry — A hybrid approach // Expert Systems with Applications — 2009. — Vol. 36, № 3, Part 1. — С. 4753–4764.

[7] *Лаптин Ю. П.* Одна экстремальная задача на случайных деревьях // Кибернетика — 1981. — Т. 17, № 2. — С. 262–265.

[8] *Mendez J. R., Glez-Pena D., Fdez-Riverola F., Diaz F., Corchado J. M.* Managing irrelevant knowledge in CBR models for unsolicited e-mail classification // Expert Systems with Applications — 2009. — Vol. 36, № 2. — С. 1601–1614.

[9] *Turhan B., Bener A.* Analysis of Naive Bayes’ assumptions on software fault data: An empirical study // Data & Knowledge Engineering — 2009. — Vol. 68, № 2. — С. 278–290.

[10] *Wegener I.* The Discrete Sequential Search Problem with Nonrandom Cost and Overlook Probabilities // Mathematics Of Operations Research — 1980. — Vol. 5, № 3. — С. 373–380.

Применение генетических алгоритмов в задаче классификации сигналов (приложение в BCI)*

Власова Ю. В.
julia.vlasova@bk.ru
Москва, МГУ имени М. В. Ломоносова, факультет ВМиК

В работе предложен метод построения признакового пространства в задаче классификации сигналов, основанный на генетической оптимизации. Основной целью работы является разработка, программная реализация и экспериментальное исследование данного метода.

Один из возможных подходов к решению задачи классификации сигналов состоит в синтезе информативного признакового описания сигналов и сведении проблемы к задаче классификации в признаковом пространстве.

В работах [1, 2, 3] признаки синтезируются по временным и спектральным характеристикам. Однако использование подобных методов не всегда гарантирует высокое качество классификации.

В данной работе предлагается воспользоваться аппаратом генетической оптимизации [4, 5] для построения информативного признакового описания сигнала на основе выделения из него наиболее значимой информации [6].

Постановка задачи

Будем предполагать, что длины всех сигналов совпадают и равны T.

Имеется множество объектов $\mathfrak{X}$ вида $\boldsymbol{x} = (x_1, \ldots, x_T) \in \mathbb{R}^T$ и двухэлементное множество имён классов $Y = \{-1, +1\}$. Множество $\mathfrak{X}$ разбито на две выборки — обучающую и контрольную. Существует целевая зависимость $y^*\colon \mathfrak{X} \to Y$, значения которой известны только на объектах обучающей выборки $X^l = \left(\boldsymbol{x}^i, y^i\right)_{i=1}^{l}$, $\boldsymbol{x}^i = (x_1^i, \ldots, x_T^i) \in \mathfrak{X} \subset \mathbb{R}^T$, $y^i = y^*\left(\boldsymbol{x}^i\right)$, l — количество объектов обучения. Требуется построить алгоритм классификации $a\colon \mathfrak{X} \to Y$, аппроксимирующий целевую зависимость $y^*(\boldsymbol{x})$ на всём множестве $\mathfrak{X}$.

Представление признака

Задача сводится к поиску информативного признакового пространства и выполнению классификации в нем. В работе основное внимание уделяется методу построению признакового пространства.

В контексте рассматриваемой задачи *признаком* называется вещественная функция от сигнала: $f(\boldsymbol{x}) = f(x_1, \ldots, x_T) \in \mathbb{R}$. Основная идея предлагаемого метода заключается в представлении данной функции в виде суперпозиции операторов обработки сигнала. Признак определяется набором составляющих его операторов. Поиск оптимального набора операторов осуществляется посредством генетического алгоритма.

*Работа выполнена при финансовой поддержке РФФИ, проект № 08-07-00305-а.

Введем два типа операторов обработки сигналов.

Операторы первого типа по описанию сигнала $\boldsymbol{x} = (x_1, \ldots, x_s)$ получают сигнал $\boldsymbol{y} = (y_1, \ldots, y_p)$, возможно, другой длины. Основное их назначение — преобразовать информацию (улучшить или перейти к иному описанию). Заметим, операторы первого типа должны быть достаточно простыми. Обозначим через $\mathfrak{B}$ множество операторов первого типа

$$B_i \in \mathfrak{B}\colon\ \boldsymbol{x} = (x_1, \ldots, x_s) \ \to\ \boldsymbol{y} = (y_1, \ldots, y_p).$$

Примеры операторов первого типа:
B_1 — фильтрация сигнала, устранение выбросов;
B_2 — сглаживание сигнала;
B_3 — исключение тренда;
B_4 — взятие производной (конечной разности);
B_5 — оператор взятия абсолютного значения;
B_6 — покомпонентное логарифмирование ряда;
B_7 — покомпонентное возведение в степень ряда;
B_8 — сортировка значений ряда;
B_9 — «стандартизация» значений ряда;
B_{10} — выделение фрагмента ряда;

Операторы второго типа преобразуют сигнал в число, вычисляя его некоторую характеристику. Через $\mathfrak{C}$ обозначим множество операторов второго типа

$$C_j \in \mathfrak{C} : \boldsymbol{y} = (y_1, \ldots, y_p) \to z.$$

Примеры операторов второго типа:
C_1 — среднее значение сигнала;
C_2 — максимальное (минимальное) значение сигнала;
C_3 — сумма максимального и минимального значений сигнала;
C_4 — стандартное отклонение;
C_5 — число значений сигнала определенного типа;
C_6 — число различных значений сигнала;
C_7 — «центр масс» для сигнала $\boldsymbol{x} \cdot (1, \ldots, s)^{\mathsf{T}}/s$.

Признак f, описывающий сигнал, представим в виде суперпозиции операторов:

$$f(\boldsymbol{x}; i_1, \ldots, i_k, i_0) = C_{i_0}\big(B_{i_k}(\ldots(B_{i_1}(\boldsymbol{x})))\big), \qquad (1)$$

где $k \in \{1, 2, \ldots\}$, $B_{i_k}, \ldots, B_{i_1} \in \mathfrak{B}$, $C_{i_0} \in \mathfrak{C}$.

При этом мы разрешаем, чтобы признаки содержали различное число k операторов в представлении (1).

Процедура построения информативного признакового пространства

Будем решать задачу построения информативного признакового пространства (совокупность признаков вида (1)), используя генетический алгоритм [4, 5]. Было рассмотрено два случая — построение одномерного и двумерного признаковых пространств.

Ниже будут использоваться термины из теории генетических алгоритмов.

Кодирование признакового пространства. Признак вида (1) кодируется индексами операторов, которые его определяют:

$$f(\boldsymbol{x}; i_1, \ldots, i_k, i_0) \leftrightarrow [i_1] \ldots [i_k][i_0]. \qquad (2)$$

Введем обозначение для конечного множества признаков вида (1):

$$\mathfrak{F} = \big\{[i_1] \ldots [i_k][i_0] \colon k \in \mathbb{N}, B_{i_k}, \ldots, B_{i_1} \in \mathfrak{B}, C_{i_0} \in \mathfrak{C}\big\}.$$

Популяцией назовем некоторое подмножество $F = \{f_1, \ldots, f_S\}$ множества $\mathfrak{F}$. Элементы данного множества будем называть особями популяции.

В случае построения двумерного признакового пространства популяцией будет называться S-элементое подмножество декартова произведения $\mathfrak{F} \times \mathfrak{F}$, обозначим его также через F. Элементы множества $F \subset \mathfrak{F}$ называются особями данной популяции F.

Функционал информативности признакового пространства. Информативность отдельного признака оценивается критерием AUC-ROC [7], а совокупности признаков — методом *скользящего контроля с исключением объектов по одному* (leave-one-out, LOO) по алгоритму k *ближайших соседей* (k nearest neighbors, kNN). При использовании алгоритма k ближайших соседей осуществляется подбор оптимальной взвешенной евклидовой метрики, то есть в качестве функции расстояния между двумя объектами $u = (u^1, u^2)$ и $v = (v^1, v^2)$ в двумерном признаковом пространстве используется метрика:

$$\rho(u, v) = \sqrt{(u^1 - v^1)^2 + \mathrm{coeff}^2 \cdot (u^2 - v^2)^2}, \qquad (3)$$

где coeff — настраиваемый параметр. Смысл заключается в растяжении либо сжатии одного из признаков для повышения качества классификации. При этом происходит поиск пары признаков вида $(f_1, \mathrm{coeff} \cdot f_2)$.

Обозначим функционал информативности особи через $Q(\cdot)$.

Алгоритм построения признакового пространства. Алгоритм работает пошагово. На каждом шаге i алгоритма популяция F^i состоит из одного и того же заданного числа особей S. Это параметр алгоритма.

Начальная популяция F^1 формируется случайным образом. Шаг алгоритма состоит из трех стадий: отбор особей из предыдущего поколения, скрещивание особей с помощью оператора кроссовера, мутация особей. Выполнение трех этапов приводит к формированию нового поколения. Количество шагов задается экспертом.

Для создания нового поколения используется принцип *элитизма* [4, 5]. В новое поколение обязательно включаются две самые лучшие (с максимальным значением функционала $Q(\cdot)$) особи предыдущего поколения, которые не участвуют в размножении, и к которым не применяется оператор мутации.

В качестве стратегии отбора родительских хромосом выбран *пропорциональный отбор* [4, 5]. Вероятность каждой особи попасть в промежуточную популяцию пропорциональна ее приспособленности $Q(\cdot)$. В англоязычной литературе данный тип отбора известен под названием *stochastic sampling*.

После отбора особи промежуточной популяции случайным образом разбиваются на пары. Каждая из них с определенной вероятностью скрещивается, в результате чего образуется два потомка. Они заносятся в новое поколение. Если же паре не выпало скрещиваться, то сами особи этой пары остаются в популяции. Опишем оператор скрещивания двух особей — $[i_1] \ldots [i_k][i_0]$ и $[j_1] \ldots [j_l][j_0]$. Без ограничения общности можно считать, что $l \leqslant k$. Преобразуем код второго признака в эквивалентный ему вид $[j_1] \ldots [j_k][j_0]$, где $j_{l+1} = \cdots = j_k = 0$. Здесь 0 — индекс, обозначающий отсутствие гена. Использовались два различных оператора скрещивания:

1. *Одноточечный оператор кроссовера.* Для родительских хромосом случайным образом выбирается точка раздела $z \in \{1, \ldots, k\}$, и они обмениваются отсеченными частями. В результате получаются два потомка:

$$[i_1] \ldots [i_z][j_{z+1}] \ldots [j_k][j_0],$$
$$[j_1] \ldots [j_z][i_{z+1}] \ldots [i_k][i_0].$$

2. *Однородный кроссовер.* В данном случае каждый ген хромосом родителей с определенной вероятностью $p = 0.5$ переходит к тому или иному потомку. Потомки имеют вид:

$$[u_1] \ldots [u_k][u_0],$$
$$[v_1] \ldots [v_k][v_0],$$

где с вероятностью 0.5 либо $u_z = i_z$ и $v_z = j_z$, либо $u_z = j_z$ и $v_z = i_z$.

Из кода потомков удаляются индексы 0.

Для всякой особи применяется оператор мутации. Каждый ген с некоторой вероятностью заменяется на произвольный новый ген. В случае поиска оптимальной пары признаков мутация может быть нескольких типов: обмен признаков в особи местами, либо замена (с определенной вероятностью) одного из операторов в составе любого признака данной особи на произвольный.

Далее выбираем особь популяции с наибольшим значением функционала $Q(\cdot)$.

Классификатор. После того, как оптимальный признак (оптимальная пара признаков) найден, выполняем классификацию объектов в построенном признаковом пространстве. В качестве классификатора используем метод k ближайших соседей со взвешенной евклидовой метрикой (3).

Вычислительный эксперимент

Предложенный метод был протестирован при решении реальных задач классификации сигналов из области BCI (Brain-Computer Interface) [8, 9, 10]. В этой области занимаются проблемой коммуникации человек–компьютер с помощью сигналов головного мозга. Предполагается, что настроив классифицирующий алгоритм на обучающей выборке (сигналы, снятые во время известных ментальных действий), ошибка на контроле будет удовлетворительной. Приложения BCI: контроль протезов и управление механизмами, организация общения с людьми с ограниченными физическими возможностями, компьютерные игры, исследование активности головного мозга человека в научных целях.

Данные эксперимента взяты с проводившихся соревнований BCI Competition III [10]. Испытуемому предлагалось выполнять два различных ментальных действия. Электрическая активность головного мозга фиксировалась с помощью ECoG-платиновой сетки из 64 электродов размера 8×8 см. Сигналы записывались в течение 3 секунд с частотой 1000 МГц. Обучающую выборку составили 278 объектов, которые описывались с помощью 64 временных рядов по 3000 точек в каждом. Контрольная выборка состояла из 100 объектов. Для проведения исследований было выбрано 7 электродов, которые наиболее хорошо помогали решать задачу.

Характерной особенностью данных соревнований BCI Competition III является то, что обучающая и контрольная выборки были сформированы в различные дни (с интервалом в 1 неделю). Различная степень усталости испытуемого, различие в положении электродов привели к тому, что обучение и контроль в построенном признаковом пространстве оказались «разнесены» (похожие по форме кластеры обучения и контроля расположены на некотором расстоянии друг относительно друга). Данная проблема решается путем «совмещения» множеств обучения и контроля [11].

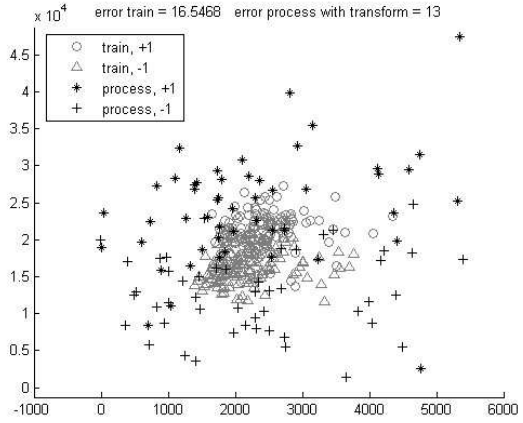


Рис. 1. Пример работы генетического подхода на данных задачи BCI. Построено признаковое пространство: $f_1 = [B_6][B_4][C_2],\ f_2 = [B_4][B_5][C_1]$.

Результаты эксперимента. В качестве примера работы изложенного подхода можно привести пару признаков $f_1 = [B_6][B_4][C_2],\ f_2 = [B_4][B_5][C_1]$ (см. стр. 96). Ошибка классификации на контроле составляет 13%. Данная пара признаков визуализирована на рис. 1.

Для анализа предложенного подхода к решению задачи классификации сигналов было проведено следующее исследование. Все признаки, содержащие в своем представлении (1) не более четырех операторов первого типа, были упорядочены по убыванию показателя AUC. Результат представлен на рис. 2. На диаграмме серым цветом показаны соответствующие показатели AUC для контроля. Анализируя поведение графиков, можно сделать несколько выводов:

— Виден эффект переобучения — серый график, соответствующий значениям показателя AUC на контроле, лежит ниже черной жирной кривой. Это означает, что в среднем ошибка на контроле больше ошибки на обучении;
— Заметна широкая полоса хороших признаков, для которых показатель AUC на обучении высокий, а переобучение минимально. С большой степенью вероятности мы попадем в данную область. Поэтому генетический алгоритм довольно быстро сходится;
— Графики совпадают при значениях показателя AUC = 0.5, что соответствует «константам» — «бесполезным» классификаторам.

Выводы

В работе предложен метод генерации признаков в задаче классификации сигналов, основанный на

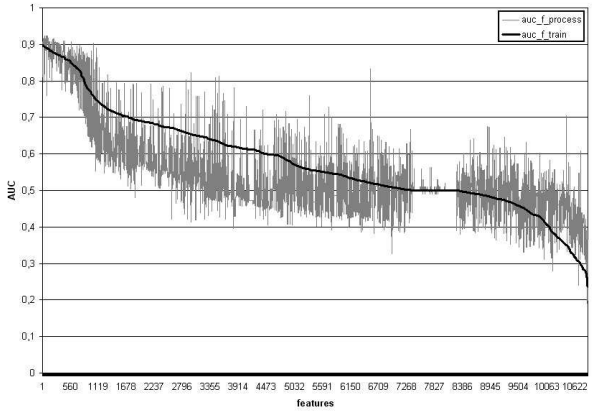


Рис. 2. Показатель AUC на обучении и контроле.

генетической оптимизации. Выполнена программная реализация предложенного метода в виде библиотеки для системы MATLAB. Проведены эксперименты на реальных данных задачи медицинской диагностики Brain-Computer Interface [8, 9, 10].

Литература

[1] *Muller K.-R., Anderson C. W., Birch G. E.* Linear and non-linear methods for brain-computer interfaces // IEEE Trans. Neural Sys. Rehab. Eng., 2003, 11(2) — P. 165–169.

[2] *Muller K.-R., Krauledat M., Dornhege G., etc.* Machine learning techniques for brain-computer interfaces // Biomed. Tech., 2004, 49(1) — P. 11–22.

[3] *Hoffmann U., Vesin J.-W., Ebrahimi T.* Recent Advances in Brain-Computer Interfaces // IEEE International Workshop on Multimedia Signal Processing, 2007.

[4] *Koza J. R.* Genetic Programming IV: Routine Human-Competitive Machine Intelligence // Springer. 2005. — pp590.

[5] *Langdon W. B., Poli. R.* Foundations of genetic programming // Springer. 2002. — pp260.

[6] *Дьяконов А. Г.* Об одном подходе к решению задач из области BCI // Сборник докладов XII Всеросийской конференции ММРО–12, Москва: ООО МАКС Пресс, 2005. — С. 95–97.

[7] `http://www.basegroup.ru/regression/logistic.htm` — Логистическая регрессия и ROC-анализ — математический аппарат.

[8] *Blankertz B., Muller K.-R., Curio G., etc.* The BCI competition 2003: Progress and perspectives in detection and discrimination of EEG single trials // IEEE Trans. Biomed. Eng., 2004, 51(6) — P. 1044–1051.

[9] *Lal T., Hinterberger T., Widman G., etc.* Methods Towards Invasive Human Brain-Computer Interfaces // Advances in Neural Information Processing Systems (NIPS), 2004. — P. 737–744.

[10] `http://ida.first.fraunhofer.de/projects/bci/competition_iii/index.html` — BCI Competition III — 2003.

[11] *Дьяконов А. Г.* Анализ кластерных конфигураций в одной проблеме фильтрации спама // Всеросс. конф. ММРО-13: сборник докладов, Москва: МАКС Пресс, 2007. — С. 476–478.

[12] *Власова Ю. В.* Применение генетических алгоритмов в задаче классификации сигналов (приложение в BCI) // Сборник тезисов XVI международной научной конференции Ломоносов — 2009, Москва: МАКС Пресс, 2009. — С. 17.

Метод построения взвешенных обучающих выборок в открытых системах распознавания

Волченко Е. В.
Lm@top.finfort.com
Донецк, Украина, Государственный университет информатики и искусственного интеллекта

Рассматривается задача сокращения размера обучающих выборок в открытых системах распознавания. Предлагается метод построения взвешенной выборки w-объектов, описывающихся набором своих признаков и дополнительным параметром — весом, что позволяет учитывать взаимное расположение объектов в пространстве признаков. Выполняется оценка временной алгоритма, приводятся результаты экспериментальных исследований.

Одним из наиболее важных требований, предъявляемых к современным открытым системам распознавания, является их адаптивность (on-line adaptation) [1], т. е. способность изменять свойства системы распознавания в соответствии с изменениями распознаваемых объектов на всем протяжении времени работы системы. В большинстве систем адаптивность реализуется за счет пополнения обучающей выборки новыми обучающими объектами.

Добавление новых объектов в обучающую выборку и, как следствие, корректировка решающих правил классификации, позволяют системе сохранять начальную эффективность классификации. Анализ ряда современных прикладных задач распознавания приводит к системам открытого типа, поэтому исследование вопросов обработки обучающих выборок и построения решающих правил в открытых системах распознавания является актуальным. Примерами таких задач являются: создание электронных библиотек, новостных серверов, модулей «спам»-фильтрации в системах обработки электронной корреспонденции и др. [2, 3].

В большинстве прикладных задач, требующих построения открытых систем распознавания, рост размера обучающих выборок является неограниченным [4] (так в задаче построения «спам»-фильтров объектами является электронная корреспонденция, а дообучение проводится при получении каждого нового письма), что делает особенно важным решение ряда следующих задач:

— корректировка решающих правил классификации при добавлении новых объектов в выборку, что требует постоянного хранения всей обучающей выборки и, соответственно, может существенно увеличивать время классификации;
— сокращение размера обучающей выборки при каждом новом добавлении объектов для уменьшения времени построения решающих правил и выполнения классификации;
— анализ новых обучающих объектов с целью определения целесообразности их добавления в обучающую выборку в зависимости от значений их признаков и уже имеющихся объектов в обучающей выборке.

Задача обработки обучающих выборок и сокращения их размера является одним из трех этапов создания обучающихся систем распознавания [1]. Наиболее известными методами сокращения выборок являются алгоритмы STOLP и ДРЭТ [5], алгоритмы NNDE (Nearest Neighbor Density Estimate) и MDCA (Multiscale Data Condensation Algorithm) [1]. Основаны эти методы на выборе некоторого подмножества «опорных» точек исходной обучающей выборки, которое наиболее полно описывает всю исходную обучающую выборку и замене исходной выборки этим подмножеством. Такой подход, на наш взгляд, является достаточно эффективным, поскольку позволяет существенно сократить объем выборки, однако сохранение информации только об «опорной» точке некоторой области признакового пространства не позволяет в полной мере сохранить информацию обо всех объектах исходной обучающей выборки. Анализ распределения «опорных» точек в многомерном пространстве признаков показывает, что рассматриваемые алгоритмы выполняют сглаживание закона распределения. Количество «опорных» точек в некоторой рассматриваемой области признакового пространства, найденных этими методами, зависит только от их близости к объектам других классов и не зависит от количества объектов «своего» класса в этой области. Кроме того, эти алгоритмы предназначены для обработки фиксированных обучающих выборок и не ориентированы на случай добавления новых объектов в процессе работы системы.

Для учета взаимного расположения объектов обучающей выборки предложено семейство алгоритмов ReliefF [6]. Расстояние между объектами «своего» и «чужого» классов отражается в виде дополнительного параметра — веса, который в дальнейшем используется для классификации. Алгоритм строит взвешенную выборку, равную по размеру исходной обучающей выборке, а для расчёта веса используется только по одному объекту «своего» и «чужого» классов.

В данной работе предлагается новый метод построения сокращенной взвешенной обучающей выборки, в котором определённые подмножества

объектов заменяются одним новым объектом, при этом учитывается взаимное расположение объектов в пространстве признаков. Значения признаков каждого объекта новой выборки являются центрами масс значений признаков объектов исходной выборки, которые он заменяет. Введенный дополнительный параметр — вес определяется как количество объектов исходной выборки, которые были заменены одним объектом новой выборки. Предлагаемый метод ориентирован как на сокращение исходной обучающей выборки, так и на анализ необходимости корректировки выборки и быстрое выполнение такой корректировки при пополнении выборки в процессе работы системы.

Для описания метода построения взвешенных обучающих выборок далее введем ряд понятий и обозначений.

Основные определения и постановка задачи сокращения выборки

Пусть дано некоторое множество объектов M, представленное в виде объединения непересекающихся классов $M = \bigcup_{i=1}^{l} V_i$. Каждый объект W_i из M описывается системой признаков, т. е. $W_i = \{w_{i1}, w_{i2}, \ldots, w_{in}\}$, и представляется точкой в линейном пространстве признаков, т. е. $W_i \in \mathbb{R}^n$.

Пусть имеется конечный набор объектов $W = \{W_1, \ldots, W_s\}$, $W \subset M$, называемый обучающей выборкой, о каждом из которых известно, к какому классу он принадлежит [7].

Пусть на объектах исходной обучающей выборки W задана некоторая метрика $R(W_i, W_j)$.

Пусть имеется некоторый алгоритм $A(W)$ преобразования исходной обучающей выборки W во взвешенную обучающую выборку MW.

Взвешенная обучающая выборка MW строится следующим образом. Вначале каждый класс объектов исходной обучающей выборки представляется с помощью алгоритма $A(W)$ в виде покрытия блоками W_f. Эти блоки в дальнейшем будем называть *образующими множествами*.

Далее каждое образующее множество W_f с помощью алгоритма $A(W)$ заменяется одним вектором $MW_f = \{w_{f1}, \ldots, w_{fn}, p_f\}$. Значения признаков $\{w_{f1}, \ldots, w_{fn}\}$ рассчитываются по значениям признаков объектов образующего множества W_f, а p_f рассчитывается как его мощность. Получаемый объект MW_f назовем w-объектом. Элемент p_f назовем весом w-объекта.

Построенную выборку w-объектов назовем взвешенной обучающей выборкой MW.

Очевидно, что построение объектов взвешенной обучающей существенно зависит от выбранной метрики R и алгоритма $A(W)$. Конкретный вид метрики и особенности реализации алгоритма будут рассмотрены далее.

Используя введенные определения и обозначения, задачу сокращения размера обучающей выборки сформулируем следующим образом.

Необходимо построить новую взвешенную обучающую выборку w-объектов MW, отвечающую следующим критериям:

1) размер выборки w-объектов должен быть меньше размера исходной обучающей выборки, т. е. $ms = |MW| < s = |W|$;
2) каждый w-объект MW_f должен относиться к тому же классу, что и объекты образующего множества W_{fi}, по которым он сформирован; ;
3) качество $N(Z, U(W', MW))$ взвешенного решающего правила, оцениваемого по контрольной выборке Z, не должно ухудшаться по сравнению с решающим правилом, построенным по исходной выборке, т. е. $N(Z, U(W', MW)) \geqslant N(Z, U(W', W))$ (при этом в общем случае под качеством классификации будем понимать количество неверно классифицированных объектов контрольной выборки Z оцениваемыми решающими правилами $U(W', MW)$ и $U(W', MW)$ соответственно).

Построение выборки w-объектов

В данном разделе для решения задачи построения выборки w-объектов опишем особенности алгоритма $A(W)$ и используемую метрику R. При описании алгоритмов без потери общности получаемых решений применим стандартный для теории распознавания подход, заключающийся в рассмотрении двухклассовых систем.

Построение каждого w-объекта состоит из трех последовательных этапов:

1) построение образующего множества W_f:
2) формирование вектора $\{w_{f1}, \ldots, w_{fn}\}$ значений признаков w-объекта и расчет веса p_f;
3) корректировка исходной обучающей выборки — удаление объектов, включенных в образующее множество $W = W \setminus W_f$).

Построение образующего множества W_f состоит в нахождении начальной точки W_{f1} формирования w-объекта, определении конкурирующей точки W_{f2} и выборе в образующее множество таких объектов исходной выборки, расстояние до каждого из которых от начальной точки меньше, чем расстояние от них до конкурирующей точки.

В данной работе на основании исследования, проведенного в [8], в качестве начальной точки W_{f1} формирования w-объекта предлагается использовать объект исходной обучающей выборки, наиболее удаленный от всех объектов другого класса.

Выбор конкурирующей точки W_{f2} также основан на исследовании, проведенном в [8], и его пред-

лагается выбирать путем нахождения ближайшего к W_{f1} объекта, не принадлежащего тому же классу, что и сам W_{f1}.

По аналогии с одним из наиболее известных методов классификации — методом k ближайших соседей — в данной работе предлагается два алгоритма построения образующих множеств:

1) алгоритм с одной конкурирующей точкой;
2) алгоритм с k конкурирующими точками.

Использование k конкурирующих точек, на наш взгляд, может быть более эффективно по сравнению с использованием одной конкурирующей точки для обучающих выборок, объекты которых расположены близко друг другу в признаковом пространстве.

Выбор объектов $\{W_{f3}, \ldots, W_{fd}\}$ образующего множества осуществляется по следующим правилам: объект W_i включается в W_f,если:

1) он принадлежит тому же классу, что и начальная точка W_{f1};
2) расстояние от рассматриваемого объекта до начальной точки W_{f1} меньше, чем до конкурирующей точки W_{f2};
3) расстояние от рассматриваемого объекта до конкурирующей и начальной точек меньше, чем расстояние между ними.

Таким образом, образующее множество формируется по правилу:

$$W_f = W_{f1} \cup W_{f2} \cup \left\{W_i \;\middle|\; R_{i,1} < R_{i,2} < R_{1,2}\right\};$$

где $f1 = \arg\min\limits_{j=1,\ldots,d} \sum\limits_{j=1}^{d} R(W_i, W_j)$;

$$f2 = \arg\max_{\substack{j=1,\ldots,d \\ W_i \notin V_k:\ W_{f1} \in V_k}} R(W_i, W_{f1});$$

$$R^2_{a,b} = R^2(W_a, W_b) = \sum_{l=1}^{n} (w_{al} - w_{bl})^2.$$

Значения признаков $\{w_{f1}, \ldots, w_{fn}\}$ нового w-объекта MW_f формируются по образующему множеству W_f и рассчитываются как координаты центра масс системы из $p_f = |W_f|$ материальных точек (примем, что объекты исходной обучающей выборки, являющиеся в признаковом пространстве материальными точками, имеют массу, равную 1):

$$w_{fl} = \frac{1}{p_f} \sum_{i:\ W_i \in W_f} w_{il}, \quad l = 1, \ldots, n.$$

После формирования очередного w-объекта, все объекты образующего его множества удаляются из исходной обучающей выборки, т. е. $W = W \setminus W_f$. Непосредственно из данного правила вытекает следующая теорема о сходимости предлагаемого алгоритма.

Теорема 1. *Предлагаемый алгоритм $A(W_f)$ построения взвешенной обучающей выборки w-объектов требует выполнения не более чем s шагов.*

На основании анализа предложенного алгоритма также можно сформулировать следующие утверждения.

Утверждение 2. *Каждый w-объект содержит от 1 до $|V_i|$, $i = 1, \ldots, m$ объектов исходной обучающей выборки.*

Утверждение 3. *Никакие два и более w-объектов не могут содержать один и тот же объект исходной выборки.*

Утверждение 4. *Каждый w-объект принадлежит тому же классу, что и все объекты образующего его множества.*

Ниже приводится псевдокод алгоритма построения взвешенной выборки w-объектов с одной конкурирующей точкой.

Отличием алгоритма построения выборки w-объектов с k конкурирующими точками является выбор k ближайших к начальной точке объектов и добавление в образующее множество рассматриваемого объекта, если расстояние от него до начальной точки меньше расстояния до каждого из k конкурирующих объектов.

Оценка сложности метода

Оценим временную сложность алгоритма построения выборки w-объектов с одной конкурирующей точкой, если исходная обучающая выборка состоит из n объектов.

Для нахождения расстояний между всем парами объектов выборки и суммарного расстояния от каждого объекта до всех объектов другого класса требуется $O(n^2)$ шагов. Для формирования образующего множества и расчета значений признаков w-объекта требуется $O(n)$ шагов. Таким образом, временная сложность алгоритма построения выборки w-объектов с одной конкурирующей точкой составляет $O(n^2)$ шагов.

Аналогично можно показать, что построение выборки w-объектов с k конкурирующими точками требует также $O(n^2)$ шагов.

Утверждение 5. *Временная сложность метода построения выборки w-объектов составляет $O(n^2)$ шагов.*

Пополнение взвешенной обучающей выборки w-объектов

Для решения задачи пополнения обучающей выборки w-объектов в данной работе предлагается следующая схема.

Алгоритм 1. Построение выборки w-объектов с одной конкурирующей точкой.

Вход: $W = \{W_1, \ldots, W_s\}$, $V = V_1, \ldots, V_s$;
Выход: $MW = \{MW_1, \ldots, MW_{sm}\}$;

для $i = 1, \ldots, s$
  для $j = 1, \ldots, s$
    // расстояние между всеми объектами
    $r[i,j] := \sqrt{\sum_{l=1}^{n}(w[i,l] - w[j,l])^2}$;
  // пока исходная выборка не пуста
$ms := 1$;
пока $W \neq \varnothing$
  // найти начальную точку
  для $i = 1, \ldots, s$
    $vv := V[i]$;
    для $j = 1, \ldots, s$
      если $V[j] = vv$ **то**
        $Rs[i] := Rs[i] + r[i,j]$;
  $f1 := \arg \max\limits_{i=1,\ldots,s} Rs[i]$;
  // найти конкурирующую точку
  $f2 := \arg \max\limits_{\substack{i=1,\ldots,d \\ V[i] \neq V[f1]}} R[f_1, i]$;
  // сформировать образующее множество
  $Wf[1] := W_{f1}$; $Wf[2] := W_{f2}$;
  для $i = 1, \ldots, s$
    если $R[i,f_1] < R[i,f_2] < R[f_1,f_2]$ **то**
      $Wf[d] := W[i]$; $d := d + 1$;
  // рассчитать значения признаков
  для $i = 1, \ldots, n$
    для $j = 1, \ldots, p$
      $MW[ms,i] := MW[ms,i] + W[j,i]$;
    $MW[ms,i] := \frac{1}{p} MW[ms,i]$;
  $MW[ms, n+1] := p$;
  $W := W \setminus W_f$; $ms := ms + 1$;

1. Определяется необходимость добавления объекта W_t как нового w-объекта взвешенной обучающей выборки, для этого выполняется его классификация.
2. Если объект W_t классифицирован верно, то его добавление в общем случае не внесет существенных изменений в решающее правило, поэтому создание нового w-объекта не выполняется. Для сохранения информации о W_t в текущей взвешенной обучающей выборке выполняется поиск ближайшего к нему w-объекта того же класса,что и W_t, и вес этого w-объекта увеличивается на единицу;
3. если объект W_t классифицирован неверно, то его добавление в обучающую выборку позволит скорректировать решающее правило классификации. Во взвешенную обучающую выборку добавляется новый w-объект MW_{ms+1}, значения признаков которого равны значениям признаков объекта W_t, и вес нового w-объекта равен единице.

Выводы

В работе предложен новый метод решения актуальной задачи сокращения обучающих выборок в открытых обучающихся системах распознавания. Основной особенностью предложенного метода является переход к взвешенным обучающим выборкам, что позволяет учитывать количество заменяемых w-объектом объектов исходной обучающей выборки.

Для алгоритмов, реализующих предложенный метод, доказана сходимость и оценена временная сложность, которая является меньше временной сложности известных алгоритмов сокращения обучающих выборок, что позволяет существенно сократить время выполнения классификации и особенно важно для открытых систем распознавания. В экспериментах на тестовых данных предложенный метод позволил сократить обучающие выборки размером 2000–5000 объектов в среднем в 60 раз при пересечении областей классов в пространстве признаков на 25% без ухудшения качества распознавания, в то время как известный алгоритм STOLP сократил обучающие выборки в среднем в 35 раз, и качество распознавания по такой выборке ухудшилось на $4,8\%$.

Литература

[1] *Pal S. K., Mitra P.* Pattern Recognition Algorithms for Data Mining: Scalability, Knowledge Discovery and Soft Granular Computing. — Chapman and Hall/CRC, 2004. — 280 p.

[2] *Larose D. T.* Discovering knowledge in Data: An Introduction to Data Mining. — New Jersey: Wiley and Sons, 2005. — 224 p.

[3] *Cherkassky V., Mulier F.* Learning from data. Concept, theory and methods, 2nd ed. — New Jersey: Wiley and Sons, 2007. — 540 p.

[4] *Гаврилова Т. А., Хорошевский В. Ф.* Базы данных интеллектуальных систем. — СПб: Питер, 2000. — 384 с.

[5] *Загоруйко Н. Г.* Прикладные методы анализа знаний и данных. — Новосибирск: Издательство института математики, 1999. — 270 с.

[6] *Robnik-Sikonja M., Kononenko I.* Theoretical and Empirical Analysis of ReliefF and RReliefF. // Machine Learning. — 2003. — V. 53. — P. 23–69.

[7] *Журавлев Ю. И., Гуревич И. Б.* Распознавание образов и распознавание изображений // Распознавание. Классификация. Прогноз. Математические методы и их применение: Сб. научн. работ. — Москва: Наука, 1989. — С. 5–72.

[8] *Волченко Е. В.* Анализ эффективности выбора условий формирования обучающей выборки метаобъектов // Вестник Хмельницкого национального университета. — 2007. — № 2. — С. 85–89.

Усовершенствование алгоритма C4.5 на основе использования полных решающих деревьев*

Генрихов И. Е., Дюкова Е. В.
ingvar1485@rambler.ru, edjukova@mail.ru
Москва, Вычислительный центр им. А. А. Дородницына РАН

Рассматриваются вопросы применения решающих деревьев для задач классификации. Предлагается усовершенствованная модель алгоритма C4.5. Приведены результаты тестирования на реальных задачах.

Одним из наиболее популярных инструментов для решения задач классификации является деревья решений. Построение дерева решений представляет собой итерационный процесс. На каждой итерации выбирается признак, удовлетворяющий определенному критерию ветвления, и строится вершина дерева.

В ситуации, когда несколько признаков удовлетворяют критерию ветвления в равной или почти равной мере, выбор одного из этих признаков осуществляется случайным образом. От выбора признака существенно зависит качество распознавания. В [2] предложен подход, позволяющий учитывать все признаки, удовлетворяющие заданному критерию ветвления. Используемая в [2] конструкция названа полным решающим деревом. В этом дереве наравне с обычными вершинами встречаются особые вершины, называемые полными. Полной вершине соответствует набор признаков, состоящий из более, чем одного признака. В случае, когда при построении дерева рассматривается полная вершина с набором признаков X, ветвление происходит по каждому из признаков, входящих в X. Данный подход продемонстрирован на примере усовершенствования алгоритма построения допустимых разбиений [1]. Усовершенствованная модель алгоритма названа ПДР.

В настоящей работе полное решающее дерево построено на основе хорошо известного алгоритма C4.5. Проведено сравнение нового алгоритма с алгоритмами ПДР и C4.5. Тестирование алгоритмов проводилось на прикладных задачах прогнозирования из различных областей. Большую часть этих задач составили задачи медицинской диагностики. Наименьшую эффективность показал алгоритм C4.5. Алгоритм полный C4.5 показал лучшие результаты по сравнению с алгоритмами ПДР и C4.5 на большинстве рассмотренных задач.

Основные понятия

Рассматривается задача распознавания с целочисленными признаками $x_1, \ldots, x_n$, с непересекающимися классами $K_1, \ldots, K_l$ и обучающими объектами $S_1, \ldots, S_m$, где $S_r = (a_{r1}, \ldots, a_{rn})$, $r = 1, \ldots, m$, a_{rj} — значение признака x_j для объекта S_r, $a_{rj} \in \{0, \ldots, k-1\}$.

Ориентированное k-арное дерево будем называть k-арным решающим деревом (РД), если:

- каждой внутренней вершине дерева сопоставлен один из признаков $x_1, \ldots, x_n$;
- каждой висячей вершине сопоставлен один из классов $K_1, \ldots, K_l$;
- каждая дуга помечена числом от 0 до $k-1$, причем из каждой внутренней вершины выходят не более k дуг, помеченные разными числами.

Нетрудно видеть, что ветви РД с вершинами $x_{j_1}, \ldots, x_{j_r}$ можно сопоставить элементарную конъюнкцию (э. к.) над переменными $x_1, \ldots, x_n$ вида $x_{j_1}^{\sigma_1} \ldots x_{j_r}^{\sigma_r}$, где σ_i — метка дуги, выходящей из вершины x_{j_r}, $i = 1, \ldots, r$, и $x^\sigma = 1$, если $x = \sigma$, и $x^\sigma = 0$ в противном случае. С другой стороны, каждой висячей вершине, а, следовательно, каждой ветви РД сопоставлен один из классов $K_1, \ldots, K_l$. Таким образом, каждой ветви РД соответсвует пара (B, K), где B — э.к. над переменными $x_1, \ldots, x_n$, $K \in \{K_1, \ldots, K_l\}$. Через N_B будем обозначать интервал истинности конъюнкции B. Пусть РД имеет μ ветвей, и этим ветвям соответсвуют пары $(B_1, K_{j_1}), \ldots, (B_\mu, K_{j_\mu})$.

Пусть далее S — распознаваемый объект. Если среди конъюнкций $B_1, \ldots, B_\mu$ существует конъюнкция B_i такая, что $S \in N_{B_i}$, то объект S относится к классу K_{j_i}. В противном случае происходит отказ от распознавания объекта S.

Пусть некоторой ветви дерева сопоставлена пара (B, K). Рассматриваемая ветвь называется *корректной*, если из условия $S_i \in \{S_1, \ldots, S_m\} \cap N_B$ следует, что S_i принадлежит классу K. В этом случае интервал N_B называется допустимым. Решающее дерево, у которого каждая ветвь является корректной, называется корректным.

Выбор признака для ветвления в алгоритме C4.5 осуществляется на основе энтропийного критерия из теории информации. Рассмотрим критерий выбора признака в алгоритме C4.5.

Пусть далее U — некоторое множество объектов из T. Обозначим через $f(K_i, U)$, $i = 1, \ldots, l$ число объектов из множества U, относящихся к классу K_i. Вероятность того, что случайно выбранный

*Работа выполнена при финансовой поддержке РФФИ, проект №07-01-00516, гранта Президента РФ по поддержке ведущих научных школ НШ №5294.2008.1 и гранта Президента РФ по поддержке молодых кандидатов наук МК №6500.2008.9

объект из множества U будет принадлежать классу K_i, равна $f(K_i, U)/|U|$.

Количество информации (энтропия), необходимое для определения класса, которому принадлежит объект из множества U, вычисляется по формуле

$$\mathrm{Info}(U) = -\sum_{i=1}^{l} \frac{f(K_i, U)}{|U|} \log_2 \left(\frac{f(K_i, U)}{|U|} \right)$$

Пусть $T = \{S_1, \ldots, S_m\}$, $\{k_{j_1}, \ldots, k_{j_t}\}$ — значения признака x_t, $t = 1, \ldots, n$. Множество T разбивается на подмножества $T^{(T)}_{k_{j_1}}, \ldots, T^{(T)}_{k_{j_t}}$, где $T^{(T)}_i$ состоит из обучающих объектов, для которых значение признака x_t равно i.

Количество информации, необходимое для определения класса, которому принадлежит объект из множества T после разбиения T по значениям признака x_t, оценивается величиной

$$\mathrm{Info}_{x_t}(T) = \sum_{i=k_{j_1}}^{k_{j_t}} \left(\frac{|T^{(T)}_i|}{|T|} \mathrm{Info}\big(T^{(T)}_i\big) \right)$$

Информационный выигрыш (information gain) после выбора признака x_t вычисляется по формуле $\mathrm{Gain}(x_t) = \mathrm{Info}(T) - \mathrm{Info}_{x_t}$.

Если для ветвления выбирать признак x_j, для которого $\mathrm{Gain}(x_j)$ принимает наибольшее значение, то информативными будут «шумящие» признаки, т. е. признаки, принимающие слишком много значений. С точки зрения предсказательных способностей построенная модель становится бесполезной.

Рассмотрим величину

$$\mathrm{SplitInfo}_{x_t}(T) = -\sum_{i=k_{j_1}}^{k_{j_t}} \frac{|T^{(T)}_i|}{|T|} \log_2 \left(\frac{|T^{(T)}_i|}{|T|} \right)$$

Величина $\mathrm{SplitInfo}_{x_t}(T)$ оценивает потенциальный информационный выигрыш, получаемый при разбиении множества T по признаку x_t [3]. Существуют и другие приемы, решающие недостаток критерия ветвления [6]. Поэтому в алгоритме C4.5 выбор признака определяется не по критерию $\mathrm{Gain}(x_t)$, а на основе критерия $\mathrm{GainRatio}(x_t)$, где

$$\mathrm{GainRatio}(x_t) = \frac{\mathrm{Gain}(x_t)}{\mathrm{SplitInfo}(x_t)}, \quad t \in \{1, \ldots, n\}.$$

Для ветвления выбирается признак, для которого данная величина принимает наибольшее значение.

В [3] описан вариант алгоритма C4.5, который не является корректным. Для того чтобы алгоритм был корректным необходимо выполнение следующих условий:

1) классы $K_1, \ldots, K_l$, представляющие объекты $S_1, \ldots, S_m$ множества T, не пересекаются (ограничение на матрицу исходных данных);
2) висячая вершина строится только в том случае, если $T^{(T)}_i$ состоит из объектов, принадлежащих одному классу.

Описание корректного варианта алгоритма C4.5

Алгоритм C4.5 является рекурсивным. Множество обучающих объектов T представляется в виде матрицы $T(a_{rj})$, $r = 1, \ldots, m$, $j = 1, \ldots, n$, в которой a_{rj} — значение признака x_j для объекта S_r. Положим на первом шаге $\tilde{T} = T(a_{rj})$. Далее на каждом шаге рекурсии выполняется следующая последовательность действий:

1. Из матрицы $\tilde{T}$ вычеркиваются все столбцы с номерами j для которых $a_{1j} = \cdots = a_{mj}$. Если $\tilde{T} \neq \varnothing$, то переходим к следующему шагу, иначе рекурсия останавливается.
2. Для каждого признака вычисляется значение $\mathrm{GainRatio}(x_j)$. Пусть t — номер первого по порядку признака, для которого данная величина принимает наибольшее значение. Для каждого значения y признака x_t создается дуга, исходящая из вершины соответсвующей x_t. Для каждой ветви строится подматрица $\tilde{T}^{(T)}_y$ матрицы $\tilde{T}$, полученная удалением столбца, соответствующего признаку с номером t, и строк S_r, в которых $a_{rt} \neq y$, $r = 1, \ldots, m$. Если получившаяся подматрица $\tilde{T}^{(T)}_y$ содержит объекты одного класса, то строится висячая вершина с меткой этого класса, иначе полагается $\tilde{T} = \tilde{T}^{(T)}_y$ и осуществляется переход к следующему шагу рекурсии.

Описание алгоритма полный C4.5

Как видно из описания алгоритма C4.5, если проверяемому условию удовлетворяет более одного признака, то выбор одного из этих признаков происходит фактически случайно. Чтобы исправить указанный недостаток, предлагается несколько модифицировать решающее дерево. Для этого введем второй тип внутренних вершин — полные вершины, которым соответствует набор равноценных признаков. Пусть внутренней вершине второго типа ν соответствует набор признаков $\{x_{j_1}, \ldots, x_{j_q}\}$, $1 \leqslant q \leqslant n$, тогда из этой вершины исходит q дуг $t_1, \ldots, t_q$, каждая из которых связывает вершину ν с вершиной первого типа. Причем дуга t_u, $u = 1, \ldots, q$, входит в вершину первого типа с меткой x_{j_u}. Полученную конструкцию будем называть полным решающим деревом.

Заметим, что в полном решающем дереве, также как и в обычном, каждой висячей вершине соответствует некоторая конъюнкция. Однако в полном

решающем дереве описание распознаваемого объекта может попасть в интервалы истинности конъюнкций, соответствующих разным ветвям дерева. В случае, если таким ветвям дерева соответствуют разные классы, выбор класса осуществляется простым голосование, то есть объект зачисляется в тот класс, который соответствует большинству из указанных ветвей.

Алгоритм построения полного решающего дерева на основе C4.5 также является рекурсивным. Обозначим через $\tilde{T}$ матрицу, расматриваемую на текущем шаге алгоритма, то есть на первом шаге $\tilde{T} = T(a_{rj})$. Шаг рекурсии представляет собой следующую последовательность действий:

1. Из матрицы $\tilde{T}$ вычеркиваются все столбцы с номерами j для которых $a_{1j} = \cdots = a_{mj}$. Если $\tilde{T} \neq \varnothing$, то переходим к следующему шагу, иначе рекурсия останавливается.
2. Для каждого признака вычисляется значение $\mathrm{GainRatio}(x_j)$. Далее вызывается процедура выбора набора признаков. Пусть в результате выбран набор признаков $X = \{x_{j_1}, \ldots, x_{j_q}\}$, $1 \leqslant q \leqslant n$. Если $q = 1$, то создается вершина первого типа с меткой x_{j_1}, иначе создается вершина второго типа с меткой $\{x_{j_1}, \ldots, x_{j_q}\}$. Для каждого признака x_{j_t}, $t = 1, \ldots, q$, создается вершина первого типа, и она соединяется с построенной вершиной второго типа. Для каждого значения y признака $x_{j_t} \in X$, $t = 1, \ldots, q$, создается дуга, исходящая из вершины соответсвующей x_{j_t}. Для каждой ветви строится подматрица $\tilde{T}_y^{(j_t)}$ матрицы $\tilde{T}$, полученная удалением столбца, соответствующего признаку с номером j_t, и строк S_r, в которых $a_{rj_t} \neq y$, $r = 1, \ldots, m$. Если получившаяся подматрица $\tilde{T}_y^{(j_t)}$ содержит объекты одного класса, то строится висячая вершина с меткой этого класса, иначе полагается $\tilde{T} = \tilde{T}_y^{(j_t)}$ и осуществляется переход к следующему шагу рекурсии.

Замечание: так как при построении полного решающего дерева происходит лавинообразный рост числа вершин и ветвей, то увеличивается и время классификации объекта. Чтобы сократить это время, предлагается строить только те ветви, которые дают голос за принадлежность распознаваемого объекта S классу $K \in K_1, \ldots, K_l$ (например: время для классификации одного объекта в задаче 4 (раздел «Результаты тестирования») — 123 мс, при использовании данного предложения время сократилось до 64 мс).

Процерура выбора набора признаков X для ветвления представляет собой следующую последовательность шагов:

1. Пусть $\tilde{T}$ содержит w столбцов, соответсвующих признакам $x_{j_1}, \ldots, x_{j_w}$. Тогда $Y = \{x_{j_1}, \ldots, x_{j_w}\}$.
2. Вычисляется средний информационный выигрыш
$$q = \frac{1}{w} \sum_{i=j_1}^{j_w} \mathrm{GainRatio}(x_i).$$
3. Определяется число признаков, для которых информационный выигрыш выше среднего или равен ему $n = \sum\limits_{i=j_1}^{j_w} c_i$, где
$$c_i = \begin{cases} 1, & \text{если } \mathrm{GainRatio}(x_i) \geqslant q; \\ 0, & \text{если } \mathrm{GainRatio}(x_i) < q. \end{cases}$$
Признаки x_i, для которых $\mathrm{GainRatio}(x_i) < q$, $i = j_1, \ldots, j_w$, удаляются из Y.
4. Вычисляется $h = \min\limits_{x_i \in Y} \mathrm{GainRatio}(x_i)$.
5. Если $(q/n) + h \geqslant \max\limits_{x_i \in Y} \mathrm{GainRatio}(x_i)$, то происходит выход из процедуры и возвращается Y, иначе осуществляется переход к третьему шагу процедуры, при $q = (q/n) + h$.

Предложенная процедура выбора признаков для ветвления выбирает те признаки, информативность которых совпадает с максимальной информативностью признаков из Y, а также признаки, информативность которых близка к максимальной информативности, при этом не надо указывать меру близости по информативности, она вычисляется на основе среднего информационного выигрыша (величина q/n). Неравенство на 5 шаге данной процедуры позволяет оценить близость признаков из Y по информативности, смысл которой в том, что мы останавливаемся только тогда, когда близость между максимальной и минимальной информативностью станет меньше определенного значения q/n. При этом в противном случае средний информационный выигрыш q будет увеличиваться, количество признаков в Y будет уменьшаться, величина q/n будет расти, близость между максимальным и минимальным значением будет уменьшаться, и неравенство на 5 шаге выполнится в любом случае на p шаге процедуры выбора признаков для ветвления.

Алгоритм полный C4.5 является корректным.

Результаты тестирования

Тестирование алгоритмов осуществлялось на пяти прикладных задачах, взятых из репозитория UCI [4]. Для оценки качества работы распознающих алгоритмов C4.5, полного C4.5 и ПДР, использовалась процедура скользящего контроля. Для этой процедуры один шаг алгоритма заключается в удалении одного из объектов обучающей выборки, построении классификатора и распознавании удаленного объекта. Далее приводятся краткие описания этих задач.

Таблица 1. Эффективность алгоритмов.

Задача	C4.5	полный C4.5	ПДР
waveform	68.56	80.67	**85.58**
Задача 2	72.9	73.43	**74.09**
dermatology	93.44	**96.85**	95.08
lymphography	80.4	**81.76**	76.35
Задача 5	84.56	**87.06**	86.24

Задача 1 (waveform, UCI).
Определение вида волны по набору параметров, которыми она описывается. Обучающая выборка состоит из 300 объектов, число признаков 21, значность $k = 16$. Содержательно признаки — это некоторые параметры, используемые для описания волны. Множество объектов разделено на три класса (три вида волн).

Задача 2 (инсульт [7]).
Дифференциальная диагностика инсульта. Обучающая выборка состоит из 301 объекта, число признаков 56, значность $k = 6$. Содержательно признаки — это некоторые характеристики состояния больного, проходящего обследование. Множество объектов разделено на два класса (ишемический инсульт и геморрагический инсульт). Объекты представляют собой описания результатов обследований пациентов.

Задача 3 (dermatology, UCI).
Дифференциальная диагностика эритематосквамозных заболеваний является довольно трудной задачей в дерматологии. К таким заболеваниям относятся псориаз, себорейный дерматит, лишай плоский, лишай розовый, хронический дерматит и лишай красный волосяной. Обычно для установления диагноза проводится биопсия. Однако перечисленные болезни вызывают также похожие клеточные изменения. Еще одной трудностью является то, что на начальной стадии могут появляться признаки другого заболевания из той же группы. Обучающая выборка состоит из 366 объектов (результатов обследований больных), число признаков 34, значность $k = 4$. Множество признаков содержит 12 клинических и 22 гистопатологических признака.

Задача 4 (lymphography, UCI).
Диагностика раковых заболеваний по состоянию лимфатических узлов. Обучающая выборка состоит из 148 объектов, число признаков 18, значность $k = 8$. Множество объектов разделено на четыре класса (четыре состояния лимфатических узлов): нормальное состояние, метастазы, злокачественные образования и фиброз. Объекты представляют собой описания результатов обследований пациентов.

Задача 5.
Обучающая выборка состоит из 2060 объектов, число признаков 6, значность $k = 16$. Множество объектов разделено на три класса.

В таблице 1 представлены результаты работы всех рассмотренных алгоритмов для задач, описанных выше. Для каждой задачи и каждого алгоритма приведен процент правильно распознанных объектов на скользящем контроле.

Проведенное тестирование показало, что точность распознавания при использовании полного решающего дерева выше точности распознавания C4.5. Кроме того, на задачах 3–5 новая модель ведет себя лучше алгоритма ПДР.

Литература

[1] *Донской В. И., Башта А. И.* Дискретные модели принятия решений при неполной информации. — Симферополь: Таврия, 1992. — С. 33–74.

[2] *Дюкова Е. В., Песков Н. В.* Об алгоритме классификации на основе полного решающего дерева // Доклады 13-й всероссийской конференции ММРО. Москва: МАКС Пресс. — 2007. — С. 125–126.

[3] *Quinlan J. R.* C4.5: Programs for Machine Learning. Morgan Kaufmann. San Mateo. CA. — 1993.

[4] *Asuncion A., Newman D. J.* UCI Machine Learning Repository, University of California, Irvine. — 2007. `www.ics.uci.edu/~mlearn/MLRepository.html`.

[5] *Дюкова Е. В.* Дискретные (логические) процедуры распознавания: принципы конструирования, сложность реализации и основные модели // Учебное пособие для студентов математических факультетов педвузов. Москва: Прометей. — 2003.— С. 3–22.

[6] *Martin J. K.* An exact probability metric for decision tree splitting and stopping // Learning from Data: Artificial Intelligence and Statistics V. Edited by D. Fisher and H.-J. Lenz. SP. — 1996.— P. 399–410.

[7] *Реброва О. Ю.* Применение методов интеллектуального анализа для решения задачи медицинской диагностики // Новости искусственного интеллекта. №3 — 2004.— С. 76–80.

Об одном подходе к синтезу алгоритмов коррекции локального возмущения в конечной полуметрике

Громов И. А.
igor_gromov@mail.ru
Москва, Вычислительный центр РАН

Предлагается и исследуется трёхэтапная схема построения алгоритмов преобразования метрической информации в задачах интеллектуального анализа данных. В рамках схемы рассматриваются алгоритмы, корректирующие изменения расстояний на заданную величину между некоторыми парами объектов и гарантировано сохраняющие метрические свойства. Устанавливаются достаточные условия, при выполнении которых коррекция полуметрик в рамках трёхэтапной схемы завершается в ходе первых двух этапов, а в специальном случае — уже в ходе первого этапа. Рассматриваются функционалы сходства полуметрик, проводится согласование алгоритмов коррекции с данными функционалами.

В теории распознавания образов и практике решения прикладных задач интеллектуального анализа данных широко применяются метрические методы. Эффективность их использования существенно зависит от выбора функции сходства (например, полуметрики) на объектах распознавания. Как правило, в прикладных задачах нельзя ввести некоторую единственную «объективную» полуметрику, следовательно, сами полуметрики становятся предметом анализа и настройки. В данной работе рассматривается случай, когда настройка предварительно полученной полуметрики начинается с изменения экспертом в предметной области расстояний между выбранными объектами.

Постановка задачи

В работе используется стандартное определение метрики. Пусть V — произвольное непустое множество.

Определение 1. *Функционал $\rho\colon V \times V \to \mathbb{R}_+$ называется метрикой на V, если он удовлетворяет следующим условиям:*

1. $\rho(v, v) = 0,\ \forall\ v \in V$;
2. $\rho(v_1, v_2) = \rho(v_2, v_1),\ \forall\ v_1, v_2 \in V$;
3. $\rho(v_1, v_3) \leqslant \rho(v_1, v_2) + \rho(v_2, v_3),\ \forall\ v_1, v_2, v_3 \in V$;
4. $\rho(v_1, v_2) = 0 \Rightarrow v_1 = v_2,\ \forall\ v_1, v_2 \in V$.

Если функционал ρ удовлетворяет только условиям (1)–(3), то он называется полуметрикой на V. Если ρ удовлетворяет только условиям (1)–(2), то он называется расстоянием, или функцией расстояния, на V. Условие (3) принято называть неравенством треугольника.

В работе будет использована графовая интерпретация метрики, определённой на конечном множестве мощности N, как нагруженной клики на N вершинах. Множество вершин графа — это это множество объектов, множество рёбер — это множество попарных расстояний между объектами, а расстояния между объектами соответствуют весам, приписанным рёбрам.

Сформулируем задачу «коррекции локального возмущения». Пусть дано конечное множество объектов V с заданной на нем полуметрикой ρ. Эксперт фиксирует некоторое подмножество объектов V' (мощности не менее двух) и задаёт новые значения расстояний между парами объектов из V', не нарушая свойств полуметрики в V'. При этом для троек объектов, в которых одновременно присутствуют как объекты из V', так и из $V \setminus V'$, неравенства треугольника верны необязательно. Такую модификацию расстояний будем называть *локальным возмущением*. В результате внесения локального возмущения возникает новая функция расстояния ρ'. Эксперт также выбирает функционал сходства полуметрик, который формализует его интуитивное представление об их сходстве. Задача коррекции локального возмущения состоит в синтезе новой полуметрики $\tilde{\rho}$, которая близка к исходной ρ в смысле выбранного функционала, но сохраняет сделанные изменения.

Общие требования, предъявляемые к алгоритму решения задачи, следующие:

- **(R1)** Указанные экспертом значения расстояний на выбранных парах объектов сохраняются в полуметрике $\tilde{\rho}$.
- **(R2)** Алгоритмы должны быть *универсальны*, т. е. применимы ко всем ρ и ρ'.
- **(R3)** Алгоритмы строят полуметрику $\tilde{\rho}$, максимально близкую (в смысле выбранного функционала) к исходной полуметрике ρ.

Однако поиск точного решения задачи затрудняет реальная экстенсивная сложность полуметрических условий. В работе предлагаются некоторые вычислительно эффективные алгоритмы приближённого решения данной задачи, предполагающие ослабление указанных требований.

Существуют различные формализации задач коррекции полуметрики и подходы к её решению [1, 2]. В данной работе предложен новый подход — трёхэтапная схема построения алгоритмов коррекции возмущённых полуметрик. В общем случае данная схема требует исследования всех троек объектов и имеет сложность $O(N^3)$, где N — общее число обрабатываемых объектов. Одна-

ко в работе установлены достаточные условия, при выполнении которых и ослаблении условий (R2) и (R3) коррекция локального возмущения в рамках трёхэтапной схемы завершается в ходе первых двух этапов, а в специальном случае использования конкретного функционала — уже в ходе первого этапа, без ослабления (R3). Таким образом, сложность понижается до $O(N^2)$ и $O(N)$ соответственно. В работе приводятся конкретные алгоритмы коррекции, полученные в рамках данной схемы. Получаемые решения не всегда оптимальны. Но они всегда начинают строиться как оптимальные, после чего за счёт ослабления требований эффективно достраиваются до полуметрики.

Будем рассматривать конечные множества объектов V_N мощности N, элементы которых отождествим с их индексами $1,\dots,N$. Пусть задана полуметрика $\rho\colon V_N \times V_N \to \mathbb{R}_+$. Матрицу попарных расстояний, определённых полуметрикой ρ, обозначим R; $R = (r_{ij})_{N\times N} \in {\mathbb{R}_+}^{N\times N}$, где $r_{ij} \stackrel{\text{def}}{=} \rho(i,j)$. Функцию расстояния, полученную в результате экспертной модификации (внесения локального возмущения), обозначим ρ', а полуметрику, полученную в результате последующей коррекции, — $\tilde\rho$. R' и $\tilde R$ — матрицы попарных расстояний, соответствующие функциям ρ' и $\tilde\rho$.

Во всей работе индексами $i_1,\dots,i_M$ обозначим объекты, расстояния между которыми изменил эксперт: $r_{i_1i_2} \mapsto r'_{i_1i_2}, \dots, r_{i_{M-1}i_M} \mapsto r'_{i_{M-1}i_M} \dots$ Обозначим $V' \ni \{i_1,\dots,i_M\}$. Новые расстояния должны сохраниться в скорректированной полуметрике: $\tilde r_{i_ui_v} = r'_{i_ui_v}$, где $i_u, i_v \in V'$. Обозначим через V_{N_0} множество $V_N \setminus V'$.

Введём множества неупорядоченных пар и троек индексов

$$
\begin{aligned}
E_{N_0} &= \big\{\{i,j\}\colon i,j \in V_{N_0},\ i\neq j\big\};\\
E' &= \big\{\{i_u,i_v\}\colon i_u,i_v \in V',\ i_u\neq i_v\big\};\\
T_{N_0} &= \big\{\{i,j,k\}\colon i,j,k \in V_{N_0},\ i\neq j\neq k\neq i\big\};\\
T_N &= \big\{\{i,j,k\}\colon i,j,k \in V_N,\ i\neq j\neq k\neq i\big\}.
\end{aligned}
$$

Обозначим через P^ρ_{ij} множество (отрезок числовой оси) допустимых значений расстояния r_{ij} между объектами i и j в полуметрике ρ при фиксированных остальных расстояниях. Нетрудно видеть, что множество P^ρ_{ij} непусто, и $P^\rho_{ij} = \big[r^{\min}_{ij}, r^{\max}_{ij}\big]$, где

$$
\begin{aligned}
r^{\min}_{ij} &= \max_{k\in V_N\setminus\{i,j\}} |r_{ik} - r_{jk}|\,;\\
r^{\max}_{ij} &= \min_{k\in V_N\setminus\{i,j\}} (r_{ik} + r_{jk})\,;
\end{aligned}
$$

для всех $\{i,j\} \in E_N$. Если для всех $\{i_u,i_v\} \in E'$ значения $r'_{i_ui_v} \in P^\rho_{i_ui_v}$, то ρ' — полуметрика.

Треугольник, построенный на тройке вершин (объектов) $\{i,j,k\} \in T_N$, будем обозначать $\triangle ijk$.

Трёхэтапная схема построения алгоритмов коррекции полуметрики

Пусть эксперт модифицировал расстояния $\big\{r_{i_ui_v}\colon \{i_u,i_v\} \in E'\big\}$, т. е. указал значения $r'_{i_ui_v}$ ($r'_{i_ui_v} \geqslant 0$). Если для всех $\{i_u,i_v\} \in E'$ значения $r'_{i_ui_v} \in P^\rho_{i_ui_v}$, то ρ' есть полуметрика, и коррекция других расстояний в ρ' не является необходимой. В этом случае ρ' может быть использована как готовая экспертная полуметрика.

Если $r'_{i_ui_v} \notin P^\rho_{i_ui_v}$, то это означает нарушение неравенств треугольника в некоторых треугольниках вида $\triangle i_ui_vj_k$, где $j_k \in V_{N_0}$, и ρ' следует корректировать. (Как указывалось выше, полагаем, что возможные нарушения в треугольниках вида $\triangle i_ui_vi_w$ эксперт скорректировал сам.) Предлагается строить алгоритмы коррекции по следующей «трёхэтапной схеме».

1-й этап: коррекция $\triangle i_ui_vj_k$ для всех $j_k \in V_{N_0}$. Обрабатываются расстояния $r_{i_uj_k}, r_{i_vj_k}$, следовательно, после первого этапа неравенства треугольника могут быть нарушены в $\triangle i_uj_kj_l$ и $\triangle i_vj_kj_l$, при $j_k, j_l \in V_{N_0}$.

2-й этап: коррекция $\triangle i_uj_kj_l$ и $\triangle i_vj_kj_l$ для всех $\{j_k,j_l\} \in E_{N_0}$. Используются такие методы, которые не вызовут новых нарушений в уже скорректированных $\triangle i_ui_vj_k$.

3-й этап: коррекция $\triangle j_kj_lj_m$. Используются такие методы, которые не вызовут новых нарушений неравенств треугольника в $\triangle i_ui_vj_k$, $\triangle i_uj_kj_l$, $\triangle i_vj_kj_l$, для всех $\{j_k,j_l,j_m\} \in T_{N_0}$.

Коррекция как бы распространяется «в ширину» от локального возмущения (в качестве аналогии можно упомянуть волновой алгоритм). Сначала рассматриваются треугольники, инцидентные ребру из E', далее — треугольники, инцидентные одной вершине из V', затем — треугольники, не инцидентные вершинам из V'. Нарушения неравенств треугольника в процессе коррекции распространяются по полуметрике. На каждом следующем этапе требуется преобразовать большее число расстояний, чем на предыдущем. На 1-м этапе коррекции требуется проверить $O(N)$ треугольников, на 2-м — $O(N^2)$, на 3-м — $O(N^3)$.

Далее в разделе сформулирован ряд условий, при выполнении которых гарантируется построение полуметрики уже после первых двух этапов. Именно существование таких условий, позволяющих строить алгоритмы сложности $O(N^2)$, и обусловливает ценность трёхэтапной схемы.

В данном разделе вид преобразований, применяемых на первом этапе, не фиксируется.

Здесь и далее в работе для всех $\{j_k,j_l\} \in E_{N_0}$ для определения множеств допустимых значений $P^{\tilde\rho}_{j_kj_l}$ расстояний $\tilde r_{j_kj_l}$ будут использоваться только наборы расстояний вида $\tilde r_{i_sj_k}$, $\tilde r_{i_sj_l}$, $s =$

$= 1, \ldots, M$, полученные в результате проведения первого этапа коррекции. Обозначим

$$\tilde{r}^{\min}_{j_k j_l} = \max_{s=1,\ldots,M} |\tilde{r}_{i_s j_k} - \tilde{r}_{i_s j_l}|;$$
$$\tilde{r}^{\max}_{j_k j_l} = \min_{s=1,\ldots,M} (\tilde{r}_{i_s j_k} + \tilde{r}_{i_s j_l}).$$

Теорема 1. *Пусть в результате проведения первого этапа коррекции полуметрики ρ метрические свойства восстановлены в $\triangle i_u i_v j_k$ для всех $i_u, i_v \in V'$, $j_k \in V_{N_0}$. Если коррекция на втором этапе проводится по правилу*

$$\tilde{r}_{j_k j_l} = \tilde{r}^{\min}_{j_k j_l}, \quad \forall\, \{j_k, j_l\} \in E_{N_0}, \tag{1}$$

то $\tilde{\rho}$ есть полуметрика.

Аналогичное утверждение справедливо и в случае использования правила

$$\tilde{r}_{j_k j_l} = \tilde{r}^{\max}_{j_k j_l}, \quad \forall\, \{j_k, j_l\} \in E_{N_0}. \tag{2}$$

Следствие 1. *Пусть в результате проведения первого этапа коррекции полуметрики ρ метрические свойства восстановлены в $\triangle i_u i_v j_k$ для всех $i_u, i_v \in V'$, $j_k \in V_{N_0}$. Если коррекция на втором этапе проводится по правилу*

$$\tilde{r}_{j_k j_l} = \alpha \tilde{r}^{\min}_{j_k j_l} + (1-\alpha) \tilde{r}^{\max}_{j_k j_l}, \ \forall\, \{j_k, j_l\} \in E_{N_0}, \tag{3}$$

где α — произвольное фиксированное число из отрезка $[0, 1]$, то $\tilde{\rho}$ есть полуметрика.

Согласно данным утверждениям, какие бы процедуры не использовались на первом этапе коррекции, на втором этапе должны быть скорректированы значения всех расстояний вида $\tilde{r}_{j_k j_l}$. Однако если на процедуры первого этапа наложить ограничения, на втором этапе появится возможность сохранить некоторые исходные расстояния.

Теорема 2. *Если в алгоритме, построенном в рамках трёхэтапной схемы, коррекция на первом этапе удовлетворяет условию $\tilde{r}_{i_u j_k} \geqslant r_{i_u j_k}$ для всех $i_u \in V'$, $j_k \in V_{N_0}$, а на втором этапе проводится по правилу*

$$\tilde{r}_{j_k j_l} = \begin{cases} r_{j_k j_l}, & \text{если } \tilde{r}^{\min}_{j_k j_l} \leqslant r_{j_k j_l} \leqslant \tilde{r}^{\max}_{j_k j_l}; \\ \tilde{r}^{\min}_{j_k j_l}, & \text{если } r_{j_k j_l} < \tilde{r}^{\min}_{j_k j_l}, \end{cases} \forall\, \{j_k, j_l\} \in E_{N_0}, \tag{4}$$

то $\tilde{\rho}$ есть полуметрика.

Указанные в теоремах формулы (1)–(4) очевидным образом задают вторые этапы алгоритмов. В этих алгоритмах на втором этапе каждое расстояние корректируется не более одного раза.

Функционалы сходства полуметрик и первый этап коррекции

Остаётся открытым вопрос о построении полуметрики $\tilde{\rho}$, схожей в некотором смысле с исходной (требование (R3)). Для его решения, а также для оценки сходства полуметрик ρ и $\tilde{\rho}$ в данном разделе вводятся функционалы сходства.

Рассматриваются два основных подхода (соответственно, два функционала), отличающиеся тем, играют ли ребра $\{i_u, i_v\}$ выделенную роль (относительный подход) или нет (абсолютный подход).

Абсолютный подход. Схожими считаются полуметрики, в которых абсолютные величины соответствующих расстояний минимально различны. В качестве функционала, формализующего такое представление о сходстве, будем рассматривать

$$Q_a(R, \tilde{R}) = \frac{\sum\limits_{j_k \in V_{N_0}} \sum\limits_{i_s \in V'} (\tilde{r}_{i_s j_k} - r_{i_s j_k})^2}{\sum\limits_{j_k \in V_{N_0}} \sum\limits_{i_s \in V'} r^2_{i_s j_k}}.$$

Нормировка введена для сопоставимости масштабов предлагаемых функционалов.

Относительный подход основан на исследовании отношений расстояний между объектами в $\triangle i_u i_v j_k$ и, следовательно, применим только для метрик. В рамках относительного подхода схожими считаются метрики, в которых близки пропорции соответствующих расстояний вида $\dfrac{r_{i_u j_k}}{r_{i_v j_k}}$. Такой подход соответствует требованию эксперта сохранить по возможности отношения нескорректированных расстояний в $\triangle i_u i_v j_k$: $\dfrac{\tilde{r}_{i_u j_k}}{\tilde{r}_{i_v j_k}} = \dfrac{r_{i_u j_k}}{r_{i_v j_k}}$ и формализуется функционалом

$$Q_r(R, \tilde{R}) = \sum_{j_k \in V_{N_0}} \sum_{\{u,v\} \in E'} \left(\frac{\tilde{r}_{u j_k}}{\tilde{r}_{v j_k}} - \frac{r_{u j_k}}{r_{v j_k}} \right)^2,$$

где $u = i_u$ и $v = i_v$, если $r_{i_u j_k} \leqslant r_{i_v j_k}$, и $u = i_v$ и $v = i_u$ в противном случае.

Функционал $Q_r(R, \tilde{R})$ инвариантен относительно масштаба исходной и получаемой полуметрик. Для определённости будем дополнительно требовать выполнение одного из неравенств треугольника как равенства в каждом корректируемом треугольнике (при этом $r'_{i_u i_v}$ не участвуют в масштабировании).

В том случае, когда эксперт требует найти компромисс между сохранением относительных и абсолютных величин расстояний в $\triangle i_u i_v j_k$, предлагается использовать *взвешенный подход*:

$$Q_w\big(R, \tilde{R}(w)\big) = \\ = wQ_a\big(R, \tilde{R}(w)\big) + (1-w) Q_r\big(R, \tilde{R}(w)\big), \ w \in [0, 1].$$

На различных подмножествах множества треугольников $\big\{\triangle i_u i_v j_k \colon j_k \in V_{N_0}\big\}$ можно применять различные значения веса w.

Для случая $|V'| = 2$ (т. е. при экспертной коррекции единственного расстояния в полуметрике)

в явном виде получены формулы для определения значений $\tilde{r}_{i_s j_k}$, доставляющих минимумы Q_a, Q_r и Q_w при заданном значении w. Для $|V'| > 2$ задача поиска оптимальных значений $\tilde{r}_{i_s j_k}$ решается методами условной оптимизации.

Алгоритмы коррекции полуметрик

В рамках трёхэтапной схемы на основе предыдущих результатов могут быть построены конкретные алгоритмы коррекции локальных возмущений. В данном разделе вводятся два таких алгоритма: $\mathcal{A}$, $\mathcal{A}_l$ и рассматриваются их свойства. Оба они удовлетворяют требованию (R1) и учитывают указанный экспертом функционал сходства на первом этапе. Мы выбираем формулировки алгоритмов, максимально согласованные с функционалами сходства, хотя сформулированные в разделе 2 теоремы гарантируют получение полуметрик и с помощью алгоритмов более общего вида.

В алгоритме $\mathcal{A}$ с функционалами сходства согласован только первый этап. Второй этап фиксирован; на нём используется вычислительно эффективная процедура, «достраивающая» полуметрику $\tilde{\rho}$ и по возможности минимально деформирующая расстояния вида r_{kl}. Таким образом, в $\mathcal{A}$ требование (R3) выполнено частично: ρ и $\tilde{\rho}$ максимально близки в смысле указанного экспертом функционала на множестве расстояний вида $r_{i_s j_k}$ для всех $i_s \in V'$, $j_k \in V_{N_0}$. Выполнение третьего этапа не требуется. В алгоритме $\mathcal{A}_l$ первый этап согласован с функционалом Q_a, он завершается получением оптимального решения, и выполнение второго и третьего этапов не требуется.

1. Алгоритм коррекции полуметрики $\mathcal{A}$.

Пусть эксперт внёс в полуметрику ρ локальное возмущение: $r_{i_u i_v} \mapsto r'_{i_u i_v}$ для всех $\{i_u, i_v\} \in E'$, требует сохранить указанные им значения $r'_{i_u i_v}$ и выбрал функционал сходства. Тогда для коррекции в ρ' нарушений неравенств треугольника предлагается следующий алгоритм $\mathcal{A}$:

1-й этап: коррекция $\triangle i_u i_v j_k$ отвечает условию $\tilde{r}_{i_u j_k} \geqslant r_{i_u j_k}$, $\tilde{r}_{i_v j_k} \geqslant r_{i_v j_k}$. При этом значения $\tilde{r}_{i_u j_k}$, $\tilde{r}_{i_v j_k}$ должны минимизировать выбранный экспертом функционал сходства полуметрик.

2-й этап: коррекция $\triangle i_u j_k j_l$ проводится по следующему правилу:

$$\tilde{r}_{j_k j_l} = \begin{cases} r_{j_k j_l}, & \text{если } \tilde{r}^{\min}_{j_k j_l} \leqslant r_{j_k j_l} \leqslant \tilde{r}^{\max}_{j_k j_l}; \\ \tilde{r}^{\min}_{j_k j_l}, & \text{если } r_{j_k j_l} < \tilde{r}^{\min}_{j_k j_l}, \forall \{j_k, j_l\} \in E_{N_0}. \end{cases}$$

3-й этап: не требуется, согласно теореме 2.

На первом этапе выполнения алгоритма требуется рассмотреть $O(N)$ треугольников, на втором — $O(N^2)$ треугольников. Вычислительная сложность первого этапа зависит от выбора метода условной оптимизации; на втором этапе обработка каждого треугольника требует постоянного времени.

2. Алгоритм $\mathcal{A}_l$ коррекции полуметрики, имеющий линейную сложность.

Теорема 3. *Пусть в полуметрике ρ эксперт задал новое значение для одного расстояния $r_{i_1 i_2}$, причем $r'_{i_1 i_2} > r_{i_1 i_2}$. Если первый этап коррекции отвечает условию*

$$\left\{\tilde{r}_{i_1 k}, \tilde{r}_{i_2 k} \colon k \in V_{N_0}\right\} = \arg\min_{\tilde{R}} Q_a(R, \tilde{R}),$$

то полученная в результате матрица $\tilde{R}$ есть матрица полуметрики.

Следствие 2. *Алгоритм $\mathcal{A}_l$ не требует выполнения 2-го и 3-го этапов и имеет сложность $O(N)$, т. е. является линейным относительно числа объектов.*

В явном виде получены формулы для $\tilde{r}_{i_1 k}$, $\tilde{r}_{i_2 k}$, удовлетворяющие условиям теоремы 3.

Выводы

Предложенная схема синтеза алгоритмов коррекции локального возмущения учитывает интерактивный характер задачи экспертной настройки полуметрики. За счёт использования функционалов сходства полуметрик удаётся приблизить её решение к оптимальному в определённом экспертом смысле. Установленные достаточные условия позволяют проводить процедуру коррекции вычислительно эффективно.

Литература

[1] *Миркин Б. Г., Родин С. Н.* Графы и гены. — М.: Наука, 1997. — 314 с.

[2] *Майсурадзе А. И.* Гомогенные и ранговые базисы в пространствах метрических конфигураций // ЖВМ и МФ. — 2006. — Т. 46, № 2. — С. 344—361.

Распознавание элементов множества, представленных взаимными расстояниями и близостями*

Двоенко С. Д.
dsd@uic.tula.ru
Тульский государственный университет

Идея обучения основана на естественном предположении, что похожие явления оказываются похожими по своим основным характеристикам, поведению и т. д. В современных условиях данные об изучаемых объектах все чаще представлены результатами парных сравнений, чем традиционным способом в виде результатов измерений их отдельных характеристик. По-прежнему предполагая, что объекты погружены в некоторое признаковое пространство, мы не требуем его восстановления с целью непосредственного представления объектов в виде векторов. В случае, когда элементы обучающего множества (объекты, признаки) представлены взаимными расстояниями или близостями, рассмотрены представление оптимальной разделяющей гиперплоскости и алгоритм её построения на примере модификации алгоритма Б. Н. Козинца.

Введение

В задаче распознавания образов предполагается, что объекты $\omega_i \in \Omega,\ i = 1, \ldots, N$, погружены в n-мерное пространство признаков (обычно евклидово). Каждый объект ω_i представлен вектором $\boldsymbol{x}_i = \boldsymbol{x}(\omega_i)$, где $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{in})$, а все объекты представлены матрицей $X(N, n)$. Признаки $X_j = (x_{1j}, \ldots, x_{Nj})^{\intercal},\ j = 1, \ldots, n$ образуют координатные оси и представлены множествами наблюдений.

В случае линейной разделимости множества объектов Ω, например, на два класса Ω_1 и Ω_2, необходимо построить решающее правило

$$g(\boldsymbol{a}, \boldsymbol{x}) = \sum_{l=1}^{n} a_l x_l + a_0 = (\boldsymbol{a} \circ \boldsymbol{x}) + a_0, \qquad (1)$$

где $(\boldsymbol{a} \circ \boldsymbol{x})$ будет обозначать скалярное произведение векторов $\boldsymbol{x}$ и $\boldsymbol{a}$, $\boldsymbol{a} = (a_1, \ldots, a_n)$ — направляющий вектор разделяющей гиперплоскости, a_0 — её смещение от начала координат, $g(\boldsymbol{a}, \boldsymbol{x}) \geqslant 0$ для $\boldsymbol{x} \in \Omega_1$ и $g(\boldsymbol{a}, \boldsymbol{x}) < 0$ для $\boldsymbol{x} \in \Omega_2$.

В случае, когда элементы множества представлены только взаимными расстояниями или взаимными скалярными произведениями, то до задачи их распознавания, вообще говоря, возникает нетривиальная проблема восстановления неизвестного пространства признаков.

В общем случае это проблема восстановления интерпретируемых значений на координатных осях. Это самостоятельная проблема, решаемая, например, в задачах многомерного шкалирования и факторного анализа [1, 2].

В данном случае её не требуется решать. Пусть дана матрица расстояний $D(N, N)$ между объектами $\omega_i \in \Omega$ в неизвестном евклидовом пространстве. Скалярные произведения $c_{ij} = (\omega_i \circ \omega_j)$ можно представить расстояниями

$$c_{ij} = \tfrac{1}{2}(d_{0i}^2 + d_{0j}^2 - d_{ij}^2) \qquad (2)$$

относительно любой точки в пространстве, взятой как начало координат (обозначим её как объект ω_0), для всех пар объектов $\omega_i \in \Omega$ и $\omega_j \in \Omega$, где $d_{0p} = d(\omega_0, \omega_p)$ — расстояние от объекта ω_p до начала координат, $d_{pq} = d(\omega_p, \omega_q)$ — расстояние между объектами ω_p и ω_q.

Из (2) следует, что $c_{ii} = d_{0i}^2$. Поэтому, если задана матрица скалярных произведений $C(N, N)$, то её диагональные элементы представляют квадраты расстояний от объектов $\omega_i \in \Omega$ до начала координат ω_0, а расстояния между объектами выражаются через скалярные произведения

$$d_{ij}^2 = c_{ii} + c_{jj} - 2c_{ij}. \qquad (3)$$

Будем считать, что объект ω_0 представлен (как и всякий другой объект из Ω), если он представлен своими расстояниями до остальных объектов из обучающего множества Ω.

Представление оптимальной разделяющей гиперплоскости

Если классы Ω_1 и Ω_2 разделимы, то выпуклые оболочки множеств элементов, содержащихся в них, по крайней мере, не пересекаются. Если выпуклые оболочки не соприкасаются, то в зазоре между ними можно построить некоторое множество разделяющих гиперплоскостей.

Обычно, в отсутствие иной априорной информации, рассматривают гиперплоскость, наиболее удаленную от выпуклых оболочек разделяемых множеств. Такая оптимальная разделяющая гиперплоскость обеспечивает минимальное число ошибок при распознавании новых объектов, которые не участвовали в обучении.

Пусть $\boldsymbol{y} \in \Omega_1$ и $\boldsymbol{z} \in \Omega_2$ являются ближайшими точками выпуклых оболочек множеств Ω_1 и Ω_2. Тогда оптимальная разделяющая гиперплоскость определяется направляющим вектором $\boldsymbol{a} = \boldsymbol{y} - \boldsymbol{z}$ и смещением $a_0 = -\frac{1}{2}((\boldsymbol{y} - \boldsymbol{z}) \circ (\boldsymbol{y} + \boldsymbol{z}))$.

Скалярные произведения направляющего вектора $c_{ai} = (\boldsymbol{a} \circ \boldsymbol{x}_i)$ немедленно выражаются через

*Работа выполнена при финансовой поддержке РФФИ, проекты №08-01-12023, №08-01-99003, №09-07-00394.

расстояния в виде

$$c_{ai} = \sum_{l=1}^{n} a_l x_{il} = \sum_{l=1}^{n} y_l x_{il} - \sum_{l=1}^{n} z_l x_{il} = \\ = \tfrac{1}{2}(d_{0y}^2 - d_{0z}^2) - \tfrac{1}{2}(d_{yi}^2 - d_{zi}^2). \qquad (4)$$

Здесь присутствуют расстояния до неизвестного начала координат ω_0, относительно которого заданы скалярных произведения. Поэтому удобно определить его, например, как центр «тяжести» $\bar{\omega}$ системы точек и представить его своими расстояниями до остальных объектов из обучающего множества, согласно методу главных проекций Торгенсона [3]:

$$d^2(\bar{\omega}, \omega_i) = \frac{1}{N}\sum_{p=1}^{N} d_{ip}^2 - \frac{1}{2N^2}\sum_{p=1}^{N}\sum_{q=1}^{N} d^2 pq.$$

Определив объект $\bar{\omega}$ как начало координат, мы можем вычислить и взаимные скалярные произведения объектов ω_i и ω_j относительно него как

$$c_{ij} = \tfrac{1}{2}\big(d^2(\bar{\omega}, \omega_i) + d^2(\bar{\omega}, \omega_j) - d^2(\omega_i, \omega_j)\big).$$

Найдем смещение a_0, представив его на основе и скалярных произведений, и расстояний

$$a_0 = -\tfrac{1}{2}\sum_{l=1}^{n}(y_l + z_l)(y_l - z_l) = \\ = -\frac{1}{2}\left(\sum_{l=1}^{n} y_l^2 - \sum_{l=1}^{n} z_l^2\right) = \\ = -\tfrac{1}{2}(c_{yy} - c_{zz}) = -\tfrac{1}{2}(d_{0y}^2 - d_{0z}^2).$$

Определим направляющий объект $\omega_a = \omega(\boldsymbol{a})$ и представим решающее правило распознавания $g(\boldsymbol{a}, \boldsymbol{x}) = g\big(\omega(\boldsymbol{a}), \omega(\boldsymbol{x})\big) = g(\omega_a, \omega)$ как скалярное произведение объектов в виде

$$(\omega_a \circ \omega) + a_0 = c_{a\omega} + a_0. \qquad (5)$$

Выразим решающее правило для распознавания нового объекта ω через расстояния

$$(\omega_a \circ \omega) + a_0 = -\tfrac{1}{2}(d_{y\omega}^2 - d_{z\omega}^2),$$

где $d_{y\omega} = d(\omega_y, \omega)$ — расстояние между объектами $\omega_y = \omega(\boldsymbol{y})$ и ω, $d_{z\omega} = d(\omega_z, \omega)$ — расстояние между объектами $\omega_z = \omega(\boldsymbol{z})$ и ω. Следовательно, объект ω представлен только своими расстояниями до объектов ω_y и ω_z и принадлежит к классу, расстояние до ближайшей точки выпуклой оболочки которого меньше.

Выразим из (3) и (4) решающее правило для распознавания нового объекта ω через скалярные произведения

$$(\omega_a \circ \omega) + a_0 = c_{y\omega} - c_{z\omega} - \tfrac{1}{2}(c_{yy} - c_{zz}),$$

где $c_{y\omega} = (\omega_y \circ \omega)$ — скалярное произведение объектов ω_y и ω, $c_{z\omega} = (\omega_z \circ \omega)$ — скалярное произведение объектов ω_z и ω относительно начала координат. Для нормированных скалярных произведений получим $c_{zz} = c_{yy} = 1$, поэтому

$$(\omega_a \circ \omega) + a_0 = c_{y\omega} - c_{z\omega}.$$

Следовательно, объект ω представлен только своими нормированными скалярными произведениями $c_{y\omega}$ и $c_{z\omega}$ с объектами ω_y и ω_z и принадлежит к более похожему классу, скалярное произведение с ближайшей точкой выпуклой оболочки которого больше.

Представление линейной комбинации объектов

Пусть дана матрица $X(N, n)$, вычислены матрица расстояний $D(N, N)$ с элементами $d_{ij} = d(\omega_i, \omega_j)$ и матрица скалярных произведений $C(N, N)$ с элементами $c_{ij} = (\boldsymbol{x}_i \circ \boldsymbol{x}_j)$ относительно исходного начала координат ω_0.

Рассмотрим два вектора $\boldsymbol{x}_p$ и $\boldsymbol{x}_q$ и найдем их выпуклую линейную комбинацию $\mu\boldsymbol{x}_p + (1-\mu)\boldsymbol{x}_q$, где $0 \leqslant \mu \leqslant 1$. Обозначим её как новый объект $\omega_\mu = \omega\big(\mu\boldsymbol{x}_p + (1-\mu)\boldsymbol{x}_q\big)$.

Относительно начала координат ω_0 объект ω_μ оказывается представлен своими скалярными произведениями $c_{\mu i} = (\omega_\mu \circ \omega_i)$ со всеми остальными объектами

$$c_{\mu i} = \sum_{l=1}^{n} x_{il}(\mu x_{pl} + (1-\mu)x_{ql}) = \\ = \mu\sum_{l=1}^{n} x_{il}x_{pl} + (1-\mu)\sum_{l=1}^{n} x_{il}x_{ql} = \\ = \mu c_{ip} + (1-\mu)c_{iq}.$$

Центрируем матрицу данных относительно вектора $\mu\boldsymbol{x}_p + (1-\mu)\boldsymbol{x}_q$, то есть определим объект ω_μ как начало координат и вычислим новую матрицу скалярных произведений $C^\mu(N, N)$ с элементами

$$c_{ij}^\mu = c_{ij} - \mu c_{ip} - (1-\mu)c_{iq} - \mu c_{jp} - (1-\mu)c_{jq} + \\ + \mu^2 c_{pp} + 2\mu(1-\mu)c_{pq} + (1-\mu)^2 c_{qq}.$$

Если $i = j$, то

$$c_{ii}^\mu = c_{ii} - 2\mu c_{ip} - 2(1-\mu)c_{iq} + \\ + \mu^2 c_{pp} + 2\mu(1-\mu)c_{pq} + (1-\mu)^2 c_{qq}.$$

Представим объект ω_μ расстояниями до остальных объектов $d_{\mu i} = d(\omega_\mu, \omega_i),\ i = 1, \ldots, N$.

Возьмем объекты $\omega_p = \omega(\boldsymbol{x}_p)$ и $\omega_q = \omega(\boldsymbol{x}_q)$. Каждый из них представлен своими расстояниями до остальных объектов $d_{pi} = d(\omega_p, \omega_i)$ и

$d_{qi} = d(\omega_q, \omega_i)$, $i = 1, \ldots, N$. Так как диагональные элементы $c^{\mu}_{ii} = d^2(\omega_\mu, \omega_i)$ определяют квадраты расстояний от объекта ω_μ до остальных объектов $\omega_i \in \Omega$ обучающего множества

$$d^2(\omega_\mu, \omega_i) = \mu\, d^2(\omega_i, \omega_p) + (1-\mu)\, d^2(\omega_i, \omega_q) - \\ - \mu(1-\mu)\, d^2(\omega_p, \omega_q),$$

то он оказывается представлен своими расстояниями до них.

Обучающий алгоритм

Рассмотрим алгоритм Б. Н. Козинца [4].

В случае, когда заданы только матрицы $D(N,N)$ и $C(N,N)$, необходимо найти представление (5) оптимальной гиперплоскости (1) для разделения двух классов Ω_1 и Ω_2 или определить, что выпуклые оболочки множеств пересекаются.

Данный алгоритм должен найти за конечное число шагов такие две точки $\omega^+ \in \Omega_1$ и $\omega^- \in \Omega_2$ выпуклых оболочек множеств Ω_1 и Ω_2, что расстояние между ними $d(\omega^+, \omega^-)$ превышает расстояние $d(\omega_y, \omega_z)$ между неизвестными ближайшими точками $\omega_y \in \Omega_1$ и $\omega_z \in \Omega_2$ выпуклых оболочек множеств Ω_1 и Ω_2 не более, чем на величину $\varepsilon d(\omega^+, \omega^-)$, где $0 < \varepsilon < 1$ — достаточно малая наперед заданная величина. Для распознавания нового объекта ω достаточно вычислить на основе расстояний величину

$$(\omega_a \circ \omega) + a_0 = -\tfrac{1}{2}\big(d^2(\omega^+, \omega) - d^2(\omega^-, \omega)\big),$$

или на основе скалярных произведений другую величину

$$(\omega_a \circ \omega) + a_0 = (\omega^+ \circ \omega) - (\omega^- \circ \omega) - \\ - \tfrac{1}{2}\big((\omega^+ \circ \omega^+) - (\omega^- \circ \omega^-)\big).$$

Если $d(\omega^+, \omega^-) < \eta$, где $0 < \eta < 1$ — достаточно малая наперед заданная величина, то выпуклые оболочки разделяемых множеств пересекаются.

Построим такой алгоритм для расстояний и скалярных произведений.

Шаг 0. Определить пару объектов $\omega^+_0 \in \Omega_1$ и $\omega^-_0 \in \Omega_2$, например, как наиболее удаленных друг от друга из классов Ω_1 и Ω_2.

На последующих шагах все обучающие объекты $\omega_k \in \Omega_1 \cup \Omega_2$ циклически просматриваются в некотором порядке.

Шаг k. Пусть, например, $\omega_k \in \Omega_1$. Назначить объект ω^+_{k-1} началом координат. Не изменяя объект $\omega^-_k = \omega^-_{k-1}$, определить новый объект ω^+_k:

1. $\rho = \dfrac{(\omega_k \circ \omega^-_{k-1})}{(\omega^-_{k-1} \circ \omega^-_{k-1})} = \\ = \dfrac{d^2(\omega^+_{k-1}, \omega_k) + d^2(\omega^+_{k-1}, \omega^-_{k-1}) - d^2(\omega_k, \omega^-_{k-1})}{2d^2(\omega^+_{k-1}, \omega^-_{k-1})}$;

2. $\mu = \dfrac{(\omega_k \circ \omega^-_{k-1})}{(\omega_k \circ \omega_k)} = \\ = \dfrac{d^2(\omega^+_{k-1}, \omega_k) + d^2(\omega^+_{k-1}, \omega^-_{k-1}) - d^2(\omega_k, \omega^-_{k-1})}{2d^2(\omega^+_{k-1}, \omega_k)}$;

3. если $\rho \leqslant \varepsilon/2$, то $\omega^+_k = \omega^+_{k-1}$;
   если $\rho > \varepsilon/2$ и $\mu \geqslant 1$, то $\omega^+_k = \omega_k$;
   если $\rho > \varepsilon/2$ и $\mu < 1$, то $\omega^+_k = \omega_\mu$.

В п. 3 этого алгоритма объект $\omega^+_k = \omega_\mu$ представлен как своими расстояниями

$$d^2(\omega^+_k, \omega_i) = \mu\ d^2(\omega_i, \omega_k) + (1-\mu)\ d^2(\omega_i, \omega^+_{k-1}) - \\ = \mu\ (1-\mu)\ d^2(\omega_k, \omega^+_{k-1})$$

до объектов обучающего множества $\omega_i \in \Omega$, так и своими скалярными произведениями

$$(\omega^+_k \circ \omega_i) = \mu\ (\omega_i \circ \omega_k) + (1-\mu)\ (\omega_i \circ \omega^+_{k-1})$$

с объектами обучающего множества $\omega_i \in \Omega$ относительно начала координат ω_0.

Распознавание признаков

Пусть похожесть объектов из обучающего множества представлена матрицей $S(N,N)$ взаимных близостей, где $s_{ij} = s(\omega_i, \omega_j) \geqslant 0$.

Если такая матрица положительно полуопределена, то её можно считать матрицей скалярных произведений объектов $\omega_i \in \Omega$ обучающего множества в некотором неизвестном нам евклидовом пространстве размерности не выше N. Такая матрица близостей имеет N неотрицательных собственных чисел, некоторые из которых могут быть нулевыми.

Если новый объект ω располагается в том же пространстве, то нужно вычислить величину

$$(\omega_a \circ \omega) + a_0 = s_{y\omega} - s_{z\omega} - \tfrac{1}{2}(s_{yy} - s_{zz}). \qquad (6)$$

для его распознавания, где $s_{y\omega} = s(\omega_y, \omega)$ — близость объектов ω_y и ω, и $s_{z\omega} = s(\omega_z, \omega)$ — близость объектов ω_z и ω. Если близости нормализованы, т. е. $s'_{ij} = s_{ij}/\sqrt{s_{ii}s_{jj}}$, то $s_{yy} = s_{zz} = 1$. Тогда решающее правило (6) представлено в виде

$$(\omega_a \circ \omega) + a_0 = s_{y\omega} - s_{z\omega}.$$

Задача исследования множества признаков часто является самостоятельной в анализе данных. Обычно взаимосвязи множества признаков $X_1, \ldots, X_n = X(N,n)$ представлены корреляционной матрицей $R(n,n)$. Задачи анализа взаимосвязей признаков обычно формулируются как задачи группировки, корреляционного и факторного анализа. Их решение обычно нацелено на получение представителей групп признаков. Если элементы

обучающего множества $\Omega_1 \cup \Omega_2$ являются признаками $\omega_i \in \Omega$, которые представлены коэффициентами взаимной корреляции, то квадраты или модули коэффициентов корреляций определяют матрицу $S(n,n)$ взаимных нормированных близостей признаков. Следовательно, для распознавания нового признака ω нужно вычислить величину

$$(\omega_a \circ \omega) + a_0 = s(\omega^+, \omega) - s(\omega^-, \omega),$$

где элементы $\omega^+ \in \Omega_1$ и $\omega^- \in \Omega_2$ найдены алгоритмом Козинца на этапе обучения.

На текущем шаге k этого алгоритма вычисляются величины $\rho = \mu = s(\omega_k, \omega_{k-1}^-)$ и принимаются решения:

если $\rho \leqslant \varepsilon/2$, то $\omega_k^+ = \omega_{k-1}^+$;

если $\rho \geqslant 1$, то $\omega_k^+ = \omega_k$;

если $\varepsilon/2 < \rho < 1$, то $\omega_k^+ = \omega_\mu$.

В последнем случае на текущем шаге k новый элемент $\omega_k^+ = \omega_\mu$ представлен своими близостями

$$s(\omega_k^+, \omega_i) = \mu s(\omega_i, \omega_k) + (1-\mu) s(\omega_i, \omega_{k-1}^+)$$

к элементам $\omega_i \in \Omega$ обучающего множества.

Заключение

Исследование явлений различной природы на основе их взаимной похожести позволяет применять одинаковые методы анализа к множествам элементов разных типов. В предыдущей работе [5] в условиях отсутствия признакового пространства была рассмотрена задача кластер-анализа и показано, что алгоритм K-средних аналогичен известному алгоритму экстремальной группировки признаков «модуль» [7].

В данной работе рассмотрена задача обучаемого распознавания как «традиционных» объектов, так и их признаков. Предложен пример модификации известного алгоритма обучения в терминах взаимных расстояний и близостей.

В ряде случаев это оказывается более удобным, например, в отличие от метода опорных векторов, который всегда приводит к необходимости прямо решать задачу квадратичного программирования.

Получены новые экспериментальные результаты для признаков (решение задачи би-факторного анализа психологических тестов [2] как задачи распознавания) и для объектов (распознавание малонаполненных классов аминокислотных последовательностей [6]).

В первом случае следует отметить, что задача би-факторного анализа, предложенная в [2], является, по сути, задачей распознавания пяти классов психологических тестов, но попытка ее решения методами факторного анализа не опиралась на идеологию распознавания образов. Поэтому вместо построения общих факторов, представляющих заранее заданные группы психологических тестов (т. е. признаков), теперь решается задача распознавания признаков с оценкой качества полученного отделения каждой группы тестов от всех остальных групп стандартным методом теории распознавания образов (скользящий контроль). В отличие от задачи би-факторного анализа, все группы тестов распознаются безошибочно.

Во втором случае следует отметить, что конфигурации белковых макромолекул, как правило, похожи в больших группах эволюционно близких белков, а множество существенно различных пространственных структур белков значительно меньше множества всех известных белков. Поэтому проблема выявления пространственной структуры белковых макромолекул оказывается задачей распознавания. Представление объектов взаимными близостями и их погружение в метрическое пространство без его непосредственного восстановления позволяет существенно уменьшить трудоемкость скользящего контроля и улучшить результат распознавания, полученный ранее в [6]. Подробнее решение данной задачи рассматривается в [8, в данном сборнике].

Литература

[1] *Cox T. F., Cox M. A. A.* Multidimensional scaling. 2nd ed. — London: Chapman & Hall, 2001. — 328 p.

[2] *Харман Г.* Современный факторный анализ. — М.: Статистика, 1972. — 486 с.

[3] *Torgenson W. S.* Theory and Methods of Scaling. — N. Y.: John Wiley & Sons, 1958. — 460 p.

[4] *Козинец Б. Н.* Рекуррентный алгоритм разделения выпуклых оболочек двух множеств // Алгоритмы обучения распознаванию образов, М.: Сов. радио, 1973. — С. 43–50.

[5] *Двоенко С. Д.* Кластеризация элементов множества на основе взаимных расстояний и близостей // ММРО-13, М.: МАКС Пресс, 2007. — С. 114–117.

[6] *Mottl V. V., Dvoenko S. D., Seredin O. S., Kulikowski C. A., Muchnik I. B.* Featureless pattern recognition in an imaginary Hilbert space and its application to protein fold classification // 2nd Int. workshop on machine learning and data mining in pattern recognition (MLDM), Leipzig: Springer, 2001. — Pp. 322–336.

[7] *Браверман Э. М., Мучник И. Б.* Структурные методы обработки эмпирических данных. — М.: Наука, 1983. — 464 с.

[8] *Барчуков М. А., Двоенко С. Д.* Разделение малонаполненных классов методом скользящего контроля // Всероссийская конференция Математические методы распознавания образов (ММРО-14), г. Суздаль, 21–25 сент., 2009. — С. 495–498 (в данном сборнике).

Спектральная реализация метода наименьших квадратов*

Дедус Ф. Ф., Алёшин С. А., Двойнев А. И., Куликова Л. И., Махортых С. А., Панкратов А. Н., Пятков М. И., Тетуев Р. К.
ffdedus@impb.ru
Москва, МГУ имени М.В. Ломоносова

В задачах аналитического описания данных измерений предлагается реализация метода наименьших квадратов отрезками ортогональных рядов на основе классических ортогональных полиномов непрерывного или дискретного аргументов. Этот подход упрощает реализацию МНК за счет резкого сокращения объема вычислительных операций, существенного расширения границ его применимости, а получаемые результаты делает более прозрачными.

Метод наименьших квадратов (МНК) относится к наиболее важным методам теории ошибок, оценивающий неизвестные величины по результатам измерений, содержащих случайные ошибки. Он применяется также для приближенного описания заданной функции другой (более простой) функцией и часто оказывается полезным при обработке наблюдений. МНК был предложен К. Гауссом (1794–1795гг.) и Л. Лежандром (1805–1806гг.) и первоначально использовался для обработки результатов астрономических и геодезических наблюдений. Строгое математическое обоснование и установление границ содержательной применимости МНК даны А. А. Марковым (старшим) [1] и А. Н. Колмогоровым [2]. В настоящее время МНК представляет собой один из важнейших разделов математической статистики и широко используется для статистических выводов в различных областях науки и техники. Практическая ценность МНК заключается главным образом в том, что он дает в руки исследователям аналитические зависимости, позволяющие адекватно оценивать скрытые закономерности в наблюдаемых явлениях, а также проверять надежность получаемых оценок путем вычисления вероятностных характеристик с учетом соответствующих законов распределения. Следует при этом аккуратно устранить наиболее трудные вычислительные этапы метода, сделать их более прозрачными, существенно расширяя границы его применимости, и обеспечить повышенную точность.

Измерения всегда производятся с некоторыми погрешностями. Поэтому для снижения влияния ошибок измерений увеличивают их число. Классическая задача оценки неизвестных параметров приводит к принципу наименьших квадратов, сводящемуся к минимизации следующего функционала

$$\varphi = \int_a^b \left(y(x) - \sum_{n=0}^{N} a_n x^n \right)^2 dx,$$

где $y(x)$ — измеряемая зависимость, $\sum_{n=0}^{N} a_n x^n$ — полиномиальное приближение, a_n — искомое значение коэффициентов. Определение неизвестных коэффициентов a_n в функционале приводит к матричным преобразованиям с промежуточными матрицами размером порядка N. В случае, если приближающий многочлен степени n плохо описывает измеренные значения $y(x)$, то требуется повысить его степень. При этом необходимо всю процедуру определения коэффициентов a_n повторить заново. П. Л. Чебышев в середине 19-го столетия при решении задач по интерполяции предложил искать минимум следующего функционала

$$\varphi = \int_a^b \left(y(x) - \sum_{n=0}^{N} A_n \varphi_n(x) \right)^2 dx.$$

Здесь суть состоит в том, что приближающий многочлен представляет собой сумму многочленов повышающихся степеней, «специальным образом организованных», и коэффициентов A_n, определяемых по методу наименьших квадратов. В дальнейшем функции $\varphi_n(x)$ он определил как взаимно ортогональные, принадлежащие ортогональному базису Чебышева. Можно показать, что вариационная задача минимизации функционала приводит к системе уравнений, решение которой определяет значения коэффициентов A_n искомого приближающего многочлена: $\tilde{y}(x) = \sum_{n=0}^{N} A_n \varphi_n(x)$, где $\varphi_n(x)$ — ортогональные многочлены Чебышева; A_n — коэффициенты разложения $y(x)$ по ортогональным полиномам Чебышева.

В случае необходимости повышения точности аналитического описания $y(x)$, достаточно вычисления новых коэффициентов разложения A_{n+1}, A_{n+2}, и т. д. При этом ранее вычисление коэффициенты разложения остаются неизменными.

Этим предложением Чебышев фактически ввел в середине 19-го столетия в решение задач по МНК применение всех классических ортогональных полиномов, которые обладают основными свойствами [3].

*Работа выполнена при финансовой поддержке РФФИ, проекты №07-01-00564-а, №08-01-12030-офи.

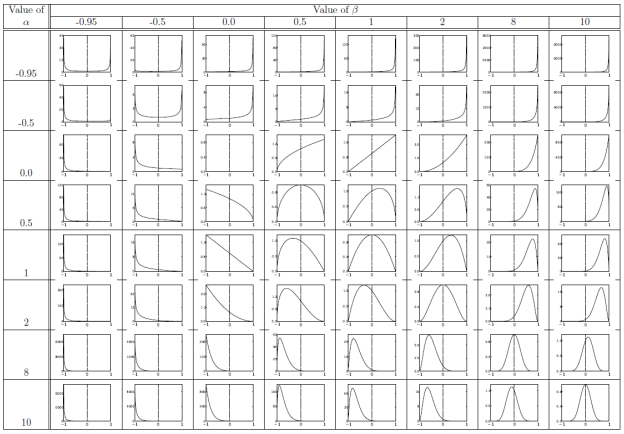


Рис. 1. Разнообразные весовые функции классических ортогональных базисов.

В настоящее время все классические ортогональные полиномы и функции непрерывного и дискретного аргументов представлены в виде таблиц и подробно описаны в работах [4, 5].

Их разнообразные свойства хорошо изучены и позволяют с помощью разработанных адаптивных процедур [4] получать аналитические описания различных измеренных характеристик с квазиоптимальной точностью.

За счет разнообразных весовых функций, определенных однозначно для каждого конкретного ортогонального базиса, можно практически всегда удовлетворять различным требованиям к особенностям описания поступающих сигналов (см. рис.1).

В диссертации [7] разработаны алгоритмы и программы сверхглубокого разложения сигналов для ряда ортогональных базисов из числа классических. При этом число членов разложения в указанных случаях достигает 10000 и выше. Глубокое разложение сигналов применяется в тех случаях, когда необходимо оценить сложный сигнал как целое. До сих пор найдено несколько приложений для глубокой аппроксимации. Одно приложение связано с оценкой фрактальности сигналов. В частности, предложен метод оценки фрактальной размерности сигналов [8]. Другое приложение, простимулировавшее развитие подхода, связано с обработкой сигналов ядерного магнитного резонанса. Сигналы ЯМР представляют собой сложные затухающие процессы, для которых наиболее подходящим базисом являются функции Лагерра.

Вопросы, связанные с устойчивым вычислением функций Лагерра и Эрмита, рассмотрены в [9]. Предложено также использовать матрицу Грама для контроля ортогональности вычисляемых функций базиса. Показано, что для функций непрерывного аргумента ортогональность имеет место при использовании квадратурных формул Гаусса. Таким образом, на практике возможность получения глубокого разложения сигналов определяется фактически возможностью устойчивого вычисления ортогональных функций.

Важная особенность аналитической аппроксимации данных на основе классических ортогональных базисов состоит в том, что члены ортогональных рядов линейно независимы, и вся информация об описываемых сигналах сосредоточена в коэффициентах разложения. Поэтому большое значение придавалось развитию алгебры для преобразований в пространстве коэффициентов разложения.

Исследование алгебраических преобразований в пространстве коэффициентов разложения на основе аппарата линейной алгебры проведено в диссертации [7]. В частности, было обосновано вычисление таких обратных операций, как деление и вычисление квадратного корня на основе последовательного рассмотрения операции умножения ортогональных рядов, для чего было введено понятие и изучены свойства оператора умножения на функцию в пространстве коэффициентов разложения [10].

Ранее применение спектральных методов часто сводилось лишь к анализу данных и хранению значений в сжатом виде. Действительно, когда требовалось преобразовать хранимую функцию, следовало восстановить все ее значения и, произведя преобразование численно, заново разложить получившуюся функцию спектрально, что, конечно же, сказывалось на скорости и точности вычислений. Указанный недостаток сдерживал распространение спектральных методов на случай практических задач, сопряженных с необходимостью быстрой и точной обработки сигналов в режиме реального времени, а также задач, связанных с обработкой значительных массивов данных. Такие требования возникают при распознавании визуальных образов, в задачах биоинформатики и др.

В диссертации [11] разработана алгебра преобразований коэффициентов разложения, которая позволяет реализовать полную обработку сигналов в пространстве коэффициентов разложения. В отличие от вычислительных приемов, применявшихся для этой цели ранее, этот способ приводит к алгоритмам с линейной сложностью относительно размера спектра и также ускоряется при реализации на ЭВМ с параллельной структурой. Преимуществом является еще то, что способ обобщен на случай всех ортогональных полиномов решений гипергеометрического уравнения и некоторые важные их модификации. Для этого потребовалось поставить и успешно разрешить задачу о нахождении рекуррентных формул особого вида, определив при этом ряд ранее неизвестных аналитических соотношений [11].

Далее мы будем использовать понятие внутриспектральных преобразований, которое мы формально определим здесь как численное изменение

спектра некоторой функции (оригинала), приводящее к спектру новой функции (результата), при этом связь между функцией-результатом и функцией-оригиналом может быть выражена аналитически. Теперь сформулируем достаточное условие существования быстрых алгоритмов линейной сложности для реализации внитриспектральных преобразований на ЭВМ:

Теорема 1. *Пусть $f(x)$ — некоторая функция пространства $L^2_\rho(a,b)$ и $\{\varphi_n(x)\}$ — система ортогональных функций того же пространства, такая что:*

$$f(x) = \sum_{n=0}^{N+1} C_n \varphi_n(x),$$

Пусть далее $A(\cdot)$ — некоторый линейный оператор, такой, что $A(f) \in L^2_\rho(a,b)$ и

$$A(f) = \sum_{n=0}^{N+q} C^*_n \varphi_n(x).$$

Тогда если для каждого $\varphi_{n+1}(x)$ существует рекуррентное соотношение вида

$$A(\varphi_{n+1}) = F_{n+1}\big(A(\varphi_n), \dots, A(\varphi_{n-d}), \varphi_{n+q}, \dots, \varphi_{n-p}\big),$$

где $F_n(\dots)$ — линейная форма, то существует алгоритм линейной временной сложности для вычисления коэффициентов разложения $\{C^*_n\}$ при известных $\{C_n\}$.

На практике для организации внутриспектральных преобразований применяется вычислительный прием названный *методом спектральных каскада и диффузии*. Весь процесс организуется в виде двух этапов: каскада и диффузии, названных так по принципу «переноса спектральных значений» вдоль ряда и способу организации вычислений. На первом этапе (каскад) требуется последовательно добавлять значения старших коэффициентов ряда к младшим начиная с самого старшего, а на втором (диффузия) значения просто «смешиваются» в соседних ячейках ряда. Иначе вычислительный процесс можно описать как организацию взвешенного суммирования в нейронных сетях.

В настоящее время ведется работа по оценке надежности получаемых результатов в виде вычисления вероятности ошибок, превосходящих определенные пределы, на основе законов распределения Стьюдента и χ^2. В частности, показано, что с ростом точности аппроксимации длина доверительных интервалов для коэффициентов разложения уменьшается и, хотя при увеличении надежности длина этих интервалов растет, но не так сильно.

В заключение необходимо отметить, что предлагаемая реализация метода наименьших квадратов и его расширение за счет алгебры не лишены трудностей. В частности, на настоящий момент нет такой реализации классических полиномиальных базисов дискретного аргумента, которая бы позволяла получать такое же глубокое разложение, как в случае базисов непрерывного аргумента. Что касается алгебры спектральных преобразований, то в настоящий момент она ограничена исследованием и реализацией преимущественно линейных операций. Можно надеяться, что эти трудности послужат стимулом для дальнейшего развития спектрально-аналитического метода.

Литература

[1] *Марков А. А.* Исчисление вероятностей. — Л: Госиздательство, 1924. — 589 с.

[2] *Колмогоров А. Н.* Теория вероятностей и математическая статистика. — М: Наука, 2005. — 582 с.

[3] *Чебышев П. Л.* Вопросы о наименьших величинах, связанные с приближенным представлением функций. — Л: Госиздательство, 1947. — 580 с.

[4] *Дедус Ф. Ф., Куликова Л. И., Панкратов А. Н., Тетуев Р. К.* Классические ортогональные базисы в задачах аналитического описания и обработки информационных сигналов. Учебное пособие. — М: Издательский отдел ВМиК МГУ, 2004. — 147 с.

[5] *Дедус Ф. Ф., Махортых С. А., Устинин М. Н., Дедус А. Ф.* Обобщенный спектрально-аналитический метод обработки информационных массивов. Задачи анализа изображений и распознавания образов. — М: Машиностроение, 1999.— 357 с.

[6] *Худсон Д.* Статистика для физиков. — М: Мир, 1970.— 243 с.

[7] *Панкратов А. Н.* Алгебраические операции над ортогональными рядами в задачах обработки данных — дисс. на соискание уч. степ. к.ф.-м.н., М: ВЦ РАН, 2004. — 105 с.

[8] *Махортых С. А., Панкратов А. Н.* О спектральном разложении нерегулярных кривых // 1-я всеросс. конф. «Спектральные методы обработки информации в научных исследованиях» (Спектр-2000, Пущино) — Москва, 2000, С. 44.

[9] *Панкратов А. Н., Бритенков А. К.* Обобщенный спектрально-аналитический метод: проблемы описания цифровых данных семействами ортогональных полиномов // Вестник НГУ им. Н. И. Лобачевского. Серия Радиофизика. Вып. 1(2). — Н. Новгород: Изд-во ННГУ, 2004. — С.5–14.

[10] *Панкратов А. Н.* О реализации алгебраических операций над рядами ортогональных функций // ЖВМиМФ, 2004. — Т. 44, № 12. — С. 2121–2127.

[11] *Тетуев Р. К.* Алгебра спектральных преобразований в задачах обработки данных — дисс. на соискание уч. степ. к.ф.-м.н., М: ВЦ РАН, 2007. — 111 с.

[12] *Тетуев Р. К., Дедус Ф. Ф.* Классические ортогональные полиномы. Применение в задачах обработки данных. — препринт, М: 11-формат, 2007. — 60 с.

Отбор эталонов, основанный на минимизации функционала полного скользящего контроля*

Иванов М. Н., Воронцов К. В.
voron@ccas.ru
Москва, ВМиК МГУ, Вычислительный центр РАН

В статье рассматривается алгоритмы отбора эталонных объектов для метода K ближайших соседей. Отбор эталонов основан на оптимизации функционала полного скользящего контроля, вычисляемого по эффективным комбинаторным формулам. Оптимизации приводит одновременно к сокращению выборки до минимального достаточного числа эталонов, устранению (цензурированию) шумовых объектов и улучшению обобщающей способности.

Выделение эталонных объектов из обучающей выборки обычно преследует несколько целей. В методе K ближайших соседей оно позволяет радикально сократить объём хранимых данных, повысить скорость классификации, найти и удалить «шумовые» (нетипичные) объекты, повысить качество классификации (обобщающую способность).

Известные методы последовательного отбора эталонов Stolp, λ-Stolp [1] и FRiS-Stolp [2], основаны, фактически, на оценивании локальных плотностей классов в каждом объекте и вычислении отношений этих оценок. Эти методы неплохо зарекомендовали себя на практике, однако остаются открытыми теоретические вопросы: какой функционал они минимизируют, почему они обладают хорошей обобщающей способностью, почему в них использованы именно такие эвристики, и какие из многочисленных возможных вариантов этих эвристик могли бы работать ещё лучше.

В [3] предлагалось отбирать эталоны путём минимизации функционала *полного скользящего контроля* (complete cross validation, CCV) [4, 5], который *по построению* характеризует обобщающую способность. Возникающие при этом вычислительные проблемы решались с помощью эффективных комбинаторных формул. В [5] такие формулы были получены для случая двух классов, $|Y| = 2$, в [3] был предложен жадный алгоритм отсева неэталонных объектов для случая $K = 1$.

В данной работе эти результаты обобщаются и улучшаются по нескольким направлениям. Предлагается способ быстрого пересчёта функционала CCV в случае произвольных K и $|Y|$. Предлагаются и сравниваются две стратегии отбора эталонов — последовательный отсев неэталонных объектов и последовательное добавление эталонов. Для эффективного построения и поиска прямых и обратных окрестностей объектов используются метрические деревья, что позволяет применять алгоритм на выборках из десятков тысяч объектов.

*Работа поддержана РФФИ (проект №08-07-00422) и программой ОМН РАН «Алгебраические и комбинаторные методы математической кибернетики и информационные системы нового поколения».

Функция вклада

Задано множество объектов X, конечное множество ответов (классов) Y, обучающая выборка $X^L = \{x_1, \ldots, x_L\}$. Существует зависимость $y^*\colon X \to Y$, известная только на объектах выборки, $y^*(x_i) = y_i,\ i = 1, \ldots, L$.

На множестве X задана функция расстояния $\rho\colon X \times X \to [0, \infty)$. Относительно произвольного объекта $u \in X$ все объекты заданного подмножества $\Omega \subseteq X^L$ можно отранжировать в порядке возрастания расстояний $\rho(u, x_i)$. Будем называть j-й по порядку элемент в этом ряду *j-м соседом объекта u среди Ω*, и обозначать его через $x_{j,u}(\Omega)$, а ответ $y^*\big(x_{j,u}(\Omega)\big)$ на этом объекте — через $y_{j,u}(\Omega)$. При этом считается, что $x_{0,u} = u$, а $y_{0,u}(\Omega) = y^*(u)$.

Алгоритм K ближайших соседей по эталонному множеству Ω относит произвольный объект $u \in X$ к тому классу, которому принадлежит большинство из K ближайших соседей u среди Ω:

$$a(u; K, \Omega) = \arg\max_{y \in Y} \sum_{j=1}^{K} \big[y_{j,u}(\Omega) = y\big]. \qquad (1)$$

Пусть $X_n^l \sqcup X_n^k = X^L,\ n = 1, \ldots, N,\ N = C_L^k$ — всевозможные разбиения полной выборки X^L на обучающую X_n^l и контрольную X_n^k подвыборки длиной l и $k = L - l$ соответственно. *Функционал полного скользящего контроля* CCV [5] определяется как средняя частота ошибок алгоритма $a(u; K, \Omega)$ на контрольных подвыборках:

$$Q(\Omega) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{k} \sum_{x \in X_n^k} \big[a(x; K, X_n^l \cap \Omega) \neq y(x)\big].$$

Значение функционала $Q(\Omega)$ показывает, насколько хорошо классифицируются контрольные объекты по обучающим объектам из эталонного множества Ω. Можно также сказать, что $Q(\Omega)$ характеризует обобщающую способность множества эталонов Ω. Отметим, что суммирование производится по *всем* возможным разбиениям выборки на обучение и контроль, а длина контроля k заранее не фиксирована и может быть выбрана произвольно. Поэтому функционал CCV наиболее устойчив среди всех разновидностей скользящего кон-

троля, включая такие распространённые на практике эмпирические оценки, как «hold-out», «leave-one-out», «q-fold CV» и др. [4]. Проблема заключается в том, что непосредственное вычисление $Q(\Omega)$ практически невозможно уже при $k \geqslant 3$, т. к. число разбиений N огромно. Однако для алгоритма K ближайших соседей существуют эффективные способы его вычисления [5, 6], а в данной работе строятся ещё и эффективные алгоритмы оптимизации $Q(\Omega)$ по множеству эталонов Ω.

Определение 1. *Окрестностью $\Psi(u, m, \Omega)$ m-го порядка объекта $u \in X^L$ назовем множество его первых m соседей $x_{1,u}(\Omega), \ldots, x_{m,u}(\Omega)$.*

Рассмотрим множество разбиений окрестности на два подмножества, $\Psi(x_i, m, \Omega) = \Psi^l \sqcup \Psi^k$:

$$\Lambda(x_i, m, \Omega) = \big\{(\Psi^l, \Psi^k) \colon |\Psi^l| = K,\ x_{m,x_i} \in \Psi^l\big\}.$$

Число таких разбиений, $|\Lambda(x_i, m, \Omega)| = C_{m-1}^{K-1}$.

Определение 2. *Функцией вклада (далее просто вкладом) объекта x_i m-го порядка называется действительная функция $T(x_i, m, \Omega)$, выражающая долю разбиений из множества $\Lambda \equiv \Lambda(x_i, m{+}K{-}1, \Omega)$, при которых ответ $a(x_i; K, \Psi^l)$ неверен:*

$$T(x_i, m, \Omega) = \frac{\#\big\{(\Psi^l, \Psi^k) \in \Lambda \colon a(x_i; K, \Psi^l) \neq y_i\big\}}{|\Lambda|}.$$

Следующая теорема раскрывает связь функции вклада и функционала CCV $Q(\Omega)$.

Теорема 1. *Для алгоритма K ближайших соседей верно соотношение*

$$Q(\Omega) = \sum_{i=1}^{L} \sum_{m=1}^{k} \tilde{C}(m)\, T(x_i, m, \Omega), \qquad (2)$$

$$\tilde{C}(m) = \frac{1}{k} \frac{C_{L-m-K}^{k-m} C_{m+K-2}^{K-1}}{C_L^l}.$$

Из определения функции вклада следует, что ее непосредственное вычисление требует полного перебора множества Λ, состоящего из C_{m+K-2}^{K-1} разбиений. Этот способ на практике вряд ли осуществим. Рассмотрим способ эффективного вычисления функции вклада.

Введём бинарную функцию, равную единице тогда и только тогда, когда ответ на m-м соседе объекта x_i не совпадает с правильным ответом y_i:

$$r_m(x_i, \Omega) = \big[y_{m,i}(\Omega) \neq y_i\big].$$

Через $n_i(y, m, \Omega)$ обозначим количество объектов среди окрестности m-го порядка объекта x_i, у которых ответ совпадает с заданным $y \in Y$:

$$n_i(y, m, \Omega) = \sum_{t=1}^{m} \big[y_{t,x_i}(\Omega) = y\big].$$

Теорема 2. *Для функции вклада $T(x_i, m, \Omega)$ верно соотношение:*

$$1 - T(x_i, m, \Omega) = \sum_{j=1}^{K} \frac{C_{n_i(y_i, m+K-2, \Omega)}^{j+r_m(x_i,\Omega)-1}}{C_{m+K-2}^{K-1}} P\big(j{-}1, Y\backslash\{y_i\}, K{-}j\big),$$

где функция $P(t, \tilde{Y}, s)$ определяется рекуррентными соотношениями:
1) если множество $\tilde{Y}$ одноэлементно, $\tilde{Y} = \{\tilde{y}\}$, то

$$P(t, \tilde{Y}, s) = [s \leqslant t]\, C_{n_i(\tilde{y}, m+K-2, \Omega)}^{s+r}.$$

2) иначе, для любого $\tilde{y} \in \tilde{Y}$

$$P(t, \tilde{Y}, s) = \sum_{v=0}^{t} C_{n_i(\tilde{y}, m+K-2, \Omega)}^{v+r} P(t, \tilde{Y}\backslash\{\tilde{y}\}, s{-}v),$$

где $r = \big[\, \tilde{y} \neq y_{m+K-1,\, x_i}(\Omega) \,\big]$.

Эффективное вычисление вкладов. Теорема 2 позволяет эффективно вычислять функцию вклада объектов. Зафиксируем объект x_i. При некотором $\tilde{y} \neq y_i$ вычислим $(K{+}1)^2$ значений

$$P(t, \{\tilde{y}\}, s) : \quad 0 \leqslant t, s \leqslant K.$$

Затем для некоторого $\tilde{y}' \in Y\backslash\{y_i, \tilde{y}\}$ вычислим по рекуррентному соотношению $(K+1)^2$ значений

$$P(t, \{\tilde{y}, \tilde{y}'\}, s) : \quad 0 \leqslant t, s \leqslant K,$$

и так далее, пока не вычислим значения функции $P(t, Y\backslash\{y_i\}, s))$. Наконец, вычислим по теореме 2 функцию вклада $T(x_i, m, \Omega)$ для всех $1 \leqslant m \leqslant k$. Сложность её вычисления для одного объекта составляет $O(|Y|{\cdot}K^2)$. Теорема 1, в свою очередь, позволяет эффективно вычислить функционал полного скользящего контроля.

Два частных случая

В частных случаях — для метода одного ближайшего соседа и для двух классов теорема 1 приводит к ранее известным результатам [3, 5, 6].

Теорема 3. *Для алгоритма одного ближайшего соседа функция вклада m-го порядка — это бинарная величина, равная единице тогда и только тогда,когда ответ на m-м соседе объекта x_i не равен ответу на самом объекте:*

$$T(x_i, m, \Omega) = r_m(x_i, \Omega) = \big[y_{m,i}(\Omega) \neq y_i\big].$$

Обозначим через $n(m, i)$ число первых $m-1$ соседей объекта x_i, у которых ответ $y_{m,i}(\Omega)$ совпадает с y_i, через $\bar{n}(m, i)$ — число первых $m-1$ соседей, у которых ответ $y_{m,i}(\Omega)$ не совпадает с y_i.

Теорема 4. *Для алгоритма K ближайших соседей при $|Y| = 2$ верно соотношение:*

$$T(x_i, m, \Omega) = \sum_{t=\frac{K+1}{2}-r_m(x_i,\Omega)}^{K-1} C^t_{\bar{n}(m+t-1,i)} C^{K-1-t}_{n(m+t-1,i)}.$$

Локальность функции вклада. Из определения функции вклада следует, что $T(x_i, m, \Omega)$ зависит только от объектов, попадающих в окрестность $(m+K-1)$-го порядка объекта x_i. Таким образом, функция вклада носит локальный характер и отражает информацию об объектах, находящихся в непосредственной близости к данному.

Определение 3. *Обратной окрестностью m-го порядка объекта x_i называется множество объектов $\bar{\Psi}(x_i, m, \Omega)$, для которых объект x_i является не далее чем m-м соседом:*

$$\bar{\Psi}(x_i, m, \Omega) = \big\{u \in X^L \colon x_i \in \Psi(u, m, \Omega)\big\}.$$

Свойство локальности функции вклада позволяет эффективно пересчитывать функционал полного скользящего контроля $Q(\Omega)$ при изменении множества эталонов Ω.

Утверждение 5. *Для пересчёта функционала $Q(\Omega)$ при удалении объекта x_i из множества эталонов Ω или при добавлении объекта x_i во множество эталонов Ω достаточно информации об объектах из его обратной окрестности $(k + K - 1)$-го порядка. Вклады остальных объектов не изменятся.*

Это утверждение следует из теоремы 1 и свойства локальности функции вклада.

Метод отбора эталонов на основе минимизации функционала полного скользящего контроля

Предлагаются два метода поиска оптимального множества эталонов Ω для алгоритма K ближайших соседей. Первый метод основан на *стратегии удаления* объектов (Алгоритм 1), второй — на *стратегии добавления* (Алгоритм 2). В обоих методах при добавлении или удалении эталона функционал полного скользящего контроля Q полностью не пересчитывается. Пересчету подлежат только вклады объектов, расположенных по близости от удаляемого или добавляемого объекта.

Последовательное удаление объектов проходит две стадии. На первой стадии из обучающей выборки исключаются *шумовые объекты*. При их удалении значение функционала Q уменьшается. На второй стадии удаляются неинформативные (периферийные) объекты, при этом значение функционала Q практически не изменяется. При переходе от первой стадии ко второй на шаге 13 Алгоритма 1 условие $\delta < 0$ необходимо заменить на $\delta \leqslant \delta_0$, где δ_0 — малый положительный параметр.

Алгоритм 1. Стратегия удаления шумовых и неинформативных объектов.

Вход: выборка X^L;
Выход: множество эталонов $\Omega \subseteq X^L$;
1: инициализация: $\Omega := X^L$; $Q := 0$;
2: **для всех** $i = 1, \ldots, L$ и $m = 1, \ldots, k$
3: вычислить функцию вклада $T(x_i, m)$;
4: $Q := Q + T(x_i, m)\tilde{C}(m)$;
5: **повторять**
6: flag := *false*;
7: **для** $i = 1, \ldots, L$
8: построить $\bar{\Psi}(x_i, k + K - 1, \Omega)$;
9: $\delta := 0$;
10: **для всех** $u \in \bar{\Psi}$, $m = 1, \ldots, k$
11: $T'(u, m)$:= вклад m-го порядка объекта u по эталонному множеству $\Omega\backslash\{u\}$;
12: $\delta := \delta + T'(u, m) - T(u, m)$;
13: **если** $\delta < 0$ **то**
14: $Q := Q + \delta$; $\Omega := \Omega\backslash\{x_i\}$; flag := *true*;
15: **для всех** $u \in \bar{\Psi}$ и $m = 1, \ldots, k$
16: $T(u, m) := T'(u, m)$;
17: **пока** flag; // пока остаются шумы.

Алгоритм 2. Стратегия добавления эталонов.

Вход: выборка X^L;
Выход: множество эталонов $\Omega \subseteq X^L$;
1: $\Omega :=$ выбрать $(K + 1)$ объектов; $Q := 0$;
2: **для всех** $i = 1, \ldots, L$ и $m = 1, \ldots, k$
3: вычислить функцию вклада $T(x_i, m)$;
4: $Q := Q + T(x_i, m)\tilde{C}(m)$;
5: **повторять**
6: flag := *false*;
7: **для** $i = 1, \ldots, L$
8: построить $\bar{\Psi}(x_i, k + K - 1, \Omega)$;
9: $\delta := 0$;
10: **для всех** $u \in \bar{\Psi}$ и $m = 1, \ldots, k$
11: $T'(u, m)$:= вклад m-го порядка объекта u по эталонному множеству $\Omega \cup \{u\}$;
12: $\delta := \delta + T'(u, m) - T(u, m)$;
13: **если** $\delta < 0$ **то**
14: $Q := Q + \delta$; $\Omega := \Omega \cup \{x_i\}$; flag := *true*;
15: **для всех** $u \in \bar{\Psi}$ и $m = 1, \ldots, k$
16: $T(u, m) := T'(u, m)$;
17: **пока** flag;

Стратегия добавления противоположна стратегии удаления. Алгоритм начинает работу с того, что заносит в текущее множество эталонов Ω некоторые $K+1$ объектов. Далее алгоритм начинает последовательное добавление объектов в Ω. Если в очередной раз невозможно добавить ни одного объекта так, чтобы функционал $Q(\Omega)$ уменьшился, то алгоритм завершает свою работу.

Метрические деревья. Для эффективного пересчета функционала качества необходимо

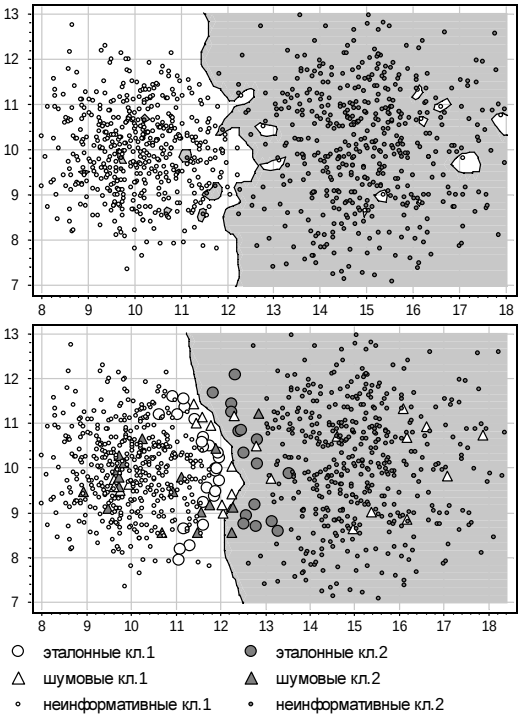


Рис. 1. Сверху: модельная задача классификации: 1000 объектов, алгоритм 1NN. Снизу: результат Алгоритма 1: отобрано 26 эталонов класса 1 и 16 — класса 2.

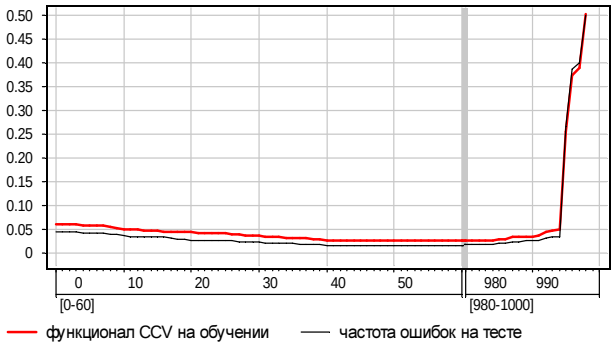


Рис. 2. Зависимость функционала CCV $Q(\Omega)$ от количества удаленных объектов $L - |\Omega|$ для Алгоритма 1.

быстро находить объекты окрестности и обратной окрестности. Для этого необходима структура данных, которая позволяла бы эффективно строить, находить и обновлять прямые и обратные окрестности k-го порядка для любого объекта $x_i \in X^L$. Наиболее подходят для этой цели *метрические деревья*, предназначенные для индексации множества объектов, заданных попарными расстояниями [7]. В среднем каждый запрос обрабатывается метрическим деревом за время порядка $O(\log L)$. Использование метрических деревьев позволяет существенно ускорить процесс отбора эталонов. На шаге 8 Алгоритмов 1 и 2 окрестность $\Psi(x_i, k+K-1, \Omega)$ строится именно с помощью метрического дерева.

Эксперименты и выводы

Модельная двумерная выборка состояла из $L = 1000$ объектов двух классов, полученных из сферических гауссовских распределений с дисперсиями 1 и 2 и расстоянием между центрами 5. Рассматривался алгоритм 1 ближайшего соседа, рис. 1. Оба Алгоритма, 1 и 2, удаляют все шумовые объекты и оставляют, соответственно 42 и 64 эталона. Эталоны выстраиваются в каждом классе вдоль границы, на некотором удалении от неё.

На рис. 2 показана зависимость функционала $Q(\Omega)$ от количества удалённых объектов $L - |\Omega|$. Левый участок [0-60] соответствует удалению шумов (около 40 объектов). Правый участок [980-1000] показывает, что число критически важных эталонов равно 6, но увеличение числа эталонов до 16 позволяет уменьшить частоту ошибок на тестовой выборке с 3,6% до 2,0%. На рис. 1 (снизу) выделено 42 эталона, которые обеспечивают минимум $Q(\Omega)$ и одновременно минимальную частоту ошибок на тестовой выборке 1,8%. Длинный средний участок [60-980], не показанный на рис. 2, соответствует удалению неинформативных объектов, на нём значение $Q(\Omega)$ почти постоянно. Тонкой линией отложена частота ошибок алгоритма $a(x; 1, \Omega)$ на независимой тестовой выборке из 2000 объектов, взятых из того же распределения. Хорошо видно, что она изменяется синхронно с функционалом $Q(\Omega)$. Это означает, что отбор эталонных объектов вообще не подвержен переобучению.

Литература

[1] *Загоруйко Н. Г.* Прикладные методы анализа данных и знаний. — Новосибирск: ИМ СО РАН, 1999.

[2] *Борисова И. А., Дюбанов В. В., Загоруйко Н. Г., Кутненко О. А.* Сходство и компактность // Всеросс. конф. ММРО-14. — М. МАКС Пресс, 2009. — 89–93 (в настоящем сборнике).

[3] *Воронцов К.В., Колосков А. В.* Профили компактности и выделение опорных объектов в метрических алгоритмах классификации // Искусственный интеллект, Донецк, 2006. — С. 30–33.

[4] *Kohavi R.* A study of cross-validation and bootstrap for accuracy estimation and model selection // 14th International Joint Conference on Artifitial Intelligence, Quebec, Canada, 1995. — Pp. 1137–1145.

[5] *Mullin M. Shukhankar R.* Complete cross-validation for nearest neighbor classifiers // International Conference on Machine Learning, 2000, Pp. 639–646.

[6] *Воронцов К. В.* Комбинаторный подход к оценке качества обучаемых алгоритмов // Математические вопросы кибернетики. — М.: Физматлит, 2004. — Т. 13. — С. 5–36.

[7] *Лившиц Ю. В.* Branch and bound: algorithms for nearest neighbour search
`simsearch.yury.name/russir/01nncourse.pdf`

О градиентном поиске логических закономерностей классов с линейными зависимостями*

Исходжанов Т. Р., Рязанов В. В.
rvv@ccas.ru
Москва, Вычислительный Центр РАН,
Московский физико-технический институт

В настоящее время достаточно хорошо развиты логические методы анализа прецедентных данных и логические методы распознавания (тестовый алгоритм, алгоритмы вычисления оценок, алгоритмы с представительными наборами, алгоритмы голосования по системам логических закономерностей, бинарные решающие деревья) [1, 2, 3, 4]. В случае дискретных признаков основная вычислительная задача состоит в нахождении тупиковых покрытий строк некоторой бинарной матрицы (матрицы сравнения) столбцами матрицы.

В случае вещественнозначных признаков основная задача поиска логических закономерностей классов (рассматривается стандартная задача распознавания по прецедентам с l классами, n признаками и m объектами обучения) состоит в нахождении таких предикатов

$$P^{\Omega\ c,d}(x) = \bigwedge_{i\in\Omega} \big[c_i \leqslant x_i \leqslant d_i\big], \quad \Omega \subset \{1,\ldots,n\},$$

(в случаях $c_i = -\infty$ или $d_i = +\infty$ рассматриваем соответствующие односторонние неравенства), которые принимают значение 1 на максимально возможном числе обучающих объектов некоторого класса K_λ, $\lambda \in \{1,\ldots,l\}$, и значение 0 на всех объектах остальных классов, при этом мощность Ω минимальна. В настоящее время созданы эффективные алгоритмы поиска систем логических закономерностей классов, основанные на поиске максимальных совместных подсистем систем линейных неравенств, комбинаторном или генетическом подходах [5]. Ясно, что случаи существования линейных зависимостей между признаками являются «неудобными» для логических алгоритмов с логическими закономерностями приведенного выше вида. Например, в случае линейной отделимости классов гиперплоскостью $\sum\limits_{j=1}^{n} c_i x_i = 1$ предикаты

$$P_1(\mathbf{x}) = \left[\sum_{j=1}^{n} c_i x_i \geqslant 1\right] \quad \text{и} \quad P_2(\mathbf{x}) = \left[\sum_{j=1}^{n} c_i x_i \leqslant 1\right]$$

будут выполняться на всех объектах своего (и только своего) класса. Если же объекты классов расположены вблизи гиперплоскости, то несложно привести пример, когда каждая логическая закономерность выполняется лишь на одном объекте, а процент правильно распознанных объектов в режиме скользящего контроля будет равен нулю. Таким образом, представляет интерес поиск логических закономерностей классов вида

$$P(\mathbf{x}) = \bigwedge_{t\in T}\left[\sum_{j=1}^{n} a_{ij}x_j + a_{t0} \geqslant 0\right], \ \ T = \{1,\ldots,N\},$$

являющихся естественным обобщением рассмотренных выше.

В основу алгоритма положим язык общей схемы многослойной нейронной сети, модель искусственного $\sum\prod$ нейрона [6] и градиентный метод обучения сети. Рассмотрим общую схему распознавания с помощью двухслойной нейронной сети прямого действия (каждый слой будем выделять соответственно множеству его связей с нейронами предыдущего слоя и их весам) и сигмоидной функцией активации. Состояния нейронов первого уровня являются линейными функциями от признаков — выходы фактически принимают бинарные значения (что можно обеспечить выбором значений параметра сигмоиды). Нейроны второго уровня принимают фактически на вход бинарные переменные как произведения выходных сигналов нейронов предыдущего уровня, функция выхода есть пороговая функция. Схема данной сети представлена на рис. 1.

Запишем основные преобразования входного сигнала $\boldsymbol{x}$ в выходной $\boldsymbol{u}$:

$$u_\alpha = f(V_\alpha) = f\Big(\sum_{\beta=1}^{M} \nu_{\beta\alpha}\,\varphi_\beta(\boldsymbol{y}) + \nu_{0\alpha}\Big),$$

где $\varphi_\beta(\boldsymbol{y}) = \prod\limits_{i\in I_\beta} y_i = \prod\limits_{i\in I_\beta} f\Big(\sum\limits_{\delta=1}^{M} \omega_{\delta i}x_\delta + \omega_{0i}\Big)$, I_β — совокупность номеров нейронов первого уровня, соединенных с нейроном β промежуточного слоя нейронов Π, $f(\cdot)$ — функция активации, например, $f(g) = (1+e^{-tg})^{-1}$ или $f(g) = \frac{1}{2}(1+\operatorname{th} g)$, $t > 0$.

Классификация некоторого вектора признаков $\boldsymbol{x}$ осуществляется по правилу: $\boldsymbol{x} \in K_\alpha$, если $u_\alpha \geqslant \frac{1}{2}$ и $\boldsymbol{x} \notin K_\alpha$ в противном случае.

Для классификации необходимо выбрать функцию активации, задать число нейронов (N) первого и (M) промежуточного слоев, множества I_β, значения весовых параметров $\omega_{\delta i}$, $\nu_{\beta\alpha}$. Пусть структурные параметры фиксированы, и задача состоит

*Работа выполнена при поддержке РФФИ (проекты №08-01-00636, №08-01-90016 бел, №08-01-90427 укр), Целевой программы №2 Президиума РАН, Целевой программы №2 Отделения математических наук РАН.

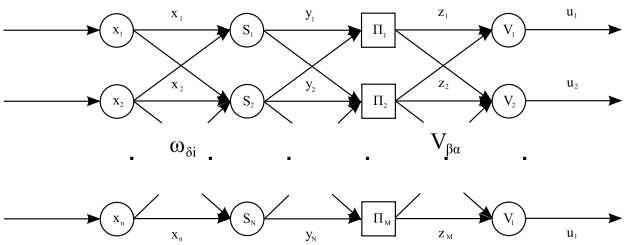


Рис. 1. Схема многослойной нейронной сети.

в поиске лишь параметров $\omega_{\delta i}$, $\nu_{\beta\alpha}$. Тогда для минимизации стандартной целевой функции ошибки

$$E(\omega_{\delta i}, \nu_{\beta\alpha}) = \frac{1}{2}\sum_{j,\alpha}\big(u_\alpha(x_j) - u^0_\alpha(x_j)\big)^2$$

при подаче сигналов обучающей выборки $\boldsymbol{x}_j$, $j = 1, \ldots, m$, может быть использован подходящий метод градиентного спуска (градиенты в данной задаче вычисляются аналитически).

К настоящему времени получены предварительные положительные экспериментальные результаты при двух вариантах выбора множеств I_β, $\beta = 1, \ldots, M$:

$$\{I_\beta\}_2 = \big\{I \subset \{1, \ldots, N\} \colon |I| \leqslant 2\big\};$$
$$\{I_\beta\}_3 = \big\{I \subset \{1, \ldots, N\} \colon |I| \leqslant 3\big\}.$$

Представляются перспективными и другие схемы, когда число сомножителей в нейронах промежуточного слоя последовательно увеличивается. Отметим, что найденные в результате обучения функции $\sum\limits_{\beta=1}^{M} \nu_{\beta\alpha}\varphi_\beta(\boldsymbol{y}) + \nu_{0\alpha}$ являются прямыми аналогами оценок за соответствующие классы в моделях голосования по системам логических закономерностей, а $\varphi_\beta(\boldsymbol{y})$ — гладкими аппроксимациями логических закономерностей искомого вида. Естественным является и другой путь практического поиска логических закономерностей с линейными функциями признаков. Сначала находятся по обучающей выборке линейные функции, разделяющие «существенную» часть объектов разных классов. Значения данных функций используются как признаки, дополнительные к исходным. В расширенном признаковом пространстве далее находятся логические закономерности классов. Данный путь является несомненно перспективным, поскольку в настоящее время имеются эффективные методы поиска логических закономерностей, а расширение признакового пространства сопровождается появлением информативных признаков. Достоинством изложенного в настоящей работе подхода является возможность в рамках одной оптимизационной задачи одновременного нахождения линейных функций признаков, информативных конъюнкций (логических закономерностей) и весовых коэффициентов при голосовании.

Литература

[1] *Дмитриев А. Н., Журавлев Ю. И., Кренделев Ф. П.* О математических принципах классификации предметов и явлений // Дискретный анализ. Вып. 7. Новосибирск: ИМ СО АН СССР, 1966. — С. 3–11.

[2] *Баскакова Л. В., Журавлев Ю. И.* Модель распознающих алгоритмов с представительными наборами и системами опорных множеств // Журн. вычисл. матем. и матем. физики. — 1981. — Т. 21, № 5. — С. 1264–1275.

[3] *Донской В. И., Башта А. И.* Дискретные модели принятия решений при неполной информации. — Симферополь: Таврия, 1992. — 166 с.

[4] *Рязанов В. В.* Логические закономерности в задачах распознавания (параметрический подход) // Журнал вычислительной математики и математической физики. — 2007. — Т. 47, № 10. — С. 1793–1808.

[5] *Ковшов Н. В., Моисеев В. Л., Рязанов В. В.* Алгоритмы поиска логических закономерностей в задачах распознавания // Журнал вычислительной математики и математической физики. — 2008. — Т. 48, № 2. — С. 329–344.

[6] *Шибзухов З. М.* Конструктивные методы обучения сигма-пи нейронных сетей // НИИ приклад. Математики и автоматизации Кабардин.-Балкар. НЦ РАН. — М.: Наука, 2006. — С. 159.

Тесты на наличие тренда общей формы во временных рядах с сезонностью и зависимостью наблюдений

Китов В. В.
vkitov@mail.ru
Москва, ЗАО «Форексис»

В статье предложены два статистических теста на наличие тренда во временном ряде, допускающем наличие сезонности и случайных ошибок, которые могут быть коррелированными во времени. Поскольку мощность теста существенно зависит от состоятельности оценивания дисперсии случайной компоненты ряда при наличии тренда и сезонности, предложены два способа для получения устойчивых оценок дисперсии (для случая коррелированных и некоррелированных ошибок), и доказана их состоятельность.

Введение

Задача выделения тренда — медленно меняющейся составляющей в статистических наблюдениях — представляет интерес как для прогнозирования, так и для ретроспективного анализа данных. В задачах прогнозирования учет и корректная спецификация тренда позволяет существенно повысить качество прогнозов, особенно долгосрочных. В задачах ретроспективного анализа данных тренд общей формы искажает вероятностную структуру рассматриваемого процесса, и должен быть учтен на самой ранней стадии анализа данных для того, чтобы последующие статистические выводы оказывались состоятельными. В связи с этим, важным начальным шагом при работе с временными рядами является проверка наличия в них тренда.

Существуют три разновидности трендов, которые исследуются в литературе.

Параметрические тренды задаются некоторой параметрической функцией известного вида. Например, предполагая квадратичный тренд, исходный временной ряд можно представить следующим образом: $y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \varepsilon_t$. Далее можно тестировать наличие тренда путем проверки гипотезы о том, что коэффициенты при тренде $\beta_0, \beta_1, \beta_2$ равны нулю.

Стохастические тренды описываются случайным процессом с заданным вероятностным законом распределения. Обычно под стохастическим трендом понимают единичный корень в модели авторегрессии. Существуют тесты Дики-Фуллера [1] и Филлипса-Перрона [2], которые тестируют нулевую гипотезу о наличии единичного корня против альтернатив, согласно которым рассматриваемый временной ряд описывается стационарной авторегрессией, которая может допускать ненулевой уровень и линейный тренд.

Непараметрические тренды представляют собой медленно меняющуюся детерминированную компоненту временного ряда неизвестной формы. Широким классом тестов, проверяющих наличие непараметрических трендов, являются так называемые непараметрические или ранговые тесты. В таких тестах вначале осуществляется упрощение рассматриваемой величины — проекция ее значений на бинарный или целочисленный набор значений, а далее строятся статистики по значениям проекции. Примерами таких тестов являются критерий числа серий знаков первых разностей [3, стр.533], тест Вальда-Волфовитца [3, стр.526], тест Манна-Кендалла [8].

Ограничением параметрических тестов на тренд является то, что форма тренда обычно является заранее неизвестной. Существует широкий класс трендов, против которых параметрические тесты обладают малой мощностью, например, тест на линейный тренд плохо распознает наличие трендов синусоидальной формы.

Аналогично, ограничением тестов на стохастический тренд является предположение об известной вероятностной структуре тренда. Зачастую предположение о случайном блуждании является несовместимым с предположениями о форме трендов, которые исследователь ожидает увидеть в данных.

Существенным предположением, используемым в непараметрических ранговых тестах, является то, что наблюдения являются независимыми и одинаково распределенными, что часто нарушается на реальных временных рядах и приводит к искаженным статистическим результатам.

В классе тестов, не подверженных указанным ограничениям, тестируется нулевая гипотеза об отсутствии тренда (ряд предполагается стационарным и допускающим зависимость наблюдений) против альтернативы, заключающейся в наличии тренда неизвестной формы, которая может быть и стохастической. Примерами подобных тестов выступают KPSS тест [3], где тестовой статистикой является нормированная сумма квадратов кумулятивных сумм регрессионных ошибок, R/S тест [4, 5], основанный на отношении разницы максимального и минимального значения кумулятивных сумм элементов выборки к стандартному отклонению, а также V/S тест [6], где анализируется отношение дисперсии кумулятивных сумм наблюдений к стандартному отклонению наблюдений.

В данной работе предлагается два новых метода проверки наличия тренда, принадлежащих к последней категории тестов. Актуальность работы заключается в том, что в исследуемом временном ряде допускается сезонность, что характерно для многих эконометрических приложений. Помимо этого, аналитическая форма тестовых статистик обеспечивают содержательную интерпретацию результатов в случае отклонения от нулевой гипотезы. Первый предложенный тест основан на величине максимального значения модуля суммы наблюдаемых величин для всевозможных окон фиксированной ширины. Он использует подход тестирования на стабильность параметров статистической модели, предложенный в [7, 8]. Большие значения тестовой статистики выявляют тренды, локализованные на одном из участков выборки. Второй тест основан на сумме квадратов величин, получающихся за счет суммирования наблюдений на непересекающихся интервалах выборки. Высокие значения данной тестовой статистики позволяют выявлять малые по величине, но устойчивые тренды, распределенные по всей выборке. Одновременное применение предложенных тестов позволяет делать содержательные выводы о форме тренда, наблюдаемого в данных. Статистики альтернативных тестов на тренд не имеют подобной содержательной интерпретации.

Мощность представленных тестов существенно зависит от устойчивости оценки долгосрочной дисперсии к присутствию тренда и сезонности в данных. Недостаточная устойчивость оценок дисперсии при отклонении от нулевой гипотезы или при наличии сезонности может существенно ухудшать точность тестирования, что показано, например, в [9]. В статье предложены два метода состоятельного оценивания долгосрочной дисперсии для этого случая — при коррелированных и некоррелированных ошибках.

Постановка задачи

Дана выборка $x_1, \dots x_n$ из некоторого случайного процесса x_t, удовлетворяющего следующим условиям:

1) случайный процесс x_t описывается моделью $x_t = \mu + m_t + s_t + \varepsilon_t$, где μ — константа, определяющая уровень временного ряда, m_t — тренд неизвестной формы, s_t — аддитивная сезонность, ε_t — случайный шум;
2) приращения тренда уменьшаются с объемом выборки $m_{t+1} - m_t = O\left(n^{-\beta}\right)$, $\beta \in \left(0, \frac{1}{2}\right]$;
3) уровень μ выбран таким образом, что $\sum\limits_{t=1}^{n} m_t = 0$;
4) сезонная компонента s_t удовлетворяет условию $s_{t+d} \equiv s_t$ для известного периода сезонности d и условию $\sum\limits_{t=1}^{d} s_t = 0$.
5) ε_t — строго стационарный процесс, удовлетворяющий условиям $\mathsf{E}\varepsilon_t = 0$, $\mathsf{E}|\varepsilon_t|^{2+\delta} < \infty$, $\delta > 0$ и условию *строгого перемешивания* [12] — для сигма-алгебр $S_a^b = \sigma\{\varepsilon_a, \dots, \varepsilon_b\}$ справедливо:

$$\sup_{A \in S_{-\infty}^{t}, B \in S_{t+\tau}^{+\infty}} \big(P(AB) - P(A)P(B)\big) = \alpha(\tau),$$

где $\alpha(\tau) < C\tau^{-\omega}$, $\omega\frac{\delta}{2+\delta} > \frac{3}{2}$.

Далее будет тестироваться нулевая гипотеза об отсутствии тренда $H_0\colon m_t \equiv 0$.

Тест на основе скользящего окна наблюдений

Рассмотрим величины:

$$S_n^0 = \sup_{0 \leqslant t \leqslant 1-h} \left| \frac{1}{\sqrt{n}\sigma_{LR}} \sum_{\tau=nt}^{n(t+h)} x_\tau \right|, \qquad (1)$$

$$S_n^1 = \sup_{0 \leqslant t \leqslant 1-h} \left| \frac{1}{\sqrt{n}\sigma_{LR}} \sum_{\tau=nt}^{n(t+h)} (x_\tau - \bar{x}) \right|, \qquad (2)$$

где σ_{LR}^2 — долгосрочная дисперсия процесса ошибок ε_t, равная $\sigma_{LR}^2 = \sum\limits_{j=-\infty}^{+\infty} \mathsf{E}\left[\varepsilon_t \varepsilon_{t+j}\right]$. В частности, при некоррелированных ошибках долгосрочная дисперсия совпадает с обычной.

Пусть $B(t)$ обозначает процесс броуновского движения, а $B^1(t)$ обозначает случайный процесс $B^1(t) = B(t) - tB(1)$ — *броуновский мост*.

Теорема 1. *При выполнении нулевой гипотезы об отсутствии тренда и при дополнительном предположении $\mu = 0$ статистика S_n^0 имеет следующее асимптотическое распределение:*

$$S_n^0 \xrightarrow{d} \sup_{0 \leqslant t \leqslant 1-h} \big|B(t+h) - B(t)\big|,\ n \to \infty, \qquad (3)$$

где символ $\xrightarrow{d}$ обозначает сходимость по распределению.

Если во временном ряде допускается присутствие ненулевого уровня, то для тестирования гипотезы об отсутствии тренда следует использовать следующий результат.

Теорема 2. *При выполнении нулевой гипотезы об отсутствии тренда статистика S_n^1 имеет следующее асимптотическое распределение:*

$$S_n^1 \xrightarrow{d} \sup_{0 \leqslant t \leqslant 1-h} \big|B^1(t+h) - B^1(t)\big|,\ n \to \infty. \qquad (4)$$

При практическом тестировании вместо значения долгосрочной дисперсии σ_{LR} в статистиках (3) и (4) следует подставлять ее состоятельную оценку. По теореме Слуцкого асимптотическое распределение тестовой статистики при этом не изменится. Оценка должна состоятельно оценивать долгосрочную дисперсию шума при наличии сезонности в располагаемой выборке и возможном наличии

тренда, чтобы статистический тест получался мощным. Если использовать стандартную оценку дисперсии, то она получится завышенной, поскольку будет учитывать детерменированные сезонные колебания и колебания, вызванные возможным трендом, что будет занижать значения тестовых статистик (3) и (4) и понижать мощность теста. Аналогичная проблема, названная *немонотонной мощностью теста*, имеет место при тестировании на структурные сдвиги [9].

В следующих теоремах предлагаются оценки дисперсии шума, устойчивые к наличию тренда и сезонности.

Теорема 3. *Если ошибки ε_t некоррелированы, то состоятельной оценкой дисперсии шума является*

$$\hat{\sigma}^2_{LR} = \frac{1}{2}\frac{1}{n-d}\sum_{t=d+1}^{n}(x_t - x_{t-d})^2.$$

Идея состоятельного оценивания σ^2_{LR}, устойчивого к тренду и сезонности, в случае коррелированных ошибок, состоит в расчете стандартной взвешенной суммы оценок моментов распределения ряда. Единственное отличие состоит в том, что моменты рассчитываются для ряда z_t, представляющего собой исходный ряд x_t, очищенный от тренда и сезонности.

Рассмотрим последовательность весов w^n_k, удовлетворяющих условиям:

1) $\sum\limits_{k=-n}^{+n} w^n_k\Big[O\left(\frac{k}{n}\right)+O\left(\frac{1}{\sqrt{n-k}}\right)\Big] \to 0$ при $n \to \infty$.

2) $\left|1 - w^n_k\right| \leqslant \left(\frac{k}{n}\right)^{\beta},\ \beta > 0.$

Примером такой последовательности весовых коэффициентов являются веса оценки Ньюи-Веста [11].

Введем обозначения среднего в точке для произвольного ряда u_t:

$$\bar{u}_t = \frac{1}{|I_t|}\sum_{i\in I_t} u_i,$$

где $I_t = \{t-p,\dots,t+p\} \cap \{1,\dots n\}$.

Обозначим $z_t = x_t - \bar{x}_t - \hat{s}_t$, где $\hat{s}_t$ — оценка сезонной компоненты, которая рассчитывается по следующей формуле ($[x]$ обозначает целую часть x):

$$\hat{s}_t = \frac{1}{[n/d]-1}\sum_{k=1}^{[n/d]-1}\left(x_{t+kd} - \bar{x}_{t+kd}\right).$$

Теорема 4. *Если ошибки ε_t коррелированы, то в качестве состоятельной оценки долгосрочной дисперсии шума $\hat{\sigma}^2_{LR}$ можно взять взвешенную оценку*

$$\hat{\sigma}^2_{LR} = \sum_{k=-K}^{K} w_{m,k}\hat{\gamma}_k, \tag{5}$$

$$\hat{\gamma}_k = \frac{1}{n}\sum_{t=1}^{n-k} z_t z_{t+k}, \quad K = \bar{o}\big(\alpha^{-1}\big),$$

$$\alpha = \max\Big\{\tfrac{1}{\sqrt{n}}\tfrac{p}{n^{\beta}}; \left(\tfrac{p}{n^{\beta}}\right)^2; \tfrac{1}{p}; \tfrac{1}{\sqrt{n}}\Big\}.$$

В теореме 4 остается открытым вопрос, каким должен выбираться параметр p. На него отвечает следующая теорема.

Теорема 5. *Наибольшая асимптотическая скорость сходимости оценки (5) достигается при выборе параметра $p = O\left(n^{\beta/2}\right)$.*

Тест на основе разбиения выборки на блоки наблюдений

Дана выборка $x_1,\dots x_n$ из некоторого случайного процесса x_t. При фиксированном параметре p при каждом n исходная выборка $x_1,\dots x_n$ разделяется на p непересекающихся отрезков длины $w = [n/p]$. Введем обозначения для каждого из отрезков выборки:

$$I_1 = \big\{1,\dots w\big\};$$
$$I_2 = \big\{w+1,\dots 2w\big\};$$
$$\dots$$
$$I_p = \big\{w(p-1)+1,\dots wp\big\}.$$

Будет рассматриваться асимптотический переход, при котором $p = \text{const}$, а $w \to \infty$.

Будем предполагать, что процесс x_t удовлетворяет условиям 1)–5) постановки задачи, но теперь процесс ошибок ε_t строго стационарен на каждом из интервалов $I_1,\dots I_p$ исходной выборки. Таким образом, процесс ошибок является кусочно стационарным, и его долгосрочная дисперсия может быть записана в виде $\sigma^2_{\varepsilon}(i) = \sigma^2_k$, где $i \in I_k$. На практике эта форма служит некоторым приближением к гетероскедастичности общей формы.

Решение

Тест на отсутствие тренда будет строиться на основе вектора

$$\boldsymbol{v} = \begin{pmatrix} \frac{1}{\sqrt{w}\sigma_1}\sum\limits_{i\in I_1}(x_i - \bar{x}) \\ \cdots \\ \frac{1}{\sqrt{w}\sigma_p}\sum\limits_{i\in I_p}(x_i - \bar{x}) \end{pmatrix}. \tag{6}$$

Очевидно, большие значения отдельных компонент вектора будут свидетельствовать о наличии тренда. Усреднение в (6) осуществляется с учетом разницы в дисперсии ошибок:

$$\bar{x} = \frac{1}{S}\left(\frac{1}{\sigma_1 w}\sum_{i\in I_1} x_i + \dots + \frac{1}{\sigma_p w}\sum_{i\in I_p} x_i\right),$$
$$S = \frac{1}{\sigma_1} + \dots + \frac{1}{\sigma_p}.$$

Теорема 6. *При нулевой гипотезе об отсутствии тренда статистика $\boldsymbol{v}^{\intercal}\boldsymbol{v}$ стремится по распределению к хи-квадрат распределению с $p-1$ степенью свободы:*

$$\boldsymbol{v}^{\intercal}\boldsymbol{v} \xrightarrow{d} \chi^2_{p-1}. \tag{7}$$

Замечание 1. По теореме Слуцкого при подстановке в (7) состоятельных оценок долгосрочной дисперсии $\hat{\sigma}_1, \dots, \hat{\sigma}_p$ ее асимптотическое распределение не изменится.

Замечание 2. При нарушении нулевой гипотезы об отсутствии тренда значения статистик (3), (4) и (7) смещаются на величину порядка $O(\sqrt{n})$.

Выбор параметров тестов

Как видно из аналитического представления, статистики (3) и (4) позволяют выявлять тренды, локализованные на отдельных участках выборки. Высокие значения параметра h позволяют обнаруживать тренды, распределенные на широкой части выборки. Более низкие значения параметра h позволяют выявлять тренды, локализованные на небольших интервалах. Таким образом, выбор h определяется тем, насколько длительные всплески значений исследователь может считать трендом, и диктуется постановкой задачи. В случае, если исследователя интересуют тренды, распределенные по всей длине выборки, следует применять статистику (7). Параметр p определяет, насколько изменчивые тренды требуется обнаружить. При медленно меняющихся трендах следует выбирать большие длины интервалов, на которые разбивается выборка, а при быстро меняющихся трендах длина интервалов должна быть меньше.

В случае сезонных трендов параметр p должен быть меньше длины сезонности. Например, при применении теста к синусоидальной волне с шумом, тест не обнаружит сезонной составляющей, когда длина интервалов разбиения равна периоду волны, и для корректной работы теста ширина интервалов должна быть меньше.

Другим параметром (7) выступает координата крайнего интервала. На рядах с коротким сезонным периодом ее рекомендуется выбирать таким образом, чтобы число интервалов выборки было наибольшим, чтобы в тестовой статистике использовалось максимальное количество информации из выборки. На рядах с длинным сезонным периодом параметры h и p следует выбирать таким образом, чтобы окна наблюдений, по которым производится суммирование, охватывали весь сезонный период, а координата крайнего интервала в (7) должна быть такой, чтобы в каждый интервал входили наблюдения как с положительной, так и с отрицательной сезонной компонентой, чтобы сезонность оказывала наименьшее влияние на результаты тестирования.

Заключение

В работе предложены две тестовых статистики для проверки нулевой гипотезы об отсутствии тренда во временном ряде, допускающем наличие аддитивной сезонности и случайных компонент, коррелированных во времени. Мощность предложенных тестов может быть существенно повышена, если использовать две предложенные оценки дисперсии случайных компонент, которые устойчивы к наличию тренда и сезонности в данных.

Литература

[1] *Dickey D. A., Fuller W. A.* Distribution of the estimators for autoregressive time series with a unit root // Journal of the American Statistical Association. — 1979. — №74. — Pp. 427–431.

[2] *Phillips P. C. B, Perron P.* Testing for a unit root in time series regression // Biometrika. — 1988. — №75. — Pp. 335–346.

[3] *Kwiatkowski D., Phillips P. C. B., Schmidt P.* Testing the Null Hypothesis of Stationarity against the Alternative of a Unit Root // Journal of Econometrics. — 1992. — №54. — Pp. 159–178.

[4] *Hurst H.* Long term storage capacity of reservoirs // Transactions of the American Society of Civil Engineers. — 1951. — №116. — Pp. 770–799.

[5] *Lo A.* Long-term memory in stock market prices // Econometrica. — 1991. — №59. — Pp. 1279–1313.

[6] *Giraitis L., Leipus R., Leipus A.* The test for stationarity versus trends and unit roots for a wide class of dependent errors // Econometric Theory. — 2006. — №22. — Pp. 989–1029.

[7] *Zeileis A., Hornik K.* Generalized M-Fluctuation Tests for Parameter Instability // Report Series SFB. — 2003. — №80. — 19 p.

[8] *Zeileis A.* Testing for Structural Change Theory, Implementation and Applications. — Dortmund: PhD thesis. Dortmund University. — 2003. — 174 p.

[9] *Perron P.* Dealing with Structural Breaks. — Boston: Boston University Working Paper, 2005. — 91 p.

[10] *Whitt W.* Stochastic-Process Limits. — New York: Springer, 2001. — 602 p.

[11] *Newey W. K., West K. D.* A simple, positive-definite, heteroskedasticity and autocorrelation consistent covariance matrix // Econometrica. — 1987. — №55. — Pp. 703–708.

[12] *Ибрагимов И. А., Линник Ю. В.* Независимые и стационарно связанные величины. — Ленинград: Наука, 1965. — 524 с.

Тест на наличие сезонности во временном ряде и условия на тренд для его применимости

Китов В. В.
vkitov@mail.ru
Москва, ЗАО «Форексис»

В статье предложен новый тест на наличие аддитивной сезонности во временном ряде при отсутствии тренда и наличии случайных ошибок общего вида, которые могут быть коррелированными во времени. Поскольку при наличии тренда предложенный и стандартный тест на сезонность могут оказаться несостоятельными, доказаны достаточные условия на тренд, гарантирующие состоятельность и несостоятельность тестов.

Введение

Учет сезонности имеет большое значение как для задач прогнозирования, так и для задач ретроспективного анализа данных. Если сезонность является статистически значимым фактором, то ее учет позволяет существенно повысить точность прогнозов. В задачах ретроспективного анализа данных оценка сезонной составляющей позволяет точнее вычислять другие параметры модели, такие, как дисперсия ошибки. С другой стороны, учет сезонности в случае, когда она не влияет на изучаемый процесс, увеличивает число параметров модели, что делает оценки других параметров менее точными, а также приводит к «переобученности» модели и затрудняет оценку качества прогнозов. Поэтому существенным этапом анализа данных является тестирование наличия сезонности в изучаемом временном ряде.

При стандартном подходе тестирования на сезонность рассматривается модель линейной регрессии, включающая сезонные фиктивные переменные. В этом случае тест на отсутствие сезонности эквивалентен тесту на одновременное равенство нулю всех регрессионных коэффициентов при сезонных регрессорах.

В статье предлагается альтернативный метод тестирования, основанный на средних величинах разностей временного ряда для различных моментов сезонного периода.

Временные ряды, исследуемые на практике, помимо сезонности могут содержать тренд — детерминированную компоненту неизвестной формы, медленно меняющуюся во времени, от которого может существенно зависеть состоятельность тестовых статистик. В работе представлены достаточные условия на тренд, при которых стандартный и предложенный методы тестирования дают состоятельные и несостоятельные (в общем случае) результаты.

Постановка задачи

Дана выборка $x_1, \dots x_T$ из некоторого случайного процесса x_t, удовлетворяющего следующим условиям:

1) случайный процесс x_t описывается моделью
$$x_t = \mu + m_t + s_t + \varepsilon_t, \qquad (1)$$
где μ — константа, определяющая уровень временного ряда, m_t — тренд общей формы, s_t — аддитивная сезонная компонента, ε_t — строго стационарный шум с нулевым средним;
2) Для ε_t справедливо предположение о *быстром забывании истории*, состоящее в том, что $\gamma_k = \mathsf{E}\left[\varepsilon_t \varepsilon_{t+k}\right] = O(k^{-\alpha}),\ \alpha > 1$.
3) сезонная компонента s_t удовлетворяет условию $s_{t+d} \equiv s_t$ для известного периода сезонности d и условию $\sum\limits_{t=1}^{d} s_t = 0$.

Рассматривается задача проверки, присутствует ли в рассматриваемом временном ряде сезонность с известным периодом сезонности d. Для этого тестируется нулевая гипотеза $H_0\colon s_t \equiv 0$.

Решение

Пусть в выборке содержится P полных сезонных периодов. Для простоты будем обозначать $u_{p,k} \equiv u_{pd+k}$, где $u_t,\ t = 1, \dots, T,$ — любая переменная, относящаяся к временному ряду. Обозначим

$$\begin{aligned} h_{p,k} &= m_{p,k} - m_{p,d}, \\ e_{p,k} &= \varepsilon_{p,k} - \varepsilon_{p,d}, \\ z_{p,k} &= x_{p,k} - x_{p,d} = h_{p,k} + e_{p,k}. \end{aligned}$$

Рассмотрим $d-1$-мерные векторы

$$\boldsymbol{\xi} = \frac{1}{\sqrt{P}} \begin{pmatrix} \sum\limits_{p=1}^{P} \left(x_{p,1} - x_{p,d}\right) \\ \dots \\ \sum\limits_{p=1}^{P} \left(x_{p,d-1} - x_{p,d}\right) \end{pmatrix},$$

$$\boldsymbol{\eta} = \frac{1}{\sqrt{P}} \begin{pmatrix} \sum\limits_{p=1}^{P} \left(\varepsilon_{p,1} - \varepsilon_{p,d}\right) \\ \dots \\ \sum\limits_{p=1}^{P} \left(\varepsilon_{p,d-1} - \varepsilon_{p,d}\right) \end{pmatrix},$$

с ковариационной матрицей Σ_P, равной

$$(\Sigma_P)_{ij} = \frac{1}{P} \mathsf{E} \left(\sum_{p=1}^{P} e_{p,i} \right) \left(\sum_{p=1}^{P} e_{p,j} \right).$$

Обозначим $\varphi_k^{i,j} = \mathsf{E}\,[e_{t,i}e_{t+k,j}]$.

Легко проверить, что при условии быстрого забывания истории стационарного шума

$$\Sigma_P \to \Sigma, \tag{2}$$

где $(\Sigma)_{ij} = \sum\limits_{k=-\infty}^{+\infty} \varphi_k^{i,j}$.

Рассмотрим последовательность весов w_k^n, удовлетворяющих условиям:

1) $\sum\limits_{k=-n}^{+n} w_k^n \left[O\left(\frac{k}{n}\right) + O\left(\frac{1}{\sqrt{n-k}}\right)\right] \to 0$ при $n \to \infty$.

2) $\left|1 - w_k^n\right| \leqslant C\left(\frac{k}{n}\right)^{\beta}$ для некоторых $C, \beta > 0$.

Данным условиям удовлетворяют, например, веса в оценке Ньюи-Веста [1, 2].

Пусть $\hat{e}_{p,k}$ обозначает оценку $e_{p,k}$.

Теорема 1. *Состоятельная оценка матрицы Σ может быть получена как матрица, в качестве элементов которой берутся взвешенные суммы*

$$(\hat{\Sigma})_{ij} = \sum_{k=-n}^{n} w_k^n\, \hat{\varphi}_k^{i,j}, \tag{3}$$

где $\hat{\varphi}_k^{i,j} = \frac{1}{P} \sum\limits_{p=1}^{P-k} \hat{e}_{p,i}\, \hat{e}_{p+k,j}$.

Для тестирования нулевой гипотезы рассмотрим вначале упрощенный случай, когда тренд m_t отсутствует.

Теорема 2. *В случае, когда в представлении (1) временного ряда отсутствует тренд, тестирующей статистикой на отсутствие сезонности является величина $\boldsymbol{\xi}^{\intercal}\hat{\Sigma}^{-1}\boldsymbol{\xi}$, которая имеет асимптотически хи-квадрат распределение с $d-1$ степенями свободы:*

$$\boldsymbol{\xi}^{\intercal}\hat{\Sigma}^{-1}\boldsymbol{\xi} \xrightarrow{d} \chi^2_{d-1}, \tag{4}$$

где символ $\xrightarrow{d}$ обозначает сходимость по распределению.

При наличии тренда вектор $\boldsymbol{\xi}$ равен

$$\boldsymbol{\xi} = \frac{1}{\sqrt{P}} \begin{pmatrix} \sum\limits_{p=1}^{P} \left(m_{p,1} - m_{p,d} + \varepsilon_{p,1} - \varepsilon_{p,d}\right) \\ \cdots \\ \sum\limits_{p=1}^{P} \left(m_{p,d-1} - m_{p,d} + \varepsilon_{p,d-1} - \varepsilon_{p,d}\right) \end{pmatrix} = \\ = \frac{1}{\sqrt{P}} \begin{pmatrix} \sum\limits_{p=1}^{P} \left(h_{p,1} + e_{p,1}\right) \\ \cdots \\ \sum\limits_{p=1}^{P} \left(h_{p,d-1} + e_{p,d-1}\right) \end{pmatrix}.$$

Если тренд изменяется достаточно медленно, то сходимость по распределению (4) при наличии тренда сохраняется. Если же изменения в тренде велики, то тест (4) перестает быть состоятельным.

Теорема 3. *Для того, чтобы при наличии тренда тест (4) оставался состоятельным, достаточно, чтобы приращения тренда равномерно убывали со скоростью $\bar{o}\left(1/\sqrt{T}\right)$:*

$$\sup_{1 \leqslant t \leqslant T} \sqrt{T}\left|m_{t+1} - m_t\right| \to 0 \quad \textit{при} \quad T \to \infty.$$

Если приращения убывают со скоростью $O\left(\frac{1}{\sqrt{T}}\right)$ или медленнее, то тестовое условие (4) в общем случае не выполнено.

Обобщение

В случае, когда изменения тренда имеют порядок $m_{t+1} - m_t = O\left(\frac{1}{n^{\beta}}\right),\ \beta \in \left[0, \frac{1}{2}\right]$, то, в соответствии с теоремой 3, тест, основанный на статистике (4), становится несостоятельным. В этом случае предлагается применять тестовую статистику (4) не к первоначальному ряду, а к его первым разностям или разностям более высокого порядка, для которых изменения тренда достаточно малы.

Предположим, тренд задается функцией $m_t = M(t/T^{\beta}),\ t = 1, \ldots, T$ с дифференцируемой производной $g(x) = M'(x)$.

В этом случае, взяв первую разность, получим понижение порядка тренда:

$$\left(m_{t+1} - m_t\right) - \left(m_t - m_{t-1}\right) = \\ = \left(M\left(\tfrac{t+1}{T^{\beta}}\right) - M\left(\tfrac{t}{T^{\beta}}\right)\right) - \left(M\left(\tfrac{t}{T^{\beta}}\right) - M\left(\tfrac{t-1}{T^{\beta}}\right)\right) = \\ = g\left(\tfrac{t+\theta_1}{T^{\beta}}\right)\tfrac{1}{T^{\beta}} - g\left(\tfrac{t-1+\theta_2}{T^{\beta}}\right)\tfrac{1}{T^{\beta}} = O\left(\tfrac{1}{T^{2\beta}}\right).$$

Предполагая наличие производных более высокого порядка, можно понижать порядок тренда и дальше.

Стандартный тест на сезонность при наличии тренда

Рассмотрим стандартный тест на сезонность. Вводятся фиктивные переменные $u_t^1, \ldots, u_t^d$, где

$$u_t^i = \begin{cases} 1, & t \bmod d = i, \\ 0, & t \bmod d \neq i, \end{cases} \qquad i = 1, \ldots, d.$$

Осуществляется оценивание регрессии

$$x_t = \beta_0 + \beta_1 u_t^1 + \beta_2 u_t^2 + \ldots + \beta_d u_t^d + \varepsilon_t, \tag{5}$$

после чего тест на наличие сезонности совпадает с тестом, проверяющим нулевую гипотезу

$$H_0 \colon \beta_1 = \beta_2 = \ldots = \beta_d = 0.$$

Определим $\boldsymbol{v}^t = \left(1, u_t^1, \ldots, u_t^d\right)^{\intercal},\ U \in \mathbb{R}^{T \times (d+1)}$, $(U)_{ti} = v_i^t,\ \hat{V}_{\beta} = T^2(U^{\intercal}U)^{-1}\hat{Q}(U^{\intercal}U)^{-1}$, где $\hat{Q}$ — состоятельная оценка долгосрочной ковариационной матрицы $Q = \sum\limits_{j=-\infty}^{+\infty} \mathsf{E}\left[\varepsilon_{t+j}\,\varepsilon_t\,\boldsymbol{v}^{t+j}\,(\boldsymbol{v}^t)^{\intercal}\right]$.

Тестирование указанной нулевой гипотезы можно осуществлять с помощью статистики Вальда [3]:

$$W = T\tilde{\boldsymbol{\beta}}^{\intercal}\tilde{V}_{\beta}^{-1}\tilde{\boldsymbol{\beta}} \xrightarrow{d} \chi^2_d, \tag{6}$$

где $\tilde{\boldsymbol{\beta}} = (\beta_1, \ldots, \beta_d)^{\intercal}$, а $\tilde{V}_{\beta}$ получается из матрицы $\hat{V}_{\beta}$ исключением первой строки и первого столбца. В случае, если в выборке присутствует тренд, справедлив следующий результат.

Теорема 4. *Для регрессии (5), примененной к x_t, справедливы следующие утверждения:*

1) *для того, чтобы при наличии тренда тест (6) оставался состоятельным, достаточно, чтобы тренд имел порядок $\bar{o}\big(\frac{1}{\sqrt{T}}\big)$ равномерно для всех точек выборки;*
2) *если тренд имеет порядок $O\big(\frac{1}{\sqrt{T}}\big)$ или ниже, то тестовое условие (6) в общем случае не выполнено.*

В случае, когда тренд не настолько мал, чтобы асимптотически его порядок можно было считать равным $\bar{o}\big(\frac{1}{\sqrt{T}}\big)$, вместо регрессии (5) следует рассматривать аналогичную регрессию, примененную к первым разностям $x_t - x_{t-1}$. Для статистики (6) в этом случае справедлив следующий результат.

Теорема 5. *Для регрессии (5), примененной к $(x_t - x_{t-1})$, справедливы утверждения:*

1) *для того, чтобы при наличии тренда тест (6) оставался состоятельным, достаточно, чтобы первые разности тренда имели порядок $\bar{o}\big(\frac{1}{\sqrt{T}}\big)$ равномерно для всех точек выборки.*
2) *если первые разности тренда имеют порядок $O\big(\frac{1}{\sqrt{T}}\big)$ или ниже, то тестовое условие (6) в общем случае не выполнено.*

Заключение

В работе был предложен новый тест на наличие сезонности и исследованы его асимптотические свойства при наличии тренда. Показано, насколько тренд влияет на результаты стандартного теста на сезонность, основанного на проверке значимого отличия от нуля регрессионных коэффициентов при сезонных фиктивных переменных.

Стандартный тест на сезонность асимптотически состоятелен, когда величина тренда имеет порядок $\bar{o}\big(1/\sqrt{n}\big)$. В случае, если приращения тренда имеют порядок $\bar{o}\big(1/\sqrt{n}\big)$, состоятельными являются стандартный тест, примененный к первым разностям исходного ряда, и тест, предложенный в данной статье. Если величины тренда (соответственно приращений тренда) имеют более низкий порядок малости, то указанные тесты не являются состоятельными. В этом случае необходимо предварительное понижение порядка тренда, например, взятием первых разностей исходного временного ряда или разностей более высокого порядка.

Полученные асимптотические выводы справедливы при условии, что сезонный период мал по сравнению с длиной выборки; например, тестируется наличие недельной сезонности в выборке дневных наблюдений на протяжении нескольких лет. Если период сезонности соизмерим с объемом выборки, то для ее приближения должны использоваться другие методы, например, сплайн-интерполяция, и методы тестирования будут другими.

Литература

[1] *Newey W. K., West K. D.* A simple, positive-definite, heteroskedasticity and autocorrelation consistent covariance matrix // Econometrica, 1987. — No. 55. — Pp. 703–708.

[2] *Hamilton J.D.* Time Series Analysis. — Princeton: Princeton University Press, 1994. — 798 p.

[3] *Анатольев С. А.* Эконометрика для продолжающих. — Москва: Российская экономическая школа, 2002–2006. — 60 с.

Обучение алгоритмов распознавания, основанных на идеях аксиоматического подхода

Коваленко Д. С., Костенко В. А.
kovalenkods@gmail.com, kost@cs.msu.su
Москва, МГУ им. Ломоносова

В данной работе рассматривается задача автоматического построения алгоритмов распознавания аномального поведения динамических систем по временным рядам, полученным с датчиков, окружающих систему. Рассмотрены основные идеи построения алгоритмов распознавания, предложен алгоритм для автоматического построения распознавателей нештатного поведения динамических систем по обучающей выборке.

Задача распознавания нештатного поведения динамических систем

Рассмотрим динамическую систему, информация о поведении которой доступна в виде фазовой траектории в пространстве показаний датчиков $X = f(t_0 + i\Delta t)$. Система может демонстрировать два типа поведения:

— штатное поведение;
— нештатное поведение; возможно несколько классов неисправностей, вызывающих данное поведение.

Каждому классу неисправности соответствует некоторая характерная траектория Y_{Anom}. Такие траектории будем называть эталонными.

Все множество траекторий, которые могут быть получены с датчиков системы, назовем допустимыми траекториями и обозначим X_{all}. Если число классов нештатного поведения системы равно L, то обозначим через $Z = \{l\}_{l=1}^{L} \cup \{0\}$ множество ответов, где 0 соответствует штатному поведению системы. Множество всех отображений из X_{all} в Z обозначим $A\colon X_{\text{all}} \to Z$.

Эталонные траектории, соответствующие различным классам нештатного поведения, могут входить в анализируемую траекторию X в искаженном виде. Искажения могут быть по амплитуде и времени. Под искажением траектории по амплитуде будем понимать изменение абсолютных значений точек траектории без изменения числа отсчетов. Под искажением траектории по времени будем понимать изменение числа отсчетов, на которых определена траектория.

Искажения могут быть нелинейными, но траектории с искажениями, соответствующие различным классам аномального поведения не должны пересекаться: $Y^1_{\text{Anom}} \cap Y^2_{\text{Anom}} = \varnothing$. При этом будем считать, что две точки траектории равны, если равны их абсолютные значения. Две траектории пересекаются, если одна из них целиком содержит другую, т. е. одна траектория содержит последовательность точек, равную всем точкам другой траектории, причем порядок следования совпадает. Результатом пересечения будем считать общую часть двух траекторий: $Y^i \cap Y^j = Y^j$, если Y^i содержит в качестве части Y^j. В противном случае будем говорить, что траектории не пересекаются: $Y^i \cap Y^j = \varnothing$.

Задача распознавания аномального поведения заключается в следующем.

Дано:

— наблюдаемая многомерная траектория X;
— набор из L классов аномального поведения системы, для каждого из которых задана эталонная траектория Y^l_{Anom}; причем траектории разных классов аномального поведения не пересекаются;
— ограничения на полноту и точность распознавания: $e_1 \leqslant \text{const}_1$ и $e_2 \leqslant \text{const}_2$, где e_1 — число ошибок распознавания первого рода, e_2 — число ошибок распознавания второго рода, const_1 и const_2 — заданные числовые ограничения.

Требуется: с учетом ограничений на полноту и точность провести распознавание и классификацию аномального поведения в работе системы на основе наблюдаемой траектории X и множества эталонных траекторий $\{Y_{\text{Anom}}\} = \bigcup\limits_{l=1}^{L} Y^l_{\text{Anom}}$.

Описание аксиоматического подхода

Идея использования аксиоматического (алгебраического) подхода для выделения трендов была предложена в работе [1], в работе [2] было предложено использование этого подхода для обнаружения нештатных режимов работы динамических систем. Основой аксиоматического подхода является разметка анализируемой траектории аксиомами. Аксиома — бинарная функция, определенная в точке и некоторой ее окрестности на траектории. Совокупность аксиом будем называть системой аксиом. Точка траектории размечается аксиомой, если в данной точке аксиома выполняется. Траектория размечается системой аксиом, если каждой точке траектории сопоставляется некоторая аксиома из системы аксиом. Таким образом, от исходной траектории $X = (x_1, x_2, \ldots, x_n)$ переходим к последовательности аксиом $J = (j_1, j_2, \ldots, j_n)$, где j_i — номер сопоставляемой отсчету i траектории X аксиомы. Размечаются эталонные траектории $\{Y_{\text{Anom}}\}$, соответствующие различным классам нештатного поведения. Далее, в ряду разметки J

ищутся последовательности аксиом, соответствующие разметкам эталонных траекторий.

Таким образом, определение нештатного поведения в работе наблюдаемой системы ведется не путем поиска эталонных траекторий $\{Y_{\text{Anom}}\}$ в наблюдаемой траектории X, а путем поиска разметок эталонных траекторий в ряду разметки J.

Для обеспечения корректности этого подхода на систему аксиом накладываются дополнительные ограничения:

— *условие полноты:* для любой точки допустимой траектории найдется аксиома из системы аксиом, её размечающая;
— *условие однозначности:* любая точка допустимой траектории может быть размечена лишь одной аксиомой из системы аксиом.

Задача построения алгоритма распознавания по обучающей выборке

Пусть задана выборка TS в виде экземпляров траекторий X, полученных в различных условиях работы системы, с различными искажениями и шумами. При этом для каждого экземпляра траектории указаны участки нештатного поведения. Всю заданную выборку можно разделить на обучающую $\check{X}$ и контрольную $\bar{X}$, где $TS = \check{X} \cup \bar{X}$.

Пусть также определена целевая функция $\psi(e_1, e_2)$, где e_1 и e_2 — число ошибок распознавания первого и второго рода.

Задача построения алгоритма распознавания нештатных ситуаций в работе динамических систем заключается в следующем:

По заданной выборке TS и заданной целевой функции $\psi(e_1, e_2)$ построить алгоритм Alg, реализующий отображение из $A : X_{\text{all}} \to Z$ и удовлетворяющий следующему набору ограничений:

— *локальные ограничения:* общее число ошибок распознавания на обучающей выборке $\check{X}$ не должно превышать заданный параметр P_{wr}: $\nu(\text{Alg}, \check{X}) \leqslant P_{wr}$.
— *требования к обобщающей способности:* алгоритм Alg должен минимизировать целевую функцию $\psi(e_1, e_2)$ на контрольной выборке $\bar{X}$;
— *ограничения на вычислительную сложность:* вычислительная сложность работы алгоритма распознавания $\Theta(\text{Alg})$ должна быть ограничена наперед заданной функцией $\theta(l, m)$, которая определяется структурой и характеристиками используемого вычислителя: $\Theta(\text{Alg}) \leqslant \theta(l, m)$, где $\theta(l, m)$ — функция от числа l эталонных траекторий и максимальной длины m эталонной траектории.

Данная задача построения алгоритма распознавания соответствует классической постановке задачи обучения по прецедентам.

Алгоритм обучения

Решением рассматриваемой задачи распознавания в рамках аксиоматического подхода является алгоритм Alg, который определяется следующими составляющими:

— алгоритм предобработки исходных данных;
— алгоритм разметки и система аксиом;
— алгоритм поиска разметок.

Все семейство алгоритмов, которые могут быть получены в результате решения рассматриваемой задачи распознавания, обозначим через S. Данное семейство является достаточно широким. Для решения поставленной задачи предложено разбить S на подсемейства.

Под *шаблоном* будем понимать следующую комбинацию:

— заданный алгоритм предобработки и диапазон допустимых значений его параметров,
— алгоритм разметки и фиксированное множество элементарных условий, из которых формируются аксиомы,
— заданный алгоритм поиска разметок и диапазон допустимых значений его параметров.

Выделение шаблонов и обучение алгоритмов распознавания в рамках выбранных шаблонов позволяет уменьшить сложность алгоритма обучения за счет поиска решения в рамках семейства, меньшего, чем S.

Для поиска решения во всем семействе S необходимо сформировать такое множество шаблонов, которое будет покрывать все семейство решений. Однако зачастую область поиска удается ограничить, отсекая некоторые подсемейства. В данной работе методы отсечения шаблонов не рассматриваются.

Общую схему алгоритма обучения распознавателей нештатного поведения с использованием шаблонов можно представить следующим образом:

1. Построение шаблонов.
2. Для каждого шаблона происходит обучение алгоритма распознавания:
   (а) построение системы аксиом;
   (б) определение значений параметров алгоритма поиска разметок.
   (в) определение значений параметров алгоритма предобработки исходных данных.
   (г) проверка критерия останова; в случае его выполнения переход на шаг 3; иначе переход на подпункт (а).
3. Выбор лучшего алгоритма распознавания из алгоритмов, полученных при обучении в рамках построенных шаблонов.

Выбор лучшего алгоритма распознавания на шаге 3 происходит на основе значений целевой функции $\psi(e_1, e_2)$ на контрольной выборке $\bar{X}$.

Построение системы аксиом. Для построения системы аксиом используется генетический алгоритм. Общую схему этого алгоритма можно представить следующим образом:

1. Создание начальной популяции.
2. Выполнение мутации особей популяции.
3. Выполнение скрещивания особей и образование расширенной популяции.
4. Селекция особей и сужение популяции.
5. Проверка критерия останова: если критерий достигнут, то останов, иначе переход к пункту 2 алгоритма.

Особью в популяции является система аксиом. На первом этапе требуется создать достаточное число существенно различных систем аксиом. Основным условием на данном этапе является создание корректных систем аксиом, то есть удовлетворяющих условиям полноты и однозначности, описанным выше.

Для автоматического создания начальной популяции была предложена схема сборки систем аксиом из отдельных аксиом, а аксиом — из элементарных условий. Элементарным условием называется предикат $P(X, j, R)$ от участка траектории X, номера отсчета этой траектории j и вектора параметров R предиката. Аксиома представляет собой дизъюнктивную нормальную форму от элементарных условий. Набор элементарных условий задается заранее.

Операции мутации и скрещивания определены на трех уровнях:

— *уровень элементарных условий:* изменение (операция мутации) и комбинация (скрещивание) параметров элементарных условий;
— *уровень аксиом:* изменение (операция мутации) составляющих аксиому элементарных условий или обмен элементарными условиями (операция скрещивания) в выбранных для этого аксиомах;
— *уровень систем аксиом:* изменение (операция мутации) аксиом или обмен аксиомами (операция скрещивания) выбранных систем аксиом.

Степень мутации и процент скрещиваемых систем аксиом в популяции определяются автоматически на основании значений функций значимости особей популяции M_s и аксиом, входящих в особи M_a. В данной работе функция значимости системы аксиом M_s определяется как

$$M_s = a_1 e_1 + a_2 e_2 + a_3 \frac{\psi_s(e_1, e_2)}{\psi_{\min}(e_1, e_2)},$$

где e_1, e_2 — число ошибок распознавания I и II рода; $\psi_s(e_1, e_2)$ — значение целевой функции для рассматриваемой системы аксиом; $\psi_{\min}(e_1, e_2)$ — значение целевой функции для системы аксиом, лучшей в популяции; a_i — некоторые положительные постоянные, определяющие вклад слагаемых.

Функция значимости аксиомы M_a определяется как:

$$M_a = b_1 M_s - b_2 \frac{\mathrm{Num}(\bar{X}) - \mathrm{num}_a}{\mathrm{Num}(\bar{X})} - b_3 \frac{L - \mathrm{eth}_a}{L},$$

где M_s — функция значимости системы аксиом, в которую входит рассматриваемая аксиома; $\mathrm{Num}(\bar{X})$ — число траекторий нештатного поведения в контрольной выборке; num_a — число срабатываний аксиомы на контрольной выборке; L — число эталонных траекторий; eth_a — число срабатываний аксиомы на эталонных траекториях; b_i — некоторые положительные постоянные, которые определяют вклад слагаемых.

При бóльших значениях функции значимости M_s параметры мутации выбираются таким образом, чтобы не сильно изменять особь, а вероятность выбора особи для скрещивания увеличивается. И, наоборот, при меньших значениях функции значимости M_s особь меняется сильнее и реже выбирается для скрещивания. Аналогично, функция значимости аксиомы M_a определяет параметры операций скрещивания и мутации на уровне аксиом и элементарных условий.

Использование функций значимости для систмы аксиом и аксиомы позволяет автоматически на каждой итерации генетического алгоритма определять параметры использования операций скрещивания и мутации. Это ускоряет процесс обучения и делает его направленным.

Селекция популяции происходит путем подсчета числа ошибок первого и второго рода e_1 и e_2 на контрольной выборке $\bar{X}$. После чего для каждой особи вычисляется значение целевой функции $\psi(e_1, e_2)$ и особи сортируются по значению целевой функции. В новую популяцию переносится заданный процент особей с лучшими значениями целевой функции, а остальные отбираются случайно из оставшихся особей.

Настройка алгоритмов предобработки и поиска разметок. Исследования показали, что рельеф целевой функции $\psi(e_1, e_2)$ алгоритма распознавания на контрольной выборке $\bar{X}$ существенно зависит от используемой системы аксиом. Подсчет целевой функции требует больших вычислительных затрат, так как требует использования алгоритма предобработки, разметки траекторий контрольной выборки, поиска разметок. Поэтому для поиска настроек алгоритмов предобработки и поиска разметок решено использовать алгоритмы локальной оптимизации, в частности алгоритм градиентного спуска.

Для предобработки использовались следующие алгоритмы: сглаживания, сжатия и интерполяции траектории на произвольный коэффициент, быстрое преобразование Фурье. Для поиска разметок

использовались следующие алгоритмы: алгоритмы на основе метрики и DTW (Dynamic Time Warping) [3], алгоритмы на основе нейросетей, кроме того был предложен алгоритм на основе расширенных разметок.

Критерием останова рассматриваемого итерационного алгоритма обучения в рамках шаблона является следующий составной критерий:

— если выполнены условия задачи обучения алгоритма распознавания, перечисленные выше, то алгоритм считается завершившимся удачно;
— если общее число итераций алгоритма превысило наперед заданное значение или число итераций без улучшения решения превысило заданный параметр, то алгоритм считается завершившимся неудачно.

Результаты численного исследования

Было проведено численное исследование на модельных данных предложенного метода построения алгоритмов распознавания и ряда известных методов: алгоритмы на основе нейросетей, алгоритмы на основе метрики и DTW (Dynamic Time Warping) [3], алгоритмы на основе преобразований Фурье и вейвлет-преобразований, алгоритм «Гусеница»-SSA (Singular Spectrum Analysis) [4]. Длина анализируемой траектории — 10000 отсчетов, средняя длина эталонной траектории — 10 отсчетов. Характеристики искажений эталонных траекторий: искажения по времени до 50%, по амплитуде до 20% от исходной траектории. Общее число траекторий нештатного поведения в модельном ряду — 50, число классов нештатного поведения — 8. Лучшие результаты показали предложенный метод построения алгоритмов распознавания и метод на основе алгоритма DTW. Число ложных распознаваний для предложенного метода составило 18, а процент нераспознанных траекторий нештатного поведения — 5%. Число ложных распознаваний для алгоритма на основе DTW составило 20, а процент нераспознанных траекторий нештатного поведения — 33%.

Литература

[1] *Рудаков К. В., Чехович Ю. В.* О проблеме синтеза обучающих алгоритмов выделения трендов (алгебраический подход) // Прикладная математика и информатика. — № 8. — М.: Издательство факультета ВМиК МГУ, 2001. С. 97–114.

[2] *Коваленко Д. С., Костенко В. А., Васин Е. А.* Исследование применимости алгебраического подхода к анализу временных рядов // Методы и средства обработки информации. Издательство факультета ВМиК МГУ, 2005. С. 553—559.

[3] *Keogh E. J., Pazzani M. J.* Derivative Dynamic Time Warping // First SIAM International Conference on Data Mining (SDM'2001), Chicago, USA. 2001. `http://www.ics.uci.edu/~pazzani/Publications/sdm01.pdf`.

[4] *Данилов Д. Л., Жиглявский А. А.* Главные компоненты временных рядов: метод «Гусеница» // СПб.: Санкт-Петербургский университет, 1997.

Отбор подмножеств взаимосвязанных признаков на основе параметрической процедуры динамического программирования*

Копылов А. В.[1]***, Середин О. С.***[1]***, Приймак А. Ю.***[2]***, Моттль В. В.***[3]
kopylov@uic.tula.ru, oseredin@yandex.ru, tramsmm@gmail.com, vmottl@yandex.ru
Тула, [1]Тульский государственный университет
Москва, [2]Московский физико-технический институт
Москва, [3]Вычислительный центр РАН

В работе предложена эффективная схема отбора подмножеств взаимосвязанных признаков на основе параметрической процедуры динамического программирования при обучении распознаванию образов. Отбор осуществляется путем введения весовых коэффициентов, характеризующих степень информативности соответствующего признака. Учет априорной информации о взаимосвязи признаков в виде штрафа на абсолютную разность весовых коэффициентов соседних признаков позволил найти новое параметрической семейство функций Беллмана и разработать на его основе бесперебороную процедуру сокращения размерности признакового пространства.

Проблема решения задачи обучения распознаванию образов в случае большого числа признаков (по сравнению с количеством объектов в обучающей выборке, так называемая ситуация «проклятия размерности»), как правило, сводится к паре стандартных подходов — сокращению размерности за счет отбора наиболее информативных признаков и наложению на решающее правило априорных ограничений (регуляризация решающего правила распознавания). В предыдущих работах авторов [1, 2, 6, 10] был предложен способ борьбы с проблемой переобученности классификаторов, комбинирующий две эти методики и заключающийся в отборе групп информативных признаков в случае их упорядоченности в задаче обучения распознаванию образов. Упорядоченность признаков характерна для задач распознавания сигналов и изображений (в этой работе мы акцентируем внимание только на одномерной упорядоченности). Действительно, большинство подходов к отбору информативных признаков рассматривает вектор признаков объектов как неупорядоченную совокупность числовых коэффициентов, более того, многие при своей постановке явно принимают гипотезу о том, что отдельные компоненты вектора признаков являются независимыми. Однако в ряде задач признаки есть суть последовательных измерений вдоль оси некоторого аргумента, например, упорядоченные отсчеты какого-либо сигнала, компоненты спектра и т. п. В ранних работах авторов [1, 2] был описан способ регуляризации решающего правила распознавания, учитывающий априорную информацию о взаимосвязи признаков. Одновременно проводились исследования по способам комбинирования модальностей в задачах интеллектуального анализа данных, которые фактически предложили эффективный инструмент отбора информативных признаков [3, 4, 5]. Учет априорной информации об одномерной упорядоченности признаков непосредственно в процедуре отбора потребовал разработки модифицированных процедур. В работе [6] уже делалась такая попытка, и модель учета взаимосвязанности признаков предлагалась как штраф в виде квадрата разности весов при информативных признаках. В этой работе мы тщательно исследуем штрафной критерий в виде функции модуля.

Отбор подмножеств взаимосвязанных признаков

Следует сразу оговориться, что, введя термин «информативные признаки», мы не подразумеваем наличие действительно некоторой информационной характеристики, описывающей совокупность признаков, как например, в информационном критерии Акаике, или энтропийном критерии Шеннона. Мы предполагаем, что на множестве измеренных характеристик существуют подмножества признаков действительно адекватных той либо иной задаче анализа данных, то есть как синоним термину «информативный признак» можно рассматривать термины «адекватный признак», «существенный признак».

Как правило, процедуры отбора признаков не учитывают специфику задач анализа сигналов и изображений. В классической постановке задачи распознавания образов предполагается, что объекты представлены своими характеристиками, и в какой последовательности они фиксировались — неважно. Грубо говоря, можно изменить порядок компонент в векторе признаков объектов, и результат построения решающего правила или отбора информативных признаков не изменится. Например, в ходе известного международного конкурса по отбору признаков NIPS 2003, признаки объектов распознавания были случайным образом переупорядочены даже в задачах класси-

*Работа выполнена при финансовой поддержке РФФИ, проекты №08-01-99003, №09-07-00394, №08-01-12023.

фикации сигналов. Мы же обращаем внимание, что для таких специфических объектов, как сигналы и изображения, можно учесть особенность их регистрации — соседство отсчетов. Наложение подобных ограничений принято называть регуляризацией (иногда стабилизацией) решающего правила распознавания. Способ учета таких структурных ограничений на направляющий вектор разделяющей гиперплоскости продемонстрирован в [1]. В [10] показано, как можно объединить эти две методики — отбор признаков и одновременный учет априорного предположения — так, что на множестве одномерно упорядоченных признаков существуют более или менее информативные группы. В сущности, предложен критерий следующего вида:

$$\begin{cases} \sum\limits_{i=1}^{n}\left[\frac{a_i^2+1/\mu}{r_i}+\left(\mu+1+\frac{1}{\mu}\right)\ln r_i\right]+ \\ +\alpha\sum\limits_{i=2}^{n} f(r_i,r_{i-1})+C\sum\limits_{j=1}^{N}\delta_j\to\min(\boldsymbol{r},\boldsymbol{a},b,\boldsymbol{\delta}), \\ g_j\Big(\sum\limits_{i=1}^{n} a_i x_{ij}+b\Big)\geqslant 1-\delta_j,\ \delta_j\geqslant 0,\ j=1,\ldots,N. \end{cases} \quad (1)$$

Здесь используются следующие обозначения: обучающая выборка $(\boldsymbol{x}_j, g_j)$, $j=1,\ldots,N$, где $\boldsymbol{x}=(x_i)_{i=1}^n\in\mathbb{R}^n$ — действительнозначный вектор признаков объекта распознавания, $g=\{\pm1\}$ — индекс классификации каждого объекта; $\boldsymbol{r}=(r_i)_{i=1}^n$ — неотрицательные весовые коэффициенты соответствующих признаков; $\boldsymbol{\delta}=(\delta_j)_{j=1}^N$ — вспомогательные переменные метода опорных векторов; $\boldsymbol{a}=(a_i)_{i=1}^n\in\mathbb{R}^n$ — направляющий вектор оптимальной разделяющей гиперплоскости и $b\in\mathbb{R}$ — её смещение; μ — неотрицательный параметр селективности; α — коэффициент глубины регуляризации.

Предлагается решать задачу минимизации критерия (1) методом Гаусса-Зайделя, разделив переменные на две группы: первая — $\boldsymbol{a}, b, \boldsymbol{\delta}$, и вторая — $\boldsymbol{r}$, и осуществлять пошаговую минимизацию критерия по одной группе параметров при фиксированной второй. Нетрудно убедиться, что при фиксированных коэффициентах $\boldsymbol{r}$ решение в двойственной форме не будет отличаться от стандартной задачи метода опорных векторов [7]. А вот для определения самих информативных весов на каждом шаге метода покоординатного спуска необходимо найти минимум следующего критерия (здесь для краткости введены обозначения: $c_i=a_i^2+1/\mu$, $i=1,\ldots,n$ и $d=\mu+1+1/\mu$; напоминаем, на этом подшаге итерационной процедуры значения $\boldsymbol{a}$ уже найдены и зафиксированы):

$$\sum_{i=1}^{n}\left[\frac{c_i}{r_i}+d\ln r_i\right]+\alpha\sum_{i=2}^{n} f(r_i,r_{i-1})\to\min(\boldsymbol{r}). \quad (2)$$

Ранее в [10] был предложен способ штрафования различий между весовыми коэффициентами, соответствующими соседним упорядоченным признакам объекта распознавания (например, сигнала) в виде квадратичной функции. В следующем разделе будет рассмотрена штрафная функция в виде функции модуля. Следует отметить, что в таком подходе априорная информация об упорядоченности признаков налагает ограничения на весовые коэффициенты, отвечающие за информативность признаков, а не на сами компоненты направляющего вектора разделяющей гиперплоскости, как, например, в процедуре, описанной в [2].

Отбор подмножества признаков с учётом модуля разности соседних весовых коэффициентов

Многочисленные эксперименты показали, что учет взаимосвязи между признаками в виде квадратичного штрафа «размывает» информативную подобласть в пространстве упорядоченных признаков. Чтобы устранить этот недостаток, было решено в качестве штрафа на отличие весовых коэффициентов использовать функцию модуля:

$$\sum_{i=1}^{n}\left[\frac{c_i}{r_i}+d\ln r_i\right]+\alpha\sum_{i=2}^{n}\left|\ln r_i-\ln r_{i-1}\right|\to\min(\boldsymbol{r}). \quad (3)$$

Поиск оптимальных значений коэффициентов направляющего вектора не изменится, а вот минимизация критерия относительно весовых коэффициентов $\boldsymbol{r}$ представляет собой отдельную проблему.

Выполним замену переменных: $\boldsymbol{u}=\{u_i=\ln r_i\}_{i=1}^n$. Тогда критерий (3) приобретает следующий вид:

$$\sum_{i=1}^{n}\left[c_i e^{-u_i}+du_i\right]+\alpha\sum_{i=2}^{n}|u_i-u_{i-1}|\to\min(\boldsymbol{u}). \quad (4)$$

Целевая функция данного критерия является парно-сепарабельной, то есть представляет собой сумму функций не более чем двух переменных.

Обозначим функции одной переменной, входящие в критерий (4), $\psi_i(u_i)=c_i e^{-u_i}+du_i$, а функции двух переменных $\gamma_i(u_i,u_{i-1})=\alpha\,|u_i-u_{i-1}|$. Тогда целевую функцию критерия (4) можно переписать в следующем виде:

$$J(\boldsymbol{u})=\sum_{i=1}^{n}\psi_i(u_i)+\sum_{i=2}^{n}\gamma_i(u_{i-1},u_i). \quad (5)$$

Для минимизации целевой функции (5) воспользуемся процедурой, основанной на принципе динамического программирования [11]. Процедура состоит в рекуррентной декомпозиции исходной задачи оптимизации функции переменных на последовательность элементарных подзадач оптимизации функции лишь одной переменной. Предполагается, что в любом случае решить такую задачу много проще, чем искать точку минимума функции многих переменных. Элементарные функции

одной переменной $\tilde{J}_i(u_i)$, подлежащие оптимизации на каждом шаге процедуры минимизации целевой функции, будем называть функциями Беллмана, как и в классическом динамическом программировании.

Процедура динамического программирования находит глобальный минимум парно-сепарабельной целевой функции за два прохода, сначала в прямом, а затем в обратном направлении.

На прямом ходе $i = 1, \ldots, n$, в соответствии с прямым рекуррентным соотношением, определяются функции Беллмана:

$$\tilde{J}_i(u_i) = \psi_i(u_i) + \min_{u_{i-1}} \Big[\gamma_i(u_{i-1}, u_i) + \tilde{J}_{i-1}(u_{i-1}) \Big]. \quad (6)$$

Последняя функция Беллмана $\tilde{J}_n(u_n)$ непосредственно показывает, как зависит минимально возможное значение критерия в целом от значения переменной u_n, и, следовательно, её оптимальное значение может быть найдено как $\tilde{u}_n = \arg\min_{u_n} \tilde{J}_n(u_n)$. Остальные элементы искомого решения $\tilde{u}_i$, $i = (n-1), \ldots, 1$, могут быть найдены путем применения обратного рекуррентного соотношения, представляющего собой обращенную форму прямого рекуррентного соотношения (6):

$$\tilde{u}_{i-1}(u_i) = \arg\min_{u_{i-1}} \Big[\gamma_i(u_{i-1}, u_i) + \tilde{J}_{i-1}(u_{i-1}) \Big]. \quad (7)$$

Применение этого соотношения на обратном ходе $i = (n-1), \ldots, 1$ процедуры очевидно.

Таким образом, каковы бы ни были функции $\psi_i(u_i)$ и $\gamma_i(u_i, u_{i-1})$ в составе парно-сепарабельной целевой функции, алгоритм динамического программирования находит точку ее глобального минимума, если, конечно, такая комбинация значений переменных существует в пределах области их варьирования, выполняя при этом заранее известное конечное число операций, пропорциональное числу переменных.

В задаче отбора взаимосвязанных признаков, переменные целевой функции являются непрерывными переменными $u_i \in \mathbb{R}^n$, и численная реализация процедуры динамического программирования возможна лишь если существует конечно-параметрическое семейство функций $\tilde{J}(u, \boldsymbol{q})$, замкнутое относительно функций $\psi_i(u_i)$ и $\gamma_i(u_i, u_{i-1})$ в том смысле, что функции Беллмана $\tilde{J}_i(u_i)$ на каждом шаге принадлежат этому семейству. В этом случае процедура оптимизации состоит в рекуррентном пересчете параметров $\tilde{\boldsymbol{q}}_i$, которые полностью представляют функции Беллмана $\tilde{J}_i(u) = \tilde{J}(u, \tilde{\boldsymbol{q}}_i)$.

В частности, как показано в [12], если функция $\gamma_i(u_i, u_{i-1})$ в целевой функции имеет вид $\gamma_i(u_i, u_{i-1}) = \alpha\,|u_i - u_{i-1}|$, $\alpha > 0$, и функция Беллмана $\tilde{J}_i(u_i)$ выпукла и дифференцируема всюду в области определения, то обратное рекуррентное соотношение (7) имеет вид:

$$\tilde{u}_{i-1}(u_i) = \begin{cases} \tilde{u}_{i-1}^{-\alpha}, & \tilde{u}_i \leqslant \tilde{u}_{i-1}^{-\alpha}, \\ \tilde{u}_i, & \tilde{u}_{i-1}^{-\alpha} < \tilde{u}_i < \tilde{u}_{i-1}^{\alpha}, \\ \tilde{u}_{i-1}^{\alpha}, & \tilde{u}_i \geqslant \tilde{u}_{i-1}^{\alpha}. \end{cases} \quad (8)$$

Здесь $\tilde{u}_{i-1}^{-\alpha}$ и $\tilde{u}_{i-1}^{\alpha}$ определяются как решения уравнений, соответственно,

$$\tilde{J}'_{i-1}(\tilde{u}_{i-1}^{-\alpha}) = \frac{d}{d\tilde{u}_{i-1}^{-\alpha}} [\tilde{J}_{i-1}(\tilde{u}_{i-1}^{-\alpha})] = -\alpha;$$

$$\tilde{J}'_{i-1}(\tilde{u}_{i-1}^{\alpha}) = \frac{d}{d\tilde{u}_{i-1}^{\alpha}} [\tilde{J}_{i-1}(\tilde{u}_{i-1}^{\alpha})] = \alpha.$$

Легко увидеть, что значение оценки $\tilde{u}_{i-1}$ полностью определяется значением соседней переменной u_i в интервале $\tilde{u}_{i-1}^{-\alpha} < u_i < \tilde{u}_{i-1}^{\alpha}$ и не зависит от него в остальной области значений переменной u_i. Этот факт и обусловливает свойство подобной процедуры сохранять резкие изменения значений параметров, и соответственно не «размывать» информативную подобласть в пространстве упорядоченных признаков.

Можно доказать, что если функции $\psi_i(u_i)$ и $\gamma_i(u_i, u_{i-1})$ выпуклы, то и все функции Беллмана также выпуклы. При неотрицательных параметрах c_i и d функции $\psi_i(u_i) = c_i e^{-u_i} + du_i$ выпуклы, и параметрическое семейство функций Беллмана, замкнутое относительно функций $\psi_i(u_i)$ и $\gamma_i(u_i, u_{i-1})$ существует, что дает возможность применить для поиска минимума неитерационную процедуру оптимизации, описанную выше.

Согласно выражению из п. 3 Алгоритма 1, производные функций Беллмана будут состоять из участков, каждый из которых представляет собой функцию вида $q_k e^{-u} + p_k$, где k — номер участка. Соответственно, параметрами производной функции Беллмана будут являться границы участков

Алгоритм 1. Вычисление оценок $\tilde{u}_i$, $i = 1, \ldots, n$.

1: начальные значения:
$\tilde{J}'_1(u_1) := -c_1 e^{-u_1} + d$;
$\tilde{u}_1^{-\alpha} := -\ln[(d+\alpha)/c_1]$;
$\tilde{u}_1^{\alpha} := -\ln[(d-\alpha)/c_1]$;
2: **для** $i := 2, \ldots, n$
3: $\tilde{J}'_i(u_i) := -c_i e^{-u_i} + d \begin{cases} -\alpha, & u_i \leqslant \tilde{u}_{i-1}^{-\alpha}; \\ \alpha, & u_i \geqslant \tilde{u}_{i-1}^{\alpha}; \\ \tilde{J}'_{i-1}(u_i), & \text{иначе}; \end{cases}$
4: найти $\tilde{u}_i^{\alpha}$: $\tilde{J}'_i(\tilde{u}_i^{\alpha}) = \alpha$;
5: найти $\tilde{u}_i^{-\alpha}$: $\tilde{J}'_i(\tilde{u}_i^{-\alpha}) = -\alpha$;
6: найти $\tilde{u}_n$: $\tilde{J}'_n(\tilde{u}_n) = 0$;
7: **для** $i := (n-1), \ldots, 1$
8: $\tilde{u}_{i-1} := \begin{cases} \tilde{u}_{i-1}^{-\alpha}, & \tilde{u}_i \leqslant \tilde{u}_{i-1}^{-\alpha}; \\ \tilde{u}_{i-1}^{\alpha}, & \tilde{u}_i \geqslant \tilde{u}_{i-1}^{\alpha}; \\ \tilde{u}_i, & \text{иначе}; \end{cases}$

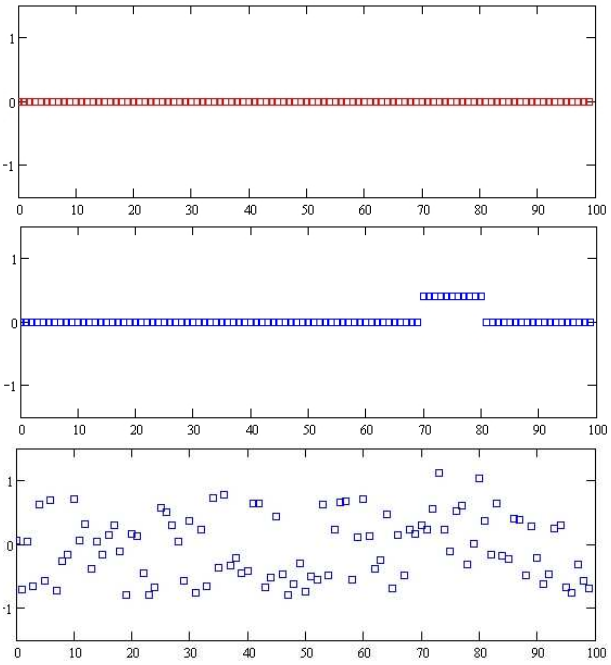


Рис. 1. Центры первого и второго классов (вверху) и пример объекта распознавания (внизу).

и параметры q_k и p_k для каждого участка k, при этом самая левая из границ участков будет соответствовать $\tilde{u}_{i-1}^{-\alpha}$, а самая правая граница будет соответствовать $\tilde{u}_{i-1}^{\alpha}$.

Экспериментальные исследования

Для экспериментального исследования предложенных алгоритмов были сгенерированы следующие тестовые данные. Два класса объектов распознавания распределены вокруг пары своих центров. Центр первого класса представляет собой 100 отсчетов со значением нуль. Центр второго класса отличается от первого тем, что на интервале от 70-го до 80-го отсчетов значения равны 0,4, см. рис. 1.

Объекты «генеральной совокупности» были сгенерированы из соответствующих центров путем добавления шума, распределенного по равномерному закону (нулевое среднее и отклонения в диапазоне $(-0{,}8; 0{,}8)$), см. рис. 1. Общее число объектов генеральной совокупности было равно 6000 (по 3000 в каждом классе). Случайным образом было отобрано по 100 обучающих выборок, содержащих по 20, 40, 60, 80, 100 и 200 объектов. Оставшиеся объекты из каждого класса использовались в качестве контрольной выборки.

На рис. 2 приведены примеры значений весовых коэффициентов для метода опорных векторов, «чистого» отбора признаков, отбора групп признаков с учетом квадратичного штрафа и штрафа в виде модуля на значения соседних весовых коэффициентов.

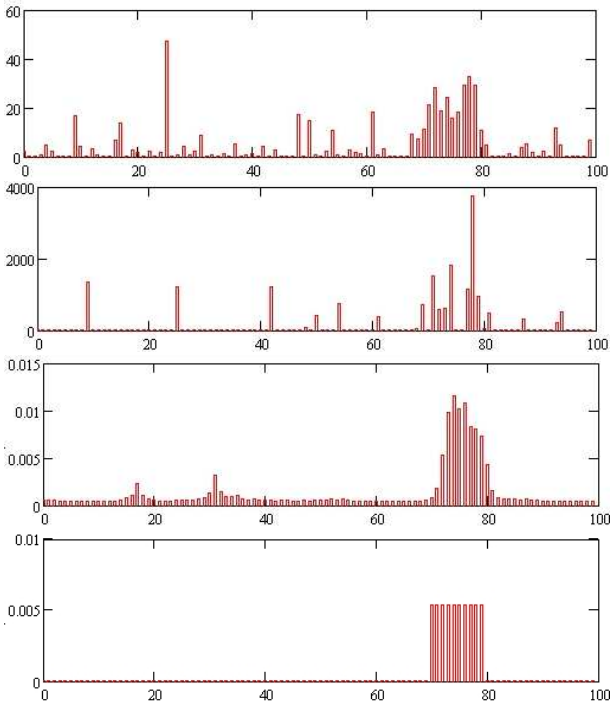


Рис. 2. Пример весовых коэффициентов для случаев (сверху вниз): SVM, SVM с отбором признаков, SVM с отбором признаков и квадратичным штрафом на несогласованность весов, SVM с отбором признаков и штрафом в виде функции модуля на несогласованность весов.

Результаты экспериментов, выраженные в виде усредненного уровня ошибки на тестовой выборке для разных размеров обучающей выборки (20-200 объектов), представлены на рис. 3. Видно, что введение регуляризации, опирающейся на поиск информативной подобласти признакового пространства (SVM с отбором признаков и квадратичным штрафом на несогласованность весов — короткий пунктир, и критерий SVM с отбором признаков и штрафом в виде функции модуля на несогласованность весов и штрафа в виде модуля — длинный пунктир) позволяет улучшить прогнозирующие свойства решающего правила по сравнению со стандартной процедурой метода опорных векторов (сплошная линия).

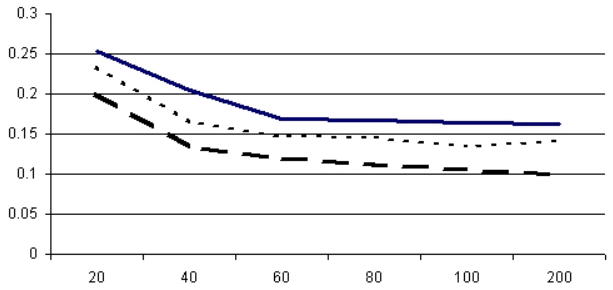


Рис. 3. Ошибка распознавания на контрольной выборке относительно числа объектов в обучающей выборке.

Остается открытым вопрос поиска оптимального значения параметра глубины регуляризации. В экспериментах использовался поиск этого параметра, основанный на процедуре скользящего контроля.

Заключение

Статья демонстрирует способ совмещения в одном критерии отбора информативных признаков и наложения априорных, разумных с точки зрения решаемой задачи, ограничений на такой отбор. Сделан обобщающий обзор ранних публикаций авторов, из которых фактически и следует предложенная в статье идея. Основная идея предложенного подхода заключается в формализации учёта информации об одномерной упорядоченности признаков, что характерно для задач анализа сигналов. Выписан критерий и предложена схема его численной оптимизации. Необходимо всесторонне рассмотреть поведение предложенного алгоритма селективного (выборочного) отбора признаков в экспериментах как на модельных, так и на реальных задачах. Так же авторам кажется разумным расширить методику на учет двумерной упорядоченности, что актуально для задач анализа изображений.

Литература

[1] *Seredin O. S., Dvoenko S. D., Krasotkina O. V., Mottl V. V.* Regularization in Image Recognition: the Principle of Decision Rule Smoothing // Pattern Recognition and Image Analysis. — 2001. — Vol. 11, № 1. P. 87–90.

[2] *Seredin O. S., Mottl V. V.* Machine Learning for Signal Recognition by the Criterion of Decision Rule Smoothness. Pattern Recognition and Image Analysis // Proceedings of the Ninth International Conference Pattern Recognition and Information Processing, Minsk, Belarus 2007 — Vol. II. P. 151–155.

[3] *Mottl V. V., Seredin O. S., Krasotkina O. V., Muchnik I. B.* Fusing of potential functions in reconstructing dependences from empirical data // Doklady Mathematics. — 2005. — Vol. 71, №. 2. P. 315–319. From Doklady Akademii Nauk. — 2005. — Vol. 401, №. 5. P. 607–612.

[4] *Mottl V. V., Seredin O. S., Krasotkina O. V., Muchnik I. B.* Principles of multi-kernel data mining // Machine Learning and Data Mining in Pattern Recognition, Springer Verlag, LNAI 3587. — 2005. P. 52–61.

[5] *Mottl V. V., Seredin O. S., Dvoenko S. D., Kulikowski C. A, Muchnik I. B.* Featureless pattern recognition in an imaginary Hilbert space // Proceedings of 16th International Conference Pattern Recognition, ICPR-2002, Quebec City, Canada. — 2002. — Vol. II. — P. 88–91.

[6] *Середин О. С., Моттль В. В.* Отбор информативных признаков при обучении распознаванию образов с упорядоченными признаками // Таврический вестник информатики и математики. — 2008. № 2. — С. 180–185.

[7] *Vapnik V.* Statistical Learning Theory. — New York: Wiley, 1998.

[8] *Mottl V., Tatarchuk A., Sulimova V., Krasotkina O., Seredin O.* Combining Pattern Recognition Modalities at the Sensor Level Via Kernel Fusion // Proceedings of 7th International Workshop Multiple Classifiers Systems, Prague, Czech Republic. — 2007. — P. 1–12.

[9] *Tatarchuk A., Mottl V., Eliseyev A., Windridge D.* Selectivity Supervision in Combining Pattern-Recognition Modalities by Feature- and Kernel-Selective Support Vector Machines // Proceedings of the 19th International Conference on Pattern Recognition, Tampa, Florida, USA. — 2008.

[10] *Seredin O., Kopylov A., Mottl V., Pryimak A.* Selection of subsets of interrelated features in pattern recognition problem // Proceedings of 9th International Conference "Pattern Recognition and Image Analysis: New Information Technologies Nizhni Novgorod. — 2008. — Vol. 2. — P. 151–154.

[11] *Mottl V., Kopylov A., Blinov A., Kostin A.* Optimization techniques on pixel neighborhood graphs for image processing // Graph-Based Representations in Pattern Recognition. Computing, Supplement 12. Wien: Springer-Verlag, 1998. — P. 135–145.

[12] *Kopylov A. V.* Parametric dynamic programming procedures for edge preserving in signal and image smoothing. // Proceedings of the 7th International Conference on Pattern Recognition and Image Analysis, St.Petersburg, 2004. — Vol. I. — P. 281–284.

Восстановление скрытой стратегии управления инвестиционным портфелем как задача оценивания нестационарной регрессии с сохранением локальных особенностей*

Красоткина О. В.[1]***, Копылов А. В.,***[1]***Моттль В. В.***[2]***, Марков М.***[3]
ko180177@yandex.ru, kav@tula.net, vmottl@yandex.ru, michael.markov@markovprocesses.com
[1]Тула, Тульский государственный университет
[2]Москва, Вычислительный центр РАН
[3]США, Markov Processes Int.

Задачи анализа сигналов практически всегда можно понимать как восстановление некоторой скрытой от наблюдателя зависимости, в общем случае нестационарной. Зачастую характер нестационарности искомой зависимости на интервале наблюдения может меняться от более к менее плавному, допуская отдельные выбросы и скачки. В данной работе предлагается алгоритм восстановления линейной регрессионной зависимости, позволяющий сохранить существенные особенности изменения коэффициентов регрессии. Предложенный метод, в отличие от существующих, является простым в настройке и обладает линейной относительно длины сигнала вычислительной сложностью. Разработанная методология позволяет следить за составом инвестиционного портфеля и анализировать стратегию инвестиционной компании с целью определения моментов резкого изменения ее инвестиционной политики.

Введение

Существует достаточно широкий класс задач анализа сигналов, в которых требуется для данного сигнала на оси дискретного аргумента $t = 1, \ldots, T$ (обычно времени) оценить скрытое значение достаточно плавно меняющегося вектора параметров $\boldsymbol{\beta} = (\beta_t)_{t=1}^N$ некоторой локальной модели наблюдаемого сигнала $Y = (y_t)_{t=1}^N$. Примерами задач подобного рода могут быть задача сглаживания сигнала, задача авторегрессионного и спектрально-временного анализа, задача восстановления структуры инвестиционного портфеля [1]. Подобные задачи в литературе принято называть задачами обобщенного сглаживания [2], и они являются частным случаем задачи нестационарной регрессии. В этой задаче анализируемый сигнал $(Y, X) = (y_t, \boldsymbol{x}_t)_{t=1}^N$ состоит из двух компонент: векторной компоненты $\boldsymbol{x}_t \in \mathbb{R}^n$, называемой регрессорами, и скалярной компоненты $y_t \in \mathbb{R}$, являющейся зашумленной линейной функцией векторной компоненты $\boldsymbol{x}_t$:

$$y_t = \sum_{i=1}^{n} \beta_t^{(i)} x_t^{(i)} + e_t = \mathbf{x}_t^{\intercal} \beta_t + e_t.$$

Естественно, что оценить мгновенную локальную модель нестационарного сигнала по его отдельному значению принципиально невозможно, поэтому оценивание нестационарной модели неизбежно должно быть основано на привлечении некоторой априорной информации. В качестве такой информации естественно принять предположение, что локальная модель изменяется в основном достаточно плавно. Однако при обработке сигналов часто возникает ситуация, когда такое априорное предположение не является вполне адекватным природе анализируемого сигнала, и необходимо допустить наличие явных разрывов в подлежащей восстановлению скрытой последовательности параметров нестационарной модели. Например, в задаче оценивания скрытой структуры инвестиционного портфеля при изменении экономических условий инвестиционная компания может довольно резко поменять стиль инвестиций, как это сделал Джорж Сорос со своим хедж-фондом Квантум, быстро продав в сентябре 1992 г. взятые им в долг британские фунты и итальянские лиры за немецкие марки, обрушив курс фунта и заработав на этой операции 2 млрд. долларов. В литературе задачи такого рода рассматриваются только для обычного сглаживания сигналов, и их принято называть задачами сглаживания с сохранением границ. В случае нестационарной регрессии методы сглаживания с сохранением границ разработаны существенно слабее и, фактически, на данный момент существует только два метода обобщенного сглаживания с сохранением границ. Один из них основан на конкурирующих фильтрах Калмана [3], другой — на итерационном применении процедуры динамического программирования [2]. Оба этих метода опираются на предположение о том, что два соседних разрыва, подлежащие обнаружению, всегда расположены не ближе, чем эффективная минимальная ширина окна, достаточного для достижения необходимой степени сглаживания. Кроме того, для обоих методов характерно наличие структурных параметров, выбор которых представляет собой плохо формализованную задачу. В данной работе предлагается алгоритм оценивания нестационарной регрессии с автоматическим оцениванием нарушений гладкости анализируемого сигнала, который лишен недостатков существующих методов сглаживания с сохранением границ.

*Работа выполнена при финансовой поддержке РФФИ, проект № 09-01-12085.

Задача оценивания структуры инвестиционного портфеля

Инвестиционная компания — это тип финансового посредника, привлекающего средства инвесторов и приобретающего на них финансовые активы такие, как, например, акции, облигации и другие ценные бумаги. Совокупность финансовых активов, которыми владеет инвестиционная компания называется ее портфелем. Целью деятельности любой инвестиционной компании является увеличение стоимости портфеля за счет естественного роста биржевых котировок составляющих его активов. Вообще говоря, финансовая инвестиционная компания не обязана информировать общественность, и даже своих акционеров, о составе своего портфеля, считая его своей профессиональной тайной. Естественно, что и собственные акционеры, и другие инвестиционные компании-конкуренты, особенно в условиях кризиса, отдали бы много за то, чтобы обладать информацией о составе интересующего их портфеля. Единственной информацией о деятельности инвестиционной компании, к которой открыт свободный доступ, является индекс ее доходности (return) в конце каждого биржевого дня, недели, квартала, года. Кроме того, известны котировки всех акций на фондовом рынке, где ведет свою деятельность Метод оценивания долей общего капитала портфеля, инвестированных в заданный набор классов ценных бумаг, представляющих собой некоторую совокупность ключевых секторов экономики, был предложен лауреатом нобелевской премии по экономике 1990 года Уильямом Шарпом под названием Returns Based Style Analysis (RBSA [4]). Этот метод, получивший огромную популярность, основан на том факте, что доходность портфеля ценных бумаг за некоторый период владения при условии, что в этот период его состав не изменялся, есть линейная комбинация доходностей тех секторов экономики, в которые предположительно вложен капитал, с коэффициентами, численно равными его долевому распределению между секторами

$$r^{(p)} \cong \alpha + \sum_{i=0}^{n} \beta^{(i)} r^{(i)} \qquad (1)$$

при естественном ограничении $\sum_{i=0}^{n} \beta^{(i)} = 1$, накладываемым на совокупность долевых коэффициентов. Фундаментальное предположение этого подхода о неизменности состава портфеля в течение рассматриваемого периода позволяет анализировать стиль формирования только на очень коротком промежутке времени. В противоположность статический модели Шарпа мы предлагаем динамическую модель, в которой долевая структура портфеля может меняться во времени $(\beta_1, \ldots, \beta_N)$, $\beta_t = \left(\beta_t^{(1)}, \ldots, \beta_t^{(n)}\right)$. Новая динамическая модель может быть записана следующим образом

$$y_t = (r_t^{(p)} - r^{(0)}) = \\ = \sum_{i=1}^{n} \beta_t^{(i)} \left(r_t^{(i)} - r^{(0)}\right) + e_t = \beta_t^{\mathsf{T}} \boldsymbol{x}_t + e_t. \qquad (2)$$

Ключевой особенностью рассматриваемого здесь подхода к динамическому анализу состава инвестиционного портфеля является предположение, что скрытая последовательность коэффициентов представляет собой марковский случайный процесс

$$\beta_t = \boldsymbol{V}_t \beta_{t-1} + \xi_t, \qquad (3)$$

где матрицы $\boldsymbol{V}_t$ определяют характер скрытой динамики коэффициентов регрессии. Для оценивания нестационарной моделей вида (2)–(3) мы используем метод, получивший в англоязычной литературе название Flexible Least Squares [5], который применительно к анализу инвестиционного портфеля выглядит следующим образом

$$J(\beta_1, \ldots, \beta_N) = \sum_{t=1}^{N} \left(y_t - \sum_{i=1}^{n} \beta_t^{(i)} x_t^{(i)}\right)^2 + \\ + \lambda \sum_{i=1}^{n} \sum_{t=1}^{N} \left(\beta_t^{(i)} - v_t^{(i)} \beta_{t-1}^{(i)}\right)^2 \to \min. \qquad (4)$$

Здесь матрицы $\boldsymbol{V}_t = \operatorname{diag}\left(v_t^{(1)}, \ldots, v_t^{(n)}\right)$ предполагаются диагональными. Что касается положительного коэффициента λ в (4), то он согласует два противоречивых требования: во-первых, требование близкой аппроксимации наблюдаемого сигнала и, во-вторых, предположение о гладкости искомой нестационарной модели. Как было показано в [6], возрастающие значения формируют последовательность почти вложенных классов моделей, и выбор подходящего значения можно осуществлять с помощью процедуры скользящего контроля. При фиксированном коэффициенте λ критерий (4) представляет собой парно-сепарабельную целевую функцию и может быть легко минимизирован с помощью метода динамического программирования с линейной относительно длины анализируемого сигнала вычислительной сложностью.

Байесовский подход к оцениванию нестационарной регрессии с переменной степенью нестационарности коэффициентов

Пусть $(\boldsymbol{x}_t)_{t=1}^N$, $\boldsymbol{x}_t = \left(x_t^{(i)}\right)_{i=1}^n$ — последовательность регрессоров, вероятностные свойства которой не изучаются. Рассмотрим анализируемую временную последовательность $(y_t)_{t=1}^N$ как наблюдаемую часть двухкомпонентного случайного процесса, чьей скрытой частью является неизвестная последовательность коэффициентов регрессии

$\big(\boldsymbol{\beta}_t = \big(\beta_t^{(i)}\big)_{i=1}^n\big)_{t=1}^N$. Главным аспектом предлагаемой здесь технологии сглаживания с сохранением границ является априорная вероятностная модель скрытого процесса коэффициентов регрессии $\boldsymbol{\beta}_t = \big(\beta_t^{(i)}\big)_{i=1}^n$. Во-первых, предположим априори, что коэффициенты регрессии независимы. Во-вторых, предположим, что начальные значения коэффициентов $\beta_0^{(i)}$ распределены нормально с нулевым математическим ожиданием $\mathsf{E}\beta_0^{(i)} = 0$ и одинаковыми дисперсиями $\mathsf{E}\big(\beta_0^{(i)}\big)^2 = \rho$. В-третьих, каждое последующее значение коэффициентов регрессии формируется как результат процесса авторегрессии $\beta_t^{(i)} = v_t^{(i)}\beta_{t-1}^{(i)} + \xi_t^{(i)}$, где $\xi_t^{(i)}$ — нормальный белый шум с нулевым средним $\mathsf{E}\xi_t^{(i)} = 0$ и дисперсией $\mathsf{E}\big(\xi_t^{(i)}\big)^2 = \frac{\rho\omega_t}{\lambda}$. Коэффициент пропорциональности ρ представляет собой дисперсию шума в модели наблюдения (2), $\mathsf{E}e_t = 0$, $\mathsf{E}(e_t)^2 = \rho$. Ясно, что если некоторые коэффициенты $\omega_t \to \infty$, то в момент времени t разрешается наличие скачка. Но мы не предполагаем значения коэффициентов $(\omega_t)_{t=1}^N$ известными. Рассмотрим совокупность независимых гамма распределений величин, обратных ω_t:

$$\gamma\big(\tfrac{1}{\omega_t} \,|\, \alpha, \vartheta\big) \propto \big(\tfrac{1}{\omega_t}\big)^{\alpha-1} \exp\big(-\vartheta\tfrac{1}{\omega_t}\big),$$

с математическим ожиданием $\mathsf{E}\big(\frac{1}{\omega_t}\big) = \frac{\alpha}{\vartheta}$ и дисперсиями $\mathsf{E}\big(\frac{1}{\omega_t}\big)^2 = \frac{\alpha}{\vartheta^2}$. Положим $\alpha = 1 + \frac{1}{(\mu^2+1)2\rho\mu}$ и $\vartheta = \frac{1}{2\rho\mu}$. Получим параметрическое семейство распределений только с одним параметром $\mu \geqslant 0$, так что $\mathsf{E}\big(\frac{1}{\omega^{(i)}}\big) = 2\rho\mu + \frac{1}{1+\mu^2}$ и $\mathsf{E}\big(\frac{1}{\omega^{(i)}}\big)^2 = (2\rho\mu)^2 + \frac{2\rho\mu}{1+\mu^2}$. Если $\mu \to 0$, практически неслучайные параметры $\frac{1}{\omega_t}$ будут близки друг к другу $\frac{1}{\omega_1} \cong \ldots \cong \frac{1}{\omega_N} \cong 1$, напротив, если $\mu \to \infty$, независимые неотрицательные значения ω_t могут быть абсолютно различными. Совместная априорная плотность распределений дисперсий $(\omega_1, \ldots, \omega_N)$ имеет вид

$$G\,(\omega_1, \ldots, \omega_N \,|\, \mu) \propto \exp\bigg(-\tfrac{1}{2\rho\mu}\sum_{t=1}^N\big(\tfrac{1}{\omega_t} + \tfrac{\ln\omega_t}{1+\mu^2}\big)\bigg).$$

Таким образом, мы определили во-первых, априорное распределение дисперсий $G\,(\omega_1, \ldots, \omega_N \,|\, \mu)$, во-вторых, условное априорное распределение скрытой последовательности коэффициентов регрессии $\Psi(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_N \,|\, \omega_1, \ldots, \omega_N)$, и, в-третьих, условное распределение наблюдаемой временного ряда $\Phi(y_1, \ldots, y_N \,|\, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_N)$. Ясно, что апостериорная совместная плотность распределения скрытой последовательности коэффициентов регрессии и дисперсии их компонент будет пропорциональна произведению

$$\begin{aligned} P(\boldsymbol{\beta}_0, \ldots, \boldsymbol{\beta}_N, \omega_1, \ldots, \omega_N \,|\, y_1, \ldots, y_N, \mu, \lambda) \propto \\ \propto \Phi(y_1, \ldots, y_N \,|\, \beta_1, \ldots, \beta_N) \times \\ \times \Psi(\beta_0, \ldots, \beta_N \,|\, \omega_1, \ldots, \omega_N, \lambda) \times \\ \times G(\omega_1, \ldots, \omega_N \,|\, \mu). \end{aligned}$$

Естественно взять максимальную точку этой апостериорной плотности распределения как оценку последовательности нестационарных коэффициентов регрессии вместе с их дисперсиями, отвечающими за наличие в модели разрывов

$$\begin{aligned} &(\hat\beta_0, \ldots, \hat\beta_N, \hat\omega_1, \ldots, \hat\omega_N \,|\, \mu, \lambda) = \\ &= \arg\max P(\beta_0, \ldots, \beta_N, \omega_1, \ldots, \omega_N \,|\, y_1, \ldots, y_N, \mu, \lambda). \end{aligned}$$

Можно показать, что максимальная точка апостериорной плотности распределения по переменным $(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_N, \omega_1, \ldots, \omega_N)$ является минимальной точкой критерия

$$\begin{aligned} &J(\boldsymbol{\beta}_0, \ldots, \boldsymbol{\beta}_N, \omega_1, \ldots, \omega_N | \mu, \lambda) = \\ &\quad = \sum_{t=1}^N\Big(y_t - \sum_{i=1}^n \beta_t^{(i)} x_t^{(i)}\Big)^2 + \sum_{i=1}^n\Big(\beta_0^{(i)}\Big)^2 + \\ &+ \sum_{t=1}^N\bigg(\tfrac{\lambda}{\omega_t}\sum_{i=1}^n\Big(\beta_t^{(i)} - v_t^{(i)}\beta_{t-1}^{(i)}\Big)^2 + \tfrac{1}{\mu\omega_t} + \tfrac{1}{\mu(\mu^2+1)}\ln\omega_t\bigg), \end{aligned}$$

Видим, что если априорные дисперсии $(\omega_1, \ldots, \omega_N)$ коэффициентов регрессии фиксированы, то результирующие критерий практически совпадает с FLS критерием (4). При практически нулевом параметре $\mu \to 0$ минимальная точка этого критерия совпадает с минимальной точкой FLS критерия. Однако наличие в модифицированном критерии дополнительных переменных $(\omega_1, \ldots, \omega_N)$ очень важно. Если какая-то из них стремится к бесконечности, $\omega_t \to \infty$, то дисперсия шума в модели состояния (3) $\mathsf{E}\big(\xi_t^{(i)}\big)^2 \to \infty$, и критерий фактически разрешает в момент t наличие разрыва первого рода. Количество точек, в которых допускается резкое изменение коэффициентов регрессии, регулируется параметром $\mu > 0$, который мы назвали параметром неровности (unevenness). Если $\mu \to \infty$, то модифицированный критерий фактически разрешает коэффициентам регрессии резко меняться практически в каждом отсчете сигнала. Для нахождения минимальной точки модифицированного FLS критерия при фиксированных параметрах μ и λ мы применяем итерационный метод Гаусса–Зайделя к двум группам переменных $(\boldsymbol{\beta}_0, \ldots, \boldsymbol{\beta}_N)$ и $(\omega_1, \ldots, \omega_N)$, начиная с некоторых значений $\omega_t^0 = 1$, $t = 1, \ldots, N$. На каждой итерации последовательность коэффициентов регрессии $(\boldsymbol{\beta}_0^k, \ldots, \boldsymbol{\beta}_N^k)$ находится посредством минимизации парно-сепарабельного критерия с помощью процедуры динамического программирования [2] для найденных значений $\big(\omega_t^k\big)_{t=1}^N$. На следующем шаге новые значения $\big(\omega_t^{k+1}\big)_{t=1}^N$ могут быть найдены по легко доказываемым соотношениям

$$\omega_t^{k+1} = (\mu^2+1)\bigg(\mu\lambda\sum_{i=1}^n\Big(\beta_t^{(i),k} - v_t^{(i)}\beta_{t-1}^{(i),k}\Big)^2 + 1\bigg).$$

Легко видеть, что если $\mu \to 0$, то все дисперсии равны друг другу и равны 1. Для нахождения под-

ходящих значений структурных параметров μ и λ, мы применяем описанный в [2] метод скользящего контроля.

Пример анализа стратегии инвестиционного портфеля

В этом разделе мы приводим пример применения предложенной методологии для восстановления инвестиционной стратегии хедж-фонда Quantum, находящегося под управлением Джорджа Сороса. В «черную среду» 16 сентября 1992 г. Сорос открыл короткую (заемную) позицию на фунт стерлингов объемом более 10 млрд, заработав за один день более 1,1 млрд. фунтов. В результате операций Сороса Банк Англии был вынужден провести массированную валютную интервенцию и, в конечном счете, вывести фунт стерлингов из механизма регулирования курсов валют европейских стран, что привело к мгновенному падению фунта по отношению к основным валютам. Это принесло колоссальную прибыль фонду Quantum — в течении месяца Джордж Сорос заработал порядка 2 миллиардов долларов, покупая на немецкие активы уже значительно подешевевшие фунты стерлингов.

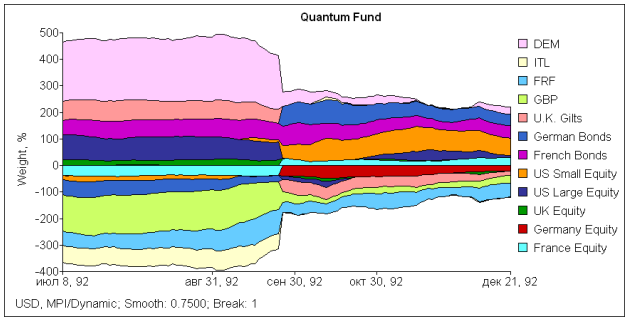


Рис. 1. Динамический анализ стратегии инвестиционного портфеля хедж-фонда Quantum в 1992 году.

Мы использовали предложенную в данной статье процедуру оценивания стратегии инвестиционного портфеля для анализа поведения динамики финансового портфеля Quantum в 1992 г. В качестве регрессоров были использованы доходности классов ценных бумаг, упомянутые Джороджем Сорсом в его известном интервью газете Таймс [7]. Результат применения методологии Flexible Least Squares с сохранением локальных особенностей к анализу портфеля Quantum показан на рис. 1. Здесь динамика долей капитала $\left(\boldsymbol{\beta}_t^{(1)}, \ldots, \boldsymbol{\beta}_t^{(n)}\right)$, инвестированных в различные валютные активы и облигации показана в виде стековой диаграммы. Отрицательные значения долей соответствуют заемным или коротким позициям. Из рис. 1 видно, что до середины сентября фонд имел значительную заемную позицию по британскому фунту, которая практически ликвидируется после «черной среды», когда Сорос произвел массированную интервенцию фунта на фондовом рынке. Из рисунка хорошо видно, что момент резкой смены политики фонда, который произошел в конце сентября, был оценен верно.

Заключение

В статье предложена процедура оценивания нестационарной регрессии с сохранением границ, позволяющая автоматически обнаруживать имеющиеся в исходном сигнале существенные неоднородности и не требующая задания таких сложных в настройке параметров, как ширина окна или величина порога. Процедура является модификацией известного в литературе метода Flexible Least Squares, и, как следствие, имеет линейную вычислительную сложность в отличие от подавляющего большинства алгоритмов сглаживания с сохранением границ, имеющих нелинейную относительно длины сигнала вычислительную сложность. Предложенная в статье методология позволяет значительно улучшить качество существующей в настоящее время методики восстановления инвестиционного портфеля RBSA, предложенной Уимьямом Шарпом.

Литература

[1] *Моттль В. В., Костин А. А., Красоткина В. В* Алгоритмы динамического программирования для анализа нестационарных сигналов. // Журнал вычислительной математики и математической физики. — 2003. — № 2. — С. 103–117.

[2] *Mottl V. V., Muchnik I. B., Kostin A. A.* Generalized edge-preserving smoothing for signal analysis // Proceedings of the IEEE Workshop on Nonlinear Signal and Image Analysis., Mackinac Island, Michigan, USA: September 7–11 1997.

[3] *Niedzwiecki M., Sethares W. A.* Smoothing of discontinuous signals: The competitive approach. // IEEE Trans. on Signal Processing, — Vol. 43, № 1, — January 1995, — Pp. 1–13.

[4] *Sharpe W. F.* Asset allocation: Management style and performance measurement. The Journal of Portfolio Management, Winter 1992. — Pp. 7–19.

[5] *Kalaba R., Tesfatsion L.* Time-varying linear regression via flexible least squares // International Journal on Computers and Mathematics with Applications — 1989. — Vol. 17. — Pp. 1215–1245.

[6] *Markov M., Krasotkina O., Mottl V., Muchnik I.* Time-varying regression model with unknown time-volatility for nonstationary signal analysis. // Proceedings of the 8th IASTED International Conference on Signal and Image Processing. Honolulu, Hawaii, USA, August 14–16, 2006.

[7] *Kaletsky A.* How Mr. Soros made a billion by betting against the pound // October 26, 1992, Monday, The Times, Copyright 1992 Times Newspapers Limited.

Сравнение эвристических алгоритмов выбора линейных регрессионных моделей*

Крымова Е. А., Стрижов В. В.
ekkrym@mail.ru, strijov@ccas.ru
Московский физико-технический институт,
Москва, Вычислительный центр РАН

В работе описан способ построения линейных регрессионных моделей, основанный на порождении и выборе признаков. Предложены эвристические алгоритмы выбора признаков. Выполнено сравнение этих алгоритмов с общеизвестными. Особенностью данного исследования является то, что задача выбора моделей поставлена для счетного набора признаков.

Введение

Процедура построения регрессионных моделей состоит из двух шагов. На первом шаге, на основе множества свободных переменных, порождается набор признаков. Один из способов построения такого набора описан в [1]. Модель-претендент есть линейная комбинация конечного подмножества признаков. На втором шаге производится выбор признаков, при этом выполняется настройка параметров модели и оценивается ее качество. Модель с настроенными параметрами, доставляющая минимум заданному функционалу качества, называется моделью оптимальной структуры.

Целью данной работы является сравнение предложенных эвристических алгоритмов с известными алгоритмами. Мотивацией работы является тот факт, что решение практических задач восстановления регрессионной зависимости требует рассмотрения большого числа порождаемых признаков. При таком условии алгоритмы, называемые «жадными», выбирают некоторый поднабор признаков, без возможности его модификации с целью улучшения структуры модели. Переборные алгоритмы, не обладая этим недостатком, имеют высокую вычислительную сложность. В данной работе предлагается компромиссный вариант алгоритма выбора регрессионных моделей и сравниваются следующие алгоритмы:

1) LARS метод наименьших углов [2];
2) полный перебор моделей [3];
3) метод группового учета аргументов [1];
4) алгоритм Лассо [4];
5) стохастическая структурная оптимизация;
6) шаговая регрессия [5, 6, 7];
7) оптимальное прореживание в шаговой регрессии [8, 9];
8) модифицированный метод наименьших углов.

Сравнение алгоритмов показано на примере прикладной задачи, связанной с моделированием волатильности опционов по реальным историческим данным торгов опционом Brent Crude Oil.

*Работа выполнена при финансовой поддержке РФФИ, проекты №07-07-00181, №08-01-12022.

Постановка задачи

Задана выборка $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$ — множество m пар, состоящих из вектора значений n свободных переменных $\boldsymbol{x}_i = (x_i^j)_{j=1}^n \in \mathbb{R}^n$, и значения одной зависимой переменной $y_i \in \mathbb{R}^1$. Индекс i элементов выборки и индекс j свободных переменных далее будем рассматривать как элементы множеств $i \in I = \{1, \ldots, m\}$ и $j \in J = \{1, \ldots, n\}$.

Задан класс регрессионных моделей $\mathcal{F} = \{f_s\}$ — параметрических функций, линейных относительно параметров,

$$y_i = f_s(\boldsymbol{\beta}_s, \boldsymbol{x}_i) = \sum_{j \in J_s} \beta_j x_i^j, \qquad (1)$$

в которой $s \in \{1, \ldots, 2^n\}$ является индексом модели, $\boldsymbol{\beta}_s = (\beta_j)_{j \in J_s}$ — вектор параметров, заданный индексом модели, $J_s \subseteq J$ — набор индексов свободных переменных. Введено ограничение на число элементов линейной комбинации (1). В множество $\mathcal{F}$ могут входить только модели с числом свободных переменных $|J_s| \leqslant R$.

Принята следующая гипотеза порождения данных. Пусть случайная аддитивная переменная ν регрессионной модели

$$y = f(\boldsymbol{\beta}, \boldsymbol{x}) + \nu$$

имеет нормальное распределение $\mathcal{N}(0, \sigma_\nu^2)$.

Тогда, с учетом гомоскедастичности регрессионных остатков, распределение зависимой переменной имеет вид

$$p(y|x, \boldsymbol{\beta}, \sigma_\nu^2, f) = \frac{\exp\left(-\frac{1}{\sigma_\nu^2} S(D \,|\, \boldsymbol{\beta}, f)\right)}{(2\pi\sigma_\nu^2)^{\frac{n}{2}}},$$

где S — сумма квадратов невязок $y_i - f(\boldsymbol{\beta}, \boldsymbol{x}_i)$. Это распределение задает указанный ниже критерий качества модели.

Дополнительно задано разбиение выборки $I = I^{\mathrm{T}} \sqcup I^{\mathrm{C}}$ на обучающую и контрольную. Для каждого набора данных, рассматриваемого в вычислительном эксперименте, наборы индексов $I^{\mathrm{T}}, I^{\mathrm{C}}$ определены до начала эксперимента. Алгоритм выбора модели определяет метод оптимизации, доставляющий оптимальное значение параметрам $\tilde{\boldsymbol{\beta}}$

модели f на обучающей выборке $\{(x_i, y_i)\colon i \in I^{\mathrm{T}}\}$. Принят критерий качества — сумма квадратов регрессионных остатков на контрольной выборке

$$S = \sum_{i \in I^{\mathrm{C}}} \Big(y_i - f(\tilde{\boldsymbol{\beta}}, \boldsymbol{x}_i)\Big)^2. \qquad (2)$$

Требуется найти такую модель $f_s \in \mathcal{F}$, которая доставляет наименьшее значение функционалу качества. Такая модель будет называться моделью оптимальной структуры.

Порождение свободных переменных

Предлагается следующий способ формирования выборки D, состоящий из двух шагов.

Шаг первый. Задано множество непорождаемых свободных переменных $\Xi = \{\xi^u\}_{u=1}^{U}$. Задано конечное множество функций $G = \{g_v\}_{v=1}^{V}$. Рассмотрим декартово произведение $G \times \Xi$, элементу (g_v, ξ^u) которого поставлена в соответствие суперпозиция $g_v(\xi^u)$, однозначно определяемая индексами v, u. Обозначим $a_i = g_v(\xi^u)$, где индекс $i = (v-1)U + u$.

Шаг второй. Назначается базовая модель порождения признаков. В качестве модели, описывающей отношение между зависимой переменной y и свободными переменными a_i, используется полином Колмогорова-Габора:

$$y = \beta_0 + \sum_{i=1}^{UV} \beta_i a_i + \sum_{i=1}^{UV}\sum_{j=1}^{UV} \beta_{ij} a_i a_j + \ldots,$$

где вектор коэффициентов

$$\boldsymbol{\beta} = (\beta_0, \beta_i, \beta_{ij}, \beta_{ijk}, \ldots)_{i,j,k,\ldots=1,\ldots,m}.$$

Запишем вышеприведенный ряд в виде

$$y = \sum_{j \in J} \beta_j x^j.$$

Переменные $\{x^j\}$ поставлены в однозначное соответствие мономам полинома.

Стандартизация данных

Выборка D стандартизирована таким образом, чтобы для $j \in J$ выполнялись условия нормировки

$$\sum_{i=1}^{m} x_i^j = 0, \quad \sum_{i=1}^{m} (x_i^j)^2 = 1, \quad \sum_{i=1}^{m} y_i = 0.$$

Предполагается, что векторы $\bar{\boldsymbol{x}}_j = \big(x_1^j, \ldots x_m^j\big)$ и $\bar{\boldsymbol{x}}_k = \big(x_1^k, \ldots x_m^k\big)$ для всех $j, k \in J$, $j \neq k$ линейно независимы.

LARS (метод наименьших углов)

LARS — алгоритм отбора признаков линейной модели [2]. На каждом шаге алгоритма происходит изменение вектора параметров модели так, чтобы доставить добавляемому признаку наибольшую корреляцию с вектором регрессионных остатков. Основным достоинством LARS является то, что он выполняется за число шагов, равное числу свободных переменных.

Лассо

Лассо — алгоритм оценивания коэффициентов линейной модели [4]. Введение ограничения на сумму абсолютных значений коэффициентов модели приводит к обращению в 0 некоторых коэффициентов модели. Ненулевые коэффициенты соответствуют признакам, входящим в модель.

Обозначим сумму модулей коэффициентов модели $T(\boldsymbol{\beta}) = \sum\limits_{j=1}^{n} |\beta_j|$. Вектор коэффициентов $\hat{\boldsymbol{\beta}}$ есть решения задачи минимизации $S(\boldsymbol{\beta})$ при ограничении: $T(\boldsymbol{\beta}) \leqslant t$, где t — параметр регуляризации. Для решения задачи используется метод квадратичного программирования.

Шаговая регрессия

Шаговая регрессия — алгоритм последовательного удаления/добавления признаков [5]. Алгоритм последовательного добавления признаков присоединяет к текущему набору A по одному признаку, который доставляет максимум нижеприведенному критерию,

$$\hat{\jmath} = \arg\max_{j \in J} \frac{S(A) - S(A \cup \bar{\boldsymbol{x}}_j)}{S(A \cup \bar{\boldsymbol{x}}_j)}.$$

Начальным считается пустой набор признаков.

Алгоритм последовательного удаления признаков начинает с самого большого набора, состоящего из всех признаков. На каждом шаге происходит удаление признака так, чтобы значение нижеприведенного критерия было как можно меньше:

$$\hat{\jmath} = \arg\min_{j \in J} \frac{S(A \backslash \bar{\boldsymbol{x}}_j) - S(A)}{S(A)}.$$

Останов алгоритма производится по выполнению условия C_p [10].

Алгоритм полного перебора

Этот алгоритм порождает все возможные множества мономов $\{\bar{\boldsymbol{x}}_j\}_{j \in J}$. Пусть сложность модели $|J| \leqslant R$. Под сложностью модели понимается число линейно входящих параметров. Алгоритм последовательно строит модели-претенденты неубывающей сложности. Параметры каждой модели настраиваются методом наименьших квадратов по обучающей выборке. Наилучшая модель выбирается исходя из минимума ошибки на контрольной выборке. Введем переменную выбора монома —

вектор $\boldsymbol{c} = (c_1, \ldots, c_n)$. Его элемент $c_j \in \{0, 1\}$ принимает значение 1, если $j \in J_s$, в противном случае 0. Базовая модель данного алгоритма имеет вид $y = \sum\limits_{j \in J_s} c_j \beta_j x^j$. Сложность этого алгоритма $\sum\limits_{i=1}^{R} C_n^i$.

Метод группового учета аргументов

Алгоритмы МГУА воспроизводят схему массовой селекции [1, 3]: последовательно порождаются модели возрастающей сложности. Каждая модель настраивается методом наименьших квадратов. Остановка алгоритма происходит, когда с увеличением номера шага начинается увеличение ошибки на контрольный выборке.

Стохастическая структурная оптимизация

Предложенный эвристический алгоритм состоит из итеративно выполняемых шагов. На первом шаге из множества признаков выбирается заданное число поднаборов, доставляющее соответствующей линейной модели наименьшее значение функционала качества. На втором шаге на заданном числе пар выполняется операция обмена признаками. На третьем шаге производится случайная замена произвольных признаков вновь полученных поднаборов. Шаги 2 и 3 итеративно повторяются. Алгоритм завершает работу, когда число шагов превысит заданное или когда ошибка оптимальной модели на контрольной выборке станет меньше заданной.

Оптимальное прореживание в шаговой регрессии

Оптимальное прореживание — это метод упрощения структуры регрессионной модели. Основная идея прореживания заключается в том, что те элементы модели, которые оказывают малое влияние на ошибку аппроксимации, можно исключить из модели без значительного ухудшения качества аппроксимации [8, 9].

Для построения регрессии требуется найти такие параметры $\hat{\boldsymbol{\beta}}$, которые доставляли бы наименьшее значение функции ошибки $S(\boldsymbol{\beta})$.

Локальная аппроксимация функции S в окрестности точки $\hat{\boldsymbol{\beta}}$ с помощью разложения в ряд Тейлора записывается в виде

$$S(\boldsymbol{\beta}+\Delta\boldsymbol{\beta}) = S(\boldsymbol{\beta})+\boldsymbol{g}^{\intercal}(\boldsymbol{\beta})\Delta\boldsymbol{\beta}+\tfrac{1}{2}\Delta\boldsymbol{\beta}^{\intercal}H\Delta\boldsymbol{\beta}+o(\|\boldsymbol{\beta}\|^3),$$

где $\Delta\boldsymbol{\beta}$ — возмущение вектора параметров $\boldsymbol{\beta}$, $\boldsymbol{g} = \frac{\partial S}{\partial \boldsymbol{\beta}}$ — градиент, $H = H(\boldsymbol{\beta}) = \frac{\partial^2 S}{\partial \boldsymbol{\beta}^2}$ — матрица вторых производных (матрица Гессе).

Функция $S(\boldsymbol{\beta})$ достигает своего максимума при $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, и ее поверхность квадратична. Таким образом, предыдущее выражение можно представить в виде $\Delta S = S(\boldsymbol{\beta} + \Delta\boldsymbol{\beta}) - S(\boldsymbol{\beta}) = \frac{1}{2}\Delta\boldsymbol{\beta}^{\intercal}H\Delta\boldsymbol{\beta}$.

Пусть исключение элемента модели есть исключение одного параметра модели, β_i. Исключенный параметр будем считать равным нулю. Исключение элемента эквивалентно выражению $\Delta\beta_i + \beta_i = 0$, иначе $\boldsymbol{e}_i^{\intercal}\Delta\boldsymbol{\beta} + \beta_i = 0$, где $\boldsymbol{e}_i$ — вектор, i-й элемент которого равен единице, все остальные элементы равны нулю.

Требуется минимизировать квадратичную форму $\Delta\boldsymbol{\beta}^{\intercal}H\Delta\boldsymbol{\beta}$ относительно $\Delta\boldsymbol{\beta}$ при ограничениях $\boldsymbol{e}_i^{\intercal}\Delta\boldsymbol{\beta} + \beta_i = 0$ для всех значений i. Задача условной минимизации решается с помощью введения Лагранжиана $L = \Delta\boldsymbol{\beta}^{\intercal}H\Delta\boldsymbol{\beta} - \lambda(\boldsymbol{e}_i^{\intercal} + w_i)$, в котором λ — множитель Лагранжа. Дифференцируя Лагранжиан по приращению параметров и приравнивая его к нулю получаем (для каждого индекса i параметра β_i)

$$\Delta\beta = -\frac{\beta_i}{[H^{-1}]_{ii}}H^{-1}\boldsymbol{e}_i.$$

Этому значению вектора приращений параметров соответствует минимальное значение Лагранжиана

$$L_i = \frac{\beta_i^2}{2[H^{-1}]_{ii}}.$$

Полученное выражение называется мерой выпуклости функции ошибки S при изменении параметра β_i.

Функция L_i зависит от квадрата параметра β_i. Это говорит о том, что параметр с малым значением будет удален из модели. Однако если величина $[H^{-1}]_{ii}$ достаточно мала, это означает, что данный параметр оказывает существенное влияние на качество аппроксимации модели.

EM+LARS

В данной работе предлагается алгоритм, сочетающий в себе жадную стратегию LARS и перебор моделей. Это позволяет улучшить структуру модели, несущественно увеличив вычислительные затраты. Происходит порождение подмножеств признаков с помощью EM-алгоритма. На этих подмножествах с помощью LARS находятся параметры моделей, производится перебор полученных моделей. Модель, доставляющая наименьшую среднеквадратичную ошибку на контрольной выборке, считается оптимальной.

Задано K натуральных чисел в порядке возрастания C_k, $k = 1, \ldots, K$, $C_k < n$. Пусть M — число моделей, получаемых на каждом шаге алгоритма. Рассмотрим k-й шаг алгоритма. Разобьем множество признаков X на C_k кластеров с помощью EM-алгоритма. M раз выбираем случайным образом из каждого кластера по одному элементу. Получаем M подмножеств из C_k признаков, принадлежащих разным классам. К каждому из подмножеств признаков на обучающей выборке приме-

Таблица 1. Результаты работы стандартных методов отбора признаков.

Алгоритм	$\frac{\mathrm{S}(I^{\mathrm{T}})}{\|\|\boldsymbol{y}_T\|\|^2}$	$\frac{\mathrm{S}(I^{\mathrm{C}})}{\|\|\boldsymbol{y}_C\|\|^2}$	AIC	k
LARS+EM	0.024	0.053	−863	9
Прореж.	0.013	0.034	−1109	15
Стохаст.	0.014	0.024	−1299	20
Lasso	0.011	0.092	−491	11
Шаг. регр.	0.032	0.086	−405	30
МГУА	0.018	0.080	−584	8
LARS	0.019	0.089	−579	5
Перебор R=5	0.024	0.036	–	5

няем алгоритм LARS. В результате будут получены векторы параметров $\boldsymbol{\beta}_{kl},\ l = 1, \ldots, M$.

За K шагов алгоритм находит KM моделей, выбирается оптимальная модель.

Вычислительный эксперимент

Сравнительный анализ алгоритмов был выполнен на исторических данных торгов опционом Brent Crude Oil [11]. Срок действия опциона — полгода, с 02.01.2001 по 26.06.2001. Тип опциона — put (право на продажу базового инструмента), символ CLG01. Базовый инструмент — нефть, символ NYM. Использовались ежедневные цены закрытия опциона и базового инструмента. Сетка цен исполнения опциона $\mathcal{K} = \{19.0, 19.5, \ldots, 28.0, 28.5\}$.

Регрессионная выборка

$$\big\{(\boldsymbol{x}_i, y_i)\big\}_{i=1}^m = \big\{(\langle K_i, t_i\rangle, \sigma_i)\big\}_{i=1}^m$$

была построена по исходным данным — историческим ценам опциона $C_{K,t}$ и базового инструмента P_t, где $K \in \mathcal{K},\ t \in T$, следующим образом. Для каждого значения K_i и t_i, $i = 1, \ldots, m$ вычислялось значение предполагаемой волатильности σ_i

$$\sigma_i = \arg\min_{\sigma\in[0,1.5]}\big(C_{K_i,t_i} - C(\sigma, P_{t_i}, B, K_i, t_i)\big),$$

где справедливая цена опциона C определялась по формуле Блэка-Шоулза [12]. Длина истории составляла 112 отсчетов времени.

Множество порождающих функций было задано следующим образом: $G = \big\{1/x, \sqrt{x}, e^x\big\}$.

Максимальная степень полинома Колмогорова-Габора была выбрана равной трём. При этом число мономов составило 84. Регрессионная выборка была случайным образом разбита на контрольную и обучающую, равные по мощности. Стандартизация контрольной и обучающей выборок была проведена раздельно. Значения ошибок на обучении и контроле были усреднены по 10 запускам алгоритмов на различных разбиениях.

Результаты экспериментов представлены в Таблице 1. Для каждого алгоритма были вычислены: ошибки $\mathrm{S}(I^{\mathrm{T}})$ и $\mathrm{S}(I^{\mathrm{C}})$ на обучении и контроле (2), отнесенные к квадрату нормы соответствующего вектора ответов $||\boldsymbol{y}_T||^2 = \sum\limits_{j\in I^{\mathrm{T}}} y_j^2$ и $||\boldsymbol{y}_C||^2 = \sum\limits_{j\in I^{\mathrm{C}}} y_j^2$; значение информационного критерия Акаике $AIC = m\ln\frac{S}{m} + 2k$; сложность модели k.

Заключение

В работе выполнено сравнение предложенных алгоритмов (стохастическая структурная оптимизация, модифицированный метод наименьших углов EM+LARS) с известными алгоритмами. Вычислительный эксперимент показал, что увеличение числа признаков позволяет добиться улучшения качества модели. Результаты экспериментов подтвердили жадность алгоритма LARS и бóльшую эффективность алгоритма EM+LARS по сравнению с LARS. По результатам экспериментов наилучшими по совокупности критериев являются EM+LARS и алгоритм оптимального прореживания.

Литература

[1] *Malada H. R., Ivakhnenko A. G.* Inductive Learning Algorithms for Complex Systems Modeling. CRC Press. 1994.

[2] *Efron B., Hastie T., Johnstone I., Tbshirani R.* Least Angle Regression // The Annals of Statistics. 2004. Vol. 32, No. 2. Pp. 407–499.

[3] *Ивахненко А. Г., Степашко В. С.* Помехоустойчивость моделирования. Киев: Наукова думка. 1985.

[4] *Tibshirani R.* Regression shrinkage and Selection via the Lasso // Journal of the Royal Statistical Society. 1996. Vol. 32, No. 1. Pp. 267–288.

[5] *Draper N., Smith H.* Applied Regression Analysis. John Wiley and Sons. 1981. Pp. 307–312.

[6] *Efroymson M. A.* Multiple regression analysis. Mathematical Methods for Digital Computers. Ralston, Wiley, New York. l960.

[7] *Rawlings J. O., Pantula S. G., Dickey D. A.* Applied Regression Analysis: A Research Tool. Springer-Verlag, New York. 1998.

[8] *Стрижов В. В.* Поиск параметрической регрессионной модели в индуктивно заданном множестве // Журнал вычислительных технологий. 2007. № 1. С. 93–102.

[9] *Хайкин С.* Нейронные сети, полный курс. М: Вильямс. 2008.

[10] *Mallows C. L.* Some Comments on C_p. Technometrics. 1973. No. 15. Pp. 661–675.

[11] *Ширяев А. Н.* Основы стохастической финансовой математики. Том 1. Факты. Модели. ФАЗИС. 2004.

[12] *Hull J. C.* Options, Futures and Other Derivatives. Prentice Hall. 2000.

Непараметрический алгоритм кластеризации для обработки больших массивов данных

Куликова Е. А., Пестунов И. А., Синявский Ю. Н.
budkinaea@mail.ru, pestunov@ict.nsc.ru, fox83@ngs.ru
Новосибирск, Институт вычислительных технологий СО РАН

В работе предлагается алгоритм кластеризации ССА, разработанный в рамках комбинации плотностного и сеточного подходов и позволяющий выделять многомодовые классы сложной формы. Результаты экспериментов на модельных и реальных данных подтверждают эффективность алгоритма. Быстродействие алгоритма позволяет проводить обработку данных в диалоговом режиме.

Введение

При решении прикладных задач в таких областях науки, как биоинформатика, генетика, астрономия, исследование Земли из космоса и т. п. часто возникает необходимость кластеризации больших массивов данных. При этом какая-либо априорная информация об искомых классах, как правило, отсутствует. В этих условиях целесообразно использовать непараметрические алгоритмы кластеризации, не требующие предположений о структуре данных. Основным недостатком таких алгоритмов является высокая вычислительная сложность (порядка $O(N^2)$, где N — объём выборки).

Одним из наиболее эффективных способов повышения быстродействия, активно развиваемым в последние годы, является переход от кластеризации отдельных объектов к обработке элементов сеточной структуры (так называемый сеточный подход) [1, 2, 3, 4]. Такие алгоритмы позволяют выделять классы сложной формы, однако они не всегда обеспечивают требуемое качество результатов при выделении классов, характеризующихся многомодовой плотностью распределения.

В работе предлагается алгоритм кластеризации ССА, разработанный в рамках комбинации плотностного и сеточного подходов и позволяющий выделять многомодовые классы сложной формы. Варьирование значения специального параметра позволяет получать результаты различной степени подробности.

Постановка задачи

Пусть множество классифицируемых объектов X состоит из векторов, лежащих в пространстве признаков $\mathbb{R}^d$: $X = \{x_i = (x_i^1, \ldots, x_i^d) \in \mathbb{R}^d, i = 1, \ldots, N\}$. Векторы x_i лежат в прямоугольном параллелепипеде

$$\mathfrak{B} = \left[\min_{x_i \in X} x_i^1, \max_{x_i \in X} x_i^1\right] \times \ldots \times \left[\min_{x_i \in X} x_i^d, \max_{x_i \in X} x_i^d\right].$$

Для формальной постановки задачи введём следующие определения.

Под *клеточной структурой* будем понимать разбиение пространства признаков гиперплоскостями:

$$x^j = \frac{i}{m}\rho^j + \min_{x_l \in X} x_l^j, \quad i = 0, \ldots, m,$$

где $\rho^j = \max\limits_{x_l \in X} x_l^j - \min\limits_{x_l \in X} x_l^j$, $j = 1, \ldots, d$, m — число разбиений $\mathfrak{B}$ по каждой размерности.

Минимальным элементом этой структуры является клетка (прямоугольный параллелепипед, ограниченный гиперплоскостями). Введём общую нумерацию клеток (последовательно от одного слоя клеток к другому).

Клетки B_i и B_j являются *смежными*, если их общая граница содержит хотя бы одну точку. Множество смежных с B клеток обозначим через A_B. *Плотностью* D_B клетки B назовём отношение

$$D_B = \frac{N_B}{V_B},$$

где N_B — количество элементов множества X, попавших в клетку B; V_B — объём клетки B.

Клетку B будем считать непустой, если $D_B \geqslant \tau$, где τ — величина заданного порога.

Непустая клетка B_i *непосредственно связна* с непустой клеткой B_j ($B_i \to B_j$), если B_j — максимальная по номеру клетка, удовлетворяющая условиям: $B_j = \arg\max\limits_{B_k \in A_{B_i}} D_{B_k}$ и $D_{B_j} > D_{B_i}$.

Непустые клетки B_i и B_j непосредственно связны ($B_i \leftrightarrow B_j$), если $B_i \to B_j$ или $B_j \to B_i$.

Непустые клетки B_i и B_j связны ($B_i \sim B_j$), если существуют $k_1, \ldots, k_l$ такие, что $k_1 = i$, $k_l = j$ и для всех $p = 1, \ldots, l-1$ выполнено $B_{k_p} \leftrightarrow B_{k_{p+1}}$.

Введение отношения связности порождает естественное разбиение множества непустых клеток на компоненты связности $\{G_i\}$. Под *компонентой связности* будем понимать максимальное множество попарно связных клеток.

Представителем компоненты связности G назовём максимальную по номеру клетку $Y(G)$, удовлетворяющую условию: $Y(G) = \arg\max\limits_{B \in G} D_B$.

Компоненты связности G' и G'' *смежные*, если существуют смежные клетки B' и B'' такие, что $B' \in G'$ и $B'' \in G''$.

Смежные компоненты связности G_i и G_j *непосредственно связны* $G_i \leftrightarrow G_j$, если существует набор клеток (путь) $P_{ij} = \{Y_i = B_{k_1}, \ldots, B_{k_l} = Y_j\}$ такой, что:

1) $B_{k_t} \in G_i \cup G_j$;

2) B_{k_t}, $B_{k_{t+1}}$ — смежные клетки;

3) $\min\limits_{B_{k_t}\in P_{ij}} D_{B_{k_t}} / \min\left(D_{B_{Y_i}}, D_{B_{Y_j}}\right) > T,\ T > 0$ — порог объединения.

Компоненты связности G_i и G_j *связны* ($G_i \sim G_j$), если существуют $k_1, \ldots, k_l$ такие, что $k_1 = i$, $k_l = j$ и для всех $p = 1, \ldots, l-1$ выполнено $G_{k_p} \leftrightarrow G_{k_{p+1}}$.

Определение 1. *Кластером C назовем максимальное множество попарно связных компонент связности, то есть*

1) *для любых компонент связности $G_i, G_j \in C$ выполнено $G_i \sim G_j$;*
2) *для любых $G_i \in C$ и $G_j \notin C : G_i \not\sim G_j$.*

Задача кластеризации заключается в разбиении исходного множества X на кластеры C_i, $i = 1, \ldots, M$, такие, что $X = \bigcup\limits_{i=1}^{M} C_i$ и $C_i \cap C_j = \varnothing$, $i \neq j$. Число кластеров M определяется в процессе обработки.

Описание алгоритма CCA

Ниже представлен алгоритм CCA, позволяющий выделять кластеры в соответствии с определением 1. В работе алгоритма можно выделить три основных этапа.

Этап 1. Формирование клеточной структуры. На этом этапе для каждой точки $x_i \in X$ определяется содержащая её клетка, вычисляются плотности клеток D_{B_i} и выявляются непустые клетки.

Этап 2. Выделение компонент связности $G_1, \ldots, G_S$ и поиск их представителей $Y(G_1), \ldots, Y(G_S)$. Краткое описание этапа приводится на рис. 1.

```
// Обозначения:
   Y — множество представителей компонент связности;
   Cell — вспомогательное множество клеток,
   принадлежащих одной компоненте;
   cNumber — номер текущей компоненты связности;
// Инициализация
   cNumber := 0;
// Формирование компонент связности и их представителей
   Для каждой непустой клетки B_i, которой не присвоен
   номер компоненты, выполнить:
      Cell := [];
      Выполнить Поиск_номера_компоненты(B_i, Cell);
Процедура Поиск_номера_компоненты(B, Cell):
   Найти непустую клетку B_j с наибольшей плотностью,
   смежную с B;
   Если B_j ≠ B и клетке B_j не присвоен номер
   компоненты, то:
      Cell ← B;
      Выполнить Поиск_номера_компоненты(B_j, Cell);
   Иначе:
      Если B = B_j, то:
         cNumber := cNumber + 1;
         Присвоить клетке B_j номер компоненты
         связности cNumber;
         Y ← B_j;
      Всем клеткам множества Cell присвоить номер
      компоненты связности, содержащей B_j;
```

Рис. 1. Краткое описание этапа 2.

```
Проверить непосредственную связность каждой пары
смежных компонент связности G_i и G_j;
// Инициализация:
   G_i ↮ G_j;
Пока G_i ↮ G_j и существуют нерассмотренные пары
смежных клеток (B', B''), где B' ∈ G_i, B'' ∈ G_j, проверить:
   Если min{D_B', D_B''} / min{D_Y_i, D_Y_j} > T, то:
      G_i ↔ G_j;
   Иначе:
      Рассмотреть следующую пару смежных клеток (B', B'');
Сформировать кластеры C_i согласно определению 1;
```

Рис. 2. Процедура обработки смежных компонент связности.

Этап 3. Формирование кластеров на основе выделенных компонент связности. Для этого каждую пару смежных компонент G_p и G_q обрабатываем по схеме, представленной на рис. 2.

В результате обработки получаем множество кластеров $\{C_1, \ldots, C_M\}$.

Вычислительная сложность алгоритма составляет $O\left(dN + dm^d\right)$.

Экспериментальное исследование

Далее приведены результаты экспериментов на модельных и реальных данных. Обработка проводилась на ПЭВМ на базе Intel Core 2 Duo E7300 с тактовой частотой 2,66 ГГц (объём оперативной памяти — 512 Мбайт).

Эксперимент 1. Использовалась модель «Бананы», состоящая из 20000 двумерных точек, сгруппированных в два линейно неразделимых класса. Модель построена с помощью инструментария [5] с параметром 0,6. На графике (рис. 3, *а*) представлена зависимость величины средней ошибки от N (для каждого значения N рассматривалось 20 случайно выбранных подмножеств X, содержащих одинаковое число представителей от каждого класса). На рис. 3 приведено подмножество, содержащее 200 точек (100 точек на класс) и результаты его обработки (*б* и *в* соответственно).

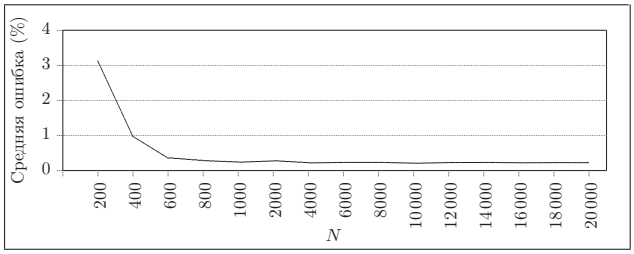
Эксперимент 2. Использовались трёхмерные данные, состоящие из 80 000 точек, сгруппированных в восемь классов C_1, …, C_8 по 10 000 точек, распределённых по нормальному закону. Векторы математических ожиданий:

$$\mu_1 = (0,0,0),\ \mu_2 = (0,-5,0),\ \mu_3 = (0,5,0),$$
$$\mu_4 = (14,-8,5),\ \mu_5 = (12,-6,6),\ \mu_6 = (10,-8,0),$$
$$\mu_7 = (13,4,-3),\ \mu_8 = (8,6,-3);$$

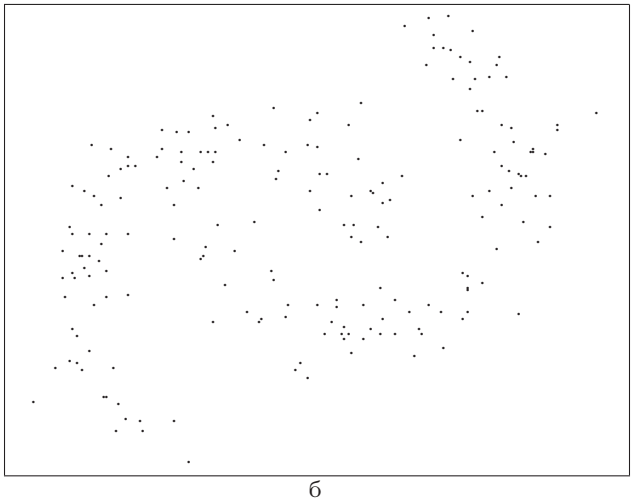
ковариационные матрицы:

$$\Sigma_i = E,\ i \in \{1,2,3,7,8\};\quad \Sigma_4 = \Sigma_5 = \Sigma_6 = E/2,$$

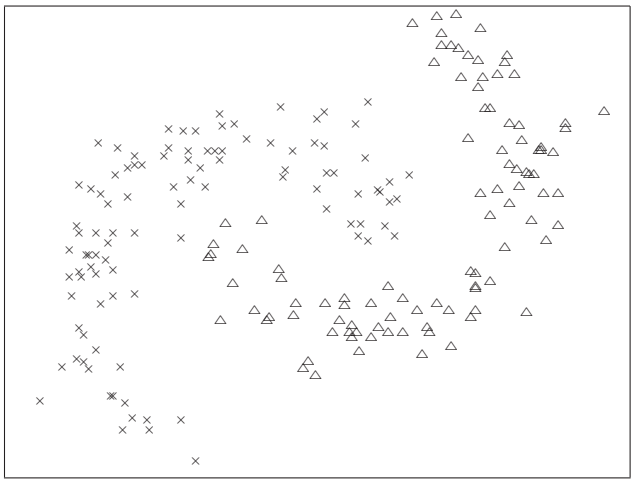
где E — трёхмерная единичная матрица. На рис. 4 представлена зависимость времени работы алгоритма CCA и величины средней ошибки от параметра m. Для каждого значения m подбиралось оптимальное значение параметра T.



а



б



в

Рис. 3. Результаты эксперимента 1. Параметры алгоритма CCA: $m = 13$, $T = 0.4$, $\tau = 0$.

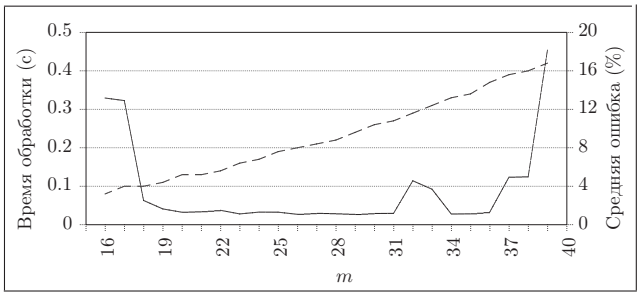


Рис. 4. Зависимость времени работы алгоритма (штриховая линия) и величины средней ошибки (сплошная линия) от параметра m.

Из рис. 4 видно, что при $m \leqslant 17$ средняя ошибка кластеризации превышает 3 %. Это объясняется слишком большим размером клеток. При $m \geqslant 32$ ошибка возрастает из-за слишком мелкого разбиения пространства, приводящего к раздробленности кластеров. При значениях $18 \leqslant m \leqslant 31$ достигается оптимальное для этой модели качество кластеризации (средняя ошибка не превышает 2 %).

Эксперимент 3. Эксперимент проводился на модели «Ирисы» [6] для оценки качества кластеризации. Множество данных состояло из 150 точек четырёхмерного пространства признаков, сгруппированных в 3 класса по 50 точек. Аналогично [2], за $|C_i^{\mathrm{O}}|$ обозначим фактическое количество точек i-го класса, $|C_i^{\mathrm{S}}|$ — число точек, содержащихся в соответствующем кластере, выделенном алго-

Таблица 1. Результаты эксперимента 3.

	$i = 1$	$i = 2$	$i = 3$
$\lvert C_i^{\mathrm{O}}\rvert$	50	50	50
$\lvert C_i^{\mathrm{S}}\rvert$	50	63	37
$\lvert C_i^{\mathrm{O}} \cap C_i^{\mathrm{S}}\rvert$	50	48	35
Точность (%)	100	76.2	94.6
Мера покрытия (%)	100	96	70

а

б

Рис. 5. Исходное изображение ($а$) и результаты обработки ($б$). Выделено 19 кластеров, которые для наглядности сгруппированы в 8 классов.

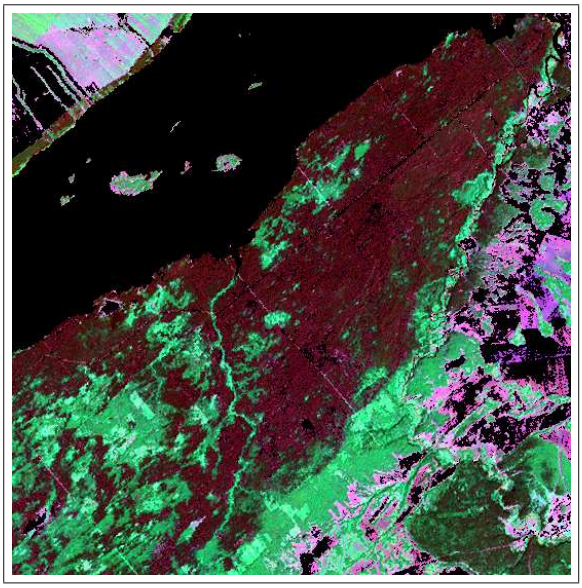
ритмом CCA. Затем вычислим точность кластеризации и меру покрытия кластерами C_i^{S} классов C_i^{O} по формулам $|C_i^{\mathrm{O}} \cap C_i^{\mathrm{S}}|/|C_i^{\mathrm{S}}|$ и $|C_i^{\mathrm{O}} \cap C_i^{\mathrm{S}}|/|C_i^{\mathrm{O}}|$ соответственно. В таблице 1 приводятся результаты выполнения алгоритма CCA на модели «Ирисы» с параметрами $m = 16$, $T = 0{,}8$, $\tau = 0$. Результаты обработки алгоритмом CCA превосходят результаты, полученные с использованием сеточного алгоритма GCOD [2] и алгоритма, описанного в [3].

Эксперимент 4. Эксперимент проводился на цветном изображении размером 1920×1200 пикселов, характеризующихся трёхмерными векторами признаков (компонентами палитры RGB). В результате кластеризации с параметрами $m = 30$, $T = 1$, $\tau = 0{,}01$ выделено 19 кластеров. Время работы алгоритма составило 0,5 с. Исходное изображение и результаты обработки представлены на рис. 5 (*а* и *б* соответственно).

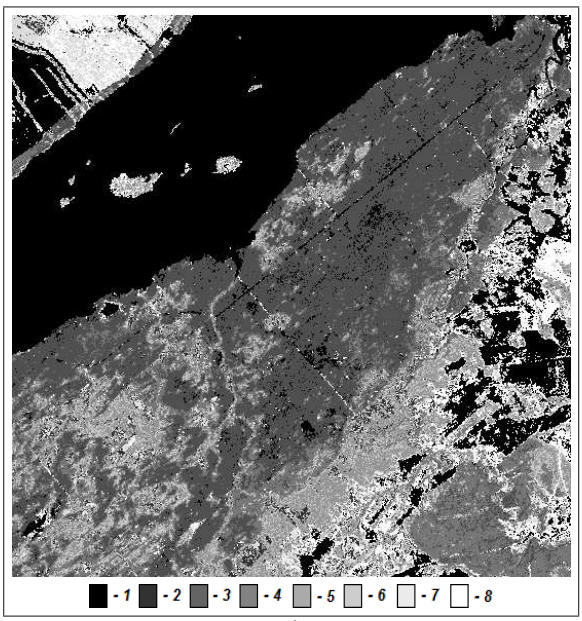
Эксперимент 5. Исследовался фрагмент снимка Караканского бора (Новосибирская область) размером 547×544 элементов разрешения, полученного со спутника LandSat 7 в июле 2002 г. Предварительно были исключены из рассмотрения не покрытые растительностью территории (с помощью пороговой сегментации по нормализованному вегетационному индексу (NDVI)). Обработка проводилась по четырём каналам $(3, 4, 5, 7)$. В результате кластеризации с параметрами $m = 45$, $T = 1$, $\tau = 0$ выделено 26 кластеров. Время работы алгоритма 1,33 с. Качество картосхемы оценивалось специалистами-дешифровщиками и было признано удовлетворительным. Исходное изображение и полученная картосхема представлены на рис. 6 (*а* и *б* соответственно).

Заключение

В работе предложен непараметрический сеточный алгоритм кластеризации CCA, позволяющий выделять кластеры сложной формы, описываемые многомодовой плотностью распределения. Результаты приведённых экспериментов на модельных и реальных данных подтверждают эффективность предложенного алгоритма. Быстродействие алгоритма позволяет проводить обработку данных в диалоговом режиме.



а



б

Рис. 6. Исходный фрагмент (*а*) и результаты кластеризации (*б*). Выделено 26 кластеров, образующих: 1 — территории, не покрытые растительностью; 2 — «сосновые насаждения»; 3 — «смешанный лес»; 4 — «вырубки»; 5 — «луг»; 6 — «березовые насаждения»; 7 — «скошенная трава»; 8 — класс-фон, составленный из кластеров, содержащих менее 0,03 % от общего числа точек.

Литература

[1] *Mercer D. P.* Clustering large datasets // Linacre College, 2003. — `http://www.stats.ox.ac.uk/~mercer/documents/Transfer.pdf`.

[2] *Qiu B.-Z., Li X.-L., Shen J.-Y.* Grid-based clustering algorithm based on intersecting partition and density estimation // PAKDD Workshops, 2007, Pp. 368–377.

[3] *Ma E. W. M., Chow T. W. S.* A new shifting grid clustering algorithm // Pattern Recognition. — 2004. — Vol. 37. — Pp. 503–514.

[4] *Lin N. P., Chang C.-I, Jan N.-Y. et al.* A deflected grid-based algorithm for clustering analysis // Inter. Journal of mathematical models and methods in applied sciences. — 2007. — Vol. 1, is. 1. — Pp. 33–39.

[5] PRTools: the Matlab Toolbox for Pattern Recognition. — `http://www.prtools.org`.

[6] *Asuncion A., Newman D. J.* UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science, 2007 — `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

Метод распознавания редких событий*

Лбов Г. С., Герасимов М. К.
lbov@math.nsc.ru, max_post@ngs.ru
Новосибирск, Институт математики СО РАН

В работе предложен метод построения модели для распознавания (прогнозирования) редких событий и явлений. Предполагается, что редкие события могут описываться как количественными, так и качественными переменными. При анализе разнотипной информации введено взвешенное расстояние (мера близости). Для нахождения оптимальных весов предложен метод адаптивного поиска приближенного значения глобального экстремума функции на симплексе.

При анализе и прогнозировании реальных редких событий и явлений следует учитывать следующие особенности:

1) методы прогнозирования должны использовать комплексное описание, содержащее как можно более полную информацию обо всех факторах, потенциально влияющих на возникновение редких событий;
2) реальные данные могут содержать не только количественную, но и качественную информацию;
3) количество соответствующих редким событиям прецедентов в эмпирической информации мало́ по отношению к общему объёму выборки.

Постановка задачи

Пусть объекты из некоторой генеральной совокупности Γ описываются набором переменных $X = \{X_1, \ldots, X_n\}$. Данный набор может одновременно содержать как количественные, так и качественные переменные. Каждому объекту исследования $a \in \Gamma$ поставлены в соответствие номер образа $Y(a) \in \{1, \ldots, K\}$ и набор значений $X(a) = \big(X_1(a), \ldots, X_n(a)\big)$, где $X_j(a)$ — значение переменной X_j для объекта a. Обозначим через D_j множество возможных значений переменной X_j. Декартово произведение $D = \prod_{j=1}^{n} D_j$ задаёт многомерное пространство разнотипных переменных.

Без ограничения общности будем предполагать, что распознаются два образа, $K = 2$, а редкие события относятся к первому образу. Предполагается, что количество наблюдений редких событий в эмпирической информации мало́ по отношению к общему объёму выборки, а возможные потери от ошибочного предсказания второго образа достаточно велики.

Обозначим множество номеров объектов первого образа через I_1, второго образа — через I_2:

$$I_1 = \big\{i\colon Y(a^i) = 1\big\}, \quad I_2 = \big\{i\colon Y(a^i) = 2\big\}.$$

Пусть для нового объекта $a^* \in \Gamma$ измерены наблюдения по набору X. Необходимо отнести объект a^* к тому или иному образу, то есть оценить $Y(a^*)$. Предполагается, что распределение $\mathsf{P}(X, Y)$ неизвестно, следовательно, необходимо определить принадлежность объекта на основе анализа обучающей выборки.

Критерий распознавания редких событий

Применение методов распознавания образов для определения редких событий в данной постановке задачи затруднительно ввиду малого числа соответствующих прецедентов в обучающей выборке. Для решения задачи будем использовать взвешенное расстояние между объектами a^i и a^l:

$$\rho^{il} = \sqrt{\sum_{j=1}^{n} \lambda_j \left(\rho_j^{il}\right)^2},$$

где $\sum_{j=1}^{n} \lambda_j = 1$, $\lambda_j \geqslant 0$, $j = 1, \ldots, n$; ρ_j^{il} — расстояние между объектами a^i и a^l по j-й компоненте набора X, задаётся в зависимости от типа переменной X_j: если X_j — количественная, то $\rho_j^{il} = \big|X_j(a^i) - X_j(a^l)\big|$; если X_j — качественная, то $\rho_j^{il} = 0$ при $X_j(a^i) = X_j(a^l)$ и $\rho_j^{il} = 1$ при $X_j(a^i) \neq X_j(a^l)$.

Будем использовать следующую гипотезу: предполагается, что существует такой набор значений коэффициентов λ_j («весов»), при котором объекты первого образа, относящиеся к редким событиям, расположены компактно в пространстве D, при этом достаточно сильно отличаясь от объектов второго образа. Исходя из этой гипотезы, можно поставить следующую оптимизационную задачу: необходимо подобрать коэффициенты λ_j таким образом, чтобы минимизировать величину

$$r_1(\lambda_1, \ldots, \lambda_n) = \sum_{i,k \in I_1\colon\, i<k} \rho^{ik},$$

одновременно максимизируя величину

$$r_2(\lambda_1, \ldots, \lambda_n) = \sum_{i \in I_1} \sum_{k \in I_2} \rho^{ik}.$$

*Работа выполнена при финансовой поддержке РФФИ, проект №07-01-00331а.

В данной работе будем минимизировать следующий критерий:

$$f(\lambda_1,\ldots,\lambda_n)=\frac{r_1(\lambda_1,\ldots,\lambda_n)}{c+r_2(\lambda_1,\ldots,\lambda_n)},\quad c=\text{const}>0.$$

Определив взвешенное расстояние в многомерном пространстве, можно по степени близости к объектам изучения оценить принадлежность новых объектов к тому или иному образу, например, методом ближайшего соседа.

Подобного рода задача оптимизации возникает также и в предложенном в работе [1] методе прогнозирования экстремальных ситуаций, которые тоже по своей сути являются редкими событиями.

Заметим, что область значений коэффициентов $\lambda_1,\ldots,\lambda_n$ (обозначим её через Λ) представляет собой многомерный симплекс

$$\Lambda=\Big\{\lambda=(\lambda_1,\ldots,\lambda_n)\colon \sum_{j=1}^{n}\lambda_j=1,\ \lambda_j\geqslant 0,\ j=1,\ldots,n\Big\}\subset\mathbb{R}^n.$$

При других способах задания расстояния (меры близости) между объектами могут быть получены другие многомерные области.

Подбор оптимальных весов

Поскольку свойства функции $f(\lambda)$ в прикладных задачах распознавания редких событий зависят от случайной выборки, то при решении указанной оптимизационной задачи можно использовать лишь общие свойства класса $\{f(\lambda)\}$ в виде некоторой «разумной» гипотезы, приведённой ниже.

Для нахождения оптимальных коэффициентов будем использовать модификацию метода поиска приближенного значения глобального экстремума функции [2, 3]. Данный метод является развитием метода СПА (случайного поиска с адаптацией) [4] для поиска наиболее информативного подпространства признаков в задачах распознавания образов. Заметим, что последующим обобщением метода СПА являются генетические алгоритмы.

Рассмотрим функцию $f(\lambda),\ \lambda\in\Lambda$. Обозначим $\lambda^*=\arg\min\limits_{\lambda\in\Lambda} f(\lambda)$. Пару $\langle\lambda^i,f(\lambda^i)\rangle$ будем называть испытанием. Под алгоритмом поиска приближенного значения глобального минимума функции понимается некоторая процедура последовательного планирования N испытаний с целью нахождения минимально возможного значения функции.

Общее число испытаний N разбивается на R групп: $N=N_1+\ldots+N_R$.

Обозначим через $N^\nu=\sum\limits_{i=1}^{\nu}N_i$ — число испытаний, проведенных за ν шагов поиска (под шагом поиска понимается проведение N_i испытаний), $\nu=1,\ldots,R$. Первая группа испытаний планируется таким образом, чтобы для любого $E\subset\Lambda$ число испытаний, соответствующее множеству E, было бы примерно равно $\frac{N_1V(E)}{V(D)}$, где $V(E)$ — объём области E. Другими словами, точки $\lambda^1,\ldots,\lambda^{N_1}$ планируем таким образом, чтобы они были бы максимально равномерно распределены по множеству Λ.

Для вычисления объёмов в предлагаемом методе используется метод Монте-Карло. Как показано в работе [5], метод Монте-Карло эффективен для вычисления многомерных интегралов, в частности — объёмов. Кроме того, метод Монте-Карло позволяет вычислять объёмы и для более сложных, чем симплекс, областей.

Пусть проведено N^ν испытаний, $\nu=1,\ldots,R-1$, и получена соответствующая таблица $v^\nu=\{\lambda^i,f(\lambda^i)\},\ i=1,\ldots,N^\nu$. С помощью алгоритма LRP (см., например, [2]) построим наилучшую регрессионную функцию $\bar{f}^\nu$ в классе логических решающих функций. Функции $\bar{f}^\nu$ соответствует некоторое разбиение множества Λ: $a_\nu=\{E_\nu^1,\ldots,E_\nu^{M_\nu}\}$. Множества E_ν^t имеют вид

$$E_\nu^t=\big\{\lambda\in\Lambda\colon c_j\leqslant\lambda_j\leqslant d_j,\ \ j=1,\ldots,n\big\}.$$

Будем использовать следующую гипотезу: вероятность достижения глобального минимума функции в области E зависит от числа проведённых испытаний, объёма множества E и результатов проведённых ранее в точках множества E испытаний. Предполагается, что чем больше объём множества E (а, значит, и неисследованность) и меньше по сравнению с другими областями полученные при испытаниях значения функции, тем больше вероятность достижения глобального минимума в этом множестве.

Введём вспомогательную функцию

$$\varkappa_b(z)=\frac{b+1}{(1+bz)^2},\quad 0\leqslant z\leqslant 1,\quad 0\leqslant b\leqslant\infty.$$

Выбор этой функции определяется следующими её свойствами:

1) при $b=0$ получаем $\varkappa_b(z)=1$;
2) при $b>0$ получаем монотонно убывающую функцию по z;
3) чем больше значение параметра b, тем больше скорость убывания функции по z;
4) $\int\limits_0^1\varkappa_b(z)dz=1$ при любом b.

Данные свойства функции $\varkappa_b(z)$ позволяют формализовать вышеуказанную гипотезу. Отметим, что в качестве функции $\varkappa_b(z)$ можно выбрать любую функцию, удовлетворяющую этим свойствам.

Зададим функцию

$$\varkappa^u = \int_0^u \varkappa_b(z)dz = \frac{(b+1)u}{1+bu}, \quad 0 \leqslant u \leqslant 1.$$

Определим вероятности p_ν^t проведения испытания в множествах E_ν^t, $t = 1, \ldots, M_\nu$. Обозначим через $f_{\min}^t$ минимальное значение функции f при проведённых испытаниях в множестве E_ν^t. Пусть $f_{\min}^{t_1} \leqslant \ldots \leqslant f_{\min}^{t_{M_\nu}}$. По оси z будем откладывать последовательно величины z^i, равные относительным объёмам соответствующих множеств:

$$z^i = \frac{V(E_\nu^{t_i})}{V(D)}, \quad i = 1, \ldots, M_\nu,$$

то есть первоначально откладывается относительный объём наилучшего множества $E_\nu^{t_1}$, которому соответствует $f_{\min}^{t_1}$, затем относительный объём второго по порядку множества $E_\nu^{t_2}$, которому соответствует $f_{\min}^{t_2}$ и т. д. Таким образом, отрезок $[0, 1]$ на оси z разбивается на M_ν отрезков. Вероятности p_ν^t определяются по формуле

$$p_\nu^t = \varkappa^{u_{t+1}} - \varkappa^{u_t} = \int_0^{u_{t+1}} \varkappa_b(z)\,dz - \int_0^{u_t} \varkappa_b(z)\,dz =$$

$$= \frac{(b+1)(u_{t+1} - u_t)}{(1+bu_{t+1})(1+bu_t)}, \quad t = 1, \ldots, M_\nu.$$

Таким образом, чем больше объём множества E_ν^t и меньше номер i, определяющий номер этого множества в порядке $E_\nu^{t_1}, \ldots, E_\nu^{t_{M_\nu}}$, тем больше вероятность p_ν^t (в соответствии с указанной выше гипотезой). Зададим линейную зависимость параметра b от числа проведенных испытаний N^ν, то есть на $\nu + 1$ шаге поиска будем использовать величину $b^\nu = \frac{N^\nu}{N} b_{\max}$. Величина $b_{\max}$ определяется из следующих соображений. После проведения всех испытаний ($N^R = N$) в соответствии с гипотезой можно указать область $E(\gamma)$ такую, что вероятность нахождения точки λ^* равна $p(\gamma) \simeq 1$. При этом область $E(\gamma)$ достаточно мала, то есть $\gamma = \frac{V(E(\gamma))}{V(D)} \simeq 0$. Например, $p(\gamma) = 0{,}95$, а $\gamma = 0{,}05$. Величина $b_{\max}$ определяется из следующего соотношения:

$$p(\gamma) = \int_0^\gamma \varkappa_b(z)\,dz = \frac{(b_{\max}+1)\gamma}{1 + b_{\max}\gamma}.$$

Предложенный метод поиска приближенного значения глобального экстремума функции может быть использован для любого множества, где возможно эффективное использование метода Монте-Карло. Таким образом, область оптимизации может представлять собой не только многомерный симплекс, но и, например, гиперсферу.

Пример решения прикладной задачи

Разработанные в рамках данного подхода методы были использованы для прогноза экстремальных ситуаций на реках Сибири [1]. В частности, оценивалось возникновение в марте экстремальной по маловодью ситуации в контрольной точке «Барнаул». Были обработаны среднемесячные данные замеров стока реки Обь за период с 1937 по 1990 гг. Использовались следующие дополнительные характеристики: температура воздуха и количество выпавших осадков (станции — «Онгудай» и «Волчиха») за сентябрь и октябрь предыдущего года. Данные с 1991 по 2000 гг. использовались для контроля. За этот период было одно маловодье (1998 г.). Был сделан верный прогноз маловодья на 1998 год и неверный на 1999 год. В остальные годы правильно предсказано отсутствие маловодья. Необходимо подчеркнуть, что имеющаяся информация представляет собой числовые (количественные) временные ряды, тогда как разработанные методы могут быть использованы для решения более общих задач, в которых информация об исследуемых событиях описывается как количественными, так и качественными характеристиками.

Выводы

В работе предложен метод распознавания (прогнозирования) редких событий. В методе используется тот факт, что количество наблюдений редких событий в эмпирической информации мало. Определив взвешенное расстояние в многомерном пространстве, можно по степени близости к объектам изучения оценить принадлежность новых объектов к тому или иному образу. Результаты решения прикладных задач показывают эффективность метода прогнозирования редких событий.

Литература

[1] *Лбов Г. С., Герасимов М. К.* Прогнозирование экстремальных ситуаций на основе совместного анализа временных рядов и экспертных высказываний // Научный вестник НГТУ. — 2007. — № 3(28). — С. 13–24.

[2] *Лбов Г. С.* Методы обработки разнотипных экспериментальных данных. — Новосибирск: Наука, 1981.

[3] *Лбов Г. С., Старцева Н. Г.* Логические решающие функции и вопросы статистической устойчивости решений. — Новосибирск: Издательство Института математики, 1999. — 212 с.

[4] *Лбов Г. С.* Выбор эффективной системы зависимых признаков // Вычислительные системы. — 1965. — Вып. 19. — С. 21–34.

[5] *Бусленко Н. П., Голенко Д. И., Соболь И. М., Срагович В. Г., Шрейдер Ю. А.* Метод статистических испытаний (метод Монте-Карло). — Москва: Физматлит, 1962. — 332 с.

О согласованной нормировке набора метрик на основе модели оптимального коллективного слагаемого*

Майсурадзе А. И.
maysuradze@cs.msu.ru
Москва, МГУ им. М. В. Ломоносова

В работе предлагается метод согласованной нормировки набора метрик над одним и тем же конечным множеством. Идея метода состоит в том, чтобы каждая метрика из набора давала активное ограничение при выделении оптимального коллективного слагаемого. Показано, что в модифицированной проблеме поиска оптимального коллективного слагаемого построение ограничений, проверка активности ограничений и вычисление нормировочного множителя могут быть выполнены вычислительно эффективным образом.

В настоящее время в распознавании образов и интеллектуальном анализе данных широкое распространение получило использование различных мер сходства, т. е. описание объектов распознавания оказывается связанным с попарным сравнением этих объектов между собой. Во многих задачах меры сходства вычисляются и используются только для конечного множества объектов распознавания, в таком случае принято говорить о конфигурациях сходства. Если результатом попарного сравнения объектов является числовая оценка, которая интерпретируется как расстояние между этими объектами, то соответствующую конфигурацию сходства называют метрической конфигурацией (МК).

Во многих теоретических и прикладных задачах распознавания образов, интеллектуального анализа данных и прогнозирования не существует некоторого единственного «объективного» способа наделить множество объектов метрикой. В подобной ситуации одним из приёмов является получение информации об одних и тех же свойствах одного и того же набора объектов из различных источников. В частности, на одном и том же множестве объектов одновременно вводится несколько метрик.

Известно, что для многих метрических методов интеллектуального анализа данных результаты обработки не зависят от масштаба используемой метрики. Однако при синтезе новых метрик из нескольких заданных бывает необходимо согласовать масштабы исходных метрик. Разумеется, можно использовать некоторые простые идеи одинаковой нормировки каждой из метрик набора. Например, нормировать метрики так, чтобы среднее расстояние в каждой метрике стало равно единице. Но подобные идеи, по сути, работают с каждой метрикой отдельно и не учитывают набор целиком. Таким образом, существует проблема разработки методов, в которых каждая метрика набора нормируется с учётом информации о всех остальных метриках этого набора. Такую нормировку можно назвать согласованной.

Модель коллективного слагаемого

Рассмотрим задачу выделения «общей» метрики, в определённом смысле содержащейся во всех исходных метриках набора. В [1] показано, что метрические конфигурации можно отождествить с элементами специального линейного векторного пространства. Таким образом, на множестве метрических конфигураций оказываются заданы операции сложения и умножения на число. Следуя [2], для формализации задачи выделения «общей» метрики предлагается рассмотреть разложение исходных метрик на коллективное слагаемое и индивидуальные поправки.

Пусть для некоторого конечного множества объектов распознавания (мощности m) имеется набор метрических конфигураций $r_1, \ldots, r_n$. Рассматривается задача разложения каждой МК r_i из заданного набора на общее «коллективное слагаемое» s и индивидуальную поправку e_i:

$$r_1 = s + e_1, \ \ldots, \ r_n = s + e_n. \qquad (1)$$

Коллективное слагаемое и индивидуальные поправки выбираются из того же пространства МК, в котором лежат исходные $r_1, \ldots, r_n$. Данному разложению можно приписать следующую интерпретацию: коллективное слагаемое соответствует прикладной области в целом, а индивидуальные поправки — отдельным источникам или способам получения исходных МК. Отметим, что нередко исходные метрики оказываются довольно похожи друг на друга, даже если имеют существенно разную интерпретацию. Это связано с тем, что сравниваются объекты из некоторой узкой предметной области, хотя метрики могли бы сравнить гораздо более широкий класс объектов. Например, во многих задачах обработки изображений требуется сравнивать не произвольные наборы пикселов, а изображения достаточно узкого круга предметов в довольно ограниченном числе ракурсов.

Очевидно, что в линейном векторном пространстве для разложения (1) в общем случае может быть использована произвольная МК s. Поэтому

*Работа выполнена при финансовой поддержке грантов МК-2252.2008.9 и РФФИ №07-01-00211-а.

на коллективное слагаемое и индивидуальные поправки следует наложить ограничения. В модели в качестве указанных ограничений на s и $e_1, \ldots, e_n$ используются требования выполнения аксиом полуметрики.

Чтобы некоторая МК удовлетворяла аксиомам полуметрики, требуется выполнение неравенств треугольника. Заметим, что в рассматриваемой в данной работе ситуации неотрицательность элементов МК следует из выполнения неравенств треугольника. Множество МК, удовлетворяющих аксиомам полуметрики, будем называть полуметрическим конусом и обозначать MET. Таким образом, рассматриваемые в работе ограничения имеют следующий вид:

$$s \in MET,\ e_1 \in MET, \ldots, e_n \in MET, \qquad (2)$$

где запись $x \in MET$ означает, что МК x удовлетворяет аксиомам полуметрики. В работе [3] получен критерий разрешимости.

Теорема 1. *Задача поиска разложения (1) с ограничениями вида (2) разрешима тогда и только тогда, когда все исходные взвешенные клики $r_1, \ldots, r_n$ удовлетворяют аксиомам полуметрики.*

Поскольку коллективное слагаемое s однозначно определяет разложение, то ниже под множеством решений задачи поиска разложения будем понимать множество значений, которые может принимать s.

Ограничения (2) представляют собой систему из $(n+1)m(m-1)(m-2)/2$ линейных неравенств. Все эти ограничения можно переформулировать как ограничения только на s. При этом существенными окажутся только $m(m-1)(m-2)$ неравенств, которые можно найти, затратив порядка $O(nm^3)$ арифметических операций. В [1] было получено описание структуры множества решений задачи (1,2).

Следуя интерпретации разложения (1), обычно при использовании современных методов интеллектуального анализа данных необходимо найти не произвольное решение задачи (1,2), а такое, в котором коллективное слагаемое является в некотором смысле наибольшим, наиболее удаленным от нулевой МК. Формализация указанного понятия неоднозначна, она обычно приводит к поиску оптимального по Парето коллективного слагаемого или введению некоторого функционала «абсолютной величины» на МК, переходу к задаче линейного или квадратичного программирования. Сложная структура экстремальных лучей полуметрического конуса приводит к тому, что решение задачи оптимизации может быть неединственным и требует тщательного выбора вычислительных методов.

Модификация модели коллективного слагаемого

Чтобы упомянутые выше сложности можно было обойти, изменим ограничения, определяющие задачу поиска разложения. В [1] описан специальный «векторный» способ представления МК, связанный с линейным разложением произвольной МК r по фиксированной системе МК K. Для некоторых семейств систем МК переход к соответствующему представлению r' от исходных расстояний и обратно может быть выполнен всего за $O(m^2)$ арифметических операций, причём достаточным условием выполнения аксиом полуметрики будет неотрицательность представления: $r' \geqslant 0 \Longrightarrow r \in MET$, которую можно тривиальным способом проверить также за $O(m^2)$ операций. Отметим, что проверка неравенств треугольника для одной МК требует $O(m^3)$ арифметических операций, а переход к векторному представлению в общем случае мог бы потребовать $O(m^4)$ операций.

Зафиксируем некоторую подходящую систему МК K и потребуем неотрицательности представления для коллективного слагаемого и индивидуальных поправок:

$$s' \geqslant 0,\ e'_1 \geqslant 0, \ldots, e'_n \geqslant 0. \qquad (3)$$

Ограничения (3) сильнее ограничений (2). Это означает, что любое решение задачи (1,3) будет решением задачи (1,2). Для задачи (1,3) в работе [3] получен критерий разрешимости.

Теорема 2. *Задача поиска разложения (1) с ограничениями вида (3) разрешима тогда и только тогда, когда представления всех исходных МК $r_1, \ldots, r_n$ неотрицательны.*

Если рассматривать систему K как базис в специальном линейном векторном пространстве МК, то множество решений (допустимых s') задачи (1,3) имеет достаточно простую структуру: это многомерный параллелепипед с рёбрами, параллельными координатным осям. Представление нулевой МК (начало координат в специальном пространстве) является вершиной этого параллелепипеда. В указанной ситуации практически очевидной представляется идея в качестве оптимального решения брать вершину параллелепипеда d', находящуюся на противоположном от представления нулевой МК конце большой диагонали указанного параллелепипеда. Действительно, d' удовлетворяет требованию наибольшей удалённости от представления нулевой МК во всех указанных выше смыслах.

Для поиска d' требуется всего порядка $O(nm^2)$ операций. Оптимальное решение s задачи (1,3) восстанавливается из представления d' за $O(m^2)$ операций и гарантированно удовлетворяет аксиомам полуметрики.

Согласованная нормировка

Можно найти координаты оптимальной вершины d'. Если через $r'(k)$ обозначить k-ю координату векторного представления МК r, то формула для вычисления $d'(k)$ примет вид

$$d'(k) = \min\bigl\{r'_1(k), \ldots, r'_n(k)\bigr\}. \qquad (4)$$

При этом множество всех допустимых представлений s' решений задачи (1,3) характеризуется условиями $0 \leqslant s' \leqslant d'$, определяющими указанный выше параллелепипед решений.

Как видно из (4), в общем случае d' может зависеть не от всех МК исходного набора, а лишь от некоторой их части, доставляющей минимум по каждой из координат. Отметим, что в прикладных задачах число исходных метрик обычно существенно меньше, чем размерность специального пространства МК.

В качестве требования согласованности нормировок потребуем, чтобы каждая МК из набора давала существенное ограничение хотя бы по одной координате. Тогда после нормировки все МК будут располагаться на границе области доминирования вершины d', т. е. на границе области $\{x\colon x \geqslant d'\}$. Быть вне этой области представления МК не могут, т. к. это противоречило бы условиям (3). Таким образом, предложенное условие согласования нормировок можно интерпретировать как требование перехода к критическим значениям индивидуальных поправок.

После того как вектор d' найден, процедура вычисления нормирующих множителей для каждой МК набора становится достаточно проста. Векторное представление r'_i МК r_i покоординатно делится на вектор d': $t_i = r'_i \oslash d'$. По построению d' гарантируется, что $t_i(k) \geqslant 1$ для всех k. Искомый нормирующий множитель n_i равен обратной величине от минимального компонента вектора t_i:

$$n_i = \frac{1}{\min\limits_k \dfrac{r'_i(k)}{d'(k)}}.$$

Для каждой МК набора нормирующий множитель можно найти за $O(m^2)$ арифметических операций.

Отметим, что после указанной нормировки оптимальное коллективное слагаемое в смысле задачи (1,3) не меняется.

Заключение

Для формализации идеи выделения «общей» метрики из набора ставится задача поиска оптимального разложения на коллективное слагаемое и индивидуальные поправки. Проводится модификация задачи, существенно снижающая сложность поиска оптимального решения. Для исходной и модифицированной задач получены критерии разрешимости, описаны множества допустимых решений, даны оценки сложности решения. Метод решения модифицированной задачи позволяет определить нормирующие множители и вычислительно эффективно найти их значения.

Литература

[1] *Майсурадзе А.И.* Гомогенные и ранговые базисы в пространствах метрических конфигураций // Ж. вычисл. матем. и матем. физ. — 2006. — Т. 46, № 2. — С. 344–361.

[2] *Майсурадзе А.И.* О построении оптимальной коллективной метрики по набору произвольных взвешенных клик // Дискретные модели в теории управляющих систем: VII Международная конф. Труды. — М.: МАКС Пресс, 2006. — С. 239–241.

[3] *Майсурадзе А.И.* О поиске оптимального коллективного слагаемого для набора метрических конфигураций // Искусственный интеллект — 2006. — № 2. — С. 146–150.

Выбор опорного множества при построении устойчивых интегральных индикаторов*

Мельников Д. И., Стрижов В. В., Андреева Е. Ю., Эденхартер Г.
strijov@ccas.ru
Москва, Вычислительный центр РАН
Берлин, Технический университет

Исследуется задача построения интегрального индикатора множества объектов, устойчивого к выбросам в описаниях объектов. Объекты описаны в линейных шкалах. Для построения интегрального индикатора из множества всех описаний с помощью критерия принадлежности выбирается множество опорных описаний. Интегральный индикатор строится методом «без учителя». Предложенный алгоритм использован для получения интегрального индикатора уровня загрязнений основных продуктов питания в регионах России.

Введение

Построение интегрального индикатора — введение отношения порядка на множестве сравнимых объектов. Выбор алгоритма построения индикатора зависит от тех свойств, которыми обладают объекты. Предполагается, что каждый объект описан вектором, компоненты которого являются результатами измерений соответствующих показателей. Все измерения выполнены в линейных шкалах. Интегральный индикатор — скаляр, поставленный в соответствие объекту. Говоря о наборе объектов, будем называть интегральным индикатором вектор, компоненты которого поставлены в соответствие сравниваемым объектам.

Распространенным алгоритмом построения интегральных индикаторов для объектов, описанных в линейных шкалах, является линейная комбинация значений показателей [1]. Веса при этом вычисляются исходя из некоторого заданного критерия информативности описаний. Принятый в данной работе критерий наибольшей информативности, введенный С. Р. Рао, рассмотрен в первом разделе в связи с методом главных компонент. Однако этот метод вызывает, при наличии выбросов в описаниях объектов, проблему адекватной сравнимости объектов. Эксперты, определяющие множество объектов, предполагают все объекты сравнимыми и ожидают от алгоритма адекватные значения интегральных индикаторов. Однако если некоторые отдельные объекты имеют значения показателей, существенно отличающиеся от значений показателей основного числа объектов, то, в рамках линейной модели, объекты-выбросы имеют большее влияние на веса показателей, чем прочие объекты. При исключении таких объектов можно наблюдать изменение значений индикаторов, существенное не только для линейных, но даже и для ранговых шкал.

Ранее были предложены алгоритмы получения устойчивых интегральных индикаторов с использованием как линейных [2], так и нелинейных моделей [3, 4].

В данной работе исследуется задача построения устойчивых интегральных индикаторов. Решением этой задачи является алгоритм построения индикатора для всего множества объектов, построенный на основе его подмножества, называемого *опорным множеством*. Алгоритм разделяет исходное множество описаний объектов на два подмножества — опорное и множество выбросов. При этом используется критерий вероятности принадлежности описаний объекта одному из двух подмножеств. По опорному множеству, с помощью метода главных компонент, вычисляются веса. Эти веса используются для получения интегральных индикаторов всей выборки.

Алгоритм построения интегральных индикаторов

Задано множество, состоящее из m объектов, которые описаны набором из n показателей. Задана матрица описаний $A \in \mathbb{R}^{m\times n}$. Элемент матрицы a_{ij} — значение j-го показателя i-го объекта. Вектор $\boldsymbol{a}_i = (a_{i1}, \ldots, a_{in})$ — описание i-го объекта.

Интегральный индикатор объекта — это свертка вида

$$q_i = \sum_{j=1}^{n} w_j g_j(a_{ij}), \tag{1}$$

где g_j — функция приведения показателей в единую шкалу:

$$g_j : a_{ij} \mapsto (a_{ij} - \min_i a_{ij})(\max_i a_{ij} - \min_i a_{ij})^{-1}, \quad i = 1, \ldots, m,\ j = 1, \ldots, n. \tag{2}$$

Если в формуле (2) знаменатель равен нулю, то это означает, что значения j-го показателя для всех объектов равны. При этом показатель не может быть использован для построения интегрального индикатора и должен быть исключен из дальнейшего рассмотрения.

Без ограничения общности будем считать, что выполнено условие монотонности такое, что

*Работа выполнена при финансовой поддержке РФФИ, проекты №07-07-00181, 08-01-12022.

из $a_{ij} \geqslant a_{\xi j}$ следует $q_i \geqslant q_\xi$ для $j = 1, \ldots, n$. Выполнение этого условия вместе с выполнением (2) влечет неотрицательность значений $w_1, \ldots, w_n$. Так как на практике выставляется требование инвариантности интегрального индикатора к линейным преобразованиям, введем еще одно условие, накладываемое на веса: $\sum\limits_{j=1}^{n} w_j^2 = 1$.

Выполнение вышеперечисленных условий включено в предварительную обработку данных с целью их приведения в соответствие с принципом «чем больше, тем лучше». Исходя из этого принципа, эксперт ожидает, что увеличение значения некоторого показателя объекта приведет к увеличению его интегрального индикатора. Объект, имеющий максимальный по значению интегральный индикатор, называется наилучшим, а показатель, имеющий максимальный по значению вес, называется важнейшим в произвольных подмножествах соответственно объектов и показателей.

Результатом работы алгоритма построения интегрального индикатора методом «без учителя» является отыскание оптимального, по отношению к критерию информативности, вектора весов $\boldsymbol{w} = (w_1, \ldots, w_n)^\intercal$ свертки (1). Рассмотрим алгоритм получения интегрального индикатора «без учителя». Метод главных компонент, используемый для вычисления интегральных индикаторов [5], заключается в том, что к множеству описаний объектов применяется преобразование вращения, которое соответствует критерию *наибольшей информативности* С. Р. Рао [6]. Согласно этому критерию, наибольшая информативность есть минимальное значение суммы квадратов расстояния от описаний объектов до их проекций на первую главную компоненту.

Наилучшим выбором линейных функций, для которых остаточная дисперсия, предсказания с помощью линейного предиктора, минимальна, является выбор первых k главных компонент случайной величины A.

Для нахождения первой главной компоненты требуется найти такие линейные комбинации $Z^\intercal = WA^\intercal$ векторов-столбцов матрицы A, что векторы-столбцы $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ матрицы Z обладали бы наибольшей дисперсией: $\max \sum\limits_{j=1}^{n} \mathsf{D}\boldsymbol{z}_j$ при ограничениях нормировки $WW^\intercal = I$ — единичная матрица. Рао было показано, что строки матрицы W есть собственные векторы ковариационной матрицы $\Sigma = A^\intercal A$. Значение интегрального индикатора $\boldsymbol{q}$ вычисляется как проекция векторов-строк матрицы A на первую главную компоненту, $\boldsymbol{q} = A\boldsymbol{w}$, где $\boldsymbol{w}$ — вектор-столбец матрицы $W^\intercal$, соответствующий наибольшему собственному значению матрицы Σ.

Поиск устойчивых интегральных индикаторов

Для получения интегральных индикаторов, устойчивых к выбросам, в рамках линейной модели ранее было предложено использовать регуляризацию. А. М. Шурыгин в работе [2] рассмотрел два способа регуляризации ковариационной матрицы Σ. Первый способ — регуляризация посредством ридж-регрессии, $\Sigma_{r\beta} = \Sigma + \beta I$, где β — регуляризующий множитель. Второй способ — диагональная регуляризация $\Sigma_{d\nu} = (1 - \nu)\Sigma + \nu \mathrm{diag}(\Sigma)$, где $\nu \in [0, 1]$ — регуляризующий множитель. Было показано, что второй способ дает лучшую устойчивость к выбросам.

Использование регуляризации приводит к потере информативности. Поставим задачу так, чтобы сохранить значение критерия наибольшей информативности на опорном множестве описаний.

Задано множество описаний объектов, $S_0 = \{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m\}$. Обозначим $\mathcal{S} = \{S_1, \ldots, S_l\}$ — множество всех подмножеств S_0, в котором число элементов $l = 2^m$. Алгоритм, вычисляющий наиболее информативный линейный предиктор, использует множество S_ξ, отыскивает веса $\boldsymbol{w}_\xi = \boldsymbol{w}(S_\xi) \in \mathbb{R}^n$ и возвращает интегральный индикатор $\boldsymbol{q}_\xi = A\boldsymbol{w}_\xi \in \mathbb{R}^m$. Обозначим $\bar{S}_\xi$ дополнение S_ξ до S_0. Исключим из рассмотрения тривиальные пары $(S_\xi, \bar{S}_\xi)$, в которых $\#S_\xi = 1$ и $\bar{S}_\xi = \varnothing$. Будем считать, что значения показателей объектов являются независимыми случайными величинами и принята гипотеза Гауссовского распределения этих величин.

Пусть $p_\xi = P(\boldsymbol{a}_i \in S_\xi)$ обозначает вероятность принадлежности некоторого объекта из S_0 множеству S_ξ, и $\bar{p}_\xi$ — вероятность того, что этот объект принадлежит дополнению до S_0. Найдем в $\mathcal{S}$ такое опорное множество S_ξ, для которого отношение $f_\xi = p_\xi / \bar{p}_\xi$ максимально.

Рассмотрим суммарные дисперсии σ_ξ и $\bar{\sigma}_\xi$ проекций $\boldsymbol{p}_i$ элементов $\boldsymbol{a}_i$ множеств S_ξ и $\bar{S}_\xi$ на первые главные компоненты, определяемые матрицей S_ξ. Обозначим $n_\xi, \bar{n}_\xi, n_0$ — число элементов во множествах $S_\xi, \bar{S}_\xi, S_0$ соответственно. Суммарная дисперсия проекций $\boldsymbol{p}_i$ элементов множеств S_ξ и $\bar{S}_\xi$ всей выборки $\sigma^2(S_0)$ равна сумме дисперсий каждой выборки, взвешенных вероятностями принадлежности вектора $\boldsymbol{a}_i$ с проекцией $\boldsymbol{p}_i$ множествам $S_\xi, \bar{S}_\xi$,

$$\sigma^2(S_0) = p_\xi^2 \sigma^2(S_\xi) + p_\xi^2 \sigma^2(\bar{S}_\xi) = \frac{p_\xi^2 \sigma_\xi^2}{n_\xi} + \frac{\bar{p}_\xi^2 \bar{\sigma}_\xi^2}{\bar{n}_\xi}. \quad (3)$$

Для получения выражения отношения вероятностей f_ξ минимизируем дисперсию $\sigma^2(S_0)$. Так как выражение (3) должно удовлетворять равенству $n_\xi + \bar{n}_\xi = n_0$, при дифференцировании используем метод множителей Лагранжа, обозначив мно-

житель λ. Тогда

$$L = \sigma^2(S_0) + \lambda(n_\xi + \bar{n}_\xi - n_0) = \\ = \frac{p_\xi^2\sigma_\xi^2}{n_\xi} + \frac{\bar{p}_\xi^2\bar{\sigma}_\xi^2}{\bar{n}_\xi} + \lambda(n_\xi + \bar{n}_\xi - n_0).$$

Приравняв частные производные по λ и по n_ξ к нулю, получаем

$$\frac{\partial L}{\partial n_\xi} = -\frac{p_\xi^2\sigma_\xi^2}{n_\xi^2} + \lambda = 0, \quad \frac{\partial L}{\partial \lambda} = n_\xi + \bar{n}_\xi - n_0 = 0,$$

откуда получаем $p_\xi\sigma_\xi = n_\xi\sqrt{\lambda}$. Из двух последних выражений $n_0\sqrt{\lambda} = (p_\xi\sigma_\xi + \bar{p}_\xi\bar{\sigma}_\xi)$ и $p_\xi = n_\xi(p_\xi\sigma_\xi + \bar{p}_\xi\bar{\sigma}_\xi)(n_0\sigma_\xi)^{-1}$. Продифференцировав лагранжиан L по $\bar{n}_\xi$, получим аналогичное отношение для вероятности $\bar{p}_\xi$. Искомое отношение вероятностей равно

$$\frac{p_\xi}{\bar{p}_\xi} = \frac{n_\xi\bar{\sigma}_\xi}{\bar{n}_\xi\sigma_\xi}. \tag{4}$$

Таким образом, вероятность принадлежности описания объекта одному из множеств прямо пропорциональна мощности этого множества и обратно пропорциональна среднеквадратичному отклонению. Искомый интегральный индикатор $\boldsymbol{q}_\xi = A\boldsymbol{w}_\xi$ доставляется таким множеством S_ξ, для которого отношение $f_\xi = \frac{n_\xi\bar{\sigma}_\xi}{\bar{n}_\xi\sigma_\xi}$ максимально.

Результаты

Был выполнен сравнительный анализ регионов России по уровню загрязнения ртутью основных продуктов питания. Каждому региону был поставлен в соответствие интегральный индикатор, указывающий на загрязненность продуктов. Были рассмотрены три показателя загрязненности: мясные продукты, молочные продукты и хлебобулочные изделия. Использовались данные 29 регионов. Данные нормированы следующим образом. В каждом регионе для каждого из трех показателей был проведен ряд стандартизованных измерений. Элемент a_{ij} матрицы описаний — величина загрязнения j-го продукта в i-м регионе. Его значение есть отношение квантиля уровня 0,9 распределения содержания ртути в серии измерений к величине предельно допустимой концентрации ртути в данном продукте.

Предложенный алгоритм отыскивает опорное множество S_ξ с целью вычисления весов показателей $\boldsymbol{w}_\xi$ для получения интегральных индикаторов, устойчивых к выбросам. Алгоритм состоит из трех шагов: назначения ядра опорного множества, отыскания опорного множества и вычисления интегрального индикатора.

1. Отыскивается центр исходного множества. Для этого находится вектор-среднее по всем компонентам векторов $\boldsymbol{a}_i$, вошедших в выборку S_0, и изымается вектор, наиболее удаленный в евклидовой метрике. Это действие производится итеративно, до получения последнего вектора, который и является центром. Для сокращения времени работы алгоритма, две трети описаний объектов, наименее удаленных от центра, были занесены в ядро опорного множества.

2. Исходное множества S_0 разбивается на множества S_ξ и $\bar{S}_\xi$ таких, что S_ξ включает ядро опорного множества в качестве собственного подмножества, а $\bar{S}_\xi$ являются объектами-выбросами. Для каждого разбиения вычисляется целевая функция $f_\xi = \frac{n_\xi\bar{\sigma}_\xi}{\bar{n}_\xi\sigma_\xi}$, где $n_\xi, \bar{n}_\xi$ — мощности множеств $S_\xi, \bar{S}_\xi$; и $\sigma_\xi, \bar{\sigma}_\xi$ — суммарная дисперсия проекций объектов множеств $S_\xi, \bar{S}_\xi$ на собственные векторы ковариационной матрицы, определяемой множествами $S_\xi, \bar{S}_\xi$. Из множества полученных функций f_ξ выбираем функцию, на которой достигается максимум.

3. Объекты выбранного опорного множества S_ξ задают матрицу «объект–показатель» A_ξ. Для нее вычисляется ковариационная матрица $\Sigma = A_\xi^\intercal A_\xi$. Первый собственный вектор матрицы Σ определяет веса $\boldsymbol{w}_\xi$ показателей исходного множества [7]. Интегральный индикатор объектов, вычисленный с помощью предложенного алгоритма, есть $\boldsymbol{q}_\xi = A\boldsymbol{w}_\xi$.

Множество исходных данных — описаний регионов — содержит три выброса по второму показателю (молочные продукты) в трех регионах: республика Карелия, г. Санкт-Петербург, Московская область. Данные Карелии, кроме того, содержат выброс по всем трем показателям. Эти три региона не вошли в опорное множество объектов.

Таблица 1. Веса показателей до и после применения алгоритма.

$\boldsymbol{w}$	Без регуляризации	С регуляризацией	С опорным множеством
w_1	0,0204	0,2264	0,4693
w_2	0,9983	0,7687	0,7706
w_3	0,0548	0,5982	0,4312

В таблице 1 показано распределение весов показателей, полученных для трех алгоритмов построения интегральных индикаторов. Первый алгоритм — применение метода главных компонент к исходным данным без использования регуляризации. Второй алгоритм — метод главных компонент с регуляризацией. Был выбран метод диагональной регуляризации, так как полученные с помощью его результаты доставили большее значение критерию устойчивости, чем результаты, полученные с помощью регуляризации посредством ридж-регрессии. Третий алгоритм — метод главных компонент для опорного множества описаний объектов. При использовании первого алгоритма выбро-

сы по второму показателю приводили к неадекватному увеличению вклада этого показателя в интегральный индикатор. Предложенный метод доставляет более адекватные значения весов показателей, как показано в последнем столбце таблицы.

Для иллюстрации результатов работы алгоритмов был введен критерий устойчивости $\varphi = \arg\min_{\Phi} \|\boldsymbol{w}_A - \boldsymbol{w}_{A^*}\|_2$, где множество Φ определено как

$$\Phi = \left\{\boldsymbol{a}^* \colon \|\boldsymbol{a}^*\|_2 = \max \|\boldsymbol{a}_i\|_2,\ i = 1, \ldots, m\right\}.$$

Вектор $\boldsymbol{w}_A$ был получен с помощью метода главных компонент для исходной матрицы A. Вектор $\boldsymbol{w}_{A^*}$ получен был получен с помощью метода главных компонент для матрицы A с присоединенным вектором-столбцом $\boldsymbol{a}^*$, который рассматривался как выброс. Значение критерия устойчивости было вычислено для трех алгоритмов: без использования регуляризации, с диагональной регуляризацией и с предложенным алгоритмом выбора опорного множества. В первом случае значение критерия устойчивости составило $\varphi = 0{,}4727$, во втором $\varphi = 0{,}0962$ и в третьем $\varphi = 0{,}0$.

Следует отметить, что алгоритм, использующий диагональную регуляризацию, позволяет получить адекватный индикатор, но тем не менее влияние объектов-выбросов на индикатор полностью не исключено. Вектор $\boldsymbol{q}_2$ — индикатор, полученный с помощью диагональной регуляризации, вектор $\boldsymbol{q}_3$ — индикатор, полученный с помощью алгоритма выбора опорного множества описаний объектов. Коэффициент ранговой корреляции был использован для сравнения в связи с тем, что он инвариантен относительно монотонных преобразований интегральных индикаторов и учитывает только порядок их значений, игнорируя при этом величину выбросов.

Алгоритм, не использующий регуляризацию, вычисляет интегральный индикатор, который существенно зависит от наличия в выборке объектов-выбросов. Коэффициент ранговой корреляции между интегральным индикатором, полученным посредством такого алгоритма, и между интегральным индикатором, полученным с помощью опорного множества, равен 0,82. Это означает, что у 37 пар, из всех возможных пар элементов двух индикаторов, порядок следования объектов отличается. В таблице 2 приведены примеры таких пар. В столбцах $\boldsymbol{q}_1$ и $\boldsymbol{q}_3$ приведены значения интегральных индикаторов указанных регионов. В столбцах $r(\boldsymbol{q}_1)$ и $r(\boldsymbol{q}_3)$ приведены ранговые номера регионов.

Таблица 2. Значения интегрального индикатора без регуляризации и интегрального индикатора, построенного на основе опорного множества.

Регион РФ	$\boldsymbol{q}_1$	$r(\boldsymbol{q}_1)$	$\boldsymbol{q}_3$	$r(\boldsymbol{q}_3)$
Архангельская обл.	0,5367	19	0,8356	23
Хабаровский край	0,7986	21	0,6165	19
...	...	...	...	...
Владимирская обл.	0,0324	12	0,3577	14
Краснодарский край	0,0449	16	0,1578	10

Заключение

В работе рассмотрена задача построения устойчивых интегральных индикаторов. При построении индикаторов предлагается выбирать из заданного множества описаний объектов опорное множество, используя предложенный критерий вероятности принадлежности описаний объектов этому множеству. Алгоритм построения интегральных индикаторов с выбором опорного множества является альтернативой алгоритмам, которые используют регуляризацию. В отличие от них, в предложенном алгоритме влияние объектов-выбросов на интегральный индикатор исключено. Предложенный алгоритм был использован для получения интегральных индикаторов регионов России по уровню загрязнения основных продуктов питания.

Литература

[1] *Орлов А. И.* Современный этап развития теории экспертных оценок. Заводская лаборатория, 1996, № 1.

[2] *Шурыгин А. М.* Прикладная стохастика: робастность, оценивание, прогноз. — М.: Финансы и статистика, 2000. — С. 99.

[3] *Nabney I. T.* NETLAB: Algorithms for pattern recognition. Springer, 2004. — Pp. 330.

[4] *Зубаревич Н. В., Тикунов В. С., Крепец В. В., Стрижов В. В., Шакин В. В.* Многовариантные методы интегральной оценки развития человеческого потенциала в регионах Российской Федерации // ГИС для устойчивого развития территорий. — Петропавловск-Камчатский, 2001. — С. 84–105.

[5] *Strijov V., Shakin V.* Index construction: the expert-statistical method. Environmental research, engineering and management. 2003. — № 4(26). — Pp. 51–55.

[6] *Рао С. Р.* Линейные статистические методы и их применения. — М.: Наука, 1968. — С. 530–533.

[7] *Jolliffe I. T.* Principal Component Analysis, 2nd ed., Springer, 2002.

Распознавание по прецедентам при наличии пропусков значений признаков*

Михайлова Е. И., Рязанов В. В., Штаюра В. А.
katya.mikh@gmail.com, rvvccas@mail.ru, vadim@musigy.com
Москва, ВМК МГУ, ВЦ РАН, МФТИ

Рассматривается задача распознавания по прецедентам при наличии пропусков значений признаков. Предложен и апробирован алгоритм восстановления неизвестных значений признаков в описаниях распознаваемых объектов. Алгоритм основан на решении оптимизационной задачи: находятся такие значения неизвестных параметров, при которых метрические отношения распознаваемого объекта с объектами обучающей выборки в полном и частичном признаковом пространствах максимально соответствуют друг другу.

С проблемой обработки пропусков в данных приходится сталкиваться при проведении разнообразных социологических, экономических, статистических, медицинских и других исследований. Традиционными причинами, приводящими к появлению пропусков, являются невозможность получения или обработки, искажение или сокрытие информации. В результате для анализа собранных данных поступают неполные сведения.

Предполагается, что таблица обучения не содержит пропусков (либо задача восстановления значений признаков в объектах обучения была решена предварительно).

В работе рассматривается стандартная задача распознавания образов по прецедентам. Задача состоит в создании метода заполнения пропусков числовых признаков в таблице распознавания для последующей классификации её объектов, его реализации и проверки работоспособности на модельных и реальных данных.

Алгоритм заполнения пропусков

Пусть дана обучающая выборка без пропусков в виде n-мерных векторов признаков $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{in})$, $i = 1, \ldots, m$, записанная в таблице обучения T_{nml}, где l — количество классов в выборке. Пусть для распознавания поступает объект $\boldsymbol{y} = (y_1, \ldots, y_n)$ с пропусками. Для простоты считаем, что первые k компонент вектора неизвестны, а остальные известны, то есть $y_r = \Delta$, $r = 1, \ldots, k$, $y_r \neq \Delta$, $r = k+1, \ldots, n$.

По обучающей выборке для каждого признака оценивается интервал его допустимых значений. Пусть $y_r \in M_r$, где $M_r = [a_r, b_r]$,

$$a_r = \min_{i=1,\ldots,m} x_{ir}, \quad b_r = \max_{i=1,\ldots,m} x_{ir}, \quad r = 1, \ldots, k.$$

Основная гипотеза заключается в том, что расстояния между объектами в частичном признаковом пространстве соответствуют их расстояниям в полном признаковом пространстве. Поэтому будем искать такие неизвестные $\tilde{\boldsymbol{y}} = (y_1, \ldots, y_k)$, при которых расстояния от распознаваемого объекта $\boldsymbol{y}$ в подпространстве известных признаков без пропусков до объектов выборки $\boldsymbol{x}_i$ будут соответствовать расстояниям от $\boldsymbol{y}$ до $\boldsymbol{x}_i$ в пространстве всех признаков. Данный подход формализуем в виде следующей оптимизационной задачи:

$$J(\tilde{\boldsymbol{y}}) \to \min_{\tilde{\boldsymbol{y}} \in M_1 \times \cdots \times M_k}.$$

В качестве J выберем одну из функций J_z, при некотором $z \in [0, 2]$:

$$J_z(\tilde{\boldsymbol{y}}) = \frac{1}{\sum_{i=1}^m \delta_i^z} \sum_{i=1}^m \frac{\big(d_i(\tilde{\boldsymbol{y}}) - \delta_i\big)^2}{\delta_i^{2-z}};$$

$$d_i^2(\tilde{\boldsymbol{y}}) = \big\|\boldsymbol{y} - \boldsymbol{x}_i\big\|^2 = \sum_{r=1}^n (y_r - x_{ir})^2;$$

$$\delta_i^2 = \sum_{r=k+1}^n (y_r - x_{ir})^2.$$

При $z = 2$ функция J_z есть среднеквадратичное отклонение расстояний. По мере уменьшения z минимизация J_z поощряет более точную аппроксимацию меньших расстояний, которые, в силу «гипотезы компактности», более важны для решения задачи распознавания [1].

Оптимальные значения $\tilde{\boldsymbol{y}}$ будем искать методом градиентного спуска [2]. Обозначим через t номер итерации, λ_t — шаг градиентного спуска. Вектор на следующей итерации вычисляется по формуле

$$\tilde{\boldsymbol{y}}^{(t+1)} = \tilde{\boldsymbol{y}}^{(t)} - \lambda_t \nabla J(\tilde{\boldsymbol{y}}^{(t)}), \quad t = 1, 2, \ldots$$

Используется метод наискорейшего спуска, при котором при переходе из точки $\tilde{\boldsymbol{y}}^{(t)}$ в точку $\tilde{\boldsymbol{y}}^{(t+1)}$ функция $J\big(\tilde{\boldsymbol{y}}^{(t)} - \lambda_t \nabla J(\tilde{\boldsymbol{y}}^{(t)})\big)$ минимизируется по λ_t. Градиент функции J_z вычисляется по формуле

$$\nabla J_z(\tilde{\boldsymbol{y}}) = \frac{2}{\sum_{i=1}^m \delta_i^z} \sum_{i=1}^m \left(\frac{d_i - \delta_i}{\delta_i^{2-z} d_i}\right)\big(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{x}}_i\big).$$

Для нахождения на очередной итерации шага градиентного спуска λ_t решается уравнение

$$E(\lambda_t) = \frac{dJ}{d\lambda_t} = -\big\langle \nabla J(\tilde{\boldsymbol{y}}^{(t+1)}), \nabla J(\tilde{\boldsymbol{y}}^{(t)}) \big\rangle = 0.$$

*Работа выполнена при поддержке проектов РФФИ № 08-01-00636, № 08-01-90016 бел, № 08-01-90427 укр, Целевой программы № 2 Президиума РАН, Целевой программы № 2 отделения математических наук РАН.

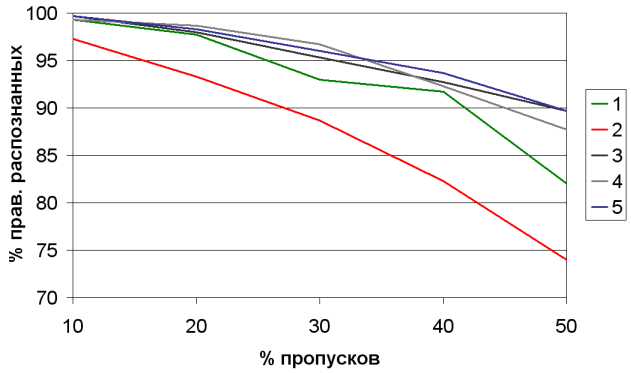


Рис. 1. Результаты распознавания: 1 — заполнение средним, 2 — k ближайших соседей, 3, 4, 5 — с использованием функций J_2, J_0, J_1 соответственно.

Для поиска градиентного шага λ_t используется конечно-разностная модификация метода Ньютона [3], показавшая хорошую сходимость:

$$\lambda_{t+1} = \lambda_t - \frac{E(\lambda_t)}{E(\lambda_t + h) - E(\lambda_t)} h,$$

где h — фиксированный параметр метода.

Для оценки справедливости основной гипотезы предложен следующий критерий, значение которого вычисляется как сумма $J_2(\boldsymbol{x}_j)$ по всем объектам обучающей выборки:

$$\varphi(T_{nml}) = \sum_{j=1}^{m} \frac{\sum_{i=1}^{m} (d_{ij} - \alpha\delta_{ij})^2}{\sum_{i=1}^{m} \delta_{ij}^2};$$

$$d_{ij}^2 = \sum_{r=1}^{n} (x_{ir} - x_{jr})^2 = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2;$$

$$\delta_{ij}^2 = \sum_{r=k+1}^{n} (x_{ir} - x_{jr})^2;$$

где $\alpha = \frac{n}{n-k}$. Заметим, что в случае $k = n$ задача восстановления пропусков не решается, а объект классифицируется согласно априорным вероятностям классов.

Численные эксперименты на модельных данных подтвердили наличие корреляции между значениями данного критерия и результатами распознавания. При высоких значениях критерия основная гипотеза не выполняется, признаки восстанавливаются неточно и ошибки распознавания становятся более частыми, и наоборот.

Результаты апробации алгоритма на модельных данных

Решалась модельная задача с тремя классами, каждый из которых является случайной выборкой, сформированной по нормальному закону распределения с независимыми признаками. В каждом классе по 100 объектов, использовалось 5 признаков. Пропуски восстанавливались с помощью предложенного подхода, метода k ближайших соседей и заполнения средним по классу. Распознавание проводилось с помощью метода линейной машины в системе Recognition [4]. На рис. 1 графически изображены результаты распознавания.

На этой модельной задаче основная гипотеза (соответствие метрических отношений в частичном и полном признаковых пространствах) выполняется, поскольку все признаки имеют равную информативность. Данный факт подтверждается более высокой точностью распознавания предложенным методом относительно альтернативных.

Задача выбора способа лечения

Предложенный подход апробирован при решении задачи выбора способа лечения мочекаменной болезни. Данные для задачи были предоставлены специалистами Урологической клиники ММА им. И. М. Сеченова. Данные клинического обследования этих больных были систематизированы по параметрам (признакам для распознавания). В представленных данных для обучения было выделено 4 типа лечения, а именно: камень отошел, литотрипсия, открытая литотрипсия, контактная и перкутанная литотрипсия, которые соответствуют классам для задачи распознавания. Выборка состояла из $m = 266$ пациентов (по классам, соответственно: 34, 189, 26, 17), описанных $n = 40$ признаками. Особенность задачи состояла в наличии большого числа пропусков. Распознавание проводилось в системе Recognition [4] с процентом деления 0,5, уровень значимости доверительного интервала 95%.

Таблица 1. Доля правильно распознанных объектов.

метод	доля, %
заполнение средним	59,4
k ближайших соседей	60,2
J_0	62,0
J_1	65,0
J_2	65,4

Применение методов заполнения пропусков позволило снизить количество отказов и ошибок при распознавании. В таблице 1 представлены результаты распознавания выборок, заполненных методом k ближайших соседей, методом заполнения средним значением и методом, предложенным в данной работе.

Литература

[1] *Дуда Р., Харт П.* Распознавание образов и анализ сцен — М.: Изд-во Мир, 1976. — 511 с.

[2] *Моисеев Н. Н., Иванилов Ю. П., Столярова Е. М.* Методы оптимизации — М.: Наука, 1978. — 352 с.

[3] *Самарский А. А., Гулин А. В.* Численные методы: Учеб. пособие для вузов — М.: Наука, 1989. — 432 с.

[4] http://www.solutions-center.ru/

Динамический синдромный анализ*

Переверзев-Орлов В. С., Трунов В. Г.
peror@iitp.ru
Москва, Институт проблем передачи информации им. А. А. Харкевича РАН

Рассматриваются результаты разработки и предварительного исследования метода обучения распознаванию и прогнозирования на основе динамического синдромного анализа (dSA). Метод является результатом развития синдромного анализа, ранее исследовавшегося нами при решении плохо формализованных «стационарных» медицинских задач. Обобщение метода оказалось возможным благодаря удивительной пластичности синдромных моделей, выявленной при их исследовании в качестве инструмента представления понятий. Теперь эта пластичность оказалась ключевым свойством синдромных сетей при их распространении на исследование сложных процессов. В качестве модельных для представления такого рода процессов рассматривались задачи узнавания типов схваточной активности беременных и прогнозирования поведения цены на бирже. Получены интересные результаты.

Введение

Синдромный анализ (SA) исходно разрабатывался нами в качестве метода формализации, уточнения и распространения медицинских знаний, являющихся, по нашему мнению, типичным примером трудно формализуемых знаний в далеких от требований «естественных наук»областях человеческой деятельности.

В этой фазе работы было выяснено, что структурной единицей такого рода знаний можно считать давно уже известную схему формального нейрона, использовавшуюся, в частности, и в персептроне. В соответствии с этой схемой входные воздействия с помощью системы порогов превращаются в бинарные «симптомы», которые затем суммируются с равными или неравными весами, формируя на выходе элементарный синдром.

Несколько позже стало ясно, что взвешивание симптомов не имеет большого смысла, и те эффекты, которые должны были бы достигаться таким взвешиванием, существенно лучше получать разумной группировкой симптомов в более сложной многоуровневой структуре синдрома, который становился в этом случае некой направленной пороговой сетью, итоговая сложность которой определялась смыслом и качеством её работы на прецедентной базе.

Целесообразность и адекватность такой схемы быстро подтвердились работой врачей, принявших и ставших применять её для формализации и уточнения своих профессиональных знаний. Это осуществлялось как в режиме автоформализации, когда врач просто расписывал такого рода структуры, делая их рабочим элементом своей деятельности, так и с помощью первых наших программ синдромного анализа — Sand и Sigm, позволявших порождать такого рода формализмы, соотнося их с реальными данными и имеющимися знаниями для развития, более простого применения и распространения.

Параллельно исследовались свойства и возможности такого рода структур знаний для представления понятий, и в качестве функциональной основы для новых методов обучения распознаванию, где, как скоро выяснилось, синдромные сети оказываются своеобразным и весьма перспективным инструментом для решения задач распознавания существенно более широкого класса, чем только медицинские. Однако всё это относилось к «стационарному» синдромному анализу.

Ситуация кардинально изменилась, когда стало ясно, что SA может эффективно использоваться и для анализа описаний сложных процессов с целью их классификации и прогнозирования динамики. К нашему удивлению, таких задач оказалось намного больше, чем «статических», и мы ранее не обращали на них внимания, возможно, именно по этой причине.

Осознав это, несложно было сообразить, что если живое существует и развивается в мире процессов, то синдромно-сетевые представления человека, о котором мы хоть что-то в этом смысле знаем, могут иметь к этому самое непосредственное отношение, и, следовательно, было бы интересно на них посмотреть с такой точки зрения.Это мы и попробовали сделать, для чего потребовалось распространить синдромный анализ на анализ процессов и выбрать что-то в качестве типичного примера, исследуя который, можно было бы получить ответы на важные вопросы.

Рассмотрение множества различных задач в качестве кандидатов для такого исследования позволило в конце концов остановиться на двух задачах, в наибольшей степени удовлетворяющих сложной комбинации требований к таким кандидатам. Ими оказался прогноз патологии в родах по характеру схваточной активности в дородовом периоде (работа над этой задачей уже идёт по гранту РФФИ) и прогноз движения цен на бирже.

*Работа выполнена при финансовой поддержке РФФИ, проект № 07-07-00407-а.

Сложность и интересность первой задачи для нас заключается в том, что в качестве источника первичного описания схваточной активности рассматривается вибрационный датчик, снимающий с поверхности живота беременной очень сложный и находящийся в широком динамическом и частотном диапазонах сигнал, порождаемый множеством источников в организме беременной. Что касается второй задачи, то биржа является наиболее простым для исследований примером систем, оказывающих активное противодействие взаимодействующим с нею системам.

Общий подход

Как отмечалось выше, основу для разработки динамического синдромного анализа (dSA) составили наши исследования синдромного анализа, в том числе и те программные средства, которые было необходимо для этого создать. Последним из них была программная система eSA для разведочного синдромного анализа статических матриц данных вида «объект-признак». Эта программа включала:

— развитые средства для предварительной обработки данных (гигиена данных, шкалирование, средства визуализации и т. п.);
— достаточно широкий набор средств для собственно разведочного анализа — одномерные и двумерные гистограммы, корреляционный и регрессионный анализ и т. п.;
— наконец, собственно синдромный анализ с многочисленными и хорошо развитыми возможностями по части манипулирования обрабатываемыми данными, критериями, классификаторами и режимами синтеза синдромных моделей.

Переход к динамическому анализу требовал расширения этого набора с учётом специфики анализа и предсказания динамики процессов. Собственно, такого рода расширение основывается на нескольких простых идеях:

— использование потенциала методов обучения распознаванию для синтеза фильтров, согласованных с классами сигналов;
— превращение функции времени, соответствующей сигналу, в многомерную векторную функцию за счет присоединения к ней множества функционалов, зависящих от времени;
— использование такого рода фильтров в качестве прогностических моделей для предсказания будущих состояний исследуемого процесса;
— применение осмысленных классификаторов состояний и критериев отнесения текущих ситуаций к рассматриваемым классам.

Стандартная фильтрация сигналов по определению — это свертка фильтруемой функции с ядром из некого базового набора. Выбор базового набора определяется пониманием того, какого рода результат фильтрации требуется. Можно считать, что синтез фильтра на основе методов обучения распознаванию — это один из способов создания нелинейных калмановских фильтров, адаптируемых под произвольно задаваемые классы сигналов. К сожалению, тут есть серьёзные проблемы с математикой,но мы надеемся, что со временем и это тоже как-то прояснится. Тем не менее, можно полагать, что более мощные процедуры обучениязпозволяют порождать и более мощные фильтры. Пока нам было достаточноэтого довольно общего утверждения.

Что касается присоединяемых функционалов,порождающих многомерную векторную функцию,анализ и прогнозирование будущих значений которой и требуется определять,то тут, к сожалению, можно апеллировать лишь к специфике решаемых задач и связанным с ними профессиональным знаниям. В частности, выбрав в качестве типичного образца биржевую задачу прогнозирования цен, мы не можем не выбрать в качестве дополнительных координат общепринятые в биржевой практике индикаторы, фактически являющиеся некими усредненными оценками ценового процесса в разной глубине его истории. На первый взгляд кажется очевидным, что в настоящее время они уже никак не выполняют той функции реальных индикаторов интересующих будущих событий, которыми они, несомненно, были в момент их изобретения. Это принципиально обусловлено тем, что биржа как самоорганизующаяся система однозначно реагирует на любое касающееся её изобретение, которое могло бы позволить начать выигрывать. Тем не менее, вполне можно предположить, что хотя индивидуально такие индикаторы и утратили своё первоначальное значение, в каких-то комбинациях исходно присущая им информативность всё же сохраняется. И тогда дело лишь в том, чтобы достаточно мощный алгоритм обучения распознаванию сумел такие комбинации обнаружить и использовать для построения прогностических фильтров. Нам представляется, что синдромный анализ в его динамической реализации такого рода возможностями вполне обладает. Более того, практическая проверка этих соображений на выбранных задачах это тоже полностью подтверждает.

Что касается настройки синдромных моделей, то мы исследовали разные критерии качества настройки, но остановились на том из них, который минимизирует суммарные потери при распознавании выбранного класса, включающие как «пропуски цели», так и «ложные тревоги», исходя при этом из того, что и в медицине, и в биржевых прогнозах интересен именно персональный результат, а не усреднение по множеству.Но надо заметить, что

отказ от этого формального требованияв целом мало влиял на окончательные результаты.

Заключение

Нам удалось преобразовать «стационарный» синдромный анализ в «динамический», используя для этого подходы цифровой фильтрации, в том числе и адаптивной.В результате возник весьма интересный и перспективный метод и инструмент для его реализации, которые в попытке применения к анализу и прогнозу будущих состояний двух сложных процессов из качественно различных областей (медицина и биржа) показали весьма обнадеживающие результаты, однозначно свидетельствующие о перспективности динамического синдромного анализа и возможности получения на его основе значимых практических результатов.

Литература

[1] *Переверзев-Орлов В. С., Трунов В. Г.* Синдромный анализ: новые вызовы // Информационные процессы (Information Processes), Электронный научный журнал. — 2008. — Т. 88, № 4. — С. 235–239.

[2] *Переверзев-Орлов В. С.* Моделирование срока родов по данным сигнала наружного датчика вибраций. Состояние проекта // ММРО-13, М.: МАКС Пресс, 2007. — С. 515–519.

[3] *Ващенко Е. А., Витушко М. А., Переверзев-Орлов В. С., Стенина И. И., Трунов В. Г.* Синдромный анализ и системы активного противодействия. Постановка задачи // Вторая Международная конференция «Системный анализ и информационные технологии» САИТ-2007, Обнинск, 2007. — Т. 2. — С. 52–56.

[4] *Vitushko M., Gurov N., Pereverzev-Orlov V.* A Syndrome As a Tool for Presenting Concepts // Pattern Recognition and Image Analysis. — 2002. — Vol. 12, № 2. — Pp. 194–202.

[5] *Витушко М. А., Гуров Н. Д., Переверзев-Орлов В. С.* Синдромное прогнозирование изменчивости // ММРО-10, М.: МАКС Пресс, 2001. — С. 28–30.

[6] *Pereverzev-Orlov V. S., Stenina I. I., Trunov V. G.* Syndrome Analysis of Data // Pattern Recognition and Image Analysis. — 1993. — Vol. 3, № 4. — Pp. 500–507.

Восстановление зависимостей по прецедентам на основе применения методов распознавания и динамического программирования*

Рязанов В. В., Тишин К. В., Щичко А. С.
rvv@ccas.ru, kirill.tishin@gmail.com, toxec@mail.ru
Москва, Вычислительный центр РАН, факультет ВМиК МГУ им. М. В. Ломоносова

Рассматривается задача восстановления зависимости между вектором независимых переменных (признаков) и зависимой скалярной величиной по данным обучающей выборки. Априорные ограничения на вид функции не накладываются. Предлагается подход к восстановлению функциональной зависимости, основанный на объединении идей распознавания и динамического программирования. Формулируется задача поиска оптимального разбиения области изменения зависимой величины на конечное число интервалов, которая сводится к задаче динамического программирования. Рассмотрены два типа зависимостей — кусочно-линейная и общий непараметрический случай. Во втором случае для произвольного фиксированного метода распознавания находятся оптимальные число и границы для простого сведения исходной задачи к задаче распознавания.

Во многих областях научной и производственной деятельности возникает задача восстановления регрессионной зависимости между вектором переменных $\boldsymbol{x} = (x_1, \ldots, x_n)$, $x_j \in M_j$, $j = 1, \ldots, n$, где M_j — множества произвольной природы, и скалярной величиной y по выборке $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$. Предполагая существование между ними функциональной связи $y = f(\boldsymbol{x})$, по выборке подбирается алгоритм (функция из некоторого параметрического класса функций), позволяющий вычислять для вектора переменных (признаков) $\boldsymbol{x}$ соответствующее значение зависимой величины y.

В настоящее время существуют различные параметрические и непараметрические подходы к восстановлению регрессии [1, 2]. Параметрические подходы требуют задания параметрической модели зависимости. Непараметрические подходы используют, как правило, методы частотной оценки y и функции расстояния. Для того, чтобы учитывать такие важные особенности реальных данных, как разнотипность признаков, различная их информативность, согласование метрик отдельных признаков, перечисленные подходы требуют дополнительных затрат. В то же время, для случая дискретной величины $y \in \{1, \ldots, l\}$ (стандартной задачи распознавания [3, 4]) такого рода ограничения не являются критическими. Перечисленные выше трудности успешно преодолеваются, например, в логических моделях распознавания [3–8], не требующих решения дополнительных задач предобработки частично противоречивых, разнотипных, непредставительных данных.

В настоящей работе рассматривается приближенный подход к решению задачи восстановления регрессии, основанный на прямом ее сведении к решению задачи распознавания: область изменения y разбивается на конечное число интервалов, и задача прогноза значения $y = f(\boldsymbol{x})$, соответствующего некоторому $\boldsymbol{x}$, сводится к решению задачи распознавания принадлежности y к одному из l интервалов. При всей примитивности данной схемы и приближенности полученного решения, основные трудности, связанные со сравнением объектов в признаковом пространстве, переносятся на уровень решения задач распознавания, а практический пользователь имеет надежный (хотя и приближенный) инструмент решения исходной задачи. Реализация данного подхода требует нахождения оптимального числа интервалов и самих границ разбиения, соответствующих используемой модели распознавания.

Рассматривались две модели восстановления зависимостей. Первая из них основана на применении логических моделей распознавания [7, 8]. Вторая модель в качестве основной гипотезы предполагает существование линейной зависимости регрессионной переменной от вектора признакового описания объектов для каждого из интервалов изменения y и состоит в итоге в восстановлении кусочно-линейной зависимости.

Без ограничения общности будем считать, что все y_i, $i = 1, \ldots, m$, упорядочены по возрастанию: $y_i \leqslant y_{i+1}$, $i = 1, \ldots, m-1$. Задача восстановления зависимости в обоих случаях формулируется как задача динамического программирования, в которой дискретные переменные

$$z_i \in \{y_2, \ldots, y_{m-1}\}, \quad i = 1, \ldots, l-1, \qquad (1)$$

определяют интервалы разбиения (каждый интервал содержит хотя бы один объект $(\boldsymbol{x}, y)$), минимизируемая функция имеет вид

$$\Phi(z_1, \ldots, z_{l-1}) = \\ = f_0(y_1, z_1) + \sum_{i=1}^{l-2} f_i(z_i, z_{i+1}) + f_{l-1}(z_{l-1}, y_m), \quad (2)$$

*Работа выполнена при поддержке РФФИ, проекты № 08-01-00636, № 08-01-90016 бел, № 08-01-90427 укр, Целевой программы № 2 Президиума РАН, Целевой программы № 2 Отделения математических наук РАН.

при ограничениях

$$y_1 < z_i < y_m, i = 1, \ldots, l-1, \quad (3)$$

$$z_i < z_{i+1}, i = 1, \ldots, l-2. \quad (4)$$

Функции $f_i(z_i, z_{i+1})$ выражают «качество распознавания» класса, соответствующего i-му интервалу.

Для модели восстановления зависимости, основанной на логических алгоритмах распознавания [7, 8], предложен алгоритм вычисления начальных приближений для числа интервалов l и границ z_i, $i = 1, \ldots, l-1$.

Предложенные модели успешно апробированы на модельных и реальных данных.

Модель восстановления регрессии по выборке прецедентов

Предлагается модель, основанная на сведении исходной задачи к задаче распознавания.

Считаем, что задана непротиворечивая обучающая выборка $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$ в виде таблицы обучения T_{mn}, где каждая строка содержит вектор значений признаков и ему соответствующее значение зависимой величины y,

$$T_{mn} = \begin{pmatrix} x_{11} & \ldots & x_{1n} & y_1 \\ \vdots & \ddots & \vdots & \vdots \\ x_{m1} & \ldots & x_{mn} & y_m \end{pmatrix}.$$

Основная задача состоит в вычислении $y(\boldsymbol{x})$ для произвольного допустимого $\boldsymbol{x}$.

Разобьем интервал $[y_1, y_m]$ изменения $y(\boldsymbol{x})$ на $l \geqslant 2$ интервалов

$$\Delta_1 = [y_1, y_{i_1}), \ldots, \Delta_l = [y_{i_{l-1}}, y_m].$$

Тогда таблице T_{mn} можно сопоставить таблицу обучения в стандартной постановке задачи распознавания с l классами. По таблице T_{mn} решается задача обучения и строится алгоритм распознавания A. Для произвольного допустимого $\boldsymbol{x}$ с помощью алгоритма A вычисляется класс K_i (т. е. считается, что $\boldsymbol{x} \in K_i$, а значит $y(\boldsymbol{x}) \in \Delta_i$). Окончательно полагаем, например, $y(\boldsymbol{x}) = \frac{(z_i - z_{i-1})}{2}$. Возможны и другие эвристические варианты вычисления $y(\boldsymbol{x})$ по информации «$y \in \Delta_i$», основанные на оценивании «степени принадлежности» $\boldsymbol{x}$ к классу K_i.

Ясно однако, что если подобных интервалов (т. е. классов) будет много, то ошибка решения задачи распознавания интервала будет велика, хотя значение $y(\boldsymbol{x})$ при правильном распознавании будет указано относительно точно. Если же число интервалов будет невелико (а сами интервалы, соответственно, будут большими), тогда задача распознавания будет решена с максимально возможной точностью, но оценка значения $y(\boldsymbol{x})$ внутри интервала будет приближенной. Здесь вопрос состоит в нахождении «золотой середины».

Пусть границы (1) определяют разбиение

$$\Delta(z_1, \ldots, z_{l-1}) = \{\Delta_1, \ldots, \Delta_l\},$$

а функция качества разбиения имеет вид (2).

Минимизация критерия (2) при ограничениях (1), (3) является известной задачей динамического программирования, для решения которой существует полиномиальный алгоритм [9], поэтому основная проблема здесь представляется в удачной формализации критерия (2). Речь идет о выборе вида функций f_i. Далее из решения задачи динамического программирования находятся оптимальные границы интервалов.

Возможны различные пути практической реализации данной модели. Например, фиксируется некоторый критерий (2), монотонно убывающий с ростом числа классов (например, оценка точности распознавания в скользящем контроле), и решается следующая задача:

$$l \to \max, \quad (5)$$

$$\Phi(z_1, \ldots, z_{l-1}) \geqslant \Phi_0 - \varepsilon, \quad (6)$$

где $\Phi_0 = \min \Phi(z_1)$, ε — допустимый порог падения точности распознавания.

Отметим, что если использовать модель с логическими закономерностями [7], то вычисление логических закономерностей класса K_i определяется границами z_{i-1}, z_i отрезка Δ_i и не зависит от значений остальных параметров z_i. Это непосредственно согласуется с представлением (2), т. к. в нём каждая функция $f_{i-1}(z_{i-1}, z_i)$ также не зависит от значений остальных параметров z_i.

Восстановление кусочно-линейных зависимостей

Основная идея осуществления сформулированного перехода к задаче распознавания состоит в разбиении обучающей выборки на классы по значению зависимой переменной y. Она хорошо вписывается в принципы кусочно-линейной модели (исходная задача — восстановление регрессии). Такой подход не решает обозначенных выше проблем разнотипности признаков, малоинформативности и т. д., но позволяет довольно легко осуществить предлагаемый переход к задаче распознавания.

Имеем обучающую выборку. Для произвольного отрезка Δ_i по данным

$$\{(y_j, \boldsymbol{x}_j) \colon j = 1, \ldots, m,\ y_j \in \Delta_i\}$$

находится функция

$$g_i(\boldsymbol{x}) = \sum_{t=1}^{n} a_t^i x_t + b^i,$$

наилучшим образом аппроксимирующая данную подвыборку (например, с помощью метода наименьших квадратов). Тогда в представлении (2) функции $f_{i-1}(z_{i-1}, z_i)$ принимают вид

$$f_{i-1}(z_{i-1}, z_i) = \sum_{\boldsymbol{x} \in K_i} (g_i(\boldsymbol{x}) - y)^2.$$

Путём решения задачи динамического программирования находится оптимальная кусочно-линейная зависимость. Заметим, что «склейка» на границах интервалов не предполагается, так как это, во-первых, уменьшает точность аппроксимации зависимости внутри отдельного класса, а во-вторых — менее четко осуществляется разбиение на классы. При этом определяются подвыборки

$$\big\{(y_j, \boldsymbol{x}_j),\ j = 1, \ldots, m\big\}, \quad y_j \in \Delta_i,$$

которые мы рассматриваем как описания соответствующих классов K_i. В признаковом пространстве объекты одного класса могут находиться «вперемешку» с объектами другого, поскольку разбиение производилось по значению регрессионной переменной y. По описаниям классов $K_i,\ i = 1, \ldots, l$, строится некоторый алгоритм распознавания A, который применяется для классификации произвольных новых объектов $\boldsymbol{x}$. Окончательно, задача вычисления $y = f(\boldsymbol{x})$ для произвольного $\boldsymbol{x}$ решается в два этапа: решается задача классификации $A\colon \boldsymbol{x} \to K_i$, далее вычисляется

$$y = f(\boldsymbol{x}) = g_i(\boldsymbol{x}) = \sum_{t=1}^{n} a_t^i x_t + b^i.$$

Метод вычисления начальных приближений

Будем вычислять начальные приближения при решении задачи восстановления регрессии на основе логических закономерностей классов.

Пусть $d_{i1} < \cdots < d_{ik_i}$ — множество всех значений i-го признака в таблице обучения, $i = 1, \ldots, n$. Поставим во взаимно однозначное соответствие произвольному $\boldsymbol{x}_j$ вектор $\boldsymbol{z}_j = (z_{j1}, \ldots, z_{jn})$ следующим образом: $z_{jk} = t$, если $x_{jk} = d_{kt}$. На множестве $\boldsymbol{z}_j,\ j = 1, \ldots, m$, определим полуметрику

$$\rho(\boldsymbol{z}_i, \boldsymbol{z}_j) = \sum_{k=1}^{n} |z_{ik} - z_{jk}|.$$

Для произвольного $\boldsymbol{z}_j$ определим окрестность

$$O(\boldsymbol{z}_j) = \big\{\boldsymbol{z}_i\colon i = 1, \ldots, m,\ \rho(\boldsymbol{z}_i, \boldsymbol{z}_j) \leqslant \delta\big\},$$

где δ — параметр. Точку $\boldsymbol{x}_\tau$ назовем точкой локального экстремума функции $y = f(\boldsymbol{x})$ на обучающей выборке, если

$$f(\boldsymbol{x}_\tau) \geqslant f(\boldsymbol{x}_t), \quad \forall \boldsymbol{z}_t \in O(\boldsymbol{z}_\tau), \quad t = 1, \ldots, m$$

или

$$f(\boldsymbol{x}_\tau) \leqslant f(\boldsymbol{x}_t), \quad \forall \boldsymbol{z}_t \in O(\boldsymbol{z}_\tau), \quad t = 1, \ldots, m.$$

Каждому положительному δ соответствует конечное число $l{+}1$ локальных экстремумов, включая максимальное и минимальное значение множества $\{f(\boldsymbol{x}_i)\}_{i=1}^m$),

$$y_{i_t} = f\left(\boldsymbol{x}_{i_t}\right),\ t = 1, \ldots, l+1.$$

Пусть для простоты $y_{i_1} < \cdots < y_{i_t}$. При помощи выбора числа δ зафиксируем соответствующее ему число l. Таким образом, будут зафиксированы границы интервалов разбиения

$$\Delta(z_1, \ldots, z_{l-1}) = \{\Delta_1, \ldots, \Delta_l\}.$$

Определим разбиение на классы обучающей выборки следующим образом:

$$K_j = \big\{\boldsymbol{x}_t\colon y_{i_j} < f(\boldsymbol{x}_t) \leqslant y_{i_{j+1}}\big\}, \quad j = 1, \ldots, l;$$
$$\big\{\boldsymbol{x}_t\colon f(\boldsymbol{x}_t) = y_{i_1}\big\} \subset K_1, \quad t = 1, \ldots, m.$$

Таким образом, возникает однопараметрическое множество (с параметром δ) задач распознавания, среди которых находится начальное приближение для задачи (5)–(6). «Содержательное» обоснование данного метода состоит в том, что находится такое разбиение обучающей выборки, при котором на обучающей выборке близким значениям величины y соответствуют значения $\boldsymbol{x}$, «благоприятные» с позиции метода голосования по логическим закономерностям [8]: классы построены таким образом, чтобы число логических закономерностей для каждого класса было как можно меньше. Для подтверждения «благоприятности» разбиения были проведены эксперименты на модельных и реальных данных. Сравнивались результаты работы алгоритма классификации методом голосования по логическим закономерностям для трех различных способов разбиения обучающей выборки: разбиение на равные интервалы, разбиение на классы одинаковой мощности и предложенное разбиение с помощью точек локального экстремума. Оценка результатов работы алгоритмов осуществлялась по скользящему контролю (10-fold CV).

В качестве модельной задачи была рассмотрена функция

$$y(x) = \sin x + \frac{x}{3},\ x \in [1, 25],$$

обучающая выборка состоит из 100 объектов, расположенных равномерно от 1 до 25. Параметры: $\delta = 16,\ l = 6$. Результаты: разбиение «отрезки одинаковой длины» — 85%, разбиение «классы одинаковой мощности» — 79%, «экстремальное» разбиение — 86% объектов правильно классифицированы.

Также была рассмотрена реальная задача восстановления цены продукта по числовым признакам. Число объектов в обучающей выборке 366. Число признаков 13. Параметры: $\delta = 300$, $l = 5$. Результаты: разбиение «отрезки одинаковой длины» — 68%, разбиение «классы одинаковой мощности» — 59.3%, «экстремальное» разбиение — 70.2% объектов правильно классифицированы.

Литература

[1] *Дрейпер Н., Смит Г.* Прикладной регрессионный анализ // М.: Издательский дом Вильямс. 2007.

[2] *Хардле В.* Прикладная непараметрическая регресия // М.: Мир. 1993.

[3] *Журавлев Ю. И.* Коректные алгебры над множествами некорректных (эвритических) алгоритмов // I. Кибернетика. 1977. № 4. — С. 5–17., // II. Кибернетика. 1977. № 6. — С. 21–27, // III. Кибернетика. 1978. № 2. — С. 35–43.

[4] *Журавлев Ю. И.* Об алгебраическом подходе к решению задач распознавания или классификации. Проблемы кибернетики // М.: Наука, 1978. Вып. 33. — С. 5–68.

[5] *Дмитриев А. Н., Журавлев Ю. И.* О математических принципах классификации предметов и явлений // Сб. «Дискретный анализ». Вып. 7. Новосибирск. ИМ СО АН СССР. 1966. — С. 3–11.

[6] *Баскакова Л. В., Журавлев Ю. И.* Модель распознающих алгоритмов с представительнымим наборами и системами опорных множеств // Журн. вычисл. матем. и физики. 1981. Т. 21, № 5. — С. 1264–1275.

[7] *Рязанов В. В.* Логические закономерности в задачах распознавания (параметрический подход) // Журн. вычисл. матем. и физики. 2007. Т. 47, № 10. — С. 1793–1808.

[8] *Журавлев Ю. И., Рязанов В. В., Сенько О. В.* РАСПОЗНАВАНИЕ. Математические методы. Программная система. Применения. // М.: Фазис. 2006. — С. 176.

[9] *Сигал И. Х., Иванова А. П.* Введение в прикладное дискретное программирование // М.: Физматлит. — 2007.

Решение задачи восстановления зависимости коллективами распознающих алгоритмов*

Рязанов В. В., Ткачев Ю. И.
rvv@ccas.ru, tkachevy@gmail.com
Москва, МГУ им. М. В. Ломоносова

Рассматривается задача установления зависимости между вектором независимых переменных и зависимой скалярной величиной по данным обучающей выборки. Априорные ограничения на вид функции не накладываются. Предлагается подход к восстановлению функциональной зависимости, основанный на решении конечного набора специальных, построенных по обучающей выборке, задач распознавания и последующем вычислении прогнозного значения зависимой величины как коллективного решения. При этом используется статистическая модель объединения результатов распознавания с использованием формулы Байеса. Предложен общий алгоритм построения регрессии при различных подходах к выбору исходного коллектива распознающих алгоритмов и оценке их вероятностных характеристик. Приводятся результаты сравнения настоящего подхода с известными моделями восстановления зависимости.

Введение

Рассматривается стандартная задача восстановления зависимости (регрессии) между вектором независимых переменных $\boldsymbol{x} = (x^{(1)}, \ldots, x^{(k)})$, $x^{(i)} \in M_i$, M_i — множество произвольной природы, и скалярной величиной y по выборке прецедентов $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^m$, предполагая существование между ними функциональной связи $y = f(\boldsymbol{x})$. Вектор $\boldsymbol{x}$ является признаковым описанием некоторого объекта, ситуации, явления или процесса, а $y \in \mathbb{R}$ — значение некоторой скалярной характеристики $\boldsymbol{x}$. Данная задача в статистической постановке известна как задача восстановления регрессии — функции условного математического ожидания, при этом предполагается существование условной плотности $p(y \,|\, \boldsymbol{x})$.

В настоящее время существуют различные параметрические и непараметрические подходы к восстановлению регрессии [2, 7]. Следует отметить существенные ограничения регрессионных подходов. Параметрические подходы требуют априорного знания аналитического вида функций. Наличие разнотипных признаков требует привлечения дополнительных средств описания объектов в единой шкале. Непараметрические подходы используют, как правило, методы частотной оценки в некоторой окрестности, при этом возникают проблемы выбора окрестности и функций расстояния, учета фактора различной важности признаков и т. п. В регрессионном анализе важно правильно выделить причинно-следственные связи между различными факторами и заложить эти соотношения в модель. Построение функций множественной нелинейной регрессии с помощью аналитических методов математической статистики в большинстве случаев невозможно. В то же время, случай дискретной величины $y = \{1, \ldots, l\}$ (стандартная задача распознавания) в настоящее время достаточно хорошо изучен. Более 20 лет тому назад Ю. И. Журавлевым было отмечено, что задача распознавания может рассматриваться как задача экстраполяции специальных функций, когда независимые переменные (значения признаков) могут быть фактически произвольны, а зависимая величина принимает конечное число значений. В настоящее время существуют различные модели и конкретные алгоритмы для решения задач распознавания [1, 3, 5, 6, 8]. С середины 70-х годов 20-го века были разработаны подходы для решения задач распознавания коллективами эвристических распознающих алгоритмов (алгебраический подход [4], логические корректоры [5]). В [9] описан Байесовский подход к синтезу коллективных решений, получивший определенное развитие за рубежом. В настоящей статье предлагается новый подход к решению задачи восстановления зависимостей, основанный на решении задач распознавания и байесовской коррекции. При этом основные трудности, связанные со сравнением объектов в признаковом пространстве (разнотипность и различная информативность признаков, согласование метрик для отдельных признаков, и др.) переносятся на уровень решения задач распознавания. В отличие от стандартного решения задачи распознавания коллективом алгоритмов, здесь каждый алгоритм решает различные задачи распознавания при равном числе классов. Классы в каждой задаче распознавания соответствуют различным разбиениям области значений переменной на интервалы. Это позволяет оценивать в итоге вероятности принадлежности y к достаточно малым интервалам, и вычислять прогнозные значения y.

Описывается общий подход восстановления зависимостей с использованием байесовской коррекции, различные способы оценки требуемых условных вероятностей. Предложена модель восстановления зависимостей с использованием коллектива логических алгоритмов, доказана корректность мо-

*Работа выполнена при финансовой поддержке РФФИ № 08-01-00636, № 08-01-90016 бел, № 08-01-90427 укр, Целевой программы № 2 Президиума РАН, Целевой программы № 2 Отделения математических наук РАН.

дели, приведены результаты сравнительных экспериментов.

Общая модель восстановления зависимости

Пусть задана непротиворечивая обучающая выборка $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^m$, т. е. для любых $i, j = 1, \ldots, m$ если $\boldsymbol{x}_i \neq \boldsymbol{x}_j$, то $y_i \neq y_j$. Возьмем число $n \leqslant m$ и $n+1$ точку из отрезка $R = [\min\limits_{i=1,\ldots,m} y_i,\ \max\limits_{i=1,\ldots,m} y_i]$: $d_0 = \min\limits_{i=1,\ldots,m} y_i,\ d_0 < d_1 < \ldots < d_n,\ d_n = \max\limits_{i=1,\ldots,m} y_i$. Получаем разбиение множества R на n интервалов $\Delta_1 = [d_0, d1),\ \Delta_2 = [d_1, d_2),\ \ldots,\ \Delta_n = [d_{n-1}, d_n]$, $\boldsymbol{\Delta} = \{\Delta_1, \ldots, \Delta_n\}$. Обозначим $c_k = \frac{1}{2}(d_{k-1} + d_k)$ — центр интервала $\Delta_k,\ k = 1, \ldots, n$.

Возьмем число $l,\ 2 \leqslant l \leqslant n$, и определим N разбиений отрезка R на l интервалов, поставив каждому разбиению в соответствие вектор с целочисленными положительными компонентами $\boldsymbol{k}_i = (k_i^{(1)}, \ldots, k_i^{(l-1)}),\ i = 1, \ldots, N,\ k_i^{(j)} < k_i^{(j+1)} < n$. Этот вектор задает «метки» интервалам Δ_k: интервалы $\Delta_1, \ldots, \Delta_{k_i^{(1)}}$ «помечены» меткой «1», $\Delta_{k_i^{(1)}+1}, \ldots, \Delta_{k_i^{(2)}}$ — меткой «2», $\ldots$, $\Delta_{k_i^{(l-2)}+1}, \ldots, \Delta_{k_i^{(l-1)}}$ — меткой «l−1», $\Delta_{k_i^{(l-1)}+1}, \ldots, \Delta_n$ — меткой «l». Каждое разбиение отрезка R определяет разбиение множества $\boldsymbol{M} = M_1 \times \ldots \times M_k$ на l подмножеств $K_1^i, \ldots, K_l^i$ (классов): $\boldsymbol{M} = \bigcup\limits_{t=1}^{l} K_t^i,\ \nu \neq \mu \Rightarrow K_\nu^i \cap K_\mu^i = \varnothing$ согласно правилу: объект $\boldsymbol{x}$ принадлежит классу K_j^i разбиения $\boldsymbol{K}^i = \{K_1^i, \ldots, K_l^i\}$ тогда и только тогда, когда $y = f(\boldsymbol{x}) \in \Delta_k$, и интервал Δ_k помечен меткой «j». Каждое разбиение $\boldsymbol{K}^i$ определяет стандартную задачу распознавания Z_i с l классами. Пусть A_i — некоторый алгоритм решения задачи распознавания Z_i, относящий произвольный объект $\boldsymbol{x}$ к одному из классов $K_{a_i}^i$.

Считаем декартово произведение $\boldsymbol{K}^1 \times \ldots \times \boldsymbol{K}^l \times \boldsymbol{\Delta}$ множеством элементарных событий, событие $(K_{a_1}^1, \ldots, K_{a_N}^N, \Delta_j)$ означает: «объект $\boldsymbol{x}$ отнесен алгоритмом A_1 в класс a_1, $\ldots$, алгоритмом A_N — в класс $a_N,\ y = f(\boldsymbol{x}) \in \Delta_j$». Вероятность такого события будем обозначать $\mathsf{P}(a_1, \ldots, a_N, \Delta_j)$.

По формуле Байеса имеем

$$\mathsf{P}(\Delta_j \,|\, a_1, \ldots, a_N) = \frac{\mathsf{P}(a_1, \ldots, a_N, \Delta_j)}{\mathsf{P}(a_1, \ldots, a_N)} = \\ = \frac{\mathsf{P}(\Delta_j)}{\mathsf{P}(a_1, \ldots, a_N)}\, \mathsf{P}(a_1, \ldots, a_N \,|\, \Delta_j). \quad (1)$$

Пусть классификаторы статистически независимы, тогда формулу (1) можно записать в виде:

$$\mathsf{P}(\Delta_j \,|\, a_1, \ldots, a_N) = \frac{\mathsf{P}(\Delta_j)}{\prod\limits_{i=1}^{N} \mathsf{P}(K_{a_i}^i)} \prod_{i=1}^{N} \mathsf{P}(K_{a_i}^i \,|\, \Delta_j). \quad (2)$$

Обозначим $p_k = \mathsf{P}(\Delta_k \,|\, a_1, \ldots, a_N),\ k = 1, \ldots, n$.

Определение 1. *Байесовским корректором называется функция* $F\colon (p_1, \ldots, p_n) \to \mathbb{R}$, *где* $p_1, \ldots, p_n$ *получены из формулы* (1).

Определение 2. *Наивным байесовским корректором называется функция* $F\colon (p_1, \ldots, p_n) \to \mathbb{R}$, *где* $p_1, \ldots, p_n$ *получены из формулы* (2).

Далее будем рассматривать именно наивный байесовский корректор.

Замечание 1. В случае задачи распознавания y — метка класса при классификации объекта $\boldsymbol{x}$, а все разбиения $\boldsymbol{K}^i$ совпадают). Модель наивного байесовского корректора в задаче распознавания описана, например, в [9].

Определение 3. *«Ответом по среднему» байесовского корректора для объекта* $\boldsymbol{x}$ *назовем величину* $\tilde{y} = \sum\limits_{k=1}^{n} p_k c_k$.

Определение 4. *«Ответом по максимуму» байесовского корректора для объекта* $\boldsymbol{x}$ *назовем величину* $\hat{y} = c_k$, *где* $k = \arg\max\limits_j p_j$.

Алгоритм восстановления зависимости

1) формирование разбиений вещественной оси на интервалы $\Delta_k,\ k = 1, \ldots, n$;
2) задание разбиений $\boldsymbol{K}^i,\ i = 1, \ldots, N$, формирование обучающих выборок задач распознавания $Z_i,\ i = 1, \ldots, N$, выбор классификаторов $A_i,\ i = 1, \ldots, N$, их обучение;
3) вычисление оценок вероятностей $\mathsf{P}(K_j^i \,|\, \Delta_k)$, $\mathsf{P}(\Delta_k)$, $\mathsf{P}(K_j^i),\ i = 1, \ldots, N,\ j = 1, \ldots, l,\ k = 1, \ldots, n$.

Алгоритм вычисления значения зависимой величины (значения регрессии)

1) классификация алгоритмами $A_1, \ldots, A_N$ объекта $\boldsymbol{x}$;
2) вычисление значений условных вероятностей $p_1, \ldots, p_n$ по формулам (2);
3) вычисление «ответа по среднему» $\tilde{y}$ или «ответа по максимуму» $\hat{y}$.

Таким образом, конкретизация модели восстановления зависимости требует конкретизации всех параметров алгоритма восстановления зависимости. Для вычисления оценок вероятностей могут быть использованы подходы математической статистики и статистической теории распознавания. Далее будут рассмотрены два нестандартных подхода к решению данных вопросов, основанных на использовании идей минимизации эмпирического

риска, логических моделей распознавания и эвристического оценивания некоторых вероятностей.

Модель восстановления зависимости, основанная на применении логических алгоритмов распознавания

Далее везде под классификатором будем понимать логические классификаторы: тестовый алгоритм, алгоритм голосования по представительным наборам или алгоритмы голосования по системам логических закономерностей. Особенностью данных алгоритмов является присущее им свойство корректности — безошибочное распознавание объектов непротиворечивой обучающей выборки.

Далее рассматривается задача с числом классов $l = 2$. Будем считать, что в обучающей выборке все y_i различны, а $y_i < y_{i+1}$, $i = 1, \ldots, m-1$. Рассмотрим два способа построения интервалов.

Первый способ построения интервалов Δ_k

Возьмем $n = m$, $N = n - 1$. Положим $d_0 = y_1$, $d_1 = \frac{y_1+y_2}{2}$, $\ldots$, $d_{m-1} = \frac{y_{m-1}+y_m}{2}$, $d_m = y_m$. Для классификатора A_i будем метить интервалы следующим образом: интервалы $\Delta_1, \ldots, \Delta_i$ будут иметь метки «1», остальные — «2».

Второй способ построения интервалов Δ_k

Найдем минимальное число $\varepsilon = y_{i+1} - y_i$, $i = 1, \ldots, m-1$. Возьмем $n = 2m-2$, $N = n-1 = 2m-3$. Положим $d_0 = y_1 - \frac{\varepsilon}{2}$, $d_1 = y_1 + \frac{\varepsilon}{2}$, $d_2 = y_2 - \frac{\varepsilon}{2}$, $d_3 = y_2 + \frac{\varepsilon}{2}$, $\ldots$, $d_{2i} = y_{i+1} - \frac{\varepsilon}{2}$, $d_{2i+1} = y_{i+1} + \frac{\varepsilon}{2}$, $\ldots$, $d_{n-1} = y_m - \frac{\varepsilon}{2}$, $d_n = y_m + \frac{\varepsilon}{2}$. Для классификатора A_i будем метить интервалы следующим образом: интервалы $\Delta_1, \ldots, \Delta_i$ будут иметь метки «1», остальные — «2».

Эмпирическое оценивание вероятностей

Частотной оценкой для $\mathsf{P}(K_j^i)$, $i = 1, \ldots, N$, $j = 1, \ldots, l$, назовем долю объектов, принадлежащих классу K_j^i в задаче распознавания Z_i.

Частотными оценками для $\mathsf{P}(K_j^i \,|\, \Delta_k)$, $i = 1, \ldots, N$, $j = 1, \ldots, l$, $k = 1, \ldots, n$, назовем отношение $\frac{m_{ij}^{(k)}}{m_{ij}}$, где m_{ij} — количество объектов обучения в классе K_j^i в задаче распознавания Z_i, $m_{ij}^{(k)}$ — количество объектов обучения в классе K_j^i в задаче распознавания Z_i, значение целевого признака которых принадлежит интервалу Δ_k.

Частотной оценкой $\mathsf{P}(\Delta_k)$, $k = 1, \ldots, n$, будем называть долю объектов обучающей выборки, значение целевого признака которых принадлежит интервалу Δ_k.

Далее будем использовать данные оценки.

Определение 5. *Будем говорить, что модель обладает свойством* $(*)$, *если для объекта* $\boldsymbol{x}_i$, $i = 1, \ldots, m$, *обучающей выборки* $p_k \neq 0 \Leftrightarrow y_i \in \Delta_k$.

Утверждение 1. *Модель, в которой интервалы построены первым способом и использовано эмпирическое оценивание вероятностей, обладает свойством* $(*)$.

Определение 6. *Модель восстановления зависимости называется корректной, если для обучающей выборки* $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^m$ $\tilde{y}_i = y_i$, $i = 1, \ldots, m$.

Определение 7. *Моделью* $\mathbb{A}_1$ *называется байесовский корректор над логическими классификаторами с частотными оценками вероятностей при использовании второго способа построения интервалов и «ответа по максимуму».*

Теорема 2. *Модель* $\mathbb{A}_1$ *при непротиворечивой обучающей информации является корректной.*

Модель восстановления зависимости, основанная на минимизации среднего риска

Будем считать, что задано некое разбиение $\boldsymbol{\Delta} = \{\Delta_1, \ldots, \Delta_n\}$ отрезка R. Обозначим $e_{i,j}^{(k)} = \mathsf{P}(K_j^i \,|\, \Delta_k)$. Напомним, что $\tilde{y} = \sum\limits_{k=1}^{n} p_k c_k$, c_k — центр интервала Δ_k, $p_k = \frac{\mathsf{P}(\Delta_k)}{\prod\limits_{i=1}^{N} \mathsf{P}(K_{a_i}^i)} \prod\limits_{i=1}^{N} e_{i,a_i}^{(k)}$.

Поставим оптимизационную задачу

$$F = \sum_{i=1}^{m} (y_i - \tilde{y}_i)^2 \to \min_{e_{i,j}^{(k)}, \mathsf{P}(\Delta_k), \mathsf{P}(K_j^i)} \qquad (3)$$

при ограничениях:

1) $\sum\limits_{j=1}^{l} \mathsf{P}(K_j^i) = 1$, $i = 1, \ldots, N$;
2) $\sum\limits_{k=1}^{n} \mathsf{P}(\Delta_k) = 1$, $\mathsf{P}(\Delta_k) > 0$, $k = 1, \ldots, n$;
3) $\sum\limits_{k=1}^{n} \mathsf{P}(\Delta_k) e_{i,j}^{(k)} = \mathsf{P}(K_j^i)$, $e_{i,j}^{(k)} \geqslant 0$, $i = 1, \ldots, N$, $j = 1, \ldots, l$.

Частным случаем задачи (3) является задача с фиксированными $\mathsf{P}(\Delta_k)$, $\mathsf{P}(K_j^i)$, оценками которых, например, являются оценки, описанные в разделе «Эмпирическое оценивание вероятностей». В данной задаче функционал — гладкая, выпуклая функция переменных $e_{i,j}^{(k)}$. Множество, по которому осуществляется оптимизация, являются выпуклым, что позволяет использовать для решения задачи метод проекции градиента.

Практическое сравнение моделей восстановления зависимости

Одномерные задачи

Парабола. $\dim \boldsymbol{x}_i = 1$.
Обучающая выборка:

$$\boldsymbol{x}_i^{(1)} = 2i,\ y_i = \big(\boldsymbol{x}_i^{(1)}\big)^2,\ i = 1, \ldots, 50.$$

Прогнозная выборка:

$$\boldsymbol{x}_i^{(1)} = i,\ y_i = \big(\boldsymbol{x}_i^{(1)}\big)^2,\ i = 1, \ldots, 100.$$

Синусоида. $\dim \boldsymbol{x}_i = 1$.
Обучающая выборка:

$$\boldsymbol{x}_i^{(1)} = 2i,\ y_i = \sin\big(100\frac{\boldsymbol{x}_i^{(1)}}{2\pi}\big),\ i = 1, \ldots, 50.$$

Прогнозная выборка:

$$\boldsymbol{x}_i^{(1)} = i,\ y_i = \sin\big(100\frac{\boldsymbol{x}_i^{(1)}}{2\pi}\big),\ i = 1, \ldots, 100.$$

Таблица 1. Нормы векторов ошибок алгоритмов.

Задача	Байесовский корректор	Линейная регрессия	Ядерное сглаживание
Парабола	0,31	6,52	0,64
Синусоида	810,36	11321,49	1138,75

Многомерные задачи

Зашумленные данные. Была проведена серия из 10 экспериментов. В каждом эксперименте были взяты 10 различных двумерных нормальных распределений (мат. ожидания и дисперсии взяты из равномерного распределения), для каждого распределения была сгенерирована выборка из 12 точек, ему подчиненных. Объединением этих данных было создано множество объектов: $\big\{\big(x^{(1)}, x^{(2)}\big)\big\}_{i=1}^{120}$, $x^{(1)}, x^{(2)} \in \mathbb{R}$. Была рассмотрена целевая зависимость y, совпадающая с точностью до знака с плотностью распределения для каждой выборки. Для первых 5 выборок значением являлась плотность, для остальных 5 — плотность со знаком минус. Данные были зашумлены 47 признаками, подчиненных равномерному распределению. Вклад в целевую зависимость они не вносили. Таким образом были получены описания 120 объектов, состоящих из 49 признаков и целевого признака. Данные были разбиты на непересекающиеся обучающую и прогнозируемую выборки: 90 и 30 объектов соответственно.

Задачи с порядковыми признаками. Была проведена серия из 10 экспериментов. В каждом эксперименте была сгенерирована выборка из 200 объектов: $\dim \boldsymbol{x}_i = 3$, $\boldsymbol{x}_i^{(1)}, \boldsymbol{x}_i^{(2)}$ — реализации случайной величины $\xi = 0, \ldots, 10$ с вероятностями $\frac{1}{11}, \ldots, \frac{1}{11}$ соответственно; $\boldsymbol{x}_i^{(3)} = i$, $i = 1, \ldots, 200$. Была рассмотрена целевая зависимость y:

$$y_i = \begin{cases} \boldsymbol{x}_i^{(3)}, & \text{если } \boldsymbol{x}_i^{(1)} \geqslant \boldsymbol{x}_i^{(2)}; \\ -\boldsymbol{x}_i^{(3)}, & \text{если } \boldsymbol{x}_i^{(1)} < \boldsymbol{x}_i^{(2)}. \end{cases}$$

Данные были разбиты поровну между обучающей и прогнозируемой выборками, т. е. по 100 объектов в каждой.

Таблица 2. Нормы векторов ошибок алгоритмов

Задача	Байесовский корректор	Линейная регрессия	Ядерное сглаживание
Зашумленные данные	3,625	4,4890	4,1500
Порядковые признаки	552,077	850,741	606,203

В таблице приведены усредненные данные по сериям экспериментов.

Литература

[1] *Дмитриев А. Н., Журавлев Ю. И., Кренделев Ф. П.* О математических принципах классификации предметов и явлений. // Сб. «Дискретный анализ». — Вып. 7. — Новосибирск: ИМ СО АН СССР, 1966. — С. 3–11.

[2] *Дрейпер Н., Смит Г.* Прикладной регрессионный анализ. — М.: Издательский дом «Вильямс», 2007.

[3] *Дуда Р., Харт П.* Распознавание образов и анализ сцен. — М.: Мир, 1976. — 511 с.

[4] *Журавлев Ю. И.* Корректные алгебры над множествами не корректных (эвристических) алгоритмов. //
I. Кибернетика. — 1977. — № 4. — С. 5–17.
II. Кибернетика — 1977. — № 6.
III. Кибернетика. — 1978. — № 2. — С. 35–43.

[5] *Журавлев Ю. И., Рязанов В. В., Сенько О. В.* Распознавание. Математические методы. Программная система. Практические применения. — М.: Фазис, 2005.

[6] *Рязанов В. В.* Логические закономерности в задачах распознавания (параметрический подход) // ЖВМиМФ. — 2007. — Т. 47, № 10. — С. 1793–1808.

[7] *Хардле В.* Прикладная непараметрическая регрессия. — М.: Мир, 1993.

[8] *Christopher J. C. Burges. A* Tutorial on Support Vector Machines for Pattern Recognition. // Data Mining and Knowledge Discovery 2 — 1998. — P. 121–167.

[9] *Domingos P., Pazzani M.* On the optimality of the simple Bayesian classifier under zero-one loss. // Machine Learning, 29 — 1997. — P. 103–130.

Оптимальные выпуклые корректирующие процедуры в задачах высокой размерности*

Сенько О. В., Докукин А. А.
senkoov@mail.ru
Москва, Вычислительный центр РАН

В работе исследуются свойства выпуклых корректирующих процедур (ВКП) над множествами предикторов. Показано, что задача минимизации обобщённой ошибки ВКП сводится к задаче квадратичного программирования. Исследованы условия невозможности сокращения наборов предикторов без потери точности соответствующей оптимальной ВКП. Проведены экспериментальные исследования прогностических свойств ВКП на наборах одномерных линейных регрессий, которые показали, что оптимизация ВКП может являться эффективным инструментом селекции наборов регрессоров.

Введение

При решении задач эмпирического прогнозирования или распознавания часто возникает ситуация, когда в распоряжении исследователя оказывается целая совокупность различных прогностических функций (решающих правил), каждая из которых по отдельности не позволяет обеспечить требуемую точность прогноза. Как правило, такие прогностические функции, которые далее будут называться исходными предикторами, получаются с помощью предварительного обучения по выборкам данных. Одним из эффективных способов повышения точности прогноза является использование коллективных решающих правил, вычисляющих прогноз в виде выпуклой комбинации прогнозов, получаемых отдельными предикторами.

Предположим, что у нас имеется набор из l предикторов, прогнозирующих значения некоторой переменной Y. Прогноз, вычисляемый i-ым предиктором для некоторого объекта ω, далее будет обозначаться $z_i(\omega)$. Пусть $(c_1,\ldots,c_l)$ — вектор действительных неотрицательных коэффициентов, удовлетворяющий условию $\sum\limits_{i=1}^{l} c_i = 1$. В работе рассматриваются выпуклые корректирующие процедуры (ВКП), вычисляющие коллективный прогноз $Z(\omega,\boldsymbol{c})$ в виде

$$Z(\omega,\mathbf{c}) = \sum_{i=1}^{l} c_i z_i(\omega).$$

Частным случаем выпуклых корректирующих процедур является использование в качестве прогнозов средних значений. Выпуклые корректирующие процедуры достаточно широко используются в теории распознавания и прогнозирования по эмпирическим данным. В качестве примера можно привести методы распознавания, основанные на голосовании по системам закономерностей [1, 2, 3], методы вычисления коллективных решений по наборам алгоритмов распознавания [1, 2, 5]. Выпуклые корректирующие процедуры над множествами линейных регрессий малой размерности рассмотрены в [4], где показана их высокая прогностическая эффективность по сравнению со стандартной регрессией с коэффициентами, найденными с помощью метода наименьших квадратов (МНК).

*Работа выполнена при поддержке грантов РФФИ, проекты № 08-01-00636-а, № 08-07-00437-а, а также гранта Президента РФ, НШ–5294.2008.1.

Ошибка прогнозирования для ВКП

Нетрудно показать, что

$$\sum_{i=1}^{l} c_i\big(Y(\omega)-z_i(\omega)\big)^2 =$$
$$= \sum_{i=1}^{l} c_i\big(Y(\omega)-Z(\omega,\boldsymbol{c})+Z(\omega,\boldsymbol{c})-z_i(\omega)\big)^2 =$$
$$= \sum_{i=1}^{l} c_i\big(Y(\omega)-Z(\omega,\boldsymbol{c})\big)^2 + \sum_{i=1}^{l} c_i\big(z_i(\omega)-Z(\omega,\boldsymbol{c})\big)^2 =$$
$$= \big(Y(\omega)-Z(\omega,\boldsymbol{c})\big)^2 + \sum_{i=1}^{l} c_i\big(z_i(\omega)-Z(\omega,\boldsymbol{c})\big)^2.$$

Таким образом, ошибка ВКП при прогнозировании Y для объекта ω может быть выражена как разность

$$\big(Y(\omega)-Z(\omega,\boldsymbol{c})\big)^2 =$$
$$= \sum_{i=1}^{l} c_i\big(Y(\omega)-z_i(\omega)\big)^2 - \sum_{i=1}^{l} c_i\big(z_i(\omega)-Z(\omega,\boldsymbol{c})\big)^2.$$

Задача поиска оптимальной ВКП может быть представлена как задача минимизации математического ожидания ошибки на пространстве всевозможных прогнозируемых объектов (генеральной совокупности).

$$\Delta_{\text{ссп}} = \mathsf{E}_\Omega\Big(\sum_{i=1}^{l} c_i\big(Y(\omega)-z_i(\omega)\big)^2 - \sum_{i=1}^{l} c_i\big(z_i(\omega)-Z(\omega,\boldsymbol{c})\big)^2\Big),$$

при ограничениях $\sum\limits_{i=1}^{l} c_i = 1,\ c_i \geqslant 0,\ i = 1,\ldots,l.$

Обозначим через δ_i математическое ожидание ошибки индивидуального прогностического алгоритма, $\delta_i = \mathsf{E}_\Omega\big(Y(\omega) - z_i(\omega)\big)^2$. Из неотрицательности вариационной компоненты

$$\mathsf{E}_\Omega \sum_{i=1}^{l} c_i\big(z_i(\omega) - Z(\omega, \boldsymbol{c})\big)^2$$

следует, что ошибка $\Delta_{\text{сср}}$ никогда не превышает взвешенной с помощью коэффициентов c_i суммы ошибок индивидуальных прогностических алгоритмов $\mathsf{E}_\Omega \sum\limits_{i=1}^{l} c_i\big(Y(\omega) - z_i(\omega)\big)^2$. Для математического ожидания $\mathsf{E}_\Omega\big(z_i(\omega) - z_j(\omega)\big)^2$, характеризующего степень расхождения i-го и j-го прогностических алгоритмов, введём обозначение ϱ_{ij}.

Минимизация ошибки ВКП как задача квадратичного программирования

Функционал обобщённой ошибки может быть представлен в виде

$$\Delta_{\text{сср}} = \sum_{i=1}^{l} c_i\delta_i + \mathsf{E}_\Omega Z^2(\omega, \boldsymbol{c}) - \sum_{i=1}^{l} c_i \mathsf{E}_\Omega z_i^2(\omega) =$$
$$= \sum_{i=1}^{l} c_i\delta_i + \sum_{i'=1}^{l}\sum_{i''=1}^{l} c_{i'}c_{i''}\mathsf{E}_\Omega\big(z_{i'}(\omega)z_{i''}(\omega)\big) - \sum_{i=1}^{l} c_i \mathsf{E}_\Omega z_i^2(\omega).$$

Принимая во внимание равенство

$$z_i z_{i'} = \frac{1}{2}\big(z_i^2 + z_{i'}^2 - (z_i - z_{i'})^2\big),$$

получаем, что

$$\sum_{i'=1}^{l}\sum_{i''=1}^{l} c_{i'}c_{i''}\mathsf{E}_\Omega\big(z_{i'}(\omega)z_{i''}(\omega)\big) - \sum_{i=1}^{l} c_i \mathsf{E}_\Omega z_i^2(\omega) =$$
$$= -\frac{1}{2}\sum_{i'=1}^{l}\sum_{i''=1}^{l} c_{i'}c_{i''}\mathsf{E}_\Omega\big(z_{i'}(\omega) - z_{i''}(\omega)\big)^2 +$$
$$+\frac{1}{2}\sum_{i'=1}^{l} c_{i'}\mathsf{E}_\Omega z_{i'}^2(\omega)\sum_{i''=1}^{l} c_{i''} + \frac{1}{2}\sum_{i'=1}^{l} c_{i'}\mathsf{E}_\Omega z_{i''}^2(\omega)\sum_{i''=1}^{l} c_{i''} -$$
$$- \sum_{i=1}^{l} c_i \mathsf{E}_\Omega z_i^2(\omega) = -\frac{1}{2}\sum_{i'=1}^{l}\sum_{i''=1}^{l} c_{i'}c_{i''}\varrho_{i'i''}.$$

Таким образом, задача минимизации обобщённой ошибки сводится к задаче квадратичного программирования

$$\sum_{i=1}^{l} c_i\delta_i - \frac{1}{2}\sum_{i'=1}^{l}\sum_{i''=1}^{l} c_{i'}c_{i''}\varrho_{i'i''} \to \min \qquad (1)$$

при ограничениях

$$\sum_{i=1}^{l} c_i = 1, \quad c_i \geqslant 0, \quad i = 1, \ldots, l.$$

Разработан метод решения задачи квадратчного программирования (1), основанный на постепенном наращивании числа предикторов в наборах с проверкой условия несократимости. Под условием несократимости набора предикторов понимается условие невозможности удаления из набора какого-либо элемента без уменьшения точности соответствующей оптимальной ВКП [6].

Множественная линейная регрессия, основанная на ВКП

Может быть предложена модель множественной линейной регрессии, основанной на выпуклой линейной комбинации простых одномерных регрессий вида $Y = \alpha_i + \beta_i X_i + \varepsilon_i$. Параметры α_i, β_i, оцениваются по обучающей выборке с помощью стандартного МНК для каждой независимой переменной из исходного множества $\tilde{X}$. В результате мы получаем множество из $l = |\tilde{X}|$ предикторов:

$$\big\{z_i(\omega) = \alpha_i + \beta_i X_i(\omega) \;\big|\; i = 1, \ldots, l\big\}.$$

Далее, с использованием метода скользящего контроля (режим leave-one-out), производится оценка обобщённой ошибки отдельных предикторов, а также параметров расхождения между предикторами ϱ_{ij}. Затем оптимальная ВКП ищется как решение задачи квадратичного программирования (1). Пусть $\boldsymbol{c}^0 = (c_1, \ldots, c_l)$ — вектор оптимальных коэффициентов ВКП. В результате получаем линейную регрессионную функцию:

$$Z(\omega, \boldsymbol{c}^0) = \sum_{i=1}^{l} c_i^0\alpha_i + \sum_{i=1}^{l} c_i^0\beta_i X_i(\omega).$$

Обычно большинство компонент $\boldsymbol{c}^0$ в задачах высокой размерности оказываются равными 0. Поэтому задача оптимизации ВКП естественным образом включает другую важную задачу регрессионного анализа: отбор множеств информативных предикторов.

Прогноз $Z(\omega)$, вычисляемый ВКП, может сильно коррелировать с Y, но при этом ошибка прогнозирования может оказаться высокой из-за того, что ВКП снижают общую дисперсию прогнозов по отношению к дисперсии индивидуальных предикторов. Поэтому необходимыми являются дополнительные линейные трансформации $Z(\omega)$. Параметры линейной регрессионной модели

$$Y = \alpha_{\text{сср}} + \beta_{\text{сср}} Z + \varepsilon_{\text{сср}}$$

оцениваются по обучающей выборке с помощью МНК. В результате получается окончательная мультифакторная линейная модель:

$$Y = \alpha_{\text{ccp}} + \sum_{i=1}^{l} c_i^0 \alpha_i \beta_{\text{ccp}} \sum_{i=1}^{l} c_i^0 \beta_i \beta_{\text{ccp}} X_i + \varepsilon_{\text{ccp}}.$$

Эксперименты

Проводилось сравнение множественной линейной регрессии, основанной на ВКП, с пошаговой линейной регрессией (далее обозначается ПР), максимизирующей оценку точности прогноза в режиме скользящего контроля.

Во всех экспериментах зависимая переменная Y и прогностические переменные X являются стохастическими функциями трёх латентных переменных $U_1,\ U_2,\ U_3$. Переменные U_k представляют собой независимые нормально распределённые переменные с математическим ожиданием 0 и дисперсией 1. Значение y_j на j-м объекте генерируется по формуле $y_j = \sum\limits_{k=1}^{3} u_{jk} + e_j^{\text{y}}$, где u_{jk} является значением латентной переменной U_k, e_j^{y} — ошибка, распределённая $\mathcal{N}(0,1)$. Значение связанной с Y (релевантной) переменной X_i генерируется по бинарному вектору $\boldsymbol{\beta^i} = (\beta_1^i, \beta_2^i, \beta_3^i)$. Для j-го объекта $x_{ij} = \sum\limits_{k=1}^{3} u_{jk}\beta_k^i + e_j^{i\text{x}}$, где u_{jk} является значением латентной переменной U_k, $\sum\limits_{k=1}^{3} \beta_k^i = 2$, $e_j^{i\text{x}}$ — ошибка, распределённая по $\mathcal{N}\left(0, \frac{1}{2}\right)$. Значение «мешающей» (не связанной с Y) переменной X_i на j-м объекте вычисляется по формуле $x_{ij} = v_j^{i\text{x}}$, где $v_j^{i\text{x}}$ — ошибка, распределённая по $\mathcal{N}(0, d_{i\text{x}})$.

Во всех экспериментах с помощью генератора случайных чисел вычислялось 100 пар выборок по одному сценарию. Единственным исключением стали задачи с 50-ю элементами и размерностью 100, поскольку иначе ПР требует слишком большого объема вычислений. Таким образом, только 50 пар выборок было создано для этого случая (результаты отмечены звездочкой в таблицах). Одна выборка из пары использовалась для вычисления переменных и построения оптимальных регрессий, в то время как вторая — для оценки прогнозирующих свойств.

Во всех выборках число релевантных переменных n_{rel} было фиксировано и равнялось пяти: две вычислялись при $\beta = \{1,1,0\}$, две — при $\beta = \{1,0,1\}$, одна — при $\beta = \{0,1,1\}$. Число мешающих переменных n_{irrel} варьировалось и было равным 5, 20, 45, 95. Результаты экспериментов представлены в таблицах 1 и 2. В таблице 1 для каждой пары выборок размера m, для каждого числа переменных n_{full} представлены три значения: средние

Таблица 1. Коэффициенты корреляции между прогнозом и реальными значениями.

	$m=20$		$m=30$		$m=50$	
	ВКП	ПР	ВКП	ПР	ВКП	ПР
10	0.75	0.75	0.77	0.79	0.80	0.82
	0.43		0.30		0.36	
25	0.78	0.64	0.78	0.72	0.79	0.77
	0.76		0.65		0.57	
50	0.73	0.5	0.77	0.57	0.80	0.69
	0.83		0.90		0.84	
100	0.75	0.5	0.76	0.53	0.79*	0.57*
	0.92		0.95		0.98*	

Таблица 2. Количество верно и ошибочно отобранных переменных.

	$m=20$		$m=30$		$m=50$	
	ВКП	ПР	ВКП	ПР	ВКП	ПР
$n_{\text{full}}=10$	235	246	258	275	290	303
	3	60	1	47	0	52
$n_{\text{full}}=25$	236	233	255	272	287	300
	11	272	3	239	0	197
$n_{\text{full}}=50$	227	211	259	265	279	303
	28	603	4	719	0	565
$n_{\text{full}}=100$	218	172	244	230	139*	153*
	37	725	6	1185	0*	946*

значения коэффициентов корреляции между прогнозом и реальным значением переменной Y для ВКП (вверху слева) и ПР (вверху справа); доля таблиц, в которых оценка прогнозирующей способности ВКП превышала оценку ПР (внизу). В таблице 2 представлено число правильно (вверху) и ошибочно (внизу) отобранных переменных для обоих методов.

Из таблиц видно, что эффективность ПР значительно уменьшается, когда число переменных сильно превышает число прецедентов в выборках. Прогнозирующая способность ПР уменьшается с 0,75–0,82 для $n_{\text{full}} = 10$ до 0,50–0,56 для $n_{\text{full}} = 100$. Доля ошибочно отобранных переменных превышает 50%.

В то же время ВКП сохраняет эффективность на всех выборках. Наблюдается только незначительное уменьшение прогнозирующей способности с 0,75–0,80 для $n_{\text{full}} = 10$ до 0,75–0,795 для $n_{\text{full}} = 100$. Доля лишних переменных мала во всех экспериментах.

Выводы

Показано, что коэффициенты оптимального ВКП зависят только от двух величин: обобщен-

ной ошибки отдельных слагаемых и матрицы математических ожиданий квадратов расстояний между прогнозами пары предикторов. Поиск коэффициентов оптимального ВКП сводится к задаче квадратичного программирования, которая решается в терминах избыточности набора переменных. Была рассмотрена линейная регрессия, основанная на оптимизации ВКП, которая естественно включает в себя отбор значимых переменных.

Результаты тестирования показывают значительное превосходство ВКП над пошаговой регрессией в задачах высокой размерности. Метод сохраняет эффективность отбора переменных и прогностическую способность в задачах, в которых число переменных превосходит число прецедентов в разы. Описанные результаты могут быть использованы в различных задачах регрессионного анализа, прогнозирования и распознавания.

Литература

[1] *Zhuravlev Yu. I., Kuznetsova A. V., Ryazanov V. V., Senko O. V., Botvin M. A.* The use of pattern recognition methods in tasks of biomedical diagnostics and forecasting // Pattern Recognition and Image Analysis. — 2008. — Vol. 18, № 2. — Pp. 195–200.

[2] *Журавлёв Ю. И., Рязанов В. В., Сенько О. В.* РАСПОЗНАВАНИЕ. Математческие методы. Программная система. Приложения. — М.: Фазис, 2006. 176 с.

[3] *Kuznetsov V. A., Senko O. V. et all.* Recognition of fuzzy systems by method of statistically weighed syndromes and its using for immunological and hematological norm and chronic pathology // Chemical Physics. — 1996. — Vol. 15, № 1. — Pp. 81–100.

[4] *Senko O. V.* The use of collective method for improvement of regression modeling stability // InterStat. Statistics on the Internet, 2004. `http://statjournals.net/`.

[5] *Kuncheva L. I.* Combining pattern classifiers. Methods and algorithms. — New Jersey: Wiley Interscience, 2004.

[6] *Senko O. V.* Optimal ensembles of predictors in convex correcting procedures // Pattern Recognition and Image Analysis (in print).

Метод распознавания по закономерностям в моделях оптимальных разбиений*

Сенько О. В., Кузнецова А. В.
senkoov@mail.ru, azfor@narod.ru
Москва, Вычислительный центр им. А. А. Дородницына РАН;
Институт Биохимической Физики им. Н. М. Эмануэля РАН

Представлен метод распознавания, основанный на взвешенном голосовании по системам подобластей признакового пространстве с преимущественным содержанием объектов одного из классов («синдромов»). При этом «синдромы» являются элементами оптимальных разбиений областей допустимых значений сочетаний признаков, которые ищутся в рамках моделей различного уровня сложности. В работе рассматриваются методы формирования оптимального множества синдромов, а также влияние на точность распознавания параметра, регулирующего отбор признаков в зависимости от сложности модели.

Методы, основанные на принятии коллективных решений по системам закономерностей, получили достаточно широкое распространение в современной теории распознавания [1, 2]. В настоящей работе рассматривается метод, использующий процедуру взвешенного голосования по системам синдромов — подобластей признакового пространства с преимущественным содержанием объектов одного из классов. Настоящий метод является дальнейшим развитием метода *статистически взвешенных синдромов* (СВС), предложенного в [3, 4]. Метод СВС использует следующую схему распознавания. Предположим, что нам требуется построить алгоритм, распознающий некоторое множество объектов из классов $K_1, \ldots, K_L$. На начальной стадии для каждого из классов K_j строится множество синдромов $\widetilde{Q}_j$ по некоторой обучающей информации $\widetilde{S}_0$. Предположим, что объект s^* распознаётся по его векторному описанию $\boldsymbol{x}^*$, принадлежащему пересечению синдромов $Q_1, \ldots, Q_p$ из системы $\widetilde{Q}_j$. Оценка объекта s^* за класс K_j вычисляется как прогноз значения индикаторной функции I_j класса K_j в точке $\boldsymbol{x}^*$. Прогноз $\theta(I_j, \boldsymbol{x})$ вычисляется с помощью процедуры статистически взвешенного голосования:

$$\theta(I_j, \boldsymbol{x}^*) = \frac{\sum_{i=1}^{p} \nu_i w_i}{\sum_{i=1}^{p} w_i}, \qquad (1)$$

где ν_i^j — доля объектов K_j из $\widetilde{S}_0$ с векторными описаниями $\boldsymbol{x}$, принадлежащими Q_i; вес w_i^j i-го синдрома вычисляется как $w_i^j = \frac{1}{m_i+1}\frac{1}{\hat{D}(I_j|Q_i)}$, где $\hat{D}(I_j|Q_i) = \nu_i^j(1-\nu_i^j)$ — оценка дисперсии индикаторной функции I_j на синдроме Q_i, m_i — число объектов из $\widetilde{S}_0$ с векторными описаниями $\boldsymbol{x}$, принадлежащими Q_i. Объект s^* будет отнесён классу K_{j^*}, для которого значение $\theta(I_{j^*}, \boldsymbol{x}^*)$ является максимальным из $\theta(I_1, \boldsymbol{x}^*), \ldots, \theta(I_L, \boldsymbol{x}^*)$. Синдромы для класса K_j строятся путём оптимальных разбиений интервалов допустимых значений признаков. Разбиение $R = \{r_1, \ldots, r_{T_R}\}$ с максимальным значением

$$F_g(R, \widetilde{S}_0, K_j) = \sum_{i=1}^{T_r} (\nu_i^j - \nu_0^j)^2 k_i$$

ищется для признака X, где T_R — число подынтервалов в разбиении R, k_i — число объектов из $\widetilde{S}_0$ со значением X из i-го подынтервала, ν_0^j — доля объектов класса K_j в $\widetilde{S}_0$. Метод СВС включает отбор выявленных закономерностей (а также соответствующих признаков) с помощью установки порогового значения для функционала качества F_g. Обычно множество синдромов $\widetilde{Q}_j$ в методе СВС включает одномерные и двумерные синдромы, задаваемые оптимальными разбиениями. Одномерный синдром, соответствующий подынтервалу r_i признака X включает объекты с $X \in r_i$. Двумерный синдром, соответствующий оптимальному подынтервалу r_{i_1} для признака X_{i_1} и оптимальному подынтервалу r_{i_2} для признака X_{i_2}, включает объекты с $X_{i_1} \in r_{i_1}$ и $X_{i_2} \in r_{i_2}$.

Очевидно, что выражение (1) для $\theta(I_j, \boldsymbol{x}^*)$ представляет собой выпуклую линейную комбинацию частот встречаемости K_j в синдромах, содержащих точку $\boldsymbol{x}^*$. Это обеспечивает снижение вариационной составляющей обобщённой ошибки коллективного решения (1) по отношению к аналогичной выпуклой линейной комбинации вариационных составляющих ошибок прогнозов с использованием отдельных синдромов. Отметим также, что высокая стабильность одномерных разбиений приводит к высокой стабильности прогнозов на отдельных синдромах. Эти два обстоятельства обеспечивают высокую стабильность прогноза коллективного решения в методе СВС и его эффективность в условиях, когда достижение высокой стабильности является принципиальным: высокая размерность данных, ограниченный объём выборок, наличие пропусков и выпадающих наблюдений. При этом несомненным недостатком является низкая аппроксимационная возможность моделей, основанных только на одномерных разбиениях.

*Работа выполнена при финансовой поддержке РФФИ, проекты №08-07-00437,09-01-00409.

Метод распознавания, основанный на мультимодельных разбиениях

Несомненно, что в реальных задачах часто возникают ситуации, когда реально существующие синдромы не могут быть обнаружены с помощью одномерных моделей. Поэтому для поиска таких синдромов естественно использовать более сложные модели. Однако, как это было показано в [7], использование более сложных моделей может приводить к ошибочному обнаружению ложных или частично ложных закономерностей из-за эффекта переобучения. Включение в голосование синдромов, соответствующих частично ложным закономерностям, может приводить к снижению устойчивости и точности распознавания. Представляется разумным промежуточный подход, предполагающий одновременное привлечение нескольких моделей различного уровня сложности. При этом окончательное использование более сложных моделей является оправданным, когда оно позволяет добиваться существенного улучшения разделения объектов из разных классов. В настоящей работе рассматривается метод *мультимодельных статистически взвешенных синдромов* (МСВС), который основан на той же самой процедуре голосования, что и СВС. Но синдромы ищутся путём поиска оптимальных разбиений одномерных или двумерных областей признакового пространства внутри четырёх априори заданных моделей:

1) простейшая одномерная модель с одной граничной точкой и двумя элементами разбиения (модель I);
2) одномерная модель с двумя граничными точками и тремя элементами разбиения (модель II);
3) двумерная модель с двумя линейными границами, параллельными координатным осям и 4-мя элементами (модель III);
4) двумерная модель с одной линейной границей с произвольной ориентацией относительно координатных осей и двумя элементами (модель IV).

Вместо функционала $F_g(R, \widetilde{S}_0, K_j)$, используемого в методе СВС, в методе МСВС максимизируется функционал

$$F_l(R, \widetilde{S}_0, K_j) = \max_{t \in \{1,\dots,T_r\}} (\nu_t^j - \nu_0^j)^2 k_t.$$

Использование $F_l(R, \widetilde{S}_0, K_j)$ вместо $F_g(R, \widetilde{S}_0, K_j)$ связано с необходимостью обеспечить возможность адекватного сравнения качества разбиений с различным числом элементов.

Метод МСВС включает отбор найденных закономерностей с помощью порога Δ_l для функционала $F_l(R, \widetilde{S}_0, K_j)$. Чтобы уменьшить эффект переобучения, в методе МСВС используется штрафной коэффициент для закономерностей, найденных с помощью более сложных моделей разбиений. Функционал качества для моделей (II -IV) умножается на штрафной коэффициент $\eta \in [0, 1]$. Рассмотрим детально алгоритм селекции синдромов для класса K_j.

Селекция одномерных синдромов. Предположим, что оптимальное разбиение R_o^1 было найдено в рамках модели I и $F_g(R_o^{11}, \widetilde{S}_0, K_j) > \Delta_l$. Тогда два синдрома, формируемые граничной точкой b_1 разбиения R_o^1, будут включены в набор $\widetilde{Q}_j$. Предположим, что оптимальное разбиение R_o^{11} было найдено в рамках модели II и $F_g(R_o^{11}, \widetilde{S}_0, K_j)\eta > > \Delta_l$. Тогда три синдрома, формируемые граничными точками b_1 и b_2 разбиения R_o^{11} будут включены в набор $\widetilde{Q}_j$.

Селекция двумерных синдромов с границами, параллельными координатным осям. Предположим, что двумерное оптимальное разбиение R_o^{12} было найдено для пары признаков (X_1, X_2) в рамках модели III. Пусть b_1^* и b_2^* — граничные точки разбиения R_o^{12} на признаках X_1 и X_2 соответственно. Они могут отличаться от граничных точек b_1 и b_2, полученных в рамках одномерной Модели I. В случае

$$F_g(R_o^{12}, \widetilde{S}_0, K_j)\eta > \Delta_l,$$

четыре двумерных синдрома, формируемых граничными точками b_1^* и b_2^*, будут включены в набор $\widetilde{Q}_j$. В случае, если

$$F_g(R_o^{12}, \widetilde{S}_0, K_j)\eta \leqslant \Delta_l,$$

четыре двумерных синдрома, формируемых граничными точками b_1 и b_2, будут включены в набор $\widetilde{Q}_j$, если справедливы неравенства:

$$F_g(R_o^1, \widetilde{S}_0, K_j) > \Delta_l;$$
$$F_g(R_o^2, \widetilde{S}_0, K_j) > \Delta_l.$$

Селекция двумерных синдромов с линейной границей, произвольно ориентированной относительно координатных осей. Предположим, что двумерное оптимальное разбиение R_o^{12} было найдено для пары признаков (X_1, X_2) в рамках модели IV и

$$F_g(R_o^{11}, \widetilde{S}_0, K_j)\eta > \Delta_l.$$

Тогда два синдрома, формируемых линейной границей, будут включены в набор $\widetilde{Q}_j$.

Пороговый параметр Δ_l и коэффициент η являются открытыми параметрами метода МСВС.

Эксперименты

В таблице 1 представлены результаты сравнения эффективности МСВС с эффективностью метода q ближайших соседей и метода опорных векторов на 5 прикладных задачах. Использовались

программные реализации методов q ближайших соседей, метода опорных векторов и СВС, вошедшие в программную систему РАСПОЗНАВАНИЕ [1]. Для метода q ближайших соседей оптимальное число соседей q оценивалось по скользящему контролю. Использовался вариант метода опорных векторов с использованием Гауссианы в качестве ядра, при этом размер ядра подбирался по скользящему контролю из интервала $[3, 10]$.

Рассматривались следующие задачи:

1. Задача компьютерной оценки тяжести пневмонии. Для обучения и оценки точности распознавания использовалась обучающая выборка, включающая описания 116 случаев заболевания, включающие значения 41 клинического параметра. Для каждого из случаев группой экспертов была произведена оценка тяжести в 4-х балльной шкале. Задача оценки тяжести сводится к задаче распознавания с 4-мя классами.

2. Задача диагностики злокачественности (меланома) поражений участков кожи по набору из 33 показателей, описывающему соответствующие изображения. Задача диагностики сведена к задаче распознавания с 3-мя классами. Общее число объектов в обучающей выборке — 80.

3. Задача прогноза рецидива миомы матки по 69 иммунологическим параметрам сведена к задаче распознавания с двумя классами. Общее число случаев, содержащихся в анализируемой выборке — 60. Особенностью данной задачи является высокая доля пропущенных значений.

4. Задача прогноза результатов хемиотерапии по 25 клиническим параметрам была сведена к задаче распознавания с двумя классами. В базе данных было представлено 372 случая. Задача характеризуется высокой долей пропущенных значений.

5. Задача прогноза исхода вирусного гепатита по 20 параметрам сведена к задаче распознавания с двумя классами. В выборке представлены описания 155 случаев заболевания. Задача взята из открытого репозитория UCI Machine Learning repository http://archive.ics.uci.edu/ml/.

В таблице 1 приведены данные по сравнению эффективности четырёх методов распознавания. Оценка точности распознавания производилась в режиме скользящего контроля. В таблице приведены: число правильно распознанных объектов m_{corr}, доля правильно распознанных объектов f_{corr}, средняя точность по классам $f^{\text{av}}_{\text{corr}}$. Обе оценки точности ($f_{\text{corr}}$ и $f^{\text{av}}_{\text{corr}}$) приведены для того, чтобы адекватно учесть существенные различия между содержанием объектов разных классов в некоторых задачах. Значение точности MSWS приведено для значения параметра $\eta = 0{,}5$. Значение параметра Δ_l выбиралось из множества $\{0{,}5; 5{,}0; 10{,}0\}$. В таблице 1 приведены лучшие из трёх в смысле $f^{\text{av}}_{\text{corr}}$ значения точности МСВС.

Таблица 1. Сравнение эффективности четырёх методов распознавания.

	MSWS	SWS	NN	SVM
Задача 1				
m_{corr}	80	76	49	69
f_{corr}	68.9%	65.5%	42.2%	59.5%
$f^{\text{av}}_{\text{corr}}$	69.4%	65.9%	39.9%	54.5%
Задача 2				
m_{err}	55	51	44	51
f_{err}	68.8%	63.8%	55%	63.8%
$f^{\text{av}}_{\text{err}}$	68.6%	63.7%	55.3%	63.1%
Задача 3				
m_{err}	42	43	40	45
f_{err}	70%	71.7%	66.7%	75%
$f^{\text{av}}_{\text{err}}$	75%	63.5%	51.05	53.3%
Задача 4				
m_{err}	241	233	236	245
f_{err}	64.8%	62.6%	63.4%	65.9%
$f^{\text{av}}_{\text{err}}$	64.9%	65.7%	61.7%	58.3%
Задача 5				
m_{err}	130	124	125	133
f_{err}	83.8%	80.0%	84.7%	85.8%
$f^{\text{av}}_{\text{err}}$	85.2%	81.7%	78.6%	74.9%

Из таблицы 1 видно, что точность метода МСВС заметно превышает точность методов q ближайших соседей и опорных векторов в задачах 1 и 2 как в смысле оценки точности f_{corr}, так и в смысле оценки точности $f^{\text{av}}_{\text{corr}}$. В задачах 3, 4, 5 точность метода МСВС в смысле f_{corr} оказывается несколько ниже точности альтернативных подходов. Однако точность в смысле $f^{\text{av}}_{\text{corr}}$ заметно превосходит точность альтернативных подходов также и в этих задачах. Следует отметить, что точность МСВС превосходит точность СВС в большинстве случаев. Исключениями являются чуть лучший результат для СВС в смысле f_{corr} в задаче 3 и чуть лучший результат в смысле $f^{\text{av}}_{\text{err}}$ в задаче 4.

Исследовалась также зависимость параметров точности от шрафующего параметра η. В таблице 2 представлены значения точности в смысле $f^{\text{av}}_{\text{err}}$ и f_{corr} при значениях $\eta \in \{0{,}1; 0{,}5; 0{,}7; 0{,}9\}$ и при том же значении Δ_l, что и в предыдущей серии экспериментов.

Результаты, приведённые в таблице говорят об отсутствии определённой тенденции, общей для всех задач. В задачах 1 и 5 максимальная точность достигается при значении $\eta = 0.5$. В задачах 2 и 4 точность возрастает с ростом η. В задаче 2 точность возрастает при уменьшении η.

Выводы

Таким образом видно, что представленный метод МСВС позволяет достичь высокой точности распознавания в различных задачах. Приведённые результаты говорят как о необходимости привлечения более сложных моделей разбиений,

Таблица 2. Зависимость точности распознавания от штрафующего коэффициента η.

η	0.1	0.5	0.7	0.9
Задача 1				
m_{err}	79	80	79	78
f_{err}	68.1%	68.1%	68.1%	67.2%
$f^{\text{av}}_{\text{err}}$	69.2%	69.9%	69.4%	68.6%
Задача 2				
m_{err}	52	55	56	56
f_{err}	65%	68.8%	70%	70%
$f^{\text{av}}_{\text{err}}$	64.8%	68.6%	70.2%	70.2%
Задача 3				
m_{err}	45	42	41	38
f_{err}	75%	70%	68.3%	63.3%
$f^{\text{av}}_{\text{err}}$	75%	75%	73.9%	70.8%
Задача 4				
m_{corr}	195	201	207	209
f_{corr}	59.6%	61.4%	63.3%	63.9%
$f^{\text{av}}_{\text{corr}}$	58.0%	59.4%	61.6%	61.8%
Задача 5				
m_{corr}	129	130	129	128
f_{err}	83.2%	83.8%	83.2%	82.5%
$f^{\text{av}}_{\text{err}}$	83.6%	85.2%	84.8%	84.4%

так и о необходимости применения более жёстких критериев при использовании найденных с их помощью закономерностей в коллективных решениях. Вместе с тем результаты, приведённые в таблице 2, говорят о необходимости учёта характеристик конкретных выборок при подборе штрафующего коэффициента.

Литература

[1] Журавлёв Ю. И., Рязанов В. В., Сенько О. В. РАСПОЗНАВАНИЕ. Математические методы. Программная система. Применения. — Москва: Фазис, 2006.

[2] Kuncheva L. I. Combining Pattern Classifiers. Methods and Algorithms. Wiley Interscience, New Jersey, 2004.

[3] Kuznetsov V. A., Senko O. V. et all. Recognition of fuzzy systems by method of statistically weighed syndromes and its using for immunological and hematological norm and chronic pathology // Chemical Physics. — 1996. — Vol. 15, No. 1. — Pp. 81–100.

[4] Jackson A. M., Ivshina A. V., Senko O. V., Kuznetsova A. V., et all. Prognosis intraversical bacillus calmette-guerin therapy for superficial bladder cancer by immunological urinary measurements: statistically weighted syndromes analysis // Journal of Urology. — 1998. — Vol. 159, No. 3. — Pp. 1054–1063.

[5] Ivshina A. V., George J., Senko O. V., Mow B., Putti T. C., Smeds J., Nordgren H., Bergh J., Liu E. T-B., Kuznetsov V. A., Miller L. D. Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer // Cancer Res., 66: 10292-10301, 2006.

[6] Senko O. V. The use of a weighted voting procedure on a system of basic sets in prediction problems // Comp. Maths. Math. Phys — Vol. 35, No. Pp.1249–1257, 1995.

[7] Senko O. V., Kuznetsova A. V. The optimal valid partitioning procedures // Statistics on the Internet. — April, 2006. — http://statjournals.net.

Алгоритм выбора нелинейных регрессионных моделей с анализом гиперпараметров*

Стрижов В. В., Сологуб Р. А.
strijov@ccas.ru, roman.sologub@yahoo.com
Москва, Вычислительный центр РАН, Московский физико-технический институт

Рассматривается задача порождения и выбора нелинейных регрессионных моделей. Модели индуктивно порождаются с помощью экспертно-заданного множества гладких функций. Для выбора моделей используется информация о распределении их параметров. Предлагается метод поиска моделей, комбинирующий подходы байесовского вывода и символьной регрессии.

Введение

Проблема выбора моделей является одной из наиболее актуальных и одновременно сложных при моделировании экономических, финансовых и социальных систем. Эта проблема состоит в отыскании оптимальной модели в некотором заданном классе моделей-претендентов. Этот класс задается в виде списка, либо в виде универсальной модели, из которой путем удаления элементов можно получить модели частного вида, либо с помощью правил порождения. В данной работе используется последний способ.

Критерий оптимальности модели задается исходя из гипотезы порождения данных — предположении о распределении случайной переменной при восстановлении регрессии и предположении о распределении параметров. Подход к модификации структуры моделей путем анализа параметров впервые предложен Ле Кюном и Хассиби в [1, 2]. Он состоит в исключении тех элементов моделей, мера выпуклости функции ошибки которых не превосходит заданный порог. Дальнейшее развитие методы анализа пространства параметров получили в работах Маккая [3, 4, 5, 6]. Им было предложено использовать гиперпараметры — параметры функций распределения данных и параметров для выбора моделей. В дальнейшем в работах [7, 8, 9] Бишоп предложил несколько способов оценки гиперпараметров: аппроксимацию Лапласа, ансамблевое обучение и оценку с помощью марковских цепей Монте-Карло.

Однако в рамках вышеприведенных подходов не рассматривались задачи порождения выбираемых моделей. Методы индуктивного порождения линейных регрессионных моделей описаны в работах Ивахненко [10, 11, 12]. Предложено порождать модели в виде линейных комбинаций мономов полинома Колмогорова–Габора. Методы порождения нелинейных регрессионных моделей развиты в работах Козы и Зелинки [13, 14]. Предложено порождать модели как произвольные суперпозиции заданного набора функций с помощью генетических оптимизированных алгоритмов. В работах Владиславлевой [15, 16] при выборе порождаемых моделей предлагается использовать Парето оптимальный фронт — множества моделей на плоскости, заданный функционал качества моделей и функций их сложности.

Ниже описан алгоритм, который является развитием алгоритма, опубликованного в [17, 18, 19]. Он выполняет следующие основные шаги. Задана модель начального приближения. Параметры этой модели настраиваются, вычисляются гиперпараметры, описывающие информативность элементов модели. Согласно гиперпараметрам, элементы модели модифицируются таким образом, чтобы с наибольшей вероятностью обеспечить попадание модели в Парето-оптимальный фронт.

Задача многомерной нелинейной регрессии

Задана регрессионная выборка — множество пар $D = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$, в котором $\boldsymbol{x} \in \mathbb{R}^P$ — свободная переменная, и $y \in \mathbb{R}^1$ — зависимая переменная.

Задано конечное множество порождающих функций $G = \{g \,|\, g\colon \mathbb{R} \times \ldots \times \mathbb{R} \to \mathbb{R}\}$. Функция $g = g(\boldsymbol{b}, \cdot, \ldots, \cdot)$ — гладкая параметрическая. Первый аргумент функции — вектор параметров, последующие аргументы — функции свободных переменных, принимающие значения в $\mathbb{R}^1$. Множество G индуктивно определяет набор допустимых суперпозиций $F = \{f_i\},\ i = 1, \ldots, M$. На эти суперпозиции накладывается ограничение сложности: каждая суперпозиция f_i состоит не более чем из R функций $g \in G$.

Суперпозиция f_i определяет параметрическую регрессионную модель $f_i = f_i(\boldsymbol{w}, \boldsymbol{x})$. Она зависит от независимых переменных $\boldsymbol{x}$ и вектора параметров $\boldsymbol{w}$. Вектор $\boldsymbol{w} \in \mathbb{R}^{W_i}$ состоит из присоединенных векторов — параметров функций $g_1, \ldots, g_{r_i}$, входящих в эту суперпозицию в лексикографическом порядке, то есть $\boldsymbol{w} = \boldsymbol{b}_1 \vdots \boldsymbol{b}_2 \vdots \ldots \vdots \boldsymbol{b}_{r_i}$, где $\vdots$ — знак присоединения векторов. Требуется отыскать в множестве F модель f_i, максимизирующую заданную целевую функцию $p(\boldsymbol{w} \,|\, D, A, \beta, f_i)$. Функция включает гиперпараметры A, β. Число параметров модели не должно превышать заданное число W^*. Число порождающих функций, из которых она состоит не должно превышать задан-

*Работа выполнена при финансовой поддержке РФФИ, проекты №07-07-00181, 08-01-12022.

ное число r^*. Модель, удовлетворяющую вышеперечисленным требованиям, будем называть моделью оптимальной структуры.

Распределение параметров моделей

Воспользуемся двухуровневым Байесовским выводом для оценки степени предпочтения порождаемых регрессионной моделей. Рассмотрим конечное множество моделей $f_1, \ldots, f_M$, приближающих данные D, обозначим априорную вероятность i-й модели $P(f_i)$. При появлении данных апостериорная вероятность модели $P(f_i \,|\, D)$ равна

$$P(f_i \,|\, D) = \frac{p(D \,|\, f_i)P(f_i)}{\sum_{j=1}^{M} p(D \,|\, f_j)P(f_j)}, \qquad (1)$$

где $p(D \,|\, f_i)$ — функция правдоподобия моделей, определяющая, насколько хорошо модель f_i описывает данные D. Знаменатель дроби обеспечивает выполнение условия $\sum_{i=1}^{M} P(f_i \,|\, D) = 1$.

Сравним две модели с помощью апостериорных вероятностей

$$\frac{P(f_i \,|\, D)}{P(f_j \,|\, D)} = \frac{p(D \,|\, f_i)P(f_i)}{p(D \,|\, f_j)P(f_j)}. \qquad (2)$$

Левая часть выражения называется отношением правдоподобия моделей. Отношение $P(f_i)/P(f_j)$ называется отношением апостериорных предпочтений моделей. Полагая априорные вероятности моделей одинаковыми, используем функции правдоподобия для выбора моделей.

Так как рассматриваемые модели f зависят от настраиваемых параметров, представим правдоподобие моделей в виде интеграла по пространству параметров

$$p(D \,|\, f) = \int p(D \,|\, \boldsymbol{w}, f)p(\boldsymbol{w} \,|\, f)d\boldsymbol{w}. \qquad (3)$$

Априорная плотность распределения параметров $\boldsymbol{w}$ модели f на выборке D равна

$$p(\boldsymbol{w} \,|\, D, f) = \frac{p(D \,|\, \boldsymbol{w}, f)p(\boldsymbol{w} \,|\, f)}{p(D \,|\, f)}, \qquad (4)$$

где $p(\boldsymbol{w} \,|\, f)$ — априорно заданная плотность вероятности параметров, и $p(D \,|\, \boldsymbol{w}, f)$ — функция правдоподобия параметров. Выражения (1) и (4) называются формулами Байесовского вывода первого и второго уровня.

Рассмотрим следующую гипотезу порождения данных при восстановлении регрессии

$$y = f(\boldsymbol{w}, \boldsymbol{x}) + \nu.$$

Пусть случайная величина ν имеет нормальное распределение $\mathcal{N}(0, \sigma^2)$ с нулевым матожиданием и дисперсией σ^2, которая не зависит от свободной переменной. Для фиксированной модели f плотность вероятности появления данных

$$p(y \,|\, \boldsymbol{x}, \boldsymbol{w}, \beta, f) \equiv p(D \,|\, \boldsymbol{w}, \beta, f) = \frac{\exp(-\beta E_D)}{Z_D(\beta)}, \quad (5)$$

где $\beta = \sigma^{-2}$, а коэффициент Z_D задан выражением, нормирующим функцию плотности в соответствии с гауссовым распределением

$$Z_D(\beta) = \left(\frac{2\pi}{\beta}\right)^{\frac{N}{2}}. \qquad (6)$$

Функция регрессионных невязок, согласно гипотезе порождения данных, равна

$$E_D = \frac{1}{2}\sum_{n=1}^{N} \left(f(x_n) - y_n\right)^2. \qquad (7)$$

Рассмотрим вектор параметров модели как многомерную случайную величину $\boldsymbol{w}$. Пусть плотность распределения параметров имеет вид многомерного нормального распределения $\mathcal{N}(\boldsymbol{0}, A)$ с матрицей ковариации A,

$$p(\boldsymbol{w} \,|\, A, f) = \frac{\exp(-E_{\boldsymbol{w}})}{Z_{\boldsymbol{w}}(A)}, \qquad (8)$$

где A — ковариационная матрица случайной величины $\boldsymbol{w}$. Нормирующая константа $Z_{\boldsymbol{w}}(A)$ равна

$$Z_{\boldsymbol{w}}(A) = (2\pi)^{\frac{W}{2}} |A|^{\frac{1}{2}}, \qquad (9)$$

где W — число параметров модели f. Функция-штраф за большое значение параметров модели при нормальном распределении равна

$$E_{\boldsymbol{w}} = \frac{1}{2}\boldsymbol{w}^{\intercal} A \boldsymbol{w}. \qquad (10)$$

При заданной модели f и заданных значениях A и β выражение (4) принимает вид

$$p(\boldsymbol{w} \,|\, D, A, \beta, f) = \frac{p(D \,|\, \boldsymbol{w}, \beta, f)p(\boldsymbol{w} \,|\, A, f)}{p(D \,|\, A, \beta, f)}. \qquad (11)$$

Записывая функцию ошибки

$$S(\boldsymbol{w}) = \frac{1}{2}\boldsymbol{w}^{\intercal} A \boldsymbol{w} + \beta E_D, \qquad (12)$$

получаем вместо (11) выражение

$$p(\boldsymbol{w} \,|\, D, A, \beta, f) \propto \frac{\exp(-S(\boldsymbol{w}))}{Z_S},$$

где Z_S — нормирующий множитель. Символ f далее будет опущен для удобства обозначений.

Вычисление гиперпараметров

Предлагается итеративно найти параметры и гиперпараметры модели по отдельности. На каждой итерации сначала при фиксированных гиперпараметрах отыскиваются параметры путем оптимизации функционала (12). Используется алгоритм Левенберга–Марквардта. Затем по формулам, предложенным ниже, вычисляются гиперпараметры.

Предположим, что после очередного шага итерации нам известен локальный максимум (12) и он находится в точке $\boldsymbol{w}_0$. Для нахождения гиперпараметров приблизим (11) методом Лапласа. Для этого построим ряд Тейлора второго порядка логарифма числителя (11) в окрестности $\boldsymbol{w}_0$

$$-S(\boldsymbol{w}) \approx -S(\boldsymbol{w}_0) - \frac{1}{2}\Delta\boldsymbol{w}^{\intercal} H \Delta\boldsymbol{w}, \qquad (13)$$

где $\Delta\boldsymbol{w} = \boldsymbol{w} - \boldsymbol{w}_0$. В выражении (13) нет слагаемого первого порядка, так как предполагается, что $\boldsymbol{w}_0$ доставляет локальный минимум функции ошибки

$$\left.\frac{\partial S(\boldsymbol{w})}{\partial \boldsymbol{w}}\right|_{\boldsymbol{w}=\boldsymbol{w}_0} = \mathbf{0}.$$

Матрица H — матрица Гессе функции ошибок

$$H = -\nabla\nabla S(\boldsymbol{w})|_{\boldsymbol{w}=\boldsymbol{w}_0}. \qquad (14)$$

Применяя экспоненту к обеим частям выражения (13) получаем требуемое приближение числителя (11)

$$\exp(-S(\boldsymbol{w})) \approx \exp(-S(\boldsymbol{w}_0)) \exp\left(-\tfrac{1}{2}\Delta\boldsymbol{w}^{\intercal} H \Delta\boldsymbol{w}\right). \qquad (15)$$

Учитывая то, что интеграл выражения (11) должен равняться единице, получаем нормирующий множитель

$$Z_S = \frac{\exp(-S(\boldsymbol{w}_0))(2\pi)^{\frac{W}{2}}}{|H|^{\frac{1}{2}}}. \qquad (16)$$

Знаменатель (11) является числителем (1) и определяет выбор наиболее правдоподобной модели. Для нахождения гиперпараметров максимизируем функцию $p(D \mid A, \beta)$ относительно A и β. Запишем ее в виде

$$p(D \mid A, \beta) = \int p(D \mid \boldsymbol{w}, A, \beta) p(\boldsymbol{w} \mid A) d\boldsymbol{w}. \qquad (17)$$

Используя выражения (5) и (8) перепишем (17) в виде

$$p(D \mid \beta, A) = \frac{Z_S}{Z_{\boldsymbol{w}}(A) Z_D(\beta)}.$$

Из (6), (9)и (16), логарифмируя (17), получим

$$\ln p(D \mid A, \beta) = -\frac{1}{2}\ln|A| - \frac{N}{2}\ln 2\pi + \frac{N}{2}\ln\beta - \\ - \beta E'_D - E'_{\boldsymbol{w}} - \frac{1}{2}\ln|H|. \qquad (18)$$

Найдем максимум выражения (18) относительно гиперпараметров, приравняв его производную поочередно по A и β к нулю. Для упрощения вычислений представим $A = \operatorname{diag}(\alpha) I_W$.

$$\frac{d\ln p(D \mid A, \beta)}{d\alpha} = -E'_{\boldsymbol{w}} + \frac{\boldsymbol{w}}{2\alpha} + \frac{d}{d\alpha}\ln\det(H).$$

Производная последнего слагаемого равна

$$\frac{d}{d\alpha}\ln|H| = \sum_{j=1}^{W} \frac{1}{\lambda_j + \alpha},$$

где λ_j — собственные значения матрицы H. Приравнивая последнее выражение к нулю и преобразовывая, получаем выражение для α

$$2\alpha E'_{\boldsymbol{w}} = W - \gamma, \quad \text{где} \quad \gamma = \sum_{j=1}^{W} \frac{\alpha}{\lambda_j + \alpha}. \qquad (19)$$

Аналогично получим β

$$2\beta E'_D = N - \sum_{j=1}^{W} \frac{\lambda_j}{\lambda_j + \alpha} = N - \gamma. \qquad (20)$$

Гиперпараметры α и β_i вычисляется итеративно следующим образом

$$\beta^{\text{new}} = \frac{N - \gamma}{E'_D}, \quad \alpha^{\text{new}} = \frac{W - \gamma}{E'_{\boldsymbol{w}}}.$$

Значения функционалов ошибок $E'_{\boldsymbol{w}}$ и E'_D оптимизируются после каждого вычисления новых значений гиперпараметров.

При выборе моделей выполняется следующая процедура. Экспертно задается модель-претендент. Каждому элементу модели ставится в соответствие свой гиперпараметр α. Параметры и гиперпараметры модели последовательно настраиваются. Элемент модели, имеющий наименьшее значение гиперпараметра, исключается. Модель пополняется новым элементом из множества G согласно заданному правилу. Так как на каждом шаге такой модификации модели функционал качества не ухудшается, процедура выполняется до сходимости функционала качества (13).

Вычислительный эксперимент

Проиллюстрируем итеративное изменение параметров и гиперпараметров с помощью модели $y = f_0(\boldsymbol{w}, \boldsymbol{x}) = w_1 + w_2 \sin x_1 + w_3 \sin x_2$. Свободные переменные данной модели имеют значения $x_1, x_2 \in \{0, 0.1, \ldots, 1\}$. Зависимые переменные, полученные как $y = f_0(\boldsymbol{w}, \boldsymbol{x}) + \nu_0$, где $\nu_0 \sim \mathcal{N}(0, \pi/2)$.

На рис. 1 показаны итеративные изменения параметров w_1, w_2, w_3, латентной переменной γ и гиперпараметра β. По оси абсцисс отложен номер

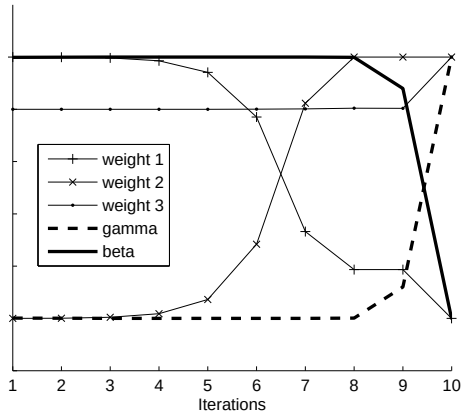


Рис. 1.

итерации. По оси ординат — нормированные значения переменных.

В результате эксперимента была получена сходящаяся последовательность, определяющая оптимальные параметры и гиперпараметры для данной модели. Итеративный алгоритм останавливается, когда значение функции ошибки текущей модели имеет значение меньше заданного, или же невозможна генерация новых моделей-претендентов.

Заключение

Данная работа описывает алгоритм выбора нелинейных регрессионных моделей из индуктивно-порождаемого множества. Для выбора моделей используются настраиваемые гиперпараметры. Каждый гиперпараметр ставится в соответствие элементы суперпозиции функций, задающих модель. По его значению определяется важность элемента суперпозиции и принимается решение о необходимости его модификации.

Литература

[1] *LeCun Y., Denker J., Solla S., Howard R. E., Jackel L. D.* Optimal brain damage // Advances in Neural Information Processing Systems II / edited by D. S. Touretzky. — San Mateo, CA: Morgan Kauffman, 1990. — P. 598–605.

[2] *Hassibi B., Stork D. G.* Second order derivatives for network pruning: Optimal brain surgeon // Advances in Neural Information Processing Systems / edited by S. J. Hanson, J. D. Cowan, C. L. Giles. — Vol. 5. — Morgan Kaufmann, San Mateo, CA, 1993. — P. 164–171.

[3] *MacKay D.* Information Theory, Inference, and Learning Algorithms. — Cambridge University Press, 2003. — 628 p.

[4] *Cavendish D. M., Mackay D. J.,C., Laboratory C.* Comparison of approximate methods for handling hyperparameters // Neural Computation. — 2003. — Vol. 11. — P. 1035–1068.

[5] *MacKay D. J.* Choice of basis for laplace approximation: Tech. rep.: Machine Learning, 1998.

[6] *Cawley G., Talbot N., Guyon I., Saffari A.* Preventing over-fitting during model selection using bayesian regularisation // Journal of Machine Learning Research. — 2007. — Vol. 8.

[7] *Bishop C.* Pattern Recognition And Machine Learning. — Springer, 2006.

[8] *Bishop C. M., Tipping M. E.* Bayesian regression and classification.

[9] *Barber D., Bishop C. M.* Ensemble learning in bayesian neural networks // Neural Networks and Machine Learning. — Springer, 1998. — P. 215–237.

[10] *Malada H. R., Ivakhnenko A. G.* Inductive Learning Algorithms for Complex Systems Modeling. — CRC Press, 1994. — 368 p.

[11] *Ивахненко А. Г., Юрачковский Ю. П.* Моделирование сложных систем по экспериментальным данным. — М.: Радио и связь, 1987. — 120 с.

[12] *Mueler J. A., Lemke F.* Sel-organising Data Mining: An Intelligent Approach To Extract Knowledge From Data. — Berlin: Dresden, 1999. — 225 p.

[13] *Koza J. R.* Genetic Programming IV: Routine Human-Competitive Machine Intelligence. — Springer, 2005.

[14] *Zelinka I., Nolle L., Oplatkova Z.* Analytic programming — symbolic regression by means of arbitrfary evolutionary algorithms // Journal of Simulation. — 2004. — Vol. 6(9). — P. 44–56.

[15] *Vladislavleva E.* Model-based Problem Solving through Symbolic Regression via Pareto Genetic Programming: PhD thesis. — Tilburg University, Tilburg, the Netherlands, 2008. — 288 p.

[16] *Vladislavleva E., Smith G., Hertog D.* Order of nonlinearity as a complexity measure for models generated by symbolic regression via pareto genetic programming // EEE Transactions on Evolutionary Computation. — 2009. — Vol. 13(2). — P. 333–349.

[17] *Стрижов В. В.* Поиск параметрической регрессионной модели в индуктивно заданном множестве // Журнал вычислительных технологий. — 2007. — № 1. — С. 93–102.

[18] *Стрижов В. В.* Методы индуктивного порождения регрессионных моделей. — М.: ВЦ РАН, 2008. — 54 с.

[19] *Стрижов В. В., Сологуб Р. А.* Индуктивное порождение регрессионных моделей предполагаемой волатильности для опционных торгов // Журнал вычислительных технологий. — 2009. — Т. 3.

Метод релевантных потенциальных функций для селективного комбинирования разнородной информации при обучении распознаванию образов на основе байесовского подхода*

Татарчук А. И.[1]***, Сулимова В. В.***[2]***, Моттль В. В.***[1]***, Уиндридж Д.***[3]
aitech@ya.ru, sulimova@uic.tula.net, vmottl@ya.ru, d.windridge@surrey.ac.uk
[1]Москва, ВЦ РАН; [2]Тула, ТулГУ; [3]Великобритания, Гилдфорд, Университет Суррей

Предлагается байесовская концепция обучения распознаванию объектов двух классов, представленных совокупностью признаков с произвольными шкалами измерения в предположении, что в шкале каждого признака определена потенциальная функция. Как следствие, шкала значений каждого признака оказывается погруженной в некоторое линейное пространство, в котором специфические линейные операции определяются видом соответствующей потенциальной функции. Задача обучения поставлена как задача поиска в декартовом произведении комбинируемых линейных пространств разделяющей гиперплоскости, оптимальной относительно обучающей совокупности в смысле В. Н. Вапника. Основная идея байесовской концепции обучения заключается в использовании параметрического семейства априорных распределения объектов двух классов в комбинированном линейном пространстве признаков вместе с априорными распределениями соответствующих компонент направляющего вектора разделяющей гиперплоскости в каждом из этих пространств. Специальный выбор априорных распределений приводит к эффекту селективности комбинирования признаков, т. е. к повышению степени участия в решающем правиле полезных потенциальных функций, называемых релевантными, и относительному подавлению остальных. В роли параметра семейства априорных распределений выступает неотрицательная переменная, названная параметром селективности обучения. При нулевом значении параметра байесовская оценка направляющего вектора разделяющей гиперплоскости содержит все исходные признаковые описания объектов, а при ее увеличении все большая часть из них практически полностью устраняется.

Стремление обеспечить требуемое качество распознавания объектов, недостижимое на основе какого-либо одного вида признакового описания, привело к идее комбинирования разнородных характеристик или, как принято их называть в англоязычной литературе, модальностей представления объектов (modalities) [1].

Каждая модальность $i = 1, \ldots, n$ выражается, как правило, в виде некоторого признакового представления $x_i(\omega) \in \mathbb{X}_i$ объектов реального мира $\omega \in \Omega$ в некоторой шкале $\mathbb{X}_i$, специфической для данного признака. Во многих практических ситуациях компьютерные представления объектов $x_i(\omega) \in \mathbb{X}_i$ не являются в исходном виде элементами линейного пространства, как, например, динамические характеристики подписи или спектральные разложения произнесений слов, представленных в виде сигналов разной длины. В простейшем случае, когда все признаки измеряются в шкале действительных чисел $\mathbb{X}_i = \mathbb{R}$, $i = 1, \ldots, n$, принято говорить об обучении распознаванию образов в конечномерном действительном линейном признаковом пространстве. Математическим инструментом погружения объектов произвольной природы в линейное пространство является *метод потенциальных функций* [2], который в сущности, стирает различие между действительными признаками и произвольными модальностями представления объектов.

Двухместная симметричная функция $K_i(x', x'')$: $\mathbb{X}_i \times \mathbb{X}_i \to \mathbb{R}$ называется *потенциальной функцией* (kernel function) на множестве $\mathbb{X}_i$, если для любой конечной совокупности элементов $\{x_1, \ldots, x_N\} \subseteq \mathbb{X}_i$ матрица ее значений $\big(K_i(x_j, x_k)\big)_{j=1}^{N}{}_{k=1}^{N}$ неотрицательно определена. Тогда существует гипотетическое линейное пространство $\tilde{\mathbb{X}}_i \supseteq \mathbb{X}_i$, в которое исходная шкала измерения рассматриваемого свойства объектов погружается в виде совокупности, быть может, изолированных точек, и в котором потенциальная функция играет роль скалярного произведения.

Для формирования общей потенциальной функции будем рассматривать декартово произведение этих линейных пространств $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{X}_1 \times \cdots \times \mathbb{X}_n \subseteq \tilde{\mathbb{X}} = \tilde{\mathbb{X}}_1 \times \cdots \times \tilde{\mathbb{X}}_n$, которое также является линейным пространством. В данной работе в качестве комбинированной потенциальной функции рассматривается сумма всех частных функций $K(\boldsymbol{x}', \boldsymbol{x}'') = \sum_{i=1}^{n} K_i(x_i', x_i'')$, определенных над шкалами исходных модальностей. Выбор вектора $\boldsymbol{a} \in \tilde{\mathbb{X}}$ и числового порога $b \in \mathbb{R}$ задает некоторую разделяющую гиперплоскость в сформированном комбинированном линейном пространстве

$$K(\boldsymbol{a}, \boldsymbol{x}) + b = \sum_{i=1}^{n} K_i(a_i, x_i) + b \gtrless 0. \qquad (1)$$

В признаковом случае $x_i(\omega) \in \mathbb{X}_i = \mathbb{R}$, потенциальная функция, заданная как произведение $K_i(x_i', x_i'') = x_i' x_i''$, определяет естественное погружение анализируемых объектов в евклидово пространство. Метод опорных векторов

*Работа выполнена при финансовой поддержке РФФИ, проекты №05-01-00679, №06-01-08042, №08-01-00695, №09-01-00573.

В. Н. Вапника [3], разработанный как метод обучения двухклассовому распознаванию образов в $\mathbb{R}^n$, осуществляет комбинирование нескольких признаков, используя $K(\boldsymbol{x}', \boldsymbol{x}'') = \sum_{i=1}^{n} x_i' x_i''$. Такая аналогия эксплуатируется модификациями метода опорных векторов для комбинирования нескольких потенциальных функций в общем виде.

Увеличение сложности модели при слишком большом числе представлений объектов неизбежно связано с опасностью переобучения, выражающейся в снижении обобщающей способности модели при обучении по недостаточно большой обучающей совокупности. В связи с этим, возникает необходимость регуляризации процесса обучения или ограничения свободы выбора модели решающего правила. Один из наиболее эффективных подходов к регуляризации основан на сокращении исходного множества представлений объектов.

В качестве методологической основы предлагаемого в данной работе *метода релевантных потенциальных функций с регулируемой селективностью* (Selective Relevance Kernel Machine) был принят классический метод опорных векторов, изначально сформулированный в сугубо детерминистских терминах. С другой стороны, использование байесовской методологии позволяет избежать необходимости «изобретения» новых эвристических алгоритмов. С целью преодоления этого противоречия предлагается специальная байесовская стратегия обучения оптимальной разделяющей гиперплоскости в линейном пространстве.

Байесовское обучение оптимальной разделяющей гиперплоскости

Будем предполагать, что линейное пространство наблюдаемых представлений объектов $\boldsymbol{x} = (x_1, \ldots, x_n) \in \tilde{\mathbb{X}} = \tilde{\mathbb{X}}_1 \times \cdots \times \tilde{\mathbb{X}}_n$ конечномерно. Пусть $K(\boldsymbol{a}, \boldsymbol{x}) + b = 0$ — некоторая гиперплоскость с направляющим вектором $\boldsymbol{a} = (a_1, \ldots, a_n) \in \tilde{\mathbb{X}}$ и параметром сдвига $b \in \mathbb{R}$. Свяжем с ней пару параметрических семейств плотностей распределения вероятностей $\varphi(\boldsymbol{x} \,|\, \boldsymbol{a}, b, y; c)$, $\boldsymbol{x} \in \tilde{\mathbb{X}}$ с параметрами $\boldsymbol{a} \in \tilde{\mathbb{X}}$, $b \in \mathbb{R}$, $y \in \{-1, 1\}$ и $c \in \mathbb{R}$:

$$\varphi(\boldsymbol{x} \,|\, \boldsymbol{a}, b, y; c) = \begin{cases} \text{const}, & yz(\boldsymbol{a}, \boldsymbol{x}) \geqslant 1, \\ e^{-c(1 - yz(\boldsymbol{a}, \boldsymbol{x}))}, & yz(\boldsymbol{a}, \boldsymbol{x}) < 1, \end{cases} \quad (2)$$

где $z(\boldsymbol{a}, \boldsymbol{x}) = K(\boldsymbol{a}, \boldsymbol{x}) + b$. Эти два семейства выражают предположение, что случайные векторы признаков объектов двух классов распределены главным образом в «своих» полупространствах $K(\boldsymbol{a}, \boldsymbol{x}) + b > 0$ и $K(\boldsymbol{a}, \boldsymbol{x}) + b < 0$, однако могут попадать и в «чужие» полупространства, причем степенью возможности «ошибочных» значений управляет параметр $c > 0$, рис. 1.

Значение «равномерной» плотности условно примем равным единице const = 1. Равномерное распределение в бесконечной области является некорректным вероятностным понятием, поскольку оно не образует единичного интеграла по всему линейному пространству. Такие плотности распределения принято называть несобственными плотностями распределения [4]. Некорректность несобственной пары плотностей (2) не приводит далее к математическим противоречиям, поскольку они участвуют только в формуле Байеса.

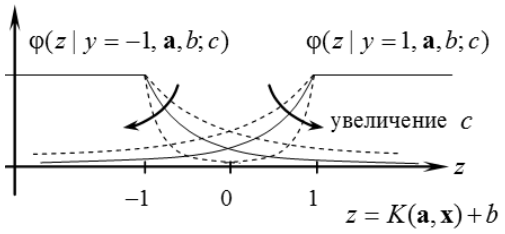


Рис. 1. Параметрическое семейство плотностей распределения в проекции на направляющий вектор гиперплоскости.

Будем далее предполагать, что получена обучающая совокупность $(X, Y) = \{(\boldsymbol{x}_j, y_j)\}_{j=1}^{N}$, образованная независимыми векторами $\boldsymbol{x}_j = (x_{ij})_{i=1}^{n}$ с известными индексами классов $y_j \in \{-1, 1\}$. Тогда условное распределение обучающей совокупности представимо в виде произведения плотностей (2):

$$\Phi(X \,|\, Y, \boldsymbol{a}, b; c) = \prod_{j=1}^{N} \varphi(\boldsymbol{x}_j \,|\, \boldsymbol{a}, b, y_j; c).$$

Другим ключевым предположением в предлагаемой вероятностной модели является суждение об априорном распределении $\Psi(\boldsymbol{a}, b)$ параметров $(\boldsymbol{a}, b)$ разделяющей гиперплоскости $K(\boldsymbol{a}, \boldsymbol{x}) + b = 0$. Будем считать, что отсутствуют какие-либо априорные предпочтения величин порога $b \in \mathbb{R}$ разделяющей гиперплоскости, тогда $\Psi(\boldsymbol{a}, b) \propto \Psi(\boldsymbol{a})$.

Апостериорная плотность распределения параметров относительно обучающей совокупности определяется формулой Байеса

$$p(\boldsymbol{a}, b \,|\, X, Y; c) \propto \Psi(\boldsymbol{a})\, \Phi(X \,|\, Y, \boldsymbol{a}, b; c). \quad (3)$$

Максимизация этой апостериорной плотности в пространстве параметров модели приводит к очевидному критерию обучения

$$(\hat{\boldsymbol{a}}, \hat{b} \,|\, X, Y; c) = \arg\max_{\boldsymbol{a}, b} \Psi(\boldsymbol{a})\, \Phi(X \,|\, Y, \boldsymbol{a}, b; c) = \\ = \arg\max_{\boldsymbol{a}, b} \big(\ln \Psi(\boldsymbol{a}) + \ln \Phi(X \,|\, Y, \boldsymbol{a}, b; c)\big). \quad (4)$$

Теорема 1. *Критерий обучения* (4) *эквивалентен минимизации целевой функции* $J(\boldsymbol{a}, b, \boldsymbol{\delta} \,|\, c)$, $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_N)$ *на выпуклом множестве, заданном набором линейных ограничений-неравенств для объектов обучающей совокупности:*

$$\begin{cases} J(\boldsymbol{a}, b, \boldsymbol{\delta} \,|\, c) = -\ln \Psi(\boldsymbol{a}) + c \sum_{j=1}^{N} \delta_j \to \min, \\ y_j \big(K(\boldsymbol{a}, \boldsymbol{x}_j) + b\big) \geqslant 1 - \delta_j,\ \delta_j \geqslant 0,\ j = 1, \ldots, N, \end{cases} \quad (5)$$

Важным здесь является тот факт, что концепция оптимальной разделяющей гиперплоскости, определяемой произвольным выбором априорных предпочтений на множестве значений направляющего вектора, сохраняет понятие активных ограничений в точке минимума критерия $y_j\,(K(\boldsymbol{a},\boldsymbol{x})+b)=1-\delta_j$, т. е. понятие *опорных векторов* в составе обучающей совокупности.

Заметим также, что параметр $c>0$ семейств распределений (2), отвечающий за априорную способность объектов генеральной совокупности нарушать границу своих классов, регламентирует приоритет между качеством разделения обучающей совокупности и априорными предпочтениями по выбору направляющего вектора.

По своей структуре оптимизационный критерий (5) представляет собой обобщение классического метода опорных векторов и отличается от него только слагаемым $-\ln\Psi(\boldsymbol{a})$ вместо квадрата нормы направляющего вектора искомой разделяющей гиперплоскости $K(\boldsymbol{a},\boldsymbol{a})=\sum\limits_{i=1}^{n}K_i(a_i,a_i)$.

Рассмотрим задачу обучения (5), приняв дополнительное предположение, что случайный направляющий вектор образован независимыми компонентами с плотностями распределения вида

$$\psi_i(a_i)\propto\exp\bigl(-\tfrac{1}{2}K_i(a_i,a_i)\bigr),\ i=1,\ldots,n.$$

Каждая такая плотность выражает круговое нормальное распределение с нулевым математическим ожиданием относительно некоторого ортонормированного базиса в соответствующем конечномерном линейном пространстве $\tilde{\mathbb{X}}$ с одинаковыми единичными дисперсиями по каждой координате. Нормирующие множители у этих плотностей опущены, поскольку дальнейший анализ инвариантен к конкретным значениям размерностей этих пространств. Тогда $\Psi(\boldsymbol{a})\propto\exp\bigl(-\frac{1}{2}K(\boldsymbol{a},\boldsymbol{a})\bigr)$,

$$\ln\Psi(\boldsymbol{a})=\mathrm{const}-\tfrac{1}{2}K(\boldsymbol{a},\boldsymbol{a}),$$

и задача (5) примет вид классического критерия обучения по методу опорных векторов при $C=2c$:

$$\begin{cases}J(\boldsymbol{a},b,\boldsymbol{\delta}\,|\,c)=K(\boldsymbol{a},\boldsymbol{a})+C\sum_{j=1}^{N}\delta_j\to\min,\\ y_j\bigl(K(\boldsymbol{a},\boldsymbol{x}_j)+b\bigr)\geqslant 1-\delta_j,\ \delta_j\geqslant 0,\ j=1,\ldots,N.\end{cases}\qquad(6)$$

Метод релевантных потенциальных функций

Будем предполагать, что компоненты направляющего вектора $\boldsymbol{a}=(a_1,\ldots,a_n)$ независимы и имеют квазинормальные априорные распределения с нулевыми математическими ожиданиями и различными дисперсиями $\boldsymbol{r}=(r_1,\ldots,r_n)$. Тогда $\psi_i(a_i\,|\,r_i)\propto\exp\bigl(-\frac{1}{2r_i}K_i(a_i,a_i)\bigr)$, $i=1,\ldots,n$ и

$$\ln\Psi(\boldsymbol{a}\,|\,\boldsymbol{r})=\mathrm{const}-\sum_{i=1}^{n}\tfrac{1}{2r_i}K_i(a_i,a_i).\qquad(7)$$

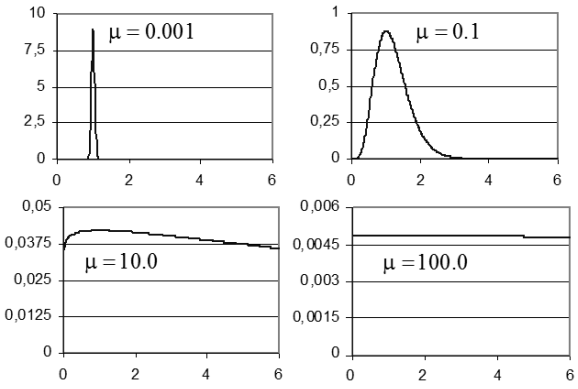


Рис. 2. Зависимость априорного гамма-распределения меры точности q_i от параметра μ.

Примем дополнительное предположение, что дисперсии $\boldsymbol{r}$ сами являются независимыми случайными величинами, характеризующимися некоторым априорным распределением, и подлежащими оцениванию в процессе обучения по байесовскому принципу вместе с параметрами разделяющей гиперплоскости $(\boldsymbol{a},b)$. Удобно оперировать обратными дисперсиями $q_i=\frac{1}{r_i}$, $i=1,\ldots,n$, называемыми мерами точности соответствующих случайных величин, полагая, что обратные дисперсии имеют одно и то же априорное гамма-распределение

$$\gamma(q_i\,|\,\alpha,\beta)\propto q_i^{\alpha-1}\exp(-\beta q_i).\qquad(8)$$

Будем выбирать значения двух параметров априорного гамма-распределения α и β, задавая один общий параметр $\mu\in(0,\infty)$ так, что $\alpha=1+\frac{1}{2\mu}$, $\beta=\frac{1}{2\mu}$. В этом случае параметрическое семейство гамма-распределений (8) будет иметь вид

$$\gamma(q_i\,|\,\mu)\propto q_i^{1/2\mu}\exp\bigl(-q_i/2\mu\bigr).$$

Нетрудно убедиться, что при увеличении параметра μ от 0 до ∞ вид априорного гамма-распределения (8) изменяется от сконцентрированного вокруг значения $q_i=1$ до почти «равномерного» на интервале $0\leqslant q_i<\infty$, рис. 2.

Обозначим через $G(\boldsymbol{q}\,|\,\mu)$ совместную априорную плотность распределения мер точности $\boldsymbol{q}=(q_1,\ldots,q_n)=\bigl(\frac{1}{r_1},\ldots,\frac{1}{r_n}\bigr)$:

$$G(\boldsymbol{q}\,|\,\mu)\propto\prod_{i=1}^{n}q_i^{1/2\mu}\exp\Bigl(-\frac{1}{2\mu}\sum_{i=1}^{n}q_i\Bigr).$$

Тогда совместная априорная плотность распределения параметров разделяющей гиперплоскости $\boldsymbol{a}$ и дисперсий $\boldsymbol{r}$ компонент направляющего вектора запишется как

$$\bar{\Psi}(\boldsymbol{a},b,\boldsymbol{q}\,|\,\mu)=\Psi(\boldsymbol{a}\,|\,\boldsymbol{q})\,G(\boldsymbol{q}\,|\,\mu).$$

Соответственно, совместное апостериорное распределение всех переменных модели относительно

обучающей совокупности, аналогично (3), пропорционально произведению

$$\bar{P}(\boldsymbol{a}, b, \boldsymbol{q} \,|\, X, Y; c, \mu) \propto \bar{\Psi}(\boldsymbol{a}, b, \boldsymbol{q} \,|\, \mu)\, \Phi(X \,|\, Y, \boldsymbol{a}, b, c).$$

Принцип максимизации совместной апостериорной плотности распределения в пространстве параметров разделяющей гиперплоскости и величин дисперсий компонент направляющего вектора приводит к оптимизационной задаче обучения

$$\bar{P}(\boldsymbol{a}, b, \boldsymbol{q} \,|\, X, Y, \mu) \to \max. \tag{9}$$

Теорема 2. *Оптимизационная задача обучения* (9) *эквивалентна критерию*

$$\begin{cases} J(\boldsymbol{a}, b, \boldsymbol{r}, \boldsymbol{\delta} \,|\, c) = \\ = \sum\limits_{i=1}^{n} \Big[\frac{1}{r_i}\Big(K_i(a_i, a_i) + \frac{1}{\mu}\Big) + \Big(\frac{1}{\mu} + 1 + \mu\Big) \ln r_i\Big] + \\ \qquad + C \sum\limits_{j=1}^{N} \delta_j \to \min, \\ y_j\Big(\sum\limits_{i=1}^{n} K_i(a_i, x_{ij}) + b\Big) \geqslant 1 - \delta_j,\ \delta_j \geqslant 0, \\ \qquad j = 1, \ldots, N,\ r_i \geqslant \varepsilon, \end{cases} \tag{10}$$

где $\varepsilon > 0$ *— небольшая неотрицательная величина.*

Предлагается оценивать априори неизвестные величины дисперсий $\boldsymbol{r}$ по обучающей выборке непосредственно в процессе обучения, как решение оптимизационной задачи (10). Для решения оптимизационной задачи (10) используется метод Гаусса-Зайделя с поочередной оптимизацией по двум группам переменных $(\boldsymbol{a}, b, \boldsymbol{\delta})$ и $\boldsymbol{r}$ с начальными условиями $\boldsymbol{r}^0 = (r_1^0, \ldots, r_n^0)$. Для фиксированных в задаче (10) значений дисперсий $\boldsymbol{r}$ оптимальные значения абстрактных компонент направляющего вектора $a_i = r_i \sum\limits_{j=1}^{N} y_j \lambda_j x_{ij},\ i = 1, \ldots, n$ выражаются в явном виде через неотрицательные множители Лагранжа $\lambda_1, \ldots, \lambda_N \geqslant 0$ соответствующей двойственной задачи

$$\begin{cases} \sum\limits_{j=1}^{N} \lambda_j - \frac{1}{2} \sum\limits_{j=1}^{N} \sum\limits_{l=1}^{N} y_j y_l \Big(\sum\limits_{i=1}^{n} r_i K_i(x_{ij} x_{il})\Big) \lambda_j \lambda_l \to \max, \\ \sum\limits_{j=1}^{N} y_j \lambda_j = 0,\ \ 0 \leqslant \lambda_j \leqslant \frac{1}{2} C,\ \ j = 1, \ldots, N. \end{cases}$$

Решающее правило (1) примет вид

$$\sum_{j:\ \lambda_j > 0} y_j \lambda_j \sum_{i=1}^{n} r_i K_i(x_{ij}, x_i) + b \geqslant 0. \tag{11}$$

Структура решающего правила (11) явно демонстрирует механизм управления участием признаков в процессе обучения. Чем больше принятое значение априорной дисперсии r_i некоторой компоненты направляющего вектора относительно дисперсий других компонент, тем большее влияние i-й признак оказывает на решение о классе объекта.

Найдя на очередном шаге решение $(\lambda_1^k, \ldots, \lambda_N^k)$ двойственной задачи для текущего приближения $\boldsymbol{r}^k = (r_1^k, \ldots, r_n^k)$, уточненные значения дисперсий $\boldsymbol{r}^{k+1} = (r_1^{k+1}, \ldots, r_n^{k+1})$ определяются выражением

$$r_i^{k+1} = (r_i^k)^2 \frac{\sum\limits_{j:\ \lambda_j^k > 0} \sum\limits_{l:\ \lambda_l^k > 0} y_j y_l K_i(x_{ij}, x_{il}) \lambda_j^k \lambda_l^k + \frac{1}{\mu}}{\frac{1}{\mu} + 1}.$$

Оптимальное значение величины порога $b \in \mathbb{R}$ определяется выражением

$$b = \frac{\sum\limits_{j:\ 0 < \lambda_j < \frac{C}{2}} \sum\limits_{l:\ \lambda_l > 0} y_l \sum\limits_{i=1}^{n} r_i K_i(x_{ij}, x_{il}) \lambda_j \lambda_l + \frac{C}{2} \sum\limits_{j:\ \lambda_j = \frac{C}{2}} y_j}{\sum\limits_{j:\ 0 < \lambda_j < \frac{C}{2}} \lambda_j}.$$

Предлагаемая итерационная процедура обычно сходится за 10–15 шагов, а сам алгоритм демонстрирует способность подавлять неинформативные признаки за счет выбора очень маленьких весов r_i в решающем правиле (11). При нулевой селективности $\mu = 0$ все оптимальные дисперсии равны единице $\hat{r}_1 = \ldots \hat{r}_n = 1$, а критерий (10) становится эквивалентным критерию (6) классического метода опорных векторов с минимальной селективностью.

При увеличении селективности $\mu \to \infty$ процедура обучения исключает из решающего правила (11) поочередно все признаки объектов, для которых оптимальные значение дисперсий стремятся к нижнему порогу $\hat{r}_i \to \varepsilon$. Оптимальное значение параметра селективности $0 \leqslant \mu < \infty$, который является структурным параметром метода обучения, предлагается оценивать по результатам проведения перекрестной проверки (cross validation).

Результаты экспериментального исследования обобщающей способности метода релевантных потенциальных функций изложены в работе [5].

Литература

[1] *Ross A., Jain A. K.* Multimodal biometrics: An overview. Proceedings of the 12th European Signal Processing Conference, 2004. — Pp. 1221–1224.

[2] *Айзерман М. А., Браверманн Э. М., Розеноэр Л. И.* Метод потенциальных функций в теории обучения машин. — М.: Наука, 1970, 384 с.

[3] *Vapnik V.* Statistical Learning Theory. — John-Wiley & Sons, Inc. 1998.

[4] *Де Гроот М.* Оптимальные статистические решения. — М.: Мир, 1974, 196 с.

[5] *Татарчук А. И., Урлов Е. Н., Ляшко А. С., Моттль В. В.* Экспериментальное исследование обобщающей способности методов селективного комбинирования потенциальных функций в задаче двухклассового распознавания образов // Всероссийская конференция ММРО-14. — М.: МАКС Пресс, 2009. — С. 196–199.

Метод опорных потенциальных функций в задаче селективного комбинирования разнородной информации при обучении распознаванию образов*

Татарчук А. И.[1]***, Урлов Е. Н.***[2]***, Моттль В. В.***[1]
aitech@ya.ru, factum_dao@mail.ru, vmottl@ya.ru
[1]Москва, Вычислительный центр РАН;
[2]Московский физико-технический институт

В работе [1] предложена байесовская концепция обучения в задаче двухклассового распознавания образов при одновременном построении оптимальной разделяющей гиперплоскости в декартовом произведении линейных пространств, образованных множествами значений каждого признака, и управляемом выборе эффективного подмножества признакового описания. Предполагается, что в каждом из линейных пространств определено априорное распределение соответствующей компоненты направляющего вектора. Параметрическое семейство априорных распределений, предложенное в [1], приводит к эффекту повышения степени участия в решающем правиле полезных признаков, т. е. связанных с ними потенциальных функций, названных релевантными, и относительному подавлению остальных. Эта же концепция селекции потенциальных функций в процессе обучения используется и в данной статье, однако рассматривается другое семейство априорных распределений, приводящее к выделению подмножества опорных потенциальных функций и к полному устранению остальных. Критерий обучения содержит параметр селективности комбинирования потенциальных функций, определяющий среди них долю опорных.

Введение

В работе [1] предложена вероятностная модель генеральной совокупности объектов реального мира $\omega \in \Omega$ в терминах значений их признаков $\boldsymbol{x}(\omega) = \big(x_1(\omega), \ldots, x_n(\omega)\big)$ с любыми шкалами значений $x_i \in \mathbb{X}_i$ в предположении, что в каждой шкале определена потенциальная функция $K_i(x_i', x_i'')$: $\mathbb{X}_i \times \mathbb{X}_i \to \mathbb{R}$. Это предположение позволяет рассматривать для каждого признака линейное пространство $\tilde{\mathbb{X}}_i \supseteq \mathbb{X}_i$, являющееся линейным замыканием его шкалы относительно линейных операций, определяемых соответствующей потенциальной функцией.

Основным элементом вероятностной модели является параметрическое семейство плотностей распределения объектов двух классов $y(\omega) \in \{-1, 1\}$, определяемое объективно существующей гиперплоскостью $\sum_{i=1}^{n} K_i(a_i, x_i) + b = 0$ в комбинированном линейном признаковом пространстве $\tilde{\mathbb{X}} = \tilde{\mathbb{X}}_1 \times \ldots \times \tilde{\mathbb{X}}_n$ с неизвестным направляющим вектором $\boldsymbol{a} = (a_1, \ldots, a_n) \in \tilde{\mathbb{X}}$ при некотором предположении о его априорной плотности распределения $\Psi(\boldsymbol{a}) = \Psi(a_1, \ldots, a_n)$ как совокупности случайных значений в линейном пространстве $\tilde{\mathbb{X}}_1 \times \ldots \times \tilde{\mathbb{X}}_n$.

Предположения приняты в такой форме, что принцип максимума апостериорной вероятности параметров $(\boldsymbol{a}, b)$ скрытой разделяющей гиперплоскости, для заданной случайной обучающей совокупности $\big\{(\boldsymbol{x}_j, y_j)\big\}_{j=1}^{N}$, где $\boldsymbol{x}_j = (x_{1j}, \ldots, x_{nj})$, $x_{ij} = x_i(\omega_j)$, $y_j = y(\omega_j)$, сводится к решению оптимизационной задачи

$$\begin{cases} J(a_1, \ldots, a_n, b, \delta_1, \ldots, \delta_N \,|\, c) = \\ \qquad = -\ln \Psi(a_1, \ldots, a_n) + c \sum\limits_{j=1}^{N} \delta_j \to \min; \\ y_j \left(\sum\limits_{i=1}^{n} K_i(a_i, x_{ij}) + b \right) \geqslant 1 - \delta_j; \\ \qquad\qquad \delta_j \geqslant 0, \;\; j = 1, \ldots, N; \end{cases} \tag{1}$$

которая по своей структуре является обобщением критерия метода опорных векторов, сформулированного В. Н. Вапником в сугубо детерминистких терминах [2], и отличается от него только слагаемым $-\ln \Psi(a_1, \ldots, a_n)$ вместо $\sum_{i=1}^{n} K_i(a_i, a_i)$.

Приняв предположение, что случайный направляющий вектор $\boldsymbol{a} \in \tilde{\mathbb{X}}$ образован независимыми компонентами с квазинормальными плотностями распределения

$$\psi_i(a_i) \propto \exp\big(-\tfrac{1}{2} K_i(a_i, a_i)\big), \;\; i = 1, \ldots, n, \tag{2}$$

получим $-\ln \Psi(a_1, \ldots, a_n) = \frac{1}{2} \sum_{i=1}^{n} K_i(a_i, a_i)$, а задача обучения (1) примет классической вид критерия обучения по методу опорных векторов.

Метод опорных потенциальных функций с управляемой селективностью

В работе [1] эффект селективности комбинирования потенциальных функций достигается тем, что, в отличие от классического метода опорных векторов (2), дисперсии их априорных распределений полагаются неизвестными

$$\psi_i(a_i \,|\, r_i) \propto \exp\big(-\tfrac{1}{2r_i} K_i(a_i, a_i)\big), \quad i = 1, \ldots, n.$$

Оценки этих дисперсий $(\hat{r}_1, \ldots, \hat{r}_n)$ определяли веса вхождения комбинируемых потенциальных функ-

*Работа выполнена при финансовой поддержке РФФИ, проекты №05-01-00679, №06-01-08042, №08-01-00695, №09-01-00573.

ций в решающее правило распознавания, внося эффект селективности признаков в процесс обучения.

Соответствующий метод обучения, предполагающий возможность управления степенью концентрации весов на малом числе признаков, назван в [1] *методом релевантных потенциальных функций с регулируемой селективностью* (Selective Relevance Kernel Machine). В данной работе рассматривается другой принцип выражения степени полезности признаков, заключающийся в жестком отборе подмножества информативных признаков, который мы называем *методом опорных потенциальных функций с регулируемой селективностью* (Selective Support Kernel Machine).

По-прежнему будем предполагать, что компоненты направляющего вектора $\boldsymbol{a} = (a_1, \ldots, a_n) \in \tilde{\mathbb{X}}$ априори независимы, однако их априорные плотности примем в виде

$$\psi(a_i) \propto \exp(-q(a_i \,|\, \mu)), \tag{3}$$

$$q(a_i \,|\, \mu) = \begin{cases} 2\mu\sqrt{K_i(a_i, a_i)}, & \text{если } \sqrt{K_i(a_i, a_i)} \leqslant \mu, \\ \mu^2 + K_i(a_i, a_i), & \text{если } \sqrt{K_i(a_i, a_i)} > \mu, \end{cases}$$

где $\mu \geqslant 0$ — неотрицательный параметр.

При таком предположении априорная плотность совместного распределения компонент направляющего вектора примет вид

$$\Psi(a_1, \ldots, a_n) \propto \exp\biggl(-\sum_{i=1}^{n} q(a_i \,|\, \mu)\biggr),$$

и поиск параметров разделяющей гиперплоскости, доставляющих максимум апостериорной плотности распределения ее параметров для заданной обучающей совокупности, сведется к решению следующей оптимизационной задачи по переменным $a_i \in \tilde{\mathbb{X}}_i$, $i = 1, \ldots, n$, $b \in \mathbb{R}$, $\delta_j \geqslant 0$, $j = 1 \ldots, N$:

$$\begin{cases} J(a_1, \ldots, a_n, b, \delta_1, \ldots, \delta_N \,|\, c) = \\ \qquad = \sum\limits_{i=1}^{n} q(a_i \,|\, \mu) + c\sum\limits_{j=1}^{N} \delta_j \to \min, \\ y_j \left(\sum\limits_{i=1}^{n} K_i(a_i, x_{ij}) + b\right) \geqslant 1 - \delta_j, \\ \qquad\qquad \delta_j \geqslant 0, \quad j = 1, \ldots, N. \end{cases} \tag{4}$$

Функция $q(a_i \,|\, \mu)$ в критерии обучения зависит от параметра $0 \leqslant \mu < \infty$, определяющего желаемый уровень селективности комбинирования потенциальных функций. При $\mu = 0$ эта функция приобретает вид $q(a_i \,|\, \mu) = K_i(a_i, a_i)$, и критерий (4) совпадает с обычным критерием метода опорных векторов, сохраняющим все потенциальные функции. При увеличении $\mu \to \infty$ всё большее число элементов направляющего вектора, как мы увидим, будут приобретать строго нулевые значения $a_i = \phi_i \in \tilde{\mathbb{X}}_i$, в смысле совпадения с нулевым элементом линейного пространства, в которое i-я потенциальная функция погружает шкалу значений соответствующего признака $\mathbb{X}_i \subseteq \tilde{\mathbb{X}}_i$.

Оптимизационная задача (4) выпукла, следовательно, любое ее решение будет глобальным. Минимизируемая целевая функция критерия не является строго выпуклой, и поэтому решение может быть не единственным.

Двойственная задача обучения

Рассмотрим следующую задачу оптимизации, которая будет играть роль двойственной по отношению к задаче обучения (4):

$$\begin{cases} \frac{1}{2}\sum\limits_{i=1}^{n} \xi_i + c\sum\limits_{j=1}^{N} \lambda_j \to \max, \\ \xi_i \leqslant \mu^2 - \sum\limits_{j=1}^{N}\sum\limits_{l=1}^{N} y_l\, y_j\, K_i(x_{ij}, x_{il})\, \lambda_j \lambda_l, \\ \xi_i \leqslant 0, \quad i = 1, \ldots, n, \\ 0 \leqslant \lambda_j \leqslant \frac{1}{2}c, \quad j = 1, \ldots, N, \quad \sum\limits_{j=1}^{N} \lambda_j y_j = 0. \end{cases} \tag{5}$$

Это задача максимизации линейной функции переменных $\xi_1, \ldots, \xi_n$, соответствующих комбинируемым потенциальным функциям, и $\lambda_1, \ldots, \lambda_N$, соответствующих объектам обучающей совокупности, при квадратичных и линейных ограничениях типа неравенств [3]. Для подобных задач существуют алгоритмы и программные системы, способные эффективно находить численное решение.

Пусть найдено решение задачи (5) $\xi_1, \ldots, \xi_n$ и $\lambda_1, \ldots, \lambda_n$. Заметим, что невозможна ситуация $\mu^2 - \sum\limits_{j=1}^{N}\sum\limits_{l=1}^{N} y_l y_j K_i(x_{ij}, x_{il})\lambda_j\lambda_l > \xi_i$ и $\xi_i < 0$. Отсюда следует, что найденное решение разбивает множество потенциальных функций $I = \{1, \ldots, n\}$ на три непересекающиеся подмножества:

$$\begin{aligned} I^+ &= \Bigl\{i \in I\colon \sum_{j=1}^{N}\sum_{l=1}^{N} y_j y_l K_i(x_{ij}, x_{il})\lambda_j\lambda_l > \mu^2\Bigr\}, \\ I^0 &= \Bigl\{i \in I\colon \sum_{j=1}^{N}\sum_{l=1}^{N} y_j y_l K_i(x_{ij}, x_{il})\lambda_j\lambda_l = \mu^2\Bigr\}, \\ I^- &= \Bigl\{i \in I\colon \sum_{j=1}^{N}\sum_{l=1}^{N} y_j y_l K_i(x_{ij}, x_{il})\lambda_j\lambda_l < \mu^2\Bigr\}. \end{aligned} \tag{6}$$

Теорема 1. *Компоненты направляющего вектора разделяющей гиперплоскости $a_i \in \tilde{\mathbb{X}}_i$ в решении задачи обучения (4) имеют вид*

$$a_i = \begin{cases} \sum\limits_{j:\lambda_j > 0} y_j \lambda_j x_{ij}, & i \in I^+, \\ \eta_i \sum\limits_{j:\lambda_j > 0} y_j \lambda_j x_{ij}, & 0 \leqslant \eta_i \leqslant 1, \quad i \in I^0, \\ \phi_i, & i \in I^-. \end{cases} \tag{7}$$

Таким образом, решение двойственной задачи полностью определяет значения компонент направляющего вектора a_i для потенциальных функций, попадающих в множества I^+ и I^-. Что же касается компонент, попадающих в множество I^0, то для них

пока не определены коэффициенты η_i в (7). Кроме того, не определено значение сдвига разделяющей гиперплоскости в (4).

Теорема 2. *Коэффициенты η_i для $i \in I^0$ и параметр сдвига гиперплоскости b находятся как решение задачи линейного программирования:*

$$\begin{cases} \mu\eta_i + c\sum\limits_{j=1}^{n}\delta_j \to \min(i \in I^0;\, b;\, \delta_1, \ldots, \delta_N), \\ \sum\limits_{i\in I^0} g_{ij}\eta_i + y_j b + \delta_j \geqslant h_j, \;\; j = 1, \ldots, N, \\ 0 \leqslant \eta_i \leqslant 1, \;\; i \in I^0, \;\; \delta_j \geqslant 0, \;\; j = 1, \ldots, N, \end{cases} \tag{8}$$

где

$$g_{ij} = \sum_{l=1}^{N} y_j y_l K_i(x_{ij}, x_{il})\lambda_l,$$

$$h_j = 1 - \sum_{i\in I^+}\sum_{l=1}^{N} y_j y_l K_i(x_{ij}, x_{il})\lambda_l.$$

Опорные потенциальные функции

После того, как найдены решения двойственной задачи (5) и дополнительной задачи (8), исходная задача обучения (4) решена полностью, т. е. найдены коэффициенты λ_j, $j = 1, \ldots, N$, и η_i, $i \in I^0$, определяющие компоненты направляющего вектора оптимальной разделяющей гиперплоскости a_i согласно (7), и ее параметр сдвига b.

Правда, элементы направляющего вектора не могут быть выражены в явной форме, поскольку содержатся, вообще говоря, в гипотетическим линейных замыканиях шкал измерения соответствующих признаков $a_i \in \hat{\mathbb{X}}_i \supseteq \mathbb{X}_i$, но, как правило, не в самих шкалах $a_i \notin \mathbb{X}_i$, однако этого и не требуется.

Согласно (1) и (4), оптимальное решающее правило $\sum\limits_{i=1}^{n} K_i(a_i, x_{ij}) + b \gtrless 0$, применимое к новому объекту с вектором признаков $(x_1, \ldots, x_n)$, после подстановки (7) примет вид:

$$\sum_{j:\,\lambda_j>0} y_j\lambda_j\Big(\sum_{i\in I^+} K_i(x_{ij}, x_i) + \sum_{\substack{i\in I^0\\ \eta_i>0}} \eta_i K_i(x_{ij}, x_i)\Big) + b \gtrless 0. \tag{9}$$

В решающее правило не вошли признаки $i \in I^-$, точнее, определенные на их шкалах потенциальные функции, поскольку соответствующие элементы оптимальной разделяющей гиперплоскости оказались нулевыми. Заведомо входят в него потенциальные функции $i \in I^+$, что же касается потенциальных функций $i \in I^0$, то в решающем правиле остаются только те из них, для которых значения коэффициентов η_i, найденные при решении задачи (8), оказались ненулевыми.

Потенциальные функции, определяющие оптимальное решающее правило

$$I_{\text{supp}} = I^+ \cup \big\{i \in I^0 \colon \eta_i > 0\big\} \subseteq I, \tag{10}$$

естественно назвать опорными. То обстоятельство, что изложенный метод обучения жестко выделяет некоторое подмножество потенциальных функций, полностью устраняя остальные, позволил назвать его методом опорных потенциальных функций в отличие от метода релевантных потенциальных функций, рассмотренного в работе [1], который лишь снабжал потенциальные функции неотрицательными весами в соответствии с их оцененной степенью информативности.

Зависимость множества опорных потенциальных функций от параметра селективности

Структура подмножеств потенциальных функций (6) и вид решающего правила (9) явно указывают механизм зависимости состава множества опорных потенциальных функций от параметра μ в критерии обучения (4).

Если $\mu = 0$, то множество заведомо опорных потенциальных функций $I^+ \subseteq I = \{1, \ldots, n\}$ совпадает со всем множеством I. В этом частном случае функция (3) квадратична $q(a_i \,|\, \mu) = K_i(a_i, a_i)$ при всех $a_i \in \hat{\mathbb{X}}_i$, и критерий обучения (4) не отличается от обычного критерия в методе опорных векторов, не обладающего свойством селективности, так что все исходные потенциальные функции входят в получаемое решающее правило распознавания, являясь опорными.

Увеличение значения параметра μ приводит к тому, что все большее число потенциальных функций попадает в множество I^- заведомо неопорных (6), соответственно уменьшается множество опорных потенциальных функций (10). При неограниченном росте параметра селективности $\mu \to \infty$ в конце концов все потенциальные функции оказываются в множестве I^-, и опорных вообще не остается $I_{\text{supp}} = \varnothing$.

Выбор значения параметра селективности

Параметр селективности $0 \leqslant \mu < \infty$ является структурным параметром метода обучения, определяя последовательность вложенных классов моделей уменьшающейся размерности, аппроксимирующих обучающую совокупность. Принципиально невозможно «оценить» его значение, исходя непосредственно из результата обучения.

Наиболее эффективным методом выбора значения параметра селективности является перекрестная проверка (cross validation), основанная на непосредственном оценивании обобщающей способности обучения.

Результаты сравнительного экспериментального исследования обобщающей способности обучения методов релевантных и опорных потенциальных функций изложены в работе [4].

Литература

[1] *Татарчук А. И., Сулимова В. В., Моттль В. В., Уиндридж Д.* Метод релевантных потенциальных функций для селективного комбинирования разнородной информации при обучении распознаванию образов на основе байесовского подхода // Всероссийская конференция ММРО-14. — М.: МАКС Пресс, 2009. — С. 188–191.

[2] *Vapnik V.* Statistical Learning Theory. — John-Wiley & Sons, Inc. 1998, 736 p.

[3] *Boyd S., Vandenberghe L.* Convex Optimization. — Cambridge: Cambridge University Press, 2007, 716 p.

[4] *Татарчук А. И., Урлов Е. Н., Ляшко А. С., Моттль В. В.* Экспериментальное исследование обобщающей способности методов селективного комбинирования потенциальных функций в задаче двухклассового распознавания образов // Всероссийская конференция ММРО-14. — М.: МАКС Пресс, 2009. — С. 196–199.

Экспериментальное исследование обобщающей способности методов селективного комбинирования потенциальных функций в задаче двухклассового распознавания образов*

Татарчук А. И.[1]***, Урлов Е. Н.***[2]***, Ляшко А. С.***[2]***, Моттль В. В.***[1]
aitech@ya.ru, factum_dao@mail.ru, andrewlyashko@gmail.com, vmottl@ya.ru
Москва, [1]Вычислительный центр РАН; [2]Московский физико-технический институт

Излагаются результаты экспериментального исследования обобщающей способности методов отбора признаков непосредственно в процессе обучения распознаванию двух классов объектов на модельных и реальных данных. Сравниваются два варианта алгоритмической реализации байесовской концепции выбора эффективного подмножества признаковых описаний объектов, каждое из которых представлено потенциальной функцией в шкале значений соответствующего признака. Исследуемые варианты байесовского обучения предложены в работах [1] и [2] как методы, соответственно, релевантных и опорных потенциальных функций. Эти методы различаются только принятыми классами априорных распределений компонент искомого направляющего вектора разделяющей гиперплоскости, что приводит, однако, к существенно разным способам выражения предпочтений одних признаков перед другими относительно обучающей совокупности — определению неотрицательных весов при всех признаках в первом методе и жесткому выделению полезного подмножества во втором. В сравнительных модельных экспериментах случайная обучающая совокупность формируется в соответствии с заданной разделяющей гиперплоскостью, что позволяет непосредственно измерять обобщающую способность обучения. В качестве источника реальных данных используется задача диагностики рака легких из репозитория UCI.

Введение

Для практики распознавания образов типична ситуация, когда объекты представлены признаками $\boldsymbol{x} = (x_1, \ldots, x_n)$, измеряемыми в шкалах $x_i \in \mathbb{X}_i$, вообще говоря, разной природы. С целью преодоления проблемы несовместимости произвольных шкал значений признаков, методы селективного комбинирования совокупности представлений объектов при обучении распознаванию образов построены в работах [1] и [2] на основе математического аппарата потенциальных функций, в значительной мере стирающего различия между разными шкалами измерения признаков объектов.

Целью данной работы является экспериментальное сравнение этих двух существенно разных методов, названных *методом релевантных потенциальных функций* (Relevance Kernel Machine — RKM) и *методом опорных потенциальных функций* (Support Kernel Machine — SKM). Первый из них присваивает неотрицательные веса всем потенциальным функциям, определяя тем самым их неравное участие в решающем правиле распознавания, а второй жестко выделяет подмножество полезных потенциальных функций, полностью устраняя остальные.

Оба сравниваемых метода основаны на предположении, что в шкале измерения каждого признака $x_i \in \mathbb{X}_i$ определена потенциальная функция $K_i(x_i', x_i'')$: $\mathbb{X}_i \times \mathbb{X}_i \to \mathbb{R}$, т. е. симметричная функция $K_i(x_i', x_i'') = K_i(x_i'', x_i')$, образующая неотрицательно определенную матрицу для любой конечной совокупности значений $\{x_{i1}, \ldots, x_{ik}\}$. Это предположение позволяет рассматривать для каждого признака линейное пространство $\tilde{\mathbb{X}}_i \supseteq \mathbb{X}_i$ со скалярным произведением $K_i(x_i', x_i'')$, являющееся линейным замыканием исходной шкалы $\mathbb{X}_i$ относительно линейных операций, определяемых соответствующей потенциальной функцией [1].

Такое предположение заведомо выполнено в простейшем случае действительных признаков $x_i \in \mathbb{X}_i = \mathbb{R}$, поскольку произведение двух действительных чисел $K_i(x_i', x_i'') = x_i' x_i''$ является скалярным произведением в одномерном линейном пространстве, образованном действительной осью. Более того, действительная ось сама является своим линейным замыканием, так что в этом случае $\tilde{\mathbb{X}}_i = \mathbb{X}_i$.

Будучи примененными к задаче обучения распознаванию образов в конечномерном пространстве действительных признаков $\boldsymbol{x} = (x_1, \ldots, x_n) \in \in \mathbb{R}^n$, исследуемые здесь методы релевантных и опорных потенциальных функций полностью сохраняют свое исходное назначение, двумя разными способами ранжируя признаки по их полезности относительно предъявленной обучающей совокупности.

В силу этого обстоятельства в данной работе исследуемые методы экспериментально сравниваются путем их применения к одним и тем же обучающим совокупностям объектов, представленных конечным числом действительных признаков.

Методы релевантных и опорных потенциальных функций для действительных признаков

Прежде всего заметим, что в случае действительного признака значение $\sqrt{K_i(x_i, x_i)} = |x_i|$ выражает его норму, а $K_i(x_i, x_i) = x_i^2$ — квадрат нормы.

*Работа выполнена при финансовой поддержке РФФИ, проекты №05-01-00679, №06-01-08042, №08-01-00695, №09-01-00573.

С учетом этой подстановки *метод релевантных потенциальных функций* (RKM) [1] выражается в виде критерия обучения по предъявленной обучающей совокупности $\{\boldsymbol{x}_j, y_j\}_{j=1}^N$, где $\boldsymbol{x}_j = (x_{1j}, \ldots, x_{nj})$, $y_j \in \{-1, +1\}$, представляющего собой задачу квадратичного программирования с $2n + N + 1$ действительными переменными:

$$\begin{cases} \sum\limits_{i=1}^{n} \left[\frac{1}{r_i}\left(a_i^2 + \frac{1}{\mu}\right) + \left(\frac{1}{\mu} + 1 + \mu\right)\ln r_i\right] + C\sum\limits_{j=1}^{N}\delta_j \to \\ \quad \to \min(a_1, \ldots, a_n, b, r_1, \ldots, r_n, \delta_1, \ldots, \delta_N), \\ y_j\Big(\sum\limits_{i=1}^{n} a_i x_{ij} + b\Big) \geqslant 1 - \delta_j, \quad \delta_j \geqslant 0, \\ \qquad\qquad j = 1, \ldots, N, \quad r_i \geqslant \varepsilon, \end{cases} \tag{1}$$

где $\varepsilon > 0$ — малая неотрицательная величина, а $C > 0$ — достаточно большой коэффициент.

Алгоритм решения этой задачи изложен в [1]. Значения $(a_1, \ldots, a_n, b)$, полученные в результате решения вместе с неотрицательными весами признаков $(r_1, \ldots, r_n)$, определяют оптимальную разделяющую гиперплоскость $\sum\limits_{i=1}^{n} a_i x_i + b \gtrless 0$ в $\mathbb{R}^n$. Её эквивалентное представление через векторы признаков обучающих объектов

$$\sum_{j:\ \lambda_j > 0} y_j \lambda_j \sum_{i=1}^{n} r_i x_{ij} x_i + b \gtrless 0, \tag{2}$$

где $\lambda_j \geqslant 0$ — множители Лагранжа при левых ограничениях-неравенствах в (1), непосредственно показывает, что чем больше вес i-го признака r_i, тем больше он участвует в принятии решения о классе нового объекта $(x_1, \ldots, x_n)$, что выражает признание его более релевантным по отношению к обучающей совокупности относительно остальных признаков.

При нулевом значении параметра $0 \leqslant \mu < \infty$ в критерии (1), называемого параметром селективности, веса признаков одинаковы $(r_1 = \ldots = r_n)$. Чем больше μ, тем более неравномерны значения весов $(r_1, \ldots, r_n)$, и тем меньше число существенно релевантных признаков $r_i > \varepsilon$ в (2).

Метод опорных потенциальных функций SKM в случае действительных признаков заключается в поиске точки минимума критерия в пространстве $n + N + 1$ действительных переменных [2]:

$$\begin{cases} \sum\limits_{i=1}^{n} q(a_i \,|\, \mu) + c\sum\limits_{j=1}^{N}\delta_j \to \min(a_1, \ldots, a_n, b, \delta_1, \ldots, \delta_N), \\ y_j\Big(\sum\limits_{i=1}^{n} a_i x_{ij} + b\Big) \geqslant 1 - \delta_j,\ \delta_j \geqslant 0,\ j = 1, \ldots, N, \end{cases} \tag{3}$$

где целевая функция, зависящая от параметра селективности $0 \leqslant \mu < \infty$, имеет вид

$$q(a_i \,|\, \mu) = \begin{cases} 2\mu|a_i|, & \text{если } |a_i| \leqslant \mu, \\ \mu^2 + a_i^2, & \text{если } |a_i| > \mu. \end{cases} \tag{4}$$

В [2] показано, что множители Лагранжа $\lambda_j \geqslant 0$, вычисляемые при левых ограничениях-неравенствах критерия (3) в точке его минимума, выделяют в множестве признаков $I = \{1, \ldots, n\}$ подмножество так называемых опорных признаков

$$I_{\text{supp}} = \Big\{i\colon \sum_{j=1}^{N}\sum_{l=1}^{N} y_j y_l x_{ij} x_{il} \lambda_j \lambda_l \geqslant \mu^2\Big\} \subseteq I, \tag{5}$$

таких, что только эти признаки участвуют в формировании оптимальной гиперплоскости, разделяющей пространство признаков $(x_1, \ldots, x_n) \in \mathbb{R}^n$ на области принятия решения в пользу одного либо другого класса:

$$\sum_{j:\ \lambda_j > 0} y_j \lambda_j \sum_{i \in I_{\text{supp}}} \eta_i x_{ij} x_i + b \gtrless 0. \tag{6}$$

Веса $0 \leqslant \eta_i \leqslant 1$ опорных признаков также вычисляются в процессе обучения.

При нулевой селективности $\mu = 0$, все признаки являются опорными $I_{\text{supp}} = I$, а при увеличении μ число опорных признаков уменьшается.

Структура модельных данных

Основной целью экспериментов является сравнение методов релевантных и опорных потенциальных функций по их способности сокращать признаковое описание объектов распознавания и, в конечном итоге, обеспечивать обобщающую способность обучения по малой выборке при большом исходном числе признаков. Преимущество модельных экспериментов заключается в том, что они дают возможность придать условным понятиям «полезных» и «лишних» признаков абсолютный смысл и позволяют для каждой конкретной гиперплоскости $a_i x_i + b \gtrless 0$ (2) или (6) непосредственно вычислить вероятность ошибки распознавания на генеральной совокупности.

В данной работе модельные эксперименты организованы в пространстве ста действительных признаков $n = 100$. Случайная обучающая совокупность $\{\boldsymbol{x}_j, y_j\}_{j=1}^N$ в каждом эксперименте генерировалась как две выборки независимых случайных векторов, распределенных для каждого из классов $y_j = \pm 1$ по разные стороны известной гиперплоскости $\sum\limits_{i=1}^{n} a_i x_i + b \gtrless 0$ с нулевым смещением $b = 0$ и направляющим вектором

$$\boldsymbol{a} = (a_1, \ldots, a_{100}) = (5, 4, 3, 2, 1, 0, \ldots, 0) \in \mathbb{R}^{100}, \tag{7}$$

фиксирующим 5 полезных признаков $\{1, 2, 3, 4, 5\}$ с убывающей степенью вклада в различение классов и 95 заведомо лишних $\{6, \ldots, 100\}$.

Случайные векторы признаков объектов двух классов в каждой из обучающих совокупностей генерировались в соответствии с равномерным распределением вероятностей в двух гиперкубах с длиной ребра, равной 1, и одной общей гранью, совпадающей с гиперплоскостью.

Каждая из обучающих совокупностей содержала $N = 100$ точек по 50 в каждом классе. Такой размер обучающей совокупности вполне достаточен для оценивания разделяющей гиперплоскости в пространстве первых пяти полезных признаков, но слишком мал для попытки ее оценивания по всем ста признакам.

Контрольная совокупность была образована $N_{\text{test}} = 100000$ точками по 50000 точек каждого класса, что, как показали предварительные эксперименты, оказалось достаточным для использования ее как модели генеральной совокупности.

Результаты сравнения методов обучения на модельных данных

Массив модельных данных включал 100 обучающих совокупностей, содержащих по $N = 100 = 50 + 50$ точек, и одну контрольную совокупность размера $N_{\text{test}} = 100000 = 50000 + 50000$, полученных независимо друг от друга в пространстве $n = 100$ признаков случайным генератором, описанным в предыдущем разделе, для одной и той же разделяющей гиперплоскости (7).

Каждая из 100 обучающих совокупностей подвергалась обработке двумя алгоритмами, реализующими методы релевантных (RKM) и опорных потенциальных функций (SKM) согласно критериям (1) и (3), всякий раз многократно с разными значениями параметра селективности, изменявшегося с логарифмическим шагом в диапазоне $0 \leqslant \mu \leqslant 10^6$.

В каждом эксперименте были проведены следующие действия:

1. Соответствующий алгоритм обучения, RKM (1) либо SKM (3), однократно применялся к обучающей совокупности в целом с принятым значением селективности μ. Для полученной оценки разделяющей гиперплоскости вычислялась доля неправильно классифицированных точек контрольной совокупности $p^{\text{RKM}}_{\text{test}}(\mu)$ либо $p^{\text{SKM}}_{\text{test}}(\mu)$ как оценка вероятности ошибки распознавания на генеральной совокупности. Для алгоритма RKM запоминалась совокупность весов признаков $\big(r_1(\mu), \ldots, r_{100}(\mu)\big)$ в полученной разделяющей гиперплоскости (2), а для алгоритма SKM запоминалось подмножество опорных признаков I_{supp} (5), активно участвующих в разделяющей гиперплоскости, и их веса η_i (6).

2. Алгоритм многократно применялся к обучающей совокупности в режиме скользящего контроля (Leave-one-out Cross Validation), и запоминалась доля контрольных объектов, класс которых распознавался ошибочно $p^{\text{RKM}}_{\text{CV}}(\mu)$ либо $p^{\text{SKM}}_{\text{CV}}(\mu)$.

3. К обучающей совокупности применялся классический алгоритм опорных векторов в предположении, что известно подмножество полезных признаков, соответствующих пяти ненулевым компонентам направляющего вектора $(a_1, \ldots, a_5)$ (7).

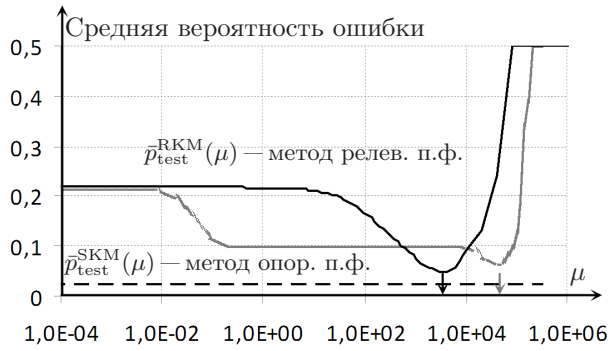


Рис. 1. Зависимость средней ошибки классификации тестовой совокупности от параметра селективности, при усреднении по 100 обучающим совокупностям.

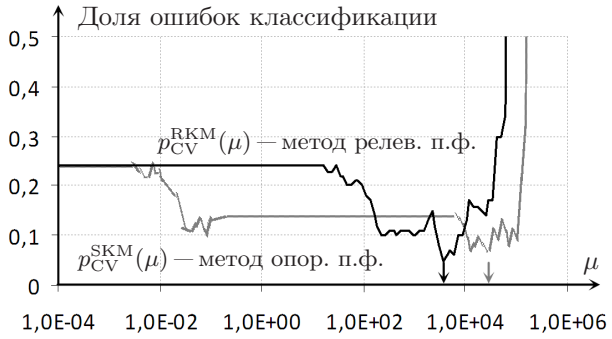


Рис. 2. Пример скользящего контроля на одной модельной обучающей совокупности для возрастающих значений параметра селективности.

На рис. 1 показаны зависимости средней ошибки распознавания классов объектов в тестовой совокупности по 100 обучающим совокупностям $\bar{p}^{\text{RKM}}_{\text{test}}(\mu)$ и $\bar{p}^{\text{SKM}}_{\text{test}}(\mu)$ для возрастающих значений селективности μ при обучении по методам релевантных и опорных потенциальных функций.

Нижняя пунктирная линия показывает среднюю ошибку метода опорных векторов при известном подмножестве информативных признаков $\{1, 2, 3, 4, 5\}$, равную 0,0239.

При нулевом значении селективности сравниваемые методы эквивалентны и дают одну и ту же среднюю ошибку $\bar{p}^{\text{SKM}}_{\text{test}}(0) = \bar{p}^{\text{SKM}}_{\text{test}}(0) = 0,2174$, равную ошибке обычного метода опорных векторов при обучении по выборке 100 объектов, явно недостаточной для пространства всех признаков размерности $n = 100$. При увеличении селективности ошибка уменьшается и достигает минимума, поскольку алгоритмы подавляют влияние лишних признаков, компенсируя эффект переобучения. Затем ошибка снова возрастает, т. к. алгоритмы начинают устранять полезные признаки. Наконец, при слишком большом значении селективности оба алгоритма подавляют все признаки, и средняя ошибка увеличивается до «безразличного» значения 0,5.

Однако минимальные значения вероятности ошибки на тестовой совокупности, достигаемые

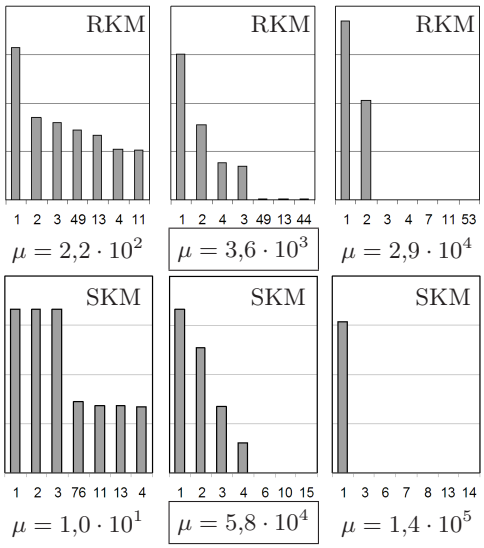


Рис. 3. Пример выделения полезных признаков методами RKM и SKM. Рамки показывают оптимальные значения селективности по скользящему контролю.

при некоторых оптимальных значениях селективности, существенно различаются. Для RKM она равна 0,0462 и примерно вдвое превышает ошибку 0,0239 при известном подмножестве полезных признаков, в то время как ошибка 0,0668 для SKM заметно больше. Меньшая ошибка, достигаемая первым методом, по-видимому, объясняется эффектом тонкой балансировки весов признаков для обрабатываемой обучающей совокупности по сравнению с жестким выбором их подмножества.

Сравнивать значения селективности, обеспечивающие минимальное значение средней ошибки, нет смысла, поскольку механизм влияния этого параметра различен в критериях (1) и (3).

На рис. 2 приведены примеры зависимости ошибки скользящего контроля от значения селективности, полученные на одной из обучающих совокупностей для каждого из двух методов. Сравнение точек минимума этих зависимостей с рис. 1 показывает, что оценки оптимальных значений селективности по одной реализации в обоих случаях обеспечивают вполне приемлемую ошибку на генеральной совокупности.

Для этой же обучающей совокупности на рис. 3 показаны относительные значения 7 наибольших весов признаков в порядке убывания, полученные в результате обучения по методам RKM и SKM, для трех значений параметра селективности. Заметим, что при оптимальном выборе μ по скользящему контролю оба метода правильно выделяют 4 главных признака из 5 информативных (7).

Эксперименты на реальных данных

Для эксперимента использовались данные по диагностике рака легких из репозитория UCI [3]. Массив образован векторами числовых признаков

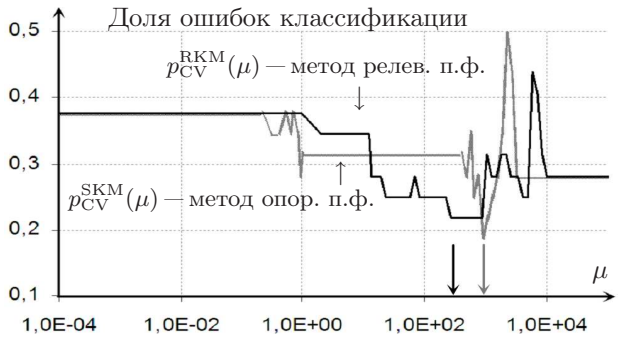


Рис. 4. Скользящий контроль в задаче распознавания рака легких методами RKM и SKM для возрастающих значений параметра селективности μ.

$N = 32$ пациентов, разбитых на два подмножества $N_{+1} = 9$ и $N_{-1} = 23$, у которых, соответственно, диагностирован и не диагностирован рак легких. Каждый вектор состоит из $n = 56$ признаков, число которых существенно превышает объем обучающей совокупности. Для сокращения признакового описания пациентов мы использовали методы релевантных и опорных потенциальных функций.

Поскольку массив данных не содержит тестовой совокупности, оценка зависимости обобщающей способности обучения от параметра селективности μ проводилась по скользящему контролю. Результаты экспериментов представлены на рис. 4.

При малых значениях μ оба метода эквивалентны обычному методу опорных векторов, учитывающему все признаки, что дает ошибку 0,38. При оптимальных μ оба метода уменьшают ошибку примерно вдвое, однако RKM выделяет 4 признака из 56, а SKM оставляет лишь 2, что, по-видимому, является причиной некоторого преимущества RKM с ошибкой 0,187 против 0,219. Наконец, слишком большие значения μ подавляют все признаки, что эквивалентно принятию решения в пользу класса, более представленного в обучающей совокупности, и ошибка стабилизируется на уровне $9/32 = 0,281$.

Литература

[1] *Татарчук А. И., Сулимова В. В., Моттль В. В., Уиндридж Д.* Метод релевантных потенциальных функций для селективного комбинирования разнородной информации при обучении распознаванию образов на основе байесовского подхода // Всероссийская конференция ММРО-14. — М.: МАКС Пресс, 2009. — С. 188–191.

[2] *Татарчук А. И., Урлов Е. Н., Моттль В. В.* Метод опорных потенциальных функций в задаче селективного комбинирования разнородной информации при обучении распознавания образов // Всероссийская конференция ММРО-14. — М.: МАКС Пресс, 2009. — С. 192–195.

[3] `http://archive.ics.uci.edu/ml/datasets/Lung+Cancer` — UCI Machine Learning Repository: Lung Cancer Data Set.

Об одной модели модифицированных алгоритмов распознавания типа потенциальных функций

Фазылов Ш. Х., Мирзаев Н. М., Мирзаев О. Н.
omirzaev@mail.ru
Ташкент, Институт математики и информационных технологий АН РУз

В работе рассмотрены вопросы построения алгоритмов распознавания образов, заданных в пространстве взаимосвязанных признаков. Предложена модель модифицированных алгоритмов распознавания, основанных на методе потенциальных функций. При их построении используется подход, основанный на оценке взаимосвязанности признаков. Рассмотренные алгоритмы используется в тех случаях, когда между признаками объектов распознавания существует некоторая зависимость. При слабом выражении зависимости между признаками используется традиционная модель алгоритмов распознавания.

К настоящему времени построен и изучен ряд моделей, из которых можно выделить достаточно известные алгоритмы распознавания образов [1–8]. К ним относятся следующие модели, основанные на использовании: принципа разделения, теории статистических решений, аппарата математической логики, теории формальных языков, принципа потенциалов, алгоритмов вычислении оценок.

Анализ этих моделей показывает, что они, в основном, ориентированы на распознавание образов, описанных в пространстве независимых (или слабозависимых) признаков [9]. Известно, что на практике часто встречаются задачи распознавания образов, заданных в пространстве признаков большой размерности (например, задачи компьютерного зрения, задачи распознавание личности по голосу и др.). В этих условиях большинство признаков взаимосвязаны, что затрудняет использование многих известных алгоритмов распознавания. Поэтому вопросы построения алгоритмов распознавания, основанных на оценке взаимосвязанности признаков, являются актуальными.

Целью данной работы является разработка модели модифицированных алгоритмов распознавания, основанных на принципе потенциалов, с использованием подхода, базирующегося на оценке взаимосвязанности признаков.

Следует отметить, что в прикладных задачах распознавания, где объекты заданы в пространстве признаков небольшой размерности, предложенный алгоритм работает неэффективно. Это связанно с тем, что в подобной ситуации большинство признаков связанно достаточно слабо. Однако, в случае взаимосвязанности признаков данный алгоритм работает эффективнее, чем исходный.

Постановка задачи

Для простоты, но без ограничения общности, рассмотрим задачу распознавания образов в случае двух классов. Пусть дано множество допустимых объектов $\{S\}$, которые заданы в пространстве признаков $X = (x_1, \ldots, x_n)$. В пространстве X каждому объекту S из $\{S\}$ соответствует описание объекта $I(S) = (a_1, \ldots, a_n)$. При этом предполагается, что множество $\{S\}$ состоит из непересекающихся подмножеств (классов) K_1 и K_2 ($K_1 \cap K_2 = \varnothing$): $\{S\} = \bigcup_{i=1}^{2} K_i$. Разбиение $\{S\}$ определено не полностью, а имеется только некоторая начальная информация I_0 о классах [1].

Рассмотрим множество m объектов $S_1, \ldots, S_m$ из $\{S\}$ в пространстве исходных признаков X:

$$S_1 = (a_{11}, \ldots, a_{1n}),$$
$$\ldots$$
$$S_m = (a_{m1}, \ldots, a_{mn}).$$

Введем следующие обозначения: $\tilde{S}^m = \{S_1, \ldots, S_m\}$, $\tilde{K}_j = \tilde{S}^m \cap K_j$, $C\tilde{K}_j = \tilde{S}^m \backslash K_j$. Тогда начальную информацию I_0 можно задать в виде

$$I_0 = \{S_1, \ldots, S_m; \lambda(S_1), \ldots, \lambda(S_m)\},$$

где $\lambda(S_i)$ — характеристическая функция объекта S_i, которая имеет вид:

$$\lambda(S_i) = \begin{cases} -1, & \text{если } S_i \in K_1; \\ 1, & \text{если } S_i \in K_2. \end{cases}$$

Дан набор объектов $\tilde{S}^q = \{S'_1, \ldots, S'_q\}$ из $\{S\}$, которые заданы в пространстве признаков X высокой размерности $X = (x_1, \ldots, x_n)$, $n > 200$. При этом предполагается, что многие признаки взаимосвязаны. В этих условиях требуется построить такой алгоритм распознавания A, который вычисляет значения функции $\lambda(S_i)$, $i = 1, \ldots, q$ по начальной информации I_0:

$$A(I_0, \tilde{S}^q) = (\hbar_1, \ldots, \hbar_q)^{\intercal}, \quad \hbar_i \in \{-1, 0, 1\}$$

Здесь $\hbar_i \in \{-1, 0, 1\}$ есть значение характеристической функции $\lambda(S'_i)$ на допустимом объекте S'_i, вычисленное алгоритмом A. Если $\hbar_i = -1$, то объект S'_i входит в класс K_1, если $\hbar_i = 1$, то объект S'_i входит в класс K_2, если $\hbar_i = 0$, то считается, что алгоритм A не вычислил значение характеристической функции $\lambda(S'_i)$.

Метод решения

В данной работе для решения задачи распознавания образов в пространстве признаков

большой размерности предлагается подход, который является логическим продолжением работ Ю. И. Журавлёва и его учеников. На базе этого подхода разработана модель модифицированных алгоритмов распознавания, основанных на выявлении независимых подмножеств взаимосвязанных признаков и выделении предпочтительной модели зависимости для каждого подмножества сильносвязанных признаков. В качестве исходной модели рассматривается модель алгоритмов типа потенциальных функций. Рассмотрим основные этапы построения алгоритмов распознавания типа потенциальных функций, основанных на оценке взаимосвязанности признаков.

1. Выделение подмножеств сильносвязанных признаков. На этом этапе определяется система «независимых» подмножеств признаков, состав которой будет зависеть от параметра n'. Задавая различные целочисленные значения этого параметра, получим различные алгоритмы. Значение параметра n' определяется на основе анализа исходных данных и, в некоторых случаях, может задаваться априорно.

Подмножества сильносвязанных признаков выделяется следующим образом. Рассматриваемая совокупность признаков объединяется в одно подмножество, если они достаточно схожи друг с другом. В противном случае они считаются различными, и их относят к разным подмножествам. При выделении сильносвязанных признаков рассматриваются объекты, принадлежащие только одному классу K_j. Если в результате выполнения данного этапа формируются одинаковые подмножества для K_1 и K_2, то это указывает на отсутствие внутренней взаимозависимости между признаками, присущей каждому классу. Очевидно, что в подобной ситуации применение предложенной модели алгоритмов распознавания нецелесообразно [10].

В зависимости от способа задания меры близости между подмножествами сильносвязанных признаков (Ω_p и Ω_q) и функционала качества разделения можно получить разнообразные алгоритмы выделения независимых подмножеств сильносвязанных признаков [11, 12].

2. Формирование набора репрезентативных признаков. Основная идея выбора репрезентативных признаков заключается в их различии (несходстве) в формируемом наборе репрезентативных признаков. В процессе формирования набора репрезентативных признаков требуется, чтобы каждый выделенный признак был типичным представителем выделенного подмножества сильносвязанных признаков. В результате выполнения данного этапа получаем сокращенное пространство признаков, размерность которого намного меньше исходного ($n' < n$). Сформированное пространство признаков обозначим через $Y = (y_1, \ldots, y_{n'})$.

Процедура выделения набора репрезентативных признаков более подробно рассмотрена в [13].

3. Определение моделей зависимости в каждом подмножестве признаков для класса K_j, $j = 1, 2$. На данном этапе предполагается, что на основе анализа исходных признаков объектов, которые принадлежат классу K_j, выделены подмножества сильносвязанных признаков Ω_q, $q \in [1, n']$, и определены соответствующие репрезентативные признаки y_q.

Пусть x_i — произвольный признак, принадлежащий подмножеству Ω_q. Предполагается, что элементы Ω_q линейно упорядочены по индексу признаков (т. е. $x_i < x_j$, если $i < j$). Далее, нулевым элементом (x_0) подмножества Ω_q считается y_q, остальные элементы обозначаются через x_i, $i = 1, \ldots, N_q - 1$; $N_q = |\Omega_q|$. Тогда модель зависимости в Ω_q принимает вид

$$x_i = F(\bar{c}, y_q), \quad x_i \in \Omega_q \setminus \{y_q\},$$

где $\bar{c}$ — вектор неизвестных параметров, F — функция из некоторого заданного класса $\{F\}$.

Вычисленные значения вектора неизвестных параметров $\bar{c}$ определяют модель зависимости в подмножестве признаков Ω_q для класса K_j, $j = 1, 2$. В зависимости от задания параметрического вида $F(\bar{c}, x)$ и метода определения $\bar{c}$ получаем разнообразные модели зависимости в множестве признаков Ω_q, $q = 1, \ldots, n'$.

Пример. В качестве заданного множества $\{F\}$ рассмотрим линейные модели. При этом предполагаем, что признак y_q из Ω_q является независимой переменной, а признак x_i из $\Omega_q \setminus \{y_q\}$ — зависимой переменной. Тогда модель зависимости в Ω_q принимает вид

$$x_i = c_{i_1} y_q + c_{i_0},$$

где c_{i_1}, c_{i_0} — параметры, которые определяются на основе критерия наименьших квадратов [14].

4. Выделение предпочтительных моделей зависимости. Пусть N_q — мощность подмножества сильносвязанных признаков Ω_q. Предполагается, что в Ω_q определено $(N_q - 1)$ моделей зависимости для объектов класса K_1:

$$x_i = F(\bar{c}, y_q), \quad x_i \in \Omega_q \setminus \{y_q\}, \quad i = 1, \ldots, N_q - 1,$$

где y_q — репрезентативный признак из Ω_q.

Поиск предпочтительной модели зависимости в Ω_q осуществляется на основе оценки доминированности рассматриваемых моделей для объектов, которые относятся к множеству I_0:

$$T_i = \frac{L_1 \sum\limits_{S \in \tilde{K}_2} \big(x_i - F(\bar{c}, y_q)\big)^2}{L_2 \sum\limits_{S \in \tilde{K}_1} \big(x_i - F(\bar{c}, y_q)\big)^2},$$

где $L_1 = |\tilde{K}_1|$, $L_2 = |\tilde{K}_2|$. Чем больше величина T_i, тем больше отдаётся предпочтение i-й модели зависимости. Если несколько моделей получают одинаковое предпочтение, то выбирается любая из них.

В результате выполнения данного этапа определяется предпочтительная модель зависимости для подмножества признаков Ω_q, которая обозначается через $x_{i_0} = F(\bar{c}, y_q)$. Далее рассматриваются только эти (т. е. предпочтительные) модели зависимости.

5. Определение функции близости между объектами S_u и S. Рассмотрим два объекта, заданных в пространстве признаков $X = (x_1, \ldots, x_n)$:

$$S_u = (a_{u1}, \ldots, a_{un}) \text{ и } S = (b_1, \ldots, b_n).$$

Пусть расстояние между объектами S_u и S по подмножеству Ω_q, $q = 1, \ldots, n'$ задано в виде:

$$d_q(S_u, S) = \tau_q\Big(\big|a_{ui_0} - F(\bar{c}, a_{ui_q})\big| + \big|b_{i_0} - F(\bar{c}, b_{i_q})\big|\Big).$$

где τ_q — параметр алгоритма; a_{ui_0}, b_{i_0} — значения признака x_{i_0}, соответственно, объектов S_u и S; a_{ui_q}, b_{i_q} — значения репрезентативного признака y_q, соответственно, объектов S_u и S.

Тогда функция близости между объектами S_u и S принимает вид

$$U(S_u, S) = \frac{1}{1 + \alpha \sum\limits_{q=1}^{n'} d_q(S_u, S)},$$

где α — параметр алгоритма; n' — число подмножеств взаимосвязанных признаков.

Следует отметить, что функция $U(S_u, S)$ характеризует отклонение объектов S_u и S от моделей зависимостей, присущих рассматриваемому классу K_j.

6. Вычисление оценки принадлежности к классу K_j. Пусть вычислены оценки близости между объектами S_u и S, $u = 1, \ldots, m$:

$$U(S_1, S), \ldots, U(S_u, S), \ldots, U(S_m, S).$$

Тогда оценка принадлежности объекта S к классу K_j $(j = 1, 2)$ определяется как функция от оценки близости между объектами [15], например:

$$\mu_j(S) = \sum_{S_u \in \tilde{K}_j} \gamma_u U(S_u, S) \; + \sum_{S_u \in C\tilde{K}_j} \gamma_u U(S_u, S),$$

где γ_u — параметр алгоритма.

7. Решающее правило. Последним этапом задания модели алгоритмов является задание решающего правила в виде [1] :

$$\beta = C\big(\mu_j(S)\big) = \begin{cases} -1, & \text{если } \mu_j(S) < -c_1; \\ 0, & \text{если } -c_1 \leqslant \mu_j(S) \leqslant c_2; \\ 1, & \text{если } c_2 < \mu_j(S). \end{cases}$$

где c_1, c_2 — параметры алгоритма $(c_1, c_2 \geqslant 0)$.

Таким образом, определена модель распознающих алгоритмов типа потенциальных функций, основанных на оценке взаимосвязанности признаков. Произвольный алгоритм A из этой модели полностью определяется заданием набора параметров $\pi = \big(n', \{\bar{c}\}, \{\tau_q\}, \alpha, \{\gamma_u\}, c_1, c_2\big)$. Совокупность всех распознающих алгоритмов из предлагаемой модели обозначаем через $A(\pi, S)$. Определение наилучшего алгоритма в рамках рассмотренной модели осуществляется в пространстве параметров π [16].

Рассмотренные алгоритмы отличаются от традиционных алгоритмов распознавания, основанных на принципе потенциалов, тем, что они основаны на оценке взаимосвязанности признаков. Поэтому эти алгоритмы используются в том случае, когда между признаками обнаруживается какая-нибудь зависимость. Очевидно, что эта зависимость должна отличаться в каждом классе. Это позволяет описать объекты каждого класса индивидуальной моделью.

Если зависимость между признаками слаба, то используется классическая модель алгоритмов распознавания (например, модель, рассмотренная в работе [5]). Следовательно, предложенная модель алгоритмов распознавания не является альтернативной моделям, основанным на принципе потенциалов, а только дополняет их.

В случае, когда между признаками всех рассматриваемых объектов обнаруживается достаточно сильная зависимость, то в процессе формирования набора репрезентативных признаков (описанных на первом и втором этапах) исключаются признаки, повторяющие одну и ту же информацию, и обеспечивают выбор признаков, достаточно хорошо представляющих все те признаки, которые не содержатся в данном наборе [10].

Экспериментальная проверка

В целях практического использования и проверки работоспособности рассмотренной модели алгоритмов разработаны функциональные схемы программ распознавания. Программная реализация разработанных алгоритмов осуществлена на языке Object Pascal в среде Delphi. Работа разработаных программ апробирована на модельном примере.

Исходные данные распознаваемых объектов для модельного примера сгенерированы в пространстве зависимых признаков. Количество классов в данном эксперименте равно двум. Объем обучающей выборки — 100 реализаций (по 50 реализаций для объектов каждого класса). Объём контрольной выборки — 100 реализаций (по 50 реализаций для объектов каждого класса). Количество признаков в модельном примере равно 20. Число подмножеств сильносвязанных признаков — 5. Вид распределения — нормальный.

Проведённые экспериментальные исследования показали высокую точность разработанной модели алгоритмов при решении данного модельного примера. В результате эксперимента выявлены все зависимые признаки и на их основе построен эффективный алгоритм. Анализ результатов решения модельного примера с помощью предлагаемых алгоритмов показывает преимущество этих алгоритмов в быстродействии и точности распознавания в случаях описания объектов в пространстве взаимосвязанных признаков.

Выводы

В настоящее время решение задач распознавания образов, заданных в пространствах признаков большой размерности, связанно со значительными вычислительными трудностями. Наиболее подходящим для сокращения вычислительных операций является выделение репрезентативных признаков в условиях взаимосвязанности признаков в рамках любой выборки из генеральной совокупности. Однако, при отличии зависимости между признаками объектов в каждом классе выделение репрезентативных признаков не может обеспечить существенного сокращения вычислительных операций. В этом случае алгоритмы распознавания необходимо модифицировать с учётом взаимосвязанности признаков.

Разработана модель алгоритмов распознавания, основанных на оценке взаимосвязанности признаков, в рамках модели алгоритмов типа потенциальных функций. Предложенная схема задания модели алгоритмов распознавания является оригинальной. Данная модель алгоритмов распознавания позволяет расширить область применения метода потенциальных функций в условиях взаимосвязанности признаков.

Результаты приведённого экспериментального исследование показали, что рассмотренная модель алгоритмов улучшает точность и значительно сокращает число вычислительных операции при решении задачи распознавания образов, заданных в пространстве взаимосвязанных признаков.

Предложенная модель алгоритмов распознавания может быть использована при составлении различных программных комплексов, ориентированных на решение задач диагностики и классификации объектов в условиях взаимосвязанности признаков.

Литература

[1] *Журавлёв Ю. И.* Избранные научные труды. — М.: Магистр, 1998. — 420 с.

[2] *Журавлёв Ю. И., Рязанов В. В., Сенько О. В.* Распознавание. Математические методы. Программная система. Практические применения. — М.: Фазис, 2006. — 159 с.

[3] *Лбов Г. С., Старцева Н. Г.* Логические решающие функции и вопросы статистической устойчивости решений. — Новосибирск: ИМ СО РАН, 1999. — 211 с.

[4] *Хайкин С.* Нейронные сети: полный курс. 2-е изд., испр.: Пер. с англ. — М.: Вильямс, 2006. — 1104 с.

[5] *Айзерман М. А., Браверманн Э. М., Розоноэр Л. И.* Метод потенциальных функций в теории обучения машин. — М.: Наука, 1970. — 348 с.

[6] *Duda R. O., Hart P. E., Stork D. G.* Pattern Classification, Second edit. New York: Wiley, 2001. — 680 p.

[7] *Vapnik V. N.* Statistical Learning Theory. New York: Wiley, 1998. — 732 p.

[8] *Шлезингер М., Главач В.* Десять лекций по статистическому и структурному распознаванию. — Киев: Наукова думка, 2004.— 535 с.

[9] *Камилов М. М., Мирзаев Н. М., Раджабов С. С.* Современное состояние вопросов построения моделей алгоритмов распознавания // Химическая технология. Контроль и управление. — 2009. — № 2. — С. 67–73.

[10] *Мирзаев О. Н.* Модели алгоритмов распознавания, основанные на принципе потенциалов // Вопросы вычислительной и прикладной математики, Ташкент: ИМИТ АН РУз, 2009. — Вып. 121. — С. 126–135.

[11] *Камилов М. М., Фазылов Ш. Х., Мирзаев Н. М.* Алгоритмы распознавания, основанные на оценке взаимосвязанности признаков // Всеросс. конф. ММРО-13, М.: МАКС Пресс, 2007. — С. 140–143.

[12] *Мирзаев Н. М.* Алгоритмы выделения подмножеств сильносвязанных признаков // Вопросы кибернетики. — 2008. — Вып. 177. — С. 99–104.

[13] *Мирзаев О. Н.* Выделение репрезентативных признаков при построении алгоритмов распознавания // Проблемы информатики и энергетики. — 2008. — № 6. — С. 23–26.

[14] *Дрейпер Н., Смит Г.* Прикладной регрессионный анализ. Множественная регрессия. 3-е издание. — М.: Диалектика, 2007. — 912 с.

[15] *Фазылов Ш. Х., Мирзаев О. Н.* Алгоритмы распознавания типа потенциальных функций, основанные на взаимосвязанности признаков // Современные проблемы математики, механики и информационных технологий. — 2008. — С. 275–277.

[16] *Мирзаев Н. М., Раджабов С. С., Жумаев Т. С.* О параметризации моделей алгоритмов распознавания, основанных на оценке взаимосвязанности признаков // Проблемы информатики и энергетики. — 2008. — № 2–3. — С. 23–27.

О некоторых аспектах интеллектуального анализа пучков временных рядов

Филипенков Н. В.
filipenkov@mail.ru
Москва, Вычислительный центр РАН

В настоящей работе рассматриваются аспекты интеллектуального анализа пучков временных рядов, базирующиеся на предположении о плавном изменении закономерностей во времени. В работе алгоритм поиска плавно меняющихся закономерностей обобщен для случая различных весов функционала качества. При этом функционал качества стационарной закономерности интегрирован в общую процедуру поиска плавно меняющихся закономерностей. Рассматриваются результаты экспериментов при различном уровне шума и параметрах алгоритма.

В настоящее время интеллектуальный анализ многомерных временных рядах привлекает большое количество исследователей. При этом особенный интерес вызывает поиск изменяющихся закономерностей в многомерных временных рядах, а также вопросы близости найденных закономерностей [1–3]. Настоящая работа продолжает данное направление исследований.

Введение

В работах [4], [5] был предложен метод поиска плавно меняющихся закономерностей в пучках временных рядов. Пучком временных рядов называется многомерный временной ряд, в котором предполагается существование взаимосвязей между различными одномерными рядами. Метод базируется на предположении о том, что закономерности, определяющие поведение пучка временных рядов, могут плавно изменяться во времени.

Основная идея предложенного метода состоит в разделении процесса поиска закономерностей на следующие этапы. На первом этапе исходный пучок временных рядов разбивается на отрезки по времени, и на каждом из отрезков происходит поиск стационарных закономерностей. Затем различные закономерности соседних отрезков «склеиваются» в одну плавно меняющуюся закономерность. При этом «склеивание» происходит по принципу близости закономерностей друг к другу.

В статье [4] близость закономерностей определяется исключительно на основе предложенной меры сходства закономерностей. В работе [5] рассматриваются функционалы качества изменяющихся закономерностей. В настоящей работе представлен подход к поиску плавно меняющихся закономерностей, где близость закономерностей определяется не только на основе меры сходства, но и с использованием функционалов качества стационарных закономерностей. Такой подход позволяет добиться лучшего качества на практике при поиске плавно меняющихся закономерностей по сравнению с предложенным в [4]. Описанный в настоящей работе подход позволяет параметризовать влияние различных функционалов на конечный вид плавно меняющейся закономерности. Вместе с тем предлагаемый подход интегрирует процесс поиска оптимальной плавно меняющейся закономерности и функционалы качества, рассмотренные в [5].

В данной работе также приводятся результаты поиска плавно меняющихся закономерностей при различных параметрах функционала, на основе которого определяется близость закономерностей.

Граф закономерностей

Пусть $\mathfrak{S}$ — пучок дискретнозначных временных рядов, то есть многомерный временной ряд, в котором предполагается наличие взаимосвязей между рядами. Пучок временных рядов представим в виде матрицы размера $N \times T$, элементами которой являются числа из $0, \ldots, k-1$, N — количество временных рядов, T — длина пучка.

Обозначим через $\mathfrak{S}^1, \ldots, \mathfrak{S}^m$ отрезки пучка временных рядов $\mathfrak{S}$ — пучки временных рядов, составленные из последовательных столбцов $\mathfrak{S}$. В работе рассматриваются отрезки, образующие разбиение оси времени. Однако это требование не является обязательным, в общем случае отрезки могут и пересекаться.

Маской ω на прямоугольнике $N \times \Delta$ назовем булеву матрицу размера $N \times \Delta$.

Закономерностью R называется набор $\langle p, \omega, f \rangle$, состоящий из следующих элементов:

1) индекс $p \in \{1, 2, \ldots, N\}$ указывает на целевой ряд (ряд, значения которого «прогнозируются» закономерностью R);
2) маска ω «выбирает» из значений всех рядов в моменты времени $t - \Delta, \ldots, t - 1$ элементы, являющиеся аргументами функции f;
3) частично определенная функция f задает зависимость значений целевого ряда от переменных, на которые указывает маска ω.

Закономерность прогнозирует значение p-го ряда в момент времени t по значениям всех рядов в моменты времени $t - \Delta, \ldots, t - 1$. Параметр Δ определяет максимальный отступ по времени.

Достоверностью $\mathrm{Conf}(R, \mathfrak{S})$ *закономерности* $R = \langle p, \omega, f \rangle$ *на пучке временных рядов* $\mathfrak{S}$ называется доля правильных прогнозов закономерности R

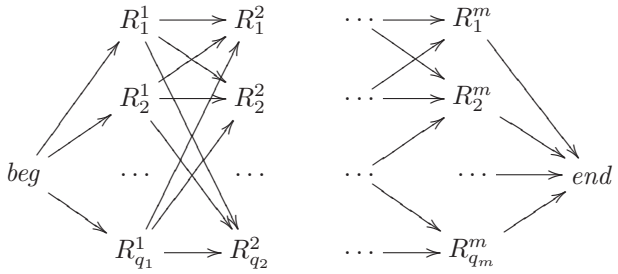


Рис. 1. Граф закономерностей.

на пучке временных рядов $\mathfrak{S}$. Из определения следует справедливость неравенства:

$$0 \leqslant \mathrm{Conf}(R, \mathfrak{S}) \leqslant 1, \tag{1}$$

Системой закономерностей называется произвольная непустая совокупность закономерностей $R_1 = \langle p_1, \omega_1, f_1\rangle, \ldots, R_n = \langle p_n, \omega_n, f_n\rangle$, у которых совпадает целевой ряд ($p_1 = \ldots = p_n$) и размерность масок ($\omega_1, \ldots, \omega_n \in \{0,1\}^{N\times\Delta}$).

В соответствии с алгоритмом, описанным в [4], на каждом из отрезков $\mathfrak{S}^1, \ldots, \mathfrak{S}^m$ происходит поиск стационарных закономерностей. Пусть закономерности $R_1^j, \ldots, R_{q_j}^j$ найдены на отрезке $\mathfrak{S}^j$, для всех $j = 1, \ldots, m$, где q_j — число закономерностей, найденных на отрезке $\mathfrak{S}^j$ алгоритмом поиска постоянных закономерностей. Все закономерности представляются на графе закономерностей, рис. 1.

В работе [4] описана процедура построения меры сходства закономерностей $\rho(R^j, R^{j+1})$. Данная мера сходства обладает следующим свойством:

$$0 \leqslant \rho(R_1, R_2) \leqslant 1, \tag{2}$$

где R_1 и R_2 — произвольные закономерности. При этом увеличение меры $\rho(R_1, R_2)$ отражает растущее различие между закономерностями.

Изменяющейся закономерностью $\tilde{R}$ для последовательности отрезков $\mathfrak{S}^1, \ldots, \mathfrak{S}^m$ на пучке временных рядов $\mathfrak{S}$ называется система закономерностей $R^1, \ldots, R^m$, где каждая закономерность взаимно однозначно соответствует некоторому отрезку $\mathfrak{S}^i$, $i = 1, \ldots, m$. Будем называть стационарные закономерности $R^1, \ldots, R^m$ *шагами*, которые *составляют* изменяющуюся закономерность $\tilde{R}$.

Длиной $l(\tilde{R})$ изменяющейся закономерности $\tilde{R}$ называется сумма мер сходства «соседних» шагов — закономерностей, составляющих меняющуюся закономерность:

$$l(\tilde{R}) = \sum_{j=1}^{m-1} \rho(R^j, R^{j+1}).$$

Пусть каждый из отрезков $\mathfrak{S}^j$, $j = 1, \ldots, m$, разбит на две части: обучение $\mathfrak{S}^j_{\mathrm{train}}$ и валидацию $\mathfrak{S}^j_{\mathrm{valid}}$.

Тогда алгоритм поиска постоянных закономерностей, примененный к каждому из отрезков, порождает наборы закономерностей $R_1^j, \ldots, R_{q_j}^j$ на каждом из отрезков $\mathfrak{S}^j$, $j = 1, \ldots, m$. Для каждой закономерности R_i^j, $i = 1, \ldots, q_j$, $j = 1, \ldots, m$, определены значения показателей качества: достоверность на обучении $\mathrm{Conf}(R_i^j, \mathfrak{S}^j_{\mathrm{train}})$ и достоверность на валидации $\mathrm{Conf}(R_i^j, \mathfrak{S}^j_{\mathrm{valid}})$.

Найденные закономерности можно представить в виде графа закономерностей (рис. 1). Вершинами графа являются стационарные закономерности, найденные на каждом из отрезков, а также две дополнительные вершины: *beg* и *end*. С каждой вершиной ассоциированы показатели качества закономерности. Дугами на графе связаны закономерности соседних отрезков, что отражает факт возможного «превращения» одной закономерности в другую. С каждой дугой ассоциирован вес — мера сходства соответствующих закономерностей. Веса дуг, соединяющие закономерности крайних отрезков с вершинами *beg* и *end*, полагаются равными нулю.

Задача выделения наилучшей изменяющейся закономерности состоит в поиске пути между вершинами *beg* и *end* на ориентированном графе, который максимизирует показатели качества закономерностей вершин, входящих в него, и минимизирует суммарный вес ребер.

Эта задача сводится к стандартной задаче поиска кратчайшего пути на графе, если использовать в качестве веса вершины величину $(1 - Q_{\mathrm{step}})$, где Q_{step} — *функционал качества шага изменяющейся закономерности* $\tilde{R}$, который задается следующим образом:

$$\begin{aligned}
Q_{\mathrm{step}}(R_i^j, R_l^{j+1}) &= w_{\mathrm{conf}}\,\mathrm{Conf}(R_i^j, \mathfrak{S}^j_{\mathrm{valid}}) + \\
&\quad + w_{\mathrm{similarity}}\big(1 - \rho(R_i^j, R_l^{j+1})\big);\\
Q_{\mathrm{step}}(beg, R_i^j) &= 0;\\
Q_{\mathrm{step}}(R_i^j, end) &= w_{\mathrm{conf}}\,\mathrm{Conf}(R_i^j, \mathfrak{S}^j_{\mathrm{valid}});\\
j = 1, \ldots, m-1,&\ \ i = 1, \ldots, q_j,\ \ l = 1, \ldots, q_{j+1};
\end{aligned}$$

где $\mathrm{Conf}(R_i^j, \mathfrak{S}^j_{\mathrm{valid}})$ — достоверность закономерности, $\rho(R_i^j, R_l^{j+1})$ — мера сходства закономерностей. Веса w_{conf}, $w_{\mathrm{similarity}}$ функционала качества шага удовлетворяют следующим условиям:

$$\begin{gathered}
0 \leqslant w_{\mathrm{conf}} \leqslant 1;\\
0 \leqslant w_{\mathrm{similarity}} \leqslant 1;\\
w_{\mathrm{conf}} + w_{\mathrm{similarity}} = 1.
\end{gathered}$$

Так как справедливы неравенства (1), (2), то для произвольных закономерностей R_1, R_2 верно неравенство

$$0 \leqslant Q_{\mathrm{step}}(R_1, R_2) \leqslant 1.$$

Таким образом, вес вершины $(1-Q_{\text{step}})$ является неотрицательным, и для решения задачи поиска кратчайшего пути на графе удобно использовать стандартные алгоритмы:

1) алгоритм Дейкстры [6] со сложностью $\underline{\underline{O}}(n^2)$, где n — число вершин графа;
2) алгоритм поиска кратчайшего расстояния в топологически отсортированном графе [6] со сложностью $\underline{\underline{O}}(n^2)$, где n — число вершин графа.

Изменяющуюся закономерность $\tilde{R}$ будем называть *плавно меняющейся*, если она составлена из закономерностей, лежащих на кратчайшем пути из вершины *beg* в вершину *end*, и выполнено неравенство $w_{\text{similarity}} > 0$ для веса меры сходства закономерностей в функционале качества шага.

Показатели качества плавно меняющихся закономерностей

Пусть $\tilde{R}_0$ — плавно меняющаяся закономерность, составленная из закономерностей $R_0^1, \ldots, R_0^m$, соответственно, на отрезках $\mathfrak{S}^1, \ldots, \mathfrak{S}^m$ пучка временных рядов $\mathfrak{S}$. Одним из основных показателей качества плавно меняющейся закономерности является ее длина $l(\tilde{R}_0)$, введённая выше. Рассмотрим обобщение для случая плавно меняющихся закономерностей понятия достоверности, введенного для постоянных закономерностей.

Достоверностью $\widetilde{\text{Conf}}(\tilde{R}_0, \mathfrak{S})$ *плавно меняющейся закономерности* $\tilde{R}_0$ на пучке временных рядов $\mathfrak{S}$ называется средневзвешеннная по длине отрезков достоверность закономерностей, составляющих плавно меняющуюся закономерность

$$\widetilde{\text{Conf}}(\tilde{R}_0, \mathfrak{S}) = \sum_{j=1}^{m} \frac{\theta_j}{T} \text{Conf}(R_0^j, \mathfrak{S}),$$

где $\theta_1, \ldots, \theta_m$ — длины отрезков $\mathfrak{S}^1, \ldots, \mathfrak{S}^m$ соответственно.

Поиск оптимальной плавно меняющейся закономерности на графе, где весом ребра является функционал качества шага изменяющейся закономерности, минимизирует следующую величину:

$$\sum_{j=1}^{m-1} \left(1 - Q_{\text{step}}(R_0^j, R_0^{j+1})\right) \to \min.$$

В свою очередь, это означает, что плавно меняющаяся закономерность доставляет максимум функционала качества:

$$\sum_{j=1}^{m-1} Q_{\text{step}}(R_0^j, R_0^{j+1}) \to \max.$$

Таким образом, плавно меняющаяся закономерность $\tilde{R}_0$ максимизирует достоверность $\widetilde{\text{Conf}}(\tilde{R}_0)$ и минимизирует длину закономерности $l(\tilde{R}_0)$.

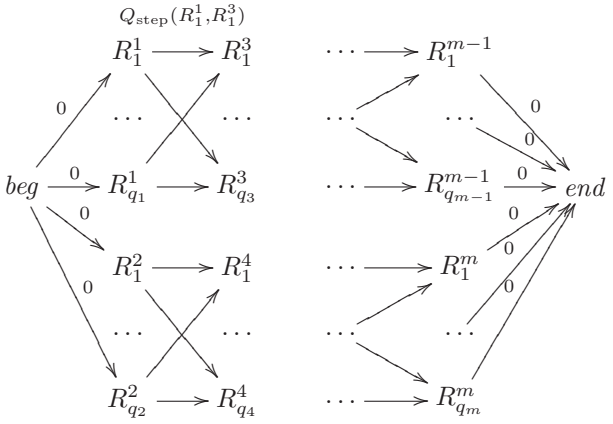


Рис. 2. Граф закономерностей для задачи поиска периодических закономерностей.

Баланс между оптимумами определяется весами w_{conf}, $w_{\text{similarity}}$ функционала качества шага изменяющейся закономерности.

Поиск периодических закономерностей

Структура графа закономерностей не обязательно должна совпадать с изображенной на рис. 1. Число ребер графа можно уменьшить, например, удалив те из них, которым приписан большой вес. И, наоборот, структуру графа можно усложнить, добавив ребра, соединяющие закономерности, полученные не на соседних отрезках. Такой подход опирается на то, что некоторые закономерности могут не проявляться в определенные моменты времени. Но, модифицируя структуру графа, крайне важно следить за делением пучка временных рядов на отрезки, так как оба этих процесса тесно взаимосвязаны и способны существенно повлиять на качество найденных закономерностей.

В зависимости от конкретных особенностей задачи можно формировать различные последовательности отрезков пучка временных рядов и составлять граф, на котором осуществляется поиск изменяющихся закономерностей. Например, во многих прикладных задачах имеет место непосредственная зависимость от времени года, месяца и т. п. В соответствии с этим естественным делением предлагается разбивать временные ряды на отрезки. Но, чтобы иметь возможность проследить периодические закономерности, необходимо модифицировать граф закономерностей, как показано на рис. 2.

Таким образом, априорную информацию о закономерностях можно использовать при выборе отрезков и при конструировании графа закономерностей.

Результаты экспериментов

Предложенный подход к поиску плавно меняющихся закономерностей был апробирован на модельных пучках временных рядов.

С целью испытания предложенного подхода для решения практических задач был подготовлен экспериментальный стенд. Стенд позволяет генерировать временные ряды и проводить поиск стационарных и меняющихся закономерностей.

Проводились две серии экспериментов на модельных рядах с целью выявить особенности для наиболее эффективного применения предложенных алгоритмов анализа временных рядов. Первая серия проводилась с целью определения эффективности алгоритма поиска постоянных закономерностей. Вторая серия экспериментов позволила оценить влияние меры сходства закономерностей при поиске плавно меняющихся закономерностей.

В первой серии модельных экспериментов было проведено исследование влияния уровня «белого шума» в моделируемых пучках временных рядов на качество распознавания. Для каждого значения уровня шума проводилась серия из 100 экспериментов. В каждом эксперименте генерировалась случайная изменяющаяся закономерность, удовлетворяющая условиям на размерность маски и индекс целевого ряда. На основе закономерности генерировался пучок временных рядов, в котором затем происходил поиск изменяющихся закономерностей.

Критерии успешного эксперимента были определены следующим образом. Генерируемая и найденная стационарные закономерности называются *совпадающими*, если полностью совпадают их маски и доля различных значений функции не превышает 5% от общего числа значений. Изменяющиеся закономерности называются совпадающими, если совпадают все их соответствующие шаги — стационарные закономерности. Эксперимент признается успешным, если генерируемая и найденная изменяющиеся закономерности совпадают.

Генерируемые пучки временных рядов состояли из 10 рядов по 1000 значений в каждом. Алфавит каждого из рядов состоял из четырех значений.

Результаты моделирования в первой серии экспериментов представлены на рис. 3. Результаты показывают, что алгоритм является достаточно стабильным и проводит вполне эффективный интеллектуальный анализ данных даже для зашумленных пучков временных рядов.

Во второй серии экспериментов исследовалось влияние меры сходства закономерностей на качество распознавания. С этой целью проводился поиск изменяющихся закономерностей для разных весов достоверности w_{conf} и меры сходства закономерностей $w_{\text{similarity}}$ функционала качества шага закономерности Q_{step}.

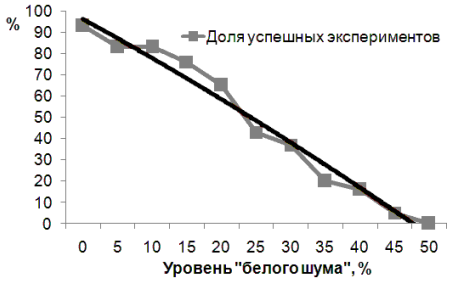


Рис. 3. Доля успешных экспериментов при различных параметрах функционала качества.

Варьируя веса закономерности, можно оценить эффект от добавления меры сходства закономерностей в процесс распознавания. Результаты второй серии экспериментов показали, что добавление меры сходства закономерностей в функционал качества позволяет повысить точность распознавания на 15–20%.

Выводы

Представленный в настоящей работе алгоритм анализа пучков временных рядов позволяет проводить поиск плавно меняющихся закономерностей. Алгоритм позволяет выбрать оптимальное соотношение между точностью найденных закономерностей для каждого отрезка пучка временных рядов и предположением о малом изменении закономерностей с течением времени.

Результаты экспериментов на модельных пучках временных рядов позволяют утверждать, что предложенный алгоритм является достаточно стабильным к зашумленным временным рядам. При этом использование меры сходства закономерностей позволяет повысить точность прогнозирования.

Литература

[1] *Morchen F., Ultsch A.* Efficient mining of understandable patterns from multivariate interval time series // Data Mining and Knowledge Discovery. — 2007. — Vol. 15, No. 2. — Pp. 181–215.

[2] *Caraca-valente J. P, Lopez-Chavarrias I.* Discovering similar patterns in time series // KDD-2000, Boston, 2000. — Pp. 497–505.

[3] *Xuan X., Murphy K.* Modeling changing dependency structure in multivariate time series // ICML-2007, Corvallis, 2007. — Pp. 1055–1062.

[4] *Филипенков Н. В.* О задачах анализа пучков временных рядов с изменяющимися закономерностями // Искусственный интеллект. — 2006. — № 2. — С. 125–129.

[5] *Филипенков Н. В.* Об оптимальном выборе закономерностей, составляющих плавно меняющуюся закономерность // Всеросс. конф. ММРО-13, М.:МАКС Пресс, 2007. — С. 223–225.

[6] *Кристофидес Н.* Теория графов. Алгоритмический подход. — М.: Мир, 1978. — 432 с.

Ускорение бустинга параметрических классификаторов с использованием генетических алгоритмов*

Янгель Б. К.
hr0nix@acm.org
Москва, МГУ им. М. В. Ломоносова

В работе предлагается подход к ускорению бустинга классификаторов, представимых в специальном параметрическом виде. Для этого используется генетический алгоритм. Проведенные эксперименты показывают, что предложенный подход позволяет значительно снизить время обучения без потери качества классификации.

Введение

Одним из широко используемых в настоящее время методов обучения классификаторов является бустинг [1]. Бустинг представляет из себя метаалгоритм машинного обучения, позволяющий обучить классификатор в виде линейной комбинации так называемых *слабых* или *базовых* классификаторов из некоторого семейства W. В случае задачи классификации на два класса (которую мы далее и будем рассматривать) такой классификатор обычно имеет вид

$$s(y) = \operatorname{sign}\left(\sum_{i=1}^{N} \alpha_i w_i(y)\right), \qquad (1)$$

где $y \in Y$ — объект для классификации, $w_i \in W$ — обученные базовые классификаторы, α_i — соответствующие им веса, $w_i(y) \in \{-1, 1\}$, $s(y) \in \{-1, 1\}$.

На каждом шаге алгоритма бустинга из семейства W необходимо выбрать (обучить) базовый классификатор, доставляющий минимум специальному функционалу ошибки, учитывающему веса каждого объекта обучения. При этом нередко в качестве базового классификатора на i-м шаге используется простейший пороговый классификатор

$$w_i(y) = g_i \operatorname{sign}(\varphi_i(y) - t_i), \qquad (2)$$

где $\varphi_i(y)$ — значение некоторого признака $\varphi_i \in \Phi$ объекта y, t_i — значение порога, а $g_i \in \{-1, 1\}$ контролирует знак результата. Если для выбора g_i и t_i обычно (для большинства широко используемых функционалов ошибки) существуют эффективные алгоритмы, то признак φ_i выбирают полным перебором по множеству признаков Φ. При этом временная сложность алгоритма обучения базового классификатора линейно зависит от мощности множества Φ.

Пороговые базовые классификаторы широко применяются, например, в задачах распознавания объектов на изображениях. При этом признаками зачастую являются значения некоторой числовой характеристики на различных прямоугольных регионах изображения. Даже для изображения небольшого размера число возможных прямоугольных регионов в нем огромно. Следовательно, велико и число признаков, пропорционально которому растет время обучения классификатора. К примеру, процесс обучения классификатора на основе семейства признаков Хаара в известной работе [2] методом AdaBoost занял несколько недель. Тем не менее, поскольку значения таких признаков для сильно перекрывающихся регионов изображения обычно коррелируют, задачу поиска наилучшего признака можно решать, используя и отличные от полного перебора методы оптимизации.

Одним из широко известных методов оптимизации является генетический алгоритм [3], основанный на идеях биологической эволюции. Оптимизационная задача кодируется таким образом, чтобы её решение могло быть представлено в виде вектора — *хромосомы*. Случайным образом создается некоторое начальное множество векторов, называемое *начальной популяцией*. Они оцениваются с использованием *функции приспособленности*, в результате чего каждому вектору ставится в соответствие значение, которое определяет вероятность выживания организма, им представленного. После этого с использованием значений приспособленности производится *селекция* — выбирается подмножество векторов, к которым применяются *генетические операторы* (обычно это скрещивание и мутация), создавая таким образом следующее поколение векторов. Особи следующего поколения также оцениваются, производится селекция, применяются генетические операторы и т. д. Так моделируется эволюционный процесс, продолжающийся несколько жизненных циклов (*поколений*), пока не будет выполнен критерий остановки алгоритма. Таким критерием может быть нахождение глобального, либо субоптимального решения, исчерпание числа поколений или времени, отпущенного на эволюцию. Генетические алгоритмы служат главным образом для поиска глобальных экстремумов мультимодальных дискретных функций. Это делает генетический алгоритм хорошим кандидатом на роль алгоритма выбора признака для порогового классификатора.

*Работа выполнена при поддержке РФФИ, гранты № 08-07-445-а, № 08-07-12081

Обзор существующих работ

В ряде работ уже было предложено использовать генетический алгоритм для обучения пороговых базовых классификаторов. Так, в работе [5] была рассмотрена модификация алгоритма AdaBoost, в которой для обучения порогового классификатора на основе признака Хаара использовался генетический алгоритм со специальными операторами скрещивания и мутации. В работе [6] генетический алгоритм применялся для предварительного выбора нескольких тысяч лучших классификаторов, которые затем уже использовались для бустинга. В работе [7] процесс бустинга был непосредственно интегрирован с генетическим алгоритмом. Из текущего поколения выбиралось несколько лучших особей, на основе которых обучался очередной базовый классификатор. Выбранные особи затем давали потомство для следующей итерации бустинга. В работе [4] использовался специальный эволюционный алгоритм, названный авторами Evolutionary Hill-Climbing. В их реализации оператор скрещивания не использовался вовсе, а новые особи генерировались случайным образом. Зато на каждом шаге алгоритма ко всем членам популяции последовательно применялось 5 различных мутаций, и результат каждой из них отбрасывался, если эта мутация не улучшала значение функции приспособленности.

Во всех этих работах число признаков было столь большим, что выбор признака перебором был невозможен в принципе. Это заставляло авторов искать альтернативные методы дискретной оптимизации. Следует заметить, что в каждой из работ генетический алгоритм был использован некоторым специальным образом. В данной работе предлагается более общий подход.

Предлагаемый метод

Итак, нас интересует общий подход к решению оптимизационной задачи обучения базового порогового классификатора для случая, когда различные признаки не являются полностью независимыми. При этом подход должен обеспечивать значительное уменьшение времени обучения по сравнению с методом полного перебора признаков. Ниже будет предложен такой подход на основе генетического алгоритма. Для этого необходимо определить, что будет являться особью популяции, как будет вычисляться функция приспособленности и генетические операторы.

Выбор особи популяции. Пусть W — некоторое параметрическое семейство базовых классификаторов, то есть классификатор $w \in W$ однозначно определяется набором из n своих вещественных параметров $x_1, \ldots, x_n$. При этом будем, не теряя общности, считать, что для последних l параметров $x_{n-l+1}, \ldots, x_n$ (в дальнейшем называемых *связанными*) существует эффективный алгоритм обучения $L_E \colon \mathbb{R}^{n-l} \to \mathbb{R}^l$, который для фиксированных значений *свободных* параметров $x_1, \ldots, x_{n-l}$ находит оптимальные значения связанных, доставляющие минимум функционала ошибки $E \colon \mathbb{R}^n \to \mathbb{R}^+$. При этом оптимизировать при помощи генетического алгоритма необходимо только свободные параметры. Их совокупность и будет представлять особь популяции.

Для простейшего порогового классификатора (2) связанными параметрами будут g_i и t_i, а свободными — параметрическое описание признака φ_i.

Построение функции приспособленности. Естественно считать, что чем больше ошибка классификатора на обучающем множестве, тем меньше вероятность того, что он может принести пользу в дальнейшем. Такой классификатор скорее всего не должен перейти в следующее поколение. Это позволяет нам ввести функцию приспособленности $F \colon \mathbb{R}^{n-l} \to \mathbb{R}^+$ следующим образом:

$$F(x_1, \ldots, x_{n-l}) = \\ = 1/E[x_1, \ldots, x_{n-l}, L_E(x_1, \ldots, x_{n-l})]. \quad (3)$$

Случай $E = 0$ мы не рассматриваем, так как в этом случае базовый классификатор правильно классифицирует все примеры обучающей выборки и, следовательно, уже является искомым.

Представление генотипа. Для представления *генотипа* члена популяции (то есть его описания в терминах генетических алгоритмов) можно использовать любой метод, позволяющий закодировать множество свободных параметров. В данной работе было выбрано представление вектора параметров в виде битовой строки, эффективность которого в задачах оптимизации функций была не раз подтверждена. Множество альтернативных вариантов представления описано в работе [3].

В случае, если некоторым точкам из $\mathbb{R}^{n-l}$ не соответствует никакого базового классификатора, соответствующие им особи можно штрафовать, явно задавая для них $F = 0$ (так делалось в данной работе), или же вообще выбирать генотип и генетические операторы так, чтобы избежать их появления.

Выбор генетических операторов. В данной работе мы использовали два наиболее распространенных генетических оператора: скрещивание и мутацию. Для битовых строк они обычно определяются так:

- оператор скрещивания меняет местами у двух хромосом все биты справа от выбранной позиции (*одноточечное* скрещивание);
- оператор мутации изменяет случайный бит в хромосоме на противоположное значение.

Алгоритм 1. Обучение базового классификатора

1: Сгенерировать начальную популяцию из N случайных битовых строк;
2: **для** $i = 1, \ldots, K_{\max}$
3: Добавить в популяцию еще $\lceil NR_c \rceil$ особей, образованных применением оператора одноточечного скрещивания к парам лучших особей;
4: Применить оператор однобитовой мутации к $\lceil NR_m \rceil$ случайным членам популяции;
5: Для каждой особи популяции вычислить значение функции (3) от ее фенотипа;
6: Удалить из текущего поколения все особи, кроме N лучших;
7: Выбрать фенотип лучшей особи последнего поколения в качестве базового классификатора.

Общий вид алгоритма. Алгоритм 1 использует элитарную стратегию селекции. Этот алгоритм имеет следующие параметры:

— $N > 0$ — количество членов популяции;
— $K_{\max} > 0$ — число поколений генетического алгоритма;
— $R_c \in [0, 1]$ — интенсивность скрещивания;
— $R_m \in [0, 1]$ — интенсивность мутации.

Замечания. Преимущество предложенного подхода состоит в том, что вычислительная сложность алгоритма 1 не зависит от мощности семейства базовых классификаторов, и, соответственно, мощности семейства признаков. Необходимого компромисса между временем обучения и ошибкой можно достичь, только варьируя значения параметров N, $K_{\max}$ и S. Похожего эффекта можно достичь, сокращая само семейство базовых классификаторов. Однако в подавляющем большинстве случаев априорные знания о том, какие из них лучше подойдут для бустинга, недоступны.

Недостаток предложенного подхода, по видимому, состоит в узкой области его применимости. В параметрическом виде может быть представлено не так много видов базовых классификаторов. Такое представление хорошо подходит для пороговых классификаторов с признаками, имеющими некоторую пространственную связь. Для случая независимых признаков это представление уже подходит плохо, т. к. генетическая оптимизация свободных параметров в этом случае превращается в случайный поиск.

Результаты экспериментов

Для сравнения подходов к обучению базового порогового классификатора в данной работе были реализованы алгоритмы Viola-Jones [2] и Face alignment via BRM [8]. Они используют различные варианты реализации процедуры бустинга (AdaBoost и GentleBoost). При этом в [2] решается задача классификации, а в [8] — ранжирования. Совокупность этих факторов делает задачи весьма разнородными. Однако, поскольку основу базового классификатора в обоих задачах составляет признак Хаара, необходимы дополнительные эксперименты, чтобы удостовериться в эффективности метода.

В обеих задачах базовый классификатор задается параметрически в виде

$$w_i{=}(X_i, Y_i, W_i, H_i, type_i, g_i, t_i). \qquad (4)$$

При этом параметры g_i и t_i являются связанными, т. к. для них существует эффективный алгоритм обучения. Параметр $type_i$, кодирующий тип признака Хаара, также был сделан связанным, т. к. его изменение в ходе работы генетического алгоритма может изменять значение функции приспособленности непредсказуемым образом. Вместо этого для каждого типа признака (всего, как и в [8], использовалось 5 типов) оптимизация проводилась отдельно, а потом выбирался лучший результат.

Дополнительно было проведено сравнение двух различных схем запуска генетического алгоритма: в одной алгоритм запускался один раз с большим значением N, а в другой проводилось большое количество запусков (далее будем обозначать количество отдельных запусков алгоритма S) с небольшим размером популяции и выбирался лучший из классификаторов, найденных на каждом из запусков. Несмотря на то, что вторая схема показала результат чуть хуже, большое количество запусков с небольшим размером популяции может оказаться предпочтительнее из соображений производительности, так как в этом случае становится возможным запускать сразу несколько экземпляров алгоритма параллельно. При этом, поскольку затраты на коммуникации между экземплярами алгоритма равны нулю, достигается идеальное ускорение.

Для обучения классификатора [2], как и в работе [5], использовалась база изображений лиц [9], разбитая на обучающее и тестовое множество приблизительно пополам. Для обучения классификатора [8] использовалась база размеченных изображений лиц FG-NET, уменьшенных до размера в 40×40 пикселов. К каждому из 400 выбранных из базы лиц было применено по 10 случайных направленных изменений позиций контрольных точек, по 6 шагов в каждом. Тестовое множество было образовано таким же образом на основе еще 200 изображений.

В таблицах 1 и 3 для каждого из режимов обучения классификатора приведено время, затраченное в среднем на одну итерацию, а так же ускорение по сравнению с вариантом обучения, использующим перебор всех признаков. В таблицах 2 и 4 приведена доля ошибочно распознанных итоговыми классификаторами примеров в обучающем и тесто-

Таблица 1. *Viola-Jones*, ускорение.

Режим			Время (сек)	Ускорение
S	N	$K_{\max}$		
1	50	10	**2.82**	**329.38**
1	100	20	9.40	98.77
1	400	40	100.29	9.26
10	10	20	4.00	231.94
20	20	40	28.74	32.31
Полный перебор			928.52	1.00

Таблица 2. *Viola-Jones*, ошибка.

Режим			Ошибка	
S	N	$K_{\max}$	Обучение	Тест
1	50	10	0.0005	0.0356
1	100	20	0.0002	0.0380
1	400	40	0.0000	**0.0328**
10	10	20	0.0003	0.0378
20	20	40	0.0000	0.0391
Полный перебор			0.0000	0.0349

Таблица 3. *Face alignment via BRM*, ускорение.

Режим			Время (сек)	Ускорение
S	N	$K_{\max}$		
1	25	10	**68.15**	**5195.88**
1	50	10	173.33	2043.09
2	75	15	909.55	389.34
4	100	20	3582.37	98.85

Таблица 4. *Face alignment via BRM*, ошибка.

Режим			Ошибка	
S	N	$K_{\max}$	Обучение	Тест
1	25	10	0.0278	0.0317
1	50	10	0.0246	0.0297
2	75	15	0.0199	0.0268
4	100	20	**0.0173**	**0.0259**

вом множествах. Для алгоритма [8] обучение классификатора методом перебора признаков не проводилось, так как на выбранном обучающем множестве процесс бы занял около года.

Эксперименты с алгоритмом [2] показывают, что классификатор, обученный при помощи генетического алгоритма, может иметь обобщающую способность не хуже (или даже лучше, как в случае с N=400), чем у классификатора, обученного методом полного перебора. При этом классификатор, обученный с параметрами S=1, N=50 и $K_{\max}$=10, весьма незначительно проигрывает в обобщающей способности оригинальному, обеспечивая ускорение процесса обучения более чем в 300 раз. Больше всего ошибка у классификаторов, обученных с маленьким значением параметра N и большим значением S, хотя разница и не столь велика. Однако на многопроцессорных системах в таком режиме обучение можно производить значительно быстрее. В экспериментах с алгоритмом [8] хорошо видно, что качество итогового классификатора напрямую зависит от сложности генетического алгоритма обучения, образованной совокупностью параметров S, N и $K_{\max}$. Однако классификатор, полученный в режиме обучения S=1, N=25 и $K_{\max}$=10, имея ошибку на тестовой выборке всего в 1.2 раза больше, чем у самого качественного из полученных классификаторов, обеспечивает ускорение процесса обучения более чем в 50 раз по сравнению с ним.

Заключение

В работе был представлен подход к ускорению процесса обучения некоторых пороговых базовых классификаторов в рамках процедуры бустинга. Для этого был использован генетический алгоритм с хромосомами в виде битовых строк. При этом задача обучения базового классификатора свелась к поиску глобального минимума функционала ошибки на взвешенном обучающем множестве. Подход был обобщен для других классификаторов, эффективно представимых в параметрическом виде. Эксперименты показали, что предложенный метод позволяет значительно ускорить процесс обучения. При этом баланс между временем обучения и качеством получившегося классификатора хорошо поддается контролю. В дальнейшем предполагается обобщить предложенный подход для бустинга деревьев.

Литература

[1] *Schapire R. E.* The boosting approach to machine learning: an overview // MSRI Workshop on Nonlinear Estimation and Classification, 2002.

[2] *Viola P.* Robust real-time object detection // International Journal of Computer Vision, 2001.

[3] *Goldberg D. E.* Genetic algorithms in search, optimization, and machine learning // Addison-Wesley Professional, 1989.

[4] *Abramson Y., Moutarde F., Stanciulescu B., Steux B.* Combining adaboost with a hill-climbing evolutionary feature search for efficient training of performant visual object detectors // FLINS, 2006.

[5] *Treptow A., Zell A.* Combining AdaBoost learning and evolutionary search to select features for real-time object detection // IEEE Congress on Evolutionary Computation (CEC 2004). — Vol. 2. — Pp. 2107–2113.

[6] *Ramirez G. A.* Face and street detection with asymmetric haar features // `http://www.cs.utep.edu/tsolorio/VLLL/ramirez_fuentes07.pdf`.

[7] *Masada K., Chen Q., Wu H., Wada T.* GA based feature generation for training cascade object detector // Int. Conf. on Pattern Recognition, 2008. — С. 1–4.

[8] *Wu H., Liu X., Doretto G.* Face alignment via boosted ranking model // Computer Vision and Pattern Recognition, 2008. — С. 1–8.

[9] *Carbonetto P.* Viola-Jones training data. — `http://www.cs.ubc.ca/~pcarbo/viola-traindata.tar.gz`.

Развитие алгоритма многокритериального выбора оптимального подмножества диагностических тестов*

Янковская А. Е., Петелин А. Е.
yank@tsuab.ru, pae@sibmail.com
Томский государственный архитектурно-строительный университет

Рассматривается задача логико-комбинаторного построения оптимального подмножества (ОП) безызбыточных безусловных диагностических тестов (ББДТ). Описывается алгоритм многокритериального выбора ОП ББДТ и способы его развития. Излагаются идеи редукции критериев и методы сокращения перебора за счёт сжатия матрицы тестов и неполного построения дерева решений. Приводится иллюстративный пример и краткое описание программной реализации алгоритма многокритериального выбора ОП ББДТ.

Построение оптимального подмножества безусловных безызбыточных диагностических тестов (ББДТ) [1, 2], а не просто «хороших» тестов, весьма актуально при принятии решений в интеллектуальных системах. Применение «хороших» ББДТ не всегда приводит к оптимальному решению, поскольку общее количество признаков в выбранном множестве тестов может быть слишком большим. Кроме того, большими могут оказаться и временные и стоимостные затраты или ущерб (риск) [3], наносимый в результате выявления значений признаков исследуемого объекта, например, в медицине.

В публикации [1] впервые были предложены идея и алгоритм многокритериального выбора ОП ББДТ.

В данном докладе приводятся основные понятия и определения, излагаются идеи редукции критериев, описываются способы развития рассматриваемого алгоритма, приводится иллюстративный пример.

Основные понятия и определения

Воспользуемся определениями и обозначениями, введенными в публикациях [4–6].

Диагностическим тестом (тестом) называется совокупность бинарных признаков, различающих любые пары объектов, принадлежащих разным классам (образам) [4–6].

Тест называется *безызбыточным* [6] (тупиковым [4]), если никакая его часть не является тестом.

Безусловный тест характеризуется одновременным предъявлением всех входящих в него признаков исследуемого объекта при принятии решений.

Безусловным безызбыточным диагностическим тестом назовем тест, любое собственное подмножество которого не является тестом.

Для построения ББДТ используется матричное представление данных и знаний: матрицы описания объектов в пространстве характеристических признаков и матрицы различений, задающей различные разбиения объектов, представленных в матрице описаний, на классы эквивалентности [6].

Признак называется *обязательным* [5], если он содержится во всех ББДТ. Признак называется *псевдообязательным*, если он не является обязательным и входит во множество используемых при принятии решений ББДТ.

Пусть $T = \{t_{ij}\}_{i=1}^{n}{}_{j=1}^{m}$ — матрица ББДТ, n — количество ББДТ, m — количество характеристических признаков, T_i — i-я строка матрицы T;

$N = \{N_i \colon i = 1, \ldots, n\}$ — множество тестов;

$Z = \{z_j \colon j = 1, \ldots, m\}$ — множество характеристических признаков, представленное T_i, причем $t_{ij} = 1 \leftrightarrow z_j \in N_i$;

L_i — множество признаков, входящих в тест N_i;

n_0 — число используемых для принятия решений тестов, задаваемое извне и определяемое экспериментально;

T' — подматрица матрицы T, содержащая все столбцы матрицы T, за исключением единичных (константных) столбцов и столбцов, число единичных значений в которых меньше n_0;

T_0 — подматрица матрицы T';

N_0 — множество тестов, соответствующих строкам матрицы T_0, $N_0 \subseteq N$;

r — число псевдообязательных признаков.

Введем бинарную операцию β над выбранными столбцами матрицы T', результатом которой является двоичный вектор-столбец, компоненты которого принимают значения одноимённых компонент (элементов) выделенной пары столбцов матрицы T', если те совпадают между собой, и значение 0 в противном случае. Пример операции β приведем в разделе «Иллюстративный пример».

Введем множество P_k, $k = 1, \ldots, r$, элементами которого являются только те комбинации столбцов матрицы T' длиной k, в результате применения операции β над которыми будет получен вектор-столбец с числом единичных значений не меньше, чем n_0.

Введем подмножество $P_{k,i}$ множества P_k, $k = 1, \ldots, r$, $1 \leqslant i < t_{k-1}$, где t_i — мощность множества P_i. Подмножество $P_{k,i}$ содержит лишь те элементы множества P_k, которые являются резуль-

*Работа выполнена при финансовой поддержке РФФИ, проекты №07-01-00452 и №09-01-99014-р_офи.

татом выполнения операции β над i-м элементом множества P_{k-1} с другими элементами этого множества, начиная с $(i{+}1)$-го (см. иллюстративный пример).

Обозначим через M^+ множество, элементами которого являются подмножества строк матрицы T' мощности n_0, сопоставленные единичным значениям вектора-столбца, полученного в результате выполнения операции β над элементами множества P_r.

Обозначим для всех признаков $z_j,\ j = 1, \ldots, m$:

w_j^r — весовой коэффициент, определяемый как разделяющая способность признака z_j [5];

w_j^g — весовой коэффициент, характеризующий информационный вес и определяемый как отношение количества единичных значений признака z_j в рассматриваемом множестве тестов к мощности данного множества [3];

w_j^s — коэффициент стоимости признака z_j;

w_j^u — ущерб (риск), наносимый в результате измерения значения признака z_j (например, в медицинской практике риск может быть связан с проводимым инвазивным исследованием пациента).

Отметим, что по Ю. И. Журавлёву [4] информационный вес признака z_j равен количеству единичных значений этого признака для всех ББДТ.

Каждому тесту $N_i \in N\ \ i = 1, \ldots, n$ соответствуют:

- вес теста $W_i^r = \sum_{j \in L_i} w_j^r$;
- вес теста $W_i^g = \sum_{j \in L_i} w_j^g$;
- стоимость теста $W_i^s = \sum_{j \in L_i} w_j^s$;
- риск теста $W_i^u = \sum_{j \in L_i} w_j^u$.

Постановка задачи

Дано множество тестов N, представленное булевой матрицей T, множество признаков Z, каждый из которых содержится хотя бы в одном тесте из N, а также коэффициенты признаков w_j^r, w_j^g, w_j^s, w_j^u и коэффициенты тестов W_j^r, W_j^g, W_j^s, W_j^u.

Необходимо выделить из матрицы T такую подматрицу T_0, содержащую n_0 строк, чтобы соответствующее ей множество тестов N_0 обеспечивало выполнение следующих критериев, в порядке их следования:

K_1 содержало максимальное число псевдообязательных признаков;

K_2 содержало минимальное общее число признаков;

K_3 имело максимальный суммарный вес W_0^r;

K_4 имело максимальный суммарный вес W_0^g;

K_5 имело наименьшую суммарную стоимость W_0^s;

K_6 обеспечивало наименьший ущерб (риск) W_0^u.

Следует отметить, что последовательность критериев зависит от конкретной проблемной области и решаемой в ней задачи.

Редукция многокритериальной задачи построения подмножества тестов

Для уменьшения количества операций и, соответственно, увеличения скорости выполнения алгоритма выбора ОП ББДТ операция скаляризации критериев, т. е. переход к единственному критерию при редукции многокритериальной задачи построения ОП ББДТ, была бы наилучшим решением задачи. Однако, для данной задачи скаляризация невозможна, поскольку имеется противоречивость оценок подмножеств ББДТ (оценки, лучшие по одним критериям, не являются таковыми по другим [7]).

Проблеме редукции критериев посвящено большое количество работ [7–9]. Согласно [10] можно выделить следующие способы редукции критериев:

- на основе обобщенного критерия;
- с помощью «искусственного» отношения предпочтения;
- человеко-машинные процедуры;
- с использованием свойств отношения предпочтения ЛПР;
- с использованием «квантов информации».

При решении поставленной задачи формирования ОП ББДТ можно предположить, что критерии K_1–K_4 могут быть объединены с использованием первого способа, т. е. линейной сверткой:

$$\sum_i \lambda_i p_i,\ \ \lambda_i \geqslant 0,\ \sum_i \lambda_i = 1,$$

где p_i — степень удовлетворенности i-го критерия (i-й целевой функции).

Развитие алгоритма

Алгоритм 1 обеспечивает выполнение критериев K_1–K_4 в порядке их следования.

Повышение скорости выполнения алгоритма достигается за счёт сокращения размера исходной матрицы ББДТ путём удаления столбцов, число единичных значений в которых меньше n_0, а также, согласно следующей теореме, удаление столбцов, сопоставленных обязательным признакам.

Теорема 1. *Соотношения между весовыми коэффициентами ББДТ не изменяются при исключении из каждого теста обязательных признаков.*

Перебор, осуществляемый на этапах 2 и 3, можно представить в виде построения дерева поиска, которое строится в ширину. Сокращение перебора достигается за счёт построения дерева поиска в глубину. При этом первоначально находится решение по одной из ветвей дерева. Далее при построении каждой вершины последующих ветвей дерева анализируется достижимость листа (решения) на соответствующем уровне дерева.

Алгоритм 1. Алгоритм построения подматрицы T_0, удовлетворяющей критериям $К_1$–$К_4$.

Вход: матрица T;
Выход: матрица T_0;

1: построение сокращенной матрицы T' из матрицы T; построение множества P_1, сопоставленного столбцам матрицы T'; $k := 2$; $d := 2$;
2: построение подмножеств $P_{k,i}$, $i = 1, \ldots, t_{k-1}$;
3: $P_k = \bigcup\limits_i P_{k,i}$;
4: **если** $P_k \neq \varnothing$, **то**
5: $k := k + 1$ и переход к шагу 2;
6: построение множества M^+;
7: **если** $|M^+| = 1$, **то**
8: переход к шагу 19;
9: удаление из множества M^+ элементов, сопоставленные которым тесты не удовлетворяют критерию d; $d := d + 1$;
10: **если** $|M^+| = 1$, **то**
11: переход к шагу 19;
12: **если** $d = 3$, **то**
13: вычисление суммарного веса тестов, сопоставленных элементам множества M^+;
14: **если** $d = 4$, **то**
15: вычисление суммарного стоимости тестов, сопоставленных элементам множества M^+.
16: **если** $d < 5$, **то**
17: переход к шагу 9;
18: удаление всех элементов множества M^+, кроме первого;
19: построение подматрицы T_0, строками которой являются строки матрицы T', сопоставленные элементу множества M^+;

Иллюстративный пример

Пусть задано множество тестов $N = \{N_i\}_{i=1}^{9}$, представленное сокращенной матрицей тестов T' (рис. 1).

		1	2	6	7	8	9	10	W_i	W'_i
	1	**1**	**0**	1	1	1	1	0	2,4	32
	2	**0**	**1**	0	0	0	0	1	3,1	34
	3	**1**	**0**	1	1	0	1	1	2,2	29
	4	**0**	**1**	0	0	1	1	1	3	42
T =	5	**1**	**0**	1	0	0	0	0	3,5	48
	6	**0**	**0**	1	0	1	1	1	2,05	23
	7	**1**	**1**	1	1	1	1	0	2,6	35
	8	**0**	**1**	1	1	0	0	1	2,8	38
	9	**0**	**0**	1	1	1	1	1	1,85	18

Рис. 1. Сокращенная матрица тестов T'.

Для иллюстрации операции β 1-й и 2-й столбцы матрицы T' выделены жирно.

Результатом применения операции β над 1-м и 2-м столбцами матрицы T' (рис. 1) является вектор-столбец с единственным единичными значением (000000100).

Число n_0 тестов, используемых для принятия решения равно 3. По матрице T' построим множество $P_1 = \{1, 2, 6, 7, 8, 9, 10\}$ и представим его в виде корня дерева (рис. 2).

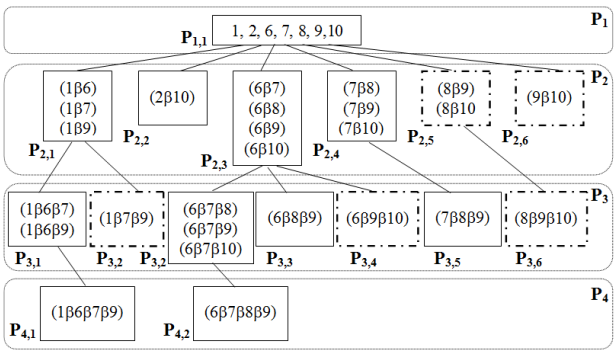


Рис. 2. Дерево реализации алгоритма.

Построим множества

$$P_{2,1} = \{(1\beta6),\ (1\beta7),\ (1\beta9)\},$$
$$P_{3,1} = \{(1\beta6\beta7),\ (1\beta6\beta9)\},$$
$$P_{4,1} = \{(1\beta6\beta7\beta9)\}.$$

Поскольку $|P_{4,1}| = 1$, то $r := 4$, и в множество M^+ включаем подмножество строк $\{1, 3, 7\}$, соответствующее $P_{4,1}$.

Заметим, что множество $P_{3,2} = \{(1\beta7\beta9)\}$ строить нет необходимости, поскольку соответствующее множество тестов не будет удовлетворять критерию $К_1$. На рис. 2 для наглядности приведены все множества. Множества, обведенные штрихпунктирной линией (5 множеств из 16), строить нет необходимости.

Рассуждая аналогичным образом, получим $P_{4,2} = \{(6\beta7\beta8\beta9)\}$, причём $|P_{4,2}| = 1$, $r = (k-1) = (5-1) = 4$. Следовательно, в множество M^+ включаем подмножество строк $\{1, 7, 9\}$, соответствующие $P_{4,2}$.

В результате $M^+ = \{\{1, 3, 7\}, \{1, 7, 9\}\}$, $|M^+|$=2, а следовательно, критерию $К_1$ удовлетворяет два подмножества тестов. Применение критерия $К_2$ не уменьшает мощность множества M^+, поскольку для подмножества тестов {1,3,7}, {1,7,9} множества M^+ число признаков одинаково.

Проверим выполнимость критерия $К_3$. Элемент $\{1, 7, 9\}$ множества M^+ не удовлетворяет этому критерию, поскольку W_1=7,2 > W_2=6,85. Таким образом, после выполнения критерия с получаем $M^+ = \{1, 3, 7\}$.

Поскольку $|M^+| = 1$, проверка на выполнимость критерия $К_4$ не имеет смысла.

Для данного примера при $n_0 = 3$ выбрано оптимальное подмножество тестов $N_0 = \{1, 3, 7\}$.

Программная реализации

Подсистема многокритериального выбора ОП ББДТ реализована на языке программирования C++ в виде динамически подключаемого модуля (плагина) к интеллектуальному инструментальному средству (ИИС) ИМСЛОГ [11].

Плагину передается ряд входных параметров: целочисленная матрица тестов, число используемых для принятия решений тестов, а также весовые коэффициенты и стоимости тестов. Выходным параметром является целочисленная матрица, строки которой сопоставлены тестам, входящим в оптимальное подмножество тестов.

Программная реализация плагина выполнена с использованием системы программирования Microsoft Visual C++ 2008 Express Edition.

Построение дерева реализовано в виде рекурсивной функции, предназначенной для построения множества $P_{k,i}$.

Для вычислений и хранения промежуточных данных используется встроенная библиотека классов ИИС ИМСЛОГ, а также стандартная библиотека STL.

Данная подсистема, включенная в ИИС ИМСЛОГ, может быть использована в прикладной интеллектуальной системе, сконструированной на базе ИИС ИМСЛОГ для различных проблемных и междисциплинарных областей: медицина, экономика, строительство, экогеология, экология, генетика, социология и др.

Заключение

Развит алгоритм многокритериального выбора ОП ББДТ путём сокращения перебора, достигаемого снижением размерности исходной матрицы ББДТ и за счёт неполного обхода дерева поиска решения. Для увеличения быстродействия алгоритма предложено выполнение редукции критериев оптимизации.

Алгоритм реализован в виде подсистемы, подключенной к ИИС ИМСЛОГ и исследован на ряде примеров.

Развитие алгоритма оптимального многокритериального выбора диагностических тестов связано с увеличением числа критериев и оптимизацией выбора ОП ББДТ.

Дальнейшие исследования будут направлены на реализацию процедуры редукции критериев в подпрограмме многокритериального выбора ОП ББДТ, увеличения количества рассматриваемых критериев, а также на проведение проверки эффективности реализованных сокращений перебора при решении практических задач.

Литература

[1] *Янковская А. Е.* Построение логических тестов с заданными свойствами и логико-комбинаторное распознавание на них // ИОИ-2002, Симферополь, 2002.— С. 100–102.

[2] *Yankovskaya A. E., Mozheiko V. I.* Optimization of a set of tests selection satisfying the criteria prescribed // 7th Int. Conf. PRIA, St. Petersburg: SPbETU, 2004. — Vol. I.— Pp. 145–148.

[3] *Янковская А. Е.* Критерии оптимизации выбора безызбыточных диагностических тестов для принятия решений в интеллектуальных диагностических системах // ММРО-13, М.: МАКС Пресс, 2007. — С. 73–76.

[4] *Журавлёв Ю. И., Гуревич И. Б.* Распознавание образов и анализ изображений // Искусственный интеллект: В 3-х кн. Кн.2. Модели и методы: Справ. / Под ред. Д.А. Поспелова. М.: Радио и связь, 1990. — С. 149–191.

[5] *Yankovskaya A. E.* Test pattern recognition with the use of genetic algorithms // Pattern recognition and image analysis. — 1999. — Vol. 9, № 1. — Pp. 121–123.

[6] *Янковская А. Е.* Логические тесты и средства когнитивной графики в интеллектуальной системе // Новые информационные технологии в исследовании дискретных структур: докл. 3-ей Всерос. конф., Томск: Изд-во СО РАН, 2000. — С. 163–168.

[7] *Авен П. О., Ослон А. А., Мучник И. Б.* Функциональное шкалирование. — М.: Наука, 1988.

[8] *Подиновский В. В.* Введение в теорию важности критериев в многокритериальных задачах принятия решения. — М.: Физматлит, 2007.

[9] *Подиновский В. В., Ногин В. Д.* Парето-оптимальные решения многокритериальных задач. — Н.: Наука, 2007.

[10] *Ногин В. Д.* Проблема сужения множества Парето: подходы к решению. Искусственный интеллект и принятие решений. — 2008. — № 1, С. 98–112.

[11] *Yankovskaya A. E., Gedike A. I., Ametov R. V., Bleikher A. M.* IMSLOG-2002 Software Tool for Supporting Information Technologies of Test Pattern Recognition // Pattern Recognition and Image Analysis. — 2003. — Vol. 13, № 2. — Pp. 243–246.

Проблемы эффективности вычислений и оптимизации

Код раздела: CO (Computation and Optimization)

- Проблемы алгоритмической и вычислительной сложности.
- Численные методы оптимизации в задачах интеллектуального анализа данных.
- Эффективные алгоритмы в цифровой обработке сигналов и изображений.

Построение эффективных линейных локальных признаков с использованием алгоритмов глобальной оптимизации*

Баврина А. Ю., Мясников В. В.
alina@smr.ru, vmyas@smr.ru
Самара, Институт систем обработки изображений РАН

В работе предлагаются два способа построения эффективных линейных локальных признаков одномерного сигнала, использующих для определения параметров признаков алгоритмы глобальной оптимизации. Производится их взаимное сравнение и сравнение со способом, использующим псевдоградиентный алгоритм. Показателем качества, используемым для сравнения, является величина, интегрально характеризующая степень коррелированности последовательностей, задающих импульсные характеристики алгоритма вычисления признаков, и точность представления с их помощью заданного набора сигналов.

Эффективные линейные локальные признаки и задачи их построения

Линейный локальный признак (ЛЛП) цифрового сигнала — это пара, включающая:

— последовательность отсчётов конечной импульсной характеристики (КИХ);
— алгоритм линейной локальной фильтрации цифрового сигнала для этой КИХ, называемый *алгоритмом вычисления признака* [1].

Понятие ЛЛП вводится в работе [1]. Эффективный ЛЛП обнаруживает оптимальное поведение в том смысле, что алгоритм вычисления признака обладает наименьшей вычислительной сложностью при одновременно экстремальном значении выбранного «показателя качества» для КИХ признака. *Общая задача* построения отдельного эффективного ЛЛП может быть сформулирована как задача построения последовательности (КИХ признака), порождающей алгоритм вычисления признака с наименьшей сложностью, которая наилучшим образом согласована с заданным производящим функционалом, численно определяющим «показатель качества» для КИХ. В работах [1, 2] показано, что более корректной (в смысле Адамара) является *частная задача* построения эффективного ЛЛП. В её постановке множество конечных последовательностей, среди которых ищется решение, ограничивается семейством НМС-последовательностей (НМС — нормализованная с минимальной сложностью [1]), обозначаемым $\wp(K, M, \bar{a})$. Здесь M — длина последовательности, K — порядок линейного (в общем случае — неоднородного) рекуррентного соотношения (ЛРС), $\bar{a} = (a_1, \ldots, a_K)^{\intercal}$ — вектор коэффициентов ЛРС [3]. Учитывая конечность мощности конкретного множества $\wp(K, M, \bar{a})$ [1], частная задача построения эффективного ЛЛП может быть решена перебором за конечное время с использованием алгоритма, приведенного в [2].

Определённым недостатком эффективных ЛЛП, полученных как решение частной задачи, и соответствующего способа их построения является то, что ограничение класса последовательностей отдельным семейством $\wp(K, M, \bar{a})$ является слишком «жёстким». Для практического использования более естественной является постановка, в которой фиксируются только ключевые параметры (K, M) семейства, определяющие сложность алгоритма вычисления признака, а тип семейства, характеризующийся вектором коэффициентов $\bar{a}$, остаётся неизвестным. Такую задачу далее назовем *расширенной частной задачей* построения эффективных ЛЛП. Более формальное определение этой задачи дано в работе [4]. Вычислительная сложность $u(A)$ алгоритма A вычисления ЛЛП в этом случае определяется так же, как для полностью фиксированного семейства НМС-последовательностей $\wp(K, M, \bar{a})$ и задаётся величиной [1, 2]:

$$u(A)\,\tfrac{N-M+1}{N} = 2K, \qquad (1)$$

здесь N — длина сигнала, для которого рассчитывается ЛЛП.

Целью настоящей работы является разработка и исследование (сравнение) различных способов решения расширенной частной задачи построения эффективных ЛЛП. *Результатом работы* должны стать выводы о целесообразности использования каждого конкретного алгоритма для решения рассматриваемой задачи.

В работе предлагаются два новых способа решения указанной задачи, использующих для определения параметров признаков алгоритмы глобальной оптимизации — генетический и имитации отжига. Производится экспериментальное сравнение этих способов по качеству получаемых решений. Дополнительно производится экспериментальное сравнение результатов построения с результатами, полученными с использованием псевдоградиентного алгоритма, предлагаемого в работе [4]. Показателем качества, выбранным для сравнения, явля-

*Работа выполнена при финансовой поддержке РФФИ, проекты №06-01-00616-а и №09-01-00434-а, Программы фундаментальных исследований Президиума РАН «Фундаментальные проблемы информатики и информационных технологий», проект 2.12.

ется величина, интегрально характеризующая степень коррелированности последовательностей, задающих импульсные характеристики ЛЛП, и точность представления с их помощью заданного набора сигналов. Выбор именно такого показателя качества в данном исследовании связан с тем, что в ситуации, когда на класс последовательностей (и, следовательно, вычислительную сложность алгоритма вычисления ЛЛП) не накладываются ограничения, рассматриваемая задача имеет известное оптимальное решение в виде разложения Карунена-Лоэва. Таким образом, в результате исследования можно сопоставить показатель качества, получаемый для конструируемых эффективных ЛЛП, с его оптимальным значением, получаемым для разложения Карунена-Лоэва.

Показатель качества признаков

Выбор показателя качества признаков тесно связан с той прикладной задачей, которая решается в конкретной системе обработки и анализа цифровых сигналов и изображений. В настоящей работе мы используем один из возможных показателей, который формализует требования, часто выдвигаемые к конструируемым ЛЛП на практике. Дадим вначале формулировку соответствующей прикладной задачи.

Пусть $\{X_i\}_{i=0}^{I-1}$ — набор из I известных конечных последовательностей (сигналов) длины M над $\mathbb{R}$. Требуется построить T $(T \leqslant M)$ последовательностей $\{h_t\}_{t=0}^{T-1}$ над $\mathbb{R}$, которые представляют набор $\{X_i\}_{i=0}^{I-1}$ с наименьшей среднеквадратической ошибкой и коррелированны взаимно в наименьшей степени. Без ограничений на класс последовательностей эта задача является хорошо известной и, как было указано выше, имеет известное решение в виде разложения Карунена-Лоэва.

Формализуя требования к последовательностям $\{h_t\}_{t=0}^{T-1}$ (в рамках настоящей работы класс рассматриваемых последовательностей ограничен НМС-последовательностями), зададим производящий функционал (показатель качества признаков) J таким образом, чтобы характеризовать одновременно их взаимную коррелированность и ошибку представления с их помощью последовательностей из набора $\{X_i\}_{i=0}^{I-1}$. Это может быть сделано следующим образом:

$$J = \alpha \cdot \frac{\sum\limits_{i=0}^{I-1} \|\Delta X_i\|^2}{\sum\limits_{i=0}^{I-1} \|X_i\|^2} + \\ + (1-\alpha) \cdot \frac{2}{T(T-1)} \sum_{t=0}^{T-2} \sum_{q=t+1}^{T-1} \frac{\langle h_t, h_q \rangle}{\sqrt{\|h_t\| \, \|h_q\|}}, \quad (2)$$

где ΔX_i — абсолютная ошибка представления последовательности X_i с помощью НМС-последовательностей $\{h_t\}_{t=0}^{T-1}$.

Первое слагаемое в (2) определяет относительную ошибку представления последовательностей $\{X_i\}_{i=0}^{I-1}$ с помощью конструируемых последовательностей. Второе слагаемое, равное квадрату нормы Гильберта-Шмидта [5], отражает степень коррелированности конструируемых последовательностей. Величина $\alpha \in [0,1]$ характеризует требуемый баланс между точностью представления последовательностей исходного множества и некоррелированностью формируемого набора последовательностей.

Учитывая характер введенного производящего функционала (2), набор из T НМС-последовательностей и соответствующих им ЛЛП считаем тем лучше, чем меньше значение функционала.

Общая идея метода решения расширенной частной задачи построения эффективных ЛЛП

Задача построения набора $\{h_t\}_{t=0}^{T-1}$ НМС-последовательностей является оптимизационной задачей, в процессе решения которой необходимо определить:

- коэффициенты ЛРС $\bar{a}_t$ для каждой НМС-последовательности $h_t(\cdot)$, которые фиксируют её семейство $\wp(K, M, \bar{a}_t)$;
- собственно НМС-последовательность $\{h_t(m)\}_{m=0}^{M-1}$ (её отсчёты) семейства $\wp(K, M, \bar{a}_t)$, минимизирующую значение производящего функционала (2).

Вторая подзадача в этом списке — это частная задача построения эффективного ЛЛП, решение которой может быть точно найдено с помощью алгоритма, указанного в работе [2]. Для решения первой подзадачи (учитывая, что оно может быть не единственным) предлагается использовать алгоритмы глобальной оптимизации. В настоящей работе используются генетический алгоритм [6, 7] и алгоритм имитации отжига [8].

Заметим также, что для снижения сложности поисковой процедуры искомые НМС-последовательности $\{h_t\}_{t=0}^{T-1}$ предлагается конструировать последовательно (путём последовательного присоединения). Тогда независимо от используемого алгоритма оптимизации первая подзадача разбивается на T последовательно решаемых задач поиска в K-мерном пространстве коэффициентов ЛРС $\bar{a}_t$. При этом внутри алгоритма глобальной оптимизации, отвечающего за поиск коэффициентов ЛРС, для каждого потенциального вектора коэффициентов $\bar{a}_t$ производится решение частной задачи построения эффективного ЛЛП известным способом [2].

Ниже дополнительно приведены комментарии по используемым в настоящей работе алгоритмам глобальной оптимизации.

Генетический алгоритм. Основными операциями генетического алгоритма являются селекция, скрещивание и мутация [6, 7]. В качестве хромосомы естественно использовать вектор $\bar{a}_t$ вещественных коэффициентов ЛРС. При таком выборе операции селекции и мутации реализуются наиболее простым и естественным образом. Детали реализации и параметры алгоритма представлены в докладе.

Алгоритм имитации отжига. Основная идея алгоритма имитации отжига заключается в недетерминированном механизме принятии нового решения (задаваемого вектором параметров), полученного из старого случайной коррекцией. А именно, на текущей итерации новое решение принимается однозначно, если значение производящего функционала на нём меньше. Если же значение функционала увеличилось, то новое решение принимается с некоторой вероятностью, зависящей как от величины «ухудшения» решения (чем больше величина ухудшения, тем ниже вероятность принятия), так и от номера итерации (чем больше номер, тем меньше вероятность). Детали реализации и параметры алгоритма представлены в докладе.

Результаты экспериментальных исследований

Проводимое экспериментальное исследование имело целью сравнить результаты решения расширенной частной задачи построения эффективных ЛЛП, получаемые с использованием алгоритмов глобальной оптимизации, а также сопоставить эти результаты с результатами, получаемыми с использованием псевдоградиентного алгоритма. Описание последнего (использующего метод деформируемого многогранника) дано в работе [4].

В качестве исходных данных для проведения исследования были синтезированы $I = 50$ последовательностей $\{X_i\}_{i=0}^{I-1}$ длины $M = 21$. Последовательности представляли собой реализации дискретного стационарного случайного процесса с автокорреляционной функцией $B_X(m) = D_X\rho^m$, где $D_X = 1,\ \rho = 0{,}9$.

Для указанного набора последовательностей были построены $T = 4$ НМС-последовательностей $\{h_t\}_{t=0}^{T-1}$ с использованием указанных выше способов (генетический алгоритм, алгоритм отжига и псевдоградиентный алгоритм) для случаев $K = 2, 3, 4, 5$. Для каждого построенного набора НМС-последовательностей рассчитывалось значение производящего функционала (2). Параметр α, характеризующий баланс между точностью представления последовательностей исходного множества и некоррелированностью формируемого набора последовательностей, полагался равным 0,5. Учитывая, что используемые алгоритмы имеют стохастическую природу и получаемые значения производящего функционала, вообще говоря, случайны, эксперименты по построению каждым конкретным алгоритмом повторялись 10 раз для каждого конкретного значения K. По полученным 10 значениям функционала J вычислялись среднее J_{mean} и среднеквадратическое J_{MSR} значения производящего функционала. Результаты проведенного вычислительного эксперимента представлены в приведенной ниже таблице 1.

Таблица 1. Показатели качества линейных локальных признаков, конструируемых с использованием различных алгоритмов.

K	$2KT$	Алгоритм	J_{mean}	J_{MSR}
2	16	генетический	0,265	7e-02
		отжига	0,466	2e-01
		псевдоградиентный	0,325	6e-02
3	24	генетический	0,137	1e-02
		отжига	0,219	6e-02
		псевдоградиентный	0,264	5e-02
4	32	генетический	0,124	4e-03
		отжига	0,183	5e-02
		псевдоградиентный	0,160	2e-02
5	40	генетический	0,117	2e-03
		отжига	0,161	5e-02
		псевдоградиентный	0,139	1e-02

Значение критерия J для $T = 4$ последовательностей, соответствующих разложению Карунена-Лоэва, составило $J = 0{,}099$. При этом вычислительная сложность расчета ЛЛП (при использовании прямого алгоритма вычисления линейной свертки), составляет $MT = 84$ операции в отличие от $2KT$ операций, необходимых для вычисления эффективных ЛЛП (см. таблицу).

В дополнении к указанной таблице в докладе представлены результаты экспериментального исследования для случаев других значений параметра α, а также для случая введения ограничений на вектор параметров ЛРС $\bar{a}$.

Заметим, что время, затрачиваемое на построение соотвествующего набора тем или иным алгоритмом, в рамках данной работы не представляет интереса, поскольку построение эффективных ЛЛП выполняется на этапе проектирования и построения системы обработки и анализа изображений и на практике не связано с жесткими ограничениями по времени (в отличие от процесса вычисления ЛЛП, выполняемого в момент функционирования такой системы).

На основании полученных численных результатов экспериментального исследования, а также на основании качественного анализа используемых алгоритмов можно сделать следующие выводы:

— наилучшее среднее качество набора НМС-последовательностей и, соответственно, признаков позволяет получить способ построения, основанный на генетическом алгоритме;
— для конкретной задачи генетический алгоритм дает наименьший разброс величины качества;
— наихудшие показатели демонстрирует алгоритм имитации отжига;
— алгоритмы глобальной оптимизации имеют преимущество перед псевдоградиентным алгоритмом в плане возможного учета дополнительных ограничений задачи, выражаемых, например, в виде ограничений на значения вектора ЛРС $\bar{a}$;
— хотя значение показателя качества J для конструируемых эффективных ЛЛП не может быть лучше значения, получаемого для разложения Карунена-Лоэва, видно, что даже для небольших K ($K = 4, 5$) значение этого показателя уже близко к оптимальному. При этом вычислительная сложность алгоритмов вычисления ЛЛП оказывается существенно ниже.

Исходя из представленных выводов в качестве метода решения расширенной частной задачи построения эффективных ЛЛП следует рекомендовать способ, использующий генетический алгоритм в качестве алгоритма численной оптимизации.

Выводы

В работе предложены и исследованы два способа построения эффективных линейных локальный признаков одномерного сигнала. Предложенные способы используют генетический алгоритм и алгоритм имитации отжига в качестве алгоритмов численной глобальной оптимизации. Производится сравнение качественных показателей признаков, построенных указанными способами и способом, использующим псевдоградиентный алгоритм. Результаты экспериментального исследования показали однозначное преимущество способа построения, использующего генетический алгоритм оптимизации.

Литература

[1] *Мясников В. В.* Эффективные локальные линейные признаки цифровых сигналов и изображений // Компьютерная оптика. — 2007. — Т. 31, № 4. — С. 58–76.

[2] *Мясников В. В.* О постановке и решении задачи построения эффективных линейных локальных признаков цифровых сигналов // Всеросс. конф. ММРО-14. — М.: МАКС Пресс, 2009 — С. 268–271.

[3] *Лидл Р., Нидеррайтер Г.* Конечные поля: В 2-х т., Т. 1. — М.: Мир, 1988. — 430 с.

[4] *Титова О. А., Мясников В. В.* Псевдоградиентный алгоритм построения эффективных линейных локальных признаков // Всеросс. конф. ММРО-14. — М.: МАКС Пресс, 2009 — С. **??**–**??**.

[5] *Виттих В. А., Сергеев В. В., Сойфер В. А.* Обработка изображений в автоматизированных системах научных исследований. — М.: Наука, 1982. — 214 с.

[6] *Goldberg D. E.* Genetic algorithms in search, optimization, and machine learning. — Reading, MA: Addison–Wesley, 1989.

[7] *Holland J. H.* Adaptation in natural and artificial systems. — Ann Arbor: University of Michigan Press, 1975.

[8] *Metropolis N., Rosenbluth A. W., Rosenbluth M. N., Teller A. H., Teller E.* Equation of state calculations by fast computing machines // J. Chem. Phys. — 1953. — Vol. 21.— Pp. 1078–1092.

О границах однозначной реконструкции слов и структуре слов, неразличимых по фрагментарной информации*

Власов П. С., Жданов С. А.
i_am_vlasov@mail.ru, zjdanov.mpgu@ru.net
Москва, Московский педагогический государственный университет

Рассматривается задача восстановления слов над конечным алфавитом по информации в виде мультимножеств фрагментов слов. Для случая полного мультимножества фрагментов определенной длины изучены границы длины, при которых реконструкция однозначна. В ходе проведенных компьютерных экспериментов изучались значения функции $n(k)$, которая по длине фрагментов k возвращает длину n минимальных слов, не различимых по мультимножеству фрагментов длины k. Получены значения $n(k)$ для более широкого диапазона k, чем было известно ранее. Описаны структурные особенности неразличимых слов.

Теория реконструкции слов возникла в результате интеграции алгебраической комбинаторики слов с методами теории кодирования и распознавания образов. Спецификой теории реконструкции как раздела комбинаторики слов является ее основная задача: синтез слов по частичным представлениям. Задачи реконструкции слов тесно связаны с задачами распознавания образов, интеллектуального анализа данных, кодирования, символической динамики. Кодирование и распознавание образов являются в некотором смысле предельными случаями реконструкции слов: задачу построения кодов можно считать задачей определения множеств слов, восстанавливаемых по единственному искаженному образцу при искажениях заданного вида, в то время как задачи реконструкции с неформальным описанием классов слов относятся к классу задач распознавания. В символической динамике возникает задача исследования различимости символических динамических систем по подсловам и фрагментам, и разработки методов, позволяющих восстановить символическую динамическую систему по некоторому набору фрагментов слов, входящих в систему.

Рассматриваемая задача заключалась в исследовании возможности восстановления слова длины n по мультимножеству его фрагментов длины k в произвольном конечном алфавите. Для каждой длины фрагмента k надо было найти длину n слова, начиная с которой слово уже однозначно не восстанавливается, т. е., начиная с которой существуют как минимум 2 слова с одинаковыми мультимножествами фрагментов.

Обозначим через $n(k)$ функцию, которая по длине фрагментов k возвращает минимальную длину n слова, которое не восстанавливается однозначно по мультимножеству фрагментов. Обратную к ней функцию обозначим через $k(n)$.

В настоящее время известные границы для $k(n)$:

$$\left\lceil \frac{\log n}{\log \alpha} + 1 \right\rceil \leqslant k(n) \leqslant \left\lfloor \frac{16}{7}\sqrt{n} \right\rfloor + 5, \quad \alpha = \frac{\sqrt{5}+1}{2}.$$

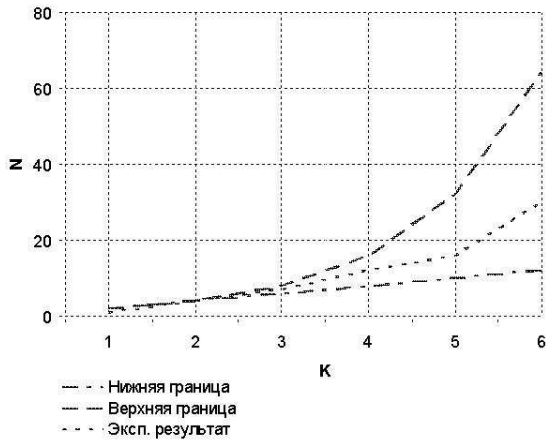


Рис. 1. Известные границы для $n(k)$ и полученные точные экспериментальные значения.

Нижняя граница получена построением конкретных примеров [1, 2]. Пары неразличимых слов — $\binom{01}{10}$, $\binom{0110}{1001}$, $\binom{0110001}{1000110}$, $\binom{011000100110}{100010110001}$. Верхняя граница получена неконструктивно, с помощью перехода от слов к многочленам, значения которых в точках $0, 1, \ldots, n-1$ определяются значениями координат фрагментов [3]. Пусть $X = x_0 x_1 \ldots x_{n-1}$ — неизвестное слово, $S_j(X)$ — число единиц, появляющихся на j-м месте в совокупности фрагментов. Тогда

$$S_j(X) = \sum_{i=0}^{n-1} C_i^j C_{n-i-1}^{k-j-1} x_i, \quad j = 0, \ldots, k-1.$$

Легко показать, что при фиксированных n и k многочлены

$$f_j(t) = C_t^j C_{n-t-1}^{k-j-1}, \quad j = 0, \ldots, k-1,$$

образуют базис в пространстве многочленов степени $(k-1)$. В [4] на основе многочленов Чебышёва 1-го рода построен многочлен $f(r)$ степени $k-1 = \left\lfloor \frac{16}{7}\sqrt{n} \right\rfloor + 4$, для которого выполняется условие

$$f(0) > f(1) + f(2) + \ldots + f(n-1).$$

Эти подходы не дают возможности улучшить представленные границы.

*Работа поддержана грантом РФФИ № 08-01-00837-а.

Целью проведенных численных экспериментов было исследование и уточнение этих границ. Фиксировалась длина фрагментов k и постепенно увеличивалась длина слова n до того, пока не найдется хотя бы пара неразличимых слов. Предел возможностей стандартных персональных компьютеров был достигнут уже при малых k: было определено, что $n(5) = 16,\ n(6) > 22$ (проверка последнего факта заняла около 40 часов). Полученные экспериментальные результаты показывают, что для выдвижения гипотез о виде функции $n(k)$ потребуется достаточно существенное увеличение k, поэтому чисто программистскими приемами вряд ли удастся добиться полезных результатов. Для проверки больших k требуется перейти к вычислительным системам со значительно большей мощностью.

Была разработана процедура распараллеливания, использованная для проведения расчетов на 8-ядерной кластерной системе. В результаты было показано, что $n(6) = 30,\ n(7) > 38$.

Для дальнейшего продвижения проведен анализ структуры найденных неразличимых слов. На основании этого анализа выдвинуты следующие гипотезы: все слова начинаются на 100 и заканчиваются на 001; для четных k существует всего 2 пары неразличимых слов, причем одна пара — симметричное отражение другой; для нечетных k существует всего одна пара неразличимых слов, причем одно слово из пары является симметричным отражением другого; при нечетном k слово, определяющее число $n(k)$, является началом слова, определяющего число $n(k+1)$.

В ближайшей перспективе планируется запустить программу на многокластерной системе МГУ, насчитывающей порядка 10000 параллельных ядер, вычислить $n(7), \ldots$ и постараться найти закономерность в граничных числах. Предполагается также провести исследование структурных особенностей множеств неразличимых слов вблизи от границы $n(k)$. Полученные результаты, а также результаты, которые будут получены при больших k, будут использованы для решения комбинаторных задач символической динамики, возникающих при фрагментарных описаниях. Фундаментальным понятием символической динамики является понятие сдвига, то есть множества бесконечных последовательностей над конечным алфавитом, каждая из которых не включает в качестве подслов конечные слова из заданного (возможно, бесконечного) списка запрещенных слов [5]. Объектами, привлекающими наибольшее внимание, являются софические сдвиги, описываемые графами, дугам которых приписаны символы алфавита, причем разным дугам могут соответствовать одинаковые символы. Частным случаем софических сдвигов являются конечные сдвиги, для которых список запрещенных слов конечен. Предположим, что задан набор фрагментов, и требуется определить, есть ли среди последовательностей из заданного сдвига такие, которые включают весь этот набор. Эта задача сводится к обобщению рассматриваемой здесь задачи: требуется найти не однозначную реконструкцию, а проверить, есть ли среди возможных решений те, которые удовлетворяют заданным запретам. Уточнение функции $n(k)$ позволит получить оценки размера выборки и величины фрагментарных описаний, при которых возможно получение точного ответа на вопрос о принадлежности слова заданному сдвигу, а также для разработки методов построения конечных или софических сдвигов по наборам запрещенных фрагментов.

Литература

[1] *Leont'ev V. K., Smetanin Yu. G.* Recognition Model with Representation of Information in the Form of Long Sequences // Pattern Recognition and Image Analysis: Advances in Mathematical Theory and Applications. — 2002. — Vol. 12, No. 3. — p. 250.

[2] *Леонтьев В. К., Сметанин Ю. Г.* О восстановлении вектора по набору его фрагментов // Докл. АН СССР. — 1988. — Т. 302, № 6. — С. 1319–1322.

[3] *Krasikov I., Rodity Y.* On a reconstruction Problem for Sequences // Journal of Computational Theory. Seires A. — 1997. — Vol. 77. — Pp. 344–348.

[4] *Borwein P., Erdelyi T., Kos G.* Littlewood-type problems on $[0, 1]$ // Proceedings of the London Mathematical Society. — 1999. — № 79 — Pp. 22–46.

[5] *Lind D., Marcus B.* An Introduction to Symbolic Dynamics and Coding // Cambridge University Press, Cambridge, UK. — 1995.

Алгоритм помехоустойчивого распознавания последовательности, включающей повторяющийся вектор, при наличии посторонних векторов-вставок из алфавита*

Долгушев А. В., Кельманов А. В.
dolgushev@math.nsc.ru, kelm@math.nsc.ru
Новосибирск, Институт математики им. С. Л.Соболева СО РАН, Новосибирский государственный университет

Рассматривается дискретная экстремальная задача, к которой сводится один из вариантов проблемы помехоустойчивого off-line распознавания векторных последовательностей, включающих в качестве элемента квазипериодически повторяющийся вектор евклидова пространства. Обоснован эффективный алгоритм решения задачи, гарантирующий оптимальность решения по критерию максимального правдоподобия в случае, когда помеха аддитивна и является гауссовской последовательностью независимых одинаково распределённых случайных величин.

Введение

Объект исследования работы — проблемы оптимизации в задачах анализа данных и распознавания образов. Предмет исследования — дискретная экстремальная задача, к которой сводится один из вариантов проблемы помехоустойчивого off-line распознавания векторной последовательности, как последовательности, включающей квазипериодически повторяющийся вектор, совпадающий с некоторым вектором из заданного алфавита векторов евклидова пространства. Цель работы — обоснование алгоритма решения этой задачи. На протяжении ряда лет статус сложности этой оптимизационной задачи был неизвестен и какие-либо алгоритмы с оценками точности также были неизвестны. Рассматриваемая задача является обобщением задачи, изученной в [1].

Одна из возможных содержательных трактовок задачи состоит в следующем. Источник сообщений через канал связи с помехой передает информацию об активном и пассивном состояниях некоторого физического объекта в виде упорядоченного набора — вектора — измеряемых характеристик. В пассивном состоянии значения каждой компоненты этого вектора равны нулю. Имеется конечная совокупность физических объектов. Каждому объекту соответствует уникальный набор измеряемых информационно важных характеристик. На приёмную сторону поступает зашумлённая последовательность квазипериодически перемежающихся векторов, в которой кроме информационно значимого вектора, соответствующего активному состоянию объекта, имеются посторонние векторы-вставки, принадлежащие конечному алфавиту вставок. Термин «квазипериодически» означает, что интервал между двумя последовательными ненулевыми векторами не одинаков, а лишь ограничен сверху и снизу некоторыми константами. Требуется определить (распознать), от какого из объектов была принята последовательность. Ситуации, в которых требуется решение подобной задачи, характерны, в частности, для геофизики, технической диагностики, электронной разведки и других приложений (см., например, [4] и цитированные там работы).

Постановка задачи

Пусть векторная последовательность $x_n \in \mathbb{R}^q$, $n \in \mathcal{N}$, где $\mathcal{N} = \{1, \ldots, N\}$, обладает свойством

$$x_n = \begin{cases} u, & n \in \mathcal{M}_1; \\ w_n, & n \in \mathcal{M}_2; \\ 0, & n \in \mathcal{N} \setminus (\mathcal{M}_1 \cup \mathcal{M}_2), \end{cases} \qquad (1)$$

где $\mathcal{M}_1 \cup \mathcal{M}_2 \subseteq \mathcal{N}$, $\mathcal{M}_1 \cap \mathcal{M}_2 = \varnothing$; $u \in \mathcal{A}_1$, $|\mathcal{A}_1| = K_1$; $w_n \in \mathcal{A}_2$, $n \in \mathcal{M}_2$, $|\mathcal{A}_2| = K_2$; $\mathcal{A}_1 \cap \mathcal{A}_2 = \varnothing$, $\mathcal{A}_1, \mathcal{A}_2 \subset \{u \,|\, u \in \mathbb{R}^q,\ 0 < \|u\|^2 < \infty\}$, где $\|\cdot\|$ — евклидова норма.

Положим $M_j = |\mathcal{M}_j|$, $j = 1, 2$, и $M = M_1 + M_2$. Вектор u будем интерпретировать как информационно значимый вектор, множество $\mathcal{A}_1$ — как алфавит информационно значимых векторов, вектор $w_n \in \mathcal{A}_2$, $n \in \mathcal{M}_2$, — как вектор-вставку, $\mathcal{A}_2$ — как алфавит векторов-вставок, а M_1 и M_2 — соответственно как число повторов информационно значимого вектора и число векторов-вставок в последовательности (1). Допустим, кроме того, что элементы набора $(n_1, \ldots, n_M)$, образующего совокупность $\{n_1, \ldots, n_M\} = \mathcal{M}_1 \cup \mathcal{M}_2$, удовлетворяют ограничениям

$$1 \leqslant T_{\min} \leqslant n_m - n_{m-1} \leqslant T_{\max} \leqslant N - 1 \qquad (2)$$

при каждом $m = 2, \ldots, M$. Ограничения (2), в которых $T_{\min}$ и $T_{\max}$ — константы, задают допустимый интервал между ближайшими номерами двух ненулевых векторов последовательности (1).

Доступной для анализа будем считать последовательность

$$y_n = x_n + e_n,\ n \in \mathcal{N}, \qquad (3)$$

где e_n — вектор помехи (ошибки измерения), независимый от вектора x_n. Заметим, что x_n =

*Работа выполнена при финансовой поддержке РФФИ, проекты № 09-01-00032, № 07-07-00022 и гранта АВЦП Рособразования, проект № 2.1.1/3235.

$= x_n(\mathcal{M}_1, \mathcal{M}_2, u, \{w_i, i \in \mathcal{M}_2\})$, $n \in \mathcal{N}$. Положим

$$S(\mathcal{M}_1, \mathcal{M}_2, u, \{w_i, i \in \mathcal{M}_2\}) = \sum_{n \in \mathcal{N}} \|y_n - x_n\|^2 \quad (4)$$

и рассмотрим следующую задачу.

Задача 1. *Дано*: последовательность $y_n \in \mathbb{R}^q$, $n \in \mathcal{N}$, и множества (алфавиты) $\mathcal{A}_1$, $\mathcal{A}_2$, $|\mathcal{A}_1| = K_1$, $|\mathcal{A}_2| = K_2$. *Найти*: вектор $u \in \mathcal{A}_1$, непересекающиеся подмножества $\mathcal{M}_1$ и $\mathcal{M}_2$ множества $\mathcal{N}$, а также множество $\{w_i \in \mathcal{A}_2,\ i \in \mathcal{M}_2\}$ такие, что целевая функция (4) минимальна при ограничениях (2) на элементы набора $(n_1, \ldots, n_M)$, которые образуют совокупность $\{n_1, \ldots, n_M\} = \mathcal{M}_1 \cup \mathcal{M}_2$.

К этой задаче сводится один из вариантов проблемы помехоустойчивого распознавания последовательности (1), как структуры, включающей повторяющийся ненулевой вектор, совпадающий с некоторым элементом из заданного алфавита векторов, которая кроме этого вектора содержит векторы-вставки из заданного алфавита векторов-вставок. В [3] показано, что к решению задачи 1 приводит статистическая формулировка проблемы, если считать, что $\{e_n\}$ в формуле (3) есть выборка из q-мерного нормального распределения с параметрами $(0, \sigma^2 I)$, где I — единичная матрица, а в качестве критерия решения задачи использовать максимум функционала правдоподобия.

Редуцированная задача

Раскрыв квадрат нормы разности векторов в правой части (4), используя (1), найдём

$$S = \sum_{n \in \mathcal{N}} \|y_n\|^2 - \sum_{n \in \mathcal{M}_1} \{2\langle y_n, u\rangle - \|u\|^2\} - \sum_{n \in \mathcal{M}_2} \{2\langle y_n, w_n\rangle - \|w_n\|^2\}, \quad (5)$$

где $\langle\cdot\rangle$ — скалярное произведение векторов. Первый член в правой части равенства (5) является константой. Поэтому имеем следующую оптимизационную задачу, к которой сводится минимизация целевой функции (4).

Задача 2. *Дано*: последовательность $y_n \in \mathbb{R}^q$, $n \in \mathcal{N}$, и множества (алфавиты) $\mathcal{A}_1$, $|\mathcal{A}_1| = K_1$, $\mathcal{A}_2$, $|\mathcal{A}_1| = K_2$, векторов из $\mathbb{R}^q$. *Найти*: вектор $u \in \mathcal{A}_1$, непересекающиеся подмножества $\mathcal{M}_1$ и $\mathcal{M}_2$ множества $\mathcal{N}$, а также множество $\{w_n \in \mathcal{A}_2,\ n \in \mathcal{M}_2\}$, доставляющие максимум целевой функции

$$G(\mathcal{M}_1, \mathcal{M}_2, u, \{w_n, n \in \mathcal{M}_2\}) = \sum_{n \in \mathcal{M}_1} \{2\langle y_n, u\rangle - \|u\|^2\} + \sum_{n \in \mathcal{M}_2} \{2\langle y_n, w_n\rangle - \|w_n\|^2\}, \quad (6)$$

при тех же ограничениях, что и в задаче 1.

Алгоритм решения задачи

Положим

$$g_1(n, u) = 2\langle y_n, u\rangle - \|u\|^2,$$

$$g_2(n, w_n) = 2\langle y_n, w_n\rangle - \|w_n\|^2,$$

$$u \in \mathcal{A}_1,\ w_n \in \mathcal{A}_2,\ n \in \mathcal{N}.$$

Тогда целевую функцию (6) можно переписать в виде

$$G = \sum_{n \in \mathcal{M}_1} g_1(n, u) + \sum_{n \in \mathcal{M}_2} g_2(n, w_n).$$

Кроме того, положим

$$d(n) = \max_{w_n \in \mathcal{A}_2} g_2(n, w_n),\ n \in \mathcal{N}.$$

В настоящей работе показано, что в случае, когда мощности M_1 и M_2 подмножеств $\mathcal{M}_1$ и $\mathcal{M}_2$ фиксированы, максимум $G_{\max}$ целевой функции G вычисляется по формуле

$$G_{\max} = \max_{u \in \mathcal{A}_1} G_{\max}(u), \quad (7)$$

где $G_{\max}(u) = \max_{\substack{\mathcal{M}_1, \mathcal{M}_2, \\ w_n \in \mathcal{A}_2}} G(\mathcal{M}_1, \mathcal{M}_2, w_n \in \mathcal{A}_2 \,|\, u)$ — условный максимум функции G для каждого фиксированного $u \in \mathcal{A}_1$, который находится по правилу

$$G_{\max}(u) = \max_{n \in \omega_M(M)} \max_{t \in \{M_1, \ldots, M\}} G_n(M_1, t, M \,|\, u). \quad (8)$$

Значения функции $G_n(M_1, t, M|u)$, $n \in \omega_M(M)$, $t \in \{M_1, \ldots, M\}$, вычисляются по рекуррентным формулам

$$G_n(l, t, m \,|\, u) = \begin{cases} g_1(n \,|\, u), & l = 1,\ t = 1,\ m = 1; \\ g_1(n \,|\, u) + \max\limits_{j \in \gamma^-_{m-1}(n)} F_j(m-1), & \\ & l = 1,\ t = 2, \ldots, M,\ m = t; \\ g_1(n \,|\, u) + & \\ + \max\limits_{j \in \gamma^-_{m-1}(n)} \max\limits_{s \in \{l-1, \ldots, m-1\}} G_j(l-1, s, m-1 \,|\, u), & \\ & l = 2, \ldots, M_1,\ t = l, \ldots, M,\ m = t; \\ d(n) + \max\limits_{j \in \gamma^-_{m-1}(n)} G_j(l, t, m-1 \,|\, u), & \\ l = 1, \ldots, M_1,\ t = l, \ldots, M,\ m = t+1, \ldots, M, \end{cases}$$

где $F_n(m) =$

$$= \begin{cases} d(n), & m = 1; \\ d(n) + \max\limits_{j \in \gamma^-_{m-1}(n)} F_j(m-1), & m = 2, \ldots, M, \end{cases}$$

при каждом $n \in \omega_m(M)$, причём

$$\omega_m(M) = \\ = \bigl\{n \,|\, 1+(m-1)T_{\min} \leqslant n \leqslant N-(M-m)T_{\min}\bigr\}$$

для каждого $m = 1, \ldots, M$ и

$$\gamma^-_{m-1}(n) = \\ = \bigl\{j \mid \max\{1+(m-2)T_{\min}, n-T_{\max}\} \leqslant j \leqslant n-T_{\max}\bigr\}$$

для каждого $n \in \omega_m(M)$ при фиксированном $m = 2, \ldots, M$.

Так как $\{n_1, \ldots, n_M\} = \mathcal{M}_1 \cup \mathcal{M}_2$, поиск непересекающихся подмножеств

$$\mathcal{M}_1 = \{n_{\mu_1}, \ldots, n_{\mu_{M_1}}\},\ \mathcal{M}_2 = \{n_{\nu_1}, \ldots, n_{\nu_{M_2}}\}$$

множества $\mathcal{N}$ эквивалентен поиску объединённого набора $(n_1, \ldots, n_M)$ и одного из подмножеств $\{\mu_1, \ldots, \mu_{M_1}\}$ или $\{\nu_1, \ldots, \nu_{M_2}\}$ множества $\{1, \ldots, M\}$; пусть для определённости искомым является подмножество $\{\mu_1, \ldots, \mu_{M_1}\}$.

Для каждого $u \in \mathcal{A}_1$ определим функции

$$\widetilde{G}_j(l, m \,|\, u) = \max_{s \in \{l-1, \ldots, m-1\}} G_j(l-1, s, m-1 \,|\, u),$$

$$L_j(l, m \,|\, u) = \operatorname*{arg\,max}_{s \in \{l-1, \ldots, m-1\}} G_j(l-1, s, m-1 \,|\, u),$$

где $l = 2, \ldots, M_1,\ \ m = l, \ldots, M,\ \ j \in \gamma^-_{m-1}(n)$, при каждом фиксированном $n \in \omega_m(M)$;

$$I_n(l, t, m \,|\, u) = \\ = \begin{cases} n, & l = 1,\ \ t = 1,\ \ m = 1; \\ \operatorname*{arg\,max}\limits_{j \in \gamma^-_{m-1}(n)} F_j(m-1), & \\ & l = 1,\ t = 2, \ldots, M,\ m = t; \\ \operatorname*{arg\,max}\limits_{j \in \gamma^-_{m-1}(n)} \widetilde{G}_j(l, m \,|\, u), & \\ & l = 2, \ldots, M_1,\ t = l, \ldots, M,\ \ m = t; \\ \operatorname*{arg\,max}\limits_{j \in \gamma^-_{m-1}(n)} G_j(l, t, m-1 \,|\, u), & \\ l = 1, \ldots, M_1,\ t = l, \ldots, M,\ m = t+1, \ldots, M, & \end{cases}$$

где $n \in \omega_m(M)$.

Тогда в соответствии с (7), искомый вектор $\hat{u}$ находится по правилу

$$\hat{u} = \operatorname*{arg\,max}_{u \in \mathcal{A}_1} G_{\max}(u),$$

а последние компоненты оптимальных наборов $(\hat{n}_1, \ldots, \hat{n}_M)$ и $(\hat{\mu}_1, \ldots, \hat{\mu}_{M_1})$ согласно (8) определяются по формуле

$$(\hat{n}_M, \hat{\mu}_{M_1}) = \\ = \arg \max_{n \in \omega_M(M)} \max_{t \in \{M_1, M_1+1, \ldots, M\}} G_n(M_1, t, M \,|\, \hat{u}).$$

Остальные компоненты (при $M_1 > 1$) оптимальных наборов находятся по правилу:

$$\hat{n}_{m-1} = \\ = \begin{cases} I_{\hat{n}_m}(M_1, \hat{\mu}_{M_1}, m \,|\, \hat{u}), & m = M, M-1, \ldots, \hat{\mu}_{M_1}; \\ I_{\hat{n}_m}(l, \hat{\mu}_l, m \,|\, \hat{u}), & m = \hat{\mu}_{l+1}-1, \ldots, \hat{\mu}_l, \\ & l = M_1-1, M_1-2, \ldots, 2; \end{cases}$$

$$\hat{\mu}_{l-1} = J_{\hat{n}_{\hat{\mu}_l}}(l, \hat{\mu}_l \,|\, \hat{u}),\ \ l = M_1, M_1-1, \ldots, 2,$$

где

$$J_n(l, m \,|\, u) = L_{I_n(l,m,m \,|\, u)}(l, m \,|\, u), \\ l = 2, \ldots, M_1,\ m = l, \ldots, M.$$

При этом, если $\hat{\mu}_2 - \hat{\mu}_1 > 1$, то

$$\hat{n}_{m-1} = I_{\hat{n}_m}(1, \hat{\mu}_1, m \,|\, \hat{u}),\ m = \hat{\mu}_2 - 1, \ldots, \hat{\mu}_1 + 1,$$

а если $\hat{\mu}_1 > 1$, то

$$\hat{n}_{m-1} = I_{\hat{n}_m}(1, m, m \,|\, \hat{u}),\ m = \hat{\mu}_1, \ldots, 2.$$

Векторы-вставки, доставляющие максимум целевой функции, находятся по правилу

$$\hat{w}_{\hat{\nu}_i} = B(\hat{n}_{\hat{\nu}_i}),\ i = 1, \ldots, M_2,$$

где

$$B(n) = \operatorname*{arg\,max}_{w_n \in \mathcal{A}_2} g_2(n, w_n),\ n \in \mathcal{N},$$

$$\{\hat{\nu}_1, \ldots, \hat{\nu}_{M_2}\} = \{1, \ldots, M\} \setminus \{\hat{\mu}_1, \ldots, \hat{\mu}_{M_1}\}.$$

Временная сложность алгоритма решения задачи 2 есть величина

$$O\bigl(K_1 M_1 M^2 (T_{\max} - T_{\min} + q)N + K_2 qN\bigr)$$

в случае, когда числа $T_{\max}$ и $T_{\min}$ являются частью входа задачи, и

$$O\bigl(K_1 M_1 M^2 (N + q)N + K_2 qN\bigr),$$

когда эти числа не являются частью входа.

Алгоритм решения редуцированной задачи лежит в основе алгоритма помехоустойчивого анализа и распознавания векторных последовательностей, как последовательностей, включающих квазипериодически повторяющийся вектор, совпадающий с некоторым вектором из заданного алфавита векторов евклидова пространства. Этот алгоритм гарантируют оптимальность решения как по критерию максимального правдоподобия в случае, когда помеха аддитивна и является гауссовской последовательностью независимых одинаково распределенных величин, так и по критерию минимума суммы квадратов уклонений.

Заключение

Рассмотренная задача входит в большое семейство актуальных задач [3, 4, 5], к которым сводятся

типовые проблемы помехоустойчивого off-line анализа и распознавания структурированных данных в виде числовых и векторных последовательностей, включающих повторяющиеся, чередующиеся и перемежающиеся информационно значимые векторы или фрагменты. В настоящей работе представлено алгоритмическое решение одной из таких ранее неизученных задач: обоснован точный полиномиальный алгоритм, который является ядром помехоустойчивого алгоритма распознавания.

Литература

[1] *Kel'manov A. V., Khamidullin S. A.* Recognizing a Quasiperiodic Sequence Composed of a Given Number of Identical Subsequences // Pattern Recognition and Image Analysis, 2000. — Vol. 10, № 1. — P. 127–142.

[2] *Kel'manov A. V., Jeon B.* A Posteriori Joint Detection and Discrimination of Pulses in a Quasiperiodic Pulse Train // IEEE Transactions on Signal Processing. — 2008. — Vol. 52, № 3. — P. 1–12.

[3] *Кельманов А. В.* Полиномиально разрешимые и NP-трудные варианты задачи оптимального обнаружения в числовой последовательности повторяющегося фрагмента // Материалы Росс. конф. «Дискретная оптимизация и исследование операций», Новосибирск: Изд-во Института математики СО РАН, 2007. — С. 46–50.
`http://math.nsc.ru/conference/door07/DOOR_abstracts.pdf`

[4] *Кельманов А. В.* О некоторых полиномиально разрешимых и NP-трудных задачах анализа и распознавания последовательностей с квазипериодической структурой // Сб. докл. 13-й Всеросс. конф. «Математические методы распознавания образов», Москва.: МАКС Пресс, 2007 — С. 261–264.

[5] `http://math.nsc.ru/~serge/qpsl/` — Система QPSLab для решения задач компьютерного анализа и распознавания числовых последовательностей с квазипериодической структурой. — 2008.

Непрерывные аппроксимации решения задачи ВЫПОЛНИМОСТЬ применительно к задачам факторизации и дискретного логарифмирования*

Дулькейт В. И., Файзуллин Р. Т., Хныкин И. Г.
r.t.faizullin@mail.ru, hig82@rambler.ru
Омск, Омский государственный технический университет

Одной из наиболее интересных задач дискретной математики является задача поиска решающего набора в задаче SAT [4]. Перспективным направлением в построении методов решения представляется сведение задачи к непрерывному поиску точек глобального минимума, ассоциированного с конъюнктивной нормальной формой (КНФ) функционала. В данной работе обосновывается выбор функционала специального вида и предлагается применить к решению системы нелинейных алгебраических уравнений, определяющих стационарные точки функционала, модифицированный метод последовательных приближений. В работе показано, что метод поддается распараллеливанию. Рассматривается схема применения метода к важным задачам криптографического анализа несимметричных шифров, в том числе для определения некоторых бит двоичного представления неизвестных сомножителей в задачах факторизации и дискретного логарифмирования больших размерностей.

Переход от КНФ к ассоциированным функционалам

Пусть $L(y) = \bigwedge_{i=1}^{M} C_i(y)$ — КНФ. Переход от задачи SAT к задаче поиска глобального минимума функционала осуществляется по формуле:

$$\min_{x\in E^n} F(x) = \min_{x\in E^n} \sum_{i=1}^{M}\prod_{j=1}^{N} Q_{i,j}(x) = 0; \qquad (1)$$

$$Q_{i,j}(x) = \begin{cases} (1-x_j)^2, & \text{если } x_j \in C_i(x), \\ x_j^2, & \text{если } \bar{x}_j \in C_i(x), \\ 1, & \text{иначе.} \end{cases}$$

Легко заметить, что $\min\limits_{x\in E^n} F(x) = 0$ соответствует достижению значения ИСТИНА на исходной КНФ.

Дифференцируя функционал по всем x_i, получим систему уравнений:

$$\sum_{\xi\in\Xi}\prod_{j\neq i}^{N} Q_{i,j}(x)\cdot x_i = \sum_{\xi\in\Lambda}\prod_{j\neq i}^{N} Q_{i,j}(x); \qquad (2)$$

$$\Xi = \{\xi, k\in\xi : x_i \text{ или } \bar{x}_i \in C_k(x)\};$$

$$\Lambda = \{\xi, k\in\xi : x_i \in C_k(x)\};$$

$$i = 1,\ldots,N.$$

Для ее решения предлагается применить метод последовательных приближений с «инерцией»:

$$\left[\sum_{p=0}^{K}\alpha_p \sum_{\xi\in\Xi}\rho_\xi \prod_{j\neq i}^{N} Q_{i,j}(x(t-p))\right]\cdot x_i(t+1) = \qquad (3)$$

$$\sum_{\xi\in\Lambda}\prod_{j\neq i}^{N} Q_{i,j}(x(t-p)) \sim A^i\cdot x_i(t+1) = B^i;$$

$$\sum_{p=0}^{K}\alpha_p = 1, \quad \alpha_p \geqslant 0, \quad \rho_\xi \geqslant 0.$$

Положив в (3) $K = 0$, $\rho_\xi = 1$, получим простой метод последовательных приближений.

Гибридизация алгоритма

Исходная КНФ преобразуется методом резолюции [3], что позволяет получить КНФ с меньшим количеством дизъюнктов и литералов, эквивалентную исходной.

Два дизъюнкта бинарно-разрешимы, если они совпадают хотя бы по одной переменной, которая входит в один дизъюнкт с отрицанием, а в другой без. Бинарно разрешимые дизъюнкты имеют вид: $x\vee P$, $\bar{x}\vee Q$.

Бинарной резольвентой называется дизъюнкт $P\vee Q$. Все возможные бинарные резольвенты с помощью операции дизъюнкции добавляются к КНФ и используются для вычисления других резольвент. Процедура ограничивается глубиной рекурсии 1. Дублирующие конъюнкты и тавтологии удаляются. Вычислительная сложность процедуры $O(n\log n)$.

Основная процедура состоит из последовательных итераций, которые совмещают метод последовательных приближений и сдвиг по антиградиенту, т. к. правая часть (2) — это хотя и градиент исходного функционала, но решения (2) — это всего лишь стационарные точки функционала. Например, если генерировать КНФ по заданной строке бит, случайно строя скобки, так, чтобы строка бит была решающим набором итоговой КНФ, то представительство литералов и их отрицаний будет одинаковым. Это означает, что ассоциированный функционал имеет «квазистационарную» точку с координатами 0,5 для каждой переменной, т. к. $A_i \approx 2B_i$. В случае же, когда представительство литералов неравное, то подобные «квазистационарные» точки могут быть произвольными. Например, генерируя случайную систему уравнений и сводя задачу к поиску решающего набора КНФ,

*

мы получаем уже существенно неравное представительство литералов. В этом случае квазистационарным точкам будут соответствовать решения неопределенных систем, получаемые из исходной системы исключением всего нескольких уравнений. Число таких точек растет экспоненциально с ростом размерности системы, и итерационная процедура поиска стационарной точки интересующей нас, как отвечающей точке минимума, практически перестает сходиться, что и подтверждается экспериментально.

Итерация состоит из двух блоков. Первый блок определяется формулой (3), используется схема Зейделя. Второй блок — реализация сдвига по антиградиенту: $x(t+1) = 2x(t) - B/A$ [1].

При приближении к решению скорость сходимости может сильно уменьшаться, одна из возможных причин этого в том, что траектория, образованная последовательными приближениями, «зацикливается» в областях локальных минимумов функционала. Метод смены траектории позволяет выйти из локального минимума с помощью формирования нового вектора приближения, который обладает свойствами не худшими, чем текущий вектор приближения, но позволяет продолжить поиск решения [1].

Распараллеливание алгоритма

Гибридный алгоритм допускает целый набор способов распараллеливания, приведем один из них. ДНФ, эквивалентная исходной КНФ, делится на две независимые части (подформулы). Векторы решений для подформул определяют точки в n-мерном пространстве. Между полученными точками проводится отрезок прямой. «Двигаясь» по этой прямой с некоторым шагом l, вычисляются векторы $x^l = (x_i^l)$:

$$x_i^l = \min(x_i^1, x_i^2) + l\big|x_i^1 - x_i^2\big|/k.$$

Вектор x^l, при котором значение функционала (1) минимально, становится новым начальным приближением для итерационной процедуры, которая запускается для функционала, ассоциированного со всей формулой.

Описанная процедура позволяет максимально приблизиться к решению. В формуле остаются невыполненными до 2% дизъюнктов. При этом около 2,5% переменных остаются неопределенными, т. е. независимо от того, какое значение они будут принимать, выполнимые скобки будут попрежнему принимать значение ИСТИНА.

Применение метода к криптографическому анализу асимметричных шифров

Подробные результаты тестирования однопроцессорной и многопроцессорной реализаций алгоритма для различных типов задач (например, [5, 6]) представлены в [1].

Была исследована схема применения метода для решения задач криптографического анализа. КНФ, ассоциированные с задачами факторизации, дискретного логарифмирования и дискретного логарифмирования на эллиптической кривой, рассматриваются в работах [2]. Оценка роста числа дизъюнктов и скобок в зависимости от размерности задачи (N) дает нам величины порядка $10N^2$ для задачи факторизации и $100N^3$, $1000N^3$ для задач дискретного логарифмирования и дискретного логарифмирования на эллиптической кривой. Метод резолюций, в применении к КНФ для факторизации, уменьшает число дизъюнктов более, чем в 2 раза, а в применении к задаче дискретного логарифмирования, позволяет уменьшить число переменных на два порядка, что приводит к относительно приемлемым цифрам для массива данных.

В рамках работы проводились исследования близости формируемых методом векторов приближений к вектору решения для задач факторизации больших размерностей. В качестве исходного материала для тестирования были выбраны по 20 независимых примеров размерностей 1024, 2048, 3072 бит. На каждой итерации метода последовательных приближений с «инерцией» проводилось сравнение компонент вектора приближений с соответствующими компонентами вектора решений с целью подсчета числа совпадающих компонент (битов). При этом производилось по 20 стартов со случайно сформированного вектора начального приближения. На рис. 1 показано поведение процентного отношения верно сформированных бит на соответствующей итерации.

Результаты показывают стабильное формирование 68% верных бит при росте размерности задачи. Максимальное (минимальное) число совпадающих бит так же стабильно: 68,3% (67,7%). При этом число верно определенных бит, отвечающих именно битам сомножителей, приблизительно равно 67,9%. Примечательно то, что результат достигается всего за 500–1000 итераций, стартуя со случайно сформиованного приближения.

На рис. 2 представлены результаты формирования среднего и максимального числа верно определенных бит при увеличении длины ключа. Отметим, что найденные переменные являются ключевыми для решения задачи, т. е. после подстановки их верных значений в исходную КНФ, формула оказывается легко разрешимой относительно оставшихся переменных.

Среднеквадратичное отклонение статистических данных не превышает 10^{-2}. Это говорит о стабильном поведении метода на данном типе задач.

Таким образом, для КНФ, эквивалентных задачам факторизации больших размерностей, метод

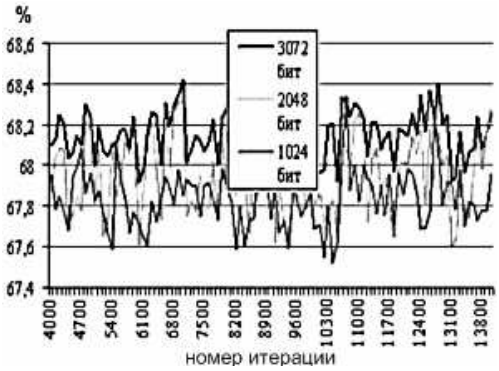


Рис. 1. Процент совпадающих (верных) бит вектора приближения и вектора решения в зависимости от итерации.

Рис. 2. Процент совпадающих (верных) бит вектора приближения и вектора решения в зависимости от размерности задачи.

формирует различные векторы приближений, каждый из которых совпадает с решением приблизительно на 68%. В качестве развития данного результата предлагается специально разработанная система тестов, которая позволяет с высокой степенью вероятности определять биты непосредственно сомножителей.

Одним из таких тестов может служить проверка обстоятельства кластеризации ненулевых строк в матрице умножения классическим «столбиком» числа p на число q в двоичной системе счисления. Каждая строка матрицы умножения состоит либо из нулей, либо из бит числа p. Аналогично, столбец матрицы умножения будет или нулевым столбцом, или столбцом, в котором записано число q. Подставляя в данную матрицу найденные векторы приближений и сравнивая строки и столбцы получившейся матрицы соответственно с p, q или нулевым вектором, можно с определенной долей вероятности выявлять значения неизвестных. Повторяя данную процедуру с различными векторами приближений можно строить методы голосования, повышающие вероятность определения верных значений неизвестных.

В таблице 1 представлены значения вероятностей верного определения значений бит для 31 независимых примеров с помощью данного теста

Таблица 1. Определение наиболее вероятных бит в сомножителях факторизуемого числа размерности 512 бит. Тест 1 — Кластеризация.

Доля тестов (из 31), в которых биты были определены, %	Число верно определенных бит	Доля верно определенных бит (из 512), %
100,00	2	0,4
96,77	1	0,2
80,65	2	0,4
77,42	2	0,4
74,19	8	1,6
70,97	9	1,8
67,74	21	4,1
64,52	50	9,8
61,29	66	12,9
Итого	161	31,6

Таблица 2. Определение наиболее вероятных бит в сомножителях факторизуемого числа размерности 512 бит. Тест 2 — Монотонность функционала.

тестируемый бит	Значение функционала при подстановке значения тестируемого бита:		разница значений функционалов
	верного значения	неверного значения	
13	261,2	263,7	−2,5
46	260,8	263,5	−2,7
73	263,0	265,0	−2,0
86	254,5	256,7	−2,2
101	255,0	257,3	−2,3
142	263,2	259,8	+3,4
217	263,7	266,9	−3,2

(при размерности чисел сомножителей 256 бит). Так, для задачи факторизации числа длиной 512 бит с вероятностью большей или равной 0,8 определяются биты 1, 13, 46, 73, 86, 101, 142, 217, 255 каждого из сомножителей.

Дополнительным тестом является то обстоятельство, что функционал (1) после подстановки верных значений указанных бит в исходную КНФ, принимает значение меньшее, чем после подстановки их неверных значений. Это позволяет практически точно определять расстановку нулей и единиц в позициях 1, 13, 46, 73, 86, 101, 217, 255. В таблице 2 приведены результаты численных экспериментов по данному тесту.

Аналогичные результаты получены и при выборках из двух и более бит.

Была исследована устойчивость верного формирования бит экспоненты в задачах дискретного логарифмирования различных размерностей. Сводные результаты приведены в таблице 3.

В качестве развития данной работы предполагается дальнейшее исследование дополнительных тестов и построение различных методов голосования для определения конкретных битов неизвестных.

Таблица 3. Определение наиболее вероятных бит экспоненты в задаче дискретного логарифмирования. Тест 3 — Устойчивость формирования бит.

Доля тестов (из 50), в которых биты были определены, %	Число верно определенных бит в экспоненте для задачи размерности:		
	48 бит	64 бит	88 бит
82	0	1	0
80	0	0	0
78	0	0	1
76	0	0	0
74	1	0	0
72	0	0	1
70	0	1	0
68	0	0	1
66	4	1	2
64	0	2	1
62	3	3	5
60	3	6	8
58	5	11	5
56	12	11	13
54	6	12	20
52	9	11	19
50	5	5	12
Итого	48	64	88

Выводы

1. Разработана модификация метода последовательных приближений с «инерцией», и обоснованы методики, равномерно улучшающие сходимость метода на всех типах задач «ВЫПОЛНИМОСТЬ». Предложены способы распараллеливания алгоритма.

2. Исследована применимость метода последовательных приближений с «инерцией» к задачам криптоанализа асимметричных шифров.

3. Для КНФ, эквивалентных задаче факторизации (с соблюдением всех условий криптостойкости RSA) размерностью до 72 бит, были получены точные решения. При этом эффективность предложенного метода превосходит многие известные нам решатели задачи «ВЫПОЛНИМОСТЬ». Приближения, формируемые методом, более чем на 68% совпадают с решением, независимо от размерности задачи (до 3072 бит включительно). Разработана система дополнительных тестов, позволяющая с высокой долей вероятности определять конкретные биты сомножителей в задаче факторизации.

Полученные результаты могут поставить под сомнение криптографическую стойкость алгоритма RSA, т. к. распараллеливание по вариантам и выбор тех вариантов расстановки нулей и единиц в указанных позициях, при которых значение функционала минимально, позволяет практически точно определять биты сомножителей с указанными номерами.

Литература

[1] *Дулькейт В.И., Файзуллин Р.Т., Хныкин И.Г.* Минимизация функционалов, ассоциированных с задачами криптографического анализа // Дифференциальные уравнения. Функциональные пространства. Теория приближений. Межд. конф., посв. 100-летию со дня рождения С. Л. Соболева, Новосибирск: Ин-т мат-ки СО РАН, 2008. — С. 484-485.

[2] *Дулькейт В. И., Файзуллин Р. Т., Хныкин И. Г.* Сведение задач криптоанализа асимметричных шифров к решению ассоциированных задач ВЫПОЛНИМОСТЬ // Всеросс. конф. ММРО-13, М.: МАКС Пресс, 2007. — С. 249–251.

[3] *Хныкин И. Г.* Модификации КНФ, эквивалентным задачам криптоанализа асимметричных шифров методом резолюции // Информационные технологии моделирования и управления. — 2007. № 2. — С. 328–337.

[4] *Cook S. A.* The complexity of theorem proving procedures // Proceedings of the Third Annual ACM Symposium on Theory of Computing, 1971. — Pp. 151–158.

[5] `http://www.lri.fr/~simon/contest05/results/` — SAT 2005 Competition results — 2005.

[6] `www.satlive.org` — SAT Live! — 2005.

Об асимптотически оптимальном построении элементарных классификаторов*

Дюкова Е. В., Инякин А. С., Колесниченко А. С., Нефёдов В. Ю.
edjukova@mail.ru, andre_w@mail.ru, whestt@gmail.com, nefedov85@mail.ru
Москва, Вычислительный центр РАН

Представлены результаты, полученные авторами в области синтеза асимптотически оптимальных алгоритмов построения тупиковых покрытий булевых и целочисленных матриц. В распознавании образов рассматриваемая задача возникает при конструировании логических процедур распознавания и классификации на этапе поиска информативных фрагментов признаковых описаний объектов (элементарных классификаторов) и может быть сформулирована как задача преобразования нормальных форм логических функций.

Один из подходов к решению задач распознавания и классификации сводится к логическому (комбинаторному) анализу исходных признаковых описаний объектов. Рассматриваемый подход имеет целый ряд достоинств, к числу которых, прежде всего, следует отнести возможность получения результата при отсутствии сведений о функциях распределения и при наличии малых обучающих выборок. Не требуется также задание метрики в пространстве описаний объектов. В данном случае для каждого признака определяется бинарная функция близости между его значениями, позволяющая различать объекты и их подописания. Особенно эффективен логический подход в случае дискретной информации низкой значности, например бинарной [1, 3, 8].

Вместо построения адекватной математической модели предметной области строятся и используются так называемые эвристические модели, выражающие по сути дела общий принцип «если описания объектов похожи, то и ответы для них похожи». Поскольку описание нового объекта, как правило, не совпадает полностью с описанием ни одного из обучающих объектов, то вопрос о классификации распознаваемого объекта решается на основе сравнения отдельных фрагментов его описания с соответствующими фрагментами описаний обучающих объектов.

Возникает конкретная фундаментальная задача построения фрагментов описаний объектов, обладающих экстремальными в том или ином смысле свойствами. Обычно такие фрагменты содержат определенную информацию о классах, например позволяют различать объекты из разных классов или отличать данный объект от объектов из других классов. Могут быть предъявлены и другие, более сложные требования информативности. Искомые фрагменты играют роль элементарных классификаторов, обеспечивая корректность распознающего алгоритма на обучающей выборке. Элементарные классификаторы, как правило, имеют содержательное описание в терминах той прикладной области, в которой решается задача, и поэтому полученные результаты распознавания также легко интерпретируются.

Комбинаторные методы привели к появлению целого класса сложно устроенных эвристик, называемых логическими процедурами распознавания. Имеются в виду прежде всего классические модели тестовых алгоритмов, алгоритмы голосования по представительным наборам (алгоритмы типа «Кора»), а также новые модели — алгоритмы голосования по антипредставительным наборам и покрытиям классов. Сюда же можно отнести и алгоритмы распознавания, основанные на построении решающих деревьев. Перечисленные алгоритмы успешно применяются при решении практических задач в таких областях, как анализ социологической информации, медицинская диагностика и прогнозирование, геологическое и техническое прогнозирование, кредитный скоринг, анализ финансовых рынков и др.

В силу большой размерности исходного пространства признаков при реализации рассматриваемого подхода возникают сложности вычислительного характера, связанные с наличием большого перебора. Формирование информативных фрагментов описаний обучающих объектов, как правило, приводит к необходимости решать известные своей трудоемкостью задачи построения покрытий булевых и целочисленных матриц, которые также могут быть сформулированы как задачи построения нормальных форм логических функций. В связи с чем, важными являются вопросы, связанные с построением эффективных процедур поиска покрытий булевых и целочисленных матриц.

Особую сложность вызывает задача построения тупиковых покрытий целочисленной матрицы [2, 4–11]. Данная задача может быть также сформулирована как задача построения максимальных конъюнкций двузначной логической функции, заданной либо множеством нулей, либо конъюнктивной нормальной формой (к. н. ф.). Поиски эффективных алгоритмов её решения ведутся с середины 1950-х годов [12].

*Работа выполнена при поддержке РФФИ проект № 07-01-00516, гранта Президента РФ по поддержке ведущих научных школ НШ № 5294.2008.1 и гранта Президента РФ по поддержке молодых кандидатов наук МК № 6500.2008.9.

Пусть M_{mn}^k — множество всех матриц размера $m \times n$ с элементами из $\{0, \ldots, k-1\}$, $k \geqslant 2$; E_k^r, $k \geqslant 2$, $r \leqslant n$, — множество всех наборов вида $(\sigma_1, \ldots, \sigma_r)$, где $\sigma_i \in \{0, \ldots, k-1\}$.

Рассмотрим $\sigma \in E_k^r$, $\sigma = (\sigma_1, \ldots, \sigma_r)$. Через $Q_p(\sigma)$, $p = 1, \ldots, r$, обозначим множество всех наборов $(\beta_1, \ldots, \beta_r)$ в E_k^r таких, что $\beta_p \neq \sigma_p$ и $\beta_j = \sigma_j$ при $j \in \{1, \ldots, r\} \setminus \{p\}$.

Пусть $L \in M_{mn}^k$. *Тупиковым σ-покрытием* матрицы L называется набор H из r различных столбцов этой матрицы такой, что подматрица L^H матрицы L, образованная столбцами набора H, обладает следующими двумя свойствами: 1) L^H не содержит строку σ; 2) если $p \in \{1, \ldots, r\}$, то L^H содержит хотя бы одну строку из множества $Q_p(\sigma)$. Если выполнено только условие 1), то набор столбцов H называется *σ-покрытием* матрицы L. Подматрица матрицы L, имеющая с точностью до перестановки строк вид

$$\begin{bmatrix} \beta_1 & \sigma_2 & \sigma_3 & \ldots & \sigma_{r-1} & \sigma_r \\ \sigma_1 & \beta_2 & \sigma_3 & \ldots & \sigma_{r-1} & \sigma_r \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \sigma_1 & \sigma_2 & \sigma_3 & \ldots & \sigma_{r-1} & \beta_r \end{bmatrix},$$

где $\beta_p \neq \sigma_p$ для $p = 1, \ldots, r$, называется *σ-подматрицей*.

Таким образом, H является тупиковым σ-покрытием тогда и только тогда, когда L^H не содержит строку σ и содержит σ-подматрицу. Понятие тупикового $(0, \ldots, 0)$-покрытия булевой матрицы совпадает с хорошо известным понятием неприводимого покрытия булевой матрицы. Отметим, что $(0, \ldots, 0)$-подматрица булевой матрицы является единичной (перестановочной) подматрицей. Единичную подматрицу назовем максимальной, если она не содержится в других единичных подматрицах.

Положим $B_r(L, \sigma)$, $\sigma \in E_k^r$, — множество всех тупиковых σ-покрытий матрицы L,

$$B_1(L) = \bigcup_{r=1}^{n} \bigcup_{\sigma \in E_k^r} B_r(L, \sigma).$$

Отметим, что задача поиска покрытий из $B_1(L)$ может быть сформулирована как задача поиска максимальных конъюнкций двузначной функции k-значной логики, заданной множеством нулей.

Через $R(\sigma)$ обозначим множество наборов $(\beta_1, \ldots, \beta_r)$ в E_k^r таких, что $\beta_j \neq \sigma_j$ при $j \in \{1, \ldots, r\}$. Набор столбцов H матрицы L называется *$R(\sigma)$-покрытием*, если в подматрице L^H матрицы L, образованной столбцами набора H, нет ни одной строки из $R(\sigma)$. Набор столбцов H матрицы L, являющийся $R(\sigma)$-покрытием, называется *тупиковым $R(\sigma)$-покрытием*, если L^H содержит подматрицу, имеющую с точностью до перестановки строк вид

$$\begin{bmatrix} \sigma_1 & \beta_{12} & \beta_{13} & \ldots & \beta_{1r-1} & \beta_{1r} \\ \beta_{21} & \sigma_2 & \beta_{23} & \ldots & \beta_{2r-1} & \beta_{2r} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \beta_{r1} & \beta_{r2} & \beta_{r3} & \ldots & \beta_{rr-1} & \sigma_r \end{bmatrix},$$

где $\beta_{ip} \neq \sigma_p$ при $i, p = 1, \ldots, r$, $i \neq p$. Такая подматрица называется $R(\sigma)$-подматрицей. Нетрудно видеть, что если $k = 2$ и $\sigma = (1, \ldots, 1)$, то понятие (тупикового) $R(\sigma)$-покрытия совпадает с понятием (неприводимого) покрытия. Аналогом единичной подматрицы является $R(\sigma)$-подматрица.

Положим $B_r(L, R(\sigma))$, $\sigma \in E_k^r$, — множество всех тупиковых $R(\sigma)$-покрытий матрицы L,

$$B_2(L) = \bigcup_{r=1}^{n} \bigcup_{\sigma \in E_k^r} B_r(L, R(\sigma)).$$

Отметим что задача поиска покрытий из $B_2(L)$ может быть сформулирована как задача поиска максимальных конъюнкций двузначной функции k-значной логики, заданной к. н. ф.

Наиболее широкую область практического применения имеют методы построения множества $P(L)$ всех неприводимых покрытий (или тупиковых $(0, \ldots, 0)$-покрытий) булевой матрицы L (методы построения сокращенной дизъюнктивной нормальной формы монотонной булевой функции, заданной конъюнктивной нормальной формой).

Далее для удобства изложения под обозначением $B(L)$ понимается либо $B_1(L)$, либо $B_2(L)$, либо $P(L)$ (в зависимости от решаемой задачи).

Для почти всех матриц L из M_{mn}^k число покрытий из $B(L)$ растет экспоненциально с ростом размера матрицы, поэтому эффективность алгоритмов построения покрытий из $B(L)$ имеет смысл оценивать в терминах полиномиальной задержки [14].

Пусть $Q(L)$ — конечная совокупность наборов столбцов матрицы L, содержащая $B(L)$. Предполагается, что каждый набор из $Q(L)$ не содержит одинаковых столбцов, и некоторые наборы столбцов в $Q(L)$ могут встречаться более одного раза.

Пусть алгоритм A строит покрытия из $B(L)$ путем последовательного просмотра всех наборов из $Q(L)$. При этом каждый набор из $Q(L)$ просматривается столько раз, сколько раз он встречается в $Q(L)$. Таким образом, на каждом шаге алгоритма A строится некоторый набор столбцов H из $Q(L)$ и проверяется принадлежность H множеству $B(L)$. Очевидно, такая проверка требует просмотра не более qm элементов матрицы L (здесь и далее $q = \min(m, n)$). Число шагов алгоритма A обозначим через $N_A(L)$.

Нас будет интересовать вычислительная сложность алгоритма A в типичном случае (для почти всех матриц из M_{mn}^k при $n \to \infty$).

Будем говорить, что алгоритм A строит $Q(L)$ *с полиномиальной задержкой*, если на каждом шаге выполняется не более $d(m,n)$ элементарных операций и $d(m,n)$ ограничено сверху полиномом от m и n. При этом под элементарной операцией понимается просмотр одного элемента матрицы L.

Алгоритм A назовем *асимптотически оптимальным*, если A строит $Q(L)$ с полиномиальной задержкой, и для почти всех матриц L из M_{mn}^k при $n \to \infty$ величина $N_A(L)$ асимптотически равна числу покрытий из $B(L)$.

В [4, 5, 6, 8] рассмотрен случай, когда число строк m булевой матрицы L имеет более низкий порядок роста, чем число столбцов n, при условии, что $n \to \infty$. Для этого случая построен асимптотически оптимальный алгоритм поиска неприводимых покрытий (алгоритм АО1). Данный алгоритм строит с задержкой, не превосходящей $O(mn)$, приближённое решение, в качестве которого рассматривается совокупность всех наборов столбцов, содержащих единичные подматрицы. Каждый такой набор алгоритм АО1 строит столько раз, сколько единичных подматриц он содержит. Показано, что если $m^\alpha \leqslant n \leqslant 2^{m^\beta},\ \alpha > 1,\ \beta < 1$, то при $n \to \infty$ число шагов данного алгоритма, равное числу единичных подматриц, почти всегда (для почти всех булевых матриц размера $m \times n$) асимптотически равно мощности $P(L)$. При конструировании алгоритма АО1 в качестве приближенного решения может быть рассмотрена совокупность наборов столбцов, содержащих максимальные единичные подматрицы (каждый такой набор столбцов строится столько раз, сколько максимальных единичных подматриц он содержит). Тогда алгоритм будет делать меньшее число шагов и работать с задержкой не превосходящей $O(qmn)$. Данная модификация алгоритма АО1 обычно используется при его реализации на ЭВМ.

В [7] построен алгоритм, основанный на переборе с задержкой не превосходящей $O(qm^2n)$ неприводимых покрытий матрицы L (алгоритм АО2). В данном случае элементами $Q(L)$ являются наборы из $P(L)$. Однако алгоритм АО2 строит каждый набор длины r из $P(L)$ столько раз, сколько единичных подматриц порядка r этот набор содержит. Из сказанного выше следует, что указанный недостаток не существенен при $m^\alpha \leqslant n \leqslant 2^{m^\beta}$, $\alpha > 1,\ \beta < 1$ (в этом случае число шагов алгоритма АО2, равное числу единичных подматриц, порождающих неприводимые покрытия, почти всегда при $n \to \infty$ асимптотически равно мощности $P(L)$).

Отметим, что проверка на повторяемость построенного на очередном шаге набора столбцов в алгоритмах АО1 и АО2 требует просмотра не более $O(qm)$ элементов матрицы L.

В [5, 6, 8] построены модификации алгоритма АО1 на случай поиска покрытий из $B_1(L)$ и $B_2(L)$, асимптотически оптимальные соответственно при условиях $m^\alpha \leqslant n \leqslant k^{m^\beta},\ \alpha > 1,\ \beta < 1,\ n \to \infty$, и при условиях $m^\alpha \leqslant n \leqslant d^{m^\beta},\ d = k/(k-1),\ \alpha > 1,\ \beta < 1,\ n \to \infty$.

В [13] построен алгоритм СМС, являющийся модификацией алгоритма АО2 для поиска покрытий из $B_2(L)$. Данный алгоритм является асимптотически оптимальным при условиях $m^\alpha \leqslant n \leqslant d^{m^\beta}$, $d = k/(k-1),\ \alpha > 1,\ \beta < 1,\ n \to \infty$ и работает с задержкой, не превосходящей $O(qu^2n)$. У алгоритма СМС, также как и у алгоритма АО2, есть «лишние» шаги, возникающие из-за того, что набор столбцов из $B_r(L, R(\sigma))$ может содержать несколько одинаковых $R(\sigma)$-подматриц. В отличие от алгоритма АО2, алгоритм СМС содержит ещё и другие «лишние» шаги, на которых не строятся покрытия из $B_2(L)$.

В [10] построен алгоритм поиска тупиковых покрытий матрицы из M_{mn}^k (алгоритм ОПТ+), асимптотически оптимальный при условиях $m^\alpha \leqslant n \leqslant k^{m^\beta},\ \alpha > 1,\ \beta < 1,\ n \to \infty$. Данный алгоритм основан на переборе с задержкой, не превосходящей $O(qmn(m+q))$, наборов столбцов матрицы, содержащих σ-подматрицы и удовлетворяющих некоторым дополнительным условиям. Результаты численных экспериментов со случайными матрицами при различных соотношениях между m и n показали, что практически во всех случаях алгоритм ОПТ+ делает существенно меньшее число шагов и поэтому работает существенно быстрее по сравнению с целочисленными модификациями алгоритмов АО1 и АО2, а также другими известными алгоритмами поиска тупиковых покрытий целочисленной матрицы. Например, при $m = 30$ и $n = 150$ алгоритм ОПТ+ имеет в среднем 4% лишних шагов, алгоритмы АО1 и АО2 при тех же условиях имеют 90% лишних шагов; при $m = 10$ и $n = 300$ соответственно — 1% лишних шагов и 50% лишних шагов. При этом в первом случае алгоритм ОПТ+ затрачивает в 10–15 раз меньше времени, чем алгоритмы АО1 и АО2, во втором случае — в 4–6 раз.

Менее исследован случай, когда $n \leqslant m$. В [2] построен алгоритм поиска тупиковых покрытий целочисленной матрицы с элементами из $\{0, \ldots, k-1\}$, $k \geqslant 2$, с полиномиальной задержкой $O(mn^2)$ на каждом шаге и такой, что логарифм по основанию k числа его шагов при $n \to \infty$ почти всегда асимптотически равен логарифму по основанию k числа тупиковых покрытий.

Работу каждого из перечисленных выше алгоритмов можно представить как односторонний обход дерева решений.

Например, алгоритмы АО1 и АО2 поиска неприводимых покрытий булевой матрицы L строят дерево решений, вершинам которого соответствуют пары (L_H, H), где H — набор столбцов матрицы L, L_H — подматрица матрицы L, образованная строками, не «покрытыми» набором H, и некоторыми из столбцов матрицы L, не входящими в набор H. Корневая вершина остается «пустой» (имеет вид $(L, \varnothing)$). Для висячих вершин проверяется принадлежность H множеству неприводимых покрытий матрицы L. Шагу алгоритма соответствует построение очередной висячей вершины. Переход по ветви дерева решений от одной вершины к следующей осуществляется путем добавления к набору H некоторого столбца матрицы L. При этом предполагается, что число элементарных операций, выполняемых на каждом шаге, ограничено сверху полиномом от размера матрицы L.

В алгоритме АО1 каждая вершина дерева решений порождается единичной подматрицей матрицы L. Пусть единичная подматрица матрицы L образована единичными элементами $a_{i_1 j_1}, \ldots, a_{i_r j_r}$, где $j_1 < \ldots < j_r$. Тогда эта подматрица порождает вершину, которой соответствует пара (L'_H, H), где H — набор столбцов с номерами $j_1, \ldots, j_r$, а L'_H образована строками, не покрытыми столбцами из H, и столбцами с номерами большими j_r, не покрытыми строками с номерами $i_1, \ldots, i_r$. Висячим вершинам соответствуют наборы столбцов, содержащие максимальные единичные подматрицы, т. е. единичные подматрицы, не содержащиеся в других единичных подматрицах.

В алгоритме АО2 каждая вершина дерева решений также порождается единичной подматрицей матрицы L. Пусть единичная подматрица матрицы L образована единичными элементами $a_{i_1 j_1}, \ldots, a_{i_r j_r}$, где $j_1 < \ldots < j_r$. Тогда эта подматрица порождает вершину, которой соответствует пара (L''_H, H), где H — набор столбцов с номерами $j_1, \ldots, j_r$, а L''_H получается из L'_H выбрасыванием так называемых охватывающих строк. Висячим вершинам соответствуют неприводимые покрытия матрицы L.

Аналогичные результаты получены для задач построения нормальных форм двузначной логической функции, заданной к. н. ф.

Литература

[1] *Баскакова Л. В., Журавлёв Ю. И.* Модель распознающих алгоритмов с представительными наборами и системами опорных множеств // Ж. вычисл. матем. и матем. физ. — 1981. — Т. 21, № 5. — С. 1264–1275.

[2] *Демьянов Е. А., Дюкова Е. В.* О построении тупиковых покрытий целочисленной матрицы // Ж. вычисл. матем. и матем. физ. — 2007. — Т. 47, № 3. — С. 539–547.

[3] *Дмитриев А. И., Журавлёв Ю. И., Кренделев Ф. П.* Асимптотически оптимальные тестовые алгоритмы в задачах распознавания // Дискретный анализ. Новосибирск: ИМ СО АН СССР. — 1966. — Вып. 7. — С. 3–17.

[4] *Дюкова Е. В.* Асимптотически оптимальные тестовые алгоритмы в задачах распознавания // Пробл. кибернетики. М.: Наука. — 1982. — Вып. 39. — С. 165–199.

[5] *Дюкова Е. В.* О сложности реализации некоторых процедур распознавания // Ж. вычисл. матем. и матем. физ. — 1987. — Т. 27, № 1. — С. 114–127.

[6] *Дюкова Е. В.* Алгоритмы распознавания типа Кора: сложность реализации и метрические свойства // Распознавание, классификация, прогноз (матем. методы и их применение). — 1989. — Вып. 2. — С. 99–125.

[7] *Дюкова Е. В.* О сложности реализации дискретных (логических) процедур распознавания // Ж. вычисл. матем. и матем. физ. — 2004. — Т. 44, № 3. — С. 550–572.

[8] *Дюкова Е. В., Журавлёв Ю. И.* Дискретный анализ признаковых описаний в задачах распознавания большой размерности // Ж. вычисл. матем. и матем. физ. — 2000. — Т. 40, № 8. — С. 1264–1278.

[9] *Дюкова Е. В., Инякин А. С.* О процедурах классификации, основанных на построении покрытий классов // Ж. вычисл. матем. и матем. физ. — 2003. — Т. 43, № 12. — С. 1910–1921.

[10] *Дюкова Е. В., Инякин А. С.* Об асимптотически оптимальном построении тупиковых покрытий целочисленной матрицы // Математические вопросы кибернетики. М.: Физматлит. — 2008. — Вып. 17. — С. 247–262.

[11] *Дюкова Е. В., Песков Н. В.* Поиск информативных фрагментов описаний объектов в дискретных процедурах распознавания // Ж. вычисл. матем. и матем. физ. — 2002. — Т. 42, № 5. — С. 741–753.

[12] *Чегис И. А., Яблонский С. В.* Логические способы контроля электрических схем // Труды Матем. ин-та им. В.А. Стеклова АН СССР. — 1958. — Т. 51. — С. 270–360.

[13] *Djukova E. V., Nefedov V. Y.* On Complexity of Logical Data Analysis in Recognition Problems // 9th International Conference “Pattern Recognition and Image Analysis: New Information Technologies” (PRIA-9-2008): Conference Proceedings. Vol. 1, Nizhni Novgorod: Диалог Культур, 2008. — С. 85–88.

[14] *Jonson D. S., Yannakakis M., Papadimitriou C. H.* On General All Maximal Independent Sets // Information processing Letters. — 1988. — V. 27. — Pp. 119–123.

О сложности преобразования нормальных форм характеристических функций классов*

Дюкова Е. В., Нефёдов В. Ю.
edjukova@mail.ru, nefedov85@mail.ru
Москва, Вычислительный центр РАН

Получены новые результаты, касающиеся вычислительной сложности процедур распознавания, основанных на построении нормальных форм характеристических функций классов.

Введение

При конструировании логических процедур распознавания используется аппарат дискретной математики, в частности, методы преобразования нормальных форм логических функций, являющихся характеристическими функциями классов [8, 9, 11, 12, 13]. При решении прикладных задач большой размерности возникают трудности вычислительного характера. Как правило, требуется построить дизъюнктивную нормальную форму (д. н. ф.) из допустимых или максимальных конъюнкций двузначной функции k-значной логики $F(x_1, \ldots, x_n)$, заданной конъюнктивной нормальной формой(к. н. ф.). Наибольшую вычислительную сложность представляет поиск всех максимальных конъюнкций функции F (т. е. построение её сокращённой д. н. ф.). Данная задача может быть сформулирована как задача построения всех тупиковых покрытий булевой или целочисленной матрицы. Исследованию её вычислительной сложности в типичном случае и построению асимптотически оптимальных алгоритмов посвящён ряд работ [2–7, 10, 14, 15]. В этих работах асимптотически оптимальные алгоритмы построены для случаев:

1) F — монотонная булева функция;
2) F задана совершенной к. н. ф.;
3) F задана к. н. ф. произвольного вида.

Случай 3) рассмотрен в [3, 6]. В [6] построен алгоритм CMC (Constructing Maximal Conjunctions), работающий с полиномиальной задержкой на каждом шаге и с числом шагов почти всегда (для почти всех к. н. ф. рассматриваемого вида) при $n \to \infty$ асимптотически равным числу максимальных конъюнкций.

В настоящей работе алгоритм CMC, построенный в [6], усовершенствован за счёт некоторых дополнительных построений на каждом шаге. Эти построения не увеличили оценку сложности шага, равную $O(qu^2n)$, где u — число элементарных дизъюнкций в к. н. ф., $q = \min(u, n)$, однако время счёта значительно сократилось (из-за уменьшения числа «лишних» шагов). Тестирование проводилось на случайных матрицах.

*Работа выполнена при поддержке проекта РФФИ №07-01-00516 и гранта Президента РФ по поддержке ведущих научных школ НШ №5294.2008.1.

Основные результаты

Пусть E_k^n — множество наборов вида $(\sigma_1, \ldots, \sigma_n)$, где $\sigma_i \in \{0, \ldots, k-1\}$ и пусть двузначная функция $F(x_1, \ldots, x_n)$ определена на наборах из E_k^n и принимает значения 1 и 0 соответственно на подмножествах наборов N_F и $N_{\bar{F}}$. Положим $x^\sigma = 1$, если $x = \sigma$, и $x^\sigma = 0$, если $x \neq \sigma$, $x, \sigma \in \{0, \ldots, k-1\}$. Обычным образом введём понятия элементарной конъюнкции и элементарной дизъюнкции.

Элементарной конъюнкцией (ЭК) над переменными $x_1, \ldots, x_n$ назовём выражение вида $x_{j_1}^{\sigma_1} \wedge \ldots \wedge x_{j_r}^{\sigma_r}$, где $x_{j_i} \in \{x_1, \ldots, x_n\}$ при $i = 1, \ldots, r$ и $x_{j_q} \neq x_{j_t}$ при $t, q \in \{1, \ldots, r\}$, $t \neq q$. ЭК принимает значение 1 тогда и только тогда, когда каждый её множитель равен 1. Через N_B будем обозначать интервал истинности ЭК B.

Элементарной дизъюнкцией (ЭД) над переменными $x_1, \ldots, x_n$ назовём выражение вида $x_{j_1}^{\sigma_1} \vee \ldots \vee x_{j_p}^{\sigma_p}$, где $x_{j_i} \in \{x_1, \ldots, x_n\}$ при $i = 1, \ldots, p$ и $x_{j_q} \neq x_{j_t}$ при $t, q \in \{1, \ldots, p\}$, $t \neq q$.

Пусть функция F задана к. н. ф. K вида

$$D_1 \wedge \ldots \wedge D_u, \tag{1}$$

где D_i, $i = 1, \ldots, u$, — ЭД над переменными $x_1, \ldots, x_n$.

Пусть B — ЭК над переменными $x_1, \ldots, x_n$, и пусть $M(B, K)$ — число дизъюнкций в к. н. ф. K, не содержащих переменных из B. ЭК B назовём *допустимой* для F, если $N_B \cap N_{\bar{F}} = \varnothing$, т. е. $M(B, K) = 0$. ЭК B назовём *неприводимой* для F, если не существует ЭК B' такой, что $N_{B'} \supset N_B$ и $M(B', K) = M(B, K)$. ЭК B назовём *максимальной* для F, если она является допустимой и неприводимой.

Через $D_C(F)$ будем обозначать сокращённую д. н. ф. функции F, то есть д. н. ф., состоящую из всех максимальных конъюнкций функции F. Рассмотрим задачу преобразования к. н. ф. K в $D_C(F)$. Нам понадобятся следующие утверждения.

Утверждение 1. *ЭК B является допустимой для F тогда и только тогда, когда каждая дизъюнкция D_i, $i \in \{1, \ldots, u\}$, содержит хотя бы один множитель из B.*

Утверждение 2. *ЭК B ранга r является неприводимой для F тогда и только тогда, когда в к. н. ф. K можно указать r дизъюнкций $D_{i_1}, \ldots, D_{i_r}$, таких, что каждая дизъюнкция содержит в точно-*

сти один множитель из B, и если $r > 1$, $p, q \in \{i_1, \ldots, i_r\}$, $p \neq q$, то дизъюнкции D_p и D_q содержат разные множители из B.

Положим $B(F)$ — множество всех максимальных конъюнкций функции F. Пусть $Q(F)$ — конечная совокупность элементарных конъюнкций над переменными $x_1, \ldots, x_n$, содержащая $B(F)$. Предполагается, что некоторые элементарные конъюнкции могут встречаться в $Q(F)$ более одного раза.

Пусть алгоритм A строит конъюнкции из $B(F)$ путём последовательного просмотра всех конъюнкций из $Q(F)$. При этом каждая конъюнкция из $Q(F)$ просматривается столько раз, сколько раз она встречается в $Q(F)$. Таким образом, на каждом шаге алгоритма строится некоторая конъюнкция B из $Q(F)$ и проверяется принадлежность B к $B(F)$. Число шагов алгоритма A обозначим через $N_A(K)$.

Нас будет интересовать вычислительная сложность алгоритма A в типичном случае (для почти всех к. н. ф. вида (1) при $n \to \infty$).

Будем говорить, что алгоритм A строит $Q(F)$ с *полиномиальной задержкой*, если на каждом шаге выполняется не более $d(u, n)$ элементарных операций и $d(u, n)$ ограничено сверху полиномом от u, n. При этом под элементарной операцией понимается просмотр одного символа переменной в к. н. ф.

Алгоритм A назовём *асимптотически оптимальным*, если A строит $Q(F)$ с полиномиальной задержкой, и для почти всех к. н. ф. вида (1) величина $N_A(K)$ асимптотически равна при $n \to \infty$ числу конъюнкций из $B(F)$.

Заметим, что если B — ЭК, то согласно утверждениям 1 и 2 для проверки условия $B \in B(F)$ требуется полиномиальное время.

Приведём описание предлагаемого в данной работе асимптотически оптимального алгоритма СМС+ построения д. н. ф. $D_C(F)$. Алгоритм СМС+ основан на построении на каждом шаге так называемой «тупиковой» конъюнкции для F и отборе среди построенных «тупиковых» конъюнкций допустимых конъюнкций.

ЭК B называется *тупиковой* для F, если она является неприводимой для F, и не существует неприводимой для F конъюнкции B' такой, что $N_B \supset N_{B'}$.

Утверждение 3. *Если ЭК является максимальной для F, то она является тупиковой для F.*

Обратное утверждение не верно (тупиковая конъюнкция может не быть допустимой). Таким образом, задача построения максимальных конъюнкций есть задача построения допустимых тупиковых конъюнкций.

Работу алгоритма СМС+ удобно представить как процесс построения дерева решений Δ_F. Корню дерева соответствует пустое множество, а каждой внутренней вершине — пара вида $(B, \tilde{K})$, где B — некоторая неприводимая конъюнкция для F, а $\tilde{K}$ — к. н. ф., которая либо совпадает с K, либо получена из K вычеркиванием некоторых дизъюнкций и некоторых слагаемых в оставшихся дизъюнкциях. При этом разным вершинам могут соответствовать одинаковые конъюнкции. Тупиковые конъюнкции порождаются висячими вершинами. На каждом шаге строится ветвь дерева Δ_F. Приведём более подробное описание данного алгоритма.

Будем считать, что K не содержит одновременно дизъюнкции x^σ и $x^{\bar{\sigma}}$, и переменные в K перенумерованы в порядке следования дизъюнкций D_i (скобок), а в каждой скобке — слева направо.

Шаг 1. Применим к K правило поглощения: $(X_1 \vee X_2) \wedge X_1 = X_1$, где X_1 и X_2 — ЭД (вычеркнем «охватывающие» дизъюнкции). Получим к. н. ф. K'. Выберем первую из оставшихся дизъюнкций D' и возьмём в ней первую по порядку переменную. Пусть это будет $x_{j_1}^{\sigma_1}$. Вычеркнем из K' все дизъюнкции, содержащие $x_{j_1}^{\sigma_1}$. Из оставшихся дизъюнкций удалим переменные, встречающиеся в D', и переменные вида $X_{j_1}^{\gamma}$, где $\gamma \neq \sigma_1$. Полученную в результате к. н. ф. обозначим через $K_1^{(1)}$. Положим $B_1^{(1)} = x_{j_1}^{\sigma_1}$ и построим вершину первого яруса первой ветви вида $(B_1^{(1)}, K')$. Возможны три следующих случая.

1. Все скобки вычеркнуты, тогда согласно утверждениям 1 и 2 конъюнкция $B_1^{(1)}$ — тупиковая и допустимая. Первая ветвь построена. Висячей вершине соответствует максимальная конъюнкция. Переходим к шагу 2.
2. Не все скобки вычеркнуты, но вычеркнуты все переменные в скобках, тогда $B_1^{(1)}$ — тупиковая конъюнкция, не являющаяся допустимой. Первая ветвь построена. На данном шаге не найдена максимальная конъюнкция. Переходим к шагу 2.
3. Не все скобки вычеркнуты, и в них ещё остались переменные, тогда процесс построения первой ветви продолжается. К к. н. ф. $K_1^{(1)}$ применяется правило поглощения и строится вершина второго яруса вида $(B_2^{(1)}, K_1^{(1)})$, где $B_2^{(1)} = x_{j_1}^{\sigma_1} \wedge x^\sigma$, x^σ — переменная, первая по порядку в $K_1^{(1)}$. Построение первой ветви продолжается до тех пор, пока не возникнет один из случаев 1 или 2.

Шаг $i + 1$, $i \geqslant 1$. Пусть на шаге i построена ветвь с висячей вершиной $(B_r^{(i)}, K_{r-1}^{(i)})$, где $B_r^{(i)} = x_{j_1}^{\sigma_1} \wedge \ldots \wedge x_{j_r}^{\sigma_r}$ и $K_{r-1}^{(i)} = K'$ при $r = 1$.

1. Если $r = 1$ и $x_{j_1}^{\sigma_1}$ не является переменной с максимальным номером, то строится ветвь, исходящая из вершины (x^σ, K') первого яруса, где x^σ непосредственно следует за $x_{j_1}^{\sigma_1}$ в K' в соответ-

ствии с установленным порядком (рассуждения те же, что и на шаге 1, с заменой $x_{j_1}^{\sigma_1}$ на x^σ, однако при построении следующей вершины рассматриваются только те переменные, номера которых больше номера переменной x^σ).

2. Если $r = 1$ и $x_{j_1}^{\sigma_1}$ является переменной с максимальным номером, то алгоритм заканчивает работу.
3. Если $r > 1$, то возвращаемся на предыдущий ярус (ярус с номером $r-1$) построенной на шаге i ветви в вершину $(B_{r-1}^{(i)}, K_{r-2}^{(i)})$, где $B_{r-1}^{(i)} = x_{j_1}^{\sigma_1} \wedge \ldots \wedge x_{j_{r-1}}^{\sigma_{r-1}}$. При этом, если $K_{r-1}^{(i)}$ содержит переменные, следующие за $x_{j_r}^{\sigma_r}$, то выбираем из них первую по порядку переменную x^σ. Начинаем строить ветвь, исходящую из вершины $(B_{r-1}^{(i)} \wedge x^\sigma, K_{r-1}^{(i)})$ (рассуждения те же, что и на шаге 1, с заменой K' на $K_{r-1}^{(i)}$ и $x_{j_1}^{\sigma_1}$ на x^σ, однако при построении следующей вершины рассматриваются только те переменные, номера которых больше номера переменной x^σ).
4. Если в $K_{r-1}^{(i)}$ нет переменных, следующих за $x_{j_r}^{\sigma_r}$, и $r > 2$, то возвращаемся на ярус с номером $r-2$ построенной на шаге i ветви в вершину $(B_{r-2}^{(i)}, K_{r-3}^{(i)})$, где $B_{r-2}^{(i)} = x_{j_1}^{\sigma_1} \wedge \ldots \wedge x_{j_{r-2}}^{\sigma_{r-2}}$, и берём переменную x^σ, непосредственно следующую за переменной $x_{j_{r-1}}^{\sigma_{r-1}}$ в $K_{r-2}^{(i)}$. Очевидно, такая переменная найдётся, так как по построению $x_{j_r}^{\sigma_r}$ следует за $x_{j_{r-1}}^{\sigma_{r-1}}$ в $K_{r-2}^{(i)}$. Строится ветвь, исходящая из вершины $(B_{r-1}^{(i)} \wedge x^\sigma, K_{r-1}^{(i)})$ (рассуждения те же, что и на шаге 1, с заменой K' на $K_{r-2}^{(i)}$ и $x_{j_1}^{\sigma_1}$ на x^σ, однако при построении следующей вершины рассматриваются только те переменные, номера которых больше номера переменной x^σ).
5. Если в $K_{r-1}^{(i)}$ нет переменных, следующих за $x_{j_r}^{\sigma_r}$, и $r = 2$, то строится ветвь, исходящая из вершины первого яруса вида (x^σ, K'), где x^σ непосредственно следует за $x_{j_1}^{\sigma_1}$ в K' (рассуждения те же, что и на шаге 1, с заменой $x_{j_1}^{\sigma_1}$ на x^σ, однако при построении следующей вершины, рассматриваются только те переменные, номера которых больше номера переменной x^σ).

Рассмотренная в данной работе задача может быть сформулирована как задача поиска тупиковых покрытий целочисленной матрицы [3]. Из результатов, приведённых в [3], следует

Утверждение 4. *Если $u^\alpha \leqslant n \leqslant d^{u^\beta}$, $d = \frac{k+1}{k}$, $\alpha > 1$, $\beta < 1$, то для почти всех к. н. ф. K вида (1) число вершин в дереве Δ_F асимптотически равно числу максимальных конъюнкций.*

Утверждение 5. *Сложность шага алгоритма СМС+ не превосходит $O(qu^2n)$, где $q = \min(u, n)$.*

Из утверждений 4 и 5 следует

Утверждение 6. *Алгоритм СМС+ является асимптотически оптимальным, и его вычислительная сложность почти всегда для почти всех функций вида 1 не превосходит $O(qu^2n)|B(F)|$.*

Замечание 1. Выше было отмечено, что алгоритм СМС+ разработан на основе усовершенствования алгоритма СМС из [6] за счёт некоторых дополнительных построений на каждом шаге. Эти дополнительные построения возникают в результате применения правила поглощения.

Замечание 2. Число шагов алгоритма СМС+ можно уменьшить, если обрывать ветвь в случае, когда ещё не все скобки вычеркнуты, но в одной из скобок уже вычеркнуты все переменные. Согласно утверждению 1 все конъюнкции, которые строятся на шагах, порождаемых рассматриваемой ветвью, не будут допустимыми. Обрывать ветвь в вершине можно и в случае, когда к. н. ф., соответствующая этой вершине, содержит одновременно дизъюнкции x^σ и $x^{\bar\sigma}$.

Замечание 3. Задача может быть рассмотрена и для случая, когда символ x^σ определяется по другому правилу, например, по правилу: $x^\sigma = 0$, если $x = \sigma$, и $x^\sigma = 1$, если $x \neq \sigma,\ x,\ \sigma \in \{0, \ldots, k-1\}$.

Замечание 4. В работе [2] построен асимптотически оптимальный алгоритм АО2 поиска максимальных конъюнкций монотонной булевой функции. Данный алгоритм на каждом шаге находит за полиномиальное время максимальную конъюнкцию, однако имеет повторяющиеся шаги. Описанный в настоящей работе алгоритм СМС+ является обобщением алгоритма АО2 на случай, когда логическая функция F задана к. н. ф. вида (1). На каждом шаге СМС+ строит за полиномиальное время тупиковую конъюнкцию, которая может не являться максимальной. Такая ситуация может возникнуть, например, в случае, когда в некоторой вершине дерева решений образуется к. н. ф. вида $(x_i^{\sigma_i} \vee X) \wedge (x_i^{\bar\sigma_i} \vee X)$, где X — некоторая дизъюнкция. После выбора $x_i^{\sigma_i}$ все переменные в $(x_i^{\bar\sigma_i} \vee X)$ оказываются вычеркнутыми. Каждая тупиковая конъюнкция может быть построена неоднократно.

Численные эксперименты

Для проведения численных экспериментов алгоритмы СМС и СМС+ были реализованы на языке С++ для ЭВМ на базе процессора Intel Core 2 Duo Processor T5500 1,66 GHz. Для сравнения алгоритмов была проведена серия экспериментов на случайных к. н. ф. от n переменных из u дизъюнкций. Для каждой пары (u, n) обсчитывалось по 20 к. н. ф. По результатам счёта вычислялась статистическая эффективность каждого алгоритма по

формулам

$$SE_1 = \frac{1}{20}\sum_{i=1}^{20}\frac{|B(K_i)|}{N_A(K_i)},$$

$$SE_2 = \frac{1}{20}\sum_{i=1}^{20}\frac{N_A^*(K_i)}{N_A(K_i)},$$

где $|B(K_i)|$ — число максимальных конъюнкций функции, заданной к. н. ф. K_i, $N_A(K_i)$ — число шагов алгоритма A при работе с к. н. ф. K_i, $N_A^*(K_i)$ — число шагов алгоритма A при работе с к. н. ф. K_i, на которых была найдена максимальная конъюнкция.

Эксперименты показали, что применение правила поглощения в алгоритме СМС+ позволяет повысить статистическую эффективность и сократить время счёта. Например, при $u = 10,\ n = 100$ время сокращается примерно в два раза (с 47,241 сек. до 25,445 сек.).

Заключение

Рассматривается задача поиска максимальных конъюнкций двузначной функции k-значной логики. Данная задача возникает при построении логических процедур распознавания. Для её решения предложен и исследован асимптотически оптимальный алгоритм, являющийся модификацией алгоритма из [6]. Данная модификация позволяет в два раза уменьшить время работы алгоритма и повысить его статистическую эффективность. Обоснование асимптотической оптимальности алгоритма базируется на результатах, полученных ранее в [3].

Литература

[1] *Демьянов Е. А., Дюкова Е. В.* О построении тупиковых покрытий целочисленной матрицы // Ж. вычисл. матем. и матем. физ. — 2007. — Т. 47, № 3. — С. 539–547.

[2] *Дюкова Е. В.* О сложности реализации дискретных (логических) процедур распознавания // Ж. вычисл. матем. и матем. физ. — 2004. — Т. 44, № 3. — С. 550–572.

[3] *Дюкова Е. В., Журавлёв Ю. И.* Дискретный анализ признаковых описаний в задачах распознавания большой размерности // Ж. вычисл. матем. и матем. физ. — 2000. — Т. 40, № 8. — С. 1264–1278.

[4] *Дюкова Е. В., Инякин А. С.* Об асимптотически оптимальном построении тупиковых покрытий целочисленной матрицы // Математические вопросы кибернетики. — 2008. — Вып. 17. — С. 247–262.

[5] *Дюкова Е. В., Песков Н. В.* Построение распознающих процедур на базе элементарных классификаторов // Математические вопросы кибернетики. — 2005. — Вып. 14. — С. 57–92.

[6] *Djukova E. V., Nefedov V. Y.* On complexity of logical data analysis in recognition problems // 9th International Conference "Pattern recognition and image analysis: new information technologies" (PRIA-9-2008). Vol. 1. — Nizhni Novgorod: Диалог Культур, 2008. — Pp. 85–88.

[7] *Андреев А. Е.* Об асимптотическом поведении числа тупиковых тестов и минимальной длины теста для почти всех таблиц // Проблемы кибернетики. Вып. 41. М.: Наука. — 1984. — С. 117–141.

[8] *Гольдберг С.И.* Об одном методе распознавания образов «Совокупный антисиндром» // Вычисл. системы. Новосибирск. — 1978. — Вып. 76. — С. 83–90.

[9] *Коган А. Ю.* О дизъюнктивных нормальных формах булевых функций с малым числом нулей // Ж. вычисл. матем. и матем. физ. — 1987. — Т. 27, № 16. — С. 924–931.

[10] *Носков В. Н.* О тупиковых и минимальных тестах для одного класса таблиц // Дискретный анализ. Новосибирск.: Институт математики СО АН СССР, — 1968. — Вып. 12. — С. 924–931.

[11] *Вайнцвайг М. Н.* Алгоритм обучения распознаванию образов «Кора» // Алгоритмы обучения распознаванию образов. М.: Сов. радио. — 1973. С. 82–91.

[12] *Баскакова Л. В., Журавлёв Ю. И.* Модель распознающих алгоритмов с представительными наборами и системами опорных множеств // Ж. вычисл. матем. и матем. физ. — 1981. — Т. 21, № 5. — С. 1264–1275.

[13] *Дмитриев А. И., Журавлёв Ю. И., Кренделев Ф. П.* Асимптотически оптимальные тестовые алгоритмы в задачах распознавания // Дискретный анализ. Новосибирск: ИМ СО АН СССР. — 1966. — Вып. 7. — С. 3–17.

[14] *Дьяконов А. Г.* Кодировки и их использование при ДНФ-реализации бинарных функций // Доклады академии наук. — 2003. — Т. 391, № 2. — С. 162–165.

[15] *Gurvich V., Khachiyan L.* On generating the irredundant conjunctive and disjunctive normal forms of monotone boolean functions // Discrete Applied Mathematics. — 1999. — Vol. 96–97, № 1. — Pp. 363–373. `http://rutcor.rutgers.edu/pub/rrr/reports95/35.ps`.

Об одном методе построения приближенного решения для задачи о покрытии*

Дюкова Е. В., Сизов А. В., Сотнезов Р. М.
edjukova@mail.ru, box.sizov@gmail.com, rom.sot@gmail.com
Москва, Вычислительный центр РАН, МГУ им. М. В. Ломоносова

Рассмотрена классическая задача о покрытии множества системой его подмножеств. В распознавании образов задача о покрытии возникает при логическом анализе данных. Экспериментально исследован алгоритм с улучшенной оценкой точности.

При логическом анализе данных в распознавании возникают задачи, для решения которых применяется аппарат дискретной оптимизации. Например, при синтезе элементарных классификаторов в логических процедурах распознавания [2], при построении логического корректора на базе элементарных классификаторов [3], при выборе параметров, характеризующих представительность обучающих объектов и опорных множеств в алгоритмах вычисления оценок [4].

В настоящей работе рассматривается одна из центральных задач дискретной оптимизации — задача о покрытии множества системой его подмножеств, которая формулируется как задача целочисленного линейного программирования. Данная задача относится к классу NP-полных, в связи с чем известные алгоритмы поиска точного решения имеют экспоненциальную вычислительную сложность и малопригодны на практике. Для задач больших размерностей ищутся приближенные решения. Как правило, хорошие результаты дает «градиентный» алгоритм. Однако в ряде случаев, например, на матрицах, разреженных по числу единиц, качество решения, выдаваемого «градиентным» алгоритмом, резко ухудшается. Поэтому актуальными являются вопросы разработки быстро работающих эвристик, дающих хорошие приближенные решения для сложных задач.

В работе рассмотрен один из вариантов реализации алгоритма General из [1], дающий приближенное решение для рассматриваемой задачи с оценкой точности $d(A)$, где $d(A)$ — максимальное число блоков из последовательных единиц в строке матрицы ограничений A. Приведены результаты экспериментального исследования алгоритма General, рассмотрено соотношение реальных и теоретических оценок точности для этого алгоритма, а также проведено сравнение его работы с другими алгоритмами, в частности, генетическими.

Постановка задачи

Сформулируем задачу Z_1 о покрытии множествами как задачу целочисленного линейного программирования. Пусть есть матрица ограничений $A = (a_{ij})_{m\times n}$, w — вектор весов столбцов матрицы A. Обозначим $I = \{1,\dots,m\}$, $J = \{1,\dots,n\}$. Задача Z_1 определяется условиями:

$$\min F(x) = \sum_{j\in J} w_j x_j, \tag{1}$$

$$\sum_{j\in J} a_{ij}x_j \geqslant 1, \quad i \in I, \tag{2}$$

$$x_j \geqslant 0, \quad j \in J, \tag{3}$$

$$x_j \in \{0,1\}, \quad j \in J. \tag{4}$$

Будем также рассматривать линейную релаксацию задачи Z_1 — задачу Z_2, которая определяется условиями (1)–(3) и условием

$$x_i \in [0,1], \quad i \in I. \tag{5}$$

Описание алгоритма General

Алгоритм основан на замене задачи Z_1 задачей Z_2 с условиями (1)–(3), (5), которая может быть решена за полиномиальное время. В результате по оптимальному решению задачи Z_2 исходная матрица ограничений A задачи Z_1 преобразуется в булеву матрицу A_1 размера $m \times n$, и рассматривается новая задача Z_1 с матрицей ограничений A_1 и тем же вектором весов. Таким образом, работу алгоритма можно разбить на три этапа.

1. Решение задачи Z_2.
2. Построение матрицы A_1 по решению Z_2.
3. Решение задачи Z_1 с матрицей ограничений A_1.

Рассмотрим первый этап — решение линейной релаксации исходной задачи. Общеизвестны полиномиальные алгоритмы для решения задачи линейного программирования: среди известных — метод эллипсоидов, алгоритм Кармаркара [7] и алгоритмы внутренних точек, которые применимы для задач больших размерностей. В данной работе используется алгоритм LIPSOL (Linear Programming Solver) [8], который является одним из алгоритмов внутренних точек.

Получив оптимальное решение $(x_1^0,\dots,x_n^0)$ задачи Z_2, будем строить новую матрицу A_1. В силу свойств задачи, исходная матрица A является булевой. Каждую строку матрицы A можно разбить на блоки единиц, идущих подряд. Пусть в её

*Работа выполнена при поддержке РФФИ проект №07-01-0516, гранта Президента РФ по поддержке ведущих научных школ НШ №5294.2008.1.

i-ой строке содержится k_i блоков последовательных единиц, пусть $U_{i,t}$ — множество номеров столбцов, пересекающихся с единичными элементами t-го блока $t = 1, \ldots, k_i$. Тогда i-ая строка матрицы A_1 будет содержать всего один блок с номером t такой, что

$$\sum_{j \in U_{i,t}} x_j^0 \geqslant \frac{1}{k_i}. \tag{6}$$

Теперь рассмотрим задачу о покрытии — Z_3 с теми же весовыми коэффициентами, но с новой матрицей A_1. Матрица A_1 обладает свойством strong C1P [6], т. е. в каждой строке матрицы есть только один блок из последовательных единиц, а, как известно, для задач с такой матрицей существуют полиномиальные алгоритмы.

Третьим этапом алгоритма может стать любой точный алгоритм решения задачи дискретного программирования, но целесообразно использовать полиномиальные алгоритмы, опирающиеся на свойство strong C1P матрицы. В работе рассматривается алгоритм DOM, описанный в [6]. Вначале к матрице A_1 применяются три известных правила упрощения, т. е. матрица A_1 проверяется на наличие охватывающих строк, охватываемых столбцов и столбцов, которые охватываются несколькими столбцами соответственно. Сформулируем эти правила.

Пусть матрица $L = (b_{ij})_{m \times n}$ — произвольная булева матрица. Будем говорить, что столбец j покрывает строку i, если $b_{ij} = 1$.

Правило 1. Если в матрице L есть две строки i_1, i_2 такие, что для любого столбца j из того, что столбец j покрывает строку i_1, следует что столбец j покрывает и строку i_2, то строка i_2 охватывается строкой i_1. Исключаем строку i_2 из матрицы.

Правило 2. Если в матрице L есть два столбца j_1, j_2 такие, что $w(j_1) \geqslant w(j_2)$, и для любой строки i из того, что столбец j_1 покрывает строку i, следует, что и столбец j_2 покрывает строку i, то столбец j_1 охватывается столбцом j_2. Исключаем столбец j_1 из матрицы.

Для третьего правила нам понадобится ввести некоторые обозначения. Обозначим через $C_{\min}(i)$ столбец матрицы L, который среди всех столбцов, покрывающих строку $i = 1, \ldots, m$, имеет наименьший вес. Пусть j_1 и j_2 — столбцы матрицы L, тогда обозначим $X(j_1, j_2) = \{C_{\min}(i) \colon\ b_{ij_1}{=}1 \wedge b_{ij_2}{=}0\}$.

Правило 3. Пусть j_1, j_2 — два столбца матрицы L, покрывающих хотя бы одну общую строку. Тогда, если $w(j_1) \geqslant w(j_2) + \sum\limits_{j \in X(j_1, j_2)} w(j)$ выполнено, то столбец j_1 охватывается группой столбцов $\{j_2\} \cup X(j_1, j_2)$. Удаляем столбец j_1 из L.

Пусть N — число единичных элементов в матрице L. В [6] доказана следующая

Теорема 1. *Матрица L может быть упрощена за время $O(Nmn)$ при помощи правил 1–3.*

Частным решением для набора строк $\{1, \ldots, i\}$, $i \in I$ называется набор столбцов такой, что он покрывает матрицу, составленную из этих строк. Алгоритм DOM представляет собой итерационную процедуру. На i-ой итерации, $i \in I$, строится множество всех частных решений S_i для набора строк $\{1, \ldots, i\}$ матрицы A_1, при этом, при $i \geqslant 2$ используются частные решения, полученные на $(i-1)$-й итерации. По множеству S_i строится матрица $B = (c_{ij})_{m \times n}$, строки которой соответствуют строкам A_1 , а столбцы соответствует частным решениям из множества S_i, причем, $c_{ij} = 1$, если j-ое частное решение из S_i содержит столбец матрицы A_1, покрывающий строку с номером i. Далее строится матрица C следующего вида

$$\begin{bmatrix} A_1 & \mathrm{B} \\ 0 & 1 \end{bmatrix}$$

Матрица C упрощается по правилам 1–3. Из множества S_i удаляются частные решения, соответсвующие удаленным из матрицы C столбцам.

По окончании работы алгоритма в S_m останется ровно одно решение, и это решение будет оптимальным. Оценка сложности алгоритма DOM, полученная в [6], приведена в сформулированной ниже теореме 2.

Теорема 2 ([6]). *Если матрица обладает свойством strong C1P, упрощена согласно правилам 1–3, её строки упорядочены в лексикографическом порядке, то время работы алгоритма DOM равно $O(M^3 n)$, где M — максимальное число единиц по строкам и столбцам.*

Оценка точности алгоритма General

Приведём теорему из [1], устанавливающую оценку точности алгоритма General и обосновывающую его корректность.

Теорема 3. *Пусть X — оптимальное решение линейной релаксации и $A' = (a'_{ij})_{m \times n}$ — булева матрица, удовлетворяющая условиям:*

1. *Множество единиц матрицы A' является подмножеством множества единиц матрицы A, т. е. $a'_{ij} \leqslant a_{ij}$ для любых i, j.*
2. *Существует число $q(A) > 0$ такое, что $\sum\limits_j a'_{ij} x_j \geqslant \frac{1}{q(A)}$ при любом $i \in I$.*
3. *Линейная релаксация задачи о покрытии с матрицей ограничений A' имеет целочисленное оптимальное решение X'. Тогда X' — допустимое решение исходной задачи, и $F(X') \leqslant q(A) F(X^*)$*

для любого допустимого решения X^* исходной задачи.

Полагая $A' = A_1$ в силу Теоремы 3, получаем оценку точности алгоритма General, равную $d(A)$, где $d(A)$ — максимальное число блоков из последовательных единиц в строке матрицы A. Серия экспериментов со случайными матрицами показала, что точность алгоритма General во многих случаях заметно выше его теоретической оценки. Из Теоремы 3 следует, что эффективность алгоритма возрастает с уменьшением $d(A)$ (если $d(A) = 1$, то полученное с помощью данного алгоритма решение будет являться точным). Построенный приближенный алгоритм также хорошо применим и к разреженным матрицам, что обеспечивается, в частности, устойчивостью алгоритма LIPSOL к разреженности матрицы ограничений .

Результаты тестирования алгоритма General

Метод протестирован на большом числе случайных матриц, с единичным вектором весов. На разреженных матрицах размера 100×100 с числом единиц примерно равным $100 \cdot 100 \cdot p$ (при $p =$ = 0,1; 0,05; 0,03) алгоритм General сравнивался с решением точного алгоритма, градиентного алгоритма и генетического алгоритма из [5]. В 50% случаев длина построенного покрытия отличалась от точного решения на 1, и в 50% случаев построенное решение оказалось точным. Чем меньше был параметр p, тем больше был процент совпадения приближенного решения с точным. Генетический алгоритм на всех матрицах выдал точные решения. Наихудшие результаты показал градиентный алгоритм. Аналогичные результаты были получены для матриц размером 200×200 и 100×300. В этом случае алгоритм General сравнивался только с генетическим и градиентным алгоритмами, поскольку для такого размера матриц точный алгоритм практически неприменим. .

Тестирование на случайных матрицах показало, что практическая точность алгоритма General во многих случаях значительно выше его теоретической оценки точности. Кроме того, данный алгоритм превосходит градиентный алгоритм по точности решения, уступая генетическому алгоритму.

Литература

[1] *Агеев А. А.* Алгоритмы с улучшенными оценками точности для задачи о покрытии множествами // Дискретный анализ и исследование операций. — 2004. — Т. 11, серия 2. — С. 3–10

[2] *Дюкова Е. В., Журавлёв Ю. И.* Дискретный анализ признаковых описаний в задачах распознавания большой размерности // Ж. вычисл. матем. и матем. физ. — 2000. — Т. 40, № 8. — С. 1264–1278.

[3] *Дюкова Е. В., Журавлёв Ю. И., Рудаков К. В.* Об алгебро-логическом синтезе корректных процедур распознавания на базе элементарных алгоритмов // Ж. вычисл. матем. и матем. физ. — 1996. — Т.36, № 8. — С. 215–223.

[4] *Журавлёв Ю. И.* Об алгебраическом подходе к решению задач распознавания или классификации // Пробл. кибернетики. — 1978. — Вып. 33. — С. 5–68.

[5] *Сотнезов Р. М.* Генетические алгоритмы для задач логического анализа данных в дискретной оптимизации и распознавании образов // Сборник тезисов лучших дипломных работ 2008 года. М.: Издательский отдел Факультета ВМиК МГУ им. М. В. Ломоносова, 2008. — С. 71–72.

[6] *Dom M.* Set cover with almost consecutive ones property // Encyclopedia of Algorithms. Springer. — 2008. — Pp. 832–834.

[7] *Karmarkar N.* A new polynomial-time algorithm for linear programming // Combinatorica. Springer Berlin. — 1984. — Vol. 4, No. 4 — Pp. 373–395.

[8] *Zhang Y.* Solving large-scale linear programs by interior-point methods under the MATLAB environment // Department of Mathematics and Statistics University of Maryland Baltimore County, Technical Report TR96-01, February, 1996.

Эффективное увеличение области притяжения глобального минимума бинарного квадратичного функционала при случайном нейросетевом поиске*

Карандашев Я. М., Крыжановский Б. В.
Ya_rad_wsem@mail.ru, iont.niisi@gmail.com
Москва, ЦОНТ НИИСИ РАН

Решается задача минимизации квадратичного функционала в конфигурационном пространстве. Для эффективного увеличения вероятности отыскания глубоких минимумов предлагается матрицу, на которой построен функционал, возводить в степень, и на полученном новом функционале решать задачу минимизации. В работе показано на примере матриц двумерной спинстекольной модели Изинга, что такая техника приводит к сдвигу спектра минимумов в более глубокую область, резко сокращает число находимых мелких минимумов и позволяет со значительно большей вероятностью (больше на 3–4 порядка) находить глобальный минимум.

В данной работе мы рассматриваем задачу нахождения глобального минимума квадратичного функционала в конфигурационном пространстве.

Пусть T — вещественная матрица размера $N \times N$, симметричная и с нулевой диагональю. Квадратичный функционал, построенный на этой матрице, имеет вид:

$$E_1 = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j}^{N} T_{ij} s_i s_j, \quad s_{i,j} = \pm 1. \qquad (1)$$

Задача состоит в том, чтобы найти такой конфигурационный вектор $S_0 = (s_1^0, s_2^0, \ldots, s_N^0)$, который даёт глобальный минимум функционала.

Нейросетевая динамика

Рассматриваемая задача является NP-полной. Её можно решить, воспользовавшись *нейросетевым спуском* со случайных стартов. Пусть у нас есть некоторая начальная (вообще говоря, случайная) конфигурация спинов $S = (s_1, s_2, \ldots, s_N)$. Найдём локальное поле, действующее на каждый её спин:

$$h_i = -\partial E_1 / \partial s_i = \sum_{j \neq i} T_{ij} s_j. \qquad (2)$$

Последовательно разворачивая спины так, чтобы они совпадали по знаку с действующим на них локальным полем, энергия конфигурации будет понижаться, до тех пор, пока мы не застрянем в одном из локальных минимумов, где все спины будут направлены вдоль действующего на них локального поля. Чтобы найти глобальный минимум, описанную *нейросетевую динамику* придётся применить многократно с различными стартовыми конфигурациями. По сути нейросетевая динамика является аналогом *покоординатного спуска* в вещественном пространстве.

*Работа выполнена при финансовой поддержке РФФИ, проект № 09-07-00159, и Гранта Президента Российской Федерации по государственной поддержке ведущих научных школ НШ-356.2008.9

Матрицы двумерной модели Изинга

В нашей работе мы будем иметь дело лишь с функционалами, построенными на матрицах спинового стекла двумерной модели Изинга. Модель Изинга — это модель взаимодействия спинов, находящихся в узлах решётки и взаимодействующих только с ближайшими соседями. Рассматриваемая размерность задачи $N = 100$.

Данный тип матриц выбран по нескольким причинам.

Во-первых, для них описанный выше обычный случайный нейросетевой поиск находит глобальный минимум менее чем один раз за миллион стартов. При этом находится порядка 990 тыс. других (более мелких) минимумов. Хотелось бы найти лучший способ поиска.

Во-вторых, для рассматриваемых матриц Изинга каждой конфигурации S соответствует ортогональная ей конфигурация S' ($SS' = 0$) с противоположной по знаку энергией $E_1(S') = -E_1(S)$. Этот факт окажется полезным в дальнейшем.

И, наконец, в-третьих, в работе [1] показано, как с помощью branch and cut метода можно найти глобальный минимум функционала на рассматриваемых матрицах небольших размерностей, используя их сильную разреженность. Значит, в руках у нас уже есть конфигурации глобальных минимумов,что помогает нам при оценке результативности наших методов.

Предлагаемый метод

Как показано в работе [2], в обобщённой модели Хопфилда при нейросетевом поиске вероятность попадания в минимум тем больше, чем больше глубина минимума. В связи с этим наша базовая идея состоит в том, чтобы видоизменить энергетическую поверхность функционала (1) таким образом, чтобы его глубокие минимумы (которые мы ищем) стали ещё глубже, а, значит, находились бы с большей вероятностью, при этом мелкие минимумы (которые для нас не представляют интереса) стали мельче или совсем исчезли из виду.

Покажем, как возведение исходной матрицы в квадрат реализует эту идею. Для этого представим нашу симметричную матрицу в виде взвешенного квазихеббовского разложения по внешним произведениям конфигурационных векторов:

$$T = \sum_{m=1}^{M} r_m S_m S_m^+, \tag{3}$$

где S_m — некоторые конфигурации, r_m — веса этих конфигураций, а M — число, достаточное для разложения. В работе [3] показано, что в качестве конфигураций в разложении (3) можно взять конфигурации экстремумов (в том числе глобальных максимумов и минимумов). В этом случае соответствующие им веса с точностью до некоторых флуктуаций будут пропорциональны энергиям этих экстремумов.

С учётом выражения (3) квадрат матрицы T примет вид:

$$T^2 = A + R, \tag{4}$$

где

$$A = N \sum_{m}^{M} r_m^2 S_m S_m^+,$$

$$R = \sum_{m=1}^{M} \sum_{n=1}^{N} (1 - \delta_{mn})\, r_m r_n S_m S_n\, (S_m, S_n). \tag{5}$$

Первое из слагаемых (матрица A) с точностью до множителя N (который уйдёт при соответствующей нормировке) совпадает с исходным разложением матрицы (3), при этом веса соответствующих конфигураций оказались возведёнными в квадрат. Именно матрица A даёт нам функционал с нужным образом видоизменённой энергетической поверхностью. Действительно, большие веса при возведении в квадрат станут ещё больше и, следовательно, увеличится глубина соответствующих минимумов и вероятность их нахождения. Малые веса при возведении в квадрат станут еще меньше, и мелкие минимумы уйдут из рассмотрения.

Нельзя забывать и про второй (перекрёстный) член R. Если конфигурации в разложении (3) были бы ортогональными, то мы имели бы R=0, и новый функционал, построенный на квадрате матрицы,

$$E_2 = -\frac{1}{2} \sum_{i=1}^{N} \sum_{j \neq i}^{N} (T^2)_{ij} s_i s_j, \tag{6}$$

имел бы конфигурации локальных минимумов, совпадающие с конфигурациями исходного функционала. Однако в общем случае это неверно. Среднее значение элементов матрицы R равно нулю, а их стандарт порядка $1/\sqrt{N}$. Это означает, что вклад матрицы R в E_2 относительно невелик. Однако его наличие приводит к тому, что конфигурации локальных минимумов функционала E_2 слегка сдвигаются относительно конфигураций S_m исходного функционала E_1.

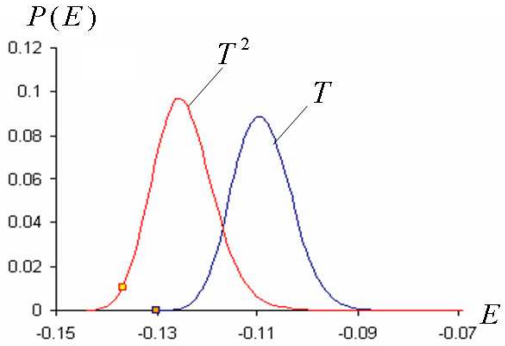


Рис. 1. Спектральные плотности минимумов функционалов построенных на матрице T и на квадрате матрицы (T^*T). Квадратиками показаны энергии конфигурации глобального минимума на обоих функционалах.

На рис. 1 приведены усреднённые по 50 матрицам графики спектральной плотности минимумов функционалов E_1 и E_2. Во всех экспериментах мы нормировали каждую матрицу (как исходную T, так и матрицу-квадрат $T^2 = T^+T$) на дисперсию её элементов. Как видно из рис.1, спектр функционала E_2 сдвинут влево по сравнению со спектром исходного функционала E_1. Также на рисунке показано, куда сдвинулся глобальный минимум исходного функционала. Он стал заметно глубже. Число обнаруженных за миллион стартов минимумов функционала E_2 порядка 600 тыс., что существенно меньше, чем для исходного (около 990 тыс.). Это показывает, что мы добились желаемого: глубокие минимумы стали глубже, а многие мелкие исчезли. В случаях, если конфигурация глобального минимума исходного функционала не являлась локальным минимумом функционала E_2, мы брали энергию минимума ближайшего к глобальному по расстоянию (число несовпадающих спинов). Число отличающихся спинов расположилось в диапазоне от 0 до 6, и в среднем по 50 матрицам оказалось равным 3.4, что подтверждает малость члена R в разложении (4).

Другим интересным подтверждением того, что при возведении матрицы в квадрат веса конфигураций тоже возводятся в квадрат, является следующее. Как упоминалось выше, спектр матрицы Изинга обладает тем свойством, что произвольному локальному минимуму функционала, построенного на этой матрице, соответствует некоторый локальный максимум этого функционала с противоположной по знаку энергией и с ортогональной конфигурацией. Заметим, что если в разложении (3) исходной матрицы минимумам соответствовали положительные веса, а максимумам — отрицательные, то возведённые в квадрат эти веса все окажутся положительными, и, значит, все они

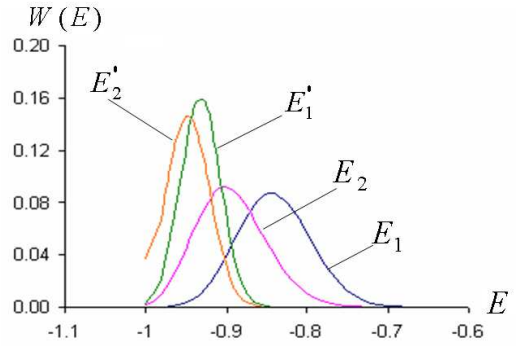


Рис. 2. Распределение плотности вероятности $W(E)$ по энергии: E_1 — для простого, E_2 — двойного спусков, E_1' — для простого спуска с отскоками, E_2' — для двойного спуска с отскоками.

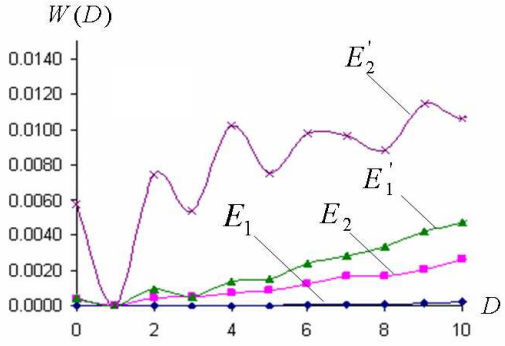


Рис. 3. Распределение вероятности попадания по расстояниям D от глобального минимума: E_1 — для простого, E_2 — двойного спусков, E_1' — для простого спуска с отскоками, E_2' — для двойного спуска с отскоками.

будут соответствовать минимумам нового функционала, или, в силу искажающего члена R, будут иметь отрицательную энергию. Эксперимент показал что, действительно, конфигурации, как максимумов, так и минимумов исходного функционала, дают на E_2 отрицательную энергию. Более того, большинство обнаруженных глубоких минимумов функционала E_2 дважды вырождены!

Поскольку наша задача состоит в том, чтобы найти глубокие минимумы исходного функционала E_1, а минимумы функционала E_2 нас, вообще говоря, не интересуют, то мы прибегли к двойному спуску. На первом этапе двойного спуска, стартуя с некоторой случайной начальной конфигурации, мы спускаемся по поверхности функционала E_2. Дойдя до минимума, мы продолжаем спуск уже по поверхности E_1, получая в итоге минимум нашего исходного функционала.

Полученные результаты

Двойной спуск даёт большую плотность вероятности попадания в области глубоких минимумов (см. рис. 2). Среднее значение энергии, получаемое при двойном спуске, равно -0.894, в то время как при обычном спуске: -0.841 (нормировка энергий такова, что глобальный минимум имеет энергию -1.000). Также двойной спуск резко уменьшил число обнаруживаемых минимумов (150 тыс. на миллион стартов) и дал увеличение вероятности попадания в глобальный минимум примерно в 160 раз.

Модификация динамики

Часто поиск останавливался в локальных минимумах, конфигурации которых расположены на расстоянии 2–3 спинов от глобального (см. рис. 3). В связи с этим мы слегка изменили нейросетевой спуск. А именно, при каждой остановке в минимуме мы случайным образом выбирали 3 спина, переворачивали их и продолжали спуск. Для каждой стартовой конфигурации мы делали

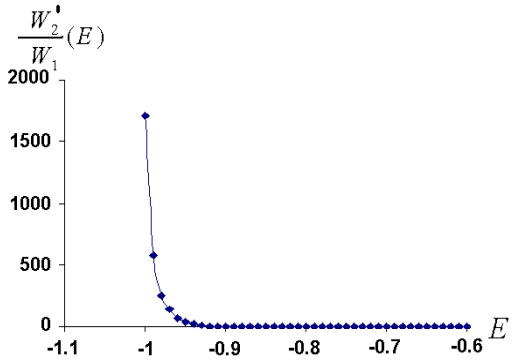


Рис. 4. Отношение плотностей вероятности попадания в заданный интервал энергий при двойном спуске с отскоками ($W_2'(E)$) и при простом спуске ($W_1(E)$).

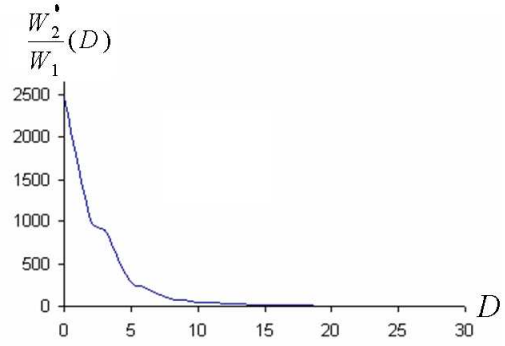


Рис. 5. Отношение вероятностей попадания на расстояние D от глобального минимума при двойном спуске с отскоками ($W_2'(D)$) и при простом спуске ($W_1(D)$).

по 100 таких «отскоков», фиксируя при этом конфигурацию самого глубокого из достигнутых минимумов. Результаты для динамики с «отскоками» (для обычного и двойного спусков) также приведены на рис. 2, 3. Среднее значение энергии при простом спуске с отскоками -0.932, а при двойном с отскоками -0.948. На рис. 4 и 5 приведены относительные характеристики этого метода по срав-

Таблица 1. Усреднённые по 50-ти матрицам результаты для двойных спусков с различных степеней матриц.

методы	вероятность попадания		среднее значение энергии
	в глобал. минимум	в интервал энергий $[-1,-0.99]$	
Без отскоков			
T(обычный)	0.00024%	0.0028%	-0.841
T^2	0.038%	0.23%	-0.894
T^3	0.2%	1.1%	-0.936
T^4	0.12%	0.56%	-0.899
T^5	0.4%	2.5%	-0.951
С отскоками			
T(обычный)	0.04%	0.3%	-0.932
T^2	0.67%	3.7%	-0.948
T^3	3.7%	14%	-0.975
T^4	1.44%	5.8%	-0.950
T^5	3.2%	14.1%	-0.974

нению с обычным. Отношение вероятности попадания в глобальный минимум при двойном спуске с отскоками к вероятности попадания в него при простом спуске без отскоков равно примерно 2400 (см. рис. 5), то есть нам удалось увеличить данный показатель на три порядка!

Разумеется, можно было бы возводить матрицу не в квадрат, а в большую степень. При этом возведение матрицы в нечётную степень приводит к тому, что веса в 1-ом члене разложения (4) тоже оказываются возведёнными в нечётную степень, а, значит, сохраняют знак. Поэтому минимумы полученной матрицы будут минимумами исходной матрицы, и мы не будем тратить время на спуск с максимумов. То есть, *возведение матрицы в нечётную степень более эффективно, чем в чётную.* Представленные в таблице 1 полученные результаты это хорошо подтверждают.

Обсуждение результатов

Не имеет смысла возводить матрицу в степень до бесконечности, ожидая роста вероятности попадания в глубокие минимумы. Как видно из рис. 1, несмотря на то, что конфигурация глобального минимума функционала E_1 имеет меньшую энергию на функционале E_2, функционал E_2 имеет какие-то новые ещё более глубокие минимумы. При дальнейшем увеличении степени матрицы эти минимумы-химеры станут ещё глубже, а, значит, и шире, пока полностью не затмят наш глобальный минимум.

Из таблицы 1 видно, что результаты для пятой степени матрицы мало отличаются от результатов для 3-ей степени.

Следует также отметить, что при возведении матрицы в степень возникают ситуации, когда глобальный минимум вообще не находится. А именно, при обычном (без отскоков) двойном спуске с куба матрицы из 50 рассматриваемых матриц для 15 мы вообще не находили глобальный минимум. Для пятой степени число матриц с ненайденным глобальным минимумом равно 34 из 50! Но даже в этой ситуации вероятность попадания в узкий интервал энергий, очень близких к глобальному минимуму (от -0.99 до -1.00), в результате замены алгоритма случайного поиска возрастала с 0.0028% до 14.1%, то есть более чем в 5000 раз (см. таблицу 1). Одновременно среднее значение энергии случайно найденного минимума понизилось с величины -0.841 до -0.974, то есть среднее расстояние до глобального минимума сократилось в 6 раз.

Более того, следует отметить, что даже в тех случаях, когда глобальный минимум не достигался, двойной спуск с большой вероятностью приводил систему в конфигурацию, очень близкую к глобальному минимуму (на расстоянии 2–5 бит от глобального минимума). Поэтому «отрицательный» результат является следствием того, что мы использовали «отскок» на три бита – наиболее простой приём выхода из мелких локальных минимумов. Однако мы не ставили перед собой цели оптимизировать выход из локального минимума — это будет сделано в дальнейшем. Нашей целью было продемонстрировать только возможность деформации потенциальной поверхности, которая приводит к существенному увеличению вероятности отыскания глобального минимума. Эта цель нами достигнута.

Литература

[1] *Hartmann A. K.* New Optimization Algorithms in Physics // Weinheim: WILEY-WCH Verlag GmbH & Co. KGaA, 2004. — P. 47—-71.

[2] *Крыжановский Б. В., Магомедов Б. М., Микаэлян А. Л.* Взаимосвязь глубины локального минимума и вероятности его обнаружения в обобщённой модели Хопфилда // ДАН. — 2005. — Т. 405, № 3, — С. 1–5.

[3] *Kryzhanovsky B. V.* Expansion of a matrix in terms of external products of configuration vectors // Optical Memory & Neural Networks (Information Optics). — 2007. № 16(4), — P. 187—-199.

Несколько актуальных проблем анализа данных*

Кельманов А. В.
kelm@math.nsc.ru
Новосибирск, Институт математики им. С.Л.Соболева СО РАН, Новосибирский государственный университет

Рассматриваются дискретные экстремальные задачи, к которым сводятся некоторые варианты проблемы поиска подмножеств векторов во множестве векторов евклидова пространства, а также некоторые варианты проблемы помехоустойчивого off-line обнаружения в числовой последовательности повторяющегося фрагмента. Изучается комбинаторная сложность редуцированных оптимизационных задач и соответствующих им задач анализа данных и распознавания образов. Анализируются алгоритмы решения этих задач.

Введение

Объект исследования работы — проблемы оптимизации в задачах анализа данных и распознавания образов. Предмет исследования — дискретные экстремальные задачи, к которым сводятся некоторые варианты проблемы поиска подмножеств «похожих» векторов во множестве векторов евклидова пространства и некоторые варианты проблемы помехоустойчивого off-line обнаружения повторяющегося фрагмента в числовой последовательности. Цель работы — обзор результатов по изучению сложности и исследованию алгоритмов решения этих задач. Данная работа дополняет сообщения [1, 2, 3].

1. Модели анализа данных

Рассмотренные ниже модели типичны для широкого спектра приложений (см., например, [4]), в которых необходимым элементом является компьютерная обработка (анализ) зашумленных структурированных данных в виде числовых или векторных последовательностей, включающих перемежающиеся информационно значимые фрагменты в одномерном случае или векторы в многомерном случае.

Пусть $x_n \in \mathbb{R}^q,\ n \in \mathcal{N}$, где $\mathcal{N} = \{1, \ldots, N\}$, — последовательность векторов. Рассмотрим две возможные структуры этой последовательности.

Структура 1. Последовательность задается формулой

$$x_n = \begin{cases} w_1,\ n \in \mathcal{M}_1, \\ \ldots,\ \ldots, \\ w_J,\ n \in \mathcal{M}_J, \\ 0,\ n \in \mathcal{N} \setminus \cup_{j=1}^{J} \mathcal{M}_j, \end{cases} \quad (1)$$

где $\cup_{j=1}^{J} \mathcal{M}_j \subseteq \mathcal{N}$, причем $\mathcal{M}_i \cap \mathcal{M}_j = \varnothing$, если $i \neq j$.

Структура 2. Последовательность обладает свойством

$$x_n = \begin{cases} w_1,\ n \in \mathcal{M}_1, \\ \ldots,\ \ldots, \\ w_J,\ n \in \mathcal{M}_J, \end{cases} \quad (2)$$

где $\bigcup_{j=1}^{J} \mathcal{M}_j = \mathcal{N}$, причем, как и в структуре 1, $\mathcal{M}_i \cap \mathcal{M}_j = \varnothing$, если $i \neq j$.

Для обеих структур положим $|\mathcal{M}_j| = M_j$, $j = 1, \ldots, J$; $M = \sum_{j=1}^{J} M_j$, $\{n_1, \ldots, n_M\} = \bigcup_{j=1}^{J} \mathcal{M}_j$. Вектор w_j будем интерпретировать как информационно значимый вектор, а M_j — как число его повторов в последовательности $x_n,\ n \in \mathcal{N}$. Доступной для анализа будем считать последовательность

$$y_n = x_n + e_n,\ n \in \mathcal{N}, \quad (3)$$

где e_n — вектор помехи (ошибки измерения), независимый от вектора x_n.

Заметим, что $x_n = x_n(\mathcal{M}_1, \ldots, \mathcal{M}_J, w_1, \ldots, w_J)$, $n \in \mathcal{N}$. Положим

$$S(\mathcal{M}_1, \ldots, \mathcal{M}_J, w_1, \ldots, w_J) = \sum_{n \in \mathcal{N}} \|y_n - x_n\|^2. \quad (4)$$

Модели анализа данных сформулируем в форме задач среднеквадратического приближения. Допустим сначала, что в отсутствие помехи данные имеют структуру 1. Сформулируем следующие задачи.

Задача 1. *Дано*: совокупность $\{y_1, \ldots, y_N\}$ векторов из $\mathbb{R}^q$. *Найти*: семейство $\{\mathcal{M}_1, \ldots, \mathcal{M}_J\}$ непустых непересекающихся подмножеств множества $\mathcal{N}$ и совокупность $\{w_1, \ldots, w_J\}$ векторов такие, что то целевая функция (4) минимальна.

Эту задачу можно трактовать как поиск семейства непересекающихся подмножеств векторов, «похожих» в среднеквадратическом смысле.

Допустим, что в рамках структуры 1 компоненты набора $(n_1, \ldots, n_M)$, соответствующие номерам ненулевых векторов в формуле (1), связаны дополнительными ограничениями

$$1 \leqslant T_{\min} \leqslant n_m - n_{m-1} \leqslant T_{\max} \leqslant N - 1, \quad (5)$$

где $m = 2, \ldots, M$; $T_{\min}$ и $T_{\max}$ — натуральные числа. Эти ограничения устанавливают допустимый интервал между двумя ближайшими номерами ненулевых векторов в последовательности (1).

Задача 2. *Дано*: последовательность $y_n \in \mathbb{R}^q$, $n \in \mathcal{N}$. *Найти*: семейство $\{\mathcal{M}_1, \ldots, \mathcal{M}_J\}$ непустых непересекающихся подмножеств множества $\mathcal{N}$ и совокупность $\{w_1, \ldots, w_J\}$ векторов такие, что целевая функция (4) минимальна при ограничениях (5) на элементы упорядоченного набора $(n_1, \ldots, n_M)$, образующие совокупность $\{n_1, \ldots, n_M\} = \bigcup_{j=1}^{J} \mathcal{M}_j$.

*Работа выполнена при финансовой поддержке РФФИ, проекты №09-01-00032, №07-07-00022 и гранта АВЦП Рособразования, проект №2.1.1/3235.

Эту задачу можно интерпретировать как оптимальное обнаружение по критерию минимума суммы квадратов уклонений ненулевых неизвестных информационно значимых векторов, повторяющихся и перемежающихся в ненаблюдаемой последовательности (1).

Предполагая, что незашумленные данные имеют структуру 2, сформулируем следующую задачу.

Задача 3. *Дано*: совокупность $\{y_1, \ldots, y_N\}$ векторов из $\mathbb{R}^q$. *Найти*: разбиение множества $\mathcal{N}$ на непустые подмножества $\mathcal{M}_1, \ldots, \mathcal{M}_J$ и совокупность $\{w_1, \ldots, w_J\}$ векторов такие, что целевая функция (4) минимальна.

Эта задача отличается от задачи 1 тем, что в ней требуется найти разбиение множества $\mathcal{N}$, а не совокупность непересекающихся подмножеств этого множества.

2. Редуцированные задачи

Легко убедиться, что во всех сформулированных задачах для любого допустимого семейства $\{\mathcal{M}_1, \ldots, \mathcal{M}_J\}$ подмножеств множества $\mathcal{N}$ минимум функционала (4) по переменным $w_1, \ldots, w_J$ доставляется векторами $\bar{w}_j = \sum\limits_{n \in \mathcal{M}_j} \frac{y_n}{|\mathcal{M}_j|}$, $j = 1, \ldots, J$. В задачах 1 и 2 в силу (1) этот минимум равен

$$S_{\min}(\mathcal{M}_1, \ldots, \mathcal{M}_J) = \\ = \sum_{n \in \mathcal{N}} \|y_n\|^2 - \sum_{j=1}^{J} \frac{1}{|\mathcal{M}_j|} \Big\| \sum_{n \in \mathcal{M}_j} y_n \Big\|^2. \quad (6)$$

Для задачи 3, учитывая (2), имеем

$$S_{\min}(\mathcal{M}_1, \ldots, \mathcal{M}_J) = \sum_{j=1}^{J} \sum_{n \in \mathcal{M}_j} \|y_n - \bar{w}_j\|^2. \quad (7)$$

Таким образом, для отыскания решений сформулированных задач необходимо решить задачи на минимум функций (6) и (7).

К идентичным оптимизационным задачам приводит статистический подход к проблеме анализа данных, если считать, что вектор e_n в формуле (3) есть выборка из q-мерного нормального распределения с параметрами $(0, \sigma^2 I)$, где I — единичная матрица, а в модели анализа данных в качестве критерия решения использовать максимум функционала правдоподобия.

Первый член в правой части равенства (6) — константа. Поэтому из задачи 1 получаем следующие редуцированные оптимизационные задачи.

Задача J-MSASVS-F (максимум суммы средних значений квадратов длин сумм векторов из подмножеств фиксированной мощности). *Дано*: множество $\mathcal{Y} = \{y_1, \ldots, y_N\}$ векторов из $\mathbb{R}^q$ и натуральные числа $M_1, \ldots, M_J$. *Найти*: семейство $\{\mathcal{B}_1, \ldots, \mathcal{B}_J\}$ непустых непересекающихся подмножеств множества $\mathcal{Y}$ такое, что

$$\sum_{j=1}^{J} \frac{1}{|\mathcal{B}_j|} \Big\| \sum_{y \in \mathcal{B}_j} y \Big\|^2 \to \max, \quad (8)$$

при ограничениях $|\mathcal{B}_j| = M_j, j = 1, \ldots, J$, на мощности искомых подмножеств.

Задача J-MSASVS-NF (максимум суммы средних значений квадратов длин сумм векторов из подмножеств, мощности которых не фиксированы). *Дано*: множество $\mathcal{Y} = \{y_1, \ldots, y_N\}$ векторов из $\mathbb{R}^q$. *Найти*: семейство $\{\mathcal{B}_1, \ldots, \mathcal{B}_J\}$ непустых непересекающихся подмножеств множества $\mathcal{Y}$ такое, что имеет место (8).

Обе задачи можно трактовать как поиск подмножеств векторов, «похожих» в среднеквадратическом смысле. Отличие задач состоит в том, что в первой из них мощности искомых подмножеств являются частью входа задачи, а во второй эти мощности — оптимизируемые величины. Аналогичным образом формулируются еще две задачи, которые следуют из задачи 2 и ориентированы на анализ последовательностей при наличии ограничений (5).

Задача J-MSASVSO-F. *Дано*: последовательность $y_n \in \mathbb{R}^q$, $n \in \mathcal{N}$, и натуральные числа $M_1, \ldots, M_J$, $T_{\min}$ и $T_{\max}$. *Найти*: семейство $\{\mathcal{M}_1, \ldots, \mathcal{M}_J\}$ непустых непересекающихся подмножеств множества $\mathcal{N}$ такое, что

$$\sum_{j=1}^{J} \frac{1}{|\mathcal{M}_j|} \Big\| \sum_{n \in \mathcal{M}_j} y_n \Big\|^2 \to \max, \quad (9)$$

при ограничениях $|\mathcal{M}_j| = M_j$, $j = 1, \ldots, J$, на мощности искомых подмножеств и при дополнительных ограничениях (5) на элементы упорядоченного набора $(n_1, \ldots, n_M)$, образующие совокупность $\{n_1, \ldots, n_M\} = \bigcup_{j=1}^{J} \mathcal{M}_j$.

Задача J-MSASVSO-NF. *Дано*: последовательность $y_n \in \mathbb{R}^q$, $n \in \mathcal{N}$, и натуральные числа $T_{\min}$ и $T_{\max}$. *Найти*: семейство $\{\mathcal{M}_1, \ldots, \mathcal{M}_J\}$ непустых непересекающихся подмножеств множества $\mathcal{N}$ такое, что имеет место (9) при ограничениях (5) на элементы упорядоченного набора $(n_1, \ldots, n_M)$, образующие совокупность $\{n_1, \ldots, n_M\} = \bigcup_{j=1}^{J} \mathcal{M}_j$.

Из задачи 3 и формулы (7) получаем хорошо известную задачу.

Задача MSSC. *Дано*: множество $\mathcal{Y} = \{y_1, \ldots, y_N\}$ векторов из $\mathbb{R}^q$ и натуральное число $J > 1$. *Найти*: разбиение множества $\mathcal{Y}$ на непустые подмножества (кластеры) $\mathcal{C}_1, \ldots, \mathcal{C}_J$ такое, что

$$\sum_{j=1}^{J} \sum_{y \in \mathcal{C}_j} \|y - \bar{w}_j\|^2 \to \min,$$

где $\bar{w}_j = \sum\limits_{y \in \mathcal{C}_j} \frac{y}{|\mathcal{C}_j|}$, $j = 1, \ldots, J$, — центры кластеров.

Эта задача является классической задачей анализа данных и распознавания образов. Сформулируем два важных специальных случая этой задачи.

Задача J**-MSSC0-F**. *Дано*: множество $\mathcal{Y} = \{y_1, \ldots, y_N\}$ векторов из $\mathbb{R}^q$ и натуральные числа $M_1, \ldots, M_J$. *Найти*: разбиение множества $\mathcal{Y}$ на непустые подмножества $\mathcal{C}_1, \ldots, \mathcal{C}_J$ такое, что

$$\sum_{j=1}^{J-1} \sum_{y \in \mathcal{C}_j} \|y - \bar{w}_j\|^2 + \sum_{y \in \mathcal{C}_J} \|y\|^2 \to \min, \qquad (10)$$

где $\bar{w}_j = \sum\limits_{y \in \mathcal{C}_j} \frac{y}{|\mathcal{C}_j|}$, $j = 1, \ldots, J-1$, при ограничениях $|\mathcal{C}_j| = M_j$, $j = 1, \ldots, J$, на мощности искомых подмножеств.

Задача J**-MSSC0-NF**. *Дано*: множество $\mathcal{Y} = \{y_1, \ldots, y_N\}$ векторов из $\mathbb{R}^q$. *Найти*: разбиение множества $\mathcal{Y}$ на непустые подмножества $\mathcal{C}_1, \ldots, \mathcal{C}_J$ такое, что имеет место (10).

Эти задачи можно трактовать как специальные случаи задачи MSSC, в которых центр одного из кластеров определять не требуется (считается, что этот центр известен и равен нулю). В первой задаче предполагается, что мощности кластеров фиксированы, а во второй число кластеров и их мощности — оптимизируемые величины.

3. Известные факты о сложности задач и алгоритмах их решения

Прежде всего заметим, что задача MSSC в силу давности постановки и широкой известности наиболее изучена в алгоритмическом плане. Имеется множество публикаций, ориентированных на построение эффективных алгоритмов с оценками точности для ее решения. Однако, лишь недавно в [5] дано корректное доказательство NP-трудности этой задачи для случая, когда $J = 2$. Все ранее опубликованные доказательства труднорешаемости этой задачи содержали ошибки [6]. Другие задачи, сформулированные в предыдущем параграфе, относятся к числу слабо изученных задач. Рассмотрим современное состояние исследований по их решению.

Алгоритмическая сложность. Относительно сложности задач поиска подмножеств векторов и специальных случаев задачи кластерного анализа получены следующие результаты. Статус NP-трудности задачи 1-MSASVS-F был установлен в [7], [8]. Из этого результата следует, что задача J-MSASVS-F при $J > 1$ также NP-трудна, как обобщение задачи 1-MSASVS-F. NP-трудность задачи 1-MSASVS-NF доказана в [9], [10]. Этот результат позволил установить труднорешаемость задачи J-MSASVS-NF при $J > 1$ в случае, когда число J является частью входа задачи. Позже в [10] была установлена труднорешаемость задачи J-MSASVS-NF для случая, когда J не является частью входа. В этой же работе было доказано, что задачи J-MSSC0-F и J-MSSC0-NF также NP-трудны.

О сложности задач анализа последовательностей с ограничением (5) на порядок выбора векторов известно следующее. Статус NP-трудности доказан [7], [8] лишь для задачи J-MSASVSO-F. Статус сложности задачи J-MSASVSO-NF пока не установлен. Скорее всего, она NP-трудна, как и задача J-MSASVS-NF.

Алгоритмы. Какие-либо алгоритмы с доказуемыми оценками точности для решения задач J-MSASVS-F и J-MSASVS-NF поиска подмножеств векторов, задач J-MSASVSO-F и J-MSASVSO-NF поиска подпоследовательностей векторов в случае, когда $J > 1$, на сегодняшний день неизвестны. То же самое можно сказать про задачи J-MSSC0-F и J-MSSC0-NF , которые имеют смысл лишь при $J > 1$.

К числу задач, для которых удалось построить алгоритмы с доказуемыми оценками точности, относятся простейшие задачи 1-MSASVS-F, 1-MSASVS-NF и 1-MSASVSO-F, в которых требуется найти лишь одно ($J = 1$) подмножество «похожих» векторов или один повторяющийся вектор в последовательности.

В [8] обоснованы приближенные асимптотически точные алгоритмы решения задач 1-MSASVS-F и 1-MSASVSO-F, имеющие временную сложность $\mathcal{O}[Nq^2(2l+1)^{q-1}]$ и $\mathcal{O}[Nq(q+M)(2l+1)^{q-1}]$ соответственно, где l — параметр алгоритма. Алгоритмы находят решение, относительная погрешность которого не превышает $(q-1)/(4l^2)$.

В [7] предложен приближенный алгоритм решения задачи 1-MSASVSO-F. Его временная сложность есть величина $\mathcal{O}[M(T_{\max} - T_{\min} + 1)N]$. К сожалению, для этого относительно «быстрого» алгоритма, хорошо зарекомендовавшего себя в численных экспериментах, гарантированная оценка точности пока не установлена.

Для решения задачи 1-MSASVS-NF в [11] предложен приближенный асимптотически точный алгоритм. Трудоемкость этого алгоритма есть величина $\mathcal{O}[Nq(q + \log N)(2l + 1)^{q-1}]$, а относительная погрешность не более $(q-1)/(4l^2)$, где l — параметр алгоритма.

В [12] доказано, что задачи 1-MSASVS-F и 1-MSASVS-NF разрешимы за время $\mathcal{O}(q^2N^{2q})$. Тем самым показано, что при фиксированной размерности q пространства эти задачи могут быть точно решены за полиномиальное время.

Для вариантов задач 1-MSASVS-F и 1-MSASVSO-F с целочисленными координатами

векторов в [13] обоснованы точные псевдополиномиальные алгоритмы. Трудоемкость этих алгоритмов есть величина $\mathcal{O}[NqM^2(2b)^{q-1}]$, где b — максимальная по абсолютной величине координата векторов из заданного множества.

Заключение

К рассмотренным NP-трудным задачам сводятся простейшие проблемы из большого семейства (насчитывающего, по крайней мере, несколько сотен элементов [14]) проблем помехоустойчивого off-line анализа и распознавания структурированных последовательностей, включающих повторяющиеся, чередующиеся и перемежающиеся информационно значимые векторы (фрагменты) в качестве структурных элементов. Очевидно, что эти труднорешаемые задачи являются частными случаями для многих еще не изученных экстремальных задач, к которым сводятся проблемы анализа данных и распознавания образов, имеющих более сложную структуру над информационно значимыми векторами. Поэтому приведенные результаты могут служить в качестве базовых (при использовании известной [15] техники полиномиальной сводимости) для доказательства NP-трудности других более сложных проблем анализа структурированных данных и распознавания образов из упомянутого семейства.

Остается заметить, что для большинства из рассмотренных экстремальных задач какие-либо алгоритмы с оценками точности на сегодняшний день неизвестны. Высокая с практической точки зрения трудоемкость существующих приближенных алгоритмов решения некоторых из рассмотренных оптимизационных задач обуславливает продолжение исследований в направлении поиска новых алгоритмических решений, а также в направлении выделения подклассов задач, для которых возможно построение алгоритмов, имеющих меньшую временную сложность.

Литература

[1] *Кельманов А. В.* Полиномиально разрешимые и NP-трудные варианты задачи оптимального обнаружения в числовой последовательности повторяющегося фрагмента // Материалы Росс. конф. «Дискретная оптимизация и исследование операций», Новосибирск: Изд-во Института математики СО РАН, 2007. `http://math.nsc.ru/conference/door07/DOOR_abstracts.pdf` — С. 46–50.

[2] *Кельманов А. В.* О некоторых полиномиально разрешимых и NP-трудных задачах анализа и распознавания последовательностей с квазипериодической структурой // Сб. докл. 13-й Всеросс. конф. «Математические методы распознавания образов», Москва.: МАКС Пресс, 2007 — С. 261–264.

[3] *Kel'manov A. V.* Off-line Detection of a Quasi-Periodically Recurring Fragment in a Numerical Sequence // Proc. of the Steklov Institute of Mathematics. — 2008. — Suppl. 2. — P. S84–S92.

[4] *Kel'manov A. V., Jeon B.* A Posteriori Joint Detection and Discrimination of Pulses in a Quasiperiodic Pulse Train // IEEE Transactions on Signal Processing. — 2008. — Vol. 52, No. 3. — P. 1–12.

[5] *Aloise D., Deshpande A., Hansen P., Popat P.* NP-Hardness of Euclidean Sum-of-Squares Clustering // Les Cahiers du GERAD, G-2008-33, 2008. — 4 p.

[6] *Aloise D., Hansen P.* On the Complexity of Minimum Sum-of-Squares Clustering // Les Cahiers du GERAD, G-2007-50, 2007. — 12 p.

[7] *Гимади Э. Х., Кельманов А. В., Кельманова М. А., Хамидуллин С. А.* Апостериорное обнаружение в числовой последовательности квазипериодического фрагмента при заданном числе повторов // Сиб. журн. индустр. математики. — 2006. — Т. 9, № 1(25). — С. 55–74.

[8] *Бабурин А. Е., Гимади Э. Х., Глебов Н. И., Пяткин А. В.* Задача отыскания подмножества векторов с максимальным суммарным весом // Дискрет. анализ и исслед. операций. Серия 2. — 2007. — Т. 14, № 1. — С. 32–42.

[9] *Кельманов А. В., Пяткин А. В.* О сложности одного из вариантов задачи выбора подмножества «похожих» векторов // Доклады РАН. — 2008. — Т. 421, № 5. — С. 590–592.

[10] *Кельманов А. В., Пяткин А. В.* Об одном варианте задачи выбора подмножества векторов // Дискрет. анализ и исслед. операций. — 2008. — Т. 15, № 5. — С. 25–40.

[11] *Кельманов А. В., Пяткин А. В.* О сложности некоторых задач поиска подмножеств векторов и кластерного анализа // Журн. вычисл. математики и мат. Физики. — 2009. — (принята в печать).

[12] *Гимади Э. Х., Пяткин А. В., Рыков И. А.,* О полиномиальной разрешимости некоторых задач выбора подмножеств векторов в евклидовом пространстве фиксированной размерности // Дискрет. анализ и исслед. операций. — 2008. — Т. 15, № 6. — С. 11–19.

[13] *Гимади Э. Х., Глазков Ю. В., Рыков И. А.,* Задача выбора подмножества векторов с целочисленными координатами в евклидовом пространстве с максимальной нормой суммы // Дискрет. анализ и исслед. операций. — 2008. — Т. 15, № 4. — С. 31–43.

[14] `http://math.nsc.ru/~serge/qpsl/` — Система QPSLab для решения задач компьютерного анализа и распознавания числовых последовательностей с квазипериодической структурой. — 2008.

[15] *Garey M. R., Johnson D. S.* Computers and Intractability: A Guide to the Theory of NP-Completeness. — San Francisco, CA: Freeman, 1979. — 345 p.

О некоторых задачах анализа и распознавания последовательностей, включающих повторяющиеся упорядоченные наборы вектор-фрагментов*

Кельманов А. В., Михайлова Л. В., Хамидуллин С. А.
kelm@math.nsc.ru, mikh@math.nsc.ru, kham@math.nsc.ru
Новосибирск, Институт математики им. С. Л. Соболева СО РАН, Новосибирский государственный университет

Рассматриваются некоторые задачи помехоустойчивого off-line анализа и распознавания числовых и векторных последовательностей, включающих повторяющиеся наборы квазипериодических фрагментов или векторов. Обоснованы эффективные алгоритмы решения этих задач, гарантирующие оптимальность решения по критерию максимального правдоподобия, в случае, когда помеха аддитивна и является гауссовской последовательностью независимых одинаково распределенных случайных величин.

Введение

В работе рассматриваются задачи анализа и распознавания структурированных данных — числовых и векторных последовательностей, в составе которых имеются повторяющиеся, чередующиеся и перемежающиеся информационно значимые фрагменты или векторы. Предмет исследования — некоторые варианты проблемы помехоустойчивого off-line анализа и распознавания последовательностей, включающих повторяющиеся упорядоченные наборы векторов или фрагментов в качестве структурных элементов, в предположении, что в отсутствие шума эти элементы совпадают с компонентами упорядоченного эталонного набора векторов, принадлежащего заданному конечному множеству (словарю). Цель работы — обоснование алгоритмов решения этих задач.

Рассмотрим две содержательные задачи. Пусть в первой из них источник сообщений передает информацию об активном состоянии некоторого физического объекта в виде эталонного набора импульсов, имеющих одну и ту же известную длительность, но различную форму. Каждому импульсу соответствует некоторое промежуточное активное состояние объекта. Порядок импульсов фиксирован. Пассивному состоянию соответствует отсутствие каких-либо импульсов. На приемную сторону через канал передачи поступает последовательность квазипериодически чередующихся импульсов, искаженная аддитивным шумом. Термин «квазипериодически» означает, что интервал между двумя последовательными импульсами не одинаков, а лишь ограничен сверху и снизу некоторыми константами. Моменты времени появления импульсов в принятой (наблюдаемой) зашумленной последовательности неизвестны. Требуется обнаружить упорядоченные наборы импульсов в наблюдаемой последовательности, т. е. определить моменты времени, в которые объект находился в активном состоянии.

Во второй содержательной задаче предполагается, что на приемную сторону поступает информация от различных физических объектов, число которых конечно. Каждому объекту однозначно соответствует известный уникальный упорядоченный векторный набор, элемент которого — результат измерения каких-либо характеристик этого объекта в промежуточном активном состоянии. Число промежуточных активных состояний у физических объектов не одинаково. В пассивном состоянии все измеряемые характеристики равны нулю. Упорядоченная совокупность промежуточных активных состояний соответствует активному состоянию этого объекта в целом. На приемную сторону поступает искаженная шумом квазипериодическая последовательность результатов измерения характеристик от неизвестного объекта. Требуется определить (распознать), от какого объекта поступила информация.

Ситуации, в которых возникают сформулированные содержательные задачи, характерны, в частности, для электронной разведки, геофизики, гидроакустики, телекоммуникации и других приложений. В обеих задачах возможны два случая, когда число принятых импульсов или число ненулевых векторных наборов в последовательности известно и неизвестно. Эти случаи для двух сформулированных содержательных задач проанализированы в настоящей работе.

1. Формальная постановка задач

Пусть $x_n \in \mathbb{R}^q,\ n \in \mathcal{N}$, где $\mathcal{N} = \{1, \ldots, N\}$, — последовательность векторов евклидова пространства. Допустим, что эта последовательность имеет следующую структуру

$$x_n = \begin{cases} u_1,\ n \in \mathcal{M}_1, \\ u_2,\ n \in \mathcal{M}_2, \\ \ldots,\ \ldots, \\ u_L,\ n \in \mathcal{M}_L, \\ 0,\ n \in \mathcal{N} \setminus \cup_{j=1}^{L} \mathcal{M}_j, \end{cases} \qquad (1)$$

где $\cup_{j=1}^{L} \mathcal{M}_j \subseteq \mathcal{N}$, причем $\mathcal{M}_i \cap \mathcal{M}_j = \varnothing$, если $i \neq j$.

Положим $|\mathcal{M}_j| = M_j,\ j = 1, \ldots, L$, и $\{n_1, \ldots, n_M\} = \bigcup_{j=1}^{L} \mathcal{M}_j$, где $M = \sum_{j=1}^{L} M_j$.

*Работа выполнена при финансовой поддержке РФФИ, проекты №09-01-00032, №07-07-00022 и гранта АВЦП Рособразования, проект №2.1.1/3235.

В дополнение к этому, допустим, что

$$\mathcal{M}_j = \{n_m \mid m \equiv j(\mathrm{mod}L),\ 1 \leqslant m \leqslant M\},\quad j = 1,\dots,L, \quad (2)$$

причем элементы набора $(n_1,\dots,n_M)$, соответствующие номерам ненулевых векторов в последовательности (1), удовлетворяют ограничениям

$$1 \leqslant T_{\min} \leqslant n_m - n_{m-1} \leqslant T_{\max} \leqslant N-1,\quad m = 2,\dots,M, \quad (3)$$

где $T_{\min}$ и $T_{\max}$ — натуральные числа.

Ограничения (3) устанавливают допустимый интервал между двумя ближайшими номерами ненулевых векторов в последовательности (1). Эти ограничения можно трактовать как условие квазипериодичности повторов ненулевых векторов в последовательности (1).

Из (1)–(3) видно, что последовательность $\{x_n\}$ включает $\lfloor M/L \rfloor$ полных повторов векторного набора $(u_1,\dots,u_L)$ и, возможно, один неполный набор. Элементы повторяющегося набора $(u_1,\dots,u_L)$ будем интерпретировать как информационно значимые векторы. Доступной для анализа будем считать последовательность

$$y_n = x_n + e_n,\ n \in \mathcal{N}, \quad (4)$$

где e_n — вектор помехи (ошибки измерения), независимый от вектора x_n. Заметим, что $x_n = x_n(n_1,\dots,n_M,u_1,\dots,u_L)$. Положим

$$S(n_1,\dots,n_M,u_1,\dots,u_L) = \sum_{n\in\mathcal{N}} \|y_n - x_n\|^2, \quad (5)$$

где $\|\cdot\|$ — норма вектора, и рассмотрим следующие задачи среднеквадратического приближения.

Задача 1. *Дано*: последовательность $y_n \in \mathbb{R}^q$, $n \in \mathcal{N}$, структура которой описывается формулами (1)–(4), набор $(u_1,\dots,u_L)$ ненулевых векторов из $\mathbb{R}^q$ и натуральное число M. *Найти*: набор $(n_1,\dots,n_M)$ номеров такой, что целевая функция (5) минимальна.

Задача 2. *Дано*: последовательность $y_n \in \mathbb{R}^q$, $n \in \mathcal{N}$, структура которой описывается формулами (1)–(4), набор $(u_1,\dots,u_L)$ ненулевых векторов из $\mathbb{R}^q$. *Найти*: набор $(n_1,\dots,n_M)$ номеров и его размерность M такие, что целевая функция (5) минимальна.

Задачи 1 и 2 отражают сущность проблемы оптимального обнаружения по критерию минимума суммы квадратов уклонений заданного повторяющегося набора информационно значимых векторов в ненаблюдаемой последовательности, структура которой описывается формулами (1)–(3). Отличие этих задач состоит в том, что в первой из них число ненулевых информационно значимых векторов считается заданным, а во второй — неизвестным, т. е. является оптимизируемой величиной.

Положим $w = (u_1,\dots,u_L)$. Допустим в дополнение к (1)–(4), что $w \in W$, причем $|W| = K$, где

$$W \subset \{(u_1,\dots,u_L) | u_j \in \mathbb{R}^q,\ 0 < \|u_j\|^2 < \infty,\ j = 1,\dots,L,\ L \in \{1,\dots,L_{\max}\}\}. \quad (6)$$

Здесь W — множество (словарь) векторных наборов (слов) мощности K, размерность которых не превосходит $L_{\max}$.

Рассмотрим еще две задачи среднеквадратического приближения.

Задача 3. *Дано*: множество W, $|W| = K$, наборов векторов из $\mathbb{R}^q$, последовательность $y_n \in \mathbb{R}^q$, $n \in \mathcal{N}$, структура которой описывается формулами (1)–(4) и (6), а также натуральное число M. *Найти*: векторный набор $w \in W$ такой, что целевая функция (5) минимальна на множестве допустимых наборов $(n_1,\dots,n_M)$.

Задача 4. *Дано*: множество W, $|W| = K$, наборов векторов из $\mathbb{R}^q$, последовательность $y_n \in \mathbb{R}^q$, $n \in \mathcal{N}$, структура которой описывается формулами (1)–(4) и (6). *Найти*: векторный набор $w \in W$ такой, что целевая функция (5) минимальна на множестве допустимых наборов $(n_1,\dots,n_M)$.

Задачи 3 и 4 соответствуют проблеме распознавания последовательностей, включающих повторяющиеся наборы чередующихся векторов, скрытых в ненаблюдаемой последовательности (1). В задаче 3 число ненулевых векторов в последовательности считается заданным, а в задаче 4 — неизвестным.

Легко установить, что к минимизации функции (5) и к таким же сформулированным выше четырем задачам приводит статистический подход к проблемам обнаружения и распознавания, если считать, что $\{e_n\}$ в формуле (4) есть выборка из q-мерного нормального распределения с параметрами $(0,\sigma^2 I)$, где I единичная матрица, а в качестве критерия решения задачи использовать максимум функционала правдоподобия.

2. Редуцированные задачи

Раскроем квадрат нормы в формуле (5):

$$S = \sum_{n\in\mathcal{N}} \|y_n\|^2 + \sum_{j=1}^{L} M_j \|u_j\|^2 - 2\sum_{j=1}^{L}\sum_{n\in\mathcal{M}_j} \langle y_n, u_j\rangle$$
$$= \sum_{n\in\mathcal{N}} \|y_n\|^2 + \sum_{m=1}^{M} \|u_{(m-1) \bmod L+1}\|^2 - 2\sum_{m=1}^{M} \langle y_{n_m}, u_{(m-1) \bmod L+1}\rangle,$$

где $\langle \cdot, \cdot \rangle$ — скалярное произведение.

Первое слагаемое в правой части полученного выражения — константа. При фиксированных M и $(u_1, \ldots, u_L)$ второе слагаемое также является константой. Поэтому имеем следующие редуцированные оптимизационные задачи, к которым сводятся задачи 1 и 2.

Задача SRTVS-F (Searching for Recurring Tuples of Vectors in a Sequence, when M is Fixed). *Дано*: последовательность $y_0, \ldots, y_{N-1}$ векторов из $\mathbb{R}^q$, набор $(u_1, \ldots, u_L)$ ненулевых векторов из $\mathbb{R}^q$ и натуральное число M. *Найти*: набор $(n_1, \ldots, n_M)$ номеров такой, что

$$\sum_{m=1}^{M} \langle y_{n_m}, u_{l(m,L)} \rangle \to \max,$$

где $l(m|L) = (m-1) \bmod L + 1$, при ограничениях (3).

Задача SRTVS-NF (Searching for Recurring Tuples of Vectors in a Sequence, when M is Not Fixed). *Дано*: последовательность $\{y_0, \ldots, y_{N-1}\}$ векторов из $\mathbb{R}^q$ и набор $(u_1, \ldots, u_L)$ ненулевых векторов из $\mathbb{R}^q$. *Найти*: набор $(n_1, \ldots, n_M)$ номеров и его размерность M такие, что

$$\sum_{m=1}^{M} \{2\langle y_{n_m}, u_{l(m,L)} \rangle - \|u_{l(m,L)}\|^2\} \to \max, \qquad (7)$$

где $l(m|L) = (m-1) \bmod L + 1$, при ограничениях (3).

Точные полиномиальные алгоритмы решения этих редуцированных оптимизационных задач обоснованы в [1, 2, 3]. Трудоемкости алгоритмов решения задач SRTVS-F и SRTVS-NF есть величины $\mathcal{O}[M(T_{\max} - T_{\min} + q)N]$ и $\mathcal{O}[L(T_{\max} - T_{\min} + q)N]$ соответственно.

Задачи 3 и 4 сводятся к решению следующих экстремальных задач.

Задача SVTVP-F (Searching for a Vector Tuple in the Vocabulary of Patterns, when M is Fixed). *Дано*: последовательность $y_0, \ldots, y_{N-1}$ векторов из $\mathbb{R}^q$, натуральное число M и множество (словарь) W, $|W| = K$, упорядоченных наборов векторов из $\mathbb{R}^q$. *Найти*: векторный набор $w \in W$ такой, что выполняется (7), при ограничениях (3).

Задача SVTVP-NF (Searching for a Vector Tuple in the Vocabulary of Patterns, when M is Not Fixed). *Дано*: последовательность $y_0, \ldots, y_{N-1}$ векторов из $\mathbb{R}^q$ и множество (словарь) W, $|W| = K$, упорядоченных наборов векторов из $\mathbb{R}^q$. *Найти*: векторный набор $w \in W$ такой, что выполняется (7), при ограничениях (3).

Точные полиномиальные алгоритмы решения этих экстремальных задач обоснованы в [4, 5].

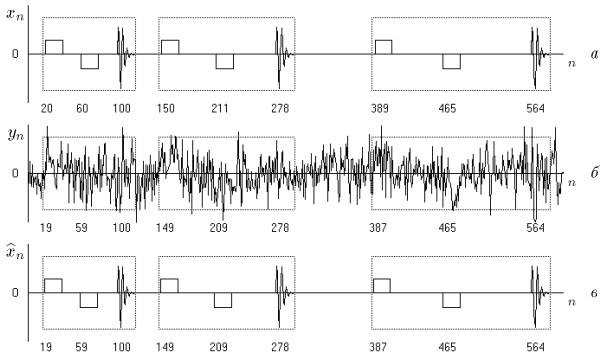


Рис. 1.

Временные сложности алгоритмов решения задач SVTVP-F и SVTVP-NF есть величины $\mathcal{O}[KM(T_{\max} - T_{\min} + q)N]$ и $\mathcal{O}[KL_{\max}(T_{\max} - T_{\min} + q)N]$ соответственно.

Алгоритмы решения приведенных редуцированных задач лежат в основе алгоритмов помехоустойчивого анализа и распознавания структурированных последовательностей, включающих повторяющиеся наборы чередующихся вектор-фрагментов. Эти алгоритмы гарантируют оптимальность решения как по критерию максимального правдоподобия в случае, когда помеха аддитивна и является гауссовской последовательностью независимых одинаково распределенных величин, так и по критерию минимума суммы квадратов уклонений.

3. Численное моделирование

Результаты численных экспериментов, представленные ниже в качестве примера, носят чисто иллюстративный характер. Они лишь демонстрируют работу алгоритмов и сущность рассмотренных задач для одномерных последовательностей.

На рис. 1 а изображена сгенерированная последовательность X, включающая 3 повтора набора фрагментов. На рис. 1 б представлена последовательность Y, подлежащая обработке (в этом примере уровень помехи превышает уровень сигнала). На рис. 1 в приведена последовательность $\hat{X}$, полученная с помощью алгоритма обнаружения, в условиях, когда число M задано. Прямоугольными рамками очерчены места расположения обнаруженного набора, найденные алгоритмом в зашумленной последовательности. Числовые данные под графиками соответствуют заданным (рис. 1 а) и найденным (рис. 1 б и 1 в) начальным номерам фрагментов. Рисунок иллюстрирует практически безупречную работу алгоритма в условиях, когда уровень сигнала ниже уровня помехи.

На рис. 2 представлены кривые оценок нормированной среднеквадратической ошибки $e(\sigma) = \mathsf{E}\|X - \hat{X}\|^2 / e^u$, где E — символ математического ожидания, e^u — оценка сверху для $\|X - \hat{X}\|^2$.

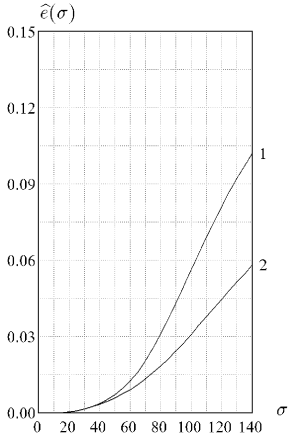


Рис. 2.

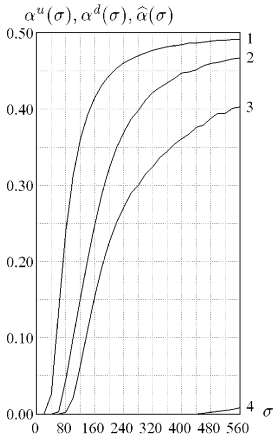


Рис. 3.

Кривая 1 получена с помощью алгоритма обнаружения при неизвестном числе M фрагментов, а кривая 2 — с помощью алгоритма, ориентированного на ситуацию, когда это число известно. Результаты получены при обработке одних и тех же 25000 сгенерированных последовательностей, в составе которых повторялся набор из трех фрагментов; места расположения фрагментов в последовательностях генерировались с помощью датчика случайных чисел.

Рис. 3 иллюстрирует зависимость от уровня помехи вероятности ошибки распознавания последовательностей, которые порождались двумя различными эталонными векторными наборами, в составе которых имелось по три вектора. Теоретические оценки верхней и нижней границ вероятности ошибки распознавания $\alpha^u(\sigma)$ и $\alpha^d(\sigma)$ в виде графиков приведены под номерами 1 и 4. Кривые 2 и 3 получены в условиях, когда число M было неизвестно и известно соответственно.

Оценка вероятности ошибки распознавания при каждом значении σ подсчитана по формуле $\hat{\alpha} = (\nu_1 + \nu_2)/2$, где ν_1 и ν_2 — числа неверно опознанных последовательностей, сгенерированных по каждому эталонному набору. Моделировалась байесовская процедура принятия решения с равновероятными гипотезами (наборами). Каждая точка экспериментальной кривой $\hat{\alpha}$ получена в результате усреднения 25000 значений. Рис. 2 и 3 демонстрируют легко доказуемый факт, что ошибка обнаружения и вероятность ошибки распознавания будут меньше в ситуации, когда число ненулевых фрагментов в последовательности известно, чем в ситуации, когда это число неизвестно.

Заключение

Рассмотренные задачи входят в большое семейство актуальных задач [6], к которым сводятся типовые проблемы помехоустойчивого off-line анализа и распознавания структурированных данных в виде числовых и векторных последовательностей, включающих повторяющиеся, чередующиеся и перемежающиеся информационно значимые векторы или фрагменты. В настоящей работе представлены эффективные алгоритмические решения четырех ранее не изученных задач из этого семейства.

Открытым остается вопрос о разрешимости обобщения рассмотренных задач обнаружения и распознавания на тот случай, когда вместо набора фрагментов, элементы которого упорядочены в соответствии с фиксированным набором векторов, требуется найти набор фрагментов с точностью до всевозможных перестановок элементов фиксированного векторного набора. Алгоритмы решения этих задач представляют значительный интерес для ряда упомянутых во введении приложений. Обоснование алгоритмов решения этих задач представляется делом ближайшей перспективы.

Литература

[1] *Кельманов А.В., Михайлова Л.В., Хамидуллин С.А.* Апостериорное обнаружение в квазипериодической последовательности повторяющегося набора эталонных фрагментов // Журн. вычисл. математики и мат. физики. — 2008. — Т. 48, № 12. — С. 1–14.

[2] *Кельманов А.В., Михайлова Л.В., Хамидуллин С.А.* Об одной задаче поиска упорядоченных наборов фрагментов в числовой последовательности // Дискретный анализ и исследование операций. — 2009 (принята в печать).

[3] *Кельманов А.В., Михайлова Л.В., Хамидуллин С.А.* Оптимальное обнаружение в квазипериодической последовательности повторяющегося набора эталонных фрагментов // Сиб. журн. вычисл. математики. — 2008. — Т. 11, № 3. — С. 311–327.

[4] *Кельманов А.В., Михайлова Л.В., Хамидуллин С.А.* Распознавание квазипериодической последовательности, включающей повторяющийся набор фрагментов // Сиб. журн. индустр. математики. — 2008, — Т. 11, № 2(34). — С. 74–87.

[5] *Кельманов А.В., Михайлова Л.В., Хамидуллин С.А.* Алгоритм распознавания квазипериодической последовательности, включающей повторяющийся набор фрагментов // Тез. докл. 15-й междунар. конф. «Проблемы теоретической кибернетики». — Казань: Отечество, — 2008. — С. 45.

[6] `http://math.nsc.ru/~serge/qpsl/` — Система QPSLab для решения задач компьютерного анализа и распознавания числовых последовательностей с квазипериодической структурой. — 2008.

Алгоритмическая сложность распознавания с использованием активного сенсора

Медников Д. И., Сергунин С. Ю., Кумсков М. И.
gamunculus@gmail.com
Москва, МГУ им. М. В. Ломоносова, мехмат, кафедра вычислительной математики

В работе описывается двухуступенчатый (гипотеза–подтверждение) метод распознавания, использующий активный сенсор, и приводятся теоретические оценки его вычислительной сложности.

Постановка задачи и определения

Задача распознавания объектов на полутоновых изображениях ставится следующим образом. Пусть дан набор полутоновых изображений проекций $\mathrm{Pr}_j(O_i)$ объектов O_i, набор полутоновых изображений эталонных фонов B_k и полутоновое изображение тестовой сцены Sc со сложным фоном. Требуется найти, какие объекты в каких проекциях находятся на сцене, и в какой части сцены они находятся, или же сообщить, что известных объектов на сцене нет, то есть построить функцию

$$F\left(Sc, O_i\right) = \{e_i,\ \mathrm{pos}_i,\ \mathrm{orient}_i,\ \mathrm{scale}_i\},$$

где

$$e_i = \begin{cases} 0, \text{если объект } O_i \text{ отсутствует на сцене } Sc, \\ 1, \text{если объект } O_i \text{ присутствует на сцене } Sc, \end{cases}$$

$\mathrm{pos}_i = (x_i, y_i)$ — координаты объекта O_i на сцене Sc (определено, если $e_i = 1$), $\mathrm{orient}_i = (\rho_i,\ \theta_i,\ \varphi_i)$ — ориентация проекции $\mathrm{Pr}(O_i)$, присутствующей на сцене Sc, относительно эталонной системы координат объекта O_i, scale_i — масштаб проекции $\mathrm{Pr}(O_i)$ присутствующей на сцене Sc, относительно эталонного масштаба объекта O_i.

Введем основные определения, с которыми мы будем иметь дело при описании предлагаемого подхода к распознаванию.

Определение 1. *Назовем объектом O_i твердый трехмерный предмет определенной текстуры с привязанной к нему эталонной сферической системой координат и эталонным масштабом.*

Определение 2. *Проекцией объекта* $\mathrm{Pr}(O_i)$ *мы будем называть полутоновое изображение объекта на однородном фоне определенного цвета в определенном масштабе и расположенного под определенным углом к эталонной системе координат объекта.*

Определение 3. *Силуэт объекта* $\mathrm{Sil}(\mathrm{Pr}(O_i))$ *— замкнутый контур, охватывающий минимальную область, содержащую объект O_i, на его проекции* $\mathrm{Pr}(O_i)$.

Определение 4. *Будем называть особой точкой (ОТ)* $\mathrm{SP}(Im)$ *некоторый характерный участок изображения. ОТ* $\mathrm{SP}(Im)$ *описывается вектором (x, y, m), где (x, y) — координаты* SP *на Im, m — числовой маркер, описывающий класс этой точки.*

Полагается, что в r-окрестности особой точки не может быть других особых точек; если они есть, то вместо этой группы ОТ ставится одна *конгломерирующая ОТ*. Конкретные реализациимогут быть пересечениями и/или изломами контуров объектов O_i на изображении Im, центрами геометрических фигур, которые можно выделить на этих контурах, или центрами цветовых сегментов изображения.

Определение 5. *Облаками особых точек мы будем называть подмножества множества всех особых точек данного изображения.*

Определение 6. *Описанием изображения* $D(Im)$ *будем называть множество особых точек* $\mathrm{SP}_i(Im)$ *такое, что по нему можно определить «похожесть» изображения на другой изображение, используя только функции сравнения облаков ОТ, но не участков изображений.*

Определение 7. *Конфигурация облака особых точек* $\mathrm{Conf}(O_i(Im))$ *— вектор определенной размерности, описывающий взаимное расположение $O_i(Im)$, находящихся на изображении Im, принадлежащих этому облаку.*

Определение 8. *Похожесть двух конфигураций особых точек* $\mathrm{Sim}(\mathrm{Conf}_i, \mathrm{Conf}_j)$ *расстояние между векторами этих конфигураций в некоторой метрике, например, евклидовой.*

Определение 9. *Рейтингом одной ОТ относительно другой* $\mathrm{Rank}(O_i, O_j)$ *будем называть число, описывающее похожесть облаков ОТ, попадающих в круговые окрестности* $\mathrm{Reg}(O_i)$ *этих O_i, друг на друга. Чем он больше, тем более похожи окрестности.*

Определение 10. *В качестве модели (шаблона)* $M(\mathrm{Pr}(O))$ *проекции* Pr *объекта O рассматривается пара* $\left(\{\mathrm{SP}_i\left(\mathrm{Pr}\left(O\right)\right)\},\ \mathrm{Sil}(\mathrm{Pr}(O))\right)$*, то есть набор ее особых точек и силуэт объекта на этой проекции.*

Определение 11. *Важнейшим для этой работы определением будет окно сенсора* $W(Im, x, y, s)$ *— прямоугольник фиксированного размера, заданный рассматриваемым изображением Im, координатами положения своего центра на этом изображении (x, y) и масштабом попадающего в него куска изображения Im.*

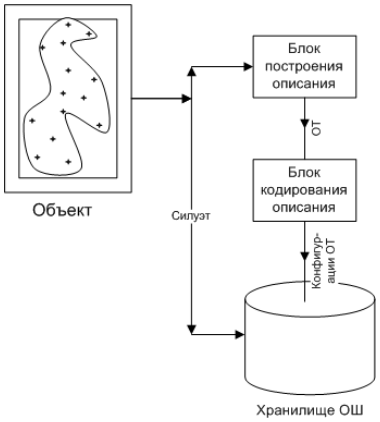


Рис. 1. Схема наполнения базы ОШ.

Система «смотрит» на изображение только через окно сенсора, в котором происходит описание ОТ их окрестностями и сравнение облаков ОТ, ОТ вне его при этом не учитываются. В каждый момент времени система работает только с частью изображения, расположенной внутри окна сенсора, и «забывает», что происходит вне окна (вернее, она помнит только, насколько похожа внешняя часть изображения на тот или иной объект, но не помнит данный об особых точках этой области). Окно сенсора может перемещаться по изображению в процессе распознавания и динамически изменять уровень «зума», то есть отображать разные по площади участки изображения в зависимости масштаба, в котором в текущий момент происходит работа.

Описание схемы распознавания

Этап наполнения базы проходит при активном участии человека-оператора, то есть является полуавтоматическим. В систему заносятся изображения объектов O_i. В *Блоке построения описания* на каждом изображении Im для каждого масштаба (количество масштабов, используемых в работе, является параметром системы) этого изображения при помощи того или иного алгоритма определяется положение особых точек $\{\mathrm{SP}_i(Im)\}$, далее им присваиваются некоторые характеристики участка изображения, составляющего их ближайшую окрестность (например, средняя яркость этого участка или количество ветвей контуров объектов, пересекающихся в особой точке), то есть каждая особая точка характеризуется своими координатами и вектором характеристик своей окрестности. Некоторые из них выделяются оператором как реперные, они должны располагаться в самых характерных участках объекта на изображении. После того, как ОТ всех объектов выделены (они составляют множество $\{\mathrm{SP}_i\}$), в том же блоке проводится их маркировка на основе кластер-анализа векторов характеристик, соответствующих особым точкам (в качестве маркера берется номер кластера i в множестве $\{\mathrm{KL}_i\}$). Таким образом, многомерное пространство векторов характеристик сводится к одномерному пространству маркеров. Полученные маркированные особые точки поступают в *Блок кодирования описания*, где описываются векторами своих конфигураций $K(\mathrm{SP}_i)$, то есть происходит обратный переход от одномерного пространства маркеров в многомерное пространство векторов конфигураций, но теперь описания ОТ зависят только от других особых точек, а не от соответствующих им участков изображения. Так получаются необходимые нам описания изображений $D(Im_i)$, абстрагированные от самих изображений, то есть двумерная матрица пикселей изображения переводится в плоский граф своих особых точек; таким образом, существенно уменьшается вычислительная сложность при дальнейшем распознавании. Полученные векторы записываются в *Хранилище объектных моделей* (базу данных), параллельно туда же для каждого объекта записывается его силуэт Sil. В хранилище вектора конфигураций особых точек индексируются для дальнейшего быстрого поиска похожих по конфигурациям ОТ по их индексам. Итак, в хранилище оказываются все модели проекций объектов $M(\mathrm{Pr}_i(O_j))$.

Подробная схема выбора активной модели выглядит следующим образом.

1. *Блок построения описания* задает первоначальные координаты центра (x, y) окна сенсора $WAS(x, y, s)$ и уровень масштаба изображения s. Сначала s выбирается максимальным, так, чтобы все изображение попало в окно.

2. *Сенсор* делает «снимок» окна с заданными координатами и передает в *Блок построения описания*, где на этом участке изображения выделяются и маркируются особые точки.

3. Набор маркированных ОТ окна сенсора передается в *Блок кодирования описания*, где строится набор его дескрипторов (векторов конфигураций ОТ), то есть представление набора в виде символьных строк.

4. Набор дескрипторов поступает в хранилище ОМ, и каждая модель в хранилище сравнивает полученные дескрипторы с собственными, хранимыми внутри модели.

5. При этом вычисляется и накапливается *рейтинг* модели $P\,(Im)$ — показатель, характеризующий количество совпавших дескрипторов, а тем самым, возможность присутствия на сцене объекта Im.

6. Модель, набравший максимальный рейтинг, становится активной моделью-кандидатом. Остальные модели становятся моделями-кандидатами (неактивными), только если их рейтинг превысит порог активации.

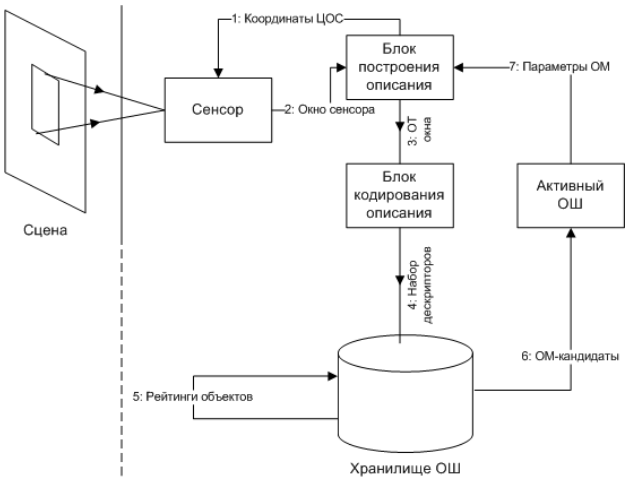


Рис. 2. Выбор активной модели.

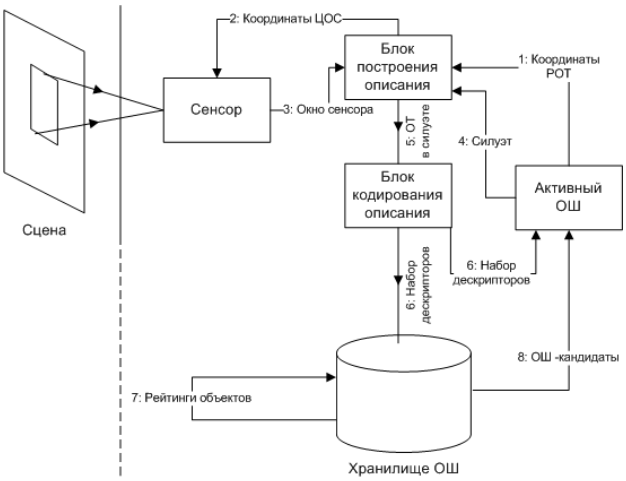


Рис. 3. Подтверждение активной модели.

7. Активная модель передает свои параметры в *Блок построения описания* для дальнейшего подтверждения.

Второй этап процесса распознавания — подтверждение активной модели-кандидата.

1. *Активный ОШ P(O)* передает координаты своих реперных особых точек $\{\mathrm{RSP}_i(P)\}$ в *Блок построения описания* для установки в них *WAS*.

2. *Блок построения описания* задает координаты *WAS* и уровень пирамиды изображения.

3. *Сенсор* передает полученное окно в *Блок построения описания*.

4. Активная модель передает силуэт объекта в *Блок построения описания*.

5. *Блок построения описания* формирует набор особых точек изображения с учетом силуэта объекта, то есть, отсекая ОТ, соответствующие фону. Полученный набор поступает в *Блок кодирования описания*.

6. *Блок кодирования описания* строит дескрипторы и передает их в *Хранилище* ОШ и в Активную модель *P(O)* для вычисления рейтингов.

7. В *Хранилище* ОМ каждая модель отдельно вычисляет свой рейтинг. Если после прохождения окном сенсора всех $\{\mathrm{RSP}_i(P)\}$ рейтинг активной модели не превышает порога подтверждения, то уровень пирамиды изображения понижается и происходит переход в шаг 1 для этого уровня. Если порог активации превышается в какой-то момент, то объект O считается найденным, и подтверждение происходит для следующей по рейтингу модели-кандидата. Иначе считается, что в данной позиции на изображении объекта O нет, и подтверждение происходит для следующей по рейтингу модели-кандидата.

8. Возможно пополнение списка моделей кандидатов, если еще какие-то модели набрали рейтинг больше порога активации.

При отсутствии моделей кандидатов производится поиск активного кандидата на более низком уровне пирамиды, чем в прошлый раз.

Итак, мы рассмотрели основные виды работы РИАС системы, определили ее базовую модульную структуру и общий алгоритм ее работы. Были показаны основные отличительные особенности РИАС системы — ограничение области видимости системы при помощи окна активного сенсора, двухуровневая система распознавания, быстро находящая точки интереса на огрубленных изображениях и проверяющая их на ограниченных силуэтами объектов участках. Первая особенность позволяет сделать распознавание более интеллектуальным — каждая модель подтверждается по-своему без каких-либо усилий со стороны человека-оператора, что улучшает его точность. Вторая позволяет сохранить вычислительные ресурсы за счет необходимости проверки меньшего числа особых точек на первом «грубом» этапе и подробного исследования лишь отдельных участков изображения ограниченных силуэтами. Использование силуэтов также повышает качество разделения объект/фон, так как при подробной верификации на втором этапе вносящие ошибки особые точки фона не используются.

Теоретические оценки

Построим теоретические оценки сложности предлагаемого алгоритма.

Лемма 1. *На прямоугольном изображении размеров $M \times N$ не может быть выделено более $\frac{2MN}{R^2}$ особых точек, если по определению ОТ в ее R-окрестности не может быть других ОТ, вне зависимости от алгоритма выделения.*

Доказательство. Покроем прямоугольник квадратами размеров $\frac{R}{\sqrt{2}} \times \frac{R}{\sqrt{2}}$. Их будет не более $\frac{\sqrt{2}M}{R}\frac{\sqrt{2}N}{R} = \frac{2MN}{R^2}$ штук. Каждый такой квадрат

имеет диагональ длины R, поэтому, по определению ОТ, в нем может располагаться не более одной особой точки (при любом положении точки внутри квадрата он целиком попадет в ее R-окрестность). Таким образом, число ОТ будет не превышать число таких квадратов, то есть $\frac{2MN}{R^2}$.

Следствие 1. *При уменьшении масштаба изображения в 2 раза максимальное число ОТ, которые могут на нем располагаться, сократится в 4 раза.*

Доказательство. Это утверждение является элементарным следствием Леммы 1. В случае уменьшения масштаба изображения в 2 раза его линейные размеры тоже уменьшатся в 2 раза ($M \times N \to \frac{M}{2} \times \frac{N}{2}$). Таким образом, по Лемме 1 вместо $\frac{2MN}{R^2}$ особых точек на нем максимально сможет поместиться $\frac{MN}{2R^2} = \frac{1}{4}\frac{2MN}{R^2}$.

Теорема 2. *При кодировании особыми точками полутонового изображения размером $M \times N$ используемый объем информации не будет превышать MN байт при $R \geqslant 5$.*

Доказательство. Запись одной маркированной особой точки потребует 12 байт (по 4 байта на маркер, x и y координаты). Поскольку особых точек на изображении не более $\frac{2MN}{R^2}$, то на запись его в виде набора ОТ потребуется не более $12\frac{2MN}{R^2} = \frac{24MN}{R^2}$ байта. Отсюда получаем $\frac{24MN}{R^2} < MN$, следовательно, $R^2 > 24$ и $R \geqslant 5$, так как оно целое.

Лемма 3. *При одноуровневом распознавании без уменьшения масштаба рассматриваемых изображений (классическая схема) потребуется $\frac{4M^2N^2}{R^4}K$ операций сравнения, где K — число изображений в базе, $M \times N$ — их размер, R — радиус пустой окрестности ОТ.*

Доказательство. Пусть в базе имеется K изображений объектов размером $M \times N$. Особых точек на них суммарно будет не более $\frac{2MN}{R^2}K$. На исследуемом изображении будет не более чем $\frac{2MN}{R^2}$ особых точек. Таким образом, для поиска объектов будет проведено $\frac{4M^2N^2}{R^4}K$ операций сравнения.

Теорема 4. *При предложенном подходе к распознаванию потребуется не более чем*

$$\frac{1}{2^{2(l+1)}}\frac{4M^2N^2}{R^4}K + \frac{\pi P^2M^2N^2S}{R^6 2^{z-3}}$$

действий для распознавания объекта на изображении, где K — число изображений в базе, $M \times N$ — их размер, R — радиус пустой окрестности ОТ, 2^l — масштаб огрубления на первом шаге алгоритма распознавания, P — радиус исследуемых окрестностей реперных ОТ, $\frac{1}{2^z}$ — доля реперных ОТ среди всех особых точек изображения, S — число изображений кандидатов, выбираемых в конце первого шага.

Доказательство. Пусть в базе имеется K изображений объектов размером $M \times N$. На первой стадии распознавания работа идет на масштабе уменьшенном в 2^l раз, поэтому особых точек на них суммарно будет не более $\frac{1}{2^{2l}}\frac{2MN}{R^2}K$. Таким образом, для сравнения всех их особых точек с точками исследуемого изображения потребуется произвести $\frac{1}{2^{2(l+1)}}\frac{4M^2N^2}{R^4}K$ сравнений особых точек.

На втором шаге сравниваются P-окрестности реперных особых точек, которых не более $\frac{1}{2^z}$ от числа ОТ исследуемого изображения. В каждой P-окрестности особых точек не более чем $\frac{2\pi P^2}{R^2}$ (вместо прямоугольника $M \times N$ покрываем квадратиками $\frac{R}{\sqrt{2}} \times \frac{R}{\sqrt{2}}$ круг радиуса P). Значит, всего придется исследовать $\frac{2\pi P^2}{R^2}\frac{MN}{R^2 2^{z-1}}$ особых точек на исследуемом изображении. На втором этапе выбирается S изображений кандидатов, которые рассматриваются уже в оригинальном масштабе, то есть на них содержится максимум $\frac{2MN}{R^2}S$. Значит, придется сделать $\frac{\pi P^2M^2N^2S}{R^6 2^{z-3}}$ сравнений. Отсюда следует утверждение теоремы.

Таким образом, подбором параметров K, l, P, S можно добиться баланса между качеством распознавания и его скоростью (числом сравнений), сделав ее выше чем у классической схемы распознавания, например, выбрав

$$l > \frac{1}{2}\log_2 \frac{\pi P^2M^2N^2SR^4}{R^6 2^{z-3}4M^2N^2K} - 1.$$

Заключение

Таким образом, показаны широкие возможности уменьшения вычислительной сложности предложенной схемы распознавания при сохранении приемлемого качества распознавания. Вопросы эффективности работы предлагаемой схемы были рассмотрены авторами в предыдущих работах [1,2,3].

Литература

[1] *Сергунин С., Миловидов А., Лозинский В., Кравченко Д.* Примеры построения алфавитов описания окна активного сенсора в РОАС системе // Докл. 12-й всеросс. конф. ММРО-12, 20-26 ноября 2005, Звенигород. — С. 447–450.

[2] *Сергунин С. Ю., Кумсков М. И.* Свойства модели объекта в системе распознавания с активным сенсором // Докл. 12-й всеросс. конф. ММРО-12, 20–26 ноября 2005, Звенигород. — С. 441–444.

[3] *Mednikov D. I., Milovidov A., Sergunin S. Yu., Kumskov M. I.* Identification of stable description elements using an active sensor // Pattern Recognition and Image Analysis. — 2009. — Vol. 19, № 1.

Задачи анализа и распознавания последовательностей, включающих серии повторяющихся вектор-фрагментов*

Михайлова Л. В.
mikh@math.nsc.ru
Новосибирск, Институт математики им. С. Л. Соболева СО РАН

Рассматриваются некоторые задачи помехоустойчивого off-line анализа и распознавания числовых и векторных последовательностей, включающих серии идентичных фрагментов или векторов. Обоснованы эффективные алгоритмы решения этих задач, гарантирующие оптимальность решения по критерию максимального правдоподобия, в случае, когда помеха аддитивна и является гауссовской последовательностью независимых одинаково распределенных случайных величин.

Введение

Предмет исследования данной работы — некоторые задачи помехоустойчивого off-line анализа и распознавания структурированных данных — числовых и векторных последовательностей, включающих серии идентичных векторов или фрагментов в качестве структурных элементов, в предположении, что в отсутствие шума эти элементы совпадают с компонентами упорядоченного эталонного набора векторов, принадлежащего заданному конечному множеству (словарю). Цель работы — обоснование алгоритмов решения этих задач.

Рассмотрим следующую содержательную задачу. Пусть источник сообщений через канал связи с помехой передает информацию о последовательности активных состояний некоторого физического объекта. Активному состоянию соответствует ненулевой вектор, пассивному — нулевой, причем различным активным состояниям соответствуют различные ненулевые векторы. Для передачи сообщения о состоянии источник порождает и посылает серию идентичных ненулевых векторов. На приемную сторону через канал передачи поступает порожденная квазипериодическая последовательность векторов, искаженная аддитивным шумом. Термин «квазипериодическая» означает, что интервал между двумя последовательными ненулевыми векторами в последовательности не одинаков, а лишь ограничен сверху и снизу некоторыми константами. Набор ненулевых векторов, соответствующий последовательности активных состояний объекта, задан. Моменты времени появления ненулевых векторов и временные границы серий в принятой (наблюдаемой) зашумленной последовательности неизвестны. Требуется определить границы серий в наблюдаемой последовательности, т. е. определить моменты времени изменения состояний объекта.

Рассмотрим еще одну задачу. Допустим, что имеется конечное число физических объектов. Каждый объект описывается уникальным набором ненулевых векторов, которые соответствуют его активным состояниям. Источник сообщений передает информацию о состояниях объекта описанным в предыдущей задаче способом. Требуется определить (распознать), от какого объекта поступила информация.

Сформулированные содержательные задачи возникают, например, в электронной разведке, геофизике, гидроакустике, телекоммуникации и других приложениях. В этих задачах возможны два случая: когда число ненулевых векторов в принятой последовательности известно и неизвестно. Эти случаи для двух сформулированных содержательных задач проанализированы в настоящей работе.

Формальная постановка задач

Допустим, что $x_n \in \mathbb{R}^q,\ n \in \mathcal{N} \equiv \{1,\dots,N\}$ — последовательность векторов евклидова пространства. Предположим, что эта последовательность имеет следующую структуру:

$$x_n = \begin{cases} u_1, & n \in \mathcal{M}_1, \\ u_2, & n \in \mathcal{M}_2, \\ \dots, & \dots, \\ u_L, & n \in \mathcal{M}_L, \\ 0, & n \in \mathcal{N} \setminus \bigcup\limits_{j=1}^{L} \mathcal{M}_j, \end{cases} \tag{1}$$

где $\bigcup\limits_{j=1}^{L} \mathcal{M}_j \subseteq \mathcal{N}$, причем $\mathcal{M}_i \cap \mathcal{M}_j = \varnothing$, если $i \neq j$.

Пусть $|\mathcal{M}_j| = M_j,\ j = 1,\dots,L$ и $\{n_1,\dots,n_M\} = \bigcup\limits_{j=1}^{L} \mathcal{M}_j$, где $M = \sum\limits_{j=1}^{L} M_j$. Здесь элементы набора $(n_1,\dots,n_M)$ соответствуют номерам ненулевых векторов в последовательности (1). Положим $\mu_0{=}0$ и $\mu_j = \mu_{j-1} + M_j,\ j = 1,\dots,L$. Допустим, что

$$\mathcal{M}_j = \{n_m \colon \mu_{j-1} + 1 \leqslant m \leqslant \mu_j\},\ j = 1,\dots,L. \tag{2}$$

Кроме того, допустим, что справедливы неравенства для всех $m = 2,\dots,M$:

$$1 \leqslant T_{\min} \leqslant n_m - n_{m-1} \leqslant T_{\max} \leqslant N - 1, \tag{3}$$

где $T_{\min}$ и $T_{\max}$ — натуральные числа.

Ограничения (3) устанавливают допустимый интервал между двумя ближайшими номерами

*Работа выполнена при финансовой поддержке РФФИ, проекты № 09-01-00032 и № 07-07-00022.

ненулевых векторов в последовательности (1). Эти ограничения можно трактовать как условие квазипериодичности повторов ненулевых векторов в последовательности (1).

Элементы набора $(u_1, \ldots, u_L)$ будем интерпретировать как информационно значимые векторы. Из (1)–(3) видно, что последовательность x_n включает L участков, каждый из которых содержит серию идентичных информационно-значимых векторов (j-я серия, $j = 1, \ldots, L$, образована вектором u_j); значения $\mu_{j-1} + 1$ и μ_j, $j = 1, \ldots, L$, определяют порядковые номера векторов, начинающих и завершающих j-ю серию.

Пусть

$$y_n = x_n + e_n,\ n \in \mathcal{N}, \tag{4}$$

где e_n — вектор помехи (ошибки измерения), независимый от вектора x_n. Будем считать, что последовательность y_n, $n \in \mathcal{N}$, доступна для наблюдения.

Из (1)–(3) следует, что последовательность x_n, $n \in \mathcal{N}$, зависит от трех наборов, а именно: $x_n{=}x_n(n_1, \ldots, n_M, \mu_1, \ldots, \mu_L, u_1, \ldots, u_L)$. Положим

$$S(n_1, \ldots, n_M, \mu_1, \ldots, \mu_L, u_1, \ldots, u_L) = \\ = \sum_{n \in \mathcal{N}} \left\| y_n - x_n \right\|^2, \tag{5}$$

где $\|\cdot\|$ — норма вектора, и рассмотрим следующие задачи среднеквадратического приближения.

Задача 1. *Дано*: последовательность $y_n \in \mathbb{R}^q$, $n \in \mathcal{N}$, структура которой описывается формулами (1)–(4), набор $(u_1, \ldots, u_L)$ ненулевых векторов из $\mathbb{R}^q$ и натуральное число M.

Найти: наборы $(n_1, \ldots, n_M)$ и $(\mu_1, \ldots, \mu_L)$ такие, что целевая функция (5) минимальна при ограничениях (2) и (3).

Задача 2. *Дано*: последовательность $y_n \in \mathbb{R}^q$, $n \in \mathcal{N}$, структура которой описывается формулами (1)–(4), набор $(u_1, \ldots, u_L)$ ненулевых векторов из $\mathbb{R}^q$.

Найти: набор $(n_1, \ldots, n_M)$ номеров и его размерность M, а также набор и $(\mu_1, \ldots, \mu_L)$ такие, что целевая функция (5) минимальна при ограничениях (2) и (3).

Задачи 1 и 2 соответствуют проблеме оптимального обнаружения по критерию минимума суммы квадратов уклонений серий повторяющихся информационно значимых векторов в ненаблюдаемой последовательности, структура которой описывается формулами (1)–(3). Отличие этих задач состоит в том, что в первой из них число ненулевых информационно значимых векторов считается заданным, а во второй — неизвестным, т. е. является оптимизируемой величиной.

Положим $w = (u_1, \ldots, u_L)$. Допустим, в дополнение к (1)–(4), что $w \in W$, причем $|W| = K$, где

$$W \subset \Big\{ (u_1, \ldots, u_L) \;\Big|\; u_j \in \mathbb{R}^q,\ 0 < \|u_j\|^2 < \infty, \\ j = 1, \ldots, L,\ L \in \{1, \ldots, L_{\max}\} \Big\}. \tag{6}$$

Здесь W — множество (словарь) векторных наборов (слов) мощности K, размерность которых не превосходит $L_{\max}$.

Рассмотрим еще две задачи среднеквадратического приближения.

Задача 3. *Дано*: множество W, $|W| = K$, наборов векторов из $\mathbb{R}^q$, последовательность $y_n \in \mathbb{R}^q$, $n \in \mathcal{N}$, структура которой описывается формулами (1)–(4) и (6), а также натуральное число M.

Найти: векторный набор $w \in W$ такой, что целевая функция (5) минимальна на множестве допустимых наборов $(n_1, \ldots, n_M)$ и $(\mu_1, \ldots, \mu_L)$.

Задача 4. *Дано*: множество W, $|W| = K$, наборов векторов из $\mathbb{R}^q$, последовательность $y_n \in \mathbb{R}^q$, $n \in \mathcal{N}$, структура которой описывается формулами (1)–(4) и (6).

Найти: векторный набор $w \in W$ такой, что целевая функция (5) минимальна на множестве допустимых наборов $(n_1, \ldots, n_M)$ и $(\mu_1, \ldots, \mu_L)$.

Задачи 3 и 4 отражают сущность проблемы распознавания последовательностей, включающих серии идентичных векторов, скрытых в ненаблюдаемой последовательности (1). В задаче 3 число ненулевых векторов в последовательности считается заданным, а в задаче 4 — неизвестным.

Легко установить, что если $\{e_n\}$ в формуле (4) есть выборка из q-мерного нормального распределения с параметрами $(0, \sigma^2 I)$, где I — единичная матрица, а в качестве критерия решения задачи использовать максимум функционала правдоподобия, то статистический подход к проблемам обнаружения и распознавания приводит к минимизации функциии (5) и к таким же сформулированным выше четырем задачам.

Редуцированные задачи

Раскрывая квадрат нормы в (5), получим

$$S{=}\sum_{n \in \mathcal{N}} \|y_n\|^2 {+} \sum_{j=1}^{L} M_j \|u_j\|^2 {-} 2 \sum_{j=1}^{L} \sum_{n \in \mathcal{M}_j} \langle y_n, u_j \rangle {=} \\ = \sum_{n \in \mathcal{N}} \|y_n\|^2 - \sum_{j=1}^{L} \sum_{m=\mu_{j-1}+1}^{\mu_j} \left(2\langle y_{n_m}, u_j \rangle - \|u_j\|^2 \right),$$

где $\langle \cdot, \cdot \rangle$ — скалярное произведение.

Поскольку первое слагаемое в правой части полученного выражения — константа, имеем следующие редуцированные оптимизационные задачи, к которым сводятся задачи 1 и 2.

Задача SSIVS-F (Searching for Series of Identical Vectors in a Sequence, when M is Fixed).

Дано: последовательность $y_0, \ldots, y_{N-1}$ векторов из $\mathbb{R}^q$, набор $(u_1, \ldots, u_L)$ ненулевых векторов из $\mathbb{R}^q$ и натуральное число M.

Найти: наборы $(n_1, \ldots, n_M)$ и $(\mu_1, \ldots, \mu_L)$ такие, что

$$\sum_{j=1}^{L} \sum_{m=\mu_{j-1}+1}^{\mu_j} \big(\|u_j\|^2 - 2\langle y_{n_m}, u_j\rangle\big) \to \max \qquad (7)$$

при ограничениях (2) и (3).

Задача SSIVS-NF (Searching for Series of Identical Vectors in a Sequence, when M is Not Fixed).

Дано: последовательность $y_0, \ldots, y_{N-1}$ векторов из $\mathbb{R}^q$ и набор $(u_1, \ldots, u_L)$ ненулевых векторов из $\mathbb{R}^q$.

Найти: набор $(n_1, \ldots, n_M)$ номеров, его размерность M и набор $(\mu_1, \ldots, \mu_L)$ такие, что выполняется (7) при ограничениях (2) и (3).

Точные полиномиальные алгоритмы решения этих редуцированных оптимизационных задач обоснованы в [1, 2]. Трудоемкости алгоритмов решения задач SSIVS-F и SSIVS-NF есть величины $\mathcal{O}\big(LM(T_{\max} - T_{\min} + q)N\big)$ и $\mathcal{O}\big(L(T_{\max} - T_{\min} + q)N\big)$ соответственно.

Задачи 3 и 4 сводятся к решению следующих экстремальных задач.

Задача SVTVP-F (Searching for a Vector Tuple in the Vocabulary of Patterns, when M is Fixed).

Дано: последовательность $y_0, \ldots, y_{N-1}$ векторов из $\mathbb{R}^q$, натуральное число M и множество (словарь) W, $|W| = K$, упорядоченных наборов векторов из $\mathbb{R}^q$.

Найти: векторный набор $w \in W$ такой, что выполняется (7), при ограничениях (2) и (3).

Задача SVTVP-NF (Searching for a Vector Tuple in the Vocabulary of Patterns, when M is Not Fixed).

Дано: последовательность $y_0, \ldots, y_{N-1}$ векторов из $\mathbb{R}^q$ и множество (словарь) W, $|W| = K$, упорядоченных наборов векторов из $\mathbb{R}^q$.

Найти: векторный набор $w \in W$ такой, что выполняется (7), при ограничениях (2) и (3).

Точные полиномиальные алгоритмы решения этих экстремальных задач обоснованы в [3, 4]. Временные сложности алгоритмов решения задач SVTVP-F и SVTVP-NF есть величины $\mathcal{O}\big(KML_{\max}(T_{\max} - T_{\min} + q)N\big)$ и $\mathcal{O}\big(KL_{\max}(T_{\max} - T_{\min} + q)N\big)$ соответственно.

На основе алгоритмов решения редуцированных задач SSIVS-NF, SSIVS-F, SVTVP-NF, а также SVTVP-F построены алгоритмы помехоустойчивого анализа и распознавания структурированных последовательностей, включающих серии идентичных вектор-фрагментов. Эти алгоритмы гарантируют оптимальность решения как по критерию минимума суммы квадратов уклонений, так и по критерию максимального правдоподобия в случае, когда помеха аддитивна и является гауссовской последовательностью независимых одинаково распределенных величин.

Численное моделирование

Ниже, в качестве примера, приведены результаты численных экспериментов, демонстрирующие работу алгоритмов и сущность рассмотренных задач для одномерных последовательностей.

На рис. 1 а изображена сгенерированная последовательность X, включающая 3 серии информационно-значимых фрагментов. На рис. 1 б представлена последовательность Y, подлежащая обработке (в этом примере уровень помехи превышает уровень сигнала). На рис. 1 в приведена последовательность $\hat{X}$, полученная с помощью алгоритма обнаружения, в условиях, когда число M неизвестно. Числовые данные под графиками соответствуют заданным (рис. 1 а) и найденным (рис. 1 в) начальным номерам фрагментов, завершающих серии, в скобках указано число фрагментов в серии. Рисунок иллюстрирует работу алгоритма в условиях интенсивных помех.

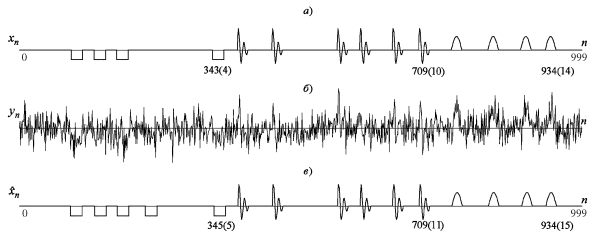


Рис. 1.

На рис. 2 и 3 приведены кривые оценок нормированной среднеквадратичной ошибки $e_M(\sigma) = \mathsf{E}\|X - \hat{X}\|^2/e^u$, где E — символ математического ожидания, e^u — оценка сверху для $\|X - \hat{X}\|^2$, а M — общее число фрагментов в обрабатываемой последовательности. Эти кривые получены при обработке одних и тех же $2 \cdot 10^4$ сгенерированных последовательностей, каждая из которых включала от 2 до 5 фрагментов, разбитых на 2 серии; начальные номера фрагментов, а также границы серий генерировались с помощью датчика случайных чисел. Каждая точка экспериментальной кривой получена по результатам обработки $5 \cdot 10^3$ последовательностей; номер кривой соответствует общему числу фрагментов в обрабатываемой последовательности.

На рис. 2 приведены кривые оценок $e_M(\sigma)$, $M = 2, \ldots, 5$, полученные с помощью алгоритма обнаружения, который использует априорную информацию о числе фрагментов в последователь-

ности. Оценки на рис. 3 получены при обработке тех же самых последовательностей с помощью алгоритма, ориентированного на обработку последовательностей в условиях, когда число фрагментов неизвестно.

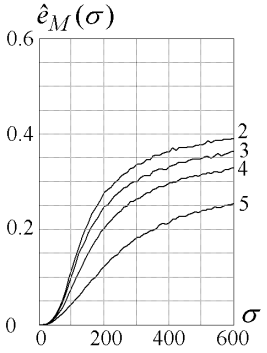


Рис. 2.

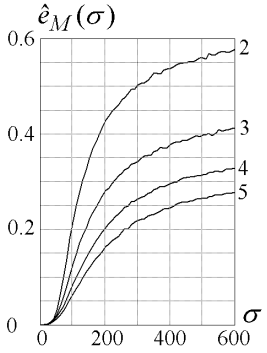


Рис. 3.

Рис. 4 и 5 иллюстрируют зависимость от уровня помехи вероятности ошибки распознавания последовательностей, порядок следования серий, в которых определялся двумя различными векторными наборами для случаев, когда общее число фрагментов является (рис. 4) и не является (рис. 5) частью входа задачи. Теоретические оценки верхней и нижней границ вероятности ошибки распознавания $\alpha^u(\sigma)$ и $\alpha^d(\sigma)$ в виде графиков приведены под номерами 1 и 6. Кривые 2–5 на рис. 4 и 5 получены при распознавании последовательностей, включавших от 2 до 5 фрагментов соответственно.

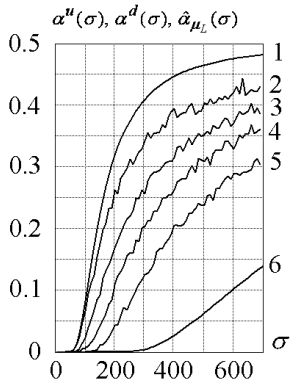


Рис. 4.

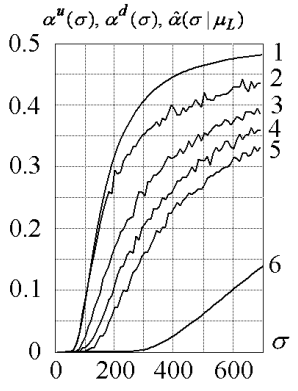


Рис. 5.

Оценка вероятности ошибки распознавания при каждом значении σ подсчитана по формуле $\hat{\alpha} = (\nu_1 + \nu_2)/2$, где ν_1 и ν_2 — числа неверно опознанных последовательностей, сгенерированных по каждому эталонному набору. Моделировалась байесовская процедура принятия решения с равновероятными гипотезами (наборами). Каждая точка экспериментальной кривой $\hat{\alpha}$ получена в результате усреднения $5 \cdot 10^3$ значений. Рис. 2–5 показывают, что ошибка обнаружения и вероятность ошибки распознавания меньше в случае, когда число ненулевых фрагментов в последовательности известно, по сравнению со случаем, когда это число неизвестно.

Заключение

Рассмотренные задачи входят в большое семейство актуальных задач [5], к которым сводятся типовые проблемы помехоустойчивого off-line анализа и распознавания структурированных данных в виде числовых и векторных последовательностей, включающих повторяющиеся, чередующиеся и перемежающиеся информационно значимые векторы или фрагменты. В настоящей работе представлены эффективные алгоритмические решения четырех ранее не изученных задач из этого семейства.

Важным, но пока не изученным, является вопрос о разрешимости обобщения рассмотренных задач обнаружения на тот случай, когда векторный набор, определяющий порядок следования серий, не фиксирован, а лишь содержит известное число векторов из заданного конечного алфавита. Алгоритмы решения этих задач представляют значительный интерес для ряда упомянутых во введении приложений.

Литература

[1] *Кельманов А. В., Михайлова Л. В.* Совместное обнаружение в квазипериодической последовательности заданного числа фрагментов из эталонного набора и ее разбиение на участки, включающие серии одинаковых фрагментов // Журн. вычисл. матем. и матем. физики. — 2006. — Т. 46, № 1. — С. 172–189.

[2] *Кельманов А. В., Михайлова Л. В.* Апостериорное обнаружение квазипериодических фрагментов из эталонного набора в числовой последовательности и ее разбиение на участки, включающие серии одинаковых фрагментов // Журн. вычисл. матем. и матем. физики. — 2008. — Т. 48, № 5. — С. 168–184.

[3] *Кельманов А. В., Михайлова Л. В.* Распознавание числовой последовательности, включающей серии квазипериодически повторяющихся эталонных фрагментов. Случай известного числа фрагментов // Сиб. журн. индустр. математики. — 2005. — Т. 8, № 3(23). — С. 69–86.

[4] *Кельманов А. В., Михайлова Л. В.* Распознавание числовой последовательности, включающей серии квазипериодически повторяющихся эталонных фрагментов // Сиб. журн. индустр. математики. — 2007, — Т. 10, № 4(32). — С. 61–75.

[5] Система QPSLab для решения задач компьютерного анализа и распознавания числовых последовательностей с квазипериодической структурой. — 2008. — `http://math.nsc.ru/~serge/qpsl`.

Эффективный алгоритм над множеством алгоритмов линейной локальной фильтрации*

Мясников В. В.
vmyas@smr.ru
Самара, Институт систем обработки изображений РАН

Рассматривается проблема построения вычислительно эффективного алгоритма линейной локальной фильтрации (ЛЛФ) сигналов и изображений. Вводятся алгебраическая система алгоритмов ЛЛФ и операция расширения опорного множества алгоритмов ЛЛФ. В качестве эффективного алгоритма ЛЛФ предлагается использовать индуцированный алгоритм, определяемый как алгоритм из расширения с наименьшей вычислительной сложностью. Устанавливаются свойства индуцированного алгоритма и требования к алгоритмам опорного множества. Представлена структура метода построения эффективного индуцированного алгоритма.

Построение вычислительно эффективных процедур ЛЛФ, то есть процедур вычисления линейной свертки входного цифрового сигнала с конечным ядром, называемым конечной импульсной характеристикой (КИХ) фильтра, — одно из наиболее исследованных направлений в теории цифровой обработки сигналов. Хорошо известны работы следующих авторов: В. А. Виттих, Л. М. Гольденберг, В. Г. Лабунец, Б. Д. Матюшкин, В. В. Сергеев, В. А. Сойфер, А. М. Трахтман, Л. П. Ярославский, R. E. Blahut, R. E. Bogner, A. G. Constantinides, B. Gold, A. V. Oppenheim, L. R. Rabiner, C. M. Rader, R. W. Schafer, D. E. Dudgeon, R. Mersereau, R. W. Hamming, G. Nussbaumer, и др. Известно, что реализация операций ЛЛФ выполняется с использованием алгоритмов одного из трех типов: прямого вычисления свертки $\mathbf{A}_{DC}$ [8, 10], быстрого вычисления свертки $\mathbf{A}_{FC}$, основанных на алгоритмах быстрых дискретных ортогональных преобразованиях типа БПФ [1, 3, 6, 7, 8, 9] и рекурсивных алгоритмов $\mathbf{A}_{RF}$ [8, 10].

К сожалению, изобилие алгоритмов вычисления сверток и методов их построения не решает основную практическую проблему: *как для конкретной задачи ЛЛФ получить наилучший алгоритм ее решения.* В целом это выражается в том, что существующие подходы не могут гарантировать, что полученный с их помощью алгоритм ЛЛФ будет обладать наилучшими вычислительными (временными) характеристиками среди всех известных и/или доступных пользователю алгоритмов.

Основным недостатком известных подходов является игнорирование доступной информации о решаемой задаче ЛЛФ. В частности, при построении алгоритмов обычно игнорируется тот факт, что КИХ на практике является заранее известной и фиксированной. Кроме того, заранее может быть доступна некоторая информация об обрабатываемом сигнале. Наконец, профессиональный разработчик обычно имеет в своем распоряжении множество (библиотеку) реализованных в виде программ алгоритмов ЛЛФ. Относительно таких алгоритмов известна вычислительная сложность их применения к конкретным задачам ЛЛФ. Эти алгоритмыпрограммы ЛЛФ как отдельно, так и совместно, могут быть использованы для построения «наилучшего» алгоритма ЛЛФ. В этом смысле использование некоторого быстрого алгоритма, лучшего для конкретной задачи ЛЛФ, может оказаться не единственно возможным и не самым лучшим решением. Предлагаемый в настоящей работе подход к построению эффективного алгоритма свободен от обозначенных выше недостатков. Для упрощения изложения описание дается в одномерном случае.

Следует также отметить, что попытки использования множеств алгоритмов ЛЛФ, реализованных в виде программ или исполняемых библиотек, для построения «наилучшего» алгоритма ЛЛФ в работах, известных автору, не предпринимались. Однако, сама идея использования набора алгоритмов для построения «наилучшего» в некотором смысле алгоритма новой не является. Около 30 лет назад она была предложена академиком Ю. И. Журавлевым [2] и используется в работах его школы [5] до настоящего времени для построения «наилучшего» (корректного) алгоритма распознавания над множеством некорректных (эвристических) алгоритмов. Следует отметить, что данная работа развивает указанную идею применительно к задаче построения вычислительно эффективных процедур ЛЛФ, но использует другое формализованное представление алгоритма и, как следствие, иную алгебраическую систему.

Алгебраическая система алгоритмов ЛЛФ

Рассмотрим задачу ЛЛФ $Z \equiv Z\big(\mathfrak{I}_0, \{x(n)\}_{n=0}^{N-1}\big)$, заключающуюся в вычислении одномерной (линейной) свертки конечного *входного сигнала* $\{x(n)\}_{n=0}^{N-1}$

*Работа выполнена при финансовой поддержке РФФИ (проекты № 06-01-00616-а, № 09-01-00434–а) и Программы фундаментальных исследований Президиума РАН «Фундаментальные проблемы информатики и информационных технологий», проект № 2.12.

длины N и КИХ $\{h(m)\}_{m=0}^{M-1}$ фильтра длины M:

$$y(n)=\sum_{m=0}^{M-1}h(m)\,x(n-m),\qquad n=M-1,\ldots,N-1. \quad (1)$$

Результатом решения задачи Z является *выходной сигнал* $\{y(n)\}_{n=0}^{N-M+1}$. Пусть отсчеты всех сигналов являются элементами некоторого коммутативного кольца $\mathbf{K}$ с единицей. Величина $\mathfrak{I}_0=\left(\{h(m)\}_{m=0}^{M-1},\mathfrak{I}_x\right)$ есть априорная информация о задаче Z, $\mathfrak{I}_x=\left(N,\mathfrak{I}_{(x)}\right)$ — априорная информация о входном сигнале, $\mathfrak{I}_{(x)}$ — априорная информация о свойствах входного сигнала, задаваемая в виде набора распределений вероятности, характеризующих согласованность входных сигналов определенным конечно-разностным уравнениям [4]. На практике задача Z формулируется при следующих *ограничениях*:

- $M<N;\ h(0)\neq 0,\ h(M-1)\neq 0$;
- отсчеты КИХ $\{h(m)\}_{m=0}^{M-1}$ и величина N известны до решения Z;
- до решения Z может быть известна информация о свойствах входного сигнала $\mathfrak{I}_{(x)}$;
- отсчеты входного сигнала $\{x(n)\}_{n=0}^{N-1}$ известны только в момент решения задачи Z.

На метод построения эффективного алгоритма ЛЛФ накладываются *дополнительные требования.* Они выражаются в том, что метод:

- учитывает ограничения задачи Z;
- использует доступную априорную информацию $\mathfrak{I}_0$ о задаче;
- использует задаваемое множество $\mathbf{A}$, называемое далее *опорным* множеством, алгоритмов ЛЛФ (на практике алгоритмы опорного множества обычно реализованы в виде программ);
- гарантирует, что конструируемый алгоритм удовлетворяет *требованию эффективности* над опорным множеством (см. ниже);
- допускает полностью автоматическое построение эффективного алгоритма.

Требование эффективности над опорными множеством по отношению к алгоритму ЛЛФ означает, что этот алгоритм в вычислительном плане:

- для любой задачи Z оказывается не хуже наилучшего алгоритма опорного множества (свойство *эффективности*);
- для некоторых задач Z оказывается лучше наилучшего алгоритма опорного множества (свойство *строгой эффективности*).

Определение процесса построения эффективного алгоритма ЛЛФ заключается в конструировании отображения

$$(\mathfrak{I}_0,\mathbf{A})\to A^{\mathfrak{I}}, \quad (2)$$

которое для фиксированного опорного множества $\mathbf{A}$ алгоритмов ЛЛФ и заданной априорной информации $\mathfrak{I}_0$ о задаче ЛЛФ (1) дает эффективный алгоритм $A^{\mathfrak{I}}$ ее решения.

Для конструирования искомого отображения (2) введем *алгебраическую систему алгоритмов* ЛЛФ. Для этого определим отношения между алгоритмами и операции с ними. В качестве *формализованного определения алгоритма* ЛЛФ используем тройку $\left(A,\aleph_A,\{U(A(Z))\}_{Z\in\aleph_A}\right)$, где:

- A — реализация алгоритма решения задач ЛЛФ (на практике — программа);
- $\aleph_A=\{Z\colon Z\in\aleph\wedge!A(Z)\}$ — *область определения* (ОО) алгоритма A, то есть множество задач, для которых алгоритм A применим;
- $U(A(Z))$ — (полная) сложность решения задачи Z алгоритмом A, задаваемая в виде:

$$U(A(Z))=\xi_{\text{add}}U_{\text{add}}(A(Z))+\xi_{\text{mul}}U_{\text{mul}}(A(Z)),\qquad(\xi_{\text{add}}+\xi_{\text{mul}}=1;\ \ \xi_{\text{add}},\xi_{\text{mul}}\in\mathbb{R}).$$

Здесь $U_{\text{add}}(A(Z))$ и $U_{\text{mul}}(A(Z))$ — число сложений и умножений, требуемых в алгоритме A для решения задачи Z. Удельная сложность алгоритма:

$$u(A(Z))=(N-M+1)^{-1}U(A(Z)).$$

Отношения между алгоритмами: тождественность, подобие (для алгоритмов совпадают их аналитические изображения), эквивалентность (совпадают сложности решения задач ЛЛФ), лучше (по вычислительной сложности), хуже и т. д.

Операции с алгоритмами:

- $E\left(A;\aleph_{\widetilde{A}}\right)$ — *распространение* алгоритма A с ОО $\aleph_A$ на ОО $\aleph_{\widetilde{A}}\supseteq\aleph_A$.
- $\tau\left(A;\aleph_{\widetilde{A}}\right)$ — *сужение алгоритма* A с ОО $\aleph_A$ на ОО $\aleph_{\widetilde{A}}\subseteq\aleph_A$ (обратная операция к операции распространения).
- $A+\widetilde{A}$ — *сумма алгоритмов*, результат которой определяется выражением:

$$\begin{cases}A(Z),\ \left(Z\in\aleph_A\setminus\aleph_{\widetilde{A}}\right)\vee\\ \vee\left(\left(Z\in\aleph_A\cap\aleph_{\widetilde{A}}\right)\wedge\left(U(A(Z))<U(\widetilde{A}(Z))\right)\right),\\ \widetilde{A}(Z),\ \left(Z\in\aleph_{\widetilde{A}}\setminus\aleph_A\right)\vee\\ \vee\left(\left(Z\in\aleph_A\cap\aleph_{\widetilde{A}}\right)\wedge\left(U(A(Z))\geqslant U(\widetilde{A}(Z))\right)\right).\end{cases}$$

Алгоритм $A^{\oplus}\equiv\sum\limits_{A\in\mathbf{A}}A$ далее назовем *компетентным* (над множеством $\mathbf{A}$). В докладе приводятся свойства введенных операций.

Расширение множества алгоритмов ЛЛФ. Индуцированный алгоритм

В дополнении к указанным операциям определим *расширение множества алгоритмов*. Построение расширения $[\mathbf{A}]$ для множества алгоритмов $\mathbf{A}$

осуществляется с использованием эквивалентного преобразования выражения (1) в виде [4]:

$$y(n) = \sum_{t=1}^{K_h+K_x-2} g(t) y(n-t) + \\ + \sum_{s=0}^{S-1} \sum_{m \in D_s} \tilde{h}_s(m) \sum_{l=0}^{K_x-1} g_x(l)\, x(n-m-l), \\ n = 0, \ldots, N-1. \quad (3)$$

Здесь $\{g_h(k)\}_{k=0}^{K_h-1}$ и $\{g_x(k)\}_{k=0}^{K_x-1}$ — отсчеты КИХ-фильтров предварительной обработки, соответственно, для КИХ $\{h(m)\}$ и входного сигнала $(g_h(0) = g_x(0) = 1)$; $g(k) = g_h(k) * g_x(k)$; $\{D_s\}_{s=0}^{S-1}$, $D_s = \overline{d_0^s, d_1^s}$ — допустимое покрытие ОО дискретно заданной функции $\{\tilde{h}(m)\}_{m=0}^{M+K_h-2}$, где $\tilde{h}(m) = h(m) * g_h(m)$. Отсчеты выходного сигнала в выражении (3) на интервале $M-1, \ldots, N-1$ совпадают с искомыми отсчетами выражения (1) в силу эквивалентности используемого преобразования.

Вычисление выражения (3) может быть произведено в рамках модели CR алгоритма ЛЛФ, формально следующей из этого выражения [4]. Модель CR определена с точностью до числовых и алгоритмических параметров. К *алгоритмическим параметрам* относятся алгоритм предварительной обработки входного сигнала A_{prep} и набор алгоритмов $\{A_s\}_{s=0}^{S-1}$, выполняющих вычисление S сверток в выражении (3) (второе слагаемое). Все алгоритмы являются элементами опорного множества $\mathbf{A}$.

Вычислительная сложность алгоритма модели CR может быть представлена в виде функции числовых и алгоритмических параметров модели [4]. Способ построения расширения опорного множества алгоритмов ЛЛФ задается определениями 1 и 2, приведенными ниже.

Определение 1. *Расширением по модели CR порядка (K_h, K_x, S) множества алгоритмов $\mathbf{A}$ на задаче Z (далее — просто расширением $CR(K_h, K_x, S)$) называется множество алгоритмов модели CR, обозначаемое $[\mathbf{A}]_{CR(K_h,K_x,S)}$, заданного порядка (K_h, K_x, S), с допустимыми значениями остальных числовых параметров и алгоритмами $A_{prep}, \{A_s\}_{s=0}^{S-1}$ из опорного множества $\mathbf{A}$, для которых задачи Z_{prep} и $\{Z_s\}_{s=0}^{S-1}$ попадают в их область определения.*

Определение 2. *Расширением по модели CR (далее — расширением) множества алгоритмов $\mathbf{A}$ на задаче Z называется множество:*

$$[\mathbf{A}] = \bigcup_{\substack{K_h=1,\ldots,M-1 \\ K_x=1,\ldots,N-1 \\ S=1,2\ldots}} [\mathbf{A}]_{CR(K_h,K_x,S)}.$$

Теперь задача построения отображения (2), то есть задача построения эффективного алгоритма, может быть конкретизирована.

Определение 3. *Алгоритм $A^{\mathfrak{I}}(Z) \in [A]$ называется алгоритмом, индуцированным априорной информацией $\mathfrak{I}_0$ задачи Z, если $A^{\mathfrak{I}}(Z)$ является наилучшим алгоритмом для задачи Z в расширении $[\mathbf{A}]$, и, кроме того, в $[\mathbf{A}]$ нет другого алгоритма меньшего порядка со сложностью, равной сложности алгоритма $A^{\mathfrak{I}}(Z)$.*

Можно доказать ряд утверждений, которые характеризуют общие свойства индуцированного алгоритма $A^{\mathfrak{I}}(Z)$ [4]:

- $U\left(A^{\mathfrak{I}}(Z)\right) = \min\limits_{A(Z) \in [\mathbf{A}]} U(A(Z))$;
- $\forall Z$ $A^{\mathfrak{I}}(Z)$ является эффективным над $\mathbf{A}$;
- $A^{\mathfrak{I}}(Z) \in [\mathbf{A}]_{CR(K_h,K_x,S)} \subseteq [\mathbf{A}]$ является строго эффективным над опорным множеством $\mathbf{A}$ для задачи Z тогда и только тогда, когда порядок модели $(K_h, K_x, S) \neq (1, 1, 1)$.

Получаем, что задача построения отображения (2) может быть сформулирована как задача нахождения параметров наилучшего алгоритма в расширении. Легко показать, что решение этой задачи — индуцированный алгоритм $A^{\mathfrak{I}}(Z)$ — по своему определению и доказанным свойствам удовлетворяет всем дополнительным требованиям, исключая требование полностью автоматического построения эффективного алгоритма. Это требование налагается на разрабатываемый метод решения задачи определения параметров алгоритма $A^{\mathfrak{I}}(Z)$. Основными вопросами, возникающими при решении этой задачи, являются:

- способ нахождения параметров индуцированного алгоритма для конкретной задачи Z и заданного опорного множества алгоритмов;
- состав опорного множества (на практике обычно достаточно, чтобы индуцированный алгоритм был эффективен над *множеством алгоритмов из основных классов* $\mathbf{A}_{DC} \cup \mathbf{A}_{FC} \cup \mathbf{A}_{RF}$: прямой, быстрой свертки и рекурсивных).

В силу ограниченности объема тезисов более подробное освещение этих вопросов приводится в докладе. Ниже приведен окончательный результат.

Эффективный алгоритм над множеством алгоритмов постоянной сложности

Определим функцию $\mathrm{par}(Z)$, которая для конкретной задачи Z возвращает пару (M, N) ее параметров, и рассмотрим подмножество алгоритмов ЛЛФ, задаваемое следующим определением.

Определение 4. *Алгоритм ЛЛФ A называется алгоритмом постоянной сложности (АПС), если выражение для сложности этого алгоритма на всей области определения $\aleph_A$ представимо в виде $u(A(Z)) = u_A(\mathrm{par}(Z))$. Алгоритмы, не являющиеся АПС, называются алгоритмами вариантной сложности.*

Область определения для АПС может быть задана путем указания множества пар индексов:

$$\aleph_{A(M,N)} = \{(M,N) : \aleph(M,N) \subseteq \aleph_A\}.$$

Для АПС естественным образом конкретизируются способы построения операций сужения и сложения. Для построения *операции распространения* АПС вводятся три *базовых способа распространения АПС*: путем разбиения КИХ, путем разбиения входного сигнала, путем решения «суперзадачи». Возможность построения распространения АПС на любую наперед заданную ОО позволяет сопоставить реальную сложность АПС в конкретной точке $(\tilde{M},\tilde{N})$ его ОО и сложность распространения $E\big(\tau\big(A;\aleph_{A(M,N)} \setminus (\tilde{M},\tilde{N})\big);\ \aleph_{A(M,N)}\big)$ в этой точке. Естественным представляется требование к АПС, в соответствие с которым реальная сложность АПС должна быть ниже сложности его распространения. Сложность АПС, удовлетворяющая этому требованию называется *корректной*. В общем случае не любой АПС имеет корректную функцию сложности. В работе для АПС вводится дополнительная операция $R(\ldots)$, называемая *итерационной операцией приведения*. Эта операция позволяет получить из существующего АПС новый АПС (называемый *приведенным*) с корректной и, как следствие, меньшей по значениям функцией сложности.

В работе доказываются утверждения, устанавливающие следующий факт: если нас интересует эффективный алгоритм над множеством алгоритмов из основных классов (прямой, быстрой свертки и рекурсивными), то для его построения достаточно построить индуцированный алгоритм над множеством из единственного алгоритма — приведенного компетентного алгоритма (ПКА), который построен над множеством АПС (прямой и быстрой свертки).

Тогда *структура метода построения эффективного алгоритма* оказывается следующей:

— *операция 1* — построение компетентного алгоритма $A^{\oplus} = \sum\limits_{A\in\mathbf{A}} A$ над предоставленным опорным множеством АПС $\mathbf{A} \subseteq \mathbf{A}_{DC} \cup \mathbf{A}_{FC}$ (в опорное множество обязательно включен АПС с требуемой областью определения, то есть $A_{DC} \in \mathbf{A}_{DC}$);

— *операция 2* — построение ПКА с помощью итерационной операции приведения компетентного алгоритма $\breve{A}^{\oplus} = R(A^{\oplus})$;

— *операция 3* — построение алгоритма $A^{\mathfrak{I}}_{\oplus}(Z) \in \big[\{\breve{A}^{\oplus}\}\big]$, индуцированного априорной информацией задачи Z, над множеством $\{\breve{A}^{\oplus}\}$ из единственного ПКА $\breve{A}^{\oplus}$.

Для заданного опорного множества $\mathbf{A}$ АПС первые две операции полностью формализованы и определены. Выполнение третьей операции производится численными методами и зависит от алгебраических и функциональных свойств сигналов.

В докладе приводятся результаты исследования предлагаемого метода, направленные на получение численных оценок вычислительного выигрыша от его использования в реальных задачах ЛЛФ.

Выводы

Проблема построения вычислительно эффективного алгоритма ЛЛФ формализована в виде задачи построения индуцированного алгоритма над множеством алгоритмов ЛЛФ постоянной сложности. Показано, что процесс построения индуцированного алгоритма может быть представлен в виде трех последовательных и хорошо формализованных операций. Построенный эффективный индуцированный алгоритм ЛЛФ в общем случае сочетает в себе свойства алгоритмов ЛЛФ основных классов: прямой, быстрой свертки и рекурсивных.

Литература

[1] *Агаян С. С.* Успехи и проблемы быстрых ортогональных преобразований // Распознавание, классификация, прогноз, М: Наука, 1990. Вып.3. — С. 146–214.

[2] *Журавлев Ю. И.* Об алгебраическом подходе к решению задач распознавания или классификации // Проблемы кибернетики. — 1978. — № 33. — С. 5–68.

[3] *Лабунец В. Г.* Единый подход к алгоритмам быстрых преобразований // Применение ортогональных методов при обработке сигналов и анализе систем, Свердловск: УПИ, 1980. — С. 4–14.

[4] *Мясников В. В.* О синтезе эффективного алгоритма над множеством алгоритмов вычисления свертки // Компьютерная оптика. — 2006. — Вып. 29. — С. 78–117.

[5] *Рудаков К. В.* Об алгебраической теории универсальных и локальных ограничений для задач классификации // Распознавание, классификация, прогноз, 1988. — Т. 1. — С. 176–200.

[6] *Чернов В. М.* Арифметические методы синтеза быстрых алгоритмов дискретных ортогональных преобразований. — М: Физматлит, 2007. — 264 с.

[7] *Ahmed N., Rao K. R.* Orthogonal Transforms for Digital Signal Processing. — New York: Springer-Verlag, 1975.

[8] *Blahut R. E.* Fast Algorithms for Digital Signal Processing. — Reading, MA: Addison-Wesley, 1984.

[9] *Nussbaumer H. J.* Fast Fourier Transform and Convolution Algorithms. — Heidelberg, Germany: Springer, 1990. — 276 с.

[10] *Rabiner L. R., Gold B.* Theory and applications of digital signal processing. — Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1975.

О постановке и решении задачи построения эффективных линейных локальных признаков цифровых сигналов*

Мясников В. В.
vmyas@smr.ru
Самара, Институт систем обработки изображений РАН

Под линейным локальным признаком цифрового сигнала понимается пара, состоящая из конечной импульсной характеристики (КИХ) и алгоритма (вычисления признака), предназначенного для вычисления линейной свертки сигнала с КИХ. Эффективный линейный локальный признак обнаруживает оптимальное поведение: алгоритм обладают наименьшей вычислительной сложностью, и КИХ признака оказывается наилучшим образом согласована с заданным критерием качества. В работе предлагаются формализованная постановка и решение задачи построения эффективного линейного локального признака.

Во многих задачах обработки и анализа цифровых сигналов и изображений одним из основных предъявляемых требований к признакам являются *требования вычислительного характера* [1]. Они выражаются в том, чтобы *алгоритм расчета признаков существовал* и был *вычислительно эффективен*. Наряду с этими требованиями содержательные постановки задач обработки, распознавания и анализа цифровых сигналов дополнительно порождают целый ряд требований и ограничений к используемым признакам, которые, совместно с указанными выше являются не только противоречивыми, но и не имеют четкой математической формализации. Поэтому до настоящего времени во многих практических задачах процесс построения признаков (указания, как именно из объекта или явления получить ту или иную характеристику, и какие именно характеристики следует считать признаками) остается в значительной степени процедурой эвристической, существенным образов зависимой и от специфики предметной области, и от опыта и квалификации разработчика.

Ниже представлена одна из возможных формализаций этой проблемы в виде задачи построения одного важного для задач компьютерного зрения класса признаков — линейных локальных признаков цифровых сигналов. Под линейным локальным признаком цифрового сигнала понимается пара, состоящая из конечной импульсной характеристики (КИХ) и алгоритма (вычисления признака) [3]. Алгоритм вычисления признака реализует вычисление линейной свертки анализируемого сигнала с КИХ признаком. Эффективный линейный локальный признак обнаруживает оптимальное поведение в том смысле, что алгоритм вычисления признака обладают наименьшей (в классе) вычислительной сложностью, а КИХ признака наилучшим образом согласована с критерием качества. Таким образом, эффективные линейные локальные признаки призваны установить рациональный баланс между двумя противоположными группами признаков. К первой относятся признаки, оптимальные в смысле критерия качества задачи, но не имеющие быстрого алгоритма их вычисления (например, полученные с использованием преобразования Карунена–Лоэва). Ко второй группе — признаки, полученные с использованием быстрых алгоритмов, которые не имеют отношения к содержательной постановке задачи и критерию качества (например, признаки, полученные с использованием БПФ).

Настоящая работа опирается на результаты, связанные с построением эффективных алгоритмов линейной локальной фильтрации (ЛЛФ), которые первоначально были получены в авторских работах [2, 5] и изложены также в работе [4], представленной на настоящей конференции. Работа состоит из двух разделов. Первый раздел содержит известные сведения об эффективных линейных локальных признаках цифровых сигналов, представленные ранее в авторских работах [2, 3, 5], в нем также содержится формальная постановка задачи их построения. Второй раздел содержит новый материал, в котором представлена вычислительная процедура, используемая для решения задачи построения эффективных линейных локальных признаков.

Эффективные линейные локальные признаки и их семейства

Пусть $\mathbf{K}$ — коммутативное кольцо с единицей.

Определение 1. *Линейным локальным признаком (ЛЛП) длины M над $\mathbf{K}$ называется пара $\left(\{h(m)\}_{m=0}^{M-1}, A\right)$, где $\{h(m)\}_{m=0}^{M-1}$ — некоторая КИХ, задаваемая в виде конечной последовательности над $\mathbf{K}$ и удовлетворяющая ограничению $h(m) \neq 0$, $h(M-1) \neq 0$, а A — алгоритм ЛЛФ входного сигнала $\{x(n)\}_{n=0}^{N-1}$ над $\mathbf{K}$ с использованием КИХ $\{h(m)\}_{m=0}^{M-1}$:*

$$y(n) = \sum_{m=0}^{M-1} h(m)\, x(n-m),\ n = M-1, \ldots, N-1.$$

*Работа выполнена при финансовой поддержке РФФИ (проекты № 06-01-00616-а, № 09-01-00434–а) и Программы фундаментальных исследований Президиума РАН «Фундаментальные проблемы информатики и информационных технологий», проект № 2.12.

Задачу построения эффективного ЛЛП удобно определить, используя введенную в работах [2, 4, 5] формализацию задачи ЛЛФ. Тогда задача построения вычислительно эффективного ЛЛП может быть определена как задача конструирования отображения

$$(\mathfrak{I}_x, \mathbf{A}) \to \left(\{h(n)\}_{n=0}^{M-1}, A^{\mathfrak{I}}\right), \qquad (1)$$

где пара $\left(\{h(n)\}_{n=0}^{M-1}, A^{\mathfrak{I}}\right)$ связана отображением

$$(\mathfrak{I}_0, \mathbf{A}) \to A^{\mathfrak{I}},$$

определяющим задачу построения эффективного (индуцированного) алгоритма ЛЛФ по априорной информации $\mathfrak{I}_0 = \left(\{h(n)\}_{n=0}^{M-1}, \mathfrak{I}_x\right)$ о задаче ЛЛФ и опорному множеству $\mathbf{A}$ алгоритмов ЛЛФ, $\mathfrak{I}_x$ — *априорная информация о свойствах входного сигнала* [2, 4, 5].

Ниже рассматривается частный случай задачи (1) при следующих ограничениях, обычно выполняемых на практике:

$$\mathbf{A} = \{A_{\text{DC}}\}, \quad \mathfrak{I}_x = \varnothing, \qquad (2)$$

где A_{DC} — единственный алгоритм прямой свертки, формирующий опорное множество алгоритмов.

Легко показать, что даже с этими ограничениями задача построения отображения (1), называемая далее *общей задачей построения эффективного ЛЛП*, является некорректной.

Заметим, что ограничения (2) приводят к тому, что алгоритмы расширения $[\{A_{\text{DC}}\}]$ оказываются подобными рекурсивному алгоритму [2, 4, 5]. В этой ситуации идея конструирования отображения (1) заключается в ограничении класса рассматриваемых последовательностей *кусочно-однородными* (КО-) последовательностями. Идея построения эффективного ЛЛП заключается в построении такой КО-последовательности, для которой вычислительная сложность порождаемого ею алгоритма достигает своей нижней границы.

А именно, для КО-последовательности оказывается возможным указать явный вид алгоритма $A_{(K,S,\bar{a})}$, порождаемого КО-последовательностью типа $(K, S, \bar{a})$, где K — порядок ЛРС, которому удовлетворяет КО-последовательность, $\bar{a} = (a_1, \ldots, a_K)^{\intercal}$ — вектор коэффициентов ЛРС, $S-1$ — число дискретных интервалов последовательности, на которых это ЛРС остается однородным. Описание алгоритма представлено ниже (алгоритм 1). Его вычислительная сложность имеет вид:

$$u\left(A_{(K,S,\bar{a})}\right) \tfrac{N-M+1}{N} = |\Theta| + K - \xi_{\text{add}},$$

где Θ — множество отсчетов, в которых однородность ЛРС для КО-последовательности нарушается. В работе [3] показана справедливость следующего неравенства:

Алгоритм 1. Алгоритм $A_{(K,S,\bar{a})}$, порождаемый КО-последовательностью

1: предварительная обработка:
$\tilde{y}(n) = \sum\limits_{m\in\Theta} x(n-m)\,\tilde{\varphi}(m),\ n = 0, \ldots, N-1;$
2: окончательная обработка:
$y(n) = \sum\limits_{k=1}^{K} a_k y(n-t) + \tilde{y}(n),\ n = 0, \ldots, N-1.$

$$\max(S,K) + (K+1) - \xi_{\text{add}} \leqslant \leqslant \tfrac{N-M+1}{N} \max_{\bar{a}} u\left(A_{(K,S,\bar{a})}\right) \leqslant (S+2)\,K - \xi_{\text{add}}.$$

Нижняя граница вычислительной сложности в этом выражении достигается для подкласса линейных рекуррентных последовательностей (ЛРП), вводимых следующими двумя определениями.

Определение 2. *ЛРП $h(0), h(1), \ldots$ порядка K над $\mathbf{K}$ называется МС-последовательностью порядка K длины M над $\mathbf{K}$, если выполняется соотношение:*

$$(h(0) \neq 0) \wedge (h(M-1) \neq 0) \wedge \wedge (\forall m \geqslant M\ h(m) = 0) \wedge (|\Theta| \leqslant K+1). \qquad (3)$$

Определение 3. *МС-последовательность порядка K длины M над $\mathbf{K}$ называется нормализованной МС-последовательностью (НМС-последовательностью) порядка K длины M, если $h(0) = 1$ и выполняется условие*

$$\sum_{m\in\Theta} 2^m I\left(\tilde{\varphi}(m) \neq 0\right) - 2^{M+K} \sum_{m\in\Theta} I\left(\tilde{\varphi}(m) = 0\right) - - \tfrac{1}{2} I\left(\tilde{\varphi}(M+K-1) = 1\right) \to \min_{\substack{\{h(m)\}_{m=0}^{M-1} \\ \{\tilde{\varphi}(m)\}_{m\in\Theta}}}. \qquad (4)$$

В соответствии с определениями 2–3 вычислительная сложность алгоритма 1, порождаемого НМС-последовательностью порядка K длины M, удовлетворяет соотношению:

$$u(A)\,\tfrac{N-M+1}{N} \leqslant 2K.$$

Очевидно, что основная составляющая величины сложности, приведенная в правой части этого выражения, зависит только от порядка НМС-последовательности. Этот факт позволяет ввести понятие семейства НМС-последовательностей.

Определение 4. *$(K, M, \bar{a})$-семейством НМС-последовательностей, обозначаемым $\wp(K, M, \bar{a})$, называется множество НМС-последовательностей порядка K длины M, удовлетворяющих линейному рекуррентному соотношению с коэффициентами $\bar{a}$ $(a_K \neq 0)$.*

На принципиальные для задачи построения эффективных ЛЛП вопросы существования и единственности НМС-последовательности, числа последовательностей в семействе отвечают приведенные ниже теорема 1 и утверждение 2 [3].

Теорема 1. *(о существовании и единственности НМС-последовательности) Пусть* $M,\ K \in \mathbb{N}$, $M \geqslant K \geqslant 1$, $\bar{a}$ $(a_K \neq 0,\ a_k \in \mathbf{F})$ *и область отсчетов неоднородности* Θ *удовлетворяет соотношениям:*

$$|\Theta| = K+1,\ \ 0 \in \Theta, \quad M+K-1 \in \Theta. \qquad (5)$$

НМС-последовательность порядка K *длины* M *над полем* $\mathbf{F}$ *с указанными параметрами либо не существует, либо существует и единственна.*

Утверждение 2. *(о количестве НМС-последовательностей в семействе) Для любых* $M,\ K \in \mathbb{N}$: $M > K \geqslant 1$ *и любого* $\bar{a}$: $a_K \neq 0,\ a_k \in \mathbf{F}$

$$|\wp(K, M, \bar{a})| \leqslant C_{M+K-2}^{K-1}.$$

Введенный аппарат НМС-последовательностей позволяет конкретизировать общую *задачу построения эффективных ЛЛП*. Для этого в работе [3] вводится дополнительно в рассмотрение *производящий функционал* $\Psi : \mathbf{K}^M \to \mathbf{R}$. Под *производящим функционалом* понимается функция, которая для последовательности $\{h(0), \ldots, h(M-1)\}$ над $\mathbf{K}$ указывает числовую величину — степень «пригодности»: последовательность считается «лучше», если значение функционала на ней меньше.

Определение 5. *Частной задачей построения эффективного ЛЛП называется следующая задача: для параметров* $N, M, K \in \mathbb{N}$ *таких, что*

$$N > M, \quad K < \tfrac{N-M+1}{2N}(M - \xi_{\text{add}}),$$

коэффициентов ЛРС $\bar{a}$: $a_K \neq 0$ *и производящего функционала* $\Psi(\ldots)$ *построить ЛЛП* $\left(\{h(m)\}_{m=0}^{M-1}, A\right)$, *в котором:*

- *последовательность* $h(0), \ldots, h(M-1)$ *является НМС-последовательностью семейства* $\wp(K, M, \bar{a})$ *с минимальным значением производящего функционала:*

$$\Psi(h(0), \ldots, h(M-1)) \to \min_{\{h(m)\}_{m=0}^{M-1} \in \wp(K,M,\bar{a})};$$

- *алгоритм* $A \in [\{A_{DC}\}]$ *порождается НМС-последовательностью* $\{h(m)\}_{m=0}^{M-1}$.

Утверждение 3. *Пусть* $\Psi(\ldots)$ — *взаимнооднозначный производящий функционал. Если решение частной задачи построения эффективного ЛЛП существует, то оно единственно.*

В силу конечности семейства НМС-последовательностей, решение частной задачи можно получить, перебрав все последовательности семейства и выбрав ту из них, которая дает наименьшее значение производящего функционала. Таким образом, центральным моментом при построении эффективного ЛЛП является процедура построения НМС-последовательности из семейства $\wp(K, M, \bar{a})$ для заданной области отсчетов неоднородности Θ, удовлетворяющей условию (5).

Вычислительная процедура построения НМС-последовательности в задаче построения эффективного ЛЛП

Доказательство теоремы 1, приведенное в работе [3], в конструктивной форме отвечает на вопрос о существовании НМС-последовательности, то есть указывает способ построения искомой последовательности. Представленная ниже процедура для заданных входных данных либо дает решение в виде НМС-последовательности, либо (в соответствие с теоремой 1) указывает на невозможность ее построения. Входными данными для процедуры являются параметры семейства $\wp(K, M, \bar{a})$ и область отсчетов неоднородности Θ, для которой выполняются ограничения (5). Выходными данными процедуры являются значения отсчетов НМС-последовательности и значения отсчетов неоднородности в области Θ.

В процедуре используется СЛАУ (в тексте алгоритма — SLAU), первоначально задаваемая в виде:

$$\begin{aligned}
&h(0) = 1,\\
&h(m) - \sum_{k=1}^{K} a_k h(m-k) = 0,\ m \in [1, M-1] \setminus \Theta,\\
&\sum_{k=1}^{K} a_k h(m-k) = 0, \quad m \in [M, M+K-1] \setminus \Theta,\\
&h(m) - \sum_{k=1}^{K} a_k h(m-k) - \tilde{\varphi}(m) = 0,\\
&\qquad\qquad m \in [1, M-1] \cap \Theta,\\
&\sum_{k=1}^{K} a_k h(m-k) + \tilde{\varphi}(m) = 0,\\
&\qquad\qquad m \in [M, M+K-2] \cap \Theta,\\
&a_K h(M-1) + \tilde{\varphi}(M+K-1) = 0.
\end{aligned} \qquad (6)$$

Решение этой СЛАУ, возможно пополненной некоторыми из следующих уравнений

$$\begin{aligned}
&\tilde{\varphi}(m) = 0,\ \ m \in \Theta \setminus \{0, M+K-1\},\\
&\tilde{\varphi}(M+K-1) = 1,
\end{aligned} \qquad (7)$$

лежит в основе процедуры построения НМС-последовательности. Наряду со SLAU в тексте алгоритма приняты следующие обозначения:

- A (SLAU) — главная матрица для SLAU;
- A_e (SLAU) — расширенная матрица для SLAU;
- INIT () — функция, возвращающая сформированную СЛАУ (6);
- SOLVE (SLAU) — некоторый метод решения Крамеровской СЛАУ (например, метод Гаусса);

Алгоритм 2. Процедура построения НМС-последовательности

1: SLAU := INIT (); *// инициализация СЛАУ*
2: **если** Rank $(A\,(\text{SLAU}))$ = Rank $(A_e\,(\text{SLAU}))$ = $= K + M$ **то**
3: $\{h\,(m)\}_{m=0}^{M-1}, \{\tilde{\varphi}\,(m)\}_{m\in\Theta}$:= SOLVE (SLAU);
4: IsSolution := $h\,(M-1) \neq 0$;
5: **если** IsSolution **то**
6: $\{h^*\,(m)\}_{m=0}^{M-1} := \{h\,(m)\}_{m=0}^{M-1}$;
7: $\{\tilde{\varphi}^*\,(m)\}_{m\in\Theta} := \{\tilde{\varphi}\,(m)\}_{m\in\Theta}$;
8: **выход**
9: **если** Rank $(A\,(\text{SLAU}))$ < Rank $(A_e\,(\text{SLAU}))$ **то**
10: IsSolution := **false**;
11: **выход**;
// делаем пополнения СЛАУ и пытаемся решать пополненную
12: IsSolution := **false**; BestRate := 2^{M+K+1};
13: **для** $k = 1, \ldots, 2^{K-1}$
// если k=0, то СЛАУ не пополняется
14: BinaryVector := ToBinaryVector (k, K);
15: SubSlau := SelectAddEquations $(K, M, \text{BinaryVector}, \Theta)$;
16: SLAU_TMP := UNION (SLAU, SubSlau);
17: **если** Rank $(A\,(\text{SLAU\_TMP})) =$ $=$ Rank $(A_e\,(\text{SLAU\_TMP})) = K + M$ **то**
18: $\{h\,(m)\}_{m=0}^{M-1}, \{\tilde{\varphi}\,(m)\}_{m\in\Theta} :=$ $:=$ SOLVE (SLAU);
19: **если** $h(M-1) \neq 0$ **то**
20: Rate := RATE$\big(K, M, \{h\,(m)\}_{m=0}^{M-1}, \{\tilde{\varphi}\,(m)\}_{m\in\Theta}\big)$;
21: **если** Rate < BestRate **то**
22: $\{h^*\,(m)\}_{m=0}^{M-1} := \{h\,(m)\}_{m=0}^{M-1}$;
23: $\{\tilde{\varphi}^*\,(m)\}_{m\in\Theta} := \{\tilde{\varphi}\,(m)\}_{m\in\Theta}$;
24: BestRate := Rate; IsSolution := **true**;
25: **если** IsSolution = **true** **то**
26: решение $\{h^*\,(m)\}_{m=0}^{M-1}$, $\{\tilde{\varphi}^*\,(m)\}_{m\in\Theta}$;
27: **иначе**
28: решение отсутствует.

— IsSolution — булева переменная, которая указывает на существование решения;
— ToBinaryVector (Number, BinaryVectorSize) — функция перевода целого положительного числа Number в его бинарное представление длины BinaryVectorSize(Number < $2^{\text{BinaryVectorSize}}$). Результатом вызова этой функции является бинарный вектор (BinaryVector) указанной длины;
— SelectAddEquations $(K, M, \text{BinaryVector}, \Theta)$ — функция, которая возвращает сформированную СЛАУ. Возвращаемая СЛАУ оказывается составлена из тех уравнений из набора (7), номера которых соответствуют положениям единиц в бинарном векторе BinaryVector длины K;
— UNION (Slau1, Slau2) — функция объединения двух СЛАУ Slau1 и Slau2 в одну СЛАУ, возвращает объединенную СЛАУ;
— RATE$\big(K, M, \{h\,(m)\}_{m=0}^{M-1}, \{\tilde{\varphi}\,(m)\}_{m\in\Theta}\big)$ — функция расчета значения минимизируемого в критерии (4) функционала;
— BestRate — текущее минимальное значение функционала из критерия (4).

В докладе приводятся примеры построенных НМС-последовательностей, выступающих в качестве КИХ для эффективных ЛЛП.

Выводы

Представлены формализованная постановка и решение задачи построения эффективных ЛЛП цифровых сигналов. Эффективный ЛЛП обнаруживает оптимальное поведение в том смысле, что алгоритм его вычисления обладают наименьшей вычислительной сложностью, а КИХ признака оказывается наилучшим образом согласована с заданным критерием качества. Показано, что построение эффективного ЛЛП приводит к задаче построения неоднородных ЛРП специального класса, названных НМС-последовательностями. Представлен явный вид вычислительной процедуры построения таких НМС-последовательностей. Приводятся примеры последовательностей.

Литература

[1] Методы компьютерной обработки изображений // Под редакцией В. А. Сойфера. 2-е изд., испр. — М: Физматлит, 2003. — 784 с.

[2] *Мясников В. В.* О синтезе эффективного алгоритма над множеством алгоритмов вычисления свертки // Компьютерная оптика. — 2006. — Вып. 29. — С. 78–117.

[3] *Мясников В. В.* Эффективные локальные линейные признаки цифровых сигналов и изображений // Компьютерная оптика. — 2007. — Т. 31, № 4. — С. 58–76.

[4] *Мясников В. В.* Эффективный алгоритм над множеством алгоритмов линейной локальной фильтрации // Сб. докл. всеросс. конф. ММРО-14, 2009. — М.: МАКС Пресс — С. 264–267 (в этом сборнике).

[5] *Myasnikov V. V.* Efficient Algorithm under the set of convolution algorithms // Proceedings of the 8-th International Conference on Pattern Recognition and Image Analysis, Yoshkar-Ola, Russia, 2007. — Vol. 2. — С. 128–132.

Псевдоградиентный алгоритм построения эффективных линейных локальных признаков*

Титова О. А., Мясников В. В.
olti@smr.ru, vmyas@smr.ru
Самара, Институт систем обработки изображений РАН

В работе рассматриваются постановка и численное решение расширенной частной задачи построения эффективных линейных локальных признаков цифровых сигналов. Численное решение строится на базе известного псевдоградиентного алгоритма. Получаемое решение сравнивается с решением частной задачи построения эффективных линейных локальных признаков по ряду критериев, характеризующих качество конструируемых признаков.

При решении прикладных задач обработки и анализа цифровых сигналов и изображений часто используются операции локальной обработки. Результатом локальной обработки сигнала или изображения, как правило, является сигнал или изображение, совпадающее по размеру с обрабатываемым. Каждый отсчет такого «выходного» изображения является результатом преобразования отсчетов исходного — «входного» — изображения, попавших в некоторую локальную пространственную окрестность обрабатываемого отсчета. Такую окрестность обычно называют «областью обработки» или «окном обработки» [1]. В процессе локальной обработки окно обработки занимает все возможные положения на обрабатываемом изображении, и для каждого его положения вычисляется соответствующее преобразование.

Учитывая, что вид преобразования может быть достаточно сложным, процесс локальной обработки оказывается вычислительно очень трудоемким: «прямая» реализация вычислений даже на современных вычислительных машинах может занимать несколько часов или дней. Одним из способов радикально снизить сложность (вычислительную и временную) обработки является разбиение преобразования на два этапа [2]. На первом этапе на основании входного изображения формируется его описание — ряд изображений, в которых каждый отсчет некоторым образом характеризует отсчеты входного изображения, попавшие в окно обработки. На втором этапе построенное описание преобразуется в требуемый выход. Изображения, полученные на первом этапе, также являются результатом локальной обработки входного изображения, но для снижения общей сложности обработки их вычисление производится с использованием вычислительно простых алгоритмов. Достаточно часто такие алгоритмы строятся на базе быстрых алгоритмов дискретных ортогональных преобразований, рекурсивных алгоритмов, вейвлет-преобразований и др. [1]. Учитывая, что каждый отсчет изображения в построенном описании характеризует окрестность входного изображения, сами эти описания (по аналогии с терминологией, принятой в теории распознавания) называют *признаками*. Для случая, когда локальная обработка, в процессе которой формируются признаки, является линейной с постоянными параметрам (то есть может быть представлена как линейная свертка входного сигнала или изображения с финитным ядром, называемым конечной импульсной характеристикой — КИХ), получаемые признаки назовем *линейными локальные признаками* исходного изображения. Каждый линейный локальный признак (ЛЛП) однозначно характеризуется конечной импульсной характеристикой, для которой рассчитывается свертка, и алгоритмом, который реализует эти вычисления.

Исходя из назначения ЛЛП, «хорошим» может считаться признак, если он удовлетворяет двум требованиям: алгоритм вычисления ЛЛП имеет низкую вычислительную сложность и КИХ признака «хорошо» согласована с решаемой прикладной задачей. Учитывая противоречивость этих требований, процесс *построения признаков* (то есть определения КИХ признака и алгоритма его вычисления), возникающий на этапе проектирования и построения прикладной системы обработки и анализа изображений, часто сводится к простому выбору одного из существующих или доступных быстрых алгоритмов. В работе [3] был предложен совершенно иной подход к построению вычислительно эффективных ЛЛП.

А именно, в работе [3] была произведена формализация задачи построения эффективного ЛЛП. *Общая задача* построения отдельного эффективного ЛЛП была сформулирована как задача построения последовательности КИХ признака, порождающей алгоритм вычисления признака с наименьшей сложностью, которая наилучшим образом согласована с заданным производящим функционалом, численно определяющим «показатель качества» для КИХ. Далее было показано, что более корректной (в смысле Адамара) является *частная задача* построения эффективного ЛЛП.

*Работа выполнена при поддержке РФФИ (проекты №06-01-00616-а, №09-01-00434-а) и Программы фундаментальных исследований Президиума РАН «Фундаментальные проблемы информатики и информационных технологий», проект 2.12.

В её постановке множество конечных последовательностей, среди которых ищется решение, ограничивается семейством НМС-последовательностей (НМС — нормализованная с минимальной сложностью [3]). Учитывая конечность мощности конкретного семейства НМС-последовательностей [3], частная задача построения эффективного ЛЛП может быть решена перебором за конечное время с использованием алгоритма, приведенного в [4].

Определённым недостатком эффективных ЛЛП, полученных как решение частной задачи, и соответствующего способа их построения является то, что ограничение класса последовательностей отдельным семейством является слишком «жёстким». Для практического использования более естественной является постановка, в которой фиксируются только ключевые параметры семейства, определяющие сложность алгоритма вычисления признака, а остальные параметры остаются неизвестными. Такая задача определяется в настоящей работе как *расширенная частная задача* построения эффективных ЛЛП.

Целью настоящей работы является разработка и исследование численного решения расширенной частной задачи построения эффективных ЛЛП, построенного на базе псевдоградиентного алгоритма. *Результатом работы* должны стать выводы о качественном изменении конструируемых ЛЛП, возникающем при переходе от решения частной к расширенной частной задаче построения эффективных ЛЛП. Качественные изменения определяются по ряду показателей, характеризующих набор построенных ЛЛП.

Эффективные линейные локальные признаки и задачи их построения

Пусть $\mathbb{R}$ и $\mathbb{N}$ — множества вещественных и натуральных чисел, $\mathfrak{K}$ — коммутативное кольцо с единицей. Рассмотрим алгоритм линейной локальной фильтрации (ЛЛФ) следующего вида [3].

Алгоритм 1. Линейная локальная фильтрация.

Вход: $\{x(n)\}_{n=0}^{N-1}$;

Выход: $\{y(n)\}_{n=M-1}^{N-1}$;

1: предварительная обработка:
$\tilde{y}(n) = \sum\limits_{m\in\Theta} x(n-m)\tilde{\varphi}(m),\ n = 0,\ldots,N-1;$

2: окончательная обработка:
$y(n) = \sum\limits_{k=1}^{M} a_k y(n-k) + \tilde{y}(n),\ n = 0,\ldots,N-1;$

В представленном алгоритме:

— значения отсчетов $x\,(n)$ и $y\,(n)$ для случая $n < 0$ полагаются равными нулю;

— величины $K \in \mathbb{N}$, $\{a_k\}_{k=1}^{K}$ $(a_k \in \mathfrak{K},\ a_K \neq 0)$ и $\{\tilde{\varphi}(m)\}_{m=0}^{M+K-1}$ $(\tilde{\varphi}(m) \in \mathfrak{K})$ определяют неоднородное *линейное рекуррентное соотношение* (ЛРС) вида

$$h(m) = \sum_{k=1}^{K} a_k h\,(m-k) + \tilde{\varphi}(m),\ m = 0,1,\ldots,$$

которому удовлетворяет последовательность отсчетов КИХ $\{h(m)\}_{m=0}^{M-1}$. В этом ЛРС значения $h\,(m)$ для случая $m < 0$ полагаются равными нулю, а для ситуации $m \geqslant M$ должны удовлетворять условию $h(m) = 0$. Будем использовать терминологию, принятую в [5]: K — *порядок ЛРС*, $\bar{a} = (a_1, a_2, \ldots, a_K)$ — *вектор коэффициентов ЛРС*, а формируемая с использованием ЛРС последовательность — *линейная рекуррентная последовательность* (ЛРП).

— величина Θ определяется как множество отсчетов, в которых однородность ЛРС нарушается: $\Theta = \big\{m \in \{0,\ldots,M{+}K{-}1\}\colon \tilde{\varphi}(m) \neq 0\big\}$.

Вычислительная сложность данного алгоритма ЛЛФ имеет вид:

$$u\,(A)\ \tfrac{N-M+1}{N} = |\Theta| + K - \xi_{\text{add}}, \qquad (1)$$

где $|\Theta|$ — мощность множества Θ. В работе [3] показано, что для заданного порядка K ЛРС минимум вычислительной сложности (1) приведенного алгоритма достигается для подкласса последовательностей, вводимых определением 2.

Определение 1. *ЛРП $h(0), h(1), \ldots$ порядка K над $\mathfrak{K}$ называется МС-последовательностью порядка K длины M над $\mathfrak{K}$, если выполняется:*

$$\big(h(0) \neq 0\big)\ \wedge\ \big(h(M-1) \neq 0\big)\ \wedge \\ \wedge\ \big(\forall m \geqslant M\ h(m) = 0\big)\ \wedge\ \big(|\Theta| \leqslant K+1\big).$$

Определение 2. *МС-последовательность порядка K длины M над $\mathfrak{K}$ называется нормализованной МС-последовательностью (НМС-последовательностью) для $\tilde{\Theta}$ порядка K длины M, если $h(0) = 1$ и выполняется условие*

$$\sum_{m\in\tilde{\Theta}} 2^m I\big[\tilde{\varphi}(m) \neq 0\big] - 2^{M+K} \sum_{m\in\tilde{\Theta}} I\big[\tilde{\varphi}(m) = 0\big] - \\ - \tfrac{1}{2} I\big[\tilde{\varphi}(M+K-1) = 1\big] \to \min_{\substack{\{h(m)\}_{m=0}^{M-1} \\ \{\tilde{\varphi}(m)\}_{m\in\tilde{\Theta}}}}. \qquad (2)$$

где $I\big[\ldots\big]$ — индикатор аргумента-выражения, принимающий значение 1, если выражения верно, и 0 — иначе.

Поскольку для НМС-последовательности выполняются соотношение $|\Theta| \leqslant K+1$ и $h(0) = \tilde{\varphi}(0) = 1$, вычислительная сложность (1) Алгоритма 1 ЛЛФ для КИХ в виде НМС-последовательности порядка K длины M удовлетворяет ограничению:

$$u\,(A)\,\frac{N-M+1}{N} \leqslant 2K. \qquad (3)$$

Множество всех НМС-последовательностей можно разбить на подклассы (семейства).

Определение 3. *$(K, M, \bar{a})$-семейством НМС-последовательностей, обозначаемым $\wp(K, M, \bar{a})$, называется множество НМС-последовательностей порядка K длины M, удовлетворяющих ЛРС с коэффициентами $\bar{a}$ $(a_K \neq 0)$.*

Задачи построения эффективных ЛЛП могут быть определены следующим образом.

Частная задача построения эффективного ЛЛП определяется в работе [3] как задача построения НМС-последовательности $h(0), \ldots, h(M-1)$ конкретного семейства $\wp(K, M, \bar{a})$ с минимальным значением заданного производящего функционала $\Psi : \mathfrak{K}^M \to \mathbb{R}$:

$$\Psi\big(h(0), \ldots, h(M-1)\big) \to \min_{\{h(m)\}_{m=0}^{M-1} \in \wp(K,M,\bar{a})}.$$

Алгоритм 1 вычисления ЛЛП для построенной НМС-последовательности имеет указанный выше вид и соответствующую сложность (1), ограниченную величиной (3). Свойства частной задачи построения эффективного ЛЛП указаны в работе [3], алгоритм ее решения указан в работе [4].

Определение 4. *Расширенной частной задачей построения эффективного ЛЛП называется задача: для заданных параметров $N, M, K \in \mathbb{N}$ таких, что $N > M$, $K < \frac{N-M+1}{2N}(M - \xi_{\text{add}})$, и заданного производящего функционала $\Psi\colon \mathfrak{K}^M \to \mathbb{R}$ построить ЛЛП $\{h(m)\}_{m=0}^{M-1}$, в котором:*

- *последовательность $h(0), \ldots, h(M-1)$ является НМС-последовательностью с минимальным значением производящего функционала:*

$$\Psi\big(h(0), \ldots, h(M-1)\big) \to \min_{\{h(m)\}_{m=0}^{M-1} \in \bigcup_{\bar{a}} \wp(K,M,\bar{a})};$$

- *параметры Алгоритма 1 определяются по найденной НМС-последовательности.*

Псевдоградиентный алгоритм построения эффективных ЛЛП

Решение расширенной частной задачи построения эффективных ЛЛП простым перебором в случае вещественнозначных компонент вектора $\bar{a}$ не представляется возможным. Учитывая это, предлагается следующий способ нахождения квазиоптимального решения задачи построения эффективного ЛЛП.

1. Определение параметра семейства (вектора $\bar{a}$) выполняется в рамках поисковой процедуры, минимизирующей значение производящего функционала (реализуется псевдоградиентным алгоритмом, описанным ниже).
2. Для каждого значения вектора $\bar{a}$, получаемого на 1-м этапе, производится построение НМС-последовательности в результате решения частной задачи построения эффективного ЛЛП (то есть перебором в семействе $\wp(K, M, \bar{a})$).

В случае построения нескольких (набора) эффективных ЛЛП их конструирование выполняется последовательно, то есть путем последовательного присоединения одного эффективного ЛЛП к набору уже существующих.

Псевдоградиентный алгоритм.

Основная идея псевдоградиентного алгоритма, известного также как метод деформируемого многогранника (МДМ) [6], заключается в последовательном перемещении и деформировании многогранника вокруг точки экстремума. Основными операторами МДМ являются сортировка, отражение, растяжение и сжатие. В качестве точки (вершины многогранника) в данном случае выступает вектор коэффициентов ЛРС $\bar{a}$. Алгоритм кратко может быть описан следующим образом.

Вначале выбираются точки-вершины многогранника и считаются значения производящего функционала в каждой из них.

Сортировка: сортируются вершины многогранника в порядке возрастания значений производящего функционала в этих точках.

Отражение: точка с наибольшим в ней значением производящего функционала «отражается» относительно центра тяжести многогранника. Координаты новой точки вычисляются по формуле:

$$\bar{a}_r = (1+\beta)\bar{a}_c - \beta\bar{a}_{\max},$$

где $\bar{a}_c$ — центр тяжести многогранника, $\bar{a}_{\max}$ — вершина с наибольшим в ней значением функционала, а β — коэффициент отражения. Если значение функционала в вычисленной точке меньше значения в точке $\bar{a}_{\max}$, то последняя точка заменяется на новую.

Растяжение: если значение функционала в новой точке $\bar{a}_r$ меньше значения в точке с наименьшим значением функционала, то делается попытка увеличить шаг и проверить точку с координатами:

$$\bar{a}_s = (1-\gamma)\bar{a}_c + \gamma\bar{a}_r,$$

где γ — коэффициент растяжения. Если значение функционала в точке $\bar{a}_s$ меньше значения в наилучшей точке, то наихудшая точка заменяется на $\bar{a}_s$.

Сжатие: строится точка $\bar{a}_p = \eta\bar{a}_{\max} + (1-\eta)\bar{a}_c$, где η — коэффициент сжатия, и в ней вычисляется значение функционала. Если полученное значение функционала меньше значения в точке $\bar{a}_{\max}$, то последняя заменяется на $\bar{a}_p$, иначе выполняется *глобальное сжатие* многогранника к точке с наименьшим значением — каждая точка перемещается на половину расстояния до наилучшей вершины.

Последовательное применение приведенных операторов МДМ позволяет решить задачу поиска при слабых требованиях к целевой функции, в качестве которой в настоящей работе выступает производящий функционал.

Экспериментальные исследования

Цель проводимых экспериментальных исследований — это определение качественных изменений в конструируемых ЛЛП, возникающих при переходе от решения частной задачи к расширенной частной задаче построения эффективных ЛЛП. Для определения этих изменений производилось построение $T = 4$ НМС-последовательностей $\{h_t\}_{t=0}^{T-1}$ с параметрами: $M = 21$, $K = 1, 2, 3, 4$. Эти последовательности конструировались как решения, соответственно, частной и расширенной частной задач построения эффективных ЛЛП. В качестве производящих функционалов выступали следующие показатели качества конструируемых наборов последовательностей, введенные в монографии [1]:

— диапазон, задаваемый величиной

$$J_1 = \max_{t=0,\ldots,T-1} \sum_{m=0}^{M-1} \big|h_t(m)\big|;$$

— число обусловленности матрицы взаимной корреляции последовательностей, задаваемое величиной (для квадратичной нормы матрицы):

$$J_2 = \sqrt{|\lambda_{\max}| \cdot |\lambda_{\min}|^{-1}},$$

где $\lambda_{\max}$ и $\lambda_{\min}$, соответственно, максимальное и минимальное по модулю собственные числа указанной матрицы;

— коэффициент сопряженности НМС-последовательностей, вычисляемый по норме Гильберта-Шмидта [1] в виде:

$$J_3 = \frac{2}{T(T-1)} \sum_{t=0}^{T-2} \sum_{q=t+1}^{T-1} \frac{\langle h_t, h_q \rangle}{\sqrt{\|h_t\| \, \|h_q\|}}.$$

Решение частной задачи производилось для семейства НМС-последовательностей, вектор $\bar{a}$ которого содержал биномиальные коэффициенты. Выбор такого вектора приводит к построению последовательностей, составленных из фрагментов многочленов. Подобные последовательности уже появлялись в различных работах в качестве удобных КИХ для алгоритмов вычисления ЛЛП с предельно низкой вычислительной сложностью [1].

Результаты экспериментальных исследований представлены в таблице 1. На основании представленных в ней результатов можно сделать следующие основные выводы:

— качество формируемых ЛЛП в результате решения расширенной частной задачи всегда выше, чем качество ЛЛП, получаемых в результате решения частной задачи; выигрыш в качестве в некоторых случаях оказывается более чем на порядок (пример — сопряженность) при одинаковой вычислительной сложности ($2KT$) вычисления T эффективных ЛЛП;

Таблица 1. Показатели качества эффективных ЛЛП для различных задач их построения.

	K	$2KT$	частная	расширенная
J_1	1	8	1,00	0,42
	2	16	0,96	0,22
	3	24	0,64	0,15
	4	32	0,46	0,12
J_2	1	8	15,00	9,34
	2	16	14,62	5,88
	3	24	12,01	5,03
	4	32	7,12	5,01
J_3	1	8	2,00	0,36
	2	16	0,91	0,005
	3	24	0,19	0,002
	4	32	0,03	0,001

— высокие качественные показатели ЛЛП достигаются уже при малом значении величины K и, как следствие, при предельно малой вычислительной сложности алгоритмов расчета ЛЛП.

В докладе представлены результаты дополнительных экспериментальных исследований псевдоградиентного алгоритма и решения расширенной частной задачи построения эффективных ЛЛП.

Выводы

В работе предложен и исследован псевдоградиентный алгоритм решения расширенной частной задачи построения эффективных линейных локальных признаков цифровых сигналов. Показано, что полученное решение в виде набора признаков по всем рассмотренным в работе показателям качества существенно (для некоторых показателей — более чем на порядок) превосходит решение частной задачи построения, решение которой производится известным алгоритмом.

Литература

[1] Методы компьютерной обработки изображений // Под редакцией В. А. Сойфера. 2-е изд., испр. — М: Физматлит, 2003. — 784 с.

[2] *Chernov A. V., Myasnikov V. V., Sergeyev V. V.* Fast Method for Local Image Processing and Analysis // Pattern Recognition and Image Analysis. — 1999. — Vol. 9, № 4. — P. 572–577.

[3] *Мясников В. В.* Эффективные локальные линейные признаки цифровых сигналов и изображений // Компьютерная оптика. — 2007. — Вып. 31, № 4. — С. 58–76.

[4] *Мясников В. В.* О постановке и решении задачи построения эффективных линейных локальных признаков цифровых сигналов // Всеросс. конф. ММРО-14. — М.: МАКС Пресс, 2009 — С. 268–271.

[5] *Лидл Р., Нидеррайтер Г.* Конечные поля: в 2-х т., Т. 1. — М.: Мир, 1988. — 430 с.

[6] *Гилл Ф., Мюррей У., Райт М.* Практическая оптимизация. — М.: Мир, 1985. — 509 с.

Распознавание алфавита векторов, порождающего последовательности с квазипериодической структурой*

Хамидуллин С. А.
kham@math.nsc.ru
Новосибирск, Институт математики им. С. Л. Соболева СО РАН

Рассматривается проблема помехоустойчивого апостериорного (off-line) распознавания алфавита векторов, порождающего последовательности, включающие квазипериодически перемежающиеся вектор-фрагменты, совпадающие с элементами из этого алфавита. Исследуются дискретные экстремальные задачи, к которым сводятся варианты этой проблемы. Обоснованы точные полиномиальные алгоритмы решения редуцированных задач, гарантирующие максимально правдоподобное принятие решения в случае, когда помеха аддитивна и является гауссовской последовательностью независимых одинаково распределенных случайных величин, а число перемежающихся вектор-фрагментов неизвестно.

Введение

Объект исследования работы — проблемы оптимизации в задачах анализа данных и распознавания образов. Предмет исследования — дискретные экстремальные задачи, к которым сводятся некоторые варианты проблемы помехоустойчивого апостериорного (off-line) распознавания алфавита векторов, порождающего последовательности, включающие квазипериодически перемежающиеся информационные фрагменты, которые совпадают с элементами из этого алфавита. Цель работы — обоснование алгоритмов решения этих задач.

Рассмотрим следующую содержательную задачу. Допустим, что имеется совокупность физических объектов, каждый из которых может находиться в пассивном и конечном множестве отличающихся активных состояний. Элементам этого множества однозначно соответствуют элементы известного алфавита импульсов одинаковой длительности, но различной формы. Пассивному состоянию соответствует отсутствие каких-либо импульсов. Алфавиты импульсов не пересекаются. Источник сообщений передает информацию о пассивных и активных состояниях некоторого физического объекта с помощью импульсов из алфавита. На приемную сторону поступает дискретный сигнал в виде зашумленной последовательности квазипериодически перемежающихся импульсов. Термин «квазипериодически» означает, что интервал между двумя последовательными импульсами не одинаков, а лишь ограничен сверху и снизу некоторыми константами. Время поступления импульсов в принятой последовательности неизвестно. Требуется определить, какому именно объекту соответствует принятая последовательность. Иными словами, задача состоит в распознавании алфавита или источника передаваемых импульсов. Ситуации, в которых возникает эта задача, характерны, в частности, для электронной разведки, геофизики, гидроакустики, телекоммуникации, распознавания речевых сигналов и других приложений. В описанной задаче возможны два случая, когда число принятых импульсов в последовательности известно и неизвестно.

*Работа выполнена при финансовой поддержке РФФИ, проекты № 07-07-00022 и № 09-01-00032.

Постановка задачи

Пусть последовательность $x_n \in \mathbb{R}^q$, $n \in \mathcal{N}$, где $\mathcal{N} = \{1, \ldots, N\}$, векторов евклидова пространства представима в виде

$$x_n = \begin{cases} u_n, & n \in \mathcal{M}, \\ 0, & n \in \mathcal{N} \setminus \mathcal{M}, \end{cases} \qquad (1)$$

где $\mathcal{M} \subseteq \mathcal{N}$.

Пусть $|\mathcal{M}| = M$. Положим $\mathcal{M} = \{n_1, \ldots, n_M\}$, и допустим, что элементы набора $(n_1, \ldots, n_M)$, соответствующие номерам ненулевых векторов в последовательности (1), удовлетворяют ограничениям

$$1 \leqslant T_{\min} \leqslant n_m - n_{m-1} \leqslant T_{\max} \leqslant N - 1, \quad m = 2, \ldots, M, \qquad (2)$$

где $T_{\min}$ и $T_{\max}$ — натуральные числа.

Допустим, что $u_n \in A$, $n \in \mathcal{M}$, где $A \subset \{u\colon u \in \mathbb{R}^q,\ 0 < \|u\|^2 < \infty\}$, причем $|A| < \infty$. Множество A будем называть *алфавитом информационных векторов*. Пусть, кроме того, $A \in \mathcal{A}$, где $\mathcal{A} = \cup_{l=1}^{L} A_l$ — совокупность алфавитов, причем $A_k \cap A_j = \varnothing$, если $k \neq j$; $|A_l| = K_l$, $l = 1, \ldots, L$, и $\sum_{l=1}^{L} K_l = K$. Будем считать, что последовательность вида (1)–(2) порождается информационными векторами из некоторого алфавита $A \in \mathcal{A}$.

Рассмотрим аддитивную модель помех (или ошибок наблюдения). Доступной для обработки (наблюдения) будем считать последовательность

$$y_n = x_n + e_n,\ n \in \mathcal{N}, \qquad (3)$$

где e_n — вектор помехи. Предположим, что векторы x_n и e_n независимы. Задача распознавания состоит в том, чтобы по наблюдаемомой последовательности y_n, $n \in \mathcal{N}$, найти алфавит $A \in \mathcal{A}$ векторов, порождающий последовательность вида (1)–(2).

В приведенной модели информационный вектор соответствует импульсу из содержательной задачи, сформулированной во введении. Положим $\eta = (n_1, \ldots, n_M)$, $w = \{u_n, n \in \mathcal{M}\}$. Тогда легко видеть, что последовательность x_n, $n \in \mathcal{N}$, зависит от наборов η, w и алфавита A, т. е. $x_n = x_n(\eta, w, A)$.

Пусть

$$S(\eta, w, A) = \sum_{n \in \mathcal{N}} \|y_n - x_n\|^2. \qquad (4)$$

Рассмотрим следующие задачи.

Задача 1. *Дано:* последовательность $y_n \in \mathbb{R}^q$, $n \in \mathcal{N}$, описываемая формулами (1)–(3), натуральное число M, совокупность $\mathcal{A} = \{A_1, \ldots, A_L\}$ конечных непересекающихся алфавитов ненулевых векторов из $\mathbb{R}^q$.

Найти: алфавит $A \in \mathcal{A}$ и наборы η и w такие, что целевая функция (4) минимальна.

Задача 2. *Дано:* последовательность $y_n \in \mathbb{R}^q$, $n \in \mathcal{N}$, описываемая формулами (1)–(3), совокупность $\mathcal{A} = \{A_1, \ldots, A_L\}$ конечных непересекающихся алфавитов ненулевых векторов из $\mathbb{R}^q$.

Найти: алфавит $A \in \mathcal{A}$, наборы η и w и их размерность M такие, что целевая функция (4) минимальна.

К сформулированным задачам приводит статистическая формулировка проблемы распознавания алфавита информационно значимых векторов, порождающего последовательности, структура которых описывается формулами (1)–(3), если считать, что $\{e_n\}$ в формуле (3) есть выборка из q-мерного нормального распределения с параметрами $(0, \sigma^2 I)$, где I — единичная матрица, а в качестве критерия решения задачи использовать максимум функционала правдоподобия. Отличие этих задач состоит в том, что в первой из них число ненулевых информационно значимых векторов в последовательности считается заданным, а во второй — неизвестным, т. е. является оптимизируемой величиной.

Следующие две задачи можно трактовать как специальные случаи двух первых задач, в которых совокупность $\mathcal{A}$ состоит из единственного алфавита. Эти задачи отражают сущность проблемы совместного помехоустойчивого обнаружения и идентификации векторов при условии, что алфавит A фиксирован.

Задача 3. *Дано:* последовательность $y_n \in \mathbb{R}^q$, $n \in \mathcal{N}$, описываемая формулами (1)–(3), натуральное число M и алфавит A ненулевых векторов из $\mathbb{R}^q$.

Найти: наборы η и w такие, что целевая функция (4) минимальна.

Задача 4. *Дано:* последовательность $y_n \in \mathbb{R}^q$, $n \in \mathcal{N}$, описываемая формулами (1)–(3) и алфавит A ненулевых векторов из $\mathbb{R}^q$.

Найти: наборы η и w и их размерность M такие, что целевая функция (4) минимальна.

Редуцированные задачи

Сумму (4) можно переписать в виде

$$S = \sum_{n \in \mathcal{N}} \|y_n\|^2 - \sum_{n \in \mathcal{M}} (2\langle y_n, u_n \rangle - \|u_n\|^2)$$

где $\langle \cdot, \cdot \rangle$ — скалярное произведение. Первый член в правой части этого выражения — константа. Следовательно, задачи 1–4 минимизации суммы S сводятся к максимизации второго члена этого выражения. Имеем следующие экстремальные задачи, к которым сводятся задачи 1 и 2.

Задача RVAGS-F (Recognition of Vector Alphabet that Generates a Sequence, when M is Fixed) — распознавание алфавита векторов при заданном числе информационных векторов.

Дано: последовательность векторов $y_n \in \mathbb{R}^q$, $n \in \mathcal{N}$, совокупность $\mathcal{A} = \{A_1, \ldots, A_L\}$ конечных непересекающихся алфавитов информационных векторов из $\mathbb{R}^q$ и натуральное число M.

Найти: алфавит $A \in \mathcal{A}$ и наборы η и w такие, что

$$\sum_{m=1}^{M} (2\langle y_{n_m}, u_{n_m} \rangle - \|u_{n_m}\|^2) \to \max, \qquad (5)$$

при ограничениях (2).

Задача RVAGS-NF (Recognition of Vector Alphabet that Generates a Sequence, when M is Not Fixed) — распознавание алфавита векторов при неизвестном числе информационных векторов.

Дано: последовательность векторов $y_n \in \mathbb{R}^q$, $n \in \mathcal{N}$, совокупность $\mathcal{A} = \{A_1, \ldots, A_L\}$ конечных непересекающихся алфавитов информационных векторов из $\mathbb{R}^q$.

Найти: алфавит $A \in \mathcal{A}$, наборы η и w и их размерность M такие, что выполняется (5) при ограничениях (2).

Точные полиномиальные алгоритмы решения редуцированных задач RVAGS-F и RVAGS-NF обоснованы в [1, 2]. Трудоемкости этих алгоритмов есть величины $O\big(M(T_{\max} - T_{\min} + Kq)(N - q + 1)\big)$ и $O\big((T_{\max} - T_{\min} + Kq)(N - q + 1)\big)$ соответственно.

Задачи 3 и 4 сводятся к решению следующих экстремальных задач.

Задача JDIVS-F (Joint Detection and Identification of Vectors in a Sequence, when M is Fixed) — совместное обнаружение и идентификация заданного числа информационных векторов.

Дано: последовательность векторов $y_n \in \mathbb{R}^q$, $n \in \mathcal{N}$, алфавит A информационных векторов из $\mathbb{R}^q$ и натуральное число M.

Найти: наборы η и w такие, что выполняется (5) при ограничениях (2).

Задача JDIVS-NF (Joint Detection and Identification of Vectors in a Sequence, when M is Not Fixed) — совместное обнаружение и идентификация неизвестного числа информационных векторов.

Дано: последовательность векторов $y_n \in \mathbb{R}^q$, $n \in \mathcal{N}$ и алфавит A информационных векторов из $\mathbb{R}^q$.

Найти: наборы η и w и их размерность M такие, что выполняется (5) при ограничениях (2).

Точные полиномиальные алгоритмы решения редуцированных задач JDIVS-F и JDIVS-NF обоснованы в [3, 4]. Трудоемкости этих алгоритмов есть величины $O\big(M(T_{\max} - T_{\min} + q)(N - q + 1)\big)$ и $O\big((T_{\max} - T_{\min} + q)(N - q + 1)\big)$ соответственно.

Алгоритмы решения приведенных редуцированных задач лежат в основе алгоритмов помехоустойчивого анализа и распознавания структурированных последовательностей, включающих вектор-фрагменты из заданного алфавита. Эти алгоритмы гарантируют оптимальность решения по критерию максимального правдоподобия в случае, когда помеха аддитивна и является гауссовской последовательностью независимых одинаково распределенных величин.

Численное моделирование

Сущность рассмотренных задач для одномерных последовательностей и работу алгоритмов иллюстрирует результат приведенного ниже численного эксперимента по распознаванию алфавита векторов при неизвестном числе информационных фрагментов в последовательности (задача RVAGS-NF).

На рис. 1 приведены 2 алфавита. Первый алфавит состоит из пяти информационных векторов, второй — из четырех. Компоненты векторов изображены в виде графиков.

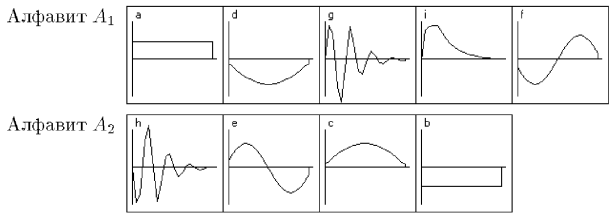


Рис. 1. Распознаваемые алфавиты.

На рис. 2 а изображена сгенерированная последовательность x_n, $n \in \mathcal{N}$, включающая 10 фрагментов, совпадающих с элементами из алфавита A_2. На рис. 2 б представлена последовательность y_n, $n \in \mathcal{N}$, подлежащая обработке (в этом примере уровень помехи соизмерим с уровнем сигнала). На рис. 2 в приведена последовательность $\hat{x}_n$, $n \in \mathcal{N}$, восстановленная с помощью предложенного алгоритма. Прямоугольными рамками очерчены места расположения обнаруженных фрагментов, найденные алгоритмом в зашумленной последовательности. Числовые данные под графиками соответствуют заданным (рис. 2 а) и найденным (рис. 2 б и 2 в) начальным номерам фрагментов. Буквы справа от них — имена информационных векторов из алфавита, порождающего последовательность (рис. 2 а) и имена информационных векторов из распознанного алфавита (рис. 2 б и 2 в). Результат распознавания — алфавит A_2.

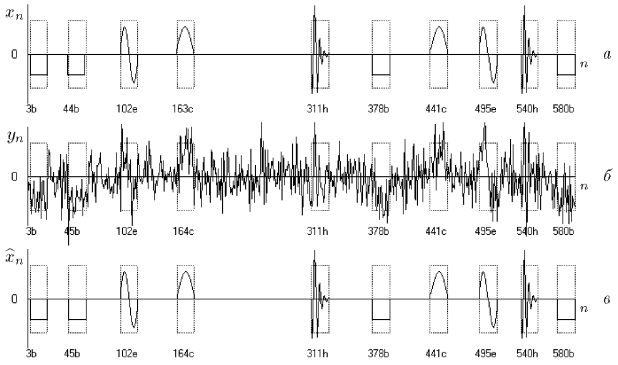


Рис. 2. Исходная ($а$), наблюдаемая ($б$) и восстановленная ($в$) последовательности.

Рисунок иллюстрирует практически безупречную работу алгоритма в условиях, когда уровень сигнала соизмерим с уровнем помехи.

Заключение

Рассмотренные задачи пополняют список (см. [5]) изученных задач дискретной оптимизации, к которым сводится помехоустойчивая off-line обработка (анализ и распознавание) числовых последовательностей, включающих какие-либо структуры над информационно-значимыми вектор-фрагментами одинаковой размерности. Открытыми остаются вопросы о разрешимости аналогов рассмотренных задач для случая, когда модель последовательности включает посторонние (мешающие) произвольные, но ненулевые фрагменты-вставки. Важной, но пока неизученной является задача распознавания алфавита информационных векторов, в которой алфавиты перекрываются. Алгоритмы решения этих задач представляют значительный интерес для ряда упомянутых приложений.

Литература

[1] *Кельманов А. В., Хамидуллин С. А.* Об одном варианте задачи распознавания алфавита векторов, порождающего последовательности с квазипериоди-

ческой структурой // Сиб. журн. вычисл. математики / РАН. Сиб. отд-ние. — Новосибирск. — 2009. — Т. 12, № 3. — С. 275–287.

[2] *Кельманов А. В., Хамидуллин С. А.* Об одном варианте задачи распознавания алфавита векторов // Тез. докл. междунар. конф. «Алгоритмический анализ неустойчивых задач», посвященной 100-летию со дня рождения В. К. Иванова, 1–6 сентября 2008 г. — Екатеринбург: изд-во Уральского университета, 2008. — С. 282–283.

[3] *Кельманов А. В., Хамидуллин С. А.* Апостериорное совместное обнаружение и различение заданного числа подпоследовательностей в квазипериодической последовательности // Сибирский журнал индустриальной математики. — 1999. — Т. 2, № 2(4). — С. 106–119.

[4] *Кельманов А. В., Окольнишникова Л. В.* Апостериорное совместное обнаружение и различение подпоследовательностей в квазипериодической последовательности // Сибирский журнал индустриальной математики. — 2000. — Т. 3, № 2(6). — С. 115–139.

[5] `http://math.nsc.ru/~serge/qpsl/` — Система QPSLab для решения задач компьютерного анализа и распознавания числовых последовательностей с квазипериодической структурой — 2008.

Вопросы аппроксимируемости задачи обучения в классе комитетных решающих правил*

Хачай М. Ю.
mkhachay@imm.uran.ru
Екатеринбург, Институт математики и механики УрО РАН

В статье обсуждаются два новых результата, связанных с вычислительной и аппроксимационной сложностью задач комбинаторной оптимизации, возникающих при обучении распознаванию образов в классе комитетных кусочно-линейных решающих правил.

Задача комбинаторной оптимизации MASC — «Минимальный аффинный разделяющий комитет» возникает на этапе обучения распознаванию в классе коллективных решающих правил, в которых согласование мнений индивидуальных классификаторов производится путем голосования большинством голосов. Интерес к изучению сложности задачи мотивируется ее тесной связью с рядом известных NP-трудных задач, таких как задача обучения простейшего двуслойного персептрона [1] и задача кусочно-линейной отделимости конечных множеств [2], частным случаем которых она является, и традиционным подходом к анализу подклассов труднорешаемых задач.

Известно, что задача MASC в общем случае NP-трудна [3] и плохо аппроксимируема [4]. Кроме того, известно [5], что задача остается труднорешаемой, будучи сформулированной в пространстве произвольной фиксированной размерности $n > 1$. Однако известные доказательства всех перечисленных результатов базировались на рассмотрении специально сконструированных вырожденных случаев исследуемой задачи. Возникает естественный вопрос, сохранит ли задача MASC свойство труднорешаемости, если такие частные случаи явно исключить из рассмотрения. Для проведения такого исключения достаточно потребовать, чтобы конечное множество (в n-мерном пространстве), определяющее условие задачи, находилось *в общем положении*, при котором каждое его подмножество из $n+1$ элемента аффинно независимо.

До последнего времени, сформулированный выше вопрос оставался открытым. Кроме того, открытым оставался вопрос уточнения оценок аппроксимируемости задачи путем сокращения разрыва между известным [5] порогом аппроксимируемости $O(\log\log\log m)$ и точностью $O(m)$ наилучшего известного [6] полиномиального приближенного алгоритма (здесь m обозначает мощность разделяемого множества). Ответы на поставленные вопросы приведены в данной работе. Фактически, статья содержит следующие результаты:

1) показано, что задача MASC, сформулированная в пространстве произвольной фиксированной размерности $n > 1$, является NP- и Max-SNP-трудной[1] при дополнительном ограничении общности положения множества, определяющего условие задачи;
2) приведен новый полиномиальный приближенный алгоритм, обладающий при естественном допущении точностью $O(\log m)$.

Известные результаты

Пусть, как обычно, $\mathbb{R}$, $\mathbb{Q}$, $\mathbb{Z}$, и $\mathbb{N}$ обозначают, соответственно, множества вещественных, рациональных, целых и натуральных чисел. Пусть $\mathbb{R}^n$, $\mathbb{Q}^n$, и $\mathbb{Z}^n$ обозначают соответствующие конечномерные векторные пространства, а $\mathbb{N}_m = \{1,\ldots,m\}$. Функцию $f\colon \mathbb{R}^n \to \mathbb{R}$ вида $f(x) = c^{\mathrm{T}}x - d$, где $c \in \mathbb{Q}^n$, $d \in \mathbb{Q}$, будем называть *аффинной функцией* (с рациональными коэффициентами).

Определение 1. *Пусть $f_1,\ldots,f_q\colon \mathbb{R}^n \to \mathbb{R}$ — аффинные функции, и A, B — конечные подмножества $\mathbb{R}^n$. Конечная последовательность $Q = (f_1,\ldots,f_q)$ называется аффинным комитетом, разделяющим A и B, если*

$$\bigl|\{i \in \mathbb{N}_q :\ f_i(a) > 0\}\bigr| > \frac{q}{2} \quad (a \in A),$$
$$\bigl|\{i \in \mathbb{N}_q :\ f_i(b) < 0\}\bigr| > \frac{q}{2} \quad (b \in B).$$

Число q называется числом элементов комитета Q, а множества A и B — отделимыми этим комитетом.

Согласно критерию Мазурова [7], множества A и B отделимы аффинным комитетом тогда и только тогда, когда $A \cap B = \varnothing$. Тем не менее, по ряду естественных причин, вызывают интерес разделяющие комитеты с наименьшим возможным (для заданных множеств) числом элементов, называемые *минимальными комитетами*.

Задача 1 (Минимальный аффинный разделяющий комитет, MASC**).** Для заданных конечных множеств $A, B \subset \mathbb{Q}^n$ требуется найти аффинный комитет Q, разделяющий эти множества, с наименьшим числом элементов.

*Работа выполнена при финансовой поддержке Президента РФ, проекты НШ-2081.2008.1 и МД-370.2008.1 и РФФИ, проект № 07-07-00168.

[1] То есть для нее не существует полиномиальной приближенной схемы (PTAS) при условии P ≠ NP.

Сложность и аппроксимируемость задачи MASC в общем случае определяются приведенными ниже теоремами.

Теорема 1 ([3]). *Задача* MASC NP*-трудна и сохраняет труднорешаемость при дополнительном ограничении* $A \cup B \subset \{x \in \{0,1,2\}^n : \|x\|_2 \leqslant 2\}$.

Теорема 2 ([5]). *Если* $P \neq NP$*, задача* MASC *не принадлежит классу* Apx*. Более того, если не выполнено условие* $\mathsf{NP} \subset \mathsf{DTIME}(2^{\mathrm{poly}(\log n)})$*, то существует константа* $D > 0$ *такая, что точность* r *произвольного полиномиального приближенного алгоритма задачи* MASC *удовлетворяет неравенству* $r \geqslant D \log\log\log m$.

Будем использовать обозначение MASC(n) для задачи MASC, сформулированной в пространстве фиксированной размерности n. Известно [7], что задача MASC(n) полиномиально разрешима при $n = 1$. При произвольном $n > 1$ задача NP-трудна. Справедливость этого факта следует из труднорешаемости следующей задачи.

Задача 2 (Аффинный разделяющий комитет на плоскости, PASC**).** Заданы множества $A = \{a_1, \ldots, a_{m_1}\}$ и $B = \{b_1, \ldots, b_{m_2}\}$, $A, B \subset \mathbb{Q}^2$, и натуральное число t. Существует ли аффинный комитет Q с не более чем t элементами, разделяющий множества A и B?

Видно, что задача PASC является модификацией задачи MASC(2) в виде задачи верификации свойства, и принадлежит NP. Доказательство ее труднорешаемости следует из полиномиальной сводимости (к ней) известной NP-полной задачи о покрытии конечного множества (точек) на плоскости множеством прямых линий, известной в литературе как задача PC.

Задача 3 (Покрытие прямыми множества на плоскости, PC**).** Заданы конечное подмножество P плоскости и натуральное число s. Существует ли такое покрытие L множества P прямыми, что $|L| \leqslant s$?

Если множество P находится в общем положении, т. е. никакие три точки из P не лежат на одной прямой, то ответ в задаче PC может быть получен тривиально («Да», если $s \geqslant \lceil |P|/2 \rceil$ и «Нет», в противном случае) за полиномиальное от логарифма размера задачи время. Тем не менее, в общем случае задача PC, как известно, труднорешаема.

Теорема 3 ([8]). *Задача* PC NP*-полна в сильном смысле.*

Договоримся далее использовать следующие обозначения: $B(x_0, \varepsilon) = \{x \in \mathbb{R}^2 : \|x - x_0\|_2 \leqslant \varepsilon\}$ — круг радиуса ε с центром в точке x_0; $\mathrm{aff}(P)$ — аффинная оболочка множества P, и $\dim$ — размерность аффинного (или линейного) многообразия.

Пусть далее множество $P = \{p_1, \ldots, p_k\} \subset \mathbb{Z}^2$ и натуральное число s задают условие задачи PC. Определим числа ρ и ε по формулам

$$\rho = \max\{\|p\|_2 : p \in P\}, \quad \varepsilon = \frac{1}{6(2\rho+1)+1}. \quad (1)$$

Зафиксируем вектор σ, $\|\sigma\|_2 = 1$, так, чтобы для произвольной пары $\{i, j\} \subset \mathbb{N}_k$, отрезки прямых $[p_i - \varepsilon\sigma, p_i + \varepsilon\sigma]$ и $[p_j - \varepsilon\sigma, p_j + \varepsilon\sigma]$ не лежали на одной прямой. Сопоставим исходной задаче PC условие (A, B, t) соответствующей задачи PASC по формулам: $A = P$, $B = (P - \varepsilon\sigma) \cup (P + \varepsilon\sigma)$ и $t = 2s + 1$. Легко показать, что описанная выше процедура сведения может быть выполнена за время, ограниченное сверху полиномом от длины записи условия исходной задачи PC. Справедлива следующая

Теорема 4 ([5]). *Множество* P *обладает покрытием из* s *прямых тогда и только тогда, когда множества* $A = P$ *и* $B = (P - \varepsilon\sigma) \cup (P + \varepsilon\sigma)$ *отделимы аффинным комитетом из* $2s + 1$ *элемента.*

Следствие 1. *Задача* PASC NP*-полна в сильном смысле. Задача* ASC(n)[2] *при произвольном* $n > 1$ *— также* NP*-полна в сильном смысле. Задача* MASC(n) NP*-трудна при произвольном* $n > 1$.

Случай общего положения

Доказательство теоремы 4 существенно опиралось на неявное допущение о возможности рассмотрения «вырожденных» условий задачи PASC, задаваемых множеством $A \cup B$ не находящимся в общем положении. В этом разделе приводится аналогичный результат, полученный без этого допущения.

Определение 2. *Говорят, что множество* $Z \subset \mathbb{R}^n$, $|Z| > n$*, находится в общем положении, если для каждого его подмножества* Z', $|Z'| = n + 1$*, справедливо равенство* $\dim \mathrm{aff}(Z') = n$.

В частности, множество $Z \subset \mathbb{R}^2$ находится в общем положении (по определению), если оно не содержит подмножества $Z' = \{z_1, z_2, z_3\} \subseteq Z$ элементы которого являются точками, лежащими на одной прямой. Очевидно, что условия задачи PASC, в которых множество $A \cup B$ не находится в общем положении, очень редки (при естественном определении вероятностной меры они составляют множество меры нуль). Таким образом, представляет интерес изучение вычислительной сложности задачи PASC при дополнительном условии общности положения множества $A \cup B$.

Задача 4 (PASC **для множеств в общем положении,** PASC-GP**).** Заданы множества $A = \{a_1, \ldots, a_{m_1}\}$ и $B = \{b_1, \ldots, b_{m_2}\}$, $A, B \subset \mathbb{Q}^2$, так,

[2] ASC(n) — версия задачи MASC(n) в виде задачи верификации свойства.

что множество $A \cup B$ находится в общем положении, и натуральное число t. Существует ли аффинный комитет Q с не более, чем t элементами, разделяющий множества A и B?

Подобно построениям, проведенным в предыдущем разделе, рассмотрим условие задачи PC, определяемое множеством P из k точек с целочисленными координатами на плоскости и некоторым натуральным числом s. Определим числа ρ и ε по формулам (1). Зафиксируем 2-мерные векторы σ и τ так, чтобы $\|\sigma\|_2 = \|\tau\|_2 = 1$, $\sigma^\intercal \tau = 0$, и для каждого $\{i, j\} \subset \mathbb{N}_k$ пары отрезков

$$[p_i - \varepsilon\sigma, p_i + \varepsilon\sigma], \quad [p_j - \varepsilon\sigma, p_j + \varepsilon\sigma];$$
$$[p_i - \varepsilon\tau, p_i + \varepsilon\tau], \quad [p_j - \varepsilon\tau, p_j + \varepsilon\tau];$$

не лежали на одной прямой. Сопоставим условию исходной задачи PC подходящее условие (A, B, t) задачи PASC-GP, определяемое равенствами: $A = \left\{p \pm \frac{\varepsilon(p)}{M}\tau \colon p \in P\right\}$, $B = \left\{p \pm \varepsilon(p)\sigma \colon p \in P\right\}$, и $t = 2s + 1$.

Здесь числа $\varepsilon(p)$ и $M > 0$ выбраны таким образом, что верно неравенство $\max\limits_{p \in P} \frac{\varepsilon(p)}{M} < \min\limits_{p \in P} \varepsilon(p)$, и множество $A \cup B$ находится в общем положении. Как и в случае с задачей PASC, легко убедиться в том, что описанная выше редукция может быть произведена за полиномиальное время.

Теорема 5. *Множество $P = \{p_1, \ldots, p_k\} \subset \mathbb{Z}^2$ обладает покрытием из s прямых тогда и только тогда, когда множества $A = \left\{p \pm \frac{\varepsilon(p)}{M}\tau \colon p \in P\right\}$ и $B = \left\{p \pm \varepsilon(p)\sigma \colon p \in P\right\}$ отделимы аффинным комитетом из $2s + 1$ элемента.*

Следствие 2. *Задача* PASC-GP NP*-полна в сильном смысле. Задачи* MASC-GP(n) *и* MASC-GP NP*-трудны*[3].

Замечание 1. Заметим, что в теоремах 4 и 5 обосновывается полиномиальная сводимость задачи MinPC[4] к задачам MASC(2) и MASC-GP(2), соответственно, *сохраняющая точность аппроксимации* (тем самым, произвольный полиномиальный приближенный алгоритм задачи MASC(2) или MASC-GP(2) индуцирует полиномиальный приближенный алгоритм для задачи MinPC с такой же точностью аппроксимации).

Доказательство труднорешаемости задачи PC в теореме 3 было получено как следствие полиномиальной сводимости к ней известной задачи 3SAT. Проведя аналогичные рассуждения для задачи Gap-3SAT$(5)_{1,\rho}$, труднорешаемость которой является следствием известной PCP-теоремы [9], получим результат, касающийся аппроксимируемости задачи MinPC.

Задача 5 (Gap-3SAT$(5)_{1,\rho}$). Для заданной 3-КНФ необходимо дать ответ «Да», если найдется разрешающий ее набор истинности, ответ «Нет», если для произвольного набора истинности, доля разрешаемых им дизъюнктов меньше ρ. В противном случае ответ не определен.

Теорема 6. *Задача* MinPC Max-SNP*-трудна.*

Следствие 3. *При произвольном $n > 1$ задача* MASC-GP(n) Max-SNP*-трудна.*

Приближенный алгоритм

В этом разделе приведен новый приближенный алгоритм для задачи MASC-GP(n). Фактически, алгоритм является модификацией известного алгоритма [6], эффективно использующей дополнительные ограничения задачи MASC-GP(n) для повышения точности аппроксимации. Введем необходимые определения и обозначения. Пусть условие задачи MASC-GP(n) задается конечным множеством $Z = A \cup B \subset \mathbb{Q}^n$.

Определение 3. *Подмножество $Z' = A' \cup B'$, в котором $A' \subseteq A$, $B' \subseteq B$, называется аффинно разделимым подмножеством (множества Z), если найдутся вектор $c \in \mathbb{R}^n$ и число $d \in \mathbb{R}$ такие, что*

$$c^\intercal a - d > 0, \ (a \in A'), \quad c^\intercal b - d < 0, \ (b \in B'). \quad (2)$$

Обозначим множество решений системы (2) через $\mathfrak{S}(Z')$.

Определение 4. *Аффинно разделимое подмножество Z' называется максимальным (по включению) аффинно разделимым подмножеством множества Z, если для каждого $z \in Z \setminus Z'$ справедливо равенство $\mathfrak{S}\big(Z' \cup \{z\}\big) = \varnothing$.*

Обозначим через $\mathfrak{M}(Z)$ множество максимальных аффинно разделимых подмножеств множества Z. Справедлива одна из следующих альтернатив:

1) множество Z аффинно разделимо, то есть $\mathfrak{M}(Z) = \{Z\}$; в этом случае произвольная пара $(c, d) \in \mathfrak{S}(Z)$ порождает конечную последовательность $Q = (c^\intercal x - d)$ (состоящую из одного элемента), являющуюся решением задачи MASC-GP(n).

2) найдется собственное подмножество $Z' \subset Z$, $Z' \neq Z$ такое, что $Z' \in \mathfrak{M}(Z)$.

Без ограничения общности можно ограничиться комитетами, порождаемыми максимальными по включению аффинно разделимыми подмножествами множества $Z = A \cup B$, свойства которых удобно описывать в терминах теории графов.

[3] Подобно PASC-GP, задачи MASC-GP(n) и MASC-GP являются модификациями задачи MASC и MASC(n) с дополнительным ограничением общности положения разделяемых множеств.

[4] MinPC — оптимизационная версия задачи PC.

Алгоритм 1. Greedy Committee.

Вход: $Z = A \cup B \subset \mathbb{Q}^n$, $|Z| = m$.
Выход: комитет Q, разделяющий мн-ва A и B.

построить граф $G_Z = (V, E)$;
если $V = \{Z\}$, **то**
  $K := (Z)$; $q_{\min} := 1$; перейти на шаг 13;
$q_{\min} := \infty$;
для всех $\zeta \in V$
  $K(\zeta) := (\zeta)$; $J := Z \setminus \zeta$; $q(\zeta) := 1$;
  пока $J \neq \varnothing$
    $\{Z', Z''\} := \arg\max\{|X_1 \cap X_2 \cap J| \colon \{X_1, X_2\} \in E\}$;
    добавить множества Z' и Z'' к последовательности $K(\zeta)$;
    $J := J \setminus (Z' \cap Z'')$ и $q(\zeta) := q(\zeta) + 2$;
  если $q(\zeta) < q_{\min}$, **то**
    $K := K(\zeta)$; $q_{\min} := q(\zeta)$;
пусть $K = (Z'_1, \ldots, Z'_{q_{\min}})$;
для всех $i \in \mathbb{N}_{q_{\min}}$
  $f_i(x) := c_i^{\intercal} x - d_i$, где (c_i, d_i) — произвольный элемент множества $\mathfrak{S}(Z'_i)$;
вернуть $Q = (f_1, \ldots, f_{q_{\min}})$.

Определение 5. *Конечный граф $G_Z = (V, E)$ называется графом максимальных аффинно разделимых подмножеств множества Z, если $V = \mathfrak{M}(Z)$ и для каждого $\{Z'_1, Z'_2\} \subset V$,*

$$\{Z'_1, Z'_2\} \in E \iff Z'_1 \cup Z'_2 = Z.$$

Введем естественное

Допущение 1. Пусть для аффинно неразделимого множества $Z = A \cup B$ найдется число t и подмножества $Z'_0, Z'_1, \ldots, Z'_{2t} \in V$ (не обязательно различные) такие, что $\{Z'_{2j-1}, Z'_{2j}\} \in E$, $j \in \mathbb{N}_t$, и для произвольных $(c_i, d_i) \in \mathfrak{S}(Z'_i)$, $i = 0, \ldots, 2t$ последовательность $Q = (c_0^{\intercal} x - d_0, c_1^{\intercal} x - d_1, \ldots, c_{2t}^{\intercal} x - d_{2t})$ — минимальный аффинный комитет, разделяющий множества A и B.

Условие, приведенное в допущении 1, может показаться на первый взгляд излишне строгим. Тем не менее, согласно результатам численных экспериментов (которые мы для краткости опускаем), случайно выбранное условие задачи MASC-GP(n) с высокой вероятностью ему удовлетворяет. Кроме того, справедливо

Утверждение 7. *Условия задач* PASC *и* PASC-GP, *построенные на этапе полиномиального сведения задачи* PC *при доказательстве теорем 4 и 5, удовлетворяют допущению 1.*

Теорема 8. *Пусть множество $Z = A \cup B \subset \mathbb{Q}^n$, $|Z| = m$, задает условие задачи* MASC-GP(n). *Сложность алгоритма «Greedy Committee» составит $O\left(\binom{m}{n}^3 + \Theta m\right)$, где Θ — сложность подзадачи нахождения решения совместной системы из не более чем m линейных неравенств от $n+1$ переменной. Точность алгоритма $O(m/n)$.*

Пусть допущение 1 верно для множества Z, тогда точность алгоритма составит $O(\log m)$.

Выводы

Задача MASC труднорешаема не только в общем случае (когда размерность n пространства признаков не ограничена), но и в пространстве произвольной фиксированной размерности $n > 1$, даже при дополнительном ограничении общности положения множества $A \cup B$. Задача MASC трудно аппроксимируема. В общем случае точность любого полиномиального приближенного алгоритма задачи находится в пределах $(\Theta(\log\log\log m), O(m))$. Однако при фиксированной размерности, дополнительном условии общности положения $A \cup B$ и допущении 1, задача может быть решена за полиномиальное время с точностью $O(\log m)$. Полученная оценка, по ряду причин, представляется неулучшаемой. Доказательство (или опровержение) этого факта пока неизвестно и является предметом дальнейших исследований.

Литература

[1] *Lin J. H., and Vitter J. S.* Complexity Results on Learning by Neural Nets // Machine Learning. 1991. No. 6. Pp. 211–230.

[2] *Megiddo N.* On the complexity of polyhedral separability // Discrete and Computational Geometry. 1988. No. 3. Pp. 325–337.

[3] *Хачай М.Ю.* О вычислительной сложности задачи о минимальном комитете и смежных задач // ДАН. 2006. Т. 406. №6. С. 742–745.

[4] *Мазуров Вл.Д., Хачай М.Ю.* Параллельные вычисления и комитетные конструкции // Автоматика и телемеханика. 2006. №5.

[5] *Khachai M. Yu.* Computational and Approximational Complexity of Combinatorial Problems Related to the Committee Polyhedral Separability of Finite Sets // Pattern Recognition and Image Analysis. 2008. Vol. 18. No. 2. Pp. 237–242.

[6] *Хачай М.Ю.* О вычислительной и аппроксимационной сложности задачи о минимальном аффинном комитете // Таврический вестник информатики и математики. 2006. № 1. С. 34–43.

[7] *Mazurov Vl.D.* Комитеты систем неравенств и задача распознавания образов // Кибернетика. 1971. №3. С. 140–146.

[8] *Megiddo N., and Tamir A.* On the complexity of locating linear facilities in the plane // Operations research letters. 1982. Vol. 1. No. 5. Pp. 194–197.

[9] *Vazirani V.* Approximation algorithms. New York: Springer. 2001.

Параллельный подход к вычислению двумерного дискретного косинусного преобразования в специальных алгебраических структурах*

Чичёва М. А.
mchi@smr.ru
Самара, Институт систем обработки изображений РАН

В работе исследуется эффективность параллельного подхода к вычислению двумерного дискретного косинусного преобразования. Предложен алгоритм, основанный на использовании представления данных в гиперкомплексной алгебре.

Двумерное дискретное косинусное преобразование (ДКП):

$$\hat{x}(m_1, m_2) = \sum_{n_1=0}^{N-1} \sum_{n_2=0}^{N-1} x(n_1, n_2) h_{m_1 m_2}(n_1, n_2), \quad (1)$$

где $h_{m_1 m_2}(n_1, n_2) =$

$$= \lambda_{m_1} \lambda_{m_2} \cos\left(\frac{\pi(n_1 + \frac{1}{2})m_1}{N}\right) \cos\left(\frac{\pi(n_2 + \frac{1}{2})m_2}{N}\right),$$

$$\lambda_m = \begin{cases} \frac{2}{\sqrt{N}}, & \text{при } m \neq 0; \\ \frac{1}{\sqrt{N}}, & \text{при } m = 0, \end{cases}$$

широко используется при обработке и анализе изображений в задачах спектрального анализа, компрессии, распознавания образов и т. п. [1].

За годы развития теории быстрых алгоритмов дискретных преобразований разработано множество эффективных алгоритмов вычисления одно- и двумерного ДКП. Часть из них обладает неулучшаемыми оценками вычислительной сложности. Однако современные задачи продолжают требовать все более высоких скоростей обработки данных. Наиболее естественным путем уменьшения времени вычисления преобразований в этой ситуации является формирование параллельных алгоритмов.

В настоящей работе исследуется эффективность параллельного подхода к вычислению двумерного ДКП, основанного на следующих основных идеях:

— сведение дискретного косинусного преобразования к дискретному преобразованию Фурье (ДПФ) того же размера [1];
— погружение данных в специальную алгебраическую структуру — четырехмерную алгебру гиперкомплексных чисел (необходимая информация изложена, например, в [2]);
— использование параллельных алгоритмов гиперкомплексного ДПФ [2, 3, 4].

*Работа выполнена при финансовой поддержке РФФИ, проекты №07-01-96612 и №09-01-00511.

Базовый алгоритм двумерного ДКП

В [1] введён алгоритм двумерного ДКП чётной длины, сводящий его к ДПФ того же размера. При этом используется представление данных в алгебре кватернионов. Очевидно, что аналогичным образом преобразование (1) может быть сведено к гиперкомплексному двумерному ДПФ:

$$G(m_1, m_2) = \sum_{n_1=0}^{N-1} \sum_{n_2=0}^{N-1} g(n_1, n_2) w_1^{m_1 n_1} w_2^{m_2 n_2}, \quad (2)$$

где корни w_1, w_2 N-й степени лежат в различных подалгебрах, изоморфных $\mathbb{C}$, четырехмерной гиперкомплексной алгебры $\mathbb{B}_2$. Её произвольный элемент имеет вид:

$$z = a + bi + cj + dij. \quad (3)$$

Правила умножения базисных элементов имеют вид:

$$i^2 = j^2 = -1, \quad ij = ji.$$

Как и алгебра кватернионов $\mathbb{H}$, алгебра гиперкомплексных чисел $\mathbb{B}_2$ имеет четыре тривиально реализуемых автоморфизма:

$$\varepsilon_o(z) = a + bi + cj + dij;$$
$$\varepsilon_i(z) = a + bi - cj - dij;$$
$$\varepsilon_j(z) = a - bi + cj - dij;$$
$$\varepsilon_{ij}(z) = a - bi - cj + dij.$$

Это позволяет говорить о симметриях гиперкомплексного спектра (2) в случае вещественного входного сигнала $x(n_1, n_2) \in \mathbb{R}$:

$$X(N - m_1, m_2) = \varepsilon_j(X(m_1, m_2));$$
$$X(m_1, N - m_2) = \varepsilon_i(X(m_1, m_2));$$
$$X(N - m_1, N - m_2) = \varepsilon_{ij}(X(m_1, m_2)).$$

С учётом вышесказанного алгоритм вычисления двумерного ДКП с использованием гиперкомплексного ДПФ принимает следующий вид.

Шаг 1. Формирование вспомогательного симметричного сигнала удвоенного размера:

$$f(n_1, n_2) = f(2N-1-n_1,\, n_2) =$$
$$= f(n_1,\, 2N-1-n_2) =$$
$$= f(2N-1-n_1,\, 2N-1-n_2) = x(n_1, n_2).$$

Шаг 2. Выделение из него отсчетов с чётными индексами:

$$g(n_1, n_2) = f(2n_1, 2n_2), \quad 0 \leqslant n_1, n_2 \leqslant N-1.$$

Шаг 3. Вычисление двумерного гиперкомплексного ДПФ (2) того же размера.

Шаг 4. Формирование косинусного спектра:

$$\hat{x}(m_1, m_2) = \mathrm{Re}\Big(G(m_1, m_2) w_{1(2N)}^{m_1/2} w_{2(2N)}^{m_2/2}\Big), \qquad (4)$$

где $w_{1(2N)} = \exp\big(\frac{2\pi i}{2N}\big)$, $w_{2(2N)} = \exp\big(\frac{2\pi j}{2N}\big)$ — корни из единицы степени $2N$.

Вычислительная сложность такого алгоритма равна

$$M_{\mathrm{dct}}(N\times N) = M_{\mathrm{dft}}(N\times N) + 4N^2;$$
$$A_{\mathrm{dct}}(N\times N) = A_{\mathrm{dft}}(N\times N) + 3N^2;$$

где $M_{\mathrm{dft}}(N\times N)$, $A_{\mathrm{dft}}(N\times N)$ — мультипликативная и аддитивная сложность используемого алгоритма гиперкомплексного ДПФ. Отсюда время работы предлагаемого алгоритма оценивается как

$$T_{\mathrm{dct}}(N) = T_{\mathrm{dft}} + 7N^2 T_{\mathrm{op}}, \qquad (5)$$

где T_{dft} — время вычисления ДПФ (2), T_{op} — время выполнения одной вещественной операции.

Параллельный алгоритм двумерного гиперкомплексного ДПФ

Для вычисления вспомогательного дискретного преобразования Фурье могут быть использованы параллельные алгоритмы, предложенные в [2, 3, 4]. Необходимо отметить, что указанные алгоритмы основаны на переходе от представления (3) гиперкомплексных чисел в алгебре $\mathbb{B}_2$ к представлению в изоморфной ей алгебре $\mathbb{C}\oplus\mathbb{C}$ (прямая сумма комплексных алгебр). Базисные элементы этой алгебры могут быть записаны в виде:

$$u_0 = 1 + ij,\ u_1 = 1 - ij,\ u_2 = i + j,\ u_3 = i - j,$$

с правилами умножения базисных элементов:

$$u_j^2 = \begin{cases} u_j, & \text{при } j = 0, 1; \\ -u_{j-2}, & \text{при } j = 2, 3; \end{cases}$$
$$u_j u_k = \begin{cases} 2u_k, & \text{при } k = j+2; \\ 0, & \text{в остальных случаях.} \end{cases}$$

В этом случае все операции над гиперкомплексными числами могут быть распараллелены на два полностью независимых процесса, а вместо представления (3) используется вид:

$$z = \tfrac{1}{2}\big((a+d)u_0 + (a-d)u_1 + (b-c)u_2 + (b+c)u_3\big). \quad (6)$$

Наиболее быстрым с точки зрения общего времени выполнения преобразования является алгоритм [4], учитывающий вещественность входных данных. Он состоит из следующих шагов.

Шаг 1. Преобразование входных данных из представления (3) к виду (6). Тривиально в силу вещественности входных данных.

Шаг 2. Рассылка данных. На первом процессоре остаются коэффициенты при u_0, u_2, на второй отсылаются коэффициенты при u_1, u_3. На самом деле, в случае вещественных входных данных коэффициенты при u_2 и u_3 равны нулю, что уменьшает объем пересылки.

Шаг 3. Вычисление преобразования на каждом процессоре. Необходим обмен результатами расчётов для учёта вещественности данных.

Шаг 4. Объединение результатов на первом процессоре.

Шаг 5. Возвращение к исходному представлению данных.

Теоретическая оценка времени работы такого алгоритма может быть получена на основании результатов работы [4] при размерности обрабатываемого сигнала $d = 2$:

$$T_{\mathrm{dft}} = \tfrac{1}{2}N^2 L_b T_m (7\log_2 N + 5) + T_l(2\log_2 N + 3) + \\ + \tfrac{1}{4}N^2 T_{\mathrm{op}}(7\log_2 N + 4), \quad (7)$$

где L_b — размер представления вещественных чисел в байтах (обычно 4 или 8 байт), T_m — время пересылки одного байта, T_l — время латентности при пересылке данных.

Объединение алгоритмов и оценка эффективности параллельной реализации

При объединении алгоритмов, описанных в предыдущих разделах, возникает возможность их некоторой модификации. Так, шаги 1 и 2 алгоритма ДКП могут быть объединены с рассылкой данных на шаге 1 параллельного алгоритма ДПФ. С точки зрения вычислений все эти шаги тривиальны.

Шаг 4 алгоритма ДКП может быть выполнен в параллельной форме сразу после шага 3 алгоритма ДПФ. То есть дополнительные умножения на $w_{1(2N)}$, $w_{2(2N)}$ в соотношении (4) могут так же быть выполнены с использованием представления (6) параллельно на двух процессорах, причем, так как заранее известен способ дальнейшего объединения результатов для перехода к ДКП, можно вычислять не все компоненты, тем самым уменьшая объем пересылки и вычислений на шагах 4 и 5 алгоритма ДПФ.

При этом оценка (7) так же изменится, поскольку на шаге 4 алгоритма ДПФ объём пересылаемых

данных уменьшается вдвое, а на шаге 5 вместо вычисления четырёх компонент исходного представления (3) гиперкомплексных чисел, достаточно вычислить только вещественную часть. Получим:

$$T'_{\text{dft}} = T_{\text{dft}} - \tfrac{1}{4}N^2 L_b T_m - \tfrac{3}{4}N^2 T_{\text{op}} =$$
$$= \tfrac{1}{4}N^2 L_b T_m (14\log_2 N + 9) + T_l(2\log_2 N + 3) +$$
$$+ \tfrac{1}{4}N^2 T_{\text{op}}(7\log_2 N + 1).$$

А оценка времени работы всего параллельного алгоритма двумерного ДКП примет вид:

$$T'_{\text{dct}}(N\times N) = T'_{\text{dft}}(N\times N) + 5N^2 T_{\text{op}}.$$

Для оценки эффективности предложенного алгоритма оценим время его работы в параллельном и последовательном варианте. При этом время работы последовательного алгоритма на основании соотношения (5) можно оценить как:

$$T_{1\text{dct}} = \big(M_{\text{dft}}(N\times N) + A_{\text{dft}}(N\times N) + 7N^2\big)T_{\text{op}},$$

используя оценки сложности, полученные в работе [5], для алгоритма гиперкомплексного ДПФ по основанию 2 с учётом вещественности входных данных:

$$M_{\text{dft}}(N\times N) = \tfrac{9}{8}N^2\log_2 N - \tfrac{23}{8}N^2,$$
$$A_{\text{dft}}(N\times N) = \tfrac{29}{8}N^2\log_2 N - \tfrac{39}{8}N^2.$$

В таблице 1 приведены оценки времени работы последовательного и параллельного алгоритма двумерного ДКП, рассчитанные при значениях (в микросекундах): $T_{\text{op}} = 4,\ T_l = 5{,}6,\ T_m = 0{,}012$.

Таблица 1. Оценка времени работы алгоритмов (с).

N	$T_{1\text{dct}}$	T'_{dct}	$U = T_{1\text{dct}}/T'_{\text{dct}}$
512	22,4	41,9	1,87
1024	97,3	187,7	1,93
2048	419,2	830,5	1,98

Высокая эффективность распараллеливания (близкая к единице) объясняется тем, что здесь использован эффективный параллельный алгоритм ДПФ, а также выполнена его модификация с учётом использования результатов преобразования для формирования косинусного спектра.

Выводы

В работе предложен высокоэффективный параллельный алгоритм двумерного дискретного косинусного преобразования. Интерес именно к этому случаю обусловлен тем, что данный алгоритм использует два процессора, а в настоящее время двухпроцессорные системы стали повседневной реальностью. Несмотря на предполагаемую высокую эффективность распараллеливания и существенное снижение оценки времени вычисления преобразования, необходимо создать исследовательский программный комплекс и провести исследование алгоритма на таких реальных системах, а не на кластерах, как это делалось автором ранее. В случае, если имеется большее количество процессоров, время обработки может быть дополнительно снижено за счёт использования внутреннего параллелизма схемы декомпозиции Кули-Тьюки. Но это также является предметом отдельного исследования.

Литература

[1] *Гашников М. В., Глумов Н. И. и др.* Методы компьютерной обработки изображений. Под ред. Сойфера В.А. — М: Физматлит, 2003. — 784 с.

[2] *Алиев М. В., Чичева М. А.* Многомерное гиперкомплексное ДПФ: параллельный подход // Компьютерная оптика. — 2005. — № 27. — С. 135–137.

[3] *Chicheva M. A.* Theoretical and experimental estimation of hypercomplex discrete Fourier transform parallelization efficiency // Proceedings of The 2007 International Workshop on Spectral Methods and Multirate Signal Processing, Moscow, 2007. — Pp. 241–248.

[4] *Chicheva M. A.* Parallel implementation of multidimensional hypercomplex DFT with regard for real type of input signal // Proc. of 9-th International Conference on Pattern Recognition and Image Analysis: New Information Technologies. (PRIA–9–2008), Nizhnii Novgorod, 2008. — Vol. 1. — Pp. 70–73.

[5] *Aliev M. V., Chernov V. M.* Two-dimensional FFT-like algorithms with overlapping // Optical memory and neural networks (Information Optics). — 2002. — Vol. 11, No 1. — Pp. 29–38.

Обработка сигналов и анализ изображений

Код раздела: SI (Signal Processing and Image Analysis)

- Теория распознавания изображений.
- Структурный подход к анализу и распознаванию изображений.
- Морфологический анализ сигналов и изображений.
- Поиск изображений.
- Сжатие изображений.
- Обработка видеоинформации.
- Анализ и синтез дискретных последовательностей.
- Идентификация сигналов.
- Исследование динамических систем методами распознавания образов.
- Распознавание речи и звуковых сигналов.

Спектральный подход к вычислению аффинных инвариантов

Алёшин С. А., Дедус Ф. Ф., Тетуев Р. К.
4memph@gmail.com, ffdedus@impb.ru, ruslan.tetuev@gmail.com
Пущино, Институт математических проблем биологии РАН

В данной работе рассматривается задача, возникающая при вычислении определенных геометрических характеристик объектов, представимых на плоскости изображений в виде контуров. Контуры, в свою очередь, рассматриваются нами как параметрически заданные замкнутые кривые. Задача, главным образом, состоит в усовершенствовании алгоритмов, ранее предложенных исследователями для вычисления аффинных инвариантов контуров. На основе применения спектральных методов удалось предложить новые вычислительные приемы, значительно повышающие эффективность алгоритмов. В работе демонстрируются результаты применения изложенного подхода к синтезированным изображениям.

Известно, что контуры визуальных объектов часто можно считать наиболее информативной частью их представления. В задачах распознавания контуры удобно представлять на плоскости параметрически, в виде пары функций: $\Gamma(t) = \{x(t), y(t)\}$. Таким образом, можно численно описать очертания, характерные для каждого из рассматриваемых объектов. При перемещении в пространстве очертания наблюдаемых объектов испытывают различные геометрические преобразования (аффинные, перспективные и др.). Такое поведение может значительно затруднить задачу распознавания объектов и сцены в целом. Однако, в некоторых случаях в задачах компьютерного зрения перспективные преобразования можно представлять приближенно как аффинные (квазиаффинные). Ввиду этого особый интерес у многих исследователей вызывают характеристики контуров, инвариантные к *аффинным* преобразованиям, т. е. нечувствительные к подобного рода трансформациям объекта.

Аффинная длина дуги

Важной и наиболее востребованной на практике характеристикой, безусловно, является длина кривой (контура):

$$l = \int_a^b \sqrt{(\dot{x})^2 + (\dot{y})^2}\, dt.$$

Относительная длина отрезков нечувствительна к жёстким преобразованиям (параллельному переносу, повороту и т. д.). Однако, к сожалению, под действием других аффинных преобразований суммарная длина отрезков контура изменяется нелинейно. Это сильно затрудняет сопоставление контуров на этапе идентификации объектов, когда предполагается также действие нежёстких преобразований: сжатие, растяжение и др. Ранее для решения данной проблемы исследователями [1] была предложена формула расчёта так называемой *аффинной* длины дуги, как некоторой величины, нечувствительной к аффинным преобразованиям:

$$\tau = \int_a^b \sqrt[3]{|\dot{x}\ddot{y} - \ddot{x}\dot{y}|}\, dt. \qquad (1)$$

Данная формула имеет строгое аналитическое обоснование и могла бы быть крайнее полезной при нормализации контуров. Однако, наличие в данной формуле вторых производных сильно затрудняет её использование на практике. Это заставило большинство исследователей отказаться от попыток точного вычисления данной характеристики в пользу более простых аналитических соотношений. Например, в работах [2, 3] предложена замена величины (1) на вычислительно более простую, не требующую знания производных выше первого порядка:

$$\sigma = \frac{1}{2}\int_a^b |x\dot{y} - \dot{x}y|\, dt. \qquad (2)$$

Эту характеристику кривой принято называть аффинной площадью. Проблема применимости параметра (2) заключается в необходимом условии замкнутости рассматриваемых фрагментов кривой. С другой стороны, любая кривая (фрагмент) может стать замкнутой путём соединения её начальной и конечной точек отрезком, но это приведёт к априорному внесению ошибок в вычисления.

В настоящей работе предлагается алгоритм для более точного вычисления вторых производных параметрически заданных кривых, основанный на использовании спектральных методов. В результате будет показана применимость аналитического соотношения (1) на практике.

Предыдущие исследования

Геометрические характеристики визуальных объектов, в частности те, которые являются устойчивыми к аффинным преобразованиям, активно изучаются исследователями в течение последних десятилетий. Разработано множество алгоритмов для математического представления объектов, подвергнутых аффинным преобразованиям.

Они могут быть условно разделены на две группы: глобальные и локальные. Глобальные методы, в большинстве своём, основаны на применении спектрального аппарата для аппроксимации/интерполяции двумерных кривых, представляющих контуры объектов: в работах [5, 6] используются так называемые дескрипторы Фурье; в статьях [7, 8] применяется парное вейвлет-разложение; разложение по B-сплайнам применяется в [9]. Локальные методы основаны на алгоритмах выделения признаков: например, в работе [4] поиск инвариантов осуществлялся на основе преобразования Хафа; статья [10] основана на сопоставлении объектов посредством нахождения «критических» точек: точек перегиба, взаимных пересечений, максимумов кривизны и др.

Параметризация кривой

Известно множество разнообразных параметризаций для представления точек плоской кривой. Как уже было отмечено, в случае аффинных преобразований обычная длина дуги изменяется нелинейно и, поэтому, не может быть использована при параметризации кривой. Формула (1) позволяет перейти к параметризации кривой, инвариантной к эквиаффинным преобразованиям (сохраняющим площадь):

$$\tau(k) = \begin{cases} 0, & k = 1, \\ \int\limits_a^{b_k} \sqrt[3]{|\dot{x}\ddot{y} - \ddot{x}\dot{y}|}\, dt, & k = 2, \ldots, m, \end{cases}$$

где b_k — равноотстоящие друг от друга точки кривой, $b_1 = a$ и $b_m = b$. Очевидно, что данный параметр может стать абсолютным инвариантом при нормировке на аффинную длину всей кривой.

Спектральная аппроксимация

Рассмотрим параметрически заданную кривую $\Gamma(t) = \{x(t), y(t)\}$, где $t \in [0, T]$ — дискретный аргумент. Под спектральной аппроксимацией кривой $\Gamma(t)$ будем понимать её разложение по базису ортогональных (ортонормированных) полиномов $T_n(t)$:

$$\begin{aligned} \big[A_n^x, A_n^y\big] &= \sum_{t=0}^{T-1} \big[x(t), y(t)\big] T_n(t); \\ \big[\tilde{x}(t), \tilde{y}(t)\big] &= \sum_{n=0}^{N} \big[A_n^x, A_n^y\big] T_n(t); \end{aligned} \tag{3}$$

где A_n^x, A_n^y — коэффициенты разложения, $\tilde{x}(t)$, $\tilde{y}(t)$ — восстановленные (аппроксимированные) составляющие кривой $\Gamma(t)$, а n и N — соответственно степень полинома и глубина разложения. Необходимо отметить, что указанный способ разложения по

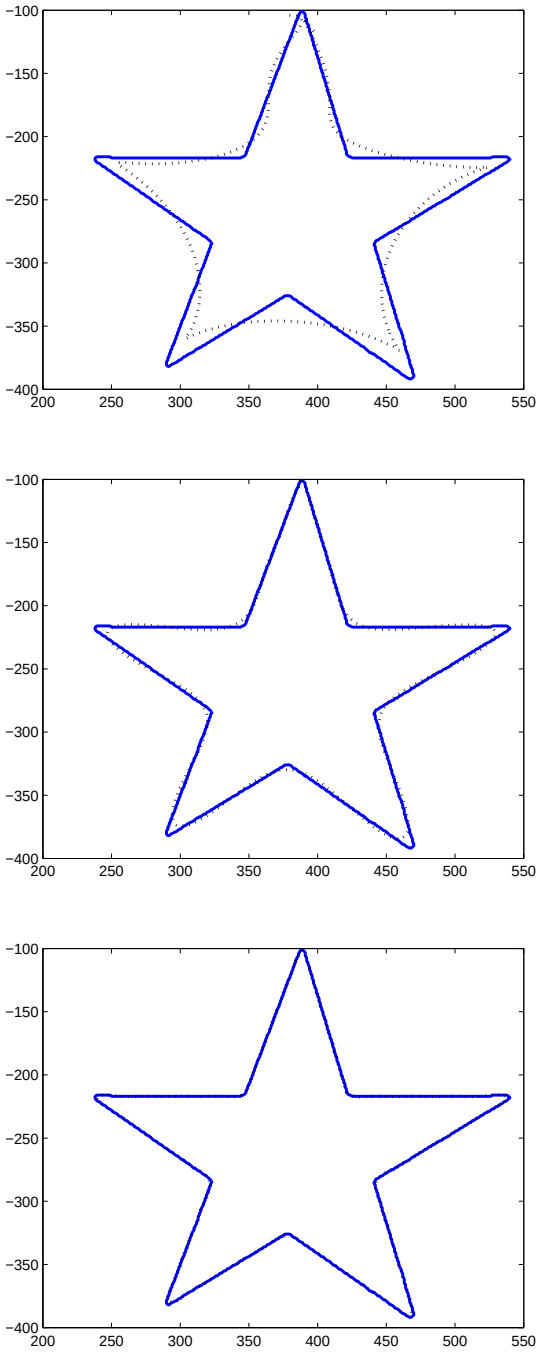


Рис. 1. Качество аппроксимации в зависимости от глубины разложения исходной замкнутой кривой. Сверху — 15 членов, в центре — 25 членов, снизу — 100 членов. Все три разложения произведены на сетке размером 1000 точек.

ортогональному базису и последующего восстановления кривой по её спектру является инвариантным относительно выбора самого базиса. Согласно [11], представление исследуемых сигналов в виде отрезков ортогональных рядов характерно тем, что структура таких описаний остается всегда неизмен-

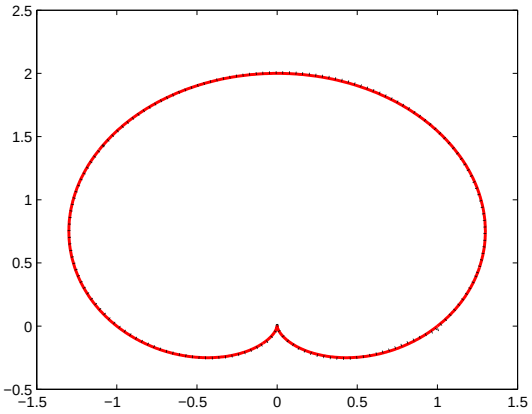


Рис. 2. Изображение кардиоиды. Сплошной линией отмечена исходная кривая, точечной линией обозначена восстановленная по 10 членам разложения.

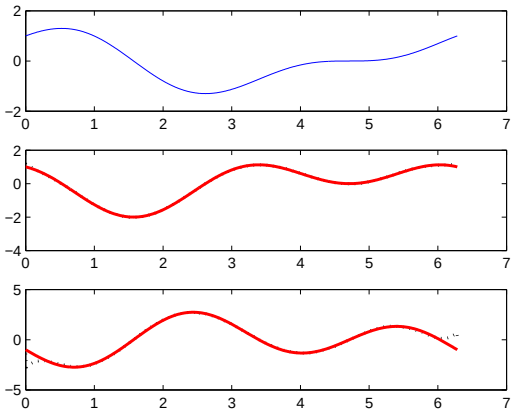


Рис. 3. Изображения $x(t)$ (сверху), $\dot{x}(t)$ (в центре), $\ddot{x}(t)$ (снизу); на центральном и нижнем графиках сплошной линией отмечена аналитически вычисленная производная, а точечной линией обозначена производная, вычисленная по методу каскадов и диффузий.

ной, а необходимая информация о сигналах содержится в коэффициентах разложения. В данной работе в качестве базиса выбраны полиномы Чебышёва дискретного аргумента, т. к. они удовлетворяют следующим свойствам:

— полиномы определены на равномерной сетке, что является крайне удобным при аналитическом описании векторизованного контура визуального образа;
— полиномы симметричны на своей области определения, что может являться дополнительным преимуществом при описании *замкнутых* кривых.

На рис. 1 представлены результаты аппроксимации контура пятиконечной звезды полиномами Чебышёва с разной глубиной разложения. Данная

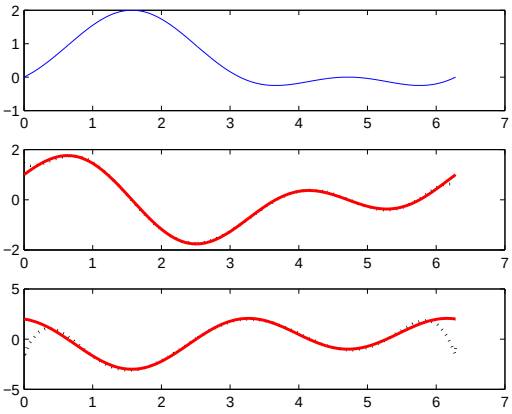


Рис. 4. Изображения $y(t)$ (сверху), $\dot{y}(t)$ (в центре), $\ddot{y}(t)$ (снизу); на центральном и нижнем графиках сплошной линией отмечена аналитически вычисленная производная, а точечной линией обозначена производная, вычисленная по методу каскадов и диффузий.

фигура была выбрана для иллюстрации аппроксимативных свойств полиномов в тех узлах сетки, где наблюдаются острые углы.

Вычисление производных

Большинство алгоритмов при оценке геометрических характеристик визуальных объектов, так или иначе, использует разностные методы вычисления производных первого и второго порядка. В нашей работе используется аналитическое описание контура выделенного на изображении объекта. В работе [12] показана возможность аналитической обработки сигнала (в нашем случае двумерной кривой), восстановленного по своему спектру, в пространстве коэффициентов разложения. Так, для аналитического вычисления производных первого и второго порядка будем пользоваться *методом каскадов и диффузий* [13], определяющим алгоритм для последовательных преобразований внутри спектра кривой. В качестве примера применения метода каскадов и диффузий на рис. 2, 3, 4 соответственно изображены гладкая кривая кардиоиды, её параметрическое представление в декартовой системе координат в виде $x(t)$, $y(t)$ вместе со своими первыми и вторыми производными.

Результаты

Как показали вычислительные эксперименты на синтезированных изображениях, удаётся получить значения производных первого и второго порядков с необходимой точностью. При этом, качество результата зависит от правильно подобранного базиса и оптимальной для данной задачи глубины разложения. В будущем авторами планируется использование данного аппарата для обработке реальных изображений.

Литература

[1] *El Oirrak A., Daoudi M., Aboutajdine D.* Affine invariant descriptors using Fourier series // Pattern Recognition Letters — 2002. — № 23 — Pp. 1109–1118.

[2] *Civi H., Ercil A.* Affine invariant 3L Fitting of Implicit Polynomials // PRIA — 2003. — Vol. 13, № 3 — Pp. 489–494.

[3] *Alferez R., Wang Y.-F.* Geometric and Illumination Invariants for Object Recognition // IEEE Transactions On PAMI. — 1999. — Vol. 21, № 6 — Pp. 505–536.

[4] *Lamdan Y., Schwartz J. T., Wolfson H. J.* Affine Invariant Model-Based Object Recognition // IEEE Trans. Robot. and Automat. — 1990. — Vol. 6, № 5 — Pp. 578–589.

[5] *Arbter K., Synder W. E., Burkhardt H., Hirzinger G.* Application of Affine Invariant Fourier Descriptors to Recognition of 3-D Objects // IEEE Transactions On PAMI. — 1990. — Vol. 12, № 7 — Pp. 640–647.

[6] *Lin C. C., Chellappa R.* Classification of partial 2-D shapes using Fourier descriptors // IEEE Transactions On PAMI. — 1987. — Vol. 9, № 5 — Pp. 686–690.

[7] *Tieng Q. M., Boles W. W.* Recognition of 2D Object Contours Using the Wavelet Transform Zero-Crossing Representation // IEEE Transactions On PAMI. — 1997. — Vol. 19, № 8 — Pp. 910–916.

[8] *Khalil M. I., Bayoumi M. M.* A Dyadic Wavelet Affine Invariant Function for 2D Shape Recognition // IEEE Transactions On PAMI. — 2001. — Vol. 23, № 10 — Pp. 1152–1154.

[9] *Huan Z. H., Cohen F. S.* Affine-invariant B-spline Moments for Curve matching // Image Processing. — 1996. — Vol. 5, № 10 — Pp. 1473–1480.

[10] *Freeman H.* Shape Description via the Use of Critical points // Pattern Recognition. — 1978. — Vol. 10, № 3 — Pp. 159–166.

[11] *Дедус Ф. Ф., Куликова Л. И., Панкратов А. Н., Тетуев Р. К.* Классические ортогональные базисы в задачах аналитического описания и обработки информационных сигналов. — М.: Издат. отд. Фак. ВМиК МГУ им. Ломоносова, 2004. — 172 с.

[12] *Новикова Д. А., Поволоцкий А. В.* Формулы для преобразования функций в пространстве коэффициентов разложения по базису Чебышева 2-го рода // Сборник статей молодых ученых факультета ВМиК МГУ. — 2007. — № 4 — С. 1–8.

[13] *Тетуев Р. К., Дедус Ф. Ф.* Классические ортогональные полиномы. Применение в задачах обработки данных. — Пущино: ИМПБ РАН, 2007.

Обнаружение и оценка частотных сдвигов в нестационарных процессах на основе многомасштабного корреляционного анализа

Анциперов В. Е.
antciperov@cplire.ru
Москва, ИРЭ им. В. А. Котельникова РАН

Рассматриваются результаты применения подхода на основе многомасштабного корреляционного анализа (МКА) к задачам обнаружения и оценки частотных сдвигов в сигналах. Показано, что предложенный подход подтверждает тезис о адекватности МКА для анализа нестационарных процессов. Найденные на основе МКА процедуры обнаружения/оценивания сдвига частот оказываются весьма простыми, что делает их перспективными с точки зрения практических применений. Показано, что имеет место альтернатива: производить оценки либо по самому распределению, либо по его Фурье преобразованию. Оценки первого типа обсуждаются далее как оценивание во временной области, а оценки второго типа — как оценивание в частотной.

Задача обнаружения (распознавания) и последующего оценивания (классификации) возможных частотных сдвигов для ряда процессов (сигналов) является актуальной, зачастую основной. В качестве примера из области техники коммуникаций можно привести так называемое частотно-сдвиговое кодирование (FSK), где отдельные символы сигнала кодируются разными частотами, и возникает проблема синхронизации приемника и передатчика. Еще более богатый набор примеров предоставляют сигналы биомедицинского происхождения. Действительно, в случае кардиологических сигналов большой интерес для специалистов представляют разного рода тахикардии, представляющие собой фрагменты учащенного сердцебиения. В области энцефалографии интерес представляют фрагменты спонтанных/вызванных ритмов ЭЭГ и их динамика, в случае эпилепсии — эпилептические разряды и их структура. Известно, что в области речевых сигналов наиболее интересными и трудными для обработки являются фрагменты резкого изменения основного тона (например на границах фонем). Список примеров может быть легко продолжен.

Основная проблема, связанная с резкими изменениями частоты колебаний некоторого сигнала, заключается в том, что он становится существенно нестационарным. По этой причине традиционные методы анализа сигналов (в первую очередь спектральный анализ) оказываются мало приспособленными к этой ситуации (отметим здесь, например, так называемый эффект «просачивания энергии» за границы участков колебательного поведения или границы изменения частоты колебаний). Поэтому в свое время большие усилия были направлены на поиск более адекватных методов анализа именно нестационарных сигналов.

Систематизации и унификации многочисленных подходов в этом направлении посвящена классическая монография [1]. В отличии от традиционных методов, использующих как правило линейные распределения (функционалы), методы анализа нестационарных процессов используют как правило квадратичные по отношению к сигналу распределения (так называемые распределения Коэновского класса). Очевидно, что при этом сложность вычислений возрастает с $\sim N$ до $\sim N^2$. В этой связи нами была предпринята попытка найти для анализа нестационарных процессов такие процедуры, которые бы совмещали достоинства процедур Коэновского класса и допускали бы быстрые реализации, типа $\sim N \ln N$. В результате был разработан подход на основе многомасштабного корреляционного анализа (МКА) [2], результатам применения которого в задачах обнаружения и оценки частотных сдвигов в нестационарных сигналах посвящена данная работа.

Корреляционное распределение МКА

Основы многомасштабного корреляционного анализа (МКА), мотивация его применения для реальных, нестационарных сигналов и обоснование деталей его алгоритмической реализации изложены в работе [2]. Формально МКА использует некоторую квадратичную по отношению к сигналу процедуру обработки с результирующим распределением $r\,(t, \vartheta)$ (см. ниже), являющимся Фурье-парой к некоторому частному частотно-временному распределению Коэновского класса [1]. По существу же, поскольку используются нормированные фрагменты сигнала, в центре внимания МКА находятся не вопросы частотно-временного распределения энергии, а вопросы структуры сигнала, в частности повторяемости и степени повторяемости его формы. По этой причине МКА ориентирован скорее на вопросы самоподобия, повторяемости сигнала (либо его отсутствия) на разных временных масштабах ϑ в окрестности каждого из анализируемых моментов времени t, нежели на исследование особенностей частотного спектра.

С целью кратко напомнить формализм МКА и согласовать обозначения, приведем конструкцию распределения $r\,(t, \vartheta)$, используемую как основное

средство анализа:

$$r(t,\vartheta) = \\ = \frac{\int S_L\big(t'+t-\frac{\vartheta}{2}\big)S_R\big(t'+t+\frac{\vartheta}{2}\big)\,dt'}{\sqrt{\int S_L^2\big(t'+t-\frac{\vartheta}{2}\big)\,dt'}\sqrt{\int S_R^2\big(t'+t+\frac{\vartheta}{2}\big)\,dt'}},$$

где

$$S_L(t') = G\big[\tfrac{2}{\vartheta}\big(t'-t+\tfrac{\vartheta}{2}\big)\big]x(t'),$$
$$S_R(t') = G\big[\tfrac{2}{\vartheta}\big(t'-t-\tfrac{\vartheta}{2}\big)\big]x(t')$$

— фрагменты анализируемого сигнала $x(t')$, выделенные слева и справа от текущего момента времени t окнами $G\big[\frac{2}{\vartheta}\big(t'-t+\frac{\vartheta}{2}\big)\big]$ и $G\big[\frac{2}{\vartheta}\big(t'-t-\frac{\vartheta}{2}\big)\big]$, см. рис. 1. Взвешивающие окна помимо смещений аргументов на $t\pm\frac{\vartheta}{2}$ зависят также от масштаба ϑ (в знаменателе), определяющего длительность фрагментов. Форма окон $G(\xi)$ симметрична, финитна на интервале $(-1,+1)$ и нормирована условием $\int G^2(\xi)\,d\xi = 1$. С учетом приведенных уточнений корреляционное распределение $r(t,\vartheta)$ имеет вид:

$$r(t,\vartheta) = \\ = \frac{\int G^2\big(\frac{2t'}{\vartheta}\big)x(t'+t_-)\,x(t'+t_+)\,dt'}{\sqrt{\int G^2\big(\frac{2t'}{\vartheta}\big)x(t'+t_-)dt'}\sqrt{\int G^2\big(\frac{2t'}{\vartheta}\big)x(t'+t_+)dt'}}, \quad (1)$$

где $t_\pm = t\pm\frac{\vartheta}{2}$. Пределы интегрирования в (1) считаются бесконечными, хотя реально, ввиду финитности $G(\xi)$, они составляют $\pm\frac{\vartheta}{2}$.

Смысл распределения (1) весьма прост — это скалярное произведение нормированных фрагментов $S_L(t')$ и $S_R(t')$. По виду $r(t,\vartheta)$ напоминает обычный коэффициент корреляции, однако, ввиду зависимости окон от масштаба, ϑ приобретает смысл не столько смещения окон, сколько размера (масштаба) сравниваемых фрагментов. Значения

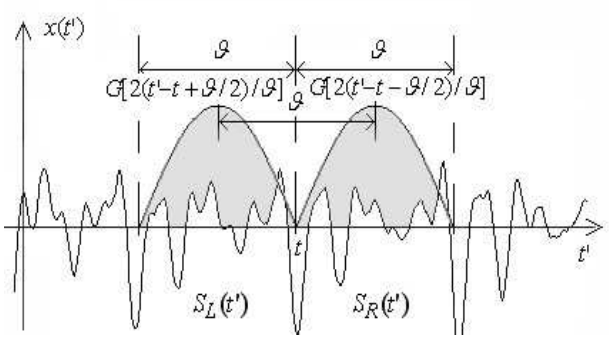


Рис. 1. Сигнал $x(t')$ и выделенные окнами $G\big[\frac{2}{\vartheta}\big(t'-t+\frac{\vartheta}{2}\big)\big]$ и $G\big[\frac{2}{\vartheta}\big(t'-t-\frac{\vartheta}{2}\big)\big]$ фрагменты $S_L(t')$ и $S_R(t')$ длительности ϑ, расположенные симметрично относительно текущего момента времени t (с взаимным смещением ϑ).

распределения, близкие к единице, говорят о хорошей повторяемости фрагментов, близкие к нулю — о их непохожести.

В работах [2–5] детально проанализированы особенности МКА и приложения метода к сигналам различной природы. Показано, что хотя распределение $r(t,\vartheta)$ не позволяет так же хорошо, как при спектральном анализе, выяснить частотный состав сигнала, на основе МКА удается гораздо лучше отслеживать основную частоту и ее динамику в случае изменения последней со временем. В частности, МКА в гораздо большей степени свободен от такого недостатка спектрального анализа, как эффект просачивания частот. Другими словами, МКА ценой огрубления информации о форме колебаний предоставляет более адекватные данные о изменениях колебательного режима, имеющих место в случае нестационарности.

В данной работе представлены результаты, связанные с исследованием особенностей МКА в отношении обнаружения заметных скачков частоты сигнала и формирования оценок их величин. Очевидно, данная ситуация подразумевает сильный характер нестационарного поведения. В этой связи изложенный ниже материал подтверждает тезис об адекватности МКА для анализа нестационарных процессов. Кроме того, найденные процедуры обнаружения/оценивания сдвига частот оказываются весьма простыми и изящными, что делает их весьма перспективными с точки зрения практических применений. Говоря о процедурах во множественном числе, здесь подразумевается, что имеется альтернатива: либо производить оценки по самому распределению $r(t,\vartheta)$, либо по его Фурье преобразованию $F(t,\lambda)$. Поэтому оценки первого типа обсуждаются далее как оценивание во временной области, а оценки второго типа — как в частотной.

Обнаружение и оценка сдвига частоты во временной области

С целью выявления особенностей МКА в вопросах обнаружения сдвигов частот рассмотрим модельный сигнал $x(t) = A\cos\varphi(t)$, имеющий на фрагменте длительностью $2T$ скачок частоты $\Delta\nu$. Пусть для определенности этот скачок имеет место при $t=0$, так, что фаза сигнала есть $\varphi(t) = 2\pi\nu_-$ при $t<0$ и $\varphi(t) = 2\pi\nu_+$ при $t>0$ и, соответственно, $\nu_+ - \nu_- = \Delta\nu$. Зависимость фазы сигнала от времени и использованные обозначения представлены на рис. 2.

Рассмотрим в данный момент $t=0$ поведение корреляционного распределения (1). Подставляя в распределение $r(t,\vartheta)$ выражения для модель-

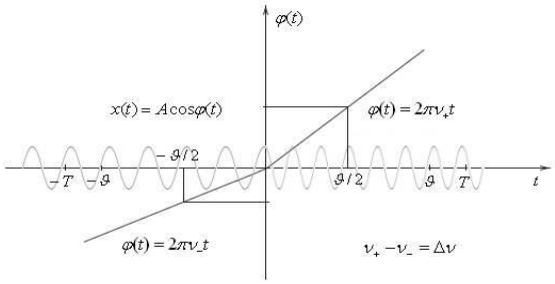


Рис. 2. Сигнал $x(t) = A\cos\varphi(t)$ и его фаза $\varphi(t)$, имеющая в нуле скачок производной $2\pi\Delta\nu$ (соответственно производные $2\pi\nu_-$ при $t<0$ и $2\pi\nu_+$ при $t>0$).

ного сигнала, получим:

$$r(0,\vartheta) = \frac{\int G^2\left(\frac{2t'}{\vartheta}\right)\cos\left(2\pi\nu_-\left[t'-\frac{\vartheta}{2}\right]\right)\cos\left(2\pi\nu_+\left[t'+\frac{\vartheta}{2}\right]\right)dt'}{\sqrt{\int G^2\left(\frac{2t'}{\vartheta}\right)\cos^2(\ldots)dt'}\sqrt{\int G^2\left(\frac{2t'}{\vartheta}\right)\cos^2(\ldots)dt'}}, \quad (2)$$

где аргументы косинусов в знаменателе повторяют соответствующие аргументы косинусов в числителе.

Если ввести функцию $H(\varsigma)$ — Фурье преобразование от $G^2(\xi)$:

$$H(\varsigma) = \int \exp(2\pi i\varsigma\xi)\, G^2(\xi)\, d\xi,$$

то после ряда преобразований (2) можно привести к виду:

$$r(0,\vartheta) = \frac{\cos(2\pi\nu\vartheta)H\left(\Delta\nu\frac{\vartheta}{2}\right) + \cos(\pi\Delta\nu\vartheta)H(\nu\vartheta)}{\sqrt{1+\cos(2\pi\nu_-\vartheta)H(\nu_-\vartheta)}\sqrt{1+\cos(2\pi\nu_+\vartheta)H(\nu_+\vartheta)}}, \quad (3)$$

где $\nu = \frac{1}{2}(\nu_- + \nu_+)$.

Поскольку $\left|H(\varsigma)\right| < \int G^2(\xi)d\xi = 1$, знаменатели в (3) можно разложить по $\cos(2\pi\nu_\pm\vartheta)H(\nu_\pm\vartheta)$:

$$\frac{1}{\sqrt{1+\cos(2\pi\nu_\pm\vartheta)\,H(\nu_\pm\vartheta)}} \approx 1-\frac{\cos(2\pi\nu_\pm\vartheta)\,H(\nu_\pm\vartheta)}{2}.$$

Применяя последнее разложение, получим окончательное представление (3) в следующем виде:

$$r(0,\vartheta) = \cos(2\pi\nu\vartheta)H\left(\Delta\nu\frac{\vartheta}{2}\right) + \varepsilon(\vartheta), \quad (4)$$

где

$$\varepsilon(\vartheta) = \cos(\pi\Delta\nu\vartheta)H(\nu\vartheta) + \cdots$$

— остаточный член. Многоточием обозначены слагаемые, содержащее по крайней мере одним из своих сомножителей либо $H(\nu\vartheta)$, либо $H(\nu_\pm\vartheta)$.

Представление (4) можно рассматривать как асимптотическое разложение при $\vartheta \gg \nu^{-1}$ в том важном случае, когда $\Delta\nu \ll \nu$ и, соответственно, $\nu_- \sim \nu_+ \sim \nu$. Действительно, поскольку эффективная ширина квадрата окна $G^2(\xi)$ порядка единицы, то тоже можно сказать и о ширине $H(\varsigma)$. Поэтому остаточный член $\varepsilon(\vartheta)$, имеющий порядок величины $H(\nu\vartheta)$, исчезающе мал при $\vartheta \gg \nu^{-1}$, а огибающая главного члена в разложении (4) вплоть до $\vartheta \sim (\Delta\nu)^{-1} \gg \nu^{-1}$ имеет порядок величины $H(0) = 1$. С учетом сделанного замечания, предполагая случай $\Delta\nu \ll \nu$, нетрудно проанализировать поведение $r(0,\vartheta)$ в области $\vartheta \gg \nu^{-1}$. Именно, ввиду малости $\varepsilon(\vartheta)$ в этой области, $r(0,\vartheta)$ описывается главным членом в (4), имеющим вид огибающей $H\left(\Delta\nu\frac{\vartheta}{2}\right)$ масштаба $2(\Delta\nu)^{-1}$ с гармоническим заполнением частоты ν. На рис. 3 представлен график распределения (4), вычисленного для случая $\nu_- = 95\,\text{Гц}$ и $\nu_+ = 105\,\text{Гц}$ (соответственно $\nu = 100\,\text{Гц}$ и $\Delta\nu = 10\,\text{Гц}$, $\Delta\nu/\nu = 0{,}1$), использованная форма окна имела вид $G(\xi) = 1-\xi^2$.

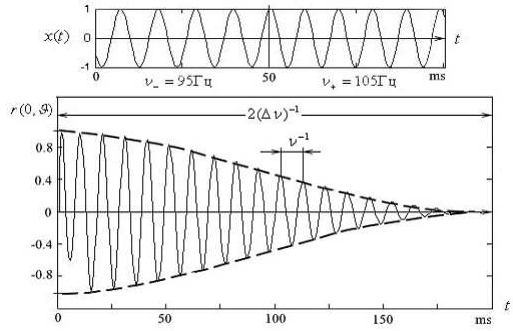


Рис. 3. Сигнал $x(t) = A\cos\varphi(t)$, фаза $\varphi(t)$ которого в момент $t = 50$мс имеет скачок производной $2\pi\Delta\nu$ ($\Delta\nu = 10\,\text{Гц}$) и соответствующее этому моменту корреляционное распределение $r(0,\vartheta)$. Пунктиром показана огибающая распределения.

Ввиду вышеизложенного теперь можно сформулировать правила обнаружения частотной модуляции и формирования оценок величины мгновенного изменения частоты $\Delta\nu$. Для обнаружения скачка частоты минимальной величины $\Delta\nu_0$ необходимо предварительно выбрать максимальный масштаб анализа $T = 2(\Delta\nu_0)^{-1}$ и по текущему времени t формировать для сигнала распределение $r(t,\vartheta)$, $0 \leqslant \vartheta \leqslant T$. При найденном $r(t,\vartheta)$ необходимо выделить его огибающую и определить, имеет ли она на интервале анализа $0 \leqslant \vartheta \leqslant T$ нули. Если обнаруживается по крайней мере один ноль, это можно рассматривать как обнаружение на данный момент скачка частоты (момент модуляции несущей частоты). При обнаружении скачка частоты можно оценить его величину в соответствии с $\Delta\nu = 2\varsigma_0/\vartheta_0$, где ϑ_0 — положение первого нуля огибающей $r(t,\vartheta)$ (в диапазоне $\nu^{-1} < \vartheta_0 < T$), а ς_0 — первый ноль (узел) функции $H(\varsigma)$: $H(\varsigma_0) = 0$.

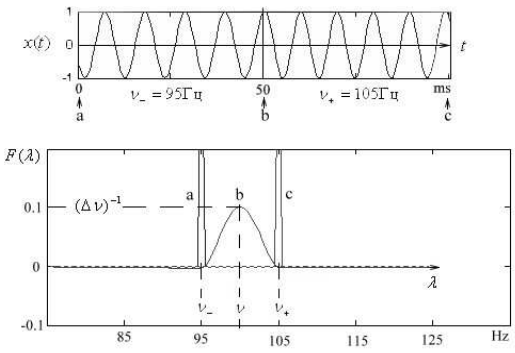


Рис. 4. Сигнал $x(t) = A\cos\varphi(t)$, фаза $\varphi(t)$ которого в момент $t = 50$ мс имеет скачок производной $2\pi\Delta\nu$ ($\Delta\nu = 10$ Гц) и соответствующие моментам времени $t = 0$ мс (a), $t = 50$ мс (b), $t = 100$ мс (c) спектры $F(t,\lambda)$.

На основе текущего распределения $r(t,\vartheta)$ можно также оценить среднюю частоту сигнала $\nu = \vartheta_{\cos}^{-1}$, оценив период колебаний $\vartheta_{\cos}$ гармонического заполнения. Отметим, что данная оценка может быть получена не только в моменты резкого изменения частоты. Если на данный момент не обнаружен скачок частот, то, с точностью до $\Delta\nu_0$, ν может интерпретироваться как текущая частота сигнала. В противно случае частоты до и после скачка оцениваются как $\nu_\pm = \nu \pm \Delta\nu/2$.

Обнаружение и оценка сдвига частоты в частотной области

Обнаружение частотной модуляции и оценивание параметров $\Delta\nu$ и ν можно осуществить также в частотной области, предоставляя процедуры, альтернативные описанным выше. Для того, чтобы сформулировать эти процедуры, рассмотрим обратное преобразование Фурье $F(0,\lambda)$ от распределения $r(t,\vartheta)$ (4):

$$F(0,\lambda) = \frac{1}{2\pi}\int \exp(-2\pi i\lambda\vartheta) r(0,\vartheta)\,d\vartheta = \quad (5)$$

$$= \frac{1}{\Delta\nu}\left[G^2\left(\frac{\lambda-\nu}{\frac{1}{2}\Delta\nu}\right) + G^2\left(\frac{\lambda+\nu}{\frac{1}{2}\Delta\nu}\right)\right] + \delta(\lambda),$$

где

$$\delta(\lambda) = \frac{1}{2\nu}\left[G^2\left(\frac{\lambda-\frac{1}{2}\Delta\nu}{\nu}\right) + G^2\left(\frac{\lambda+\frac{1}{2}\Delta\nu}{\nu}\right)\right] + \cdots$$

— остаточный член в спектре $F(0,\lambda)$.

Как следует из (5), для положительных частот λ преобразование Фурье от распределения $r(0,\vartheta)$ имеет главным членом одиночный пик высотой $\sim(\Delta\nu)^{-1}$ и шириной $\sim\Delta\nu$ на фоне остаточного члена $\delta(\lambda)$, имеющего порядок величины $\sim\nu^{-1}$ и плавно изменяющегося на масштабах $\sim\nu$. Вид спектров $F(0,\lambda)$ для фрагментов модельного сигнала $x(t) = A\cos\varphi(t)$, содержащих и не содержащих скачок частоты $\Delta\nu = 10$ Гц, $\Delta\nu/\nu = 0{,}1$ представлен на рис. 4.

В случае $\Delta\nu \ll \nu$ главный пик $F(0,\lambda)$ по высоте существенно превышает остаточный член $\delta(\lambda)$, поэтому хорошо обнаруживается. По положению главного пика, как следует из (5), оценивается средняя частота ν. Для оценки скачка частоты $\Delta\nu$ можно использовать как высоту пика $F(0,\nu)$, тогда $\Delta\nu = G^2(0)/F(0,\nu)$, так и ширину пика, например, по половинному уровню $d_{0,5}$, тогда $\Delta\nu = d_{0,5}/\xi_{0,5}$, где $\xi_{0,5}$ определяется из соотношения $G^2(\xi_{0,5}) = 0{,}5$.

Выводы

Приведенные выше результаты теоретического рассмотрения и численного моделирования позволяют заключить, что использование МКА для задач обнаружения и оценки частотных сдвигов в нестационарных процессах доставляет простые и изящные процедуры обнаружения/оценивания, практическая реализация которых не вызовет затруднений. В практически важном случае $\Delta\nu \ll \nu$ области корреляционного распределения $r(t,\vartheta)$ и спектрального $F(t,\lambda)$ распределения, используемые в процедурах обнаружения/оценивания имеют достаточно универсальный вид (зависящий только от окна $G(\xi)$) и удачно «разделяются» с областями их сложного с аналитической точки зрения поведения. Именно, во временной области используется асимптотика $\vartheta \gg \nu^{-1}$, $r(0,\vartheta) \approx \cos(2\pi\nu\vartheta)\,H\left(\frac{\Delta\nu\vartheta}{2}\right)$, а в частотной — значения $F(0,\lambda) \geqslant G^2(0)/2\Delta\nu$, при которых $F(0,\lambda) \approx G^2\left(\frac{\lambda-\nu}{\Delta\nu/2}\right)/\Delta\nu$.

Литература

[1] *Cohen L.* Time Frequency Analysis: Theory and Applications // New Jersey, Prentice Hall PTR 1995.

[2] *Анциперов В. Е.* Многомасштабный корреляционный анализ нестационарных, содержащих квазипериодические участки сигналов // Радиотехника и электроника. — 2008. — Т. 53, № 1. — С. 73–85.

[3] *Antciperov V. E., Morozov V. A., Obukhov Y. V.* Representation of Epileptic Discharge Dynamics in EEGs on the Basis of Multiscale Correlative Signal Dynamics Analysis // Pattern Recognition and Image Analysis. — 2008. — V. 18, № 2. — С. 342–346.

[4] *Анциперов В. Е., Обухов Ю. В.* Двумерное многомасштабное представление данных ЭЭГ записей эпилептических разрядов // Альманах клинической медицины. — 2008. — Т. XVII, № 1. — С. 154–157.

[5] *Анциперов В. Е., Обухов Ю. В.* ногомасштабный корреляционный анализ и основанное на нем представление сигналов медико-биологического происхождения // Доклады VIII международной научно-технической конференции ФРЭМЭ–2008 — С. 180–184.

Скелетная сегментация полутоновых линейчатых изображений*

Аргунов Д А., Местецкий Л. М.
dmiarg@gmail.com, l.mest@ru.net
Москва, МГУ имени М. В. Ломоносова

В данной работе исследуется новый метод сегментации линейчатых изображений, основанный на тринаризации исходного изображения и совместной обработке внутреннего и внешнего скелетов полученного тринарного изображения. Подходы тринаризации и совместного анализа скелетов описываются в этой работе впервые. В статье описаны численные эксперименты, которые показывают перспективность применения этих подходов в системах биометрической идентификации и машинного зрения.

В настоящее время активно расширяется область применения биометрических систем. Один из главных способов биометрической идентификации — идентификация по отпечаткам пальцев. Так, дактилоскопическая информация станет обязательным параметром личности в биометрических паспортах. Развитие дактилоскопических систем привело к тому, что разница между лучшими решениями проявляется только на сильно зашумленных образцах отпечатков. Таким образом, актуальна задача разработки алгоритмов обработки линейчатых изображений (в частности, изображений отпечатков пальцев), ориентированных на сильно зашумленные изображения.

Рис. 1. Пример зашумленного изображения отпечатка пальца.

Постановка задачи

Задача, сходная с выделением линий папиллярного узора отпечатка пальца, возникает при обработке текста, когда надо выделить линии строк текста и сориентировать их параллельно. Поэтому было решено выделить общий класс изображений, отвечающий некоторым свойствам, и разрабатывать алгоритм, исходя из этих свойств.

*Работа выполнена при финансовой поддержке РФФИ, грант № 08-01-00670.

Определение 1. *Линейчатым изображением будем называть серое полутоновое изображение, удовлетворяющее следующим критериям:*

1) *линейчатое изображение состоит из двух непересекающихся множеств точек: линий и промежутков между ними;*
2) *линии и промежутки локально параллельны;*
3) *длина линий и промежутков значительно больше ширины;*
4) *выборочные средние яркости пикселей линий и промежутков значительно различаются;*
5) *ширина линий и промежутков варьируется незначительно, причем совпадения ширины линий и промежутков не требуется.*

В случае изображения отпечатка пальца линии и промежутки — это линии папиллярного узора и промежутки между ними соответственно. В случае изображения текста это строки текста и межстрочные интервалы.

В данной статье описывается подход, основанный на скелетизации изображения. Скелетное описание линий и промежутков представляется удобным для вычисления признаков, таких как множества точек разветвлений и терминальных точек линий для проведения дактилоскопической идентификации.

Определим теперь задачу сегментации линейчатого изображения. Входными данными задачи является серое полутоновое линейчатое изображение I. Выходными данными задачи является описание двух скелетов изображения S_{line} и S_{span}, которые соответствуют линиям и промежуткам исходного изображения.

Методы, описанные в данной работе, базируются на двух основных идеях:

— проведение сегментации с использованием двух порогов (тринаризация);
— совокупная обработка внутреннего и внешнего скелетов изображения.

Сложность бинаризации линейчатых изображений

Для получения скелета изображения необходимо провести бинаризацию. Однако прямая пороговая бинаризация не всегда позволяет получить

корректное бинарное изображение, и скелет может оказаться неадекватным исходному изображению. На рис. 2 приведен пример изображения отпечатка пальца, для которого корректная пороговая бинаризация невозможна.

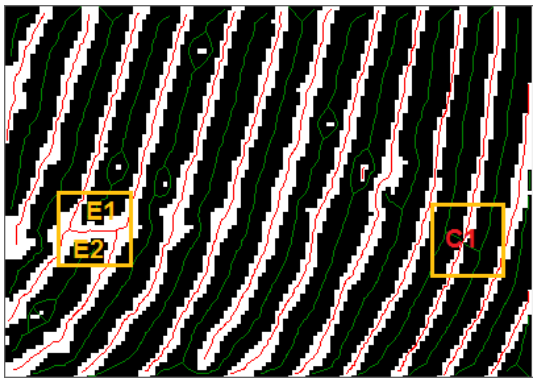


Рис. 2. Пример бинаризации линейчатого изображения и полученных скелетов.

Скелеты бинаризованного изображения на рисунке содержат две некорректности:

1) ветви скелета E_1 и E_2 разделены, хотя соответствуют одной линии изображения — такую некорректность сегментации будем называть *разрывом*;
2) скелет содержит ветвь C_1, которая не соответствует ни одной линии исходного изображения и связывает две корректные ветви — такую некорректность сегментации будем называть *перемычкой*.

Очевидно, что при повышении порога в скелете останется ветвь C_1, а при понижении останется перемычка между ветвями E_1 и E_2. Таким образом, ни при каком пороге бинаризации не будет получено корректного бинарного изображения.

Основные идеи предлагаемого метода

Итак, противоречие, возникающее при бинаризации линейчатого изображения, заключается в том, что порог надо повысить для избавления от некорректностей первого рода (разрывов) и понизить для избавления от некорректностей второго рода (перемычек). Тогда разумно рассматривать два порога, позволяющих избавиться от всех некорректностей одного класса в бинаризации изображения I. Первый порог $L^I_{\text{con}} \in \mathbb{Z}$ — это максимальный порог, исключающий появление перемычек, а второй порог $L^I_{\text{int}} \in \mathbb{Z}$ — это минимальный порог, исключающий появление разрывов. Пороги L^I_{con} и L^I_{int} будем называть *порогами ошибок бинаризации*.

Отметим, что существование разрыва в скелете бинаризованного линейчатого изображения означает также существование перемычки в его внешнем скелете. Это свойство можно рассматривать как следствие более общего свойства линейчатых изображений — инвариантности критериев линейчатого изображения относительно преобразования инвертирования яркости.

Утверждение 1. *Пусть I — линейчатое изображение размером $m \times n$, I_{neg} — изображение, полученное из изображения I преобразованием яркостей*

$$I_{neg}(i,j) = 255 - I(i,j),\ i = 0,\dots,m,\ j = 0,\dots,n.$$

Тогда I_{neg} — линейчатое изображение.

Ясно, что точки линий на инвертированном линейчатом изображении становятся точками промежутков и наоборот. Очевидна связь порогов ошибок бинаризации инвертированного и исходного изображений.

Утверждение 2.

$$L^{I_{neg}}_{con} = L^I_{int},$$

$$L^{I_{neg}}_{int} = L^I_{con}.$$

Таким образом, задачи сегментации линейчатого изображения I и инвертированного изображения I_{neg} тесно связаны.

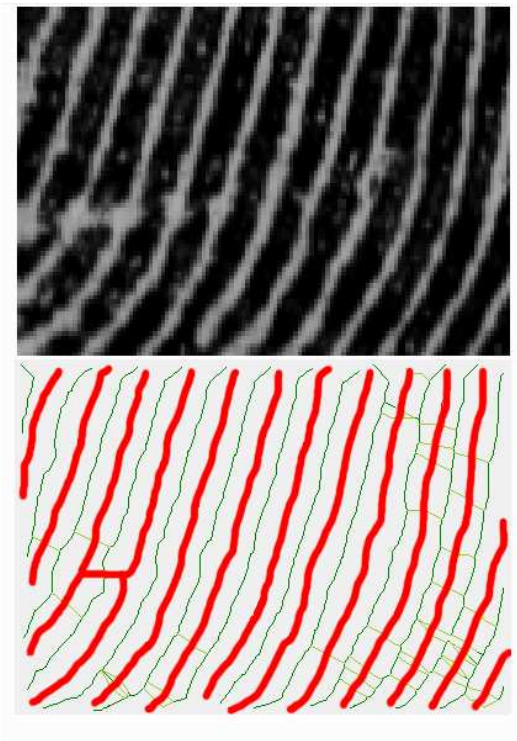


Рис. 3. Пример исходного линейчатого изображения и двух пересекающихся скелетов S и S_{neg}.

Предлагается получить искомые скелеты линий и промежутков линейчатого изображения I

на базе скелетов бинарных изображений B и B_{neg}, где B — это бинаризация изображения I по порогу L^I_{int}, и B_{neg} — это бинаризация изображения I_{neg} по порогу $L^{I_{\text{neg}}}_{\text{int}} = L^I_{\text{con}}$. Обозначим эти скелеты S и S_{neg} соответственно. Из определения этих порогов следует, что оба скелета содержат избыточные ветви-перемычки. Отсечение этих ветвей проводится на основе совместного анализа скелетов изображений B и B_{neg}.

Так как $L^I_{\text{int}} > L^I_{\text{con}}$, объекты на изображениях B и B_{neg} имеют общие точки. Поэтому скелеты этих изображений S и S_{neg} в общем случае пересекаются, как показано на рис. 3.

Из рисунка видно, что из двух пересекающихся ветвей скелетов одна ветвь — это перемычка, а вторая — корректная ветвь, соответствующая линиям исходного изображения. Также все перемычки неизбежно пересекаются с одной из ветвей другого скелета. Таким образом, точки пересечений скелетов указывают на перемычки. Поэтому предлагается принимать решение об отсечении ветвей-перемычек на основе анализа пересечений скелетов.

Тринарные изображения

Итак, в основу предложенного метода положено проведение сегментации исходного линейчатого изображения по двум порогам L^I_{int} и L^I_{con}. Пороговая бинаризация по одному порогу позволяет разделить множество точек изображения на два непересекающихся подмножества. Аналогично два порога позволяют разделить множество точек изображения на три подмножества, то есть провести *тринаризацию* изображения.

Введем необходимые определения.

Определение 2. *Тринарным изображением будем называть такое изображение, где каждый пиксель может принимать 3 разных значения:* $0, 1, 2$.

Пусть I — серое полутоновое изображение, $L_1, L_2 \in \mathbb{Z}$ — пороги тринаризации.

Определение 3. *Тринаризацией изображения I будем называть тринарное изображение I_3, полученное из изображения I при помощи преобразования:*

$$I_3(i,j) = \begin{cases} 0, & \text{если } I(i,j) < L_1; \\ 2, & \text{если } I(i,j) > L_2; \\ 1, & \text{иначе.} \end{cases}$$

Приведем пример тринаризации линейчатого изображения (рис. 4). На нем черным показаны пиксели со значением 0 (множество P_0), серым — со значением 1 (множество P_1), белым — со значением 2 (множество P_2).

На данном рисунке выбор порогов тринаризации был проведен в соответствии с определением порогов L^I_{int} и L^I_{con}. Как видно, множества P_0 и P_2

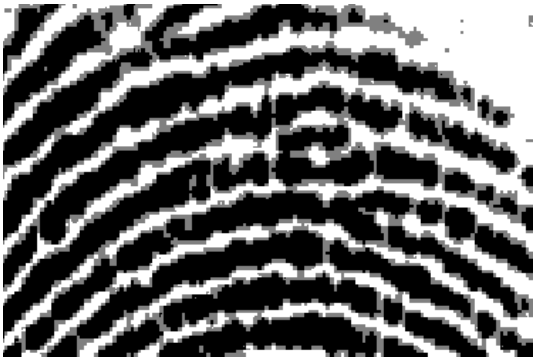


Рис. 4. Пример тринаризации линейчатого изображения.

построены так, что не содержат перемычек, а множества $P_0 \cap P_1$ и $P_1 \cap P_2$ не содержат разрывов. Множества $P_0 \cap P_1$ и $P_1 \cap P_2$ совпадают со множествами точек объектов описанных выше бинарных изображений B и B_{neg}. Построенные по ним скелеты S и S_{neg} (рис. 3) будем называть *внутренним и внешним скелетами тринарного изображения I*.

Скелеты S и S_{neg} — это первое приближение скелетов S_{line} и S_{span}, которые являются результатом работы алгоритма. Дальнейшая обработка скелетов заключается только в удалении некорректных ветвей-перемычек, поэтому они должны максимально корректно описывать линии и промежутки исходного изображения. Таким образом, пороги тринаризации должны быть выбраны таким образом, чтобы множество P_1 оказалось достаточно большим, включая все точки разрывов и перемычек, но достаточно маленьким, чтобы множества P_0 и P_2 были близки к искомым множествам линий и промежутков.

Вычисление порогов тринаризации — это нетривиальная задача, и разработка алгоритмов поиска оптимальных порогов тринаризации не входила в цели данной работы. Для всех описанных в этой статье экспериментов пороги тринаризации были выбраны экспертно.

Анализ скелетов тринарного изображения

Итак, обработка скелетов S и S_{neg} заключается в анализе точек их пересечения и удалении одной из пересекающихся ветвей скелетов. Получим правило отсечения ветвей. Анализ пересекающихся скелетов различных тринаризаций линейчатых изображений показал, что ветви перемычек, сориентированные в направлении, близком к перпендикулярному к направлению линий исходного линейчатого изображения. Это позволяет сделать заключение о длине ветви-перемычки. Рассмотрим схематичное изображение пересечения на рис. 5.

Ветвь-перемычка пересекает промежуток между линиями линейчатого изображения и две линии от границы до скелета этих линий. Так как ске-

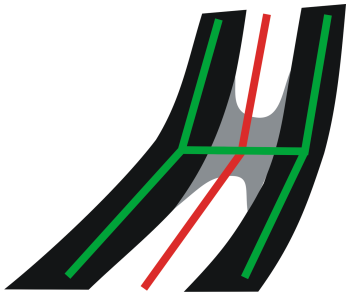


Рис. 5. Схематичное изображение пересечения скелетов.

лет представляет собой серединную линию объекта, то расстояние от границы линии до скелета равно половине ширины линии. Таким образом, длина перемычки скелета равна сумме ширины линии и промежутка. По свойствам линейчатого изображения длина линии и промежутка много больше ширины, поэтому можно утверждать, что ветвь скелета, описывающая перемычку, короче ветви, описывающей линию или промежуток.

Утверждение 3. *В случае пересечения двух ветвей скелетов S и S_{neg} более короткая из двух ветвей описывает ложную перемычку, а более длинная — корректно описывает линию или промежуток исходного изображения.*

Таким образом, отсечению подлежит более короткая из двух пересекающихся ветвей.

Вычислительные эксперименты

Для оценки качества работы алгоритма был проведен ряд вычислительных экспериментов. В качестве оценки качества работы алгоритма было принято число ложных разрывов и перемычек в результирующих скелетах S_{line} и S_{span}. Этот критерий показывает, насколько успешно предложенный алгоритм устраняет характерные для линейчатых изображений ошибки сегментации, проиллюстрированные рис. 2.

Для того, чтобы оценить новый алгоритм, его необходимо сравнить с некоторым известным алгоритмом, решающим аналогичную задачу. Представленный алгоритм работает с линейчатыми изображениями, и нет известных алгоритмов, работающих с теми же входными данными. Поэтому был выбран более общий алгоритм, составленный как композиция пороговой бинаризации и скелетизации.

В таблице приведено количество ложных перемычек и разрывов для скелетов различных линейчатых изображений, обработанных двумя методами.

Тринаризация	Бинаризация	Отношение
1	5	0,2
1	3	0,33
0	1	0
19	35	0,54
12	20	0,6
25	70	0,36

Как видно из таблицы, предложенный алгоритм позволяет значительно уменьшить количество ошибок. В среднем число ложных перемычек и разрывов сократилось в 3 раза по сравнению с алгоритмом «бинаризация+скелетизация».

Выводы

В работе описан метод сегментации, ориентированный на класс линейчатых изображений. Он основан на двух инновационных идеях — тринаризации и совместной обработке внутреннего и внешнего скелетов тринарного изображения. В ходе вычислительных экспериментов данный метод показал значительно лучшие результаты, чем метод, основанный на бинаризации и скелетизации.

Разработанный метод представляется перспективным для применения в обработке изображений отпечатков пальцев и изображений текста.

Литература

[1] *Местецкий Л. М.* Непрерывная морфология бинарных изоражений: фигуры, скелеты, циркуляры. — М.: Физматлит, 2009.

[2] *Гонзалес Р., Вудс Р.* Цифровая обработка изображений. — М.: Техносфера, 2006.

[3] *Yanushkevich S. N., Wang P. S. P.* Image pattern recognition: synthesis and analysis in biometrics. // World Scientific. — 2007.

[4] *Яне Б.* Цифровая обработка изображений. — М.: Техносфера, 2006.

[5] `http://www.neurotechnology.com`

[6] `http://www.sonda-tech.com/ru`

Метод сравнения формы ладоней при наличии артефактов*

Бакина И. Г., Местецкий Л. М.
irina_msu@mail.ru, L.Mest@ru.net
МГУ им. М. В. Ломоносова

В работе рассматривается метод сравнения ладоней для задачи распознавания личности по форме руки. Идея предлагаемого подхода заключается в «подгонке» эталонного образца ладони под тестовый экземпляр, который может обладать рядом артефактов (длинные ногти, «склеенные» пальцы, длинный рукав). «Подгонка» заключается в применении аффинных преобразований как ко всей ладони, так и отдельно к каждому пальцу. Сравнение предъявляемой ладони и модифицированного эталонного экземпляра осуществляется на основе оценки симметрической разности сопоставленных силуэтов ладоней. Области большого пальца и запястья «отсекаются» и не учитываются при сравнении ладоней.

Практически во всех существующих методах распознавания личности, основанных на анализе формы руки [1, 2, 4], присутствует требование предъявления ладони с хорошо разделенными пальцами (за исключением [3], где, наоборот, требуется предъявлять ладонь с плотно прижатыми пальцами). Данное требование является существенным для таких систем, т. к. только при его соблюдении возможно корректное определение параметров ладони, таких как длина и ширина пальцев, ширина ладони и т. д. Использование платформ со штырьками-разделителями автоматически решает эту проблему. Однако требуемое позиционирование ладони оказывается для многих людей затруднительным. Наблюдения показывают, что существуют люди, которые, в случае отсутствия специальных платформ, зачастую склонны предъявлять ладонь со «склееными» пальцами, что влечёт их неправильное распознавание системой.

При предъявлении ладони для распознавания очень часто возникают и другие артефакты, связанные с наличием длинных ногтей, колец, браслетов и длинных рукавов одежды. В работе [2] авторы решают проблему присутствия колец, а также предлагают метод восстановления/отсечения запястья. Подходы к распознаванию в случае длинных ногтей не рассматриваются.

Постановка задачи

Целью данной работы является разработка метода сравнения ладоней, который бы позволял сопоставлять ладони даже в случае следующих артефактов (рис. 1):

1) присутствие «склеенных» пальцев;
2) наличие длинных ногтей;
3) закрытое запястье.

Исходными данными для сравнения являются бинарные изображения ладоней, получаемые с помощью web-камеры (предполагается, что ладонь представлена черными пикселями на белом фоне). Существенно, что камера находится на фиксированном расстоянии над платформой, на которой человек позиционирует свою ладонь. Такое требование позволяет не учитывать масштабирование ладони, как в случае, если бы разрешалось варьирование расстояния между ладонью и камерой.

В предлагаемом подходе подразумевается, что имеется исходная база эталонных изображений ладоней, которые лишены перечисленных выше артефактов. Данная база «хороших» ладоней собирается при регистрации пользователей системы и, быть может, проходит некоторую предварительную обработку экспертами.

Тестовая ладонь может быть сопоставлена только с эталонными изображениями. В этом случае эталонная модель ладони «подгоняется» под тестовую. В сравнении ладоней участвуют только их силуэты.

Модель ладони

В предлагаемом подходе используется представление ладони в виде гибкого объекта. Эта модель была предложена в [1] и представляется удобной для проведения таких преобразований ладони, как вращение ее отдельных частей.

Сначала для исходного бинарного изображения ладони строятся осевой и циркулярный графы, как показано на рис. 2. Далее необходимо указать группу применяемых трансформаций. В [1] в качестве таких трансформаций рассматривались вращения частей циркулярного графа (фактически, пальцев

Рис. 1. Возможные артефакты: 1) «склеенные» пальцы; 2) длинные ногти; 3) закрытое запястье.

*Работа выполнена при финансовой поддержке РФФИ, проекты №08-01-00670, №08-07-00305 и №08-07-00270.

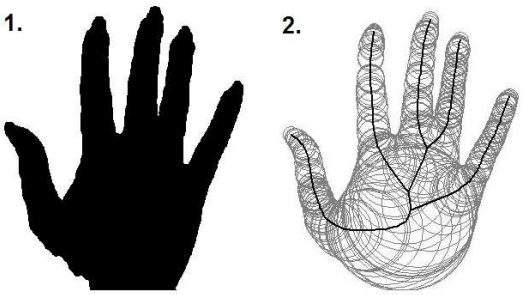


Рис. 2. 1) Бинарное изображение ладони; 2) циркулярный и осевой графы.

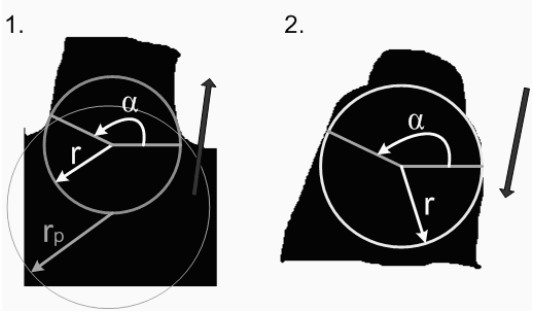


Рис. 3. Определение 1) основания пальца; 2) кончика пальца.

руки) вокруг точек изгиба. При этом точки изгиба определялись как точки пересечения осевого графа с большим кругом ладони. Однако в случае наличия артефактов, связанных с закрытым запястьем или склеенными пальцами, большой круг ладони не может быть определён корректно, что, следовательно, ведёт к неправильному выделению точек изгиба. Поэтому данный этап построения модели ладони должен быть модифицирован. В работе определение точек изгиба предлагается проводить на основе информации о положении основания и кончика для каждого из пальцев руки.

Выделение пальцев. Рассмотрим способ выделения основания и кончика для одного из пальцев ладони. Отметим, что в предлагаемом методе сначала определяется местонахождение основания пальца; и только затем, на основе полученной информации, ищется его кончик.

Рассматриваем одну из ветвей осевого графа ладони, относящуюся к анализируемому пальцу, и перебираем последовательно все соседние пары кругов циркулярного графа от точки ветвления до концевой вершины. Для каждой пары (предыдущий и текущий круги) проверяется выполнение следующих условий:

$$\begin{cases} r \leqslant r_0; \\ \left[\begin{array}{l} r - r_p \leqslant 0; \\ \alpha \geqslant \alpha_0. \end{array} \right. \end{cases}$$

Здесь α — угол между двумя радиусами, проведёнными из центра круга в точки касания этого круга с силуэтом ладони; r — радиус текущего круга, а r_p — радиус предыдущего круга в рассматриваемой паре; α_0 и r_0 — пороговые константы, определённые в результате проведения ряда экспериментов. Вершина осевого графа, соответствующая кругу из пары, удовлетворяющей всем указанным условиям, объявляется *основанием* пальца.

Далее похожим образом определяется положение *кончика* пальца. Просматриваются последовательно все круги циркулярного графа из концевой вершины в точку ветвления до тех пор, пока не найдётся круг, удовлетворяющий следующим двум условиям:

$$\begin{cases} \alpha \geqslant \alpha_0; \\ r \geqslant R(r_{\text{root}}). \end{cases}$$

В этом случае α — снова угол между двумя радиусами, проведенными из центра круга в точки касания этого круга с силуэтом ладони; r — радиус рассматриваемого круга, а r_{root} — радиус круга в основании пальца; α_0 — пороговая константа; $R(r_{\text{root}})$ — некоторая функция, зависящая от радиуса круга в основании пальца (в работе рассматривалась $R(r_{\text{root}}) = 0{,}65 r_{\text{root}}$).

Иллюстрация поиска основания и кончика пальца приведена на рис. 3.

Определение точек изгиба. При горизонтальном положении ладони на поверхности человек имеет возможность двигать пальцами руки. Для каждого пальца наиболее подвижным является основание основной фаланги (за исключением большого пальца, для которого подвижным является также место сочленения начальной и средней фаланги). Отдельным этапом процесса построения модели ладони является этап определения местоположения этого основания, называемого в дальнейшем *точкой изгиба* или *точкой вращения пальца*. Нахождение этой точки для каждого пальца ладони производится только для эталонного образца и необходимо для проведения последующей его «подгонки» под тестовую ладонь.

Точки вращения пальцев могут быть заранее указаны экспертом при подготовке базы эталонных образцов ладоней. Но они могут быть также определены без участия специалиста.

Выше указывалось, почему для этих целей не подходит метод, изложенный в [1]. Идея предлагаемого способа определения точки вращения пальца сходна с [2]. Через выделенные точку основания и кончик пальца проводится прямая, называемая *серединной осью* пальца. Точка изгиба определяется как точка, лежащая на серединной оси и отстоящая от основания пальца на 30% длины этого пальца. (длина пальца — евклидово расстояние между

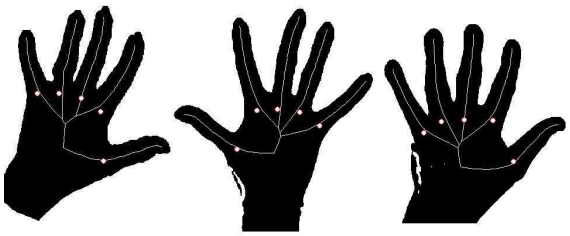


Рис. 4. Результат нахождения точек вращения пальцев (белые жирные точки) для трёх различных ладоней.

точкой основания и кончиком пальца). Пример нахождения точек изгиба представлен на рис. 4.

Сравнение ладоней

Этап сравнения эталонной и тестовой моделей ладоней включает в себя следующие шаги:

1) приведение моделей в стандартное положение;
2) «шевеление» пальцев эталонной модели с целью совмещения с тестовым образцом;
3) выделение области для сравнения моделей;
4) определение степени похожести ладоней.

Приведение в стандартное положение. Серединная ось среднего пальца руки, проведённая из кончика к основанию пальца, задаёт ось ординат локальной системы координат ладони. Локальная ось абсцисс определяется как ось, перпендикулярная локальной оси ординат и направленная в сторону мизинца (рис. 5). Стандартным положением ладони называется положение, при котором локальная система координат совпадает с глобальной. Приведение в стандартное положение осуществляется путём аффинного преобразования координат.

Совмещение эталонной и тестовой моделей. Понятно, что при предъявлении ладони человеку затруднительно расположить её таким же образом, как и при регистрации этой ладони в базе. Основная проблема заключается в том, что угол между пальцами меняется. При этом некоторые из пальцев могут оказаться «склеенными».

С целью совмещения эталонной и тестовой моделей ладоней для их дальнейшего сравнения осуществляется поворот пальцев эталонного образца вокруг точки изгиба. Углы поворота определяются следующим образом.

Рассматривается сравниваемая тестовая модель ладони, и вычисляются углы между серединными осями указательного пальца и остальных пальцев ладони (на рис. 5 это углы α_1, α_2, α_3 и α_4). Далее проводится вращение пальцев эталона в положение, при котором соответствующие углы совпадают с полученными для тестового образца. Фактически все преобразования осуществляются над осевым и циркулярным графами. После того, как исходная эталонная модель была совмещена с тестовой, для

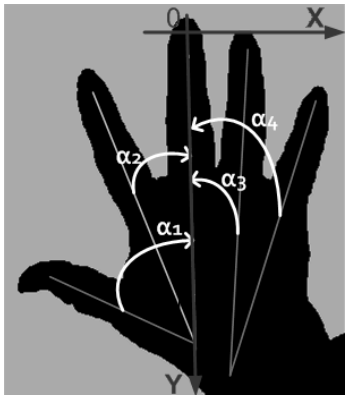


Рис. 5. Локальная система координат ладони и определение углов между пальцами.

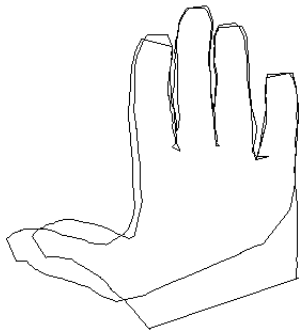


Рис. 6. Результат совмещения двух ладоней.

обоих циркулярных графов строится огибающая (семейства кругов, из которых этот граф составлен). Данная огибающая рассматривается как контур ладони.

Поскольку человек не всегда предъявляет системе одну и ту же часть ладони для распознавания, необходимо уметь выделять область, которая может быть использована для сравнения практически во всех случаях. Так, на рис. 6 видно, что ладони хорошо совмещаются в области четырёх пальцев, но значительное расхождение наблюдается в зоне запястья. Большой палец эталонной руки также оказался плохо «подогнанным», т. к. возможные движения этого пальца в горизонтальной плоскости не ограничиваются вращением в точке изгиба. Более того, кожа руки у основания этого пальца значительно деформируется при его движении.

Выделение области сравнения ладоней. В силу того, что области запястья и большого пальца руки не могут быть совмещены достаточно хорошо, они исключаются из дальнейшего рассмотрения, и сравнение ладоней в этой части не происходит.

Через точки вращения указательного пальца и мизинца эталонной модели проводится прямая линия. Область ладони, лежащая ниже этой прямой (содержащая запястье), отсекается. Отсечение

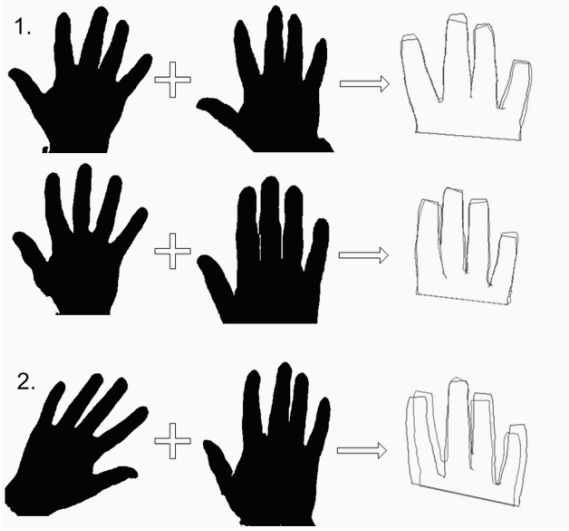


Рис. 7. Совмещение ладоней с выделением области сравнения для 1) одного человека; 2) разных людей. В левой колонке представлены эталонные ладони, в средней — тестовые.

проводится как для эталонного, так и для тестового образца.

Результат совмещения ладоней и выделения области сравнения представлен на рис. 7.

Определение похожести ладоней. В качестве меры близости ладоней в работе рассматривается площадь симметрической разности их совмещённых силуэтов.

Эксперименты по анализу отличительной способности данного признака были проведены для группы, составленной из 15 человек. Снятые изображения ладоней были разделены на две группы — эталонные и тестовые образцы. К эталонам были отнесены изображения ладоней, лишенные артефактов; тогда как тестовые ладони могли их содержать. Далее каждый тестовый образец сопоставлялся последовательно со всеми эталонными. Если сравниваемая пара содержала ладони, принадлежащие одному человеку, то она была отнесена к первому классу. Пара, представленная ладонями разных людей, — ко второму. В результате первый класс был представлен примерно 300 объектами-парами, а второй — 600. На рис. 8 представлены полученные графики плотности распределения расстояний для каждого из указанных классов.

Заключение и выводы

Представленный подход к сравнению ладоней позволяет учитывать и обрабатывать артефакты, часто возникающие в системах распознавания, основанных на геометрии и/или форме руки. Полученные графики плотности распределения внутриклассовых расстояний свидетельствуют о том, что предложенный признак может быть использован для дальнейшего решения задачи распознавания личности. Однако точность распознавания может быть недостаточно высокой, т. к. под графиками есть существенная общая область. Более аккуратное разделение классов возможно в случае уменьшения влияния следующих факторов. Во-первых, необходимо более точно выделять кончик пальца, «обрезая» ногти на изображении ладони. Второй фактор обусловлен тем, что при «слипании» пальцев деформируется кожа между ними. Моделирование «склеенных» пальцев на эталонном образце ладони приводит к тому, что пальцы начинают частично перекрывать друг друга. Это, в свою очередь, ведёт к некорректному построению огибающей циркулярного графа ладони.

В рамках дальнейшей работы планируется решение указанных проблем с построением системы верификации/идентификации личности, использующей рассмотренный подход к сравнению ладоней в комбинации с другими признаками.

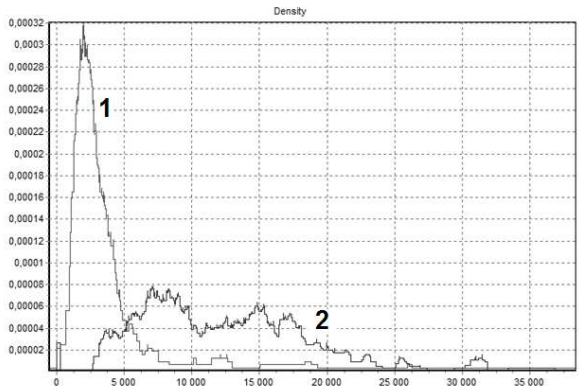


Рис. 8. Плотность распределения внутриклассовых расстояний для объектов-пар из 1) первого (пары ладоней одного человека) и 2) второго (ладони в парах от разных людей) классов.

Литература

[1] *Местецкий Л. М.* Непрерывная морфология бинарных изображений: фигуры, скелеты, циркуляры. — Москва: ФИЗМАТЛИТ, 2009. — 288 с.

[2] *Konukoglu E., Yoruk E., Darbon J., Sankur B.* Shape-Based Hand Recognition // Image Processing, IEEE, Volume 15, Issue 7. — P. 1803–1815.

[3] *Rahman A., Azad S., Anwar F.* An Efficient Technique for Human Verification Using Finger Stripes Geometry // International Journal of Soft Computing 2 (3), 2007. — P. 445–449.

[4] *Su C.-L.* Index, Middle, and Ring Finger Extraction and Identification by Index, Middle, and Ring Finger Outlines // Int'l Conf. on Computer Vision Theory and Applications (VISAPP), Volume 1, 2008. — P. 518–520.

Метод идентификации групповых телеметрических сигналов на основе частотно-рангового распределения

Балтрашевич В. Э., Васильев А. В., Жукова Н. А., Соколов И. С.
igor.s.sokolov@gmail.com
Санкт-Петербург, ОАО «НИЦ СПб ЭТУ»

В докладе рассматриваются вопросы формирования описания групповых телеметрических сигналов на основе частотно-рангового распределения и методы их идентификации с использованием алгоритмов кластерного анализа. Приводятся результаты экспериментальных исследований на модельных данных.

В самом широком смысле понятие телеметрии связано с формированием организованного потока измерений от удалённых объектов, его передачей и обработкой [1]. Телеметрическая информация может использоваться, например, для контроля за состоянием удалённого объекта или для получения данных о каких-либо физических процессах во время проведения экспериментов. Передача телеметрической информации осуществляется в виде единого потока данных (группового телеметрического сигнала, ГТС) по единому каналу связи. Извлечение информационных параметров из ГТС производится с использованием формуляра, содержащего описание структуры ГТС. Наличие и корректность формуляра определяет возможность проведения анализа и влияет на качество получаемых результатов. В случае отсутствия формуляра (его потеря, задержка при передаче), несоответствия формуляра передаваемой информации, возникновения ошибок при формировании формуляра провести оценку результатов испытаний сложного технического объекта становится практически невозможно.

В связи с крайне узкой спецификой описанной проблемы все решения, которые существуют на сегодняшний день, заключаются в ручном подборе параметров описания структуры. В рамках статьи предлагается подход к определению структуры ГТС на основании анализа его соответствия ранее накопленным ГТС с известной структурой.

Особенности формирования ГТС

В современных информационно-телеметрических системах в большинстве случаев телеметрируемые параметры передаются в дискретно-квантованном виде. Информация от датчиков системы упаковывается с использованием временного разделения данных в единый групповой телеметрический сигнал, который можно рассматривать как поток информационных слов постоянной длины.

Пусть проводятся испытания сложного технического объекта, на котором установлены датчики Д1-Д5 (рис. 1). Каждое измерение датчиками параметра – это информационное слово фиксированной длины. Последовательный опрос датчиков главным коммутатором порождает группу

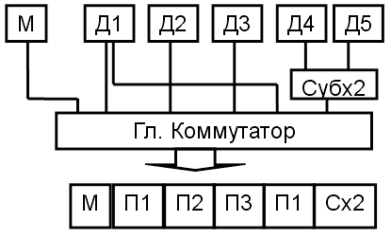


Рис. 1. Схематическое изображения коммутации показаний от датчиков в сложном техническом объекте.

информационных слов, называемых кадром. Каждый кадр начинается с постоянного фиксированного значения, называемого синхрогруппой или маркером (М). Длина кадра и длина информационного слова постоянны на всем ГТС.

Для передачи параметров, в спектре которых присутствуют частоты вдвое большие чем частота опроса главного коммутатора, применяется суперкоммутация. *Суперкоммутация* — это передача данных с частотой кратной частоте главного коммутатора. При суперкоммутации за один кадр передается несколько показаний одного параметра.

С целью минимизации информационного потока для передачи медленно изменяющихся параметров применяется субкоммутация. *Субкоммутация* — это передача значений параметра с частотой субкратной частоте главного коммутатора. При субкоммутации одно измерение параметра передается один раз в несколько кадров.

В рассматриваемом примере (рис. 1) Д4 и Д5 подключены к субкоммутатору и опрашиваются на половинной частоте главного коммутатора, а датчик Д1 на удвоенной.

Под *структурой ГТС* будем понимать описание набора передаваемых параметров и схему их расположения в результирующем телеметрическом потоке. Таким образом, ГТС обладают одной и той же структурой, если в них передается одинаковый набор параметров и они имеют одинаковые суб- и суперкоммутации.

Структура кадра представлена в нижней части рис. 1, где М — это значение маркера, П1-3 — показания параметров 1-3, Сх2 — чередующиеся через 1 кадр показания параметров 4 и 5.

В рамках данной работы для идентификации ГТС предлагается использовать частотно- ранговое представление ГТС и методы кластерного анализа. В качестве критерия качества описываемого подхода будет использоваться вероятность верной идентификации ГТС.

Частотно-ранговая зависимость и её аппроксимация

В 40-х годах 20 века американский лингвист Джордж Ципфе (George Kingsley Zipf) на основании проведённых исследований в области лингвистики предложил эмпирический закон распределения частоты слов естественного языка.

При анализе фрагмента текста формируется словарь — множество всех встречаемых слов, при этом для каждого i-го слова определяется частота его встречаемости в тексте f_i. На основании упорядоченной по убыванию последовательности частот определяется ранг слова r как порядковый номер слова в последовательности.

Закон Ципфа формулируется следующим образом: *произведение ранга некоторого слова и частоты его встречаемости в тексте является величиной постоянной, имеющей примерно одинаковое значение для любого слова из словаря рассматриваемого текста:* $f_r r = C$, где f_r — частота встречаемости в тексте слова с рангом r, C — эмпирическая постоянная величина.

В более поздний период известный математик Бенуа Мандельброт (Benoît B. Mandelbrot) сформулировал теоретическое обоснование закона Ципфа, основанное на представлении слов языка как сообщений, имеющих определенную «стоимость». В соответствии с законом о минимальной «стоимости» сообщений Мандельброт пришел к аналогичной закону Ципфа зависимости:

$$f_r r^{\gamma} = C, \tag{1}$$

где γ — величина (близкая к единице), значение которой определяется в зависимости от свойств анализируемого текста; C — константа, определяемая объёмом выборки.

Дальнейшие исследования показали, что закону Ципфа-Мандельброта удовлетворяют не только слова из текстов на естественном языке, но и практически все объекты современного информационного пространства.

Аппроксимация частотно-рангового распределения в соответствии с законом (1) осуществляется с помощью функциональной зависимости:

$$\Phi(r) = Cr^{-\gamma \exp(d \cdot r)}, \tag{2}$$

где $\Phi(r)$ — относительная частота слова с рангом r; d — параметр увеличения коэффициента частотно-рангового соотношения. Подбор параметров C, γ, d сводится к минимизации критерия суммы квадратов расстояний:

$$\sum_{r=1}^{N} \big(f_r - \Phi(r, C, \gamma, d)\big)^2 \to \min_{C,\gamma,d}. \tag{3}$$

Так как для целевой функции возможно определить градиент, то в качестве методов оптимизации могут быть использованы градиентные методы (например, метод Ньютона) или квазиньютоновские методы (например, метод Бройдена-Флетчера-Гольдфарба-Шанно).

С другой стороны для целевой функции может быть определена матрица Гессе, а следовательно для определения минимума суммы квадратов расстояний может быть применен метод Ньютона-Рафсона.

Кластерный анализ частотно-рангового представления ГТС

Групповой телеметрический сигнал может быть рассмотрен как текст, состоящий из набора слов, где под различными словами понимаются значения информационных слов телеметрического сигнала. Понятию естественного языка как определённой системе построения сообщений из словаря языка можно сопоставить понятие ГТС как некоторой системы компоновки, использующей временное разделение телеметрируемых параметров и принципов суб- и суперкоммутации. Таким образом, как и текст естественного языка, каждый телеметрический сигнал можно описать с использованием частотно-рангового представления.

В рамках предлагаемого подхода в качестве вектора признаков ГТС предлагается использовать набор коэффициентов (C, γ, d) функционала (2).

На первом шаге формируется словарь информационных слов, затем вычисляется частота встречаемости каждого слова из словаря. Полученное множество частот упорядочивается по убыванию значений, определяются ранги слов.

Ряд экспериментов позволил установить, что аппроксимацию частотно-рангового распределения (2) лучше применять для рангов начиная со второго, пропуская первый ранг, который определяет наиболее встречаемое слово в ГТС. Такая модификация позволяет в случаях резкого спада частоты встречаемости после первого ранга сократить относительную ошибку аппроксимации примерно на 20%. На рис. 2 представлены графики зависимостей ранг-частота слов типового группового телеметрического сигнала.

Потеря информации о наиболее часто встречаемом слове избегается за счёт расширения вектора признаков до 5 составляющих: (C, γ, d, a, F), где a — значение наиболее часто встречающегося слова, F — частота наиболее часто встречающегося слова.

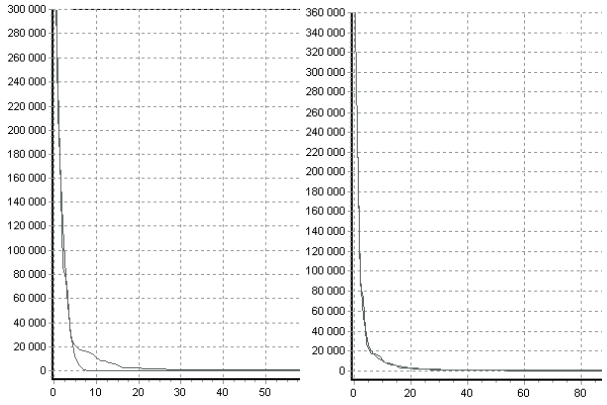


Рис. 2. Графики зависимостей ранг-частота слов типового группового телеметрического сигнала и графики аппроксимации для рангов начиная с первого (слева) и начиная со второго (справа).

К набору векторов, описывающих структуру ГТС, применяются алгоритмы кластерного анализа, которые позволяют разбить множество рассматриваемых ГТС на классы. В качестве алгоритмов кластеризации были рассмотрены EM-алгоритм [2] и алгоритм адаптивного выбора подклассов (АВП)[3].

Групповые телеметрические сигналы, отнесенные к одному кластеру, идентифицируются как имеющие одинаковую структуру. Основным достоинством предлагаемого метода является отсутствие требований к наличию априорной информации о структурных характеристиках и характеристиках параметров, входящих в анализируемый ГТС. Потенциальный недостаток метода связан с большим размером словаря, который в худшем случае составляет 2^m слов, где m — длина информационного слова в битах. К примеру, в случае размера слова, равного 32 битам, размер словаря в худшем случае составляет $4 \cdot 10^9$ слов, что ведет к потенциальным проблемам реализации.

Экспериментальные исследования

Экспериментальные исследования предложенного метода проводились на сгенерированных модельных данных. Генерация ГТС проводилась в два этапа: первый этап — генерация структуры ГТС, второй — «заполнение» ГТС измерениями параметров в соответствии со сгенерированной структурой. В качестве измерений параметров были использованы измерения из реальных ГТС. Для проведения эксперимента были построены 4 рандомизированные структуры ГТС с длиной слова 8 бит: «128-1» и «128-2» — структуры с длиной кадра 128 слов, «192-1» — соответственно, структура с длиной кадра 192 слова, «96-1» — с длиной кадра 96. Для структур «128-1», «128-2», «192-1» было сгенерировано по 10 ГТС, для структуры «96-1» было сгенерировано 5 ГТС. Во время анализа для каждого полученного ГТС был выбран участок длиной в 10000 кадров. Все участки находились в середине ГТС на одинаковом расстоянии от начала сигнала. Для каждого такого участка были подсчитаны коэффициенты (C, γ, d, a, F). В результате было получено всего 35 векторов признаков, которые приведены в таблице 1.

Таблица 1. Векторы признаков.

№	ГТС	C	γ	d	a	F
1	128-1	121405	1.153136	-0.000571	0	437592
2	128-1	122997	1.175762	-0.000690	0	437746
3	128-1	120894	1.170481	-0.000702	0	437516
4	128-1	121260	1.177205	-0.000738	0	437193
5	128-1	123508	1.183833	-0.000762	0	437781
6	128-1	124932	1.162494	-0.000583	0	437927
7	128-1	122713	1.156331	-0.000571	0	437290
8	128-1	122650	1.153792	-0.000547	0	437689
9	128-1	124728	1.158107	-0.000524	0	437925
10	128-2	123157	1.134301	-0.000381	0	437760
11	128-2	179367	1.388730	-0.001000	0	424732
12	128-2	184286	1.405204	-0.001012	0	424802
13	128-2	185375	1.401688	-0.000977	0	423611
14	128-2	183839	1.394106	-0.000953	0	423718
15	128-2	182194	1.388777	-0.000929	0	426018
16	128-2	163732	1.325167	-0.000822	0	441111
17	128-2	166616	1.332439	-0.000798	0	442402
18	128-2	168555	1.338316	-0.000810	0	440102
19	128-2	172903	1.358236	-0.000857	0	439995
20	128-2	168640	1.349081	-0.000869	0	441186
21	192-1	86452	0.920213	0.001491	0	538798
22	192-1	85952	0.918962	0.001491	0	538761
23	192-1	85989	0.921167	0.001431	0	538991
24	192-1	86052	0.923265	0.001396	0	538624
25	192-1	88366	0.935150	0.001384	0	538660
26	192-1	89364	0.941862	0.001276	0	532636
27	192-1	89390	0.953890	0.001312	0	544383
28	192-1	87263	0.934220	0.001360	0	544075
29	192-1	87245	0.933338	0.001396	0	544212
30	192-1	87470	0.932742	0.001408	0	544411
31	96-1	137370	1.174594	0.000096	127	485717
32	96-1	137460	1.179923	0.000037	127	485470
33	96-1	139224	1.288081	-0.000822	127	487801
34	96-1	139707	1.259078	-0.000595	127	488073
35	96-1	140419	1.264597	-0.000619	127	487448

Кластеризация проводилась с использованием двух методов: EM-алгоритма и АВП-алгоритма. Для EM-алгоритма параметры начального приближения были вычислены с помощью алгоритма k-средних, также для данного алгоритма не задавалось априорное число кластеров. Для алгоритма АВП было задано максимальное количество кластеров, равное 4 (количество различных структур в синтезированных данных). Как уже было ранее сказано критерием качества предложенного метода служит процент корректно идентифицированных ГТС. В данном случае под корректно

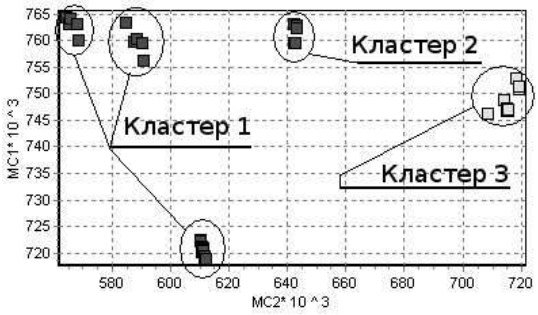


Рис. 3. Результаты кластерного анализа при пользовании EM-алгоритма.

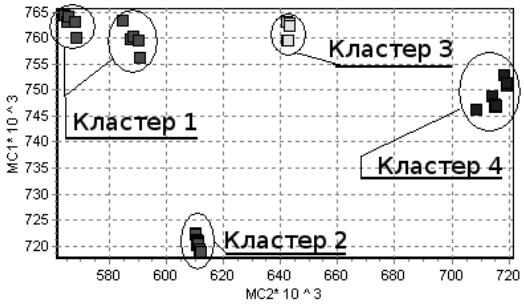


Рис. 4. Результаты кластерного анализа при использовании АВП-алгоритма.

идентифицированным ГТС понимается попадание вектора признаков в группу векторов, полученных из ГТС с идентичной структурой. Результаты выполнения указанных алгоритмов представленные в виде точек в пространстве главных компонент находятся на рис. 3 и 4

В результате работы EM-алгоритма было выбрано количество классов, равное 3. В первый класс попали вектора, характеризующие ГТС со структурой «96-1». Во второй класс были отнесены вектора всех запусков для структур «128-1» и «128-2», обладающие одинаковой длиной кадра. Третий класс составляют вектора, характеризующие ГТС со структурой «192-1».

Как было уже сказано, для АВП-алгоритма задавалось априорное число классов. В результате работы алгоритма каждый вектор попал в «свой» класс в соответствии со структурой ГТС. Таким образом EM-алгоритм верно определил классы, однако он отнес все ГТС с длиной кадра 128 к одному кластеру. Более точные результаты могут быть получены с применением АВП-алгоритма, который выявляет в представленных данных четыре кластера, что в точности соответствует содержимому модельных данных.

Выводы

В рамках работы предложен метод идентификации ГТС на основе поиска аналогов среди ранее накопленных данных. Для проведения идентификации ГТС не требуются знания о внутренней организации телеметрического потока, необходимо только знание длины информационного слова, что является основным достоинством рассмотренного метода. Потенциальный недостаток метода связан с большим объёмом словаря при больших длинах информационных слов. Полученные результаты подтверждают эффективность предлагаемого метода для идентификации таких сложных информационных структур, как групповые телеметрические сигналы.

Литература

[1] *Назаров А. В., Козырев Г. И., Шитов И. В. и др.* Современная телеметрия в теории и на практике. Учебн. курс. — СПб.: Наука и Техника, 2007. — 672 с.

[2] *Ian W., Eibe F.* Data mining. Practical machine learning tools and techniques with Java implementation — Morgan Kaufmann publishers, 1999.

[3] *Геппенер В. В., Емельянов Г. М.* Об одном подходе к задачам классификации // Известия ЛЭТИ (Известия Ленинградского электротехнического института), СПб., 1969. — С. 29–33.

Знаковое представление изображений и его информативность*

Броневич А. Г., Гончаров А. В.
brone@mail.ru, ag.tsure@gmail.com
Таганрог, Таганрогский Технологический Институт Южного Федерального Университета,
лаборатория математических методов искусственного интеллекта

В работе рассматривается аксиоматический подход к введению меры информативности на изображениях, приводящий к функционалу, обладающему свойствами энтропии Шеннона. На основе введенной меры информативности изображений вводится мера информативности знакового представления изображений, которое хорошо зарекомендовало себя при решении некоторых задач распознавания образов.

Введение

Идея знакового представления данных возникла при решении задач статистической оценки, и впервые была опубликована в виде целостной теории в монографии М. В. Болдина и соавторов [1] применительно к задачам эконометрики. Тем не менее, данный подход оказался продуктивным и при решении задач обработки и анализа изображений.

Задачи распознавания образов, такие как детекция лиц и идентификация лиц [2], локализация антропометрических точек лица [3], а также обнаружение нечетких дубликатов (нечеткость в данном случае интерпретируется как частичное совпадение изображений) в больших коллекциях изображений [4], эффективно решаются с помощью методов, основанных на знаковом представлении изображений.

Под знаковым представлением изображения в работах [2, 5, 4] понимается матрица (называемая также матрицей изменения яркостей), элементами которой являются пары чисел, соответствующие знакам частных производных от функции яркости изображения по направлениям координатных осей.

Целью данной работы является обобщение введенного ранее знакового представления изображений и изучение его свойств. Для этого рассмотрим понятие информативности и неопределенности изображения и соответствующего ему знакового представления.

Знаковое представление изображения

Под полутоновым изображением будем понимать неотрицательную целочисленную функцию $f = f(x_1, x_2)$, заданную в целочисленных точках сетки $\Omega = \{1, \dots, N_1\} \times \{1, \dots, N_2\}$, т. е. $f\colon \Omega \to \mathbb{Z}_+$, где $\mathbb{Z}_+ = \{0, 1, \dots\}$. Точку $\boldsymbol{x}_i = (x_{i_1}, x_{i_2})$ будем называть *пикселем*, тогда $f(\boldsymbol{x}_i)$ — значения яркости изображения f в пикселе $\boldsymbol{x}_i$, где $i = 1, \dots, N$, $N = |\Omega|$. Множество всех изображений на Ω обозначим через $\mathfrak{F}$. Знаковое представление изображения $f \in \mathfrak{F}$ будем описывать с помощью отношения квазипорядка τ на Ω, т. е. рефлексивного и транзитивного отношения, и получаемого с помощью транзитивного замыкания отношения

$$\sigma = \{(\boldsymbol{x}_i, \boldsymbol{x}_j)\colon f(\boldsymbol{x}_i) \geqslant f(\boldsymbol{x}_j),\ \boldsymbol{x}_i \in O(\boldsymbol{x}_j)\},$$

где $O(\boldsymbol{x}_j) = \{\boldsymbol{x}_i \in \Omega\colon \|\boldsymbol{x}_i - \boldsymbol{x}_j\| = 1\}$. Множество всех возможных рефлексивных и транзитивных отношений на Ω обозначим $\mathcal{T}$.

Очевидно, что одному знаковому представлению $\tau \in \mathcal{T}$ может соответствовать некоторое множество изображений $\mathfrak{F}_\tau \subseteq \mathfrak{F}$, поэтому возникает задача выбора из этого множества наиболее «типичного» представителя, т. е. такого изображения, которое бы содержало основную информацию об изображении и, в то же время, не содержало бы избыточной информации о градациях яркости пикселей. Проанализируем, как можно измерить количество данной информации для полутоновых изображений.

Мера информативности изображения

Меру информативности изображений будем строить в рамках аксиоматического подхода, согласно которому требуется определить некоторое конечное число аксиом (желательных свойств меры информативности), которые бы ее определяли однозначно.

Аксиома 1. Мера информативности — это функционал $U\colon \mathfrak{F} \to [0, +\infty)$.

Аксиома 2. Пусть $f \in \mathfrak{F}$ и множество значений $f(\Omega) = \{f(\omega) : \omega \in \Omega\}$ функции f является одноэлементным, т. е. $|f(\Omega)| = 1$. Тогда $U(f) = 0$.

Сформулируем аксиомы, которые бы позволили определить классы преобразований изображений, не изменяющих их информативность. Рассмотрим преобразование, связанное с «перемешиванием» пикселей на изображении. В этом случае значение информативности не меняется, мера информативности должна учитывать только градации фона.

Аксиома 3. Пусть $f\colon \Omega_1 \to \mathbb{Z}_+$ и $\psi\colon \Omega_1 \to \Omega_2$ — биекция. Тогда $U(\psi \circ f) = U(f)$.

*Работа выполнена при поддержке РФФИ, проекты №08-07-00129, №07-07-00067

Согласно следующей аксиоме, мы не теряем информацию о градациях яркости изображения, если присваиваем им новые значения с помощью биективного отображения.

Аксиома 4. Пусть $f\colon \Omega \to \mathbb{Z}_+$, и отображение $\varphi\colon f(\Omega) \to f(\Omega)$ является биекцией. Тогда $U(f \circ \varphi) = U(f)$.

Для каждого изображения f введем в рассмотрение функцию $h_f\colon \mathbb{Z}_+ \to \mathbb{Z}_+$, значение которой $h_f(i)$ дает нам число пикселей на изображении с яркостью i. Из аксиом 3–4 можно вывести следующее важное следствие.

Следствие 1. *Пусть изображения заданы функциями $f\colon \Omega_1 \to \mathbb{Z}_+$ и $g\colon \Omega_2 \to \mathbb{Z}_+$. Тогда $U(f) = U(g)$, если существует такая биекция $\varphi\colon \mathbb{Z}_+ \to \mathbb{Z}_+$, что $h_g(i) = h_f(\varphi(i))$ для всех $i \in \mathbb{Z}_+$.*

Отметим, что данное следствие позволяет упростить задачу, так как достаточно определить значение функционала U на всех возможных последовательностях вида $(h_f(0), h_f(1), \dots)$. Далее будем рассматривать также функционал $\bar{U}(f) = U(f)/|\Omega|$, показывающий среднее значение информативности пикселя для изображения $f \in \mathfrak{F}$.

Предположим, что изображение g состоит из k копий изображения f, в этом случае, очевидно, $h_g(i) = k\,h_f(i)$ для любого $i \in \mathbb{Z}_+$. Последнее условие можно выразить через частоты появления пикселей на изображениях f и g

$$p_g(i) = \frac{h_g(i)}{\sum_{j\in\mathbb{Z}_+} h_g(j)}, \quad p_f(i) = \frac{h_f(i)}{\sum_{j\in\mathbb{Z}_+} h_f(j)},$$

в виде $p_g(i) = p_f(i)$ для любого $i \in \mathbb{Z}_+$. Естественно предположить, что $\bar{U}(f) = \bar{U}(g)$ для таких изображений. Принимая во внимание аксиому 3, введем следующую аксиому.

Аксиома 5. Пусть $f, g \in \mathfrak{F}$ и $p_g(i) = p_f(i)$ для всех $i \in \mathbb{Z}_+$. Тогда $\bar{U}(f) = \bar{U}(g)$.

Пусть имеется изображение $f \in \mathfrak{F}$. Подействуем на это изображение (необязательно инъективным) отображением $\varphi\colon f(\Omega) \to f(\Omega)$. В результате получим изображение $f \circ \varphi$. Если φ не является инъекцией, то мы потеряем часть информации о градациях яркости в первоначальном изображении f, а именно, в этом случае множества $\varphi^{-1}(b) = \{a \in \varphi(\Omega)\colon \varphi(a) = b\}$ для $b \in \varphi(f(\Omega))$ не обязательно будут одноэлементными. Это означает, что исходное изображение «огрубляется» за счет присвоения «близким по яркости» пикселям одного значения. Отметим, что данное преобразование является характерным при обработке изображений, когда необходимо сократить число градаций яркости, оставляя наиболее характерные срезы функции изображения. Пусть $\varphi(f(\Omega)) = \{b_1, \dots, b_n\}$. Рассмотрим множества $\Omega_k = \{\omega \in \Omega\colon \varphi(f(\omega)) = b_k\}$, $k = 1, \dots, n$, которые, очевидно, задают разбиение множества Ω. Если отображение φ инъективно, то изображения $f_k\colon \Omega_k \to \mathbb{Z}_+$, являющиеся сужениями функции f на множества Ω_k будут иметь нулевую информативность согласно аксиоме 2, так как в данном случае $|f_k(\Omega_k)| = 1$. Когда же отображение φ не является инъективным, $|f_k(\Omega_k)| \neq 1$. Поэтому величины $\sum_{k=1}^{n} U(f_k)$ будут характеризовать суммарные потери информации при отображении φ. Предлагая такой аддитивный характер накопления неопределенности, можно ввести следующую аксиому аддитивности.

Аксиома 6. Пусть $f \in \mathfrak{F}$ и $\varphi\colon \mathbb{Z}_+ \to \mathbb{Z}_+$. Пусть $\varphi(f(\Omega)) = \{b_1, \dots, b_n\}$. Рассмотрим множества $\Omega_k = \{\omega \in \Omega\colon \varphi(f(\omega)) = b_k\}$, а также сужения $f_k\colon \Omega_k \to \mathbb{Z}_+$ функции f на множества Ω_k. Тогда $\sum_{k=1}^{n} U(f_k) + U(f \circ \varphi) = U(f)$.

Выразим рассмотренные аксиомы через введенный функционал $\bar{U}$. С учетом аксиомы 5, нам достаточно определить данный функционал для последовательности чисел $P = \big(p(i)\big)_{i\in\mathbb{Z}_+}$ таких, что $p(i) \geqslant 0$ и $\sum_{i\in\mathbb{Z}_+} p(i) = 1$. Отметим, что значение $p(i)$ можно интерпретировать как вероятность появления на изображении пикселя с яркостью i. Поэтому P можно рассматривать в качестве вероятностной меры. Тогда можно определить вероятность $P(A)$ любого подмножества $A \subseteq \mathbb{Z}_+$ выражением $P(A) = \sum_{i\in A} p(i)$. Далее будем использовать стандартные обозначения из теории вероятностей, в частности, пусть $\varphi\colon \mathbb{Z}_+ \to \mathbb{Z}_+$, тогда P^φ — это вероятностная мера, задаваемая равенством $P^\varphi(A) = \mathsf{P}\{i \in \mathbb{Z}_+\colon \varphi(i) \in A\}$. Отметим, что в рамках в поставленной задачи не требуется рассматривать всевозможные вероятностные меры. Согласно построению все $p(i)$ являются рациональными числами, и лишь конечное подмножество этих чисел отлично от нуля. Множество всех таких вероятностных мер на алгебре подмножеств $\mathbb{Z}_+$ обозначим через M_{pr}.

Следствие 2. *Функционал $\bar{U}$ на M_{pr} обладает следующими свойствами.*

1. *$\bar{U}(P) \geqslant 0$ для всех $P \in M_{\mathrm{pr}}$;*
2. *$\bar{U}(P) = 0$, если существует $i \in \mathbb{Z}_+$, что $P(i) = 1$;*
3. *Пусть отображение $\varphi\colon \mathbb{Z}_+ \to \mathbb{Z}_+$ инъективно. Тогда $\bar{U}(P^\varphi) = \bar{U}(P)$ для всех $P \in M_{\mathrm{pr}}$;*
4. *Пусть $P \in M_{\mathrm{pr}}$, $\varphi\colon \mathbb{Z}_+ \to \mathbb{Z}_+$. Рассмотрим разбиение множества $A = \{i \in \mathbb{Z}_+\colon p(i) > 0\}$ на подмножества, представляющие прообразы элементов множества $B = \{b_1, \dots, b_n\} = \varphi(A)$, т. е. разбиение состоит из множеств $A_k = \{i \in A\colon \varphi(i) = b_k\}$. Тогда*

$$\bar{U}(P) = \sum_{k=1}^{n} P(A_k)\,\bar{U}(P_{A_k}) + \bar{U}(P^\varphi),$$

где P_{A_k} *— это условные вероятностные меры, и* $P_{A_k}(C) = P(C \cap A_k)/P(A_k)$, $C \subseteq \Omega$.

Замечание 1. Отметим, что свойства, перечисленные в следствии 2, являются хорошо известными свойствами энтропии Шеннона. Свойство 2 аккумулирует в себе свойства симметричности и продолжения. Свойство 4 — это свойство аддитивности, которое формулируется следующим образом. Пусть ξ — случайная величина со значениями в $\mathbb{Z}_+$ и $\eta = \varphi(\xi)$. Тогда для энтропии Шеннона S выполняется: $S(\xi, \eta) = S(\xi) = S(\xi \,|\, \eta) + S(\xi)$. В данном случае $S(\xi \,|\, \eta) = S(\xi)$, так как значения η полностью зависят от ξ.

Утверждение 1. *Пусть функционал* $\bar{U}$ *на* M_{pr} *удовлетворяет свойствам, перечисленным в следствии 2. Тогда* $\bar{U}$ *— это энтропия Шеннона, т. е.*

$$\bar{U}(P) = -c \sum_{i \in A} p(i) \ln p(i),$$

где $P \in M_{\mathrm{pr}}$, $A = \{i \in \mathbb{Z}_+ \colon p(i) > 0\}$ *и* $c \geqslant 0$.

Можно условно считать, что $p(i) \ln p(i) = 0$, если $p(i) = 0$. Тогда для произвольного изображения $f \in \mathfrak{F}$ мера информативности

$$U(f) = -cN \sum_{i \in \mathbb{Z}_+} p_f(i) \ln p_f(i),$$

где $N = |\Omega|$ и $p_f(i) = h_f(i)/N$. Мера информативности определяется единственным образом условием нормировки. Например, будем считать, что самое информативное изображение состоящее из n пикселей имеет значение информативности, равное 1. Тогда $c = 1/(n \ln n)$.

Отметим также, что имеется следующая вероятностная интерпретация средней информативности $\bar{U}(f)$ пикселя для изображения $f \in \mathfrak{F}$. Рассмотрим вероятностную меру P на алгебре всех подмножеств Ω, задаваемую равенством $P(A) = |A|/|\Omega|$. Тогда отображение $f \in \mathfrak{F}$ можно рассматривать в качестве случайной величины и, очевидно, $S(f) = \bar{U}(f)$. Далее рассмотрим векторную случайную величину $\xi = (\xi_1, \ldots, \xi_N)$, где $N = |\Omega|$ и случайные величины независимы и одинаково распределены, как и случайная величина f. Тогда $S(\xi) = \bar{U}(f)$.

Меры информативности и неопределенности знакового представления

Рассмотрим теперь вопрос, как можно измерить информативность знаковых представлений, а также их неопределенность, вызванную потерями информации о градациях яркости изображения. Данные характеристики мы будем описывать с помощью функционалов U и $\widehat{U}$, определенных на множестве отношений квазипорядка и дающих соответственно значения их информативности и неопределенности. Данные функционалы будем определять аксиоматически, указывая их желательные свойства. Ясно, что функционалы U и $\widehat{U}$ должны быть некоторым образом связаны с ранее введенной мерой информативности, определенной на множестве изображений $\mathfrak{F}$. Эту связь дает

Аксиома 7. Пусть $\tau \in \mathcal{T}$, тогда

$$U(\tau) + \widehat{U}(\tau) = \sup\{U(f) \colon f \in \mathfrak{F}_\tau\} = U_{\max}(\tau). \quad (1)$$

Отметим, что аксиома 7 выражает известный принцип в теории информации, предложенный Д. Клиром (G. J. Klir) [5], согласно которому неопределенность связана с некоторыми потерями информации. Правая часть формулы (1) соответствует информативности наиболее информативного изображения со знаковым представлением τ. Поэтому, из формулы (1) выводим, что $\widehat{U}(\tau) = U_{\max}(\tau) - U(\tau)$. Таким образом, количество неопределенности знакового представления τ изображения f равна разности информативности изображения и информативности его знакового представления, причем изображение f выбирается из принципа максимума неопределенности. Будем считать изображения $f_1, f_2 \in \mathfrak{F}$ эквивалентными, если существует монотонно возрастающая биекция $\varphi \colon f_1(\Omega) \to f_2(\Omega)$, что $f_2 = \varphi \circ f_1$. Предположим, что эквивалентные изображения содержат одну и ту же информацию. Тогда должна выполняться следующая аксиома.

Аксиома 8. $\widehat{U}(\tau) = 0$, если отношение $\tau \in \mathcal{T}$ является связным, т. е. любые два элемента $\omega_1, \omega_2 \in \Omega$ сравнимы между собой.

Аксиома 9. Пусть $\tau_1 \subseteq \tau_2$ для $\tau_1, \tau_2 \in \mathcal{T}$, тогда $\widehat{U}(\tau_1) \geqslant \widehat{U}(\tau_2)$.

Отметим, что, когда $\tau_1 \subseteq \tau_2$, мы имеем больше информации, описывая изображение с помощью знакового представления τ_2, по сравнению со знаковым представлением τ_1. Откуда мы выводим, что аксиома 9 должна выполняться.

Аксиома 10. Пусть $G_\tau(\Omega)$ — это граф знакового представления $\tau \in \mathcal{T}$ и множества $\Omega_1, \ldots, \Omega_m$ определяют компоненты связности графа G_τ. Тогда $\sum\limits_{k=1}^{m} U(\tau_{\Omega_k}) = U(\tau)$, где $\tau_{\Omega_k} = \tau \cap \Omega_k \times \Omega_k$ — сужение отношения τ на множество Ω_k, $k = 1, \ldots, m$.

Смысл аксиомы 10 заключается в том, что компоненты связности графа G_τ представляют собой фрагменты независимой информации, поэтому информативность всего представления должна быть равной сумме информативностей данных независимых компонент. Нашей дальнейшей задачей будет теоретическое исследование свойств функционалов U, $\widehat{U}$, $U_{\max}$ на $\mathcal{T}$ и рассмотрение способов определения U и $\widehat{U}$. Пусть $\tau \in \mathcal{T}$, тогда отношение

$\theta = \tau \cap \tau^{-1}$ является отношением эквивалентности. Обозначим через $V = \{v_1, \ldots, v_n\}$ множество всех классов эквивалентности, определяемых отношением θ на Ω. Рассмотрим также продолжение τ^θ отношения θ на классы эквивалентности. Считаем, что $(v_i, v_j) \in \tau^\theta$, если существует пара $(\omega_i, \omega_k) \in \theta$, что $\omega_i \in v_i$ и $\omega_k \in v_j$. Известно, что получаемое таким образом отношение τ^θ на V является рефлексивным, антисимметричным и транзитивным отношением, т. е. отношением частичного порядка. Известно также, что всегда можно построить отношение нестрогого линейного порядка ρ так, что $\rho \supseteq \tau^\theta$. Тогда отношению ρ будет соответствовать класс изображений, в котором все v_i имеют различные градации яркости. Отсюда следует следующее утверждение.

Утверждение 2. *Пусть $\tau \in \mathcal{T}$, $\theta = \tau \cap \tau^{-1}$ и τ^θ — это продолжение отношения τ на множество V классов эквивалентности, порожденных θ, тогда*

$$U_{\max}(\tau) = -cN \sum_{i=1}^{n} p(i) \ln p(i),$$

где $p(i) = |v_i|/N$, и $N = |\Omega|$.

Следствие 3. *Пусть G_τ — это граф отношения $\tau \in \mathcal{T}$, и его компоненты связности определяются множествами $\Omega_1, \ldots, \Omega_m$, причем τ_{Ω_i}, $i = 1, \ldots, m$ — это связные отношения. Тогда*

$$\widehat{U}(\tau) = -cN \sum_{i=1}^{m} p(i) \ln p(i),$$

где $p(i) = |\Omega_i|/N$ и $N = |\Omega|$.

Нетрудно показать, что, если выполняются условия следствия 3, то отношение $\tau \cup \tau^{-1}$ является отношением эквивалентности, и с ним связано разбиение $\{\Omega_1, \ldots \Omega_m\}$ и $\widehat{U}\left(\tau \cup \tau^{-1}\right) = \widehat{U}(\tau)$.

Утверждение 3. *Пусть $\tau \in \mathcal{T}$ и $\alpha \subseteq \tau \cup \tau^{-1}$ — это отношение эквивалентности. Тогда $\widehat{U}(t) \leqslant \widehat{U}(\alpha)$.*

Утверждение 3 позволяет ввести следующую верхнюю оценку $\widehat{U}_{\text{up}}$ для меры неопределенности $\widehat{U}$. Пусть $Eq\left(\tau \cup \tau^{-1}\right)$ — это множество всех отношений эквивалентности, которые включаются в отношение $\tau \cup \tau^{-1}$. Тогда функционал $\widehat{U}_{\text{up}}$ определим следующим образом:

$$\widehat{U}_{\text{up}}(\tau) = \min\left\{\widehat{U}(\alpha) \colon \alpha \in Eq\left(\tau \cup \tau^{-1}\right)\right\},$$

и в силу утверждения 3 $\widehat{U}(\tau) \leqslant \widehat{U}_{\text{up}}(\tau)$ для всех $\tau \in \mathcal{T}$.

Утверждение 4. *Функционал $\widehat{U}_{\text{up}}(\tau)$, как мера неопределенности знакового представления, и функционал $U = U_{\max} - \widehat{U}_{\text{up}}$, как мера информативности знакового представления, на множестве знаковых представлений $\mathcal{T}$ удовлетворяют аксиомам 7–10.*

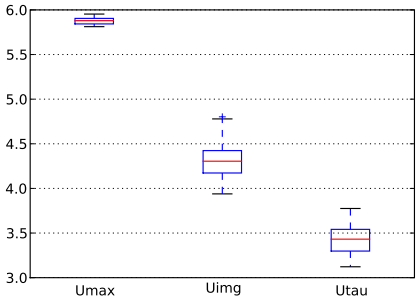


Рис. 1. Результат оценки информативности изображений лиц и соответствующих знаковых представлений.

Примеры

Рассмотрим вычисление мер информативности изображений и соответствующих им знаковых представлений на примере изображений лиц из базы BioID [6]. На рис. 1 представлены графики, характеризующие значения информативности исходных изображений, информативность соответствующих им знаковых представлений и максимальную информативность изображений.

Заключение и выводы

В рамках данной работы рассмотрено знаковое представление изображений, и предложен аксиоматический подход к введению меры его информативности.

Литература

[1] *Болдин М., Симонова Г., Тюрин Ю.* Знаковый статистический анализ линейных моделей. М.: Наука. Физматлит. — 1997.

[2] *Гончаров А., Горбань А., Каркищенко А., Лепский А.* Поиск портретных изображений по содержанию // Интернет-математика 2007: Сборник работ участников конкурса. — 2007. P. 56–64. `http://download.yandex.ru/IMAT2007/goncharov.pdf`

[3] *Goncharov A., Gubarev V.* Comparison of high-level and low-level face recognition methods // Pattern recognition and image analysis: new information technologies (PRIA-9-2008). — 2008. — P. 178–181.

[4] *Goncharov A., Melnichenko A.* Pseudometric approach to content based image retrieval and near duplicates detection. // Российский семинар по Оценке Методов Информационного Поиска. Труды РОМИП 2007–2008. — 2008.— P. 120–134.

[5] *Klir G. J.* Uncertainty and Information: Foundations of Generalized Information Theory. John Wiley & Sons, Inc. — 2006.

[6] *Jesorsky O., Kirchberg K., Frischholz R.* Robust face detection using the hausdorff distance // Audio and Video based Person Authentication — AVBPA, Springer. — 2001.

Адаптивное сжатие графической информации на базе корреляционно-экстремальных контурных методов*

Васин Ю. Г., Лебедев Л. И.
lebedev@pmk.unn.ru
Нижний Новгород, НИИ прикладной математики и кибернетики НГУ им. Н. И. Лобачевского

В работе предлагается технология решения задачи адаптивного сжатия графической информации на базе корреляционно-экстремальных контурных методов (КЭКМ) распознавания дискретных объектов. Приводится эффективный метод представления описания информации об эталоне и структурированных данных о его местоположении, ориентации, габаритах на документе, которые он должен принять, чтобы заменить исходный объект. Показывается, что эффективность сжатия обеспечивается за счет выбора метода распознавания дискретных объектов, использования стандартных наборов эталонных последовательностей, оптимального описания автоматически формируемых эталонов и повторяемости на обрабатываемом документе дискретных объектов со сходной формой. Приводятся результаты работы полученного алгоритма на примере сжатия информации фрагмента морской навигационной карты.

Введение

В последнее время заметно возросла степень использования изображений в различных областях человеческой деятельности в целях оперативного управления. Наглядность, изобразительность, тематичность и другие особенности изображения, как правило, не требуют для принятия решений дополнительных сведений о данной предметной области. А так как оперативное управление предполагает передачу информации по каналам связи, то этот факт предопределяет необходимость эффективного способа представления изображений, в частности, путем сжатия его описания. В зависимости от постановки задачи к алгоритмам и методам сжатия предъявляются различные требования к качеству восстановленного изображения — от сжатия без потерь до сжатия с приемлемой визуализацией получаемого на выходе изображения. Особенностью графической информации многих сканируемых документов является наличие большого количества хаотически расположенных дискретных объектов, значительно снижающих коэффициент сжатия известных алгоритмов. Однако по составу набор дискретных объектов на большинстве растровых документах существенно ограничен, что является основанием для оптимального представления всех дискретных объектов на документе. Поэтому представляется перспективным направление, связанное с декомпозицией сканированного документа, состоящей в выделении информации о дискретных объектах в целях получения оставшегося изображения, более эффективного с точки зрения сжатия. Полученное изображение будет содержать информацию только о точечных и линейных объектах. Робастная фильтрация точечных объектов позволяет освободиться от шумов сканирования документа. Базовым описанием графических изображений принято считать растровую модель представления информации. Альтернативной базовому описанию графических изображений служит векторная модель представления информации, в которой объектами являются ориентированные самонепересекающиеся многоугольники на плоскости (контурамы). Координаты вершин многоугольника задают метрическое описание объектов. Существует много различных методов сжатия описаний линейных объектов, однако с точки зрения единого подхода к формированию метрического описания эталона здесь выбран метод задания точек посредством приращений, то есть покоординатных разностей двух соседних точек исходного описания $(\Delta x_i, \Delta y_i) = (x_{i+1} - x_i, y_{i+1} - y_i)$. Кроме того, при сжатии изображения существует специфика передачи определенного рода объектов, которые должны быть восстановлены с необходимой точностью, отличной от точности восстановления других объектов. Таким образом, при разработке технологии сжатия изображения необходимо предусмотреть, чтобы имелась возможность адаптивного сжатия объектов.

Данная работа посвящена именно решению этих вопросов.

Постановка задачи

Переведем интуитивное представление о точечных, дискретных и линейных объектах в разряд их формального описания. Точечным будем считать объект изображения, если изменение его границы на одну единицу растра влечет существенное изменение формы контура, и поэтому их трудно отнести к одному классу. В нашем случае это объекты, имеющие размеры растра 4×4 и менее. Основная масса штриховых условных знаков на документах, собственно для сжатия которых и создавалась эта технология (топографические, морские навигационные карты, планы и планшеты) имеют физические размеры менее 10 мм. Это означает, что матрицы 256×256 пикселей достаточно для представления штриховых условных знаков при сканировании

*Работа выполнена при финансовой поддержке РФФИ, проект № 05-01-00590.

документов с разрешением 600 dpi и ниже. Поэтому любой объект, который имеет габариты меньше 256 единиц и не относится к точечным, будем считать дискретным. Очевидно, что оставшуюся массу объектов будем называть линейными. Оптимальное сжатие информации графических документов предполагает в первую очередь оптимальное представление дискретных объектов на них. Существенное и основное отличие предлагаемого метода сжатия состоит в том, что сжатию подвергаются не описания дискретных объектов, а само множество дискретных объектов, содержащихся на графическом изображении. Это означает, что все дискретные объекты необходимо автоматически разбить на группы (классы) объектов, имеющих между собой определенное сходство. Тогда каждую группу объектов, нивелируя различия в их описаниях, будет представлять один из них (эталон), а описания всех остальных объектов из групп не передавать. В этом случае для идентификации на документе каждого объекта из группы необходимо дополнительно задать только информацию о его местоположении, ориентации, размерах и номере эталона. В этом и состоит суть предлагаемого алгоритма сжатия графических изображений.

Для того, чтобы реализовать этот подход к оптимальному сжатию передаваемой информации о дискретных объектах, необходимо решить три задачи:

1) задача нахождения минимального числа эталонных групп на базе всех дискретных объектов;
2) задача получения эффективного описания самих эталонов;
3) задача эффективного представления любого дискретного объекта на базе сформированных эталонов.

Кроме того, в рамках единого подхода к формированию метрического описания эталонов необходимо решить задачу эффективного представления линейных объектов.

Методы решения

Очевидно, что решение первой задачи напрямую связано с выбором метода распознавания дискретных объектов. Так как количество групп и состав каждой группы изначально неизвестен, то очевидно, что они должны формироваться автоматически из числа тех дискретных объектов, которые имеют низкий коэффициент сходства с любым из эталонов, представляющих группы. Для того, чтобы по составу группа была максимальной, а их количество минимальным, необходимо, чтобы методы определения сходства двух объектов были инварианты к ортогональным преобразованиям и масштабированию. Это предположение давало бы возможность к одному классу отнести все объекты, описания которых отличаются местоположением, ориентаций и габаритами на документе. Поэтому в качестве метода определения сходства двух фигур был взят корреляционно-экстремальный контурный метод [1]. В качестве исходного описания этот метод использует метрическое описание отдельно взятых контуров, которые для этого метода определяются как дискретные объекты. Таким образом, если метрическое описание эталона формировать на базе дискретного объекта относительно габаритной точки с минимальными координатами, то для задания любой точки эталона достаточно двух байт оперативной памяти. Еще два байта необходимы для задания количества точек данного эталона и его типа. Таким образом, выбор КЭКМ в качестве метода распознавания дискретных объектов решает и вторую задачу. На основе полученного описания эталона теперь для восстановления на документе любого дискретного объекта из этой группы необходимо определить следующие параметры:

1) координаты точки, определяющие местоположение дискретного объекта на документе;
2) номер эталона, представляющий класс объектов;
3) ориентированные габариты эталона после его совмещения с объектом;
4) направление обхода контура.

Так как параметры совмещения объекта с эталоном корреляционно-экстремальным методом находятся аналитически, то получение необходимых данных для восстановления дискретного объекта на графическом изображении не является сложной вычислительной задачей [2, 3]. Как показывают расчеты, для описания дискретных объектов в новом формате с учетом отмеченных выше размеров достаточно 8 байт для документов, растровое изображение которых не превышает по любой из координат величину 32768.

Для формирования нового описания линейных объектов предлагается использовать ту же структуру хранения, которая разработана для эталонов дискретных объектов. В этом варианте метрическое описание линейных объектов задается координатами начальной точки, а последующие узлы — посредством приращений, не превышающих половины от максимально возможных габаритов дискретных объектов. Если покоординатные расстояния превышают этот порог, то описание линейного объекта дополняется одной или несколькими интерполированными точками. Для получения более эффективного представления линейных объектов описания последних подвергаются кусочно-линейному сжатию с адаптивным порогом, величина которого зависит от сжатия различных участков контура, в том числе зависит и от его длины. Отме-

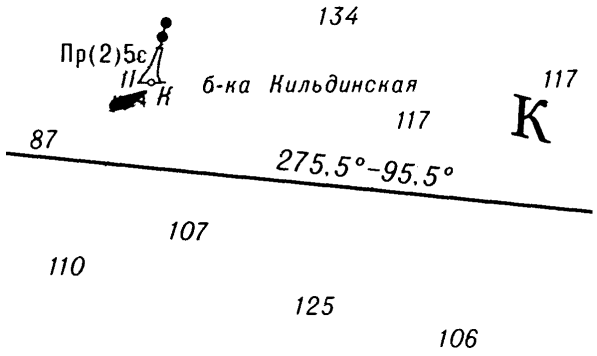


Рис. 1. Исходное изображение фрагмента МНК.

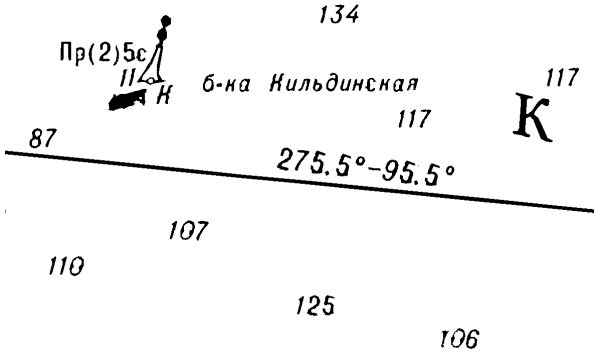


Рис. 2. Восстановленное изображение фрагмента.

тим, что для идентификации описаний линейных и дискретных объектов имеется возможность задания типа эталона.

На графическом документе не все области являются равнозначными по точности описания и восстановления дискретных объектов. Поэтому в алгоритме сжатия предусмотрена возможность изменения порога сходства в заданных областях при распознавании дискретных объектов и формировании их новых описаний. Те дискретные объекты, метрическое описание которых не должно подвергаться каким-либо искажениям, передаются как линейные объекты. Таким способом решается задача адаптивного сжатия описаний эталонов.

При описанной методике последовательного пополнения файла эталонов из числа неопознанных дискретных объектов может возникнуть ситуация, когда ранее опознанный объект будет иметь большее сходство с только что сформированным эталоном и, следовательно, желательно передавать выходное описание дискретного объекта на базе нового эталона. Этот недостаток алгоритма легко устраняется при повторении процедуры распознавания дискретных объектов на базе уже сформированного файла эталонов.

Полученные результаты

Из вышеизложенного следует, что описанный алгоритм сжатия на базе корреляционно-экстремального контурного метода позволяет решить поставленные задачи. В качестве примера применения алгоритма сжатия на базе КЭКМ можно проиллюстрировать его работу на фрагменте морской навигационной карты, изображенном на рис. 1.

Данное изображение размерами 1177×755 в формате pcx имеет объем, равный 17237 байт. Контурная модель данного изображения, являющаяся одним из вариантов векторного формата представления информации, содержит в общей сложности 85 объектов (контуров), из которых 2 точечных объекта и 1 линейный. Использование предлагаемой технологии сжатия информации изображения позволило сформировать последовательность из 31 эталона с общим объемом их метрического описания в 2350 байт. На основе этих эталонов были получены новые описания всех 83 объектов, объемом в 672 байта. Таким образом, новое описание информации изображения составило 3022 байта. Отсюда, коэффициент сжатия информации представленного изображения относительно исходного pcx-файла равен 5,7. Сжатие без потерь стандартным архиватором позволяет уменьшить объем pcx-файла до 6444 байт, а объем описаний эталонов — до 1771 байта. Сравнение этих описаний позволяет сделать вывод о сжатии передаваемой информации в 2,65 раза. На рис. 2 представлено восстановленное изображение, из анализа которого непосредственно можно сделать вывод об отсутствии потерь какой либо информации.

Еще большего коэффициента сжатия можно добиться, если при распознавании использовать стандартные шрифты для данной тематической области, которые нет необходимости передавать. На нашем изображении в этом случае будут опознан 61 символ (некоторые символы состоят из совокупности нескольких контуров), для передачи каждого из них достаточно тех же 8 байт. Кроме того, для получения нового описания условных знаков будет сформированы эталонные последовательности из нераспознанных контуров объемом 210 байт и на основе их еще 5 новых описаний объектов. В этом случае объем передаваемой информации составит всего 738 байт, что поднимет коэффициент сжатия до отметки в 8,73 раза.

Приведенный пример сжатия информации фрагмента морской навигационной карты продемонстрировал заявленные при разработке данного алгоритма положения об эффективности формирования эталонов, их описаний и нового представления объектов. Предлагаемый алгоритм также является более универсальным и эффективным, так как превосходит в несколько раз по степени сжатия графических изображений, содержащих объекты с произ-

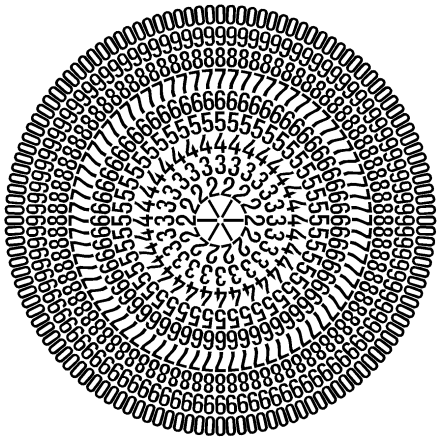


Рис. 3. Тестовое изображение

вольной ориентацией и местоположением, известные процедуры, использующие такие форматы как pdf, jpeg, png, tif и другие. Наиболее убедительно это можно продемонстрировать на сжатии тестового изображения, приведенном на рисунке 3.

Данное тестовое изображение имеет размеры 1320×1320 и содержит 1044 элементарных объектов (контуров). Совершенно очевидно, что сжатие с использованием процедур, основанных на построении квадродеревьев, форматов jpeg и других им подобных алгоритмов, крайне неэффективно. Поэтому в таблице 1 сведения об объемах тестового изображения в этих форматах отсутствуют. Первые четыре формата представления изображения являются стандартными, причем первые два являются разработками НИИ ПМК ННГУ.

формат	объем (б)	Объем zip (б)
`str`	197120	102751
`kmp`	207872	71518
`pcx`	198749	73758
`bmp`	221822	70436
`djvu`	21502	21644
`png`	89754	89891
`tif`	36680	35596

Таблица 1. Объем тестового изображения в различных форматах.

Из таблицы видно, что объемы файлов в этих форматах приблизительно одни и те же, причем использование zip-архиватора позволяет уменьшить объемы этих файлов, за исключением первого, почти в 3 раза и получить результаты сжатия, сравнимые с использованием формата png. Из таблицы также следует, что наиболее эффективными являются форматы tif и djvu. Предлагаемый алгоритм сжатия графической информации на базе КЭКМ формирует файл эталонов размером 528 байт и файл нового описания контуров размеров 8×1044 = 8352 байта. В итоге получается описание тестового изображения, представленное 8890 байтами. Более того, использование zip-архиватора позволяет уменьшить объем этих данных до 7050 байт, который в 3 раза меньше, чем у лучшего из приведенных методов сжатия.

Заключение

Коэффициенты сжатия в 3 и более раз по отношению к объемам архивированных pcx-файлов были получены в результате экспериментов, проведенных на нескольких трейсинг-кальках, представляющих случайную выборку из имеющегося набора документов такого типа, для сжатия которых собственно и разрабатывался этот алгоритм. В данных экспериментах распознавание с использованием наборов стандартных шрифтов не использовалось. Сжатие происходило за счет использования эффективных методов распознавания, использования оптимального числа эталонов, и на этой основе нового более эффективного описания дискретных объектов.

Необходимо отметить также тот факт, что предлагаемый алгоритм является процедурой, настраиваемой на уровень сходства контуров. Увеличивая или уменьшая уровень сходства формы двух контуров можно получать на восстановленных графических изображениях дискретные объекты с большей или меньшей степенью близости относительно исходных описаний, вплоть до передачи метрического описания контуров без каких-либо искажений.

Проведенные исследования показали, что данная технология адаптивного сжатия графических изображений может быть использована для решения задач оптимального представления данных в целях хранения и/или передачи информации по каналам связи.

Литература

[1] *Васин Ю. Г., Лебедев Л. И., Пучкова О. В.* Контурные корреляционно-экстремальные методы обнаружения и совмещения объектов видеоинформации // Автоматизация обработки сложной графической информации: Межвуз. темат. сб. науч. тр. / Под ред. Ю. Г. Васина. Горький: Горьков. гос. ун-т, 1987. — С. 97–112.

[2] *Васин Ю. Г., Лебедев Л. И.* Инвариантные методы определения сходства плоских форм // Информационные технологии в анализе изображений и распознавании образов: Тез. докл. 1-й межд. конф.: Львов: Физ.-мат. ин-т АН УССР, 1990. — С. 225–228.

[3] *Васин Ю. Г., Лебедев Л. И., Пучкова О. В.* Оптимизация вычислительной и емкостной сложности алгоритмов распознавания объектов видеоинформации // Автоматизация обработки сложной графической информации: Межвуз. темат. сб. науч. тр. / Под ред. Ю. Г. Васина. Нижний Новгород: Нижегород. гос. ун-т, 1987. — С. 62–86.

Критериальные проективные морфологии

Визильтер Ю. В.
viz@gosniias.ru
Москва, ФГУП Государственный научно-исследовательский институт авиационных систем

Описана критериальная проективная морфология, обобщающая свойства морфологий Пытьева и Серра. Определены способы вычисления морфологического коэффициента корреляции и морфологических спектров. Сформулирован ряд достаточных условий проективности критериальных операторов, выделены соответствующие типы морфологических проекторов. Приведены примеры критериальной проективной морфологической фильтрации одномерных функций и контуров двумерных изображений.

В области морфологического анализа изображений существуют два наиболее известных на сегодняшний день математических формализма: математическая морфология (ММ) Серра [1] и морфологический анализ Пытьева [2]. В статье [3] была предложена так называемая проективная морфология разложений, опирающаяся на структурное представление изображения в виде «моделей с однородными связями» и позволяющая единообразно описывать как операторы ММ, так и проекторы на форму разбиения кадра Пытьева. В данной работе рассматривается альтернативный подход к объединению морфологий, основанный на задании целевых критериев и построении оптимальных в смысле этих критериев проективных операторов. При этом, несмотря на отказ от обязательного выделения на изображении структурных элементов, сохраняется возможность вычисления морфологических коэффициентов корреляции, а также построения и анализа морфологических спектров.

Проективные морфологии

Пусть имеется множество образов Ω , на котором определена операция сложения «+», задающая на Ω группу с «нулевым образом» $\varnothing$, определена операция вычитания «−» и определена норма $\mu(A) = \|A\|\colon \Omega \to R$, причем норма разности двух образов обладает свойствами расстояния:

$$\forall A, B, C \in \Omega\colon\ \|A - B\| \geqslant 0,\ \|A - A\| = 0,$$
$$\|A - B\| + \|B - C\| \geqslant \|A - C\|.$$

Введем на Ω оператор проекции Pr:

$$\forall A \in \Omega\colon\ \Pr(A) \in \Omega,\ \Pr(\varnothing) = \varnothing,$$
$$\Pr(A) = \Pr(\Pr(A)). \qquad (1)$$

Алгебраическую систему $\{\Omega, +, \mu, \Pr\}$ будем называть проективной морфологией на Ω на базе проектора Pr. Множество собственных (стабильных) элементов проектора

$$M = \{A \in \Omega\colon\ \Pr(A) = A\}$$

назовем модельным множеством или моделью. Очевидно, проектор Pr имеет смысл оператора проецирования образа на модель:

$$\Pr(A) = \Pr(A, M).$$

Для сравнения ненулевого образа с моделью определим морфологический коэффициент корреляции изображения с моделью

$$K_M(A, M) = \exp\left(-\frac{\|A - \Pr(A, M)\|}{\|\Pr(A, M)\|}\right) \qquad (2)$$

со следующими стандартными свойствами:

1) $0 \leqslant K_M(A, M) \leqslant 1$;
2) $K_M(A, M) = 1 \Leftrightarrow A \in M$;
3) $K_M(A, M) = 0 \Leftrightarrow \Pr(A, M) = \varnothing$.

Заметим, что форма выражения (2) отличается от формы морфологического коэффициента корреляции, предложенного Пытьевым [2] и также использовавшегося в работе [3], поскольку в общем случае равенство нормы проекции норме исходного изображения не гарантирует их совпадения. Для ненулевых образов можно также определить морфологический коэффициент корреляции изображений

$$K_M(A, B, M) = 1 - \frac{\|\Pr(A, M) - \Pr(B, M)\|}{\max(\|A\|, \|B\|)},$$

$A, B \in \Omega$, позволяющий установить морфологическую эквивалентность образов. Пусть теперь имеются модели M_1 и M_2. Если

$$M_2 \subseteq M_1, \qquad (3)$$

то модель M_1 по отношению к M_2 является морфологически более сложной. Морфологическая сложность [2] определяет на множестве моделей отношение частичного порядка. В терминах морфологических коэффициентов корреляции условие (3) имеет вид

$$\forall A \in M_2\colon K_M(A, M_1) = 1;$$
$$\exists B \in M_1\colon K_M(B, M_2) < 1.$$

Критериальные морфологии

Пусть теперь задана целевая функция-критерий

$$\Phi(A, B)\colon \Omega \times \Omega \to R,$$

и пусть задача построения критериального морфологического фильтра имеет вид

$$\psi(A, \Phi) = \arg\min_{B \in \Omega} \Phi(A, B). \qquad (4)$$

При этом хорошо определенным критерием является такой, что

$$\forall A \in \Omega \ \exists B \in \Omega: \\ \forall C \in \Omega,\ C \neq B \Rightarrow \Phi(A,B) < \Phi(A,C), \quad (5)$$

то есть критерий Φ однозначно определяет морфологический фильтр $\psi(A,\Phi)$. Если $\psi(A,\Phi)$ удовлетворяет условию проективности (1), он может быть назван критериальным морфологическим проектором и определяет критериальную проективную морфологию на базе критерия Φ.

Определим еще несколько полезных понятий. Областью допустимых значений (ОДЗ) критерия Φ при проецировании исходного образа A назовем

$$V(A,\Phi) = \{B \in \Omega \colon \Phi(A,B) < +\infty\}.$$

Соответственно необходимое условие проективности имеет вид

$$\forall A \in \Omega \colon B \in V(A,\Phi) \Rightarrow B \in V(B,\Phi). \quad (6)$$

Определим условие монотонности ОДЗ:

$$\forall A \in \Omega,\ \forall B \in V(A,\Phi) \colon V(B,\Phi) \subseteq V(A,\Phi). \quad (7)$$

Критерии, для которых выполняется условие (7), будем называть монотонными по ОДЗ или просто монотонными. Далее будем рассматривать следующий стандартный критерий штрафа

$$\Phi(A,B) = J(A,B) + \chi(A,B) + \alpha \times Q(B), \quad (8)$$

где $J(A,B)$ — критерий соответствия проекции и проецируемого образа (matching function), обладающий следующим естественным свойством

$$\forall A \in \Omega, B \in V(A,\Phi) \colon J(A,A) \leqslant J(A,B),$$

$\chi(A,B)$ — критерий (предикат) допустимости решения (validation function) вида

$$\chi(A,B) = \begin{cases} 0, & B \in V(A,\Phi) \\ +\infty, & B \notin V(A,\Phi) \end{cases},$$

определяющий область допустимых значений; $Q(B)$ — критерий качества проекции (projection quality function), характеризующий ее принадлежность модели $M; \alpha \geqslant 0,$ — структурирующий параметр, обеспечивающий компромисс между критериями соответствия и качества. Соответствующий морфологический проектор будет проектором на базе структурирующих критериев и параметров

$$\psi(A,\Phi) = \Pr(A,J,\chi,\alpha,Q) = \\ = \arg\min_{B \in \Omega} \Phi(A,B,J,\chi,\alpha,Q). \quad (9)$$

В случае, когда $\chi(A,B) \equiv 0$ (ОДЗ неограниченна), критерий (8) принимает упрощенный вид

$$\Phi(A,B) = J(A,B) + Q(B).$$

Критерий (8) является хорошо определенным, если требования соответствия и качества оказываются противоположными:

$$\forall A \in \Omega,\ B \in V(A,\Phi),\ \Phi(A,B) < \Phi(A,A) \Rightarrow \\ \Rightarrow J(A,B) \geqslant J(A,A),\ Q(B) < Q(A),$$

то есть лучшее соответствие данным может компенсировать худшее качество фильтрации, и наоборот. Уровень равновесия здесь устанавливает параметр α, обладающий, как доказано в [4], следующим важным свойством: с увеличением значения структурирующего параметра α в выражении (8) морфологическая сложность модели (3), которую определяет проектор (9), монотонно убывает. Таким образом, структурирующий параметр также можно назвать параметром морфологической сложности модели. Более того, в силу этого свойства методика построения морфологических спектров, ранее предложенная Maragos [5] в рамках ММ Серра, может быть обобщена на случай вычисления критериальных морфологических спектров по параметру морфологической сложности:

$$\mathrm{Sp}(A,\alpha) = -\frac{\partial \| \Pr(A,J,\chi,\alpha,Q) \|}{\partial \alpha}.$$

Наконец, для любого конкретного образа A однозначно определяется коэффициент максимальной морфологической сложности по отношению к $\{J,\chi,Q\}$:

$$\alpha_{\max}(A) = \max\{\alpha \geqslant 0 \colon A = \Pr(A,J,\chi,\alpha,Q)\}.$$

Рассмотрим теперь различные способы и достаточные условия построения проективных операторов на базе критериев типа (8).

Достаточные условия проективности и типы морфологических операторов

Проекторы минимального расстояния. Пусть критерий соответствия $J(A,B)$ обладает свойствами расстояния:

$$\forall A,B,C \in \Omega \colon J(A,B) \geqslant 0,\ J(A,A) = 0, \\ J(A,B) = J(B,A), \\ J(A,B) + J(B,C) \geqslant J(A,C).$$

Назовем его критерием минимального расстояния. Пусть, кроме того, критерий $\Phi(A,B)$ (9) является монотонным в смысле условия (7). В [4] доказано, что монотонные критерии минимального расстояния (4, 6, 7) определяют морфологический проектор (9).

Проекторы максимума обобщенной нормы проекции. Рассмотрим критерий вида

$$\Phi(A,B) = -J(B) + \chi(A,B) + \alpha \times Q(B), \qquad (10)$$

отличающийся от (8) тем, что $J(B)$ не зависит от A, но при этом по-прежнему

$$\forall A \in \Omega,\ \forall B \in V(A,\Phi):\ J(A) \geqslant J(B). \qquad (11)$$

Назовем критериями максимума обобщенной нормы проекции все критерии, которые можно записать в виде (10, 11). Определим также условие уменьшающей монотонности ОДЗ

$$\exists A, \exists B\colon \chi(A,B) = +\infty \Leftrightarrow$$
$$\Leftrightarrow \forall A \in \Omega, \forall B \in V(A,\Phi)\colon V(B,\Phi) \subset V(A,\Phi). \qquad (12)$$

Для критериев с монотонно уменьшающейся ОДЗ (12) в [4] было доказано, что требование максимума обобщенной нормы (10, 11) определяет морфологический проектор (9). Частными случаями проекторов максимальной обобщенной нормы проекции являются, например, морфологические фильтры opening и closing в математической морфологии Серра [1].

Квазимонотонные проекторы. Назовем эффективным подмножеством области допустимых значений $V(A,\Phi)$ такое множество $U(A,\Phi) \subseteq V(A,\Phi)$, что

$$\forall B \in V(A,\Phi),\ B \notin U(A,\Phi)\colon$$
$$\exists C \in U(A,\Phi),\ \Phi(A,C) < V(A,\Phi). \qquad (13)$$

Понятие эффективного подмножества ОДЗ позволяет сформулировать следующее расширенное условие квазимонотонности ОДЗ:

$$\forall A \in \Omega,\ \forall B \in V(A,\Phi)\colon\ U(B,\Phi) \subseteq V(A,\Phi). \qquad (14)$$

Критерии, для которых выполняется условие (14), назовем квазимонотонными по ОДЗ или просто квазимонотонными. В [4] доказано, что для критериев, квазимонотонных по ОДЗ (13, 14), требование максимума обобщенной нормы проекции (10, 11) определяет проектор (9). Переход от монотонных к квазимонотонным критериям позволяет, в частности, обосновать существование широкого класса проективных морфологий на базе структурной интерполяции [4].

Проекторы на базе предиката качества и выпуклого критерия соответствия. Рассмотрим теперь критерий (8), предполагая, что $\chi(A,B)$ и $Q(B)$ являются штрафными предикатами (принимают значения на множестве $\{0, +\infty\}$), а критерий $J(A,B)$ является хорошо определенной функцией соответствия, то есть удовлетворяет условию

$$\forall A, B \in \Omega,\ A \neq B \Rightarrow J(A,A) < J(A,B). \qquad (15)$$

Легко показать, что если $Q(B)$ является штрафным предикатом, а критерий $J(A,B)$ является хорошо определенной функцией соответствия (15), то критерий (8) определяет оператор, обладающий проективными свойствами. Проекторы такого типа рассматриваются в рамках морфологического анализа изображений Пытьева [2].

Проекторы на базе предикатов качества и соответствия. Рассмотрим теперь критерий (8), предполагая, что все входящие в него критерии $J(A,B)$, $\chi(A,B)$ и $Q(B)$ являются штрафными предикатами. Тогда критерий (8) и оператор (9) принимают вид

$$\Phi(A,B) = \chi(A,B) + Q(B), \qquad (16)$$

$$\Pr(A,\chi,Q) = \arg\min_{B\in\Omega} \Phi(A,B,\chi,Q).$$

Предикату $\Phi(A,B)$ соответствует область допустимых значений $V(A,\Phi)$. Легко убедиться, что критерий (16) является хорошо определенным (5) и задает морфологический проектор (9), в том и только в том случае, когда для любого проецируемого образа A область допустимых значений $V(A,\Phi)$ содержит не более одного образа.

Примеры морфологической фильтрации на базе критериев

Описанные теоретические результаты можно проиллюстрировать примерами построения простейших операторов морфологической фильтрации и сегментации, реализуемых методом динамического программирования (операторов ДП-фильтрации и ДП-сегментации). На рис. 1, 2 представлены примеры одномерной проективной монотонной морфологической фильтрации. На рис. 3, 4 — примеры среднеквадратичной и монотонной проективной морфологической сегментации. На рис. 5 показан пример проективной морфологической сегментации двумерной кривой (контура двумерного бинарного образа) на базе кусочно-линейной интерполяции. На всех рисунках хорошо видна зависимость сложности формируемого морфологического описания данных от значений параметра α.

Выводы

В статье дано краткое описание критериальной проективной морфологии. Рассмотрены различные схемы построения критериальных проективных операторов анализа цифровых данных.

Наиболее общими методами алгоритмической реализации описанных критериальных морфологических проекторов являются методы рекурсивного (логического и динамического) программирования. К сожалению, многие важные с практической точки зрения типы данных (например, характерные для изображений двумерные прямоугольные решетки) не допускают решения этими методами, так

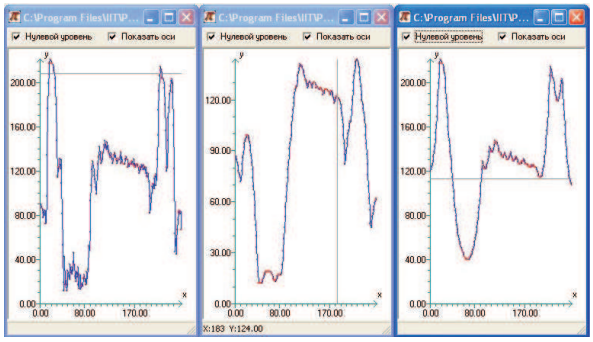


Рис. 1. Слева исходная функция, далее результаты ДП-фильтрации: DP-Open (α=200) и DP-Close (α=200).

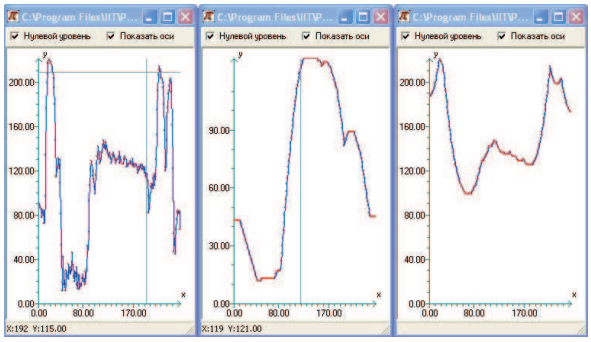


Рис. 2. Слева исходная функция, далее результаты ДП-фильтрации: DP-Open ($\alpha = 10^3$) и DP-Close ($\alpha = 10^3$).

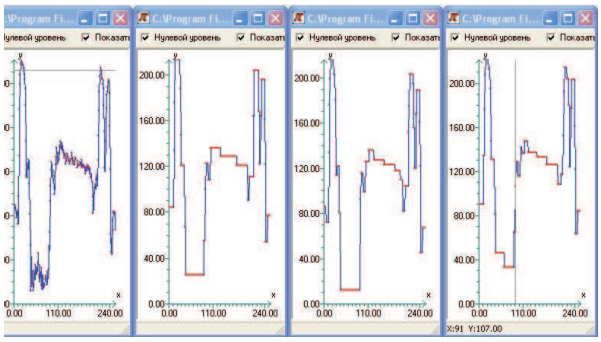


Рис. 3. Слева исходная функция, далее результаты применения операторов ДП-сегментации: DP-LSE ($\alpha = 500$), DP-Open ($\alpha = 10^4$), DP-Close ($\alpha = 10^4$).

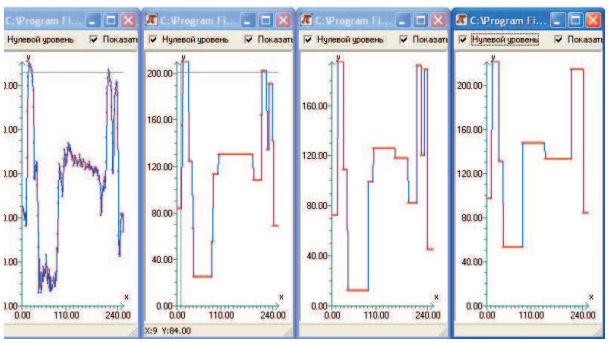


Рис. 4. Слева исходная функция, далее результаты применения операторов ДП-сегментации: DP-LSE ($\alpha = 2000$), DP-Open ($\alpha = 10^5$), DP-Close ($\alpha = 10^5$).

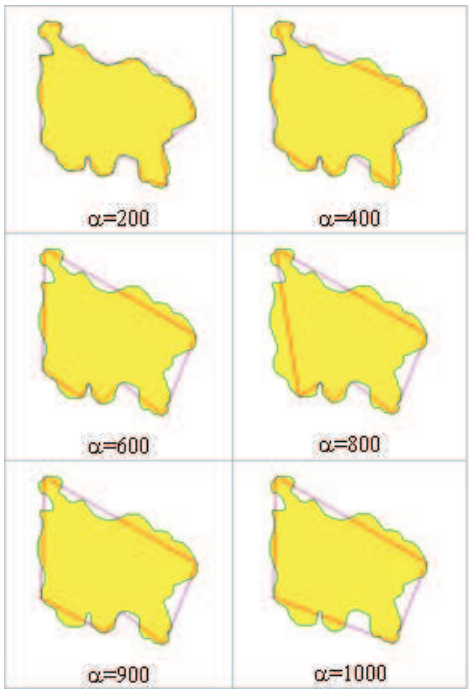


Рис. 5. Пример критериальной морфологической сегментации контура двумерного бинарного образа при различных значениях структурирующего параметра α.

как рекурсивное программирование работает лишь с ациклическими структурами данных. Поэтому основные проблемы и сложности практического использования описанного морфологического формализма в задачах анализа изображений будут, по-видимому, связаны с поиском эффективных способов представления двумерных данных ациклическими структурами.

Литература

[1] *Serra J.* Image Analysis and Mathematical Morphology. — London: Academic Press, 1982.

[2] *Пытьев Ю. П.* Морфологический анализ изображений // Доклады АН СССР, 1983. — Т. 269, № 5. — С. 1061–1064.

[3] *Визильтер Ю. В., Желтов С. Ю.* Проективные морфологии и их применение в структурном анализе цифровых изображений // Изв. РАН. ТиСУ. — 2008. — № 6. — С. 113–128.

[4] *Визильтер Ю. В.* Обобщенная проективная морфология // Компьютерная оптика. — 2008. — Т. 32. № 4. — С. 384–399.

[5] *Maragos P.* Pattern Spectrum, Multiscole Shape Representation // IEEE Trans.on pattern analysis, machine intelligence. — 1989. — Vol.II, № 7.

О распознавании образов в пространстве пирамидальных представлений*

Ганебных С. Н., Ланге М. М.
lange_mm@ccas.ru
Москва, ВЦ РАН

Рассматривается задача распознавания двумерных полутоновых объектов в пространстве их иерархически структурированных представлений. Предложен способ обучения классификатора, который состоит в отборе эталонов, порождающих модифицированную ε-сеть с многоуровневым разрешением. Разработан быстрый алгоритм поиска решения по критерию голосования. Приведены экспериментальные результаты распознавания жестов и подписей.

Постановка задачи

Рассматриваются объекты, заданные двумерными телами на полутоновых изображениях. Пусть $\mathbf{A}^L$ — множество всевозможных объектов, в котором каждый объект представлен пирамидой

$$A^L = \{a^l\}_{l=0}^L,$$

содержащей $L+1$ уровней разрешения, где a^l — представление l-го уровня [1]. Для любой пары пирамидальных представлений $(A^L, \widetilde{A}^L) \in \mathbf{A}^L$ определено семейство мер различия

$$\mathbf{D} = \big\{D_l(A^L, \widetilde{A}^L) \geqslant 0\big\}_{l=0}^L, \qquad (1)$$

в котором $D_l(A^L, \widetilde{A}^L) = D(a^l, \widetilde{a}^l)$ — мера различия пары представлений $a^l \in A^L$ и $\widetilde{a}^l \in \widetilde{A}^L$. Множество $\mathbf{A}^L$ содержит объекты, принадлежащие $K+1$ классам

$$\mathbf{A}^L = \big\{\mathbf{A}_i^L\big\}_{i=0}^K,$$

где $\mathbf{A}_i^L$ — i-й класс. Классы с ненулевым номером содержат семантически однородные объекты, идентифицируемые номером соответствующего класса, а нулевой класс включает все прочие объекты.

Пусть $\mathbf{B}^L \subset \mathbf{A}^L$ — обучающее множество объектов, включающее классы

$$\mathbf{B}_i^L = \big\{B_{ij}^L\big\}_{j=1}^{m_i}, \quad i = 1, \ldots, K, \qquad (2)$$

так что $\mathbf{B}^L = \bigcup\limits_{i=1}^K \mathbf{B}_i^L$, $\|\mathbf{B}^L\| = \sum\limits_{i=1}^K m_i = M$, и во всех классах $\mathbf{B}_i^L \subset \mathbf{B}^L$ выбраны группы эталонов

$$\widehat{\mathbf{B}}_i^L = \big\{\widehat{B}_{ij}^L\big\}_{j=1}^{\widehat{m}_i}, \quad i = 1, \ldots, K, \qquad (3)$$

объединение которых порождают множество эталонов $\widehat{\mathbf{B}}^L = \bigcup\limits_{i=1}^K \widehat{\mathbf{B}}_i^L \subset \mathbf{B}^L$, где $\widehat{m}_i \leqslant m_i$ и $\|\widehat{\mathbf{B}}^L\| = \sum\limits_{i=1}^K \widehat{m}_i = \widehat{M} \leqslant M$. Каждый эталон $\widehat{B}_{ij} \in \widehat{\mathbf{B}}^L$ на l-м уровне разрешения имеет сферу влияния с параметром $D_l^*(\widehat{B}_{ij}^L) \geqslant 0$, который вычисляется по мере $D_l(A^L, \widehat{B}_{ij}^L)$ из семейства (1). Совокупность параметров эталона $\widehat{B}_{ij}$ для всех уровней разрешения образует множество

$$\mathbf{D}^*\big(\widehat{B}_{ij}^L\big) = \big\{D_l^*(\widehat{B}_{ij}^L)\big\}_{l=0}^L. \qquad (4)$$

Используя семейство мер (1) и семейство множеств (4) для всех $\widehat{B}_{ij}^L \subset \widehat{\mathbf{B}}^L$, на любом уровне разрешения $l = 0, \ldots, L$ вводится функция сходства объекта $A^L \in \mathbf{A}^L$ с группой эталонов $\widehat{\mathbf{B}}_i^L \subset \mathbf{B}_i^L$:

$$\mu_l(A^L, \widehat{\mathbf{B}}_i^L) = \sum_{j=1}^{\widehat{m}_i} \Big[D_l(A^L, \widehat{B}_{ij}^L) \leqslant D_l^*(\widehat{B}_{ij})\Big] e^{-sD_l(A^L, \widehat{B}_{ij}^L)}, \qquad (5)$$

где $[z]$ — индикатор z, а $s \geqslant 0$ — параметр весовой функции индикатора. В терминах соотношений (1)–(5) критерий классификации (распознавания) объектов сводится к нахождению номера класса

$$n = \arg\max_{i=1}^K \mu_L(A^L, \widehat{\mathbf{B}}_i^L) \Big[\max_{i=1}^K \mu_L\big(A^L, \widehat{\mathbf{B}}_i^L\big) > 0\Big]. \qquad (6)$$

В настоящей работе предлагается процедура отбора множества эталонов $\widehat{\mathbf{B}}^L$, которое на обучающем множестве пирамидальных представлений $\mathbf{B}^L$ образует модифицированную многоуровневую ε-сеть [2] по семейству мер вида (1), и быстрый алгоритм поиска решения по критерию (6) с вычислительной сложностью порядка $O(K \log K)$.

Класс объектов и инвариантность их представлений

Ограничения, налагаемые на объекты, и способ построения их пирамидальных представлений с многоуровневым разрешением рассмотрены авторами в работах [3, 4].

Определение 1. *Класс допустимых объектов составляют двумерные тела (возможно многосвязные) с однородной или неоднородной яркостной окраской, имеющие однозначно идентифицируемую систему собственных координат.*

В [4] это утверждение сформулировано в терминах ограничений на центральные моменты объекта,

*Работа выполнена при финансовой поддержке РФФИ, проект №09-01-00573-а.

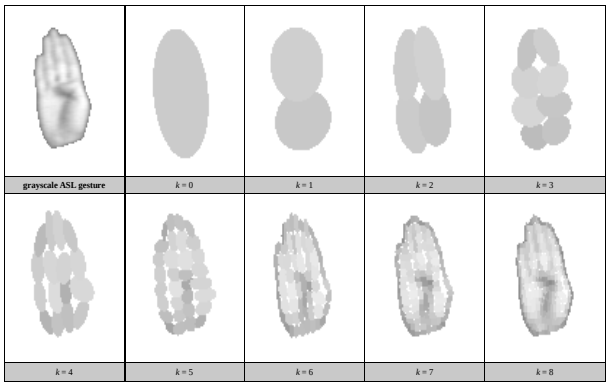


Рис. 1. Примеры пирамидальных представлений жеста руки с уровнями разрешения $l = 0, \ldots, 8$.

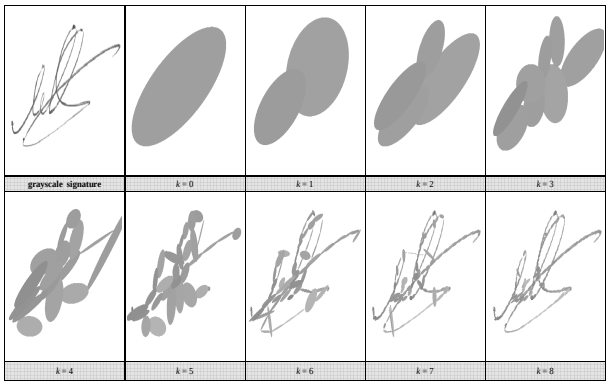


Рис. 2. Примеры пирамидальных представлений подписи с уровнями разрешения $l = 0, \ldots, 8$.

образующие матрицу тензора инерции двумерного тела, и в терминах формальных требований к асимметрии объекта. В этой же работе дан алгоритм построения представления любого объекта из указанного класса набором эллиптических примитивов, образующих бинарное дерево. Дерево примитивов, содержащее уровни с номерами $l = 0, \ldots, L$, дает пирамиду из $L+1$ представлений с многоуровневым разрешением. Разрешение L-го уровня определяется числом примитивов этого уровня, которое в общем случае не превосходит 2^l. Пирамида представлений обладает свойством инвариантности, которое сформулировано в следующем утверждении.

Утверждение 1. *Если* $(A^L_\Delta, \widetilde{A}^L_\Delta) \in \boldsymbol{A}^L_\Delta$ *— пара пирамидальных представлений для объектов, которые могут быть совмещены преобразованиями поворота, смещения, изменения масштаба и уровня яркости, где* Δ *— размер пикселя на изображении объекта, то* $\lim\limits_{\Delta\to 0} D_l(A^L_\Delta, \widetilde{A}^L_\Delta) = 0$ *при всех* $l = 0, \ldots, L$.

Класс допустимых объектов, удовлетворяющих Определению 1, включает областные (region-based) и линейчатые (line-based) объекты. Алгоритм построения древовидных построений универсален и дает пирамидальные представления, которые с нарастающим разрешением воспроизводят форму и яркостную окраску объектов от различных источников. Примеры пирамидальных представлений жеста руки и подписи, содержащие девять уровней разрешения показаны на рис. 1 и рис. 2.

Обучение классификатора

Обучение включает две процедуры. Первая состоит в получении для всех объектов $B^L_{ij} \in \mathbf{B}^L$ оценок параметров $D^*_l(B^L_{ij})$, которые при $l = 0, \ldots, L$ образуют семейство множеств вида (4); вторая — в отборе групп эталонов вида (3), которые образуют множество $\widehat{\mathbf{B}}^L \subset \mathbf{B}^L$.

Оценки параметров $D^*_l(B^L_{ij})$ строятся по каждому классу объектов $\mathbf{B}^L_i \subset \mathbf{B}^L$, $i = 1, \ldots, K$, путем просмотра множества значений $\mathbf{D}_L(B^L_{ij}) = \{D_L(B^L_{ik}, B^L_{ij})\}_{k=1}^{m_i}$ при наибольшем уровне разрешения L. Для фиксированного $B^L_{ij} \in \mathbf{B}^L_i$ и текущего значения $\widetilde{D}_L(B^L_{ij}) \in \mathbf{D}_L(B^L_{ij})$, вычисляются индикаторы ошибок

$$\lambda_{\mathrm{FAR}} = \left[B^L \notin \mathbf{B}^L_i\right] \left[D_L(B^L, B^L_{ij}) \leqslant \widetilde{D}_L(B^L_{ij})\right],$$

$$\lambda_{\mathrm{FRR}} = \left[B^L \in \mathbf{B}^L_i\right] \left[D_L(B^L, B^L_{ij}) > \widetilde{D}_L(B^L_{ij})\right],$$

которые соответствуют ложному распознаванию (FAR) и ложному отказу (FRR), и по всем объектам обучающего множества вычисляется эмпирический риск

$$\eta(\mathbf{B}^L, \widetilde{D}_L(B^L_{ij})) = \frac{1}{M} \sum_{B^L \in \mathbf{B}^L} (\lambda_{\mathrm{FAR}} + \lambda_{\mathrm{FRR}}). \quad (7)$$

Оценка параметра $D^*_L(B^L_{ij})$ выбирается из условия минимизации риска (7):

$$D^*_L(B^L_{ij}) = \operatorname*{arg\,min}_{\widetilde{D}_L(B^L_{ij}) \in \mathbf{D}_L(B^L_{ij})} \eta(\mathbf{B}^L, \widetilde{D}_L(B^L_{ij})). \quad (8)$$

Способ построения оценки (8) гарантирует существование объекта $B^L_{ik^*} \in \mathbf{B}^L_i$, такого что $D^*_L = D_L(B^L_{ik^*}, B^L_{ij})$. Поэтому оценки параметров $D^*_l(B^L_{ij})$ при уровнях разрешения $l = 0, \ldots, L-1$ следуют из условия $D^*_l = D_l(B^L_{ik^*}, B^L_{ij})$.

Группы эталонов вида (3) отбираются во всех классах вида (2) независимо. Используя меру различия $D_L(B^L, \widetilde{B}^L)$ в пространстве представлений с наибольшим разрешением ($l = L$), в каждом классе $\mathbf{B}^L_i \subset \mathbf{B}^L$, $i = 1, \ldots, K$, тестируются всевозможные группы $\widetilde{\mathbf{B}}^L_i = \{\widetilde{B}^L_{ij}\}_{j=1}^{\widetilde{m}_i}$ в порядке уменьшения их размера $\widetilde{m}_i$, путем предъявления каждой группе всех объектов обучающего множества $\mathbf{B}^L$. В результате тестирования в рассматриваемом классе отбирается группа наименьшего размера, которая обеспечивает суммарную долю ошибочных решений (FAR + FRR) не более заданной допустимой величины $\varepsilon > 0$.

Формально для каждой текущей группы $\widetilde{\mathbf{B}}_i^L \subset \mathbf{B}_i^L$ вычисляются индикаторы ошибочных решений

$$\widetilde{\lambda}_{\mathrm{FAR}} = \left[B^L \notin \mathbf{B}_i^L\right] \left[\mu_L(B^L, \widetilde{\mathbf{B}}_i^L) > 0\right],$$

$$\widetilde{\lambda}_{\mathrm{FRR}} = \left[B^L \in \mathbf{B}_i^L\right] \left[\mu_L(B^L, \widetilde{\mathbf{B}}_i^L) = 0\right],$$

и эмпирический риск

$$\widetilde{\eta}(\mathbf{B}^L, \widetilde{\mathbf{B}}_i^L) = \frac{1}{M} \sum_{B^L \in \mathbf{B}^L} (\widetilde{\lambda}_{\mathrm{FAR}} + \widetilde{\lambda}_{\mathrm{FRR}}). \qquad (9)$$

Группа $\widehat{\mathbf{B}}_i$ наименьшего размера $\widehat{m}_i(\widehat{\mathbf{B}}_i^L) = \|\widehat{\mathbf{B}}_i^L\|$ выбирается из условия минимизации числа объектов $\widetilde{m}_i(\widetilde{\mathbf{B}}_i^L) = \|\widetilde{\mathbf{B}}_i^L\|$ при ограничении сверху на величину риска (9):

$$\widehat{\mathbf{B}}_i^L = \underset{\widetilde{\mathbf{B}}_i^L : \widetilde{\eta}(\mathbf{B}^L, \widetilde{\mathbf{B}}_i^L) \leqslant \varepsilon}{\arg\min} \widetilde{m}_i(\widetilde{\mathbf{B}}_i^L). \qquad (10)$$

Допустимые значения ε в (10) ограничены снизу величиной $\varepsilon_{\min}(\mathbf{B}^L) = \max\limits_{i=1}^{K} \widetilde{\eta}(\mathbf{B}^L, \mathbf{B}_i^L)$, которая зависит от выбранного обучающего множества $\mathbf{B}^L$.

При риске (9), вычисляемом по мере $D_L(B^L, \widetilde{B}^L)$, критерий (10) дает для минимальной группы эталонов $\widehat{\mathbf{B}}_i^L$ представление $\widehat{\mathbf{b}}_i^L = \{\widehat{b}_{ij}^L\}_{j=1}^{\widehat{m}_i}$ с наибольшим разрешением ($l = L$). При $l = 0, \ldots, L-1$, представления $\widehat{\mathbf{b}}_i^l = \{\widehat{b}_{ij}^l\}_{j=1}^{\widehat{m}_i}$ этой же группы строятся "проецированием" представления $\widehat{\mathbf{b}}_i^L$ в пространство представлений с соответствующим уровнем разрешения. Последовательность $\widehat{\mathbf{B}}_i^L = \{\widehat{\mathbf{b}}_i^l\}_{l=0}^{L}$ образует группу эталонов с многоуровневым разрешением для i-го класса обучающего множества. Объединение таких групп по всем классам $i = 1, \ldots, K$ дает множество

$$\widehat{\mathbf{B}}^L = \left\{\widehat{\mathbf{B}}_i^L = \{\widehat{\mathbf{b}}_i^l\}_{l=0}^{L}\right\}_{i=1}^{K}, \qquad (11)$$

которое содержит $\widehat{M} = \|\widehat{\mathbf{B}}^L\| = \sum\limits_{i=1}^{K} \widehat{m}_i$ эталонов. Множество (11) порождает на обучающем множестве $\mathbf{B}^L$ модифицированную ε-сеть по семейству мер вида (1), которая содержит $L+1$ уровней разрешения.

Быстрый алгоритм поиска решения

Введем множества

$$\widehat{\mathbf{b}}^l = \left\{\widehat{\mathbf{b}}_i^l = \{\widehat{b}_{ij}^l\}_{j=1}^{\widehat{m}_i}\right\}_{i=1}^{K}, \qquad (12)$$

содержащие представления всех $\widehat{M}$ эталонов на уровнях разрешения $l = 0, \ldots, L$. Последовательность множеств (12) позволяет представить множество эталонов (11) в виде многоуровневой базы

$$\widehat{\mathbf{B}}^L = \left\{\widehat{\mathbf{b}}^l\right\}_{l=0}^{L}, \qquad (13)$$

в которой каждый l-й уровень $\widehat{\mathbf{b}}^l$ содержит все K групп эталонов с соответствующим уровнем разрешения.

Введем функцию

$$K_l = \lfloor K 2^{-\alpha l} \rfloor, \qquad (14)$$

которая определяет число анализируемых классов K_l на последовательных уровнях $l = 0, \ldots, L-1$ базы эталонов (13). Параметр функции (14): $\alpha = \frac{1}{L}\log_2(K/K_L)$ при заданном $K_L \geqslant 1$. С учетом (14) быстрый алгоритм поиска решения для объекта A^L по критерию (6) выполняется с помощью следующей итеративной процедуры. На каждом текущем уровне $l = 0, \ldots, L-1$ анализируется сегмент $\widehat{\mathbf{b}}_{\mathrm{seg}}^l \subset \widehat{\mathbf{b}}^l$ базы эталонов (13), который содержит K_l групп, и в нем отбираются K_{l+1} групп $\widehat{\mathbf{b}}_i^l$ с наибольшими значениями функции сходства $\mu_l(A^L, \widehat{\mathbf{B}}_i^L) = \mu(a^l, \widehat{\mathbf{b}}_i^l)$ вида (5) по мере $D_l(A^L, \widehat{B}_{ij}^L) = D(a^l, \widehat{b}_{ij}^l)$. На следующем $(l+1)$-м уровне группы, отобранные на l-м уровне, порождают сегмент $\widehat{\mathbf{b}}_{\mathrm{seg}}^{l+1} \subset \widehat{\mathbf{b}}^{l+1}$ с $(l+1)$-м уровнем разрешения. На нулевом уровне ($l = 0$): $\widehat{\mathbf{b}}_{\mathrm{seg}}^0 = \widehat{\mathbf{b}}^0$, $K_0 = K$. На последнем уровне ($l = L$) сегмент $\widehat{\mathbf{b}}_{\mathrm{seg}}^L$ содержит K_L групп вида $\widehat{\mathbf{b}}_n^L = \{\widehat{b}_{nj}^L\}_{j=1}^{\widehat{m}_i}$, которые отбираются на $L-1$ уровне и берутся с L-м уровнем разрешения. Решение для объекта A^L определяется номером n согласно (6).

Вычислительная сложность алгоритма поиска решения определяется числом обрабатываемых вершин (примитивов) в многоуровневых представлениях эталонов множества $\widehat{\mathbf{B}}^L$. Учитывая, что l-й уровень в пирамидальном представлении любого объекта содержит 2^l примитивов, при $L \leqslant \log_2 K$ ($\alpha \geqslant 1$ в (14)), вычислительная сложность рассмотренного быстрого алгоритма имеет порядок $O(K \log K)$. Для сравнения сложность алгоритма поиска решения на основе перебора всех эталонов имеет порядок $O(K^2)$.

Экспериментальные результаты

Предложенный классификатор опробован в экспериментах по распознаванию жестов и подписей, взятых из баз данных, опубликованных в [5] и [6]. База жестов [5] представлена изображениями размера 256×256 пикселей и числом уровней яркости 256. Множество жестов содержало $K^{\mathrm{ges}} = 25$ классов, которые соответствуют буквам латинского алфавита. Общее число жестов 750 = 30×25 (по 30 реализаций в каждом классе) было разбито на обучающее и тестовое множества, по 375 = 15×25 объектов в каждом. Отбор эталонов проводился с наименьшим параметром $\varepsilon_{\min}^{\mathrm{ges}} = 0{,}005$ для обучающего множества жестов. Весовые коэффициенты в функции сходства (5) взяты с параметром $s = 0$. Результаты распознавания объектов тестового множества в виде зависимостей доли ошибочных

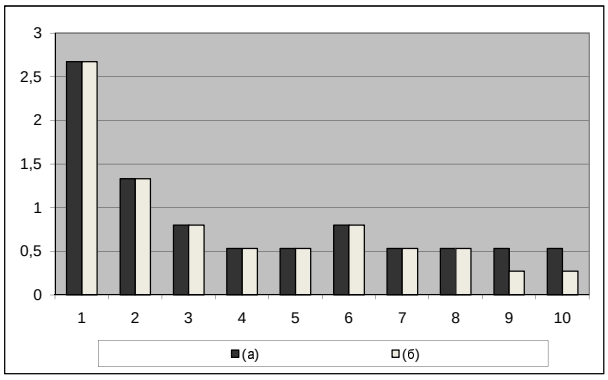


Рис. 3. Доля ошибок распознавания жестов (в %) от уровня максимального разрешения для быстрого (а) и переборного (б) алгоритмов.

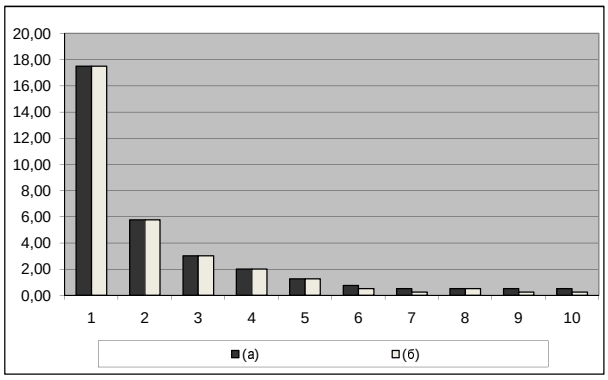


Рис. 4. Доля ошибок распознавания подписей (в %) от уровня максимального разрешения для быстрого (а) и переборного (б) алгоритмов.

решений от величины максимального уровня разрешения даны на рис. 3.

База исходных подписей [6] представлена сигналами, которые были преобразованы в изображения размера 512×512 пикселей, содержащие 256 уровней яркости. Число классов $K^{\mathrm{sig}} = 40$ определялось числом персон. Общее число подписей $800 = 20{\times}40$ (по 20 реализаций от каждой персоны) было разбито на эквивалентные обучающее и тестовое множества по $400 = 10{\times}40$ объектов в каждом. Обучение проводилось с наименьшим параметром $\varepsilon^{\mathrm{sig}}_{\min} = 0{,}005$ для обучающего множества подписей. Параметр функции сходства $s = 0$. Результаты распознавания представлены графиками на рис. 4.

Графики на рис. 3 и рис. 4 демонстрируют для обоих алгоритмов практически совпадающие показатели качества распознавании при существенном вычислительном выигрыше быстрого алгоритма по сравнению с алгоритмом перебора всех эталонов.

Выводы

Рассмотрен класс двумерных полутоновых объектов с идентифицируемой системой собственных координат и способ построения пирамидальных представлений таких объектов на основе их рекурсивной декомпозиции и аппроксимации эллиптическими примитивами. Предложено семейство аддитивно вычислимых мер различия на множестве пирамидальных представлений с многоуровневым разрешением. Используя введенные меры, разработана процедура обучения классификатора на основе принятия решения по критерию голосования. Процедура обучения сведена к построению модифицированной ε-сети эталонов с многоуровневым разрешением. При числе классов K, иерархическая структура сети эталонов позволила ускорить процедуру поиска решения приблизительно в $K/\log_2 K$ раз по сравнению с переборной процедурой. Разработанный классификатор продемонстрировал возможность распознавания жестов и подписей в пространстве унифицированных пирамидальных представлений с вероятностью ошибки порядка 0,01. В работе [7] показана возможность уменьшения доли ошибочных решений при распознавании подписей до величины 0,003 за счет объединения рассмотренного классификатора с другими классификаторами. Планируется обобщение рассмотренного иерархического пространства представлений для объектов, заданных многоканальными изображениями.

Литература

[1] *Rosenfeld A.* Multiresolution Image Processing and Analysis. — Berlin: Springer, 1984.

[2] *Колмогоров А. Н., Тихомиров В. М.* Эпсилон-энтропия и эпсилон-емкость множеств в функциональных пространствах. Теория информации и теория алгоритмов. — М.: Наука, 1987.

[3] *Lange M., Ganebnykh S., Lange A.* Moment-based Pattern Representation Using Shape and Grayscale Features. // Lecture Notes in Computer Science, Vol. 4477. — Berlin: Springer, 2007. — P. 523–530.

[4] *Ганебных С. Н., Ланге М. М.* Древовидное представление образов для распознавания полутоновых объектов. // Труды Вычислительного центра им. А. А Дородницына РАН, (отдельный выпуск). — М.: ВЦ РАН, 2007. — 32 с.

[5] Thomas Moeslund's gesture recognition database. — `vision.auc.dk/~tbm/Gestures/database.html` — 2002.

[6] First International Signature Verification Competition (SVC 2004). — `www.cs.ust.hk/svc2004/index.html`.

[7] *Mottl V., Lange M., Sulimova V., Yermakov A.* Signature Verification Based on Fusion of On-line and Off-line Kernels. // Proceedings of the 19th International Conference on Pattern Recognition (ICPR-2008, Tampa, USA). — Los Alamitos: IEEE CS Press, 2008.

Выделение характерных признаков лиц на цифровых изображениях с использованием знакового представления*

Гончаров А. В., Губарев В. В.
ag.tsure@gmail.com, vlad.gubarev@gmail.com
Таганрог, Лаборатория математических методов искусственного интеллекта

В работе приведен обзор современных методов выделения характерных признаков лиц на изображениях. Предложена модификация одного из методов, основанная на знаковом представлении изображения, позволяющем существенно снизить вычислительную сложность алгоритма. Получены оценки качества рассматриваемых методов.

Методы и алгоритмы автоматического обнаружения и распознавания лиц используются в широком спектре современных систем компьютерного зрения: биометрическая идентификация, зрение роботов, компьютерная анимация, видеоконференции, интернет-поиск, охранные системы, и т. д.

Несмотря на то, что рассмотренные технологии уже существуют и активно внедряются во многих сферах своего применения, показатели качества работы говорят о том, что данное направление целесообразно развивать дальше. Кроме точности результатов, важным показателем является скорость обработки информации.

Объектом исследования данной работы являются цифровые изображения лиц. Цель работы состоит в повышении точности и производительности имеющихся алгоритмов и методов, а так же разработке новых. В качестве предмета исследования выступают характерные антропометрические элементы лиц на изображениях (рис. 1): брови, глаза, нос, губы, контур лица.

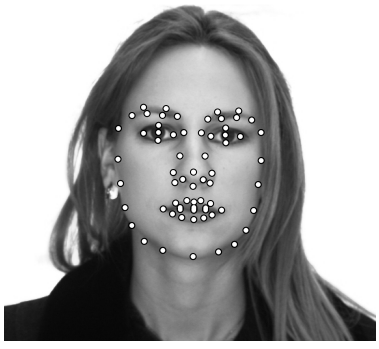


Рис. 1. Характерные элементы (исходное изображение взято из базы AR [4]).

Обзор существующих методов

Простые методы выделения характерных признаков изображений попадают под общую категорию *сегментация изображений*. Простейший подход заключается в перемещении некоторого шаблона по изображению с целью поиска совпадения. Разнообразие текстур и другие особенности изображений делают такой подход неудачным. Существует большое количество других методов, таких как выделение границ и линий, которые описаны в большинстве книг по обработке изображений.

Принципиально другим подходом является использование *деформируемых моделей*. Идея таких методов заключается в «набрасывании» некоторой модели на характерные признаки. Модель взаимодействует с изображением и «выравнивается» по искомому признаку. Существует большое количество методов, работающих по этому принципу. Часть из них основана на *активных контурах*, которые выступают в качестве деформируемой модели. Активный контур представляет собой параметрическую кривую $v(s) = \big(x(s), y(s)\big)$, обладающую собственной внутренней энергией E_{int}. На контур действует некоторая внешняя энергия E_{ext}. Идея заключается в поиске формы контура, при которой достигается минимум общей энергии:

$$E = \int_0^1 \big(E_{\text{int}}\left(v(s)\right) + E_{\text{ext}}\left(v(s)\right)\big)\, ds. \qquad (1)$$

В данной работе за основу была взята модель Active Shape Model (ASM). Эта модель является «классическим» примером методов выделения признаков на изображениях [1]. Существует множество модификаций метода ASM. Одна из таких модификаций предложена и подробно рассмотрена в [5]. Именно на ней и основана данная работа.

Модель ASM

Рассмотрим подробнее модель ASM.

Как отмечалось ранее, в основе модели лежит контур. Известно, что цифровое изображение представляет собой дискретный набор данных, заданный на равномерной сетке. Исходя из этого, и основываясь на теоретическом представлении активного контура, в дальнейшем под контуром будем понимать конечный дискретный упорядоченный набор точек (x, y) двумерного пространства.

Для контуров введено понятие расстояния между ними и определены операции поворота, масштабирования, перемещения.

*Работа выполнена при финансовой поддержке РФФИ, проект № 08-07-00129, № 07-07-00067

Алгоритм состоит из двух этапов: обучение и непосредственное применение. Обучение производится на выборке, состоящей из изображений с лицами и вручную указанными положениями искомых точек для каждого изображения.

После того, как алгоритм пройдет обучение, его можно применять для поиска характерных элементов. Основная идея метода заключается в поиске наилучшего положения для каждой отдельно взятой точки контура, а затем корректировке всего контура, исходя из расположения точек относительно друг друга. Данный метод базируется на двух моделях:

— *Модель поиска точек по шаблону.* В рамках данной модели для каждой точки контура ищется наилучшее местоположение путем сравнения значений яркости в окрестности этой точки с некоторым шаблоном. Для этой цели используется метрика Махаланобиса:

$$\rho^2(x, \bar{x}) = (x - \bar{x})^{\intercal} \Sigma^{-1} (x - \bar{x}),$$

где x — рассматриваемый вектор; $\bar{x}$ — средний вектор; Σ — ковариационная матрица. В соответствии с формулой (1), данная модель описывает влияние внешней энергии на контур со стороны изображения. Отметим, что для вычисления этой метрики требуется $O(n^2)$ операций.

— *Модель формы.* Эта модель описывает взаимное расположение всех точек в контуре и выполняет их корректировку. В качестве математической основы модели используется метод главных компонент (PCA — Principal Components Analysis). Данная модель определяет внутреннюю энергию контура, позволяющую сохранять форму в пределах класса лиц, а не каких-либо других объектов.

На рис. 2 приведен пример работы модели формы. Отметим, что модель обладает большим числом параметров, определяемых по обучающей выборке. В рамках данной работы модель формы не подвергается модификации и подробно рассматриваться не будет.

Рис. 2. Модель формы. **Слева** — результат работы без применения модели формы. **Справа** — результат работы с применением модели формы.

Поиск точек по шаблону. На этапе обучения для каждой точки контура вычисляется профиль (шаблон). Профиль строится на основе значений яркости изображения в некоторой окрестности для каждой точки контура. В классической ASM используется одномерный профиль. В модифицированной ASM применяются как одномерные, так и двумерные профили.

Одномерный профиль выделяется вдоль отрезка, направленного по нормали к контуру в рассматриваемой точке (рис. 3). Вдоль этой нормали вычисляется производная функции яркости в каждой точке отрезка. Полученный вектор нормализуется на значение суммы модулей всех его элементов. Такой вектор будем называть одномерным профилем. На этапе обучения по выборке оцениваются средние профили и ковариационные матрицы каждой точки контура. Эти оценки необходимы для вычисления метрики Махаланобиса.

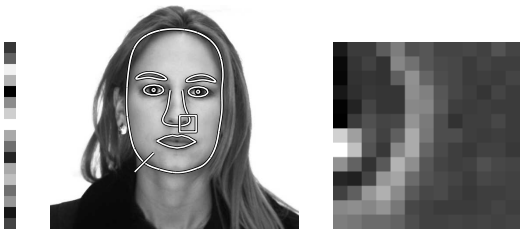


Рис. 3. Профили точки. **Слева** — одномерный профиль. **Справа** — двумерный профиль.

Двумерный профиль. В модифицированной ASM было предложено использование двумерных профилей [5]. Двумерные профили строятся для градиентного представления изображения в квадратной окрестности точки контура (рис. 3). Полученную матрицу можно представить как вектор, записав в столбец все ее элементы. Данный вектор нормализуется на значение суммы модулей всех его элементов. Такой вектор будем называть двумерным профилем. Как и в случае с одномерным профилем, на этапе обучения оцениваются средние профили и ковариационные матрицы.

Применение двумерного профиля делает алгоритм выделения характерных элементов более точным, так как учитывается большее количество информации. Однако время вычислений существенно увеличивается. Одна из причин этого — очень большой размер ковариационной матрицы, используемой в метрике Махаланобиса. Для уменьшения вычислительных затрат применяют технологию разрежения матрицы. При этом свойство положительной определенности должно сохраняться. Как правило, заполнение разреженной матрицы составляет 10%. Это позволяет добиться существенного увели-

Алгоритм 1. Локализация признаков

Вход: изображение лица;
Выход: контур, соответствующий искомым признакам;
1: установить начальный контур по результатам детекции лица;
2: **для** каждой субмодели
3: **для** каждого уровня пирамиды изображений
4: **пока** контур не сошелся
5: **для** каждой точки контура
6: построить профили в окрестности точки;
7: выявить наилучшее соответствие эталону;
8: переместить точку в наилучшее положение.
9: скорректировать контур с помощью модели формы.

чения производительности, но при этом незначительно уменьшается точность алгоритма.

В работе [5] был проведен глубокий анализ десятков параметров алгоритма. Так, например, если размер одномерного профиля равен 17 пикселей, а двумерного — $13{\times}13 = 169$, то количество элементов в ковариационных матрицах составляет 289 и 28561 соответственно.

Общий алгоритм. Эксперименты показали, что можно существенно повысить точность алгоритма, если использовать пирамиду изображений [6]. Локализация контура осуществляется последовательно для нескольких масштабов одного и того же изображения, соответствующим уровням пирамиды. Начальное позиционирование контура осуществляется с помощью детектора лиц, например, детектора Viola-Jones [7]. Так как применение двумерного профиля сопряжено с высокими вычислительными затратами, то локализация характерных признаков (см. алгоритм 1) основана на двух последовательно применяемых субмоделях:

1) локализация на 4-х уровнях пирамиды с использованием только одномерных профилей;
2) локализация на 2-х уровнях пирамиды с использованием одномерных и двумерных профилей.

Знаковое представление изображения

Рассмотрим альтернативный метод выделения профилей, основанный на знаковом представлении.

Знаковое представление изображения, заданного функцией двумерного аргумента $I(x, y)$, выглядит следующим образом [2, 8]:

$$M = \begin{pmatrix} \operatorname{sign} I'_x & \operatorname{sign} I'_y \end{pmatrix},$$

где $\operatorname{sign} I'_x$, $\operatorname{sign} I'_y$ — знаки частных производных первого порядка функции $I(x, y)$.

Таким образом знаковое представление M — матрица, элементами которой являются пары вида $(\operatorname{sign} I'_x,\ \operatorname{sign} I'_y)$. На практике для изображения размером $H{\times}W$ удобно использовать знаковое представление в виде развернутого вектора размером $2HW$. В качестве метрики между двумя такими векторами удобно использовать метрику Хэмминга:

$$\rho(x, y) = \sum_{k=1}^{2HW} |x_k - y_k|.$$

Для вычисления этой метрики требуется $O(n)$ операций. Отметим, что метрика Хэмминга на векторах признаков соответствует псевдометрике на исходных изображениях, то есть из равенства векторов признаков не следует равенство исходных изображений. Изображения, между признаками которых метрика Хэмминга равна нулю, образуют классы эквивалентности. Примечательным является тот факт, что при изменении яркости или контрастности изображения результат преобразования попадает в тот же класс эквивалентности, что и исходное изображение. Таким образом, предложенное представление изображения обеспечивает устойчивость к влиянию освещенности.

Используя знаковое представление в качестве профиля в модели ASM, можно существенно уменьшить вычислительные затраты за счет более компактного представления и метрики Хэмминга, которая является вычислительно менее сложной, чем метрика Махаланобиса.

Вычислительные эксперименты

В работе [5] качество результатов работы алгоритма оценивается на базе BioID [3]. Данная база содержит 1521 изображение лиц, размеченных вручную. Разметка одного изображения состоит из 20 точек. Критерием оценки служит среднее отклонение точек ручной разметки от результатов работы алгоритма. Количество точек в разметке (20) и в контуре модели (68) различается. Кроме того, не для всех точек ручной разметки существуют соответствующие точки в контуре модели. Таким образом, результаты оценивались по 17 общим точкам. В работе [5] оценка алгоритма проводилась точно таким же образом (me17).

В модели ASM [5] поиск осуществляется по двум субмоделям с различным числом уровней в пирамиде изображений: 1) одномерный профиль (17 элементов), 4 уровня; 2) одномерные и двумерные профили (17 и $13^2 = 169$ элементов), 2 уровня. Для обоих типов профилей использовалась метрика Махаланобиса. В случае двумерного профиля ковариационная матрица была разреженной.

В предложенной модели (MBV — Matrix of Brightness Variation) использовались только профили, основанные на знаковом представлении на 4-х уровнях пирамиды изображений. Эксперименты показали, что оптимальный размер профиля составляет $14^2 = 196$ элементов. Все остальные параметры модели были взяты такими же, как и для ASM.

В таблице 1 представлены результаты измерений времени работы алгоритмов и их точность по метрике me17, значение которой нормировано на расстояние между зрачками. Как видно из таблицы, время загрузки и инициализации модели MBV происходит быстрее модели ASM в 7 раз. Это достигается за счет более компактного представления MBV. Для модели ASM представлены два значения времени вычисления метрики — для одномерного и двумерного профилей. Вычисление метрики Хэмминга для вектора размером в 196 элементов в 6,7 раз быстрее, чем вычисление метрики Махаланобиса для вектора размером в 169 элементов. Общее время локализации признаков на одном изображении при использовании MBV в два раза меньше, чем у ASM.

По показателю точности MBV немного уступает модели ASM. Отметим, что начальный контур, построенный по результатам детекции, отклоняется в среднем от вручную размеченного контура на 0,1363 по метрике me17.

Таблица 1. Сравнение методов MBV и ASM.

Критерий	ASM	MBV
Загрузка модели, с.	7,05	1,05
Локализация, с.	4,49	2,10
Вычисление метрики, мкс.	18.,9 / 82,3	12,2
Точность, me17	0,0585	0,0527

На рис. 4 представлены распределения значений me17 для сравниваемых моделей.

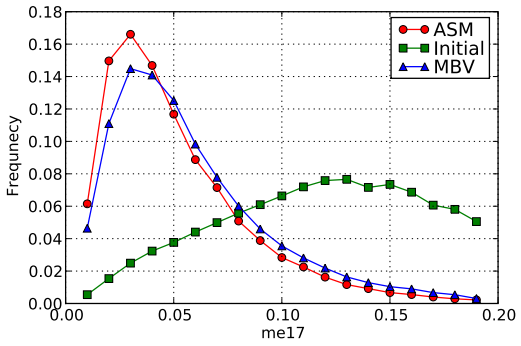


Рис. 4. Гистограмма распределения отклонений найденных точек от идеального положения.

Программная реализация обоих алгоритмов отличалась лишь типом профилей. При этом использовался пакет Matlab 7.4.0 (R2007a) под управлением ОС Ubuntu 7.10 Gutsy Gibbon (2.6.20-15-generic). Конфигурация компьютера: Intel Core 2 Duo 1,80 ГГц, DDR2 Memory 2×1024 Мб.

Выводы

В данной работе предложена модифицированная модель ASM для выделения характерных признаков лиц на цифровых изображениях. В рассматриваемой модели предлагается использовать знаковое представление в качестве профиля изображения. Применение знакового представления позволило существенно сократить временные затраты за счет более компактной формы представления профиля и вычислительно менее затратной метрики Хэмминга, вместо метрики Махаланобиса, используемой в работах [1, 5]. При этом качество результатов ухудшилось незначительно. Стоит отметить, что автором работы [5] была проведена серьезная работа по оценке большого числа параметров модели. В данной работе такой оценки не проводилось. Таким образом, при дальнейших исследованиях будет выявлен набор параметров, при которых предложенная в данной работе модель покажет более точные результаты.

Литература

[1] *Cootes T., Taylor C.* Statistical models of appearance for medical image analysis and computer vision: Tech. rep. // The University of Manchester School of Medicine. — 2004. `www.isbe.man.ac.uk/~bim/Models/app_models.pdf`.

[2] *Goncharov A., Gubarev V.* Comparison of high-level and low-level face recognition methods // Pattern recognition and image analysis: new information technologies (PRIA-9-2008). — 2008. — P. 178–181.

[3] *Jesorsky O., Kirchberg K., Frischholz R.* Robust face detection using the hausdorff distance // Audio and Video based Person Authentication — AVBPA, Springer. — 2001.

[4] *Martinez A., Benavente R.* The AR face database // CVC Technical Report. — № 24. — 1998

[5] *Milborrow S.* Locating Facial Features with Active Shape Models // Master's thesis, Faculty of Engineering, University of Cape Town. — 2007.

[6] *Adelson E., Anderson C., Bergen J.* Pyramid method in image processing // RCA Engineer. — 1984. — Vol. 29, № 6. — P. 33–41. `web.mit.edu/persci/people/adelson/pub_pdfs/RCA84.pdf`.

[7] *Viola P., Jones M.* Robust real-time face detection // Int. J. Comput. Vision. — 2004. — Vol. 57, № 2. — P. 137–154.

[8] *Гончаров А., Горбань А., Каркищенко А., Лепский А.* Поиск портретных изображений по содержанию // Интернет-математика 2007: Сборник работ участников конкурса. — 2007. `http://download.yandex.ru/IMAT2007/goncharov.pdf`.

Сегментация модели лица на статические и динамические области по трехмерной видеопоследовательности*

Гордеев Д. В., Дышкант Н. Ф.
dott1718@gmail.com, Natalia.Dyshkant@gmail.com
МГУ им. М. В. Ломоносова

Рассматривается задача сегментации трёхмерной модели лица на статические и динамические области по трёхмерному видеоряду процесса жевания. Трёхмерные модели получены методом трёхмерного сканирования в виде облаков точек. Предлагается метод, позволяющий сегментировать трёхмерную модель и описывать динамику движения подвижной части относительно статической. Предлагаемый подход основывается на методе подгонки частей моделей лиц, то есть минимизации меры различия между ними. Проведены вычислительные эксперименты на реальных данных.

Современные технологии трёхмерного сканирования позволяют не только получать точную модель лица, но и производить съёмку изменений и движений (motion capture) в режиме реального времени, захватывая движения нижней челюсти (например, во время жевания или разговора) и любые мимические движения. В результате можно получить серию последовательных трёхмерных изображений поверхности движущегося объекта — трёхмерную видеопоследовательность.

Задачи анализа механического движения челюсти человека являются актуальными и важными для исследований в таких востребованных областях медицины, как хирургическая стоматология и челюстно-лицевая хирургия. Параметрическая модель движения нижней челюсти может быть использована при медицинской диагностике и оценке результатов операций в ортодонтии.

В настоящее время трёхмерные технологии моделирования активно применяются в стоматологии и косметологии: работа с 3D моделями позволяет производить планирование операций и анализировать возможные результаты лечения. В [1] предлагается метод лазерного сканирования и система визуализации для виртуального планирования операций. В [2] предлагается использование фотограмметрической системы с высокой точностью измерений для получения 3D моделей челюстей и лица и дальнейшего определения взаимного расположения нижней и верхней челюстей.

Съёмка для настоящего исследования производилась трёхмерным сканером Broadway™ компании «Artec Group» [3], позволяющим делать до 15 снимков в секунду; полученные трёхмерные модели лица задаются облаками точек, рис. 1.

Задача сегментации 3D модели

Рассмотрим задачу сегментации модели лица человека по трёхмерному видеоряду процесса жевания — серии последовательных трёхмерных изображений, полученных при съёмке жующего человека. Под сегментацией будем понимать разбиение поверхности лица на статические и динамические (относительно исследуемого видеоряда) области. При этом сегментируется статичная трёхмерная модель лица, полученная при «нейтральном» выражении лица снимающегося.

Рис. 1. Схема сканирования модели лица.

Исходными данными в задаче являются модели, заданные в виде облаков точек:

— статичная 3D модель лица S;
— видеоряд трёхмерных изображений $D_1, \ldots, D_n$.

Анализируя данные, полученные в результате обработки последовательных движений, необходимо получить информацию о расположении статических и динамических областей модели в трёхмерном пространстве и описать деформации областей снятой поверхности.

Метод сегментации

Нормализация моделей. Каждая исходная модель задана в трёхмерном пространстве набором координат точек:

$$\{x_i, y_i, F(x_i, y_i)\}_{i=1}^{N}.$$

Обозначим набор точек плоскости, на котором задается поверхность модели через $G = \{x_i, y_i\}_{i=1}^{N}$.

На первом шаге работы происходит нормализация моделей: с каждой из моделей $S, D_1, \ldots, D_n$ связывается своя система координат (см. рис. 2): ось Oz проходит вдоль оси визирования, ось Oy идёт вдоль лица по направлению от подбородка ко лбу, ось Ox — поперёк лица от правой щеки

*Работа выполнена при финансовой поддержке РФФИ, проекты №08-07-00305 и №09-07-92652.

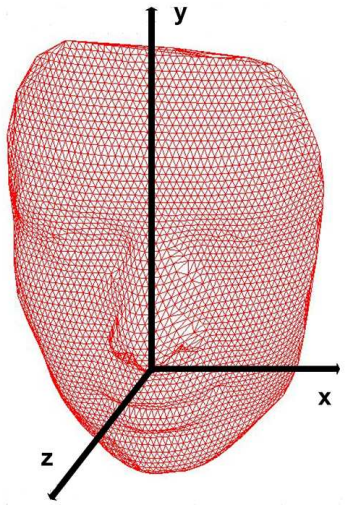


Рис. 2. Триангулированная 3D модель лица и связанная с ней система координат.

к левой, начало координат выбирается так, чтобы для модели была выполнена следующая система:

$$\begin{cases} \sum\limits_{(x,y)\in G} x = \sum\limits_{(x,y)\in G} y = 0; \\ \max\limits_{(x,y)\in G} F(x,y) = 0. \end{cases}$$

Здесь суммирование производится по всем точкам (x, y) из G. Описанную систему координат будем называть *стандартной* для конкретной модели.

Сегментация. При движении нижней челюсти на снимках видеоряда наибольшей статической частью является верхняя часть лица, к которой относятся лоб и нос; нижняя часть лица (щёки, губы и подбородок) относятся к динамической части. Для выделения верхней и нижней части лица предлагается сечение модели лица горизонтальными плоскостями, то есть плоскостями, параллельными Oxz.

Для сравнения двух моделей лица со снимков предлагается подход, заключающийся в определении статической части для обеих снимков методом подгонки и описании движения движущейся (динамической) части относительно статической.

Пусть S_B, D_B — статические (верхние) части модели S и модели D из видеоряда, а S_H, D_H — динамические части этих моделей (см. рис. 3); O_1, O_2 — системы координат, связанные с моделями S и D соответственно.

Модель D приводится к стандартной системе координат модели S, и статические части S_B, D_B сопоставляются друг с другом методом подгонки так, чтобы мера различия между ними была наименьшей. При этом динамические части моделей S_H и D_H будут отличаться. Задача описания динамики движения подвижной части модели относительно статической состоит в регистрации таких различий для каждого снимка D из видеоряда.

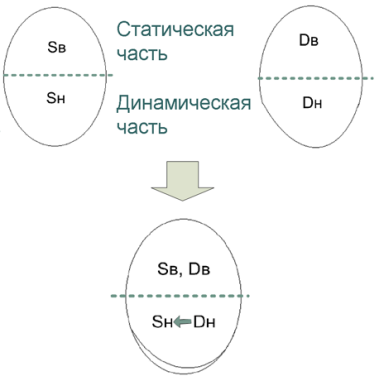


Рис. 3. Подгонка статических частей двух моделей. S_B, S_H — верхняя и нижняя части статичной модели S; D_B, D_H — верхняя и нижняя части модели D видеоряда.

Для *выделения статических частей* модели разделяются плоскостью P, первоначально совпадающей с плоскостью Oxz, верхние части моделей сравниваются, далее плоскость P смещается вдоль оси Oy с некоторым шагом, и отделяемые ею верхние части моделей снова сравниваются. При этом используется мера различия, учитывающая площадь, на которой задана верхняя часть модели, отсекаемая плоскостью P. Смещение плоскости P происходит до тех пор, пока мера различия поверхностей уменьшается.

В качестве метода сравнения поверхностей используется метод, предложенный в [4], позволяющий вычислять меру различия между поверхностями, заданными как функции на разных нерегулярных сетках. Данный метод основан на аппроксимации поверхностей кусочно-линейными функциями по триангуляциям Делоне; основная идея состоит в восполнении значений каждой из функций, соответствующих поверхностям, в точках второй сетки через построение триангуляций и локализацию их друг в друге.

Используемые меры различия между поверхностями двух моделей и методы оптимизации функционала различия между ними приведены в следующем разделе. Сопоставление (подгонка) поверхностей состоит в нахождении такого движения, при котором мера различия между поверхностями, описывающими лица, минимальна.

Функционал различия поверхностей

Определение 1. *Поверхность называется однолистной или монотонной по отношению к некоторой оси, если любая прямая, параллельная этой оси, пересекает поверхность не более чем в одной точке.*

Пусть M — движение в пространстве $\mathbb{R}^3$. Далее будем рассматривать его как композицию последовательных поворотов на углы $\alpha_M, \beta_M, \gamma_M$ вокруг осей Ox, Oy, Oz соответственно и параллельного переноса на вектор $(\Delta x_M, \Delta y_M, \Delta z_M)$.

Пусть D — множество точек пространства $\mathbb{R}^3$, а M — движение в пространстве $\mathbb{R}^3$. Тогда обозначим $M(D) \equiv \{M(d)\colon d \in D\}$, где $M(d)$ — точка, полученная в результате движения M точки d.

Определение 2. *Пусть на множестве однолистных поверхностей задана полуметрика ρ. Мерой различия двух однолистных поверхностей D_1 и D_2 назовём величину $\inf\limits_{M\in\mathscr{M}} \rho(D_1, M(D_2))$, где $\mathscr{M}$ — множество движений в пространстве, сохраняющих свойство однолистности поверхности D_2.*

Определение 3. *Внутренним подмножеством двух двумерных сеток G_1 и G_2 назовём множество* $\mathrm{Int}_{G_1,G_2} = (G_1 \cup\ G_2) \cap \mathrm{Conv}(G_1) \cap \mathrm{Conv}(G_2)$,

где $\mathrm{Conv}(G)$ — выпуклая оболочка множества G.

Определение 4. *Совместная триангуляция Делоне T двух сеток G_1 и G_2 — триангуляция Делоне, построенная на множестве узлов* Int_{G_1,G_2}. *Обозначим* $N_T = |\,\mathrm{Int}_{G_1,G_2}\,|$

Меры различия. Рассмотрим примеры мер различия двух однолистных поверхнстей D_1 и D_2, заданных в виде функций F_1 и F_2 на множестве Int_{G_1,G_2}.

Объём симметрической разности.

Рассмотрим непрерывные функции $\hat{F}_1$ и $\hat{F}_2$, заданные на $\mathbb{R}^2$ и полученные линейной интерполяцией функций F_1 и F_2 по точкам сеток G_1 и G_2 соответственно. Таким образом, $\hat{F}_1$ и $\hat{F}_2$ задают триангулированные поверхности. Введем обозначение

$$V(A,B,C,F_1,F_2) = \iint\limits_{\triangle ABC} |\hat{F}_1(x,y) - \hat{F}_2(x,y)|dxdy.$$

Тогда объём симметрической разности вычисляется по следующей формуле

$$\rho_V(D_1,D_2) = \sum_{\triangle ABC\in T} V(A,B,C,F_1,F_2),$$

где через T обозначена совместная триангуляция Делоне сеток G_1 и G_2, а суммирование происходит по всем треугольникам из T.

Среднее осевое расстояние:

$$\rho_M(D_1,D_2) = \sum_{(x,y)\in\mathrm{Int}_{G_1,G_2}} \frac{\big|F_1(x,y) - F_2(x,y)\big|}{N_T}.$$

Отсечённое осевое расстояние.

Пусть задано число $0 \leqslant \alpha \leqslant 1$. Тогда величина $\rho^\alpha_M(D_1,D_2)$ определяется исходя из пары соотношений

$$\begin{cases} \big|\{(x,y)\colon |F_1(x,y)-F_2(x,y)| \leqslant \rho^\alpha_M\}\big| \geqslant \alpha N_T; \\ \big|\{(x,y)\colon |F_1(x,y)-F_2(x,y)| \geqslant \rho^\alpha_M\}\big| \leqslant (1-\alpha) N_T. \end{cases}$$

Здесь суммирование происходит по всем точкам $(x,y) \in \mathrm{Int}_{G_1,G_2}$.

Замечание 1. Меры различия ρ_M и ρ^α_M не требуют построения объединённой триангуляции Делоне для двух сеток, на которых заданы модели, в отличие от ρ_V.

Оптимизация функционала различия.

Подгонка двух поверхностей заключается в минимизации меры различия между ними.

Рассмотрим следующую оптимизационную задачу в пространстве $\mathbb{R}^6$:

$$\rho(G_1, M(G_2)) \to \inf,$$

где инфинум берётся по всем шести параметрам движения M — $\alpha_M, \beta_M, \gamma_M, \Delta x_M, \Delta y_M, \Delta z_M$ — из возможных движений M сетки G_2.

Следует отметить, что поскольку ρ является полуметрикой, то минимум функционала равен 0 и достигается при таком сдвиге модели, когда $\mathrm{Conv}(G_1) \cap \mathrm{Conv}(G_2) = \varnothing$. Таким образом, при оптимизации необходимо использовать метод поиска локального экстремума, а не глобального. Кроме того, выбор начального приближения метода является критичным.

Описанная выше предварительная нормализация моделей — приведение их к стандартной системе координат — является эффективным решением последней проблемы. Нормализация модели производится параллельным переносом и, очевидно, не выводит F из класса однолистных поверхностей.

Предлагаются следующие методы поиска локального экстремума меры близости.

Метод покоординатного спуска. Как известно, суть метода заключается в фиксировании всех переменных за исключением одной и решении задачи минимизации для получившейся функции одной переменной. Данный метод весьма эффективен при решении поставленной задачи, однако его главным недостатком является медленная скорость работы из-за большого количества вычислений функционала ρ.

Метод Нелдера-Мида (метод деформируемого многогранника, симлекс-метод). Идея метода состоит в сравнении значений функции в вершинах начального симплекса и перемещении симплекса по направлению оптимальной точки с помощью итерационной процедуры. Как показали проведённые эксперименты, в рассматриваемой задаче этот метод проигрывает в качестве оптимизации методу покоординатного спуска, однако скорость его работы значительно выше.

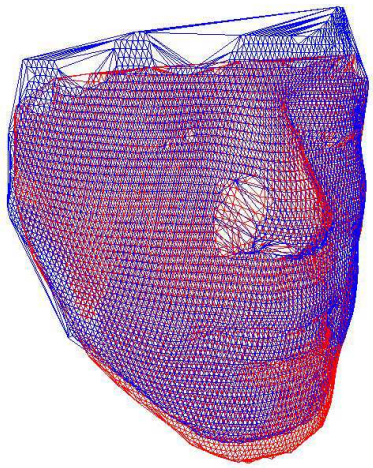


Рис. 4. Модели лиц после нормализации и оптимизации функционала среднего осевого расстояния методом Нелдера-Мида.

Таблица 1. Значение меры ρ_V и время её вычисления при разных методах оптимизации.

Метод оптимизации	ρ_V	**Время (мс)**
Без оптимизации	13 208	—
Покоординатный спуск	6 358	45 883
Нелдера-Мида	6 735	16 068

Таблица 2. Значение меры ρ_M и время её вычисления при разных методах оптимизации.

Метод оптимизации	ρ_M	**Время (мс)**
Без оптимизации	2,46	—
Покоординатный спуск	1,21	4 712
Нелдера-Мида	1,24	1 700

Таблица 3. Значение меры ρ_M^α и время её вычисления при разных методах оптимизации.

Метод оптимизации	ρ_M^α	**Время (мс)**
Без оптимизации	3,59	—
Покоординатный спуск	1,65	4 883
Нелдера-Мида	1,95	1 996

Вычислительные эксперименты

В рамках исследования были проведены вычислительные эксперименты на видеоряде из 83 моделей, полученном при съёмке жующего человека.

Количество точек в моделях варьируется от 4 000 до 5 000.

Как показывают эксперименты (см. таблицы 1–3), самая быстрая оптимизация для исследуемых моделей — оптимизация среднего осевого расстояния методом Нелдера-Мида — занимает менее 2 секунд (эксперименты проведены на машине с процессором Intel Core2Duo с частотой 2,2 ГГц и оперативной памятью 2 Гбайт).

Выводы

Сформулирована постановка задачи сегментации модели лица на статические и динамические области по трёхмерной видеопоследовательности процесса жевания. Предложен подход к её решению, позволяющий выделять статические и динамические части модели и оценивать движение нижней челюсти. Проведены вычислительные эксперименты на реальных данных. Предложенный подход базируется на методе подгонки поверхностей лиц, минимизирующем меру различия между ними. Предложено использование трёх мер различия между поверхностями и двух методов минимизации функционала различия. Для них проведены сравнительные эксперименты.

В дальнейшем планируется анализировать динамические части моделей видеоряда, и сегментировать из них наименнее изменяющуюся при движениях челюсти область (область подбородка, где мягкие ткани наиболее плотно примыкают к нижней челюсти). Преобразование системы координат, связанной с этой областью, в систему координат статической части модели и будет описывать динамику движения нижней челюсти. Формальное описание можно будет получить в виде матриц преобразования. Результат может быть визуализирован в виде анимации, показывающей движение системы координат при таком преобразовании.

Благодарности

Авторы выражают благодарность своему научному руководителю профессору Леониду Моисеевичу Местецкому и Арчилу Цискаридзе.

Литература

[1] *Koidis P., Patias P., Tsioukas V.* 3D Visualization of Dental Data for Virtual Treatment Planning // ISPRS Congress Istanbul 2004, Proceedings of Commission V, 2004. — Pp. 996–1001.

[2] *Knyaz V. A., Zheltov S. Yu.* Photogrammetric Techniques for Dentistry Analysis, Planning and Visualisation // ISPRS Congress Beijing 2008, Proceedings of Commission V, 2008. — Pp. 783–788.

[3] `http://www.artec-group.com` — Artec Group — 3D Scanning Technologies — 2007.

[4] *Дышкант Н. Ф., Местецкий Л. М.* Сравнение однолистных поверхностей полученных при 3D сканировании // Proceedings of 18th International Conference on Computer Graphics and Vision «GraphiCon'2008», 2008. — Pp. 270–277.

Формализация задачи распознавания последовательности состояний сложного источника

Грызлова Т. П.
ktntpgryzlova@mail.ru
Рыбинск, РГАТА им. П. А. Соловьева

Цель работы — разработка модели сигнала, на основе которой можно проектировать системы автоматического анализа сложных сигналов, не делая предположений, необходимых для аналитического решения задачи оптимального приема или обоснованного применения известных методов обработки данных. Результатом анализа является разложение сигнала на составные элементы и распознавание как элементов, так и сигнала в целом.

Теоретические проблемы распознавания последовательности состояний источника по цифровым нестационарным сигналам возникают в задачах распознавания речи, испытаниях авиационных двигателей и множестве других задач анализа сигналов, для которых модели, позволяющие синтезировать алгоритмы оптимального приема или обработки или не известны, или настолько сложны, что пользоваться ими неконструктивно. Кроме того, оценки последовательности состояний источника можно использовать как промежуточные данные для диагностики неисправности сложного технического объекта по почти-периодическим цифровым сигналам. Известные модели источников — комбинаторные, вероятностные, стационарные и составные [1, 2] — не отражают реальной сложности исследуемых объектов, поэтому применение классических методов спектрального, корреляционного или авторегрессионного анализа ограничено.

Проблемы анализа сигналов сложного источника

Часто методы автоматического анализа сигналов не обеспечивают ожидаемого результата, поскольку они применяются в условиях, противоречащих базовым теоретическим принципам методов. Так, анализ сложных диагностических сигналов, формируемых при функционировании технических систем и подключении к ним систем специальных датчиков, проводится в условиях неопределенности состояния диагностируемого объекта, то есть неопределенности модели функционирования. Как правило, неизвестны модели элементарных сигналов, неизвестны законы взаимодействия элементарных сигналов, недостаточна длительность элементарных сигналов для эффективного разрешения по частоте, элементы сигнала могут быть переходными процессами, недостаточна выраженность переходов между элементарными сигналами. Спектральный анализ в таких случаях не дает однозначной информации. Само по себе проведение спектрального анализа не решает задачи автоматического выбора признаков для диагностики. Для формирования признаков используются статистические характеристики сигналов или их спектров, и требуется целенаправленный отбор подходящих алгоритмов обработки спектров или формализация интерпретации спектральных данных. В каскадной модели составного источника [2], структурно близкой к модели сложного источника, предлагаемой в настоящей статье, предполагается статистическая однородность данных в пределах сегментов и используется модель стохастических линейных дифференциальных уравнений, коммутируемых дискретным источником. Анализ кусочно-стационарных сигналов спектральным методом требует либо знания моментов переключения дискретного источника, либо оценки моментов переключения таких источников.

Прикладной анализ сигналов сложных источников ставит ряд теоретических задач, которые, возможно, не могут быть решены аналитически. Модель сложного источника предназначена для построения многоуровневой системы распознавания, в простейшем случае она является двухуровневой.

Определение 1. *Распознавание последовательности состояний источника на верхнем уровне — это одновременное обнаружение моментов переключения источника и распознавание состояний.*

Определение 2. *Распознавание последовательности состояний сложного источника на нижнем уровне — одновременное обнаружение границ элементарных сигналов в сложном сигнале и распознавание элементарных сигналов.*

В состоянии Q_i создается выходной сигнал ${}^i\boldsymbol{s}_0^{T_i}$ длительности T_i, в переходных состояниях Q_{ij} регистрируются сигналы с трудно определяемыми временными границами. Например, в звуке выделяют экскурсию, выдержку и рекурсию. Удобно пользоваться этими терминами и в случаях, когда речь идет о сигналах произвольных сложных источников.

Определение 3. *Последовательность s_{t0}^{t0+n-1} длины n является элементом сигнала $\boldsymbol{s}_0^{T-1}$, если сигнал можно разложить в последовательную композицию*

$$\boldsymbol{s}_0^{T-1} = \boldsymbol{r}_0^{t_0} \cdot s_{t_0}^{t_0+n-1} \cdot \boldsymbol{q}_{t_0+n-1}^{T-1}$$

так, что для $0 \leqslant t_0 < (T-1-n),\ n < T$

$$r(t_0) = s(t_0),\ \text{а}\ \ q(t_0+n-1) = s(t_0+n-1).$$

В свою очередь, процессы $\boldsymbol{r}(t)$ и $\boldsymbol{q}(t)$, определенные на интервалах $[\,0, t_0\,]$ и $[\,t_0{+}n{-}1, T{-}1\,]$, соответственно, могут быть представлены в виде последовательной композиции элементарных процессов [3].

Определение 4. *Вспомогательные функции, выделяющие участки элементарных сигналов, соответствующие выдержке, называются стробирующими, или просто стробами.*

Трудности решения теоретических и практических задач анализа сигналов сложных источников, в том числе распознавания последовательности состояний, обусловлены следующим:

— границы элементарных сигналов очень нечеткие, элементарные сигналы интерферируют;
— вариативность элементов в выборке сигналов нескольких источников очень высока;
— вариативность элементов в пределах одного сигнала существенна (например, в сигнале «мама» формальные характеристики разных фонем в пределах слогов имеют большее сходство, чем фонемы одного класса в разных слогах);
— обучение на специально зарегистрированных изолированных элементах не приводит к успешному распознаванию;
— рекурсивность: выделение обучающих элементов из сложного сигнала для синтеза системы распознавания — задача, которую можно было бы решить, если бы система распознавания уже была построена;
— функционалы не должны быть сложными, чтобы задачу распознавания элементов сигнала можно было решить в реальном времени.

В основе предлагаемого решения лежит принцип повторяемости элементов сигнала и общности законов формирования объектов одного класса. Элементарные сигналы могут быть сильно вариативными, а законы их формирования — сложными и неизвестными.

Задача оценки последовательности состояния управляемого источника

Пусть определено множество состояний $A = \{a_i\}$ дискретного источника и по реализации $s(t)$, которой при дискретизации соответствует цифровая последовательность $\boldsymbol{s}_0^t$, требуется определить последовательность состояний источника $a(t)$. Эту оценку будем обозначать $\hat{a}(t)$. Пусть имеются образцы сигналов $s(t, a_i) = \boldsymbol{p}_0^{\tilde{t}}$, генерируемых источником в каждом фиксированном состоянии a_i или при переходе из состояния в состояние. Оценки ищутся через оценку сигнала $\hat{\boldsymbol{s}}_0^t$ и промежуточные оценки элементов сигнала $\hat{\boldsymbol{p}}_0^{\tilde{t}}$, выделенных из сложного сигнала, и оценки некоторых локальных временных параметров сигнала, необходимых для распознавания. Остановимся на обобщенных временных характеристиках модели. Не предполагается, что переходы из состояния в состояние происходят через фиксированные интервалы $\Delta\tau$. Наоборот, одновременно с оценкой $\hat{a}(t)$ оценивается последовательность моментов разладок $\hat{\tau}(t) = \hat{\tau}_1^K$ процесса $s(t, a_\tau)$. K — количество элементов в последовательности оценок $\hat{a}(t)$, оно может отличаться от количества L моментов разладок $\tau(t) = \tau_1^L$ в исходной последовательности $a(t)$. Предполагается, что момент времени $\tau(t) = \tau(t-1) + \Delta\tau_{t-1}$. При анализе используются две модели времени — дискретный аналог временного интервала $[T_0, T_1]$ длиной T и событийное время (индекс событий). Одно и то же обозначение t применяется как для индексации последовательностей, так и как параметр временных функций. В первом случае $t \in [\,0, 1, 2, \dots\,)$, во втором — $t \in [\,0, \Delta t, 2\,\Delta t, \dots\,)$, где $\Delta t = 1/f_s$ — шаг дискретизации аналогового сигнала $s(t)$, f_s — частота дискретизации. Вспомогательная задача, которую приходится решать — отбор на обучающих последовательностях $\boldsymbol{p}_0^{\tilde{t}} \leftrightarrow a_i$, ассоциированных с состоянием a_i, эффективных статистик и преобразований $\boldsymbol{F}(\boldsymbol{s}_0^{\tilde{t}})$ для решения задачи распознавания $\boldsymbol{s}_0^t \to \hat{a} \in A$. Обучающие последовательности имеют произвольную длину $\tilde{t}$, близкую к среднему интервалу между сменой состояния источника или несколько большую. Длительность наблюдений заранее не фиксируется, а определяется ходом реализации наблюдаемого процесса, то есть, речь идет о последовательных решениях, пригодных для использования в системах реального времени. Потоки событий в модели сигнала и, соответственно, в системе его анализа, описываются многоуровневой системой. На верхнем уровне это текст, последовательности в алфавите A. На нижнем — изменение некоторой физической величины.

Модель последовательности состояний распределенного источника

Модель распределенного сложного источника включает множество почти одинаковых источников $\{A_1, \dots, A_m\}$. Каждый источник формирует сигнал $x_k(t)$, который является компонентой многомерного процесса

$$\boldsymbol{X}(t) = \big(\boldsymbol{x}_1(t) \cdot \boldsymbol{b}_1(t), \dots, \boldsymbol{x}_m(t) \cdot \boldsymbol{b}_m(t)\big)^{\intercal},$$

где $\boldsymbol{b}_k(t)$ — функция включения соответствующей компоненты. Такие функции имеют область значений $\{0, 1\}$ и называются временными переключательными функциями. В среде распространения и в воспринимающей системе (датчике) компоненты взаимодействуют, в результате наблюдаемый сигнал является неизвестным сложным преобразованием $\boldsymbol{s}(t) = \boldsymbol{S}\boldsymbol{T}(\boldsymbol{X}(t))$. В частном случае это мо-

жет быть простое суммирование компонент или их коммутация. В общем случае это сложное взаимное влияние, включающее как последействие «отработавшего» процесса, так и предварительную подготовку к процессу или одновременное воздействие на датчик. Можно ввести более общую модель:

$$\boldsymbol{X}(t) = \big(\boldsymbol{x}_1(t) \cdot \mathrm{swf}_1(t), \ldots \boldsymbol{x}_m(t) \cdot \mathrm{swf}_m(t)\big)^{\mathsf{T}},$$

где $\mathrm{swf}_k(t)$ — специально сконструированные переключательные функции, например, одна компонента может нарастать, а другая — убывать. В предложенной модели сложного источника источников сигналов много, сигналы от них достаточно просты, похожи (характерны), но взаимодействуют сложным и случайным образом. Модель сигнала строится на основе чередования или наложения характерных последовательностей друг с другом или с другими элементарными последовательностями. Характерная последовательность может быть результатом взаимодействия $\chi(t) = \bigotimes\limits_i s_i$ элементарных сигналов $s_i = {}^i s_0^{T_i}$. Некоторая часть цепочки взаимодействующих сигналов может соответствовать отдельному источнику. Взаимодействие разных последовательностей в едином сигнале приводит к искажению их формы и проявляется как неаддитивные помехи. В природе, технике часто наблюдаются сигналы, в которых человек легко выделяет повторяющиеся последовательности и опознает по ним объекты или их состояние. Такие явления могут возникать в акустике, когда имеется неопределенное количество источников повторяющихся сигналов, накладывающихся друг на друга и взаимодействующих сложным образом. Почти периодически повторяется пространственная конфигурация элементарных источников вибраций относительно датчиков в задаче диагностики системы подшипников трансмиссии газотурбинных двигателей (ГТД). В [4] последовательности, достаточно часто повторяющиеся в наблюдаемом сигнале, были определены как характерные последовательности. Для предварительной обработки информации в системах распознавания (диагностики) глобального состояния сложного источника разработан и исследован метод характерных последовательностей (МХП), который позволяет избежать сегментации и автоматически найти признаковые пространства для диагностики. Последовательности полагались неизвестными, с неизвестными размерами, частотными характеристиками и частотами встречаемости в сигнале. Для подсчета повторяющихся последовательностей определены процедура сравнения последовательностей, правило вычисления расстояния между последовательностями и правила задания порога расстояния, при котором последовательности считаются равными. Последовательности проверяются относительно случайно выбранных эталонов. Если вокруг эталона образуется кластер, то он может быть переопределен как характерная последовательность. Метод характерных последовательностей показывает важность эвристических методов исследования, когда нет возможности формализовать задачу анализа сигнала, чтобы решить ее аналитически. Хотя он разработан на основе феноменологической модели сложного источника, признаки, отбираемые с помощью МХП, могут непосредственно использоваться для распознавания, служить элементами описания и модели сигнала. Недостатком его, как и большинства других методов анализа, является использование нормировки данных и достаточно большое количество вычислений.

Анализ последовательности состояний сложного источника

Обычно перед распознаванием элементов сигнала выполняется сегментация, причем для выделения элементов сигнала и их распознавания вычисляются разные функционалы. Например, сегментация выполняется по значениям функции сложности [5], а распознавание — по вектору спектральных коэффициентов. В [5] введено несколько функций, вычисляемых при жестко заданном размере инструментального блока данных b, но рекомендации по выбору этого важного для качества сегментации параметра отсутствуют. Функция сложности (ФС) типа I вычисляется по данным трех инструментальных блоков как скалярное произведение между рассматриваемым блоком отсчетов и вектором средних значений соседних блоков:

$$CF_i = \big(\boldsymbol{s}_{ib}^{(i+1)b-1}, \tfrac{1}{2}\big(\boldsymbol{s}_{(i-1)b}^{ib-1} + \boldsymbol{s}_{(i+1)b}^{(i+2)b-1}\big)\big),$$

где $i = 1, \ldots, (Nb-2)$, Nb — количество блоков, на которые разбивается сигнал, при этом фазы элементарных сигналов в блоках получаются случайными. Известно, что оптимальный корреляционный прием требует либо знания фазы сигнала, либо ее оценки, либо должен быть квадратурным. Метод сегментации, предложенный в [5], не может быть пригодным для анализа сигналов произвольных сложных источников, поскольку при построении функции разрушается информация о фазе элементов сигнала. На практике требуется не только определить границы сегментов, но и отделить сегменты, соответствующие выдержке, от переходных сегментов. Предлагается метод анализа, использующий информацию, заключенную в последовательности моментов времени пересечения сигналом нулевого уровня. Основные черты метода:

— выделение блоков данных, согласованных по фазе (полуволн);
— использование усеченного блочного кодирования полуволн переменной длины;

— вычисление временных функционалов на основе расстояний между кодами полуволн;
— многоканальный прием элементов функционалов как случайных величин с заданными средними значениями и дисперсиями (формирование стробов под участки выдержки);
— объединение выходов параллельных каналов в стробирующую последовательность, позволяющие выделить требуемые элементы сигнала.

Анализ последовательности отсчетов начинается с определения моментов времени пересечения нулевого уровня и выделения положительных полуволн $HW_i^+ = (s_{t0}^{t1})_i,\ \ i = 0, \ldots, (n^+ - 1)$, определения последовательностей моментов времени начала $t0_i$ и окончания $t1_i$ i-тых полуволн. n^+ — количество положительных полуволн — подсчитывается как функция времени. События на оси времени представляются точками, которые могут сохраняться в бинарных последовательностях

$$b_{01}(t) = \begin{cases} 1, & \text{если } s(t-1) \leqslant 0 \text{ и } s(t) > 0; \\ 0, & \text{иначе .} \end{cases}$$

Определяются интервалы между событиями пересечения нулевого уровня в одном направлении и длины полуволн. Аналогично выделяются отрицательные полуволны HW^-. После декомпозиции сигнала на полуволны и формирования потока данных о событиях пересечения нулевого уровня можно вычислить ряд полезных для распознавания последовательности состояний источника нестационарных функционалов. Полуволны кодируются:

$$HW_i^+ = s_{t0_i}^{t1_i} \to C_{t0_i}^{t1_i}; \quad M_i = \frac{1}{t1_i - t0_i + 1} \sum_{t=t0_i}^{t=t1_i} s_t;$$

$$C_{t0_i}^{t1_i} = b_{t0_i} \ldots b_{t1_i}; \quad b_{t0_i+j_i} = \begin{cases} 1, & s_{t0_i+j_i} > M_i, \\ 0, & s_{t0_i+j_i} \leqslant M_i, \end{cases}$$

где $L_i = t1_i - t0_i + 1$ — длины полуволн, $j_i = 0, \ldots, L_i - 1$ — индекс отсчетов сигналов в пределах полуволн. Для сравнения полуволн вычисляется расстояние по Хэммингу между кодами в точках $t = t0_i,\ \ i = 1, \ldots, (n^+ - 1)$

$$\Phi 1(t = t0_i) = \rho_i = \rho(HW_i^+, HW_{i-1}^+).$$

Поскольку последовательности имеют разную длину L_i, расстояние считается по индексу j_i более короткой последовательности. Такая обработка сигнала позволяет обнаружить моменты переключения. Одним из функционалов, который можно использовать для обнаружения моментов переключения почти-периодического состояния является

$$\Phi 2(t) = \varepsilon_i^2(t0_i) = \\ = \left(\rho(HW_i^+, HW_{i-1}^+) - \rho(HW_{i+1}^+, HW_i^+)\right)^2.$$

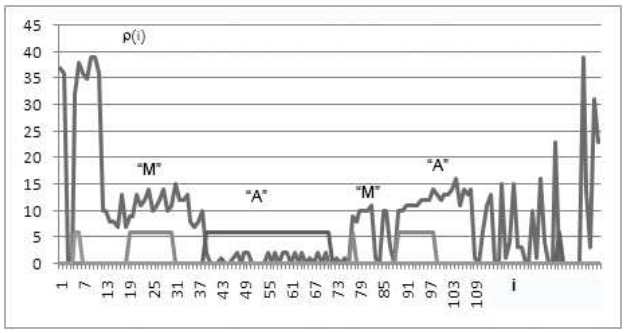


Рис. 1. Расстояние между полуволнами HW^+ и стробы для распознавания элементов сигнала «_мама_».

Усредняя его по Δi полуволнам, получим:

$$\Phi 3(t) = \bar{\varepsilon}_i^2(t0_i) = \\ = \frac{1}{\Delta i} \sum_{t=i}^{t=i+\Delta i} \left(\rho(HW_i^+, HW_{i-1}^+) - \right. \\ \left. - \rho(HW_{i+1}^+, HW_i^+)\right)^2.$$

Сопоставим решения задачи о разладке в речевом сигнале «_мама_», полученные на основе функции сложности типа I, введенной в [5], и на основе функционала $\Phi 1(t)$, введенного выше. Анализируется выборка отсчетов $\boldsymbol{s}_0^{7999}$ при частоте дискретизации $f_s = 8$ кГц. Соответствующая последовательность состояний источника:

$$\boldsymbol{A}_0^{10} = p \cdot Q_{pm} \cdot m \cdot Q_{ma} \cdot a \cdot Q_{am} \cdot m \cdot Q_{ma} \cdot a \cdot Q_{ap} \cdot p,$$

где p — пауза, m и a — состояния выдержки при артикуляции «М» и «А», Q — переходные состояния. В обобщенной модели последовательности состояний источника переходные состояния объединены с состояниями, соответствующими выдержке:

$$\boldsymbol{A}_0^5 = p \cdot \tilde{m} \cdot \tilde{a} \cdot \tilde{m} \cdot \tilde{a} \cdot \tilde{p}.$$

Моменты переключения, определенные детальным визуальным анализом характера сигнала на осциллограмме: $\Delta\tau_1^5 =$

$$= (1593 - 1642, 2617, 3815, 4730, 6000 - 6400).$$

Последовательность интервалов времени артикуляции звуков (длительности пребывания в состояниях $\tilde{m}, \tilde{a}, \tilde{m}, \tilde{a}$): $\Delta\tau_1^4 =$

$$= (1600 - 2617, 2617 - 3815, 3815 - 4730, 4730 - 6000).$$

С помощью системы многоканального приема стационарных функционалов определяются границы участков выдержки. На рис. 1 показан результат автоматического формирования стробов под участки выдержки в речевом сигнале, которые можно использовать для распознавании элементов. Стробы выставлены под интервалы, вложенные в интервалы артикуляции: $\Delta\hat{\tau}_1^4 =$

$$= (1985 - 2432, 2840 - 3499, 4232 - 4652, 4817 - 5368).$$

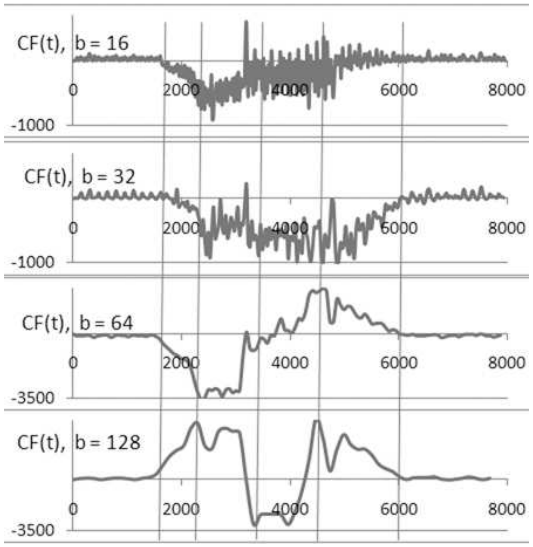


Рис. 2. Функции сложности при различных размерах блока для сигнала «_мама_».

Таким образом, автоматически исключены элементы сигнала, соответствующие переходным состояниям (рис. 1). Управляя порогами многоканальной системы приема стационарных случайных функционалов, можно расширять стробы или сужать их. Функции сложности при малых размерах блоков очень защумлены (рис. 2). Поскольку они имеют большое количество ложных экстремумов, алгоритмы автоматического сегментирования, разработанные в [5], выделяют большое количество ложных границ или трудно интерпретируемых деталей.

При больших значениях блоков визуально можно выделить участки

$$\Delta\hat{\tau}_1^5 \,|\, (b=64) = (1408-2432, 2432-3008, \\ (3008-3200, 3200-4416), 4416-6000),$$

$$\Delta\hat{\tau}_1^6 \,|\, (b=128) = \big(1408-2304, \\ (2304-3072, 3072-3328, 3328-3968), \\ 3968-4352, 4352-5888\big).$$

Это означает, что в принципе, можно разработать многоканальную систему приема, выделяющую как константы, так и участки линейного роста или спада функции сложности. Система правил визуального выделения элементов $CF(t)$ значительно сложнее, чем обработка $\Phi 1(t)$. Выделяемые элементы не являются однородными. Так, состоянию артикуляции ударного «А» соответствует сигнал с заметными участками экскурсии, выдержки, рекурсии, а состоянию артикуляции безударного «А» соответствует сигнал, в котором выделение элементов затруднительно. При этом размер блока надо подбирать. Если размер блока фиксирован, то при ускорении речи будет возрастать дисперсия $CF(t)$, увеличивать же размер блока по очевидным причинам бессмысленно. Заметим, что длительности стационарных участков в тестовом сигнале составляют 56 мс, 82 мс, 53 мс, 69 мс; а элементов, соответствующих и стационарному, и переходному состояниям источника — 128 мс, 150 мс, 114 мс и 159 мс. Длительность элементарных сигналов для $b = 64$ равна 8 мс, для $b = 128 - 16$ мс. Интервалы, на которых вычисляется функции сложности, равны $DT = 3b \cdot \Delta t$. Для размеров блоков $b = 16$, 32, 64 и 128 они составляют 6 мс, 12 мс, 24 мс и 48 мс, и в двух последних случаях соизмеримы с элементарными сигналами. Средние длины полуволн в паузе составляют $4-6$ мс, а в сигнале, генерируемом в состояниях m, a, m, a — 2,5 мс, 1 мс, 2,6 мс, 3 мс. То есть данные сравниваются малыми порциями по $2-6$ мс, на стационарное состояние источника их приходится около 30–60.

Выводы

На основе модели сложного источника, формализованной методами математической теории проектирования вычислительных систем, предложена многоуровневая временная модель системы распознавания. Предварительные исследования показывают, что простые и быстрые алгоритмы, вычисляющие нестационарные функционалы от временных характеристик сигналов, могут быть информативнее, чем распространенные функционалы от амплитудных характеристик.

Литература

[1] *Кричевский Р. Е.* Сжатие и поиск информации. — М: Радио и связь, 1989. — 168 с.

[2] *Орищенко В. И., Санников В. Г., Свириденко В. А.* Сжатие данных в системах сбора и передачи информации. — М: Радио и связь, 1985. — 184 с.

[3] *Капитонова Ю. В., Летичевский А. А.* Математическая теория проектирования вычислительных систем. — М: Наука, 1988. — 296 с.

[4] *Gorshkov A. P., Gryzlova T. P.* Family of effective features for diagnostics of condition of complex technical systems by the example of GTE transmission bearings // 9th International conference on pattern recognition and image analisys: New information technologies (PRIA-9-2008): Vol. 1. — Nizhni Novgorod: Dialog Culture, 2008. — Pp. 185–188.

[5] *Браверман Э. М., Мучник И. Б.* Структурные методы обработки эмпирических данных. — М: Наука., 1983. — 464 с.

Исследование и сравнительный анализ реализаций алгоритмов поиска лиц на изображениях*

Дегтярёв Н. А., Крестинин И. А., Середин О. С.
n.a.degtyarev@gmail.com, crown_s@rambler.ru, oseredin@yandex.ru
Тула, Тульский государственный университет

В работе предложена методика сравнения характеристик качества алгоритмов поиска лиц на изображениях. Основная идея заключается в рутинной обработке базы изображений, как содержащих лица людей, так и нет. Истинное положение лица на изображении задается экспертно в виде координат центров глаз. Однако часть алгоритмов не имеет функционала поиска глаз, поэтому был предложен способ восстановления этих координат по неразмеченной области лица. В эксперименте участвовали пять программных реализаций алгоритмов поиска лиц на изображениях; статья содержит результаты их сравнения.

В настоящее время всё большую актуальность приобретают задачи, связанные с поиском лиц на изображении. Они находят широкое применение в фото/видео технике, в системах обеспечения безопасности объектов и зданий, контроля доступа на территорию. На рынке представлено большое количество программных средств, решающих задачу поиска лиц на изображениях. Цель этой работы предложить параметры качества и методику объективного сравнения различных методов. Такой анализ будет полезен как при выборе уже готового решения, так и разработчикам новых алгоритмов.

В настоящее время проведено сравнение пяти алгоритмов: Intel OpenCV (OCV), Luxand FaceSDK (FSDK), Face Detection Library (FDL), SIFinder (SIF), University of Surrey (UniS). Эти алгоритмы сравниваются по False Rejection Rate (FRR) — коэффициенту неправильного отказа в доступе, False Acceptance Rate (FAR) — коэффициенту ложной идентификации, Total Error Rate (TER) — общему коэффициенту ошибки и по различию векторов ошибок алгоритмов.

Обзор существующих алгоритмов

Алгоритм OCV построен на основе метода, разработанного П. Виола (P. Viola), а затем улучшенного Р. Линхарт (R. Lienhart) [6, 7]. В этом алгоритме применяются каскады бинарных классификаторов, полученных в результате бустинга и работающих в пространстве признаков, получаемых при применении преобразования Хаара. Алгоритм имеет следующие параметры настройки: тип фильтра, шаг окна, шаг масштабирования, число верных срабатываний, необходимое для отнесения фрагмента изображения к заданному классу [7]. В нашей работе подвергалось изменению только число верных срабатываний. Все остальные параметры были принятыми равными параметрам по умолчанию, в частности, используется фильтр `haarcascade_frontalface_alt`, разработанный Линхартом в 2000-м году и поставляющийся в комплекте с OpenCV SDK.

Алгоритм SIF [1], разрабатываемый в Лаборатории анализа данных Тульского государственного университета, основан на методе опорных векторов и имеет только один параметр настройки коэффициент сдвига разделяющей два класса (лиц и нелиц) гиперплоскости в пространстве признаков.

Алгоритм FDL был разработан В. Киензле, Г. Бакир, М. Франз, В. Щолкоф (W. Kienzle, G. Bakir, M. Franz и B. Scholkopf) в Институте биологической кибернетики им. Макса Планка и основан на методе опорных векторов с потенциальной функцией на основе радиальной базисной функции Гаусса при с.к.о. равным 10 в пространстве признаков сепарабельных фильтров. Фрагменты выбираются для распознавания последовательно, по движению окна, а затем масштабируется к размеру сепарабельных фильтров [4, 5]. Алгоритм имеет только один параметр настройки, который позволяет изменять его «строгость».

Алгоритмы FSDK и UniS любезно предоставлены Luxand Inc. (`http://www.luxand.com`) и University of Surrey соответственно, являются коммерческими продуктами, и принципы их работы не разглашаются. FSDK имеет единственный параметр настройки, который позволяет изменять «строгость» алгоритма. UniS присваивает каждому найденному лицу «степень уверенности», которая изменяется от 0 до 100. В качестве параметра настройки принимается порог внутри этого диапазона.

Модель точности локализации лиц

Существуют различные модели представления лиц, например, центром лица и его радиусом, центром лица и размером его квадрата (OCV, FDL), координатами центров глаз (SIF, FSDK, UniS), опорными точками эллипса лица и т. д. В этой работе мы будем представлять лица в виде координат центров глаз, под которыми понимаются центры зрачков. Для такой модели описания лица, во-первых, представляется более удобным определение различия двух результатов поиска; во-вто-

*Работа выполнена при финансовой поддержке РФФИ, проект № 09-07-00394.

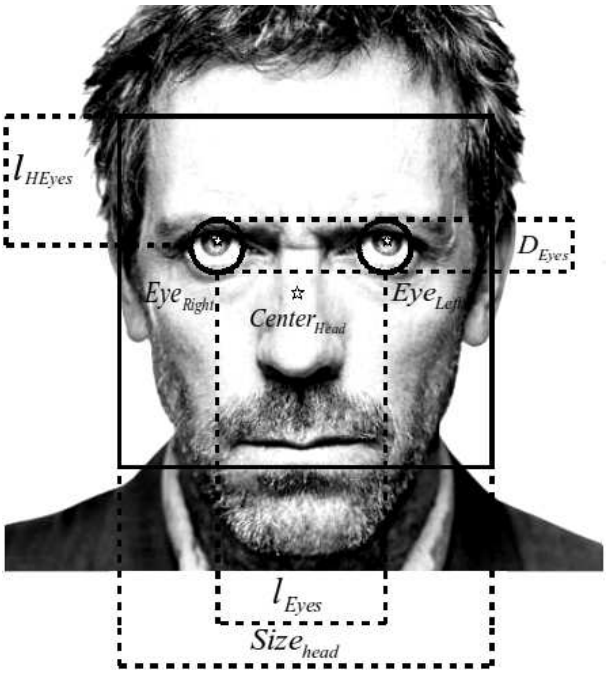


Рис. 1. Схематичное изображение лица.

рых, алгоритмы распознавания обычно требуют предварительного совмещения центров глаз обучающей выборки. Таким образом, для унификации методов сравнения мы предлагаем модель, которая, имея на входе прямоугольный фрагмент, выдавала бы координаты центров глаз. На рис. 1 показано схематичное изображения лица: мелким пунктиром представлена анатомическая разметка лица и его пропорции [3], сплошными жирными линиями — найденный фрагмент лица с его разметкой. На рис. 1 представлены следующие величины: Eye_{Left} и Eye_{Right} — абсолютные координаты правого и левого глаза соответственно; l_{Eyes} — это расстояние между центрами глаз; l_{HEyes} — расстояние от верхней границы лица до центра глаз; $Size_{Head}$ — это размер квадрата лица; D_{Eyes} — диаметр области допустимого отклонения экспериментально найденного положения глаз от действительного Eye^A_{Right} и Eye^A_{Left}; $Center_{Head}$ — абсолютная координата центра найденного лица.

Точность локализации для алгоритмов, определяющих центры глаз. В случае если алгоритм определяет центры глаз лица на изображении, будем считать, что лицо найдено правильно, если экспериментальное положение глаз попадают в области диаметром D_{Eyes}, зависящего от расстояния между глазами и параметра α, принятого в работе равным 0,25:

$$D_{Eyes} = 2\alpha \times l_{Eyes}, \quad \alpha = \text{const}\,. \tag{1}$$

Точность локализации для алгоритмов, не определяющих координаты центров глаз. Пусть имеется изображение лица в анфас, без наклона головы (рис. 1), и алгоритм определил его центр и размер — ($Center_{Head}$ и $Size_{Head}$ соответственно). Очевидно, что глаза на таком изображении располагаются симметрично относительно вертикальной оси, находясь от неё на половине расстояния между ними $l_{Eyes}/2$, и на одинаковом расстоянии l_{HEyes} от верхней границы области найденного лица. Таким образом, абсолютные координаты глаз можно найти используя соотношения:

$$\begin{aligned} Eye^y_{Right} &= Eye^y_{Left} = Center^y_{Head} + \\ &\quad + l_{HEyes} - \tfrac{1}{2} Size_{Head}, \\ Eye^x_{Right} &= Center^x_{Head} - \tfrac{1}{2} l_{Eyes}, \\ Eye^x_{Left} &= Center^x_{Head} + \tfrac{1}{2} l_{Eyes}. \end{aligned}$$

Попытаемся оценить параметры конкретного алгоритма, а именно l_{Eyes} и l_{HEyes}, как среднестатистическое для большого числа изображений, на которых экспертом были указаны координаты центров глаз. Опираясь на анализ большого числа изображений, был найден коэффициент A — среднее значение пропорции расстояния между глазами l_{Eyes} к размеру возвращаемой области лица; и коэффициент B — среднее значение пропорции расстояния от верхней границы её же до центра глаз l_{HEyes} к размеру возвращаемой области лица. Они вычисляются используя информацию об истинном положении глаз на изображении

$$A = \frac{1}{N} \sum_{i=1}^{N} \frac{l^i_{HEyes}}{Size^i_{Head}};$$

$$B = \frac{1}{N} \sum_{i=1}^{N} \frac{l^i_{Eyes}}{Size^i_{Head}};$$

где l^i_{HEyes}, l^i_{Eyes} и $Size^i_{Head}$ — соответствующие параметры, измеренные для i-го изображения в базе изображений, содержащей N объектов. Тогда координаты глаз для заданного размера лица и подобранные для конкретного алгоритма средние коэффициенты пропорций определяются как:

$$\begin{aligned} Eye^y_{Right} &= Eye^y_{Left} = \\ &= Center^y_{Head} + Size_{Head}\left(A - \tfrac{1}{2}\right); \\ Eye^x_{Right} &= Center^x_{Head} - Size_{Head}\left(\tfrac{1}{2}B\right); \\ Eye^x_{Left} &= Center^x_{Head} + Size_{Head}\left(\tfrac{1}{2}B\right). \end{aligned}$$

Когда предполагаемые координаты глаз внутри прямоугольника найдены, для определения точности локализации используется выражение (1).

В случае, если на изображении представлено лицо в анфас с наклоном головы в бок, будем определять l_{HEyes} как среднее арифметическое расстояний двух глаз от верхней границы области найденного лица, т. е.

$$l_{HEyes} = \tfrac{1}{2}\left(l^{Left}_{HEyes} + l^{Right}_{HEyes}\right).$$

Параметры оценки результатов

Если изображение содержит лицо, а алгоритм его не находит, то, очевидно, такое изображение ошибочно классифицировано (ошибка первого рода), также к ошибочно классифицированным относятся изображения, не содержащие лиц, но на которых тестируемым алгоритмом было найдено лицо (ошибка второго рода).

Результаты работы каждого алгоритма оценивались по следующим параметрам:

— *FRR* — доля ошибок первого рода, которая показывает вероятность не узнавания объекта своего класса:

$$FRR = \frac{N_{Face}^{nonFace}}{N_{Face}^{Face} + N_{Face}^{nonFace}},$$

где N_{Face}^{Face} — число объектов класса лиц, распознанных верно, $N_{Face}^{nonFace}$ — число объектов класса лиц, распознанных неверно;

— *FAR* — доля ошибки второго рода, которая показывает вероятность того, что классификатор по ошибке отнесёт объект не своего класса к объектам своего:

$$FAR = \frac{N_{nonFace}^{Face}}{N_{nonFace}^{Face} + N_{nonFace}^{nonFace}},$$

где $N_{nonFace}^{Face}$ — число объектов класса не лиц, распознанных неверно, $N_{nonFace}^{nonFace}$ — число объектов класса не лиц, распознанных верно;

— *TER* — общая доля ошибок (как первого, так и второго рода), которая показывает вероятность того, классификатор по ошибке отнесёт объект не к тому классу к которому он принадлежит:

$$TER = \frac{1}{N}\left(N_{nonFace}^{Face} + N_{Face}^{nonFace}\right).$$

Мерой различия результатов работы алгоритмов является нормированное расстояние Хемминга между парой бинарных векторов ошибок алгоритмов (компонента вектора ошибок, соответствующая одному тестовому изображению, принимает значение «единица» в том случае, когда алгоритм верно нашел область лица на этом изображении):

$$d_H(X_i, X_j) = \frac{1}{N}\sum_{s=1}^{N}\left|x_i^{(s)} - x_j^{(s)}\right|.$$

Будем называть матрицу таких величин (значения от нуля до единицы) матрицей различий алгоритмов поиска лиц.

Структура экспериментальных данных

Тестирование проводилось на следующих базах изображений:

1. *Face Place* — содержит 1247 изображений 150 человек, снятых с различных углов, разрешение 480×400 пикс., `http://www.face-place.org/`;
2. *The IMM Face Database* — 240 изображений 40 персон, разрешение 512×342 пикс., `http://www.imm.dtu.dk/~aam/`;
3. *Коллекция изображений Б. Ачерманна* (B. Achermann, Бернский университет) — содержит 300 изображений 30 персон, снятых с различных точек, разрешение 512×342 пикс., `ftp://ftp.iam.unibe.ch/pub/Images/FaceImages/`;
4. *BioID* — содержит 1520 файлов, разрешение 384×286 пикс., `http://www.humanscan.de/support/downloads/facedb.php`;
5. *Коллекция изображений лиц, созданная в Лаборатории анализа данных ТулГУ* — 4198 файлов, разрешение 320×240 пикс.;
6. *Коллекция изображений, не содержащих лиц, собранная в Лаборатории анализа данных ТулГУ* — 9510 файлов, разрешение 320×240 пикс.

Общий размер базы тестовых изображений составляет 17015 элементов.

Коэффициенты A и B определялись на базе *Georgia Tech face*, содержащей 750 изображений 50 персон, разрешение 640×480 пикселей; `http://www.anefian.com/research/face_reco.htm`.

Анализ полученных результатов

После проведения серии экспериментов, для каждого из алгоритмов были получены параметры FRR, FAR, TER (см. таблицу 3) и вектор ошибок алгоритма. Коэффициенты модели точности локализации для алгоритмов не определяющих координаты центров глаз (A и B) приводятся в таблице 1.

Таблица 1. Коэффициенты (A и B) модели точности локализации для выбранных алгоритмов.

Алгоритм	Парам. A	Парам. B
FDL	0.383	0.3332
OpenCV	0.3666	0.3858

Представим набор значений FRR и FAR, для каждого алгоритма, в виде группы точек. Каждая из таких точек будет соответствовать оценке качества распознавания изображения данным алгоритмом при определенном параметре т. е. для каждого алгоритма, мы получим ROC-кривые (рис. 2.) в стандартном для биометрических систем виде [2, 8]. ROC-кривые позволяют определить наиболее подходящий алгоритм с требуемыми параметрами работы (настройки) для конкретной ситуации. Для того, чтобы провести более детальный анализ, зафиксируем параметры работы алгоритмов, обеспечивающие наименьшую ошибку распознавания, т. е. имеющие наименьшее значение параметра TER, и построим для них матрицу различий (таблица 2).

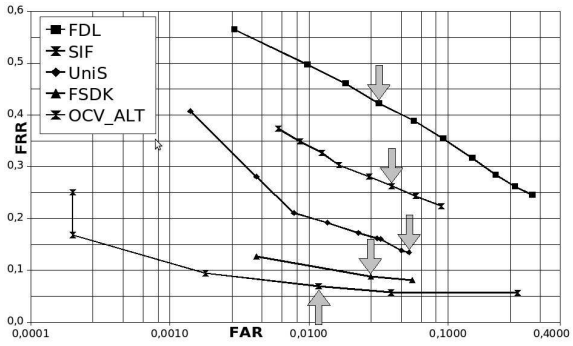


Рис. 2. Сводный график зависимости ошибок детектирования первого и второго рода протестированных алгоритмов (ROC-кривые).

Таблица 2. Матрица расстояний между алгоритмами при выбранных параметрах (в скобках).

	OCV	SIF	FDL	FSDK	UniS
OCV(3)	0	0,11	0,19	0,03	0,05
SIF(-3)		0	0,14	0,12	0,13
FDL(-3)			0	0,19	0,21
FSDK(0)				0	0,06
UniS(0)					0

Таблица 3. Значение ошибок алгоритмов при выбранных параметрах (в скобках).

Алгоритм	FRR	FAR	TER
OCV(3)	0,0695	0,0117	0,0373
SIF(-3)	0,2629	0,0393	0,1384
FDL(-3)	0,4327	0,0318	0,2041
FSDK(0)	0,0879	0,0278	0,0544
UniS(0)	0,1347	0,0523	0,0888

Значения ошибок алгоритмов при выбранных параметрах приведены в таблице 3.

Анализ элементов матрицы приводит к выводу, что наиболее похожими являются алгоритмы OCV и FSDK, а наиболее различными FPL и UniS.

Очевидно, что каждый из алгоритмов обладает уникальными особенностями детектирования. Одним из способов их числовой оценки является сравнение количества уникально правильно классифицированных изображений каждого алгоритма по сравнению с другими, с количеством «лёгких» и «трудных» случаев (см. таблицу 4).

Таблица 4. Количественные характеристики «особых» изображений в тестовой базе.

Случай	**Кол. (лиц)**	**% в базе**
«Легкие» случаи	11621 (3484)	68,29
«Трудные» случаи	171 (170)	1
только OCV	42 (41)	0,24
только SIF	14 (14)	0,08
только FDL	2 (1)	$< 0{,}01$
только FSDK	53 (53)	0,31
только UniS	106 (105)	0,62

Здесь под «легкими случаями» понимаются изображения, которые были верно классифицированы всеми методами, в обратной ситуации они считается «трудными случаями».

Заключение

В статье была разработана статистическая модель восстановления глаз по центру и размеру квадратной области лица, была предложена и продемонстрирована, на примере следующих алгоритмов: FdLib, Intel OpenCV, SIFinder, Luxand FSDK, University of Surrey, методика сравнения алгоритмов поиска лиц на изображениях. По результатам этого сравнения было обнаружено, что среди них выделяется явный лидер по качеству распознавания — OpenCV, дающий наилучшие результаты при различных параметрах. Он заметно превосходит остальные алгоритмы по качеству детектирования, но разделяет второе место в тесте на время обработки с UniS. Лидером же по этому параметру является алгоритм FDLib, но он, в свою очередь, имеет плохое качество детектирование лиц. Алгоритм SIF, разрабатываемый в ТулГУ, продемонстрировал средний уровень качества детектирования и в настоящий момент ведется работа по его улучшению. Стоит заметить, что программа UniS верно обработала больше всего «уникальных» изображений, т. е. таких, на которых другие алгоритмы не смогли верно детектировать лица. В дальнейшем планируется проведение повторных экспериментов на увеличенном объеме тестового материала, что позволит улучшить репрезентативность результатов. Также авторы надеются на сотрудничество с другими разработчиками алгоритмов поиска лиц на изображениях.

Литература

[1] *Крестинин И. А., Середин О. С.* Метод особых точек в задачах поиска лиц на графических изображениях // Известия ТулГУ, серия «Технические науки», Вып.3 — Тула: ТулГУ, 2008. — С. 218–227.

[2] *Fawcett T.* An introduction to ROC analysis //Pattern Recognition Letters. — Vol. 27, Issue 8. — P. 861–874.

[3] *Gheno D.* Construction of the head `http://www.dangheno.net/pconstruction_of_the_head1.htm`

[4] *Kienzle W.* FDLib Description `http://www.kyb.mpg.de/bs/people/kienzle/fdlib/fdlib.htm`

[5] *Kienzle W., Bakir G., Franz M., Scholkopf B.* Face detection — efficient and rank deficient // Advan. in neural inform. process. systems 17, 2005. — P. 673–680.

[6] *Landre J.* Programming with Intel IPP and Intel Open CV under GNU Linux

[7] *Pisarevsky V.* Introduction to OpenCV. Intel corp. manual.

[8] *Wechsler H.* Reliable face recognition methods: system design, implementation and evaluation // Springer, 2007. — 329 p.

Регуляризация скелета для задачи сравнения формы*

Домахина Л. Г.
Ludmila.domakhina@gmail.com
Москва, МГУ им. М. В. Ломоносова, факультет вычислительной математики и кибернетики

Данная работа имеет две основные цели: описание неустойчивых элементов (нерегулярностей) скелета и формальная постановка задачи регуляризации скелета для сравнения формы фигур. Предлагается классификация нерегулярностей скелета на три типа: рудиментные терминальные рёбра, перехлёст узлов в коротких внутренних рёбрах и рудиментные циклы. Скелетизация — некорректная задача в том смысле, что не обладает устойчивостью. Поэтому предлагается искать приближенные устойчивые скелеты. Предлагается построить регуляризирущий функционал, который использует априорную информацию о виде устойчивого решения. Задача минимизации этого функционала может быть решена на основе композиции функций, устраняющих нерегулярности скелета.

Скелетизация — некорректная задача, поэтому необходимо разработать математически строгие критерии регуляризации скелетов фигур. В приложениях это нужно для выделения фундаментальных структурных свойств формы фигур, не зависящих от шумовых эффектов и незначительных деформаций фигур. В том или ином виде в различной литературе, посвященной исследованию и использованию скелетов, рассматриваются нерегулярности (неустойчивые элементы) скелета [1, 4, 5, 6]. Тем не менее, отсутствует строгая математическая модель устойчивого скелета.

Скелет фигуры

Определение 1. *Непрерывная фигура [2] — это связная замкнутая область на плоскости, ограниченная конечным числом непересекающихся жордановых кривых.*

Определение 2. *Дискретная фигура — это максимальное связное множество черных точек на растровой решетке [4].*

Определение 3. *Скелетом непрерывной фигуры [2] называется множество центров максимальных вписанных в нее окружностей.*

Определение 4. *Непрерывным скелетом дискретной фигуры [4] называется скелет ее аппроксимирующей фигуры минимального периметра.*

Определение 5. *Фигурой F будет считаться любая дискретная или непрерывная фигура.*

Скелет можно рассматривать как планарный граф [2] — *скелетный граф*. Его вершины — центры окружностей, касающихся границы в трёх и более точках, а также терминальные точки скелета, а рёбра — серединные оси фигуры, линии, состоящие из центров окружностей, касающихся границы в двух и более точках. Вершина скелета, имеющая одно инцидентное ребро, называется *терминальной*, более одного — *узлом скелета*. Ребро, инцидентное терминальной вершине называется *терминальным*, остальные рёбра — *внутренними*.

*Работа выполнена при финансовой поддержке РФФИ, проект № 08-01-00670.

Сравнение формы на основе скелета

Одинакова ли форма двух фигур, одна из которых имеет зашумленную, а другая — гладкую границу (рис. 1а)? Одинакова ли форма фигур двух человечков, ноги и руки которых находятся в разных положениях (рис. 1б)? Одинакова ли форма двух ящериц, у одной из которых глаза выделены дыркой (рис. 1в)? Сходство форм в этих случаях видно невооруженным глазом. Возникает задача — как описать математически столь очевидное сходство? Непрерывный скелет — это тот инструмент, который хорошо подходит для описания подобного сходства форм. Тем не менее, классически определенный как множество срединных осей [4], скелет имеет ряд особенностей, которые затрудняют поставленную задачу. Для визуально сходных фигур (рис. 1) скелеты совершенно различны (рис. 2, 3, 4).

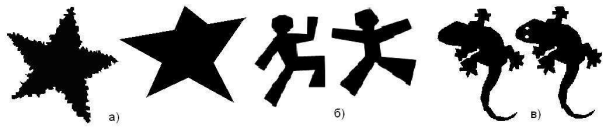


Рис. 1. Сходные или различные фигуры?

Вывод: скелетизация — *некорректная задача* [3] в том смысле, что не обладает устойчивостью. Методом решения некорректных задач является получение некоторого приближенного решения, которое было бы корректным [3]. Такой метод называется *регуляризацией по Тихонову*. Глядя на скелеты похожих фигур (рис. 2, 3, 4), можно обнаружить некоторые общие (устойчивые) элементы, откуда возникает предположение, что скелет всё-таки можно регуляризировать, то есть найти некоторый *приближенный устойчивый скелет*.

Виды нерегулярностей скелета

Необходимо определить, что такое нерегулярность в скелете. Интуитивно под нерегулярностью скелета понимается некоторый неустойчивый его элемент, который сильно влияет на скелет при незначительных изменениях фигуры. На основе описанных примеров (рис. 2, 3, 4) предлагается

Таблица 1. Виды нерегулярностей скелета.

	Вид нерегулярности	**Причины возникновения**
Ψ_1	рудиментые терминальные рёбра	неровности границы фигуры: рис. 1а, 2
Ψ_2	перехлёст внутренних вершин	короткие внутренние рёбра: рис. 1б, 3
Ψ_3	рудиментные циклы скелетного графа	изменение связности фигуры: рис. 1в, 4

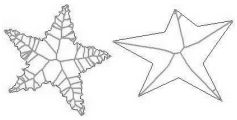


Рис. 2. Рудиментные рёбра скелетного графа.

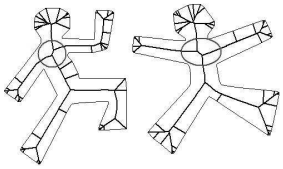


Рис. 3. Перехлёст внутренних вершин.

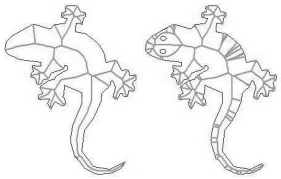


Рис. 4. Рудиментные циклы скелетного графа.

классификация нерегулярностей скелета на три типа (таблица 1).

Введём следующие обозначения: $\mathrm{ma}\colon F \to \mathrm{Sk}$ — оператор, который по фигуре строит непрерывный скелет [4]; $\mathrm{ma}(F)$ — непрерывный скелет фигуры F; $D_H(F, F_1)$ — расстояние Хаусдорфа между фигурами F и F_1.

Первая «нерегулярность» непрерывного скелета — это *терминальные шумовые рёбра*, вызванные неровностью границы фигуры (рис. 1а), не имеющие ничего общего с общей структурой фигуры (рис. 2).

Определение 6. *Терминальным рудиментным ребром скелета* $\mathrm{ma}(F)$ *фигуры* F *с точностью* ε *называется терминальное ребро* e *такое, что найдется* ε*-близкая фигура* F_1, $D_H(F, F_1) < \varepsilon$, *что в её скелете* $\mathrm{ma}(F_1)$ *отсутствует это терминальное ребро:*

$$\mathrm{ma}(F) \supseteq \mathrm{ma}(F_1), \quad \mathrm{ma}(F) \setminus \mathrm{ma}(F_1) \supseteq e.$$

Вторая «нерегулярность» кроется во внутренних рёбрах скелета при незначительных вариациях фигуры: внутренние узлы короткого ребра скелета могут поменяться местами — *перехлёст внутренних узлов скелета* (рис. 3).

Определение 7. *Рудиментным внутренним ребром скелета* $\mathrm{ma}(F)$ *фигуры* F *с точностью* ε *называется внутреннее ребро* e *такое, что найдется* ε*-близкая фигура* F_1, $D_H(F, F_1) < \varepsilon$, *что в её скелете* $\mathrm{ma}(F_1)$ *отсутствует это внутреннее ребро:*

$$\mathrm{ma}(F) \supseteq \mathrm{ma}(F_1), \quad \mathrm{ma}(F) \setminus \mathrm{ma}(F_1) \supseteq e$$

Наконец, «нерегулярность» в скелете может возникнуть из-за небольших «дырок» фигуры, которые приводят к многосвязности и серьезными изменениям топологии скелета (рис. 4) — *появлению несущественных циклов* (рис. 1в). Каждой «дырке» соответствует цикл скелетного графа. Максимальное расстояние между двумя точками «дырки» будем считать ее диаметром.

Определение 8. *Рудиментным циклом скелетного графа* $\mathrm{ma}(F)$ *фигуры* F *с точностью* ε *назовем такой цикл, который соответствует дырке, имеющей диаметр меньше фиксированного значения* ε.

Устранение нерегулярностей

Нерегулярности можно устранять на уровне исходных фигур или построенных скелетов [5], а также комбинируя эти два способа [6, 1]. Определим некоторые функции для устранения нерегулярностей скелета.

Функция устранения рудиментных терминальных рёбер. Устранение нерегулярности первого типа обычно представляет собой некоторую стрижку терминальных рёбер скелета. Например, в работе [5] выполняется стрижка всех терминальных рёбер скелета. Большинство из методов стрижки эвристические. Математически строго определенный метод — построение базового скелета с фиксированной точностью аппроксимации [1] (рис. 5). Обозначим $\Psi_1(F, \alpha)$ — оператор, который строит базовый скелет с точностью α.

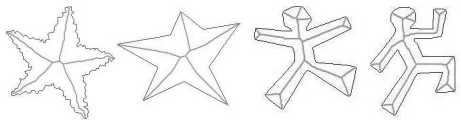


Рис. 5. Устранение рудиментных терминальных ребер.

Функция устранения перехлёстов. В работе [6] описана проблема «перехлёста» в рамках задачи поиска аппроксимирующих фигур с изоморфными скелетами. Предлагается проводить структурные изменения следующего вида (рис. 6): удаление внутренних ребер скелетного графа — так называемая склейка ребер. Обозначим $\Psi_2(F, \alpha)$ — оператор, который проводит склейку всех ребер скелета таким образом, что фигура F деформируется не более, чем на величину α в метрике Хаусдорфа (то есть склейка с точностью α).

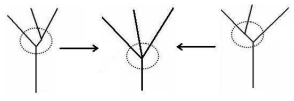


Рис. 6. Устранение внутренних коротких ребер.

Функция устранения циклов. «Очистка» от циклов в топологическом смысле выполняется довольно просто: нужно разорвать этот цикл и удалить его часть из скелета. Но по метрическим критериям этого недостаточно. Например, слон с носом потеряет часть головы со стороны разрыва цикла (рис. 7). С другой стороны, появление мелких отверстий в дискретной фигуре за счёт шумов — это обычное дело. Можно выполнять регуляризацию на уровне фигуры, для чего использовать диаметры дыр. Необходимо выставить порог по диаметру дыры и удалить контуры, окружающие мелкие отверстия, что устранит указанную нерегулярность (рис. 8). Обозначим $\Psi_3(F,\alpha)$ — оператор, который удаляет все дыры фигуры таким образом, что фигура F деформируется не более, чем на величину α в метрике Хаусдорфа (*устранение циклов с точностью* α).

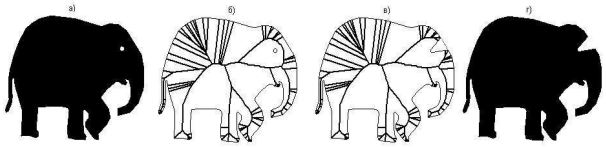


Рис. 7. Устранение циклов как преобразование скелета: а-б — скелет с циклом и фигура; в — удаление цикла; г — потеря части головы.

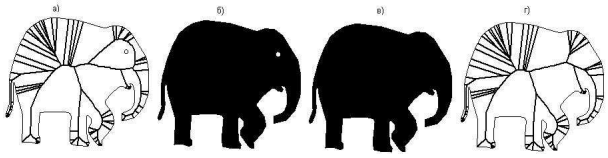


Рис. 8. Устранение циклов как преобразование фигуры: а-б — скелет с циклом и фигура; в — преобразование фигуры; г — скелет без цикла.

Задача классификации циклов на значимость непроста, так как непосредственно по скелету определить, насколько цикл значим, не очень просто. Нужно учитывать не только саму протяжённость этого цикла, но и функцию ширины (размеры кругов). Низкая алгоритмическая эффективность существующих методов работы с циклами скелета [4] добавляет трудность проведения практических исследований данного вопроса.

Регуляризация скелета по Тихонову

В фундаментальном смысле устойчивость скелета эквивалентна непрерывности оператора [7], который по фигуре строит ее скелет. Скелетный оператор должен получать на вход одну фигуру и строить устойчивый вид скелета, который при незначительных изменениях фигуры меняется незначительно.

Определение 9. *Скелет, полученный с помощью оператора* $\mathrm{Im}\colon F \to \mathrm{Sk}$ *устойчив на паре метрических пространств* (Φ,Λ) *с расстояниями* $\rho_\Phi(\cdot,\cdot)$ *и* $\rho_\Lambda(\cdot,\cdot)$*, если для всякого* $\varepsilon > 0$ *существует такое* $\delta(\varepsilon) > 0$*, что для любых двух фигур* $F_1, F_2 \in \Phi$ *из неравенства* $\rho_\Lambda(\mathrm{Sk}_1,\mathrm{Sk}_2) < \delta(\varepsilon)$ *следует неравенство* $\rho_\Phi(F_1,F_2) < \varepsilon$*, где* $\mathrm{Sk}_1 = \mathrm{Im}(F_1)$*,* $\mathrm{Sk}_2 = \mathrm{Im}(F_2)$*.*

Теорема 1. *Оператор непрерывного скелета* $\mathrm{ma}\colon F \to \mathrm{Sk}$ *неустойчив на паре метрических пространств* (Φ,Λ) *— пространство фигур и скелетных графов с расстояниями:* $\rho_\Phi(\cdot,\cdot)$ *— расстояние Хаусдорфа,* $\rho_\Lambda(\cdot,\cdot)$ *— топологическое расстояние (например, разность числа ребер скелетных графов).*

Обозначим $\mathrm{ma}^0\colon F \to \mathrm{Sk}^0$ *— однореберный скелетный оператор*, который по фигуре строит скелетный граф, состоящий из одного ребра, являющегося подграфом $\mathrm{ma}(F)$ — цепочкой максимальной длины: $\mathrm{Sk}^0 = e$ (рис. 9в).

Теорема 2. *Однореберный скелетный оператор* $\mathrm{ma}^0\colon F \to \mathrm{Sk}$ *устойчив на паре метрических пространств* (Φ,Λ) *с расстояниями* $\rho_\Phi(\cdot,\cdot)$ *и* $\rho_\Lambda(\cdot,\cdot)$ *— теми же, что и в теореме 1.*

Однореберный скелетный оператор $\mathrm{ma}^0(F)$ для задач сравнения формы не несет в себе достаточной информации, хотя может быть использован как признак фигуры. Оператор $\mathrm{ma}(F)$ неустойчив, что делает его для задач сравнения формы также непригодным. Необходимо найти какой-то промежуточный скелетный оператор (рис. 9б) между неустойчивым, содержащим в себе «лишнюю» информацию ($\mathrm{ma}(F)$ — рис. 9а) и устойчивым, но содержащим в себе мало информации ($\mathrm{ma}^0(F)$ — рис. 9в). Для этого предлагается использовать регуляризацию по Тихонову [3].

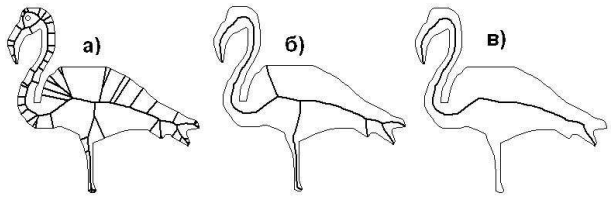


Рис. 9. Регуляризация скелета: а — непрерывный скелет; б — промежуточный скелет; в — устойчивое решение.

С каждой точкой скелета можно связать радиус максимального пустого круга, центром которого данная точка является, то есть задать *граничное скелетное представление фигуры*. По такому представлению можно однозначно восстанавливать исходную фигуру. Это дает возможность определить обратный оператор скелетизации: $\mathrm{ma}^{-1}(sk) = F$.

Определение 10. *Функционал*

$$\Omega(sk,\alpha) = \rho_\Phi(F, \mathrm{ma}^{-1}(sk))^2 + \alpha\rho_\Lambda(sk, sk^0)^2$$

называется функционалом Тихонова для задачи $\mathrm{ma}^{-1}(sk) = F$*, где* $sk \in \Lambda$ *— планарный скелетный*

граф, $sk^0 = \mathrm{ma}^0(F)$ — результат действия устойчивого одноребернoго скелетного оператора, ρ_Λ — топологическая мера сходства скелетов, ρ_Φ — расстояние Хаусдорфа.

Задача 1. Регуляризация скелета как задача минимизации тихоновского функционала:

$$\Omega(sk, \alpha) \to \min_{sk \in \Lambda}.$$

При малых значениях параметра α решение этой задачи близко к исходной некорректной задаче. При больших α решение устойчивое, но находится дальше от исходной задачи. Приближенный скелет sk^α, найденный как минимум функции $\Omega(sk, \alpha)$, будет зависеть от параметра α.

Это позволяет выдвинуть гипотезу о полноте системы функций $\Psi_1(F, \alpha)$, $\Psi_2(F, \alpha)$, $\Psi_3(F, \alpha)$, устраняющих нерегулярности трёх типов.

Гипотеза 1. Для любой фигуры F и заданного α найдутся такие параметры α_1, α_2, α_3, что решение задачи минимизации тихоновского функционала sk^α может быть найдено как комбинация функций, устраняющих нерегулярности трех типов с точностями α_1, α_2, α_3 соответственно:

$$sk^\alpha = \Psi_1(F, \alpha_1) \circ \Psi_2(F, \alpha_2) \circ \Psi_3(F, \alpha_3).$$

Регуляризация скелета для сравнения формы фигур

Задача сравнения формы фигур требует регуляризации скелета (рис. 1), но для фиксированной пары фигур можно упростить фундаментальную постановку регуляризации скелета в терминах «подгонки скелетов».

Задача 2. Регуляризация скелета для сравнения формы. Для заданных двух фигур F_1 и F_2 найти в некотором смысле наилучшие скелеты фигур F_1 и F_2, близкие в некоторой метрике к непрерывным скелетам фигур F_1 и F_2. То есть построить *регуляризирущий оператор* на основе двух фиксированных фигур $\mathrm{Re}\,(F_1, F_2) \to (\mathrm{Sk}_1, \mathrm{Sk}_2)$.

Решение похожей задачи без учета нерегулярности с циклами Ψ_3 приведено в работе [6], где поставлена и решена задача поиска аппроксимирующих фигур с изоморфными скелетами. Аналогично решается и задача 2. Например, наилучшими скелетами можно считать изоморфные скелеты некоторых двух фигур, близких по расстоянию Хаусдорфа к исходным F_1 и F_2: $\mathrm{Sk}_1 \cong \mathrm{Sk}_2$.

Возникает вопрос о необходимости решения такой сложной задачи как задача 1 для сравнения формы. Возможно, достаточно решения задачи 2? Актуальность решения задачи 1 может быть обоснована с точки зрения оптимальности вычислений для больших баз фигур.

Пусть имеется база из n фигур. Для сравнения фигур в задаче 2 необходимо построить $O\left(\frac{n^2}{2} + n\right)$ скелетов. В задаче 1 количество построенных скелетов порядка $O(2n)$. Таким образом, задача минимизации регуляризирующего оператора актуальна и для задачи сравнения формы.

Выводы

В настоящей работе описаны виды нерегулярностей скелета: рудиментные терминальные рёбра, перехлёст узлов в коротких внутренних рёбрах и рудиментные циклы. Предложены функции устранения указанных нерегулярностей как преобразование исходных фигур или их скелетов, или комбинации этих двух вариантов.

Предложена фундаментальная постановка задачи регуляризации скелета как задача минимизации функционала Тихонова с параметром регуляризации α и априорной информацией об устойчивом виде скелета. Предложена гипотеза о представлении решения поставленной задачи на основе функций устранения нерегулярностей трех типов. Нерешенными вопросами остаются подтверждение выдвинутой гипотезы либо численный метод решения поставленной задачи минимизации.

Рассмотрена упрощенная формулировка регуляризации скелета для задачи сравнения формы. В последней фиксируется пара фигур и на ее основе ищутся наилучшая пара скелетов, например, на основе их изоморфизма.

Тем не менее, фундаментальная регуляризация скелета актуальна и для задачи сравнения формы.

Литература

[1] *Местецкий Л. М., Рейер И. А.* Непрерывное скелетное представление изображения с контролируемой точностью // Труды 15 международной конференции ГРАФИКОН-2003. — С. 246–249.

[2] *Choi H. I., Choi S. W., Moon H. P.* Mathematical theory of medial axis transform // Pacific. J. of Math. 1997. — Vol. 181, № 1. — P. 57–88.

[3] *Тихонов А. Н., Арсенин В. Я.* Методы решения некорректных задач — М.: Наука, 1986.

[4] *Местецкий Л. М.* Непрерывная морфология бинарных изображений. Фигуры. Скелеты. Циркуляры. // М.: ФИЗМАТЛИТ, 2009.

[5] *Mirela Tanase* Shape Decomposition and Retrieval // PhD Thesis, Utrecht University — 2005.

[6] *Domakhina L., Okhlopkov A.* Изоморфные скелеты растровых изображений // Труды 18 международной конференции ГРАФИКОН-2003, Москва.

[7] *Виноградов И. М.* Устойчивости теория // Математическая энциклопедия, Т. 5. — 1977. — С. 551–553.

Параметрическое семейство гранично-скелетных моделей формы*

Жукова К. В., Рейер И. А.
kz@pisem.net,reyer@forecsys.ru
Москва, Вычислительный центр РАН

В работе рассматривается семейство гранично-скелетных моделей формы объекта. В основе такой модели лежит комбинация базового скелета аппроксимирующей объект многоугольной фигуры и границы множества базовых кругов, позволяющая анализировать граничные и структурные особенности формы, проявляющиеся при различных уровнях детализации. Исследуются свойства монотонности и непрерывности базового скелета и описывается идея построения множества гранично-скелетных моделей для некоторого набора или диапазона значений точности аппроксимации.

Для решения широкого круга задач машинного зрения требуется анализ свойств формы, проявляющихся на разных уровнях детализации. В настоящее время для описания и анализа свойств формы объекта часто используется концепция масштабируемой кривизны границы (curvature scale space) [1, 2, 3]. Этот подход основан на аппроксимации границы кусочно-гладкой кривой, сглаживании этой кривой и выявлении экстремумов или нулей кривизны границы при разных степенях сглаживания.

В работе предлагается подход, который позволяет работать с масштабируемым представлением формы. В его основе лежит построение параметрического семейства гранично-скелетных моделей формы, позволяющих проводить совместный анализ как контурных, так и структурных свойств. Гранично-скелетная модель представляет собой структуру, состоящую из взаимосвязанных граничного и скелетного описаний формы объекта [4]. Для получения граничного описания используется аппроксимация дискретного образа многоугольной фигурой. Скелетное описание задает форму как множество серединных осей, образованных центрами максимальных кругов, целиком лежащих внутри многоугольной фигуры (так называемых максимальных пустых кругов). В качестве скелетного описания в разработанной модели используется так называемый базовый скелет [5] — подмножество скелета многоугольной фигуры, аппроксимирующее с известной точностью фундаментальную часть скелета любой замкнутой области, близкой фигуре в смысле расстояния Хаусдорфа.

Базовый скелет многоугольной фигуры

Напомним основные моменты концепции базового скелета, представленной в [5]. При этом будем рассматривать случай, когда многоугольная фигура является односвязной.

Пусть P — односвязная многоугольная фигура, ε — некоторое неотрицательное число.

Определение 1. *Круг C называется ε-допустимым кругом для P, если:*

1) *расстояние Хаусдорфа между областями P и $P \bigcup C$: $H(P, P \bigcup C) \leqslant \varepsilon$;*
2) *расстояние Хаусдорфа между границами областей $H\big(\partial P, \partial(P \bigcup C)\big) \leqslant \varepsilon$.*

Определение 2. *Круг C называется максимальным ε-допустимым кругом для P, если:*

1) *C является ε-допустимым кругом для P;*
2) *C не содержится целиком ни в каком другом ε-допустимом для P круге.*

Справедливы следующие утверждения.

Утверждение 1. *Если $C = (p, r)$ — максимальный ε-допустимый круг для P, то $C' = (p, r - \varepsilon)$ — максимальный пустой круг для P.*

Утверждение 2. *Если $C = (p, r)$ — максимальный пустой круг для P, то $C' = (p, r + \varepsilon)$ — максимальный ε-допустимый круг для P.*

Следствием этих утверждений является

Теорема 3. *Множество центров максимальных ε-допустимых кругов для P совпадает со множеством центров максимальных пустых кругов для P.*

Определение 3. *Круг C называется базовым кругом для многоугольной фигуры P, если выполнено следующее:*

1) *круг C является максимальным ε-допустимым кругом для P;*
2) *пусть точки, в которых максимальный пустой круг C', соответствующий кругу C, касается границы фигуры, разбивают границу на фрагменты $P_1, \ldots, P_n$, $n \geqslant 2$, а радиусы круга C, проходящие через эти точки, разбивают окружность круга C на дуги $L_1, \ldots, L_n$; существуют i: $1 \leqslant i \leqslant n$, j: $1 \leqslant j \leqslant n$, $i \neq j$, такие, что $H(P_i, L_i) \geqslant \varepsilon$ и $H(P_j, L_j) \geqslant \varepsilon$ (рис. 1).*

Определение 4. *Базовым скелетом многоугольной фигуры P называется множество центров всех базовых кругов области.*

Из Теоремы 3 следует, что базовый скелет P является подмножеством скелета P.

*Работа выполнена при поддержке РФФИ, проекты № 08-07-00338 и № 08-01-00670.

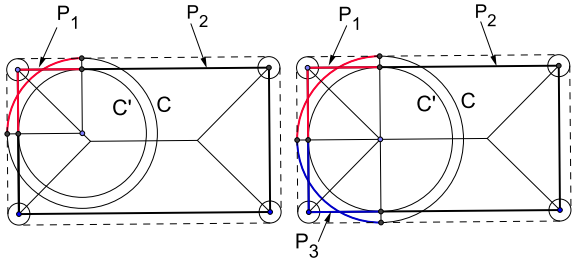


Рис. 1.

Определение 5. *Будем называть замкнутую область Ω, имеющую кусочно-гладкую границу, ε-близкой замкнутой области с кусочно-гладкой границей Ω', если выполнены следующие условия:*

1) *расстояние Хаусдорфа между областями Ω и Ω': $H(\Omega, \Omega') \leqslant \varepsilon$;*
2) *расстояние Хаусдорфа между границами областей: $H(\partial\Omega, \partial\Omega') \leqslant \varepsilon$.*

Базовый скелет односвязной многоугольной фигуры, соответствующий точности аппроксимации ε, обладает следующим свойством: для каждого ребра базового скелета существует некоторое $\delta(\varepsilon)$, такое, что в δ-окрестности этого ребра находится ветвь скелета любой замкнутой односвязной области, являющейся ε-близкой фигуре (рис. 2). Максимальные пустые окружности этой

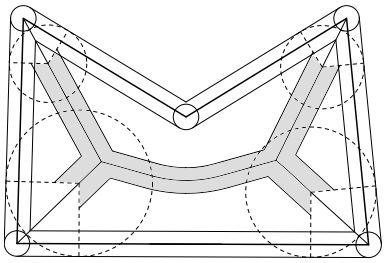


Рис. 2.

ветви касаются сегментов границы области, проходящих через ε-окрестности фрагментов границы многоугольной фигуры, бисектором которых является данное ребро базового скелета. Таким образом, можно говорить, что каждое ребро базового скелета фигуры аппроксимирует некоторую ветвь скелета любой ε-близкой этой фигуре замкнутой односвязной области.

Монотонность и непрерывность базового скелета

Скелет односвязной многоугольной фигуры представляет собой плоский граф без циклов [6]. При этом точки скелета могут быть трех типов: терминальные вершины (они совпадают с вершинами границы), нетерминальные вершины и внутренние точки ребер. Для любой точки скелета определен максимальный пустой круг с центром в этой точке. Для терминальных вершин точка касания соответствующего максимального пустого круга единственна и совпадает с самой вершиной; для внутренних точек ребер максимальный пустой круг касается границы в двух точках; для нетерминальных вершин скелета — в $k \geqslant 2$ точках. Таким образом, для каждой точки O скелета граница фигуры разбивается точками касания максимального пустого круга с центром в O на $n \geqslant 2$ фрагментов (в случае терминальной вершины единственную точку касания тоже считаем фрагментом границы). Для фиксированного ε рассмотрим максимальный ε-допустимый круг с центром в O. Радиусы, проведенные через точки касания максимального пустого круга, разбивают окружность максимального ε-допустимого круга на n дуг, соответствующих фрагментам границы. Для каждой пары соответствующих множеств «дуга-фрагмент границы» определено расстояние Хаусдорфа между ними. Если существует две пары таких множеств, для которых расстояние Хаусдорфа между элементами пары больше либо равно ε, то точка O принадлежит базовому скелету. Поскольку граница представляет собой замкнутую ломаную, то для вычисления расстояния Хаусдорфа между элементами пары «дуга-фрагмент границы» достаточно знать расстояния от дуги до вершин фрагмента границы.

Исследуем, как изменяется базовый скелет в зависимости от точности аппроксимации ε. При ε=0 все точки скелета являются базовыми. При увеличении ε процесс «выпадения» точек из базового скелета начнется от терминальных вершин скелета. Пусть O — некоторая точка скелета, C — максимальный пустой круг с центром в точке O радиуса r, C_ε — максимальный ε-допустимый круг с центром в точке O и радиусом $r + \varepsilon$.

Обозначим U_i, $i = 1, \ldots, n$ — подмножества вершин границы, принадлежащих фрагментам, на которые разбивается граница точками касания круга C. Рассмотрим максимальное расстояние от точки O до точек из множества U_i:

$$d_i = \max\bigl\{d(O, u) | u \in U_i\bigr\}.$$

Упорядочим расстояния d_i, $i = 1, \ldots, n$, по возрастанию:

$$d_{i_1} \leqslant d_{i_2} \leqslant \ldots \leqslant d_{i_{n-1}} \leqslant d_{i_n},$$

и выберем такое подмножество U_j, что соответствующее расстояние d_j является вторым по величине. Если таких подмножеств несколько, рассмотрим любое из них. Если же $d_1 = d_2 = \cdots = d_{n-1} = d_n$, то выберем любое из подмножеств U_i. В дальнейшем выбранное подмножество вершин границы будем обозначать U', а соответствующее максимальное расстояние от точки O до точек U' — d'.

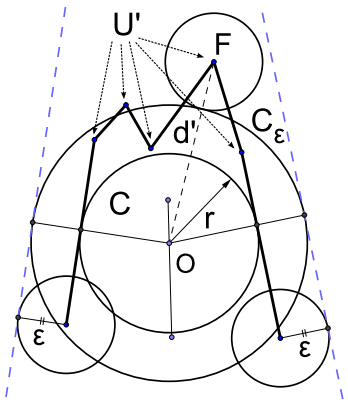


Рис. 3.

Очевидно, что при $\varepsilon > \frac{1}{2}(d' - r)$ (рис. 3) точка O не будет базовой, так как нарушается условие 2 Определения 3. Значит, существует такое ε, при котором точка O «выпадает» из базового скелета. Пусть при $\varepsilon = \varepsilon_1$ максимальный ε-допустимый круг C_{ε_1} с центром в точке O не базовый. Докажем, что для любого $\varepsilon_2 > \varepsilon_1$ соответствующий максимальный ε_2-допустимый круг C_{ε_2} с центром в O также не является базовым. Так как круг C_{ε_1} не базовый, то для $k \geqslant n-1$ дуг окружности C_{ε_1} расстояние Хаусдорфа между дугой и соответствующим фрагментом границы меньше ε_1. Рассмотрим любую из таких дуг. Обозначим эту дугу через L_1, а соответствующий фрагмент границы P — через P_{12}. Соответственно, $H(L_1, P_{12}) < \varepsilon_1$. Пусть L_2 — дуга окружности C_{ε_2}, образуемая радиусами C_{ε_2}, проходящими через те же точки касания соответствующего максимального пустого круга, что и радиусы C_{ε_1}, образующие дугу L_1. Поскольку для расстояния Хаусдорфа выполняется неравенство треугольника $H(L_2, P_{12}) \leqslant H(L_2, L_1) + H(L_1, P_{12})$, то $H(L_2, P_{12}) < (\varepsilon_2 - \varepsilon_1) + \varepsilon_1 = \varepsilon_2$. Получаем, что для $k \geqslant n-1$ дуг окружности C_{ε_2} расстояние Хаусдорфа между дугой и соответствующим фрагментом границы меньше ε_2, то есть круг C_{ε_2} тоже не является базовым. Это означает, что если ε достигло значения, при котором точка перестает быть базовой, то при всех последующих значениях ε эта точка также не будет принадлежать базовому скелету. Таким образом, базовый скелет, соответствующий точности $\varepsilon_2 > \varepsilon_1$, будет подмножеством базового скелета точности ε_1. Из этих рассуждений следует

Теорема 4. *Базовый скелет односвязной многоугольной фигуры монотонно зависит от точности аппроксимации ε.*

Итак, при $\varepsilon = \frac{1}{2}(d' - r)$ точка O является терминальной вершиной базового скелета. Исследуем изменение базового скелета при росте ε. Ребро скелета многоугольной фигуры может быть трех типов: отрезок, порожденный парой сайтов-сегментов; отрезок, порожденный парой сайтов-точек; фрагмент параболы, порожденный парой сайтов «точка-сегмент». Рассмотрим, как ведет себя терминальная точка базового скелета в каждом из этих случаев.

Пусть AB и CD — два сегмента границы, точка O принадлежит бисектору этой пары, $\varepsilon > 0$ такое, что точка O является терминальной точкой базового скелета (рис. 4). Не ограничивая общности, будем считать, что между вершинам A и C расположен фрагмент границы, вершины которого составляют множество U' для точки O. Пусть F — наиболее удаленная от точки O вершина границы из множества U'.

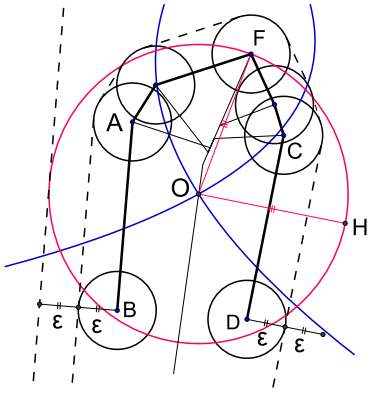


Рис. 4.

Пусть R — радиус базового круга с центром в точке O. Рассмотрим окружность радиуса $R+\varepsilon$ с центром в O. Нетрудно видеть, что эта окружность проходит через точку F. Пусть OH — радиус этой окружности, перпендикулярный сегменту CD. Так как $OF{=}OH$, то точка O лежит на параболе с фокусом F и директрисой, проходящей через точку H и параллельной сегменту границы CD (аналогично рассуждая, видим, что точка O принадлежит параболе с фокусом F и директрисой, параллельной сегменту AB и лежащей на расстоянии 2ε от AB). Таким образом, терминальная точка базового скелета лежит на пересечении параболы и ребра скелета. Увеличим ε на некоторую достаточно малую величину Δ. Пусть O_1 — терминальная вершина базового скелета точности $\varepsilon + \Delta$, R_1 — радиус базового круга с центром в O_1. Тогда $R_1 = R + \Delta$. Точка O_1 будет точкой пересечения ребра скелета и параболы с фокусом в той же точке F. Директриса этой параболы параллельна сегменту CD и лежит на расстоянии $2(\varepsilon + \Delta)$ от него. В системе координат с центром в фокусе F и осью абсцисс, параллельной сегменту CD, параболы при разной точности ε имеют один вид $y = x^2/4(c + \varepsilon) - (c + \varepsilon)$. Отсюда видим, что при увеличении ε директриса удаляется от фокуса, ветви параболы «расходятся» и ребро скелета «стирается» точкой пересечения с параболой.

Рассмотрим теперь случай, когда парой сайтов являются сегмент границы AB и вершина D (рис. 5).

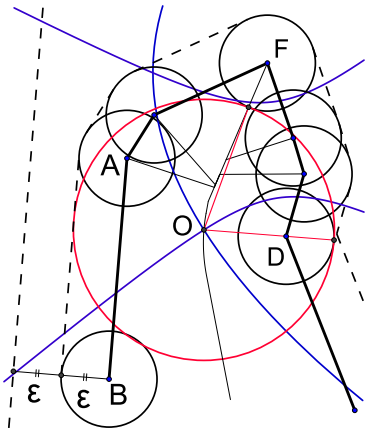


Рис. 5.

Рассуждая аналогично, получим для сегмента AB параболу с фокусом F и директрисой, параллельной AB и лежащей на расстоянии 2ε от AB. Рассмотрим базовый круг с центром в O и радиусом R. Очевидно, что $R = \mathrm{OF} - \varepsilon = \mathrm{OD} + \varepsilon$. Следовательно, $\mathrm{OF} - \mathrm{OD} = 2\varepsilon$, то есть разность расстояний постоянна. Значит, точка O лежит на гиперболе с фокусами F и D и расстоянием между вершинами 2ε.

Теперь рассмотрим ситуацию, когда оба сайта являются вершинами границы B и D (рис. 6).

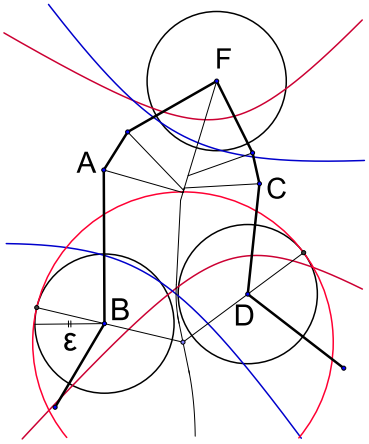


Рис. 6.

Нетрудно видеть, что в данном случае центр базового круга лежит на пересечении ветвей двух гипербол — с фокусами $\{B, F\}$ и $\{D, F\}$ соответственно. Расстояние между вершинами у гипербол одинаково и равно 2ε.

Из приведенных рассуждений следует, что при росте ε происходит непрерывное (в пределах ребра) «стирание» кривой (параболой или гиперболой) ветвей скелета.

При описании движения терминальной точки мы предполагали, что в пределах ребра самая удаленная от точки скелета вершина из подмножества вершин границы U' постоянна. Однако, возможны ситуации, когда при «стирании» ребра наиболее удаленная вершина меняется и, соответственно, происходит смена пары стирающих кривых. Так, в примере, приведенном на рис. 7, для точки O_1 скелетного ребра наиболее удаленной вершиной является вершина F, а для точки O_2 того же ребра — вершина E.

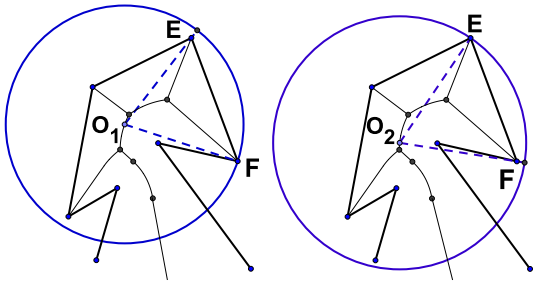


Рис. 7.

Каждой вершине границы соответствует «зона дальности» – множество точек, расстояние до которых от этой вершины больше, чем от любой другой. Такое разбиение плоскости называется диаграммой Вороного дальней точки [7].

Диаграмма Вороного дальней точки подмножества вершин границы U' определяет для каждой вершины из U' зону дальности. Если ребро скелета целиком лежит в одной зоне дальности, то для всех точек ребра вершина, определяющая стирающие кривые, единственна. В противном случае ребро разбивается на несколько фрагментов, каждый со своей определяющей вершиной, и при движении терминальной точки происходит смена стирающих кривых (рис. 8).

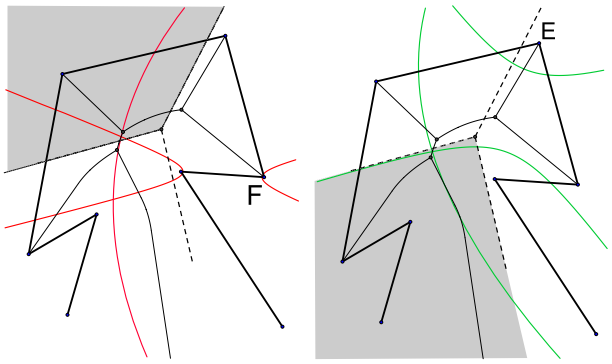


Рис. 8.

Таким образом, при росте ε скелет многоугольной фигуры «стирается» набором кривых (парабол и гипербол). При этом для скелета фигуры определен набор точек, в которых происходит смена

стирающей кривой. К ним относятся нетерминальные вершины скелета и внутренние точки скелетных ребер, лежащие на пересечении с ребрами соответствующих диаграмм Вороного дальней точки. Следует также отметить, что если в процессе «стирания» при некотором значении точности вершина скелета меняет степень (целиком стирается одно из ребер, выходящих из этой вершины), то эта точка при дальнейшем увеличении точности не «выпадет» из базового скелета до тех пор, пока не станет терминальной.

Если рассматривать расстояние Хаусдорфа между базовыми скелетами многоугольной фигуры, соответствующими различным значениям точности аппроксимации, то можно сказать, что малому изменению точности аппроксимации соответствует малое изменение базового скелета и справедлива следующая теорема:

Теорема 5. *Базовый скелет односвязной многоугольной фигуры непрерывно зависит от точности аппроксимации ε в смысле расстояния Хаусдорфа.*

Параметрическое семейство гранично-скелетных моделей

Итак, с многоугольной фигурой связано семейство базовых скелетов, монотонно и непрерывно зависящее от величины параметра ε. При этом границу объединения множества всех базовых кругов можно рассматривать в качестве модели контура фигуры, отражающей те свойства границы, которые являются существенными в пределах точности аппроксимации. Таким образом, получаем параметрическое семейство гранично-скелетных моделей. В качестве масштабируемого представления формы можно использовать множество гранично-скелетных моделей из семейства, соответствующее некоторому набору или диапазону значений точности аппроксимации.

Следует также отметить, что, аналогично модели масштабируемой кривизны границы, предложенное представление может быть использовано для нахождения экстремумов кривизны контура. В самом деле, дуга базовой окружности с центром в терминальной вершине базового скелета аппроксимирует с известной точностью соответствующий участок границы. Следовательно, этот участок можно рассматривать в качестве локального максимума кривизны контура в пределах точности аппроксимации.

Эта идея, в частности, лежит в основе решения задачи выделения линии профиля на изображениях лица для последующей биометрической идентификации [8]. Строится гранично-скелетная модель растрового образа и проводится анализ расположения выпуклых и вогнутых особенностей контура. Для анализа кривизны границы используются модели внутренней и внешней областей контура. Дуги базовых окружностей в терминальных вершинах базового скелета внутренней области соответствуют выпуклостям, а дуги базовых окружностей внешней области — вогнутостям границы.

Выводы

В работе предложена концепция представления формы объектов на изображении, которую можно использовать для анализа свойств формы, проявляющихся при различных значениях точности аппроксимации. Описано параметрическое семейство гранично-скелетных моделей, строящихся на основе аппроксимирующей объект многоугольной фигуры и состоящих из базового скелета фигуры и границы объединения множества базовых кругов. Свойства монотонности и непрерывности базового скелета позволяют получить множество гранично-скелетных моделей, соответствующих некоторому набору или диапазону значений точности. Предложенный способ представления позволяет эффективно выделять и анализировать особенности кривизны контура без аппроксимации границы кривыми высших порядков. Метод построения семейства моделей допускает обобщение на тот случай, когда аппроксимирующая фигура имеет внутренние контуры.

Литература

[1] *Abbasi S., Mokhtarian F., Kittler J.* Curvature scale space image in shape similarity retrieval, MultiMedia Systems, Vol.7, 1999, Pp. 467–476.

[2] *Dudek G., Tsotsos J. K.* Shape representation and recognition from mutliscale curvature, Computer Vision and Image Understanding, Vol. 68, No. 2, 1997, Pp. 170–189.

[3] *Ray B. K., Pandyan R.* ACORD – an adaptive corner detector for planar curves, Pattern Recognition, Vol.36, 2003, Pp. 703–708.

[4] *Местецкий Л. М.* Непрерывная морфология бинарных изображений: фигуры, скелеты, циркуляры — Москва: Физматлит, 2009. — 288 с.

[5] *Местецкий Л. М., Рейер И. А.* Непрерывное скелетное представление изображения с контролируемой точностью // Труды 13 международной конф. ГРАФИКОН-2003, Москва, 2003, С. 246–249.

[6] *Choi H. I., Choi S. W., Moon H. P.* Mathematical Theory of Medial Axis Transform // Pathific Journal of Mathematics, Vol.181, No. 1, 1997.

[7] *Препарата Ф., Шеймос М.* Вычислительная геометрия: введение — Москва: Мир, 1989. — 478 с.

[8] *Жукова К. В., Рейер И. А.* Выделение линии профиля по опорным точкам с применением базового скелета // Всеросс. конф. ММРО-13, 2007, С. 323–328.

Оценка качества JPEG2000 изображений

Зараменский Д. А., Хрящев В. В.
connect@piclab.ru
Ярославль, ЯрГУ

Описаны алгоритмы оценки качества сжатых изображений стандарта JPEG2000. Алгоритм, действующий в вейвлет-области, является неэталонным и основан на изменениях в статистической модели изображения, вызванных квантованием вейвлет-коэффициентов. Алгоритмы, действующие в пространственной области, основаны на анализе границ в изображении. Проведен сравнительный анализ предложенных алгоритмов с двумя известными эталонными методами оценки качества: пиковым отношением сигнал/шум и универсальным индексом качества. Приведены примеры оценки качества тестовых изображений.

Цель кодирования изображений состоит в минимизации искажения сжатого изображения для данного отношения бит/пиксель (или, минимизации отношения бит/пиксель при данном уровне искажения). Эта задача требует наличия методов для точного измерения искажений или качества кодированного изображения. Искажение обычно оценивается с помощью таких метрик, как пиковое отношение сигнал/шум (ПОСШ) или универсальный индекс качества (УИК) [1]. Данные метрики измеряют отличия искаженного изображения от оригинала, но не различают конкретные типы искажений. Например, зашумленное и размытое изображения могут иметь одинаковое ПОСШ. Поэтому данные метрики не всегда позволяют оценить степень конкретных искажений и служить основанием для выбора параметров кодера [2]. Кроме того стандартные метрики требуют наличия эталона изображения, что не всегда возможно. Поэтому требуются метрики количественной оценки искажений для более точной и, по возможности, неэталонной оценки качества изображений. Конечная цель подобных исследований — создание кодера оптимального с учетом современных метрик.

Алгоритм JPEG2000 [2] сжимает изображение с потерями, используя разложение по биортогональному 9/7 вейвлет-базису [3]. Коэффициенты дискретного вейвлет-преобразования (ДВП) квантуются с помощью скалярного квантователя. В результате восстановленное из квантованных ДВП-коэффициентов изображение содержит такие типы искажений, как размытые границы и звон. Размытие возникает по причине затухания высоких частот в изображении и характеризуется расплыванием границ и общей потерей детальности. Явление звона вызвано квантованием высокочастотных коэффициентов и проявляется в виде ряби около резких границ на изображении. В связи с расширяющимся ростом приложений, использующих стандарт JPEG2000, вопросы оценки искажений, вносимых при вейвлет-сжатии изображений, широко освещаются в современной англоязычной литературе в области ЦОИ.

Так, например, Огуз (Oguz) предложил метрику оценки видимых искажений около границ (Visible Ringing Measure, VRM) [4]. Алгоритм основан на создании маски, которая показывает только те области изображения, которые находятся в окрестности четко различимых границ. К сожалению, автор не рассматривает вопрос соответствия предложенной метрики VRM с субъективными визуальными оценками.

Ли (Li) разработал алгоритм [5], целью которого является неэталонное измерение нескольких искажений, представленных в изображении: общей размытости, реакции на воздействие аддитивных белого и импульсного шума, блочности и звона. Измерение искажений приводится только для одного изображения, сравнения с субъективными оценками в работе Ли не представлены.

В данной работе описаны новые алгоритмы оценки уровня размытия, звона и общего качества изображений, сжатых при помощи ДВП. Метрики размытия и звона определены в пространственной области и основаны на анализе границ в изображении, индекс качества определен в вейвлет области и основан на статистической модели изображения. В ходе психометричекого эксперимента было проведено сравнение предложенных методов с субъективным критерием качества и объективными общепринятыми критериями.

Алгоритм оценки уровня размытия

Вследствие процедуры сжатия границы изображения размываются. Поэтому, предлагаемый алгоритм измерения размытия основан на измерении ширины границ. Его схема представлена на рис. 1. Первый шаг заключается в применении детектора границ к яркостной компоненте оригинального изображения. В качестве детектора границ используется фильтр Собеля. Шум и незначительные границы удаляются путем гибкой настройки порога. На следующем шаге мы накладываем выделенные границы на искаженное (сжатое) изображение и сканируем в нем каждую строку. Для каждого пикселя границы в строке изображения определяются начальная и конечная точки границы — как точки ближайших к пикселю локальных экстремумов яркости в данной строке. В процессе определения локальных экстремумов отфиль-

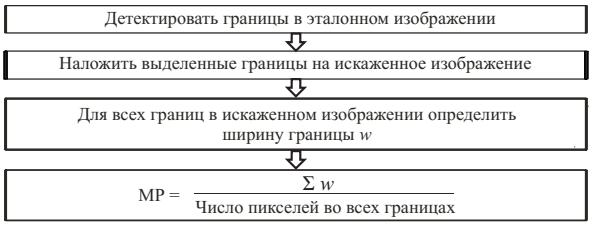


Рис. 1. Схема эталонной метрики размытия.

тровываются ошибочно детектированные границы. Для каждого пикселя границы определим ширину границы в данной строке w как расстояние между начальной и конечной точками границы и назовем локальным уровнем размытия. Метрика размытия (MP) определяется путем усреднения всех локальных уровней размытия всех границ, найденных в изображении: $\mathrm{MP} = \left(\sum w\right) \,/\, \left(\sum n\right)$, где n — число пикселей в границе.

В алгоритме, описанном выше, в целях ускорения работы рассматриваются только вертикальные границы. Таким образом, учитывается только горизонтально направленное размытие границ. Алгоритм легко может быть расширен для учета горизонтальных границ, путем фильтрации горизонтальным фильтром Собеля и сканированием каждой колонки. Тестирование алгоритма показало, что указанный прием не делает общую оценку размытия более точной.

Данный алгоритм, имеет как эталонную, так и неэталонную реализацию. В эталонной реализации положение границ определяется в оригинальном изображении. При неэталонной реализации метрики размытия, положение границ следует определять в сжатом изображении. Это в некоторой степени влияет на точность определения границ (в зависимости от степени сжатия или искажения).

Алгоритм оценки уровня звона

Схема алгоритма представлена на рис. 2. Так же, как и в случае метрики размытия границ, метрика звона определяется для каждого пикселя p выделенной границы. Алгоритм осуществляет поиск вертикальных границ в оригинальном изображении (слабые границы и шум удаляются гибкой настройкой порога) и подсчитывает начальную и конечную точки и ширину границы w для каждого пикселя границы в каждой строке сжатого изображения (аналогично метрике размытия). Затем сканируется каждая строка в сжатом изображении, и измеряется звон в окрестности пикселя границы p. Длину отрезка от начальной точки границы до p назовем левой шириной границы. Длину отрезка от p до конечной точки границы назовем правой шириной границы. Мы можем определить левый и правый звон по отношению к p. Для этого определяется ширина звона (левая и правая) как

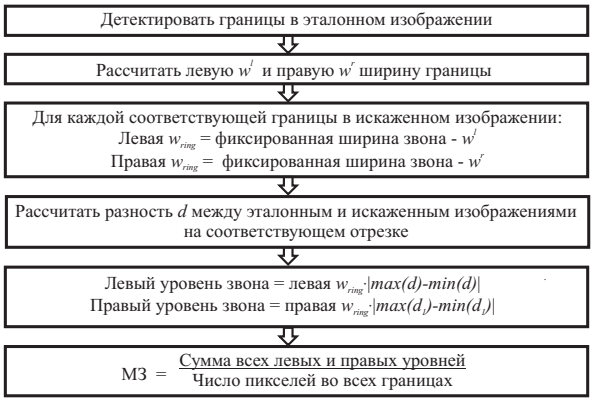


Рис. 2. Схема эталонной метрики звона.

$w_{\mathrm{ring}}^{l(r)} = w_f - w^{l(r)}$, где w_f — фиксированная ширина звона, определяемая опытным путем, $w^{l(r)}$ — левая или правая ширина границы. Локальный уровень левого звона l_{ring} для текущего p определяется по формуле $\left|\max(I_1 - I_2) - \min(I_1 - I_2)\right| \cdot \left|w_{\mathrm{ring}}^{l}\right|$, где I_1 и I_2 — значения яркости оригинального и искаженного изображений на отрезке $\left[p - w_f, p - w^l\right]$. Локальный уровень правого звона r_{ring} определяется аналогично, по отрезку $\left[p + w^r, p + w_f\right]$. Затем усредняем все локальные уровни звона (левые и правые уровни суммируются) по числу пикселей всех границ в изображении и получаем окончательную метрику звона (МЗ) для данного изображения: $\mathrm{МЗ} = \left(\sum l_{\mathrm{ring}} + r_{\mathrm{ring}}\right) \,/\, \left(\sum n\right)$, где n — число пикселей в границе.

Неэталонный индекс качества

Данный алгоритм является неэталонным алгоритмом оценки качества изображения. В нем используется статистическая модель изображения (СМИ) [6, 7] в роли эталона, с которым можно сопоставить оцениваемое изображение. Это относительно новый подход к неэталонной оценке изображений, сжатых при помощи стандарта JPEG2000 (или любым другим алгоритмом сжатия, основанном на вейвлет-разложении).

Статистическая модель изображения отображает статистические взаимозависимости вейвлет-коэффициентов изображений в каждом поддиапазоне вейвлет-разложения и их корреляцию с другими вейвлет-коэффициентами аналогичных поддиапазонов в последующих уровнях разложения. Мы использовали моделирование величины вейвлет-коэффициента C, определяемого величиной линейного предсказателя коэффициента P.

$$C = MP + N, \quad P = \sum_{i=1}^{n} l_i C_i,$$

где M и N — независимые случайные переменные с нулевым средним, C_i — n соседних с C коэффициентов в пространстве, направлении и иерархии, и l_i — коэффициенты линейного предсказателя [7].

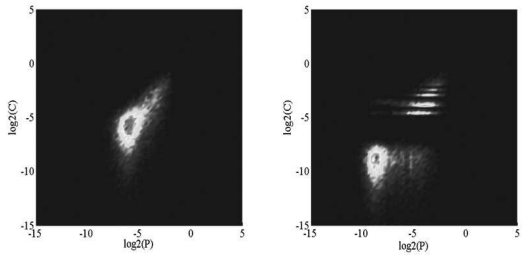


Рис. 3. Пример изменения совместных гистограмм для одного поддиапазона изображения «Лена» при сжатии алгоритмом JPEG2000.

Линейное предсказание производится по соседним коэффициентам по направлению и иерархии: 4 соседних коэффициента из текущего уровня разложения, 1 родительский коэффициент из предыдущего уровня и 1 прародительский коэффициент и пред-предыдущего уровня вейвлет-разложения.

На рис. 3 показаны совместные гистограммы одного и того же поддиапазона для исходного изображения «Лена» и его сжатой копии. Влияние эффекта квантования очевидно: происходит сдвиг коэффициентов к началу координат и нарушается зависимость между величинами C и P, демонстрируя отклонение от естественной СМИ. Мы предлагаем использовать упрощенную модель двух состояний изображения в вейвлет-области. Эти два состояния соответствуют тому, значителен или незначителен коэффициент или его предсказатель. Коэффициент или его предсказатель считаются значительными, если их значения превышают порог. Два порога (один для коэффициента C и один для предсказателя P) выбираются для каждого поддиапазона и зависят от изображения. Совместная модель двух состояний обусловлена тем, что в результате процесса квантования в JPEG2000, который проявляется во всех поддиапазонах, большое количество значений P и C являются менее значительными, чем ожидалось для естественных (без сжатия) изображений. Следовательно, хорошим показателем отклонения изображения от оригинала и наличия визуальных эффектов квантования является малая пропорция значительных P и C [7]. В результате мы получаем набор из четырех эмпирических вероятностей p_{ii}, p_{is}, p_{si}, p_{ss}, соответствующих вероятностям того, что пара предсказатель/коэффициент лежит в одной из четырех областей: незначительный P и незначительный C, незначительный P и значительный C, значительный P и незначительный C, значительный P и значительный C соответственно.

В ходе проведения исследований было замечено, что вероятности p_{ss} являются хорошими показателями потери визуального качества сжатых изображений. Были вычислены соответствующие вероятности для шести поддиапазонов: горизонтального, вертикального и диагонального для первого и второго уровней вейвлет-разложения. Сочетание всех этих параметров должно быть нелинейным, т. к. статистические зависимости отличаются для различных направлений и уровней.

Для преобразования вероятности используется следующая формула [6]:

$$q_i = K_i\left(1 - \exp\left(-\frac{p_{ss,i} - u_i}{T_i}\right)\right),$$

где q_i — преобразованная вероятность (предсказанное качество изображения) для i-го поддиапазона; $p_{ss,i}$ — вероятность p_{ss} для i-го поддиапазона; K_i, T_i и u_i — параметры, аппроксимируемые кривой и получаемые из обучающего набора данных. Параметры подбирались таким образом, чтобы корреляция между результатами алгоритма (объективными оценками) и точками кривой — результатом нелинейной регрессии субъективных и объективных оценок для изображений из обучающего набора, была наибольшей. Взвешенное среднее от преобразованных параметров используется для предсказания качества изображений. По причине схожести статистических зависимостей по горизонтальным и вертикальным поддиапазонам для данной иерархии предполагается, что их веса одинаковы. Таким образом, вектор качества поддиапазона размерности шесть $q = \{q_i \,|\, i = 1, \ldots, 6\}$ преобразуется в вектор размерности четыре q' путем усреднения предсказаний качества горизонтального и вертикального поддиапазона для данного уровня. Окончательно неэталонным индексом качества изображений формата JPEG2000 (НИК2000), лежащим в интервале $[0, 100]$, будет взвешенное среднее $\mathbf{q}'$ [8]:

$$\begin{pmatrix} q'_1 \\ q'_2 \\ q'_3 \\ q'_4 \end{pmatrix} = \begin{pmatrix} (q_1 + q_2)/2 \\ q_3 \\ (q_4 + q_5)/2 \\ q_6 \end{pmatrix}, \quad \text{НИК2000} = \mathbf{q}'^{\tau}\mathbf{w},$$

где веса $\mathbf{w}$ могут быть получены путем минимизации ошибки предсказания качества, используя данные обучающего массива.

Способ, предложенный выше, зависит не только от степени квантования, но и от изменений в содержании изображения. Например, если значения вероятностей сделать постоянными, тогда малое значение p_{ss} может определять как сильно квантованное изображение, так и относительно однородное изображение с невысоким уровнем квантования. Чтобы сделать данную характеристику менее зависимой от содержания изображения, предлагается определять пороги для каждого типа изображений, так чтобы они были ниже для однородных изображений и выше для высоко детализированных изображений.

Рис. 4. Результаты тестирования алгоритмов на изображении «Скарлетт» с различными коэффициентами сжатия. Слева: К = 17 (ПОСШ = 38,25 дБ, УИК = = 0,66, МЗ = 0,49, МР = 7,57, НИК2000 = 59,84). Справа: К = 100 (ПОСШ = 32,43 дБ, УИК = 0,43, МЗ = 0,86, МР = 12,76, НИК2000 = 40,03).

Результаты тестирования алгоритмов

Для тестирования использовались 10 полутоновых изображения с разрешением 512×512 пикселей с разной степенью детализации, сжатые алгоритмом JPEG2000 с 6 коэффициентами сжатия K. Пример изображения из тестового набора и соответствующие оценки качества и значения искажений приведены на рис. 4.

Данные изображения были предварительно оценены экспертами в ходе проведения визуального эксперимента. Задачей эксперимента было определить фиксированную ширину звона и степень согласованности предложенных метрик с субъективной визуальной оценкой DMOS (difference mean opinion score), которая вычислялась как разность между средней оценкой оригинала и средней оценкой текущего изображения (MOS — mean opinion score). В ходе эксперимента также определялись параметры алгоритма НИК2000. Для определения степени согласованности оценок с DMOS использовались следующие критерии корреляции:

— коэффициент линейной корреляции Пирсона;
— коэффициент ранговой корреляции Спирмена;
— корень из среднеквадратичной ошибки.

Результаты приведены в таблице 1.

Анализ данных показывает, что предложенные алгоритмы показывают хорошую согласованность с визуальной оценкой DMOS. НИК2000, являясь неэталонным критерием, показывает чуть меньшую корреляцию с субъективными оценками, чем известные эталонные критерии. Меньшая корреляция метрик размытия и звона с DMOS по сравнению с ПОСШ и УИК объясняется их направленностью на измерение одного конкретного типа искажения (звона или размытия), в то время как задачей экспертной оценки является комплексная оценка качества изображения. В конечной реализации в задачах определения качества изображения или пост-обработки необходимо учитывать комбинацию предложенных метрик.

Таблица 1. Коэффициенты корреляции между значениями DMOS и объективными критериями.

Критерии оценки качества	Коэффициенты корреляции		
	Пирсона	Спирмена	$\sqrt{CKO}$
ПОСШ	0,7821	0,7865	11,4558
УИК	0,7896	0,8170	11,7365
НИК2000	0,6541	0,6196	14,4668
Эталонная МР	0,7026	0,7035	13,6098
Эталонная МЗ	0,6713	0,6990	14,1750

Выводы

Предложенные алгоритмы можно использовать для измерения степени вносимых искажений в процессе сжатия и комплексной неэталонной оценки качества изображений. Алгоритм НИК2000 не требует эталона изображения, что существенно упрощает процесс оценки качества. Метрики размытия и звона разделяют типы искажений, что позволяет более точно настроить параметры кодера. Для цветного изображения, измерение звона, размытия границ и индекса качества производится для яркостной компоненты. Кроме того, низкая вычислительная сложность метрик размытия и звона позволяет адаптировать их для оценки искажений в видеопоследовательностях, сжатых по стандарту Motion JPEG2000.

Литература

[1] *Wang Z., Bovik A. A* Universal Image Quality Index // IEEE Signal Processing letters. — 2002. — V. 9, № 3.— P. 81-84. 33

[2] *Taubman D. S., Marcellin M. W.* JPEG2000: Image Compression Fundamentals, Standards, and Practice // Norwell, MA: Kluwer, 2001.

[3] *Добеши И.* Десять лекций по вейвлетам. — Ижевск: НИЦ «Регулярная и хаотическая динамика», 2001.

[4] *Oguz S. H., Hu Y. H., and Nguyen T. Q.* Image coding ringing artifact reduction using morphological post-filtering // IEEE Second Workshop on Multimedia Signal Processing. — 1998, — P. 628–633.

[5] *Li X.* Blind image quality assessment // IEEE Int. Conf. Image Process. — Rochester, — Sept. 2002.

[6] *Simoncelli E. P.* Statistical models for images: compression, restoration and synthesis // IEEE Asilomar Conf. Signals, Systems, and Computers. — Nov.1997. — P. 673–678.

[7] *Buccigrossi R. W., Simoncelli E. P.* Image compression via joint statistical characterization in the wavelet domain // IEEE Trans. Image Process. — Dec.1999. — V. 8, № 12.— P. 673–678.

[8] *Бекренев В. А., Саутов Е. Ю.* Неэталонная оценка качества изображений, сжатых алгоритмом JPEG2000 // 15-я межд. науч.-тех. конф. «Проблемы передачи и обработки информации в сетях и системах телекоммуникаций» — Рязань 2008. — Т. 1. — P. 124–126.

Метод определения на видеоряде объектов, изображения которых накладываются друг на друга

Ивановский С. А., Марьяскин Е. Л.
Санкт-Петербург, СПбГЭТУ «ЛЭТИ»

В данной статье поставлена проблема распознавания объектов, расположенных в нескольких подряд идущих кадрах видеопоследовательности один перед другим, и предложен метод, позволяющий в некоторых случаях определить в кадре оба таких объекта. В конце статьи описаны направления развития метода в сторону получения лучшего результата.

В обработке изображений зачастую приходится сталкиваться со сценами, в которых одни объекты частично или полностью перекрывают другие. В таких случаях успешная сегментация изображений с обнаружением всех скрытых объектов становится затруднённой или невозможной. Причём, если для непрозрачных изображений объектов это вполне понятно, то сегментация тех сцен, на которых изображения объектов представляют собой разреженные цветовые пятна, должна, очевидно, быть вполне решаемой задачей.

Сцены такого рода, возникают, например, при старте запускаемых стратосферных или орбитальных аппаратов. Съёмка стартов, особенно произведённых в облачную погоду, представляется в виде кадров, на которых видно множество светлых пятен, и даже человеческий мозг не всегда может определить, где за облаками находится аппарат в момент, например, отделения от него ступеней.

Данная статья описывает подход к решению подобных задач, основанный на динамической сегментации изображений сцен, то есть, на методах, использующих сам факт движения наблюдаемых объектов.

Оптический поток

Большинство современных задачи обработки изображений сводится к определению движущихся объектов, определению границ таких объектов, определению факта и параметров движения, и т. д. [3] Методы решения таких задач основаны на сегментации изображений. Термин *сегментация изображения* означает разбиение изображения на множество покрывающих его областей.

Цель многих задач анализа изображений заключается в сегментации на области, с которыми связана существенная для этой задачи информация. В качестве областей интереса могут выбираться также группы пикселей с границей определённой формы [1].

Сегментация изображений имеет две основные цели. Первая цель заключается в декомпозиции изображения на части, более удобные для дальнейшего анализа. В простых случаях удаётся организовывать сцены так, чтобы в процессе сегментации требовалось надёжно выделять небольшое количество областей, неоходимых для дальнейшей обработки. Вторая цель сегментации заключается в изменении формы описания изображения. В результате проведённой сегментации пикселы изображения преобразуются в высокоуровневые структуры, содержащие больше информации или обеспечивающие эффективную организацию дальнейших операций анализа данного изображения [2].

Перспектива разработки одного единственного метода сегментации, одинаково хорошо подходящего для всех задач, выглядит малоосуществимой. Опыт показывает, что разработчику приложений машинного зрения обычно приходится выбирать один из нескольких известных методов и, чаще всего, модифицировать метод с учётом специфических особенностей конкретной задачи.

Многие динамические алгоритмы сегментации опираются на определение оптического потока между близкими кадрами. Опеределение движения объектов в кадре начинается с определения движения каждой из точек кадра. По своей сути, поле потока между кадрами представляет собой оценку поля скоростей.

Оптический поток определяется как «поток» уровней яркости на плоскости изображений.

Традиционный подход к вычислению оптического потока предполагает накладывание дополнительно множества ограничений на исследуемые сцены, приводящие к упрощению уравнения потока, которое в литературе обычно представляется в следующем виде:

$$\frac{\delta g}{\delta t} + f^{\mathsf{T}} \nabla g = 0, \qquad (1)$$

где f определяет оптический поток, а g — значения яркостей точек.

Для вычисления оптического потока на основе уравнения (1) разработано и реализовано несколько алгоритмов. Одним из самых известных является Алгоритм Лукаса-Канаде [4].

Для решения такого уравнения приходится накладывать некоторые дополнительные ограничения. Например, пусть смещение будет одинаковым для всех точек некоторой небольшой окрестности. Пусть размер окрестности — 5×5. Тогда уравнение потока можно записать в матричном виде

для 25 точек $p_1, \ldots, p_{25}$ этой окрестности следующим образом:

$$Ad = b, \tag{2}$$

где A — матрица изменений яркости по координатам размера 25×2, d — вектор потока, b — вектор изменения яркости по времени. Таким образом, решением задачи оказывается вектор d, вычисленный как

$$d = (A^\intercal A)^{-1} A^\intercal b. \tag{3}$$

Но движение и изменение уровней яркости не эквивалентны. Два классических примера, где спроектированное поле движений и оптический поток не равны, были приведены Горном (Horn) [5]. Первым примером является вращающаяся сфера с однородной поверхностью любого вида. Такая сфера может вращаться вокруг любой оси, проходящей через её центр тяжести, не вызывая поля оптического потока. Противоположным примером является та же самая сфера в состоянии покоя, освещаемая движущимся источником света. Теперь поле движений равно нулю, но изменения в уровнях яркости, обусловленные движением источника света, вызывают ненулевое поле оптического потока.

Вычисление потока

Все предлагаемые методы вычисления поля потока предлагают на самом деле вычисление векторов изменения яркости. Поэтому на результате вычисления будут отчётливо видны только границы двигающихся объектов, а их середина, в свою очередь, будет определена как неподвижная, поскольку изменения яркости там не происходит, или значительно меньше изменений яркости на границе.

Традиционный подход к обработке результатов таких вычислений сводится к пороговому отделению точек по признаку длины вектора потока. Добавление учёта отдельно длины и отдельно направления векторов может привести к интересным результатам.

Рассмотрим случай перекрывающих друг друга изображений объектов. Каждый из них корректно описывается векторами потока на границах объектов. Поскольку внутренняя область изображения ближнего объекта определится на основе метода вычисления оптического потока по порогу как неподвижная, есть возможность увидеть на изображении потока векторы, порождённые границей дальнего объекта. Пример приведён на рисунке 1.

Рисунок сделан с помощью специального инструментального средства, в первую очередь предназначенного для покадрового просмотра информации о вычисленном в кадре оптическом потоке. На изображении отчётливо видны границы эллипсов — изображений объектов, в то время как соответствующий кадр исходного видеоряда показывает, естественно, одно белое пятно, ограниченное большим эллипсом.

Рис. 1. Пример накладывающихся изображений двух объектов эллиптической формы (слева) и оптического потока от них (справа).

Поскольку на изображении потока (правой половине рисунка) отчетливо видна эллиптическая структура двух объектов, изображения которых наложились друг на друга, данный пример показывает, что при анализе данных оптического потока может быть найден путь к решению заявленной задачи сегментации изображений от накладывающихся объектов.

Предлагаемый метод

Метод, с помощью которого предлагается решать задачу сегментации накладывающихся объектов, был разработан в ходе исследования существующих методов обработки оптического потока и является, таким образом, частью обобщённого решения проблемы сегментации изображений динамических сцен. Исследования показывают, что данный набор действий, фильтров и характеристик оказывается достаточным, с точностью до настроек алгоритмов, для успешной сегментации изображений рассматриваемого типа. Обобщённый до ближайших предстоящих модификаций, алгоритм представляется следующим образом.

1. Выделение кластеров, соответствующих частям изображений объектов и избавление от основных шумов.
2. Определение характерных свойств кластеров.
3. Чтение информации о предыдущих обработанных кадрах с определением ожиданий о форме объектов, их местоположении и характеристиках.
4. Формирование представления в указанном кадре о реальных объектах, траекториях их движений и характеристиках объектов.
5. Запись нового состояния.

Первичное выделение кластеров. Для выполнения этапа первичного выделения кластеров используется итерационная схема кластеризации и фильтрации изображения до тех пор, пока не будут достигнуты заданные пороговые значения характеристик. Целью этапа является получение кластеров, отвечающих нескольким свойствам:

1. Каждый кластер полностью является частью какого-либо объекта.

2. Ни один кластер не соответствует шуму.
3. Ни один кластер не является частью двух и более объектов.

Основные особенности этого этапа:

— кластеризация;
— фильтрация;
— определение характеристик качества результата;
— итеративная схема.

Кластеризация служит для разграничения групп точек картинной плоскости, движущихся совместно и, по сути, являющихся частями изображений одного и того же объекта. Фильтрация применяется для удаления случайных шумов и уточнения границ кластеров. Итерация позволяет автоматизировать всю последовательность обработки изображения вплоть до достижения определенных характеристик качества результата.

Основная идея заключается в повторяющейся взвешенной кластеризации векторного поля очередного кадра. Причём кластеризация обладает следующими отличительными свойствами:

1. Происходит по четырёхмерным векторам, которые составлены из двух геометрических координат векторов потока и двух координат самого вектора оптического потока в данной точке.
2. Кластеризация является взвешенной, что позволяет в зависимости от задач и условий настраивать алгоритм.
3. Изображение потока делится на кластеры, каждый из которых является частью не более чем одного исходного изображения объекта.

Таким образом, результат подобной кластеризации представляет собой векторное поле, разделённое на кластеры близких (в евклидовом смысле) векторов, причём каждый кластер представляет собой либо совокупность векторов фона, либо часть изображения реального объекта исходной сцены. Соответственно, при наложении изображений объектов, кластеры «внутреннего» объекта будут отделены от кластеров «внешнего» за счёт учёта направления векторов потока, которое у этих кластеров различно.

Определение характеристик. Прочие части метода представляют собой фильтрацию не относящихся к целевым объектам данных и определению характеристик, позволяющих определить наилучший момент прекращения итеративного процесса.

После этапа обработки на основании конфигурации выделенных кластеров зачастую представляется возможным сделать выводы об особенностях конкретных исследуемых объектов. Два самых ярких примера — это вращение и изменение размеров.

— *Вращающийся объект* будет видно по делению его на кластеры, делящие его на две доли, которые вращаются в противоположные стороны.
— *Расширяющийся/сжимающийся объект.* В этом случае кластеры представляют собой кольцо, замкнутое вокруг центра объекта, как это видно на примере выше для центрального объекта.

Определение информации предыдущих сессий. На этом этапе происходит чтение информации, записанной для предыдущих кадров и позволяющей по ожиданиям для размеров и местоположений объектов определить, какой кластер в какой объект входит. Такой метод активно использует динамическую информацию об изменениях потоков и позволяет решать многие интересные задачи, как, например, определение наложения одного объекта на другой, что невозможно обычными методами.

Этап постобработки. На данном этапе происходит окончательное формирование объектов из кластеров, с постфильтрацией и подчёркиванием формы. Алгоритм представляет собой итерационное выполнение различных, чаще всего индивидуальных для каждого кластера фильтров.

В целом фильтры основаны на существующих методах статической обработки изображений, а значения характеристик отражают основные свойства кластеров: форму, размер, расположение, заполненность точками, распределение внутри кластеров и пр.

Процесс обработки представляет собой итеративно повторяющуюся кластеризацию данных, с сопутствующим применением на каждом шаге набора фильтров, позволяющих избавиться от шума и уточнить реальные границы кластеров. Итерация происходит до тех пор, пока не будут достигнуты требуемые значения характеристик.

Этап записи результата. Результаты работы (выходные данные программного приложения) сформированы в результате исследования целей, возлагаемых на приложение анализа потоковых данных. На выходные данные накладывается несколько требований:

1. Данные должны содержать всю требуемую информацию об объекте:
   — о его координатах и форме;
   — о его разреженности и внутренних характеристиках;
   — о его траектории.
2. Данные должны быть легко используемы в дальнейшем.
3. Формат данных должен поддерживать их изменение после обработки очередного кадра.

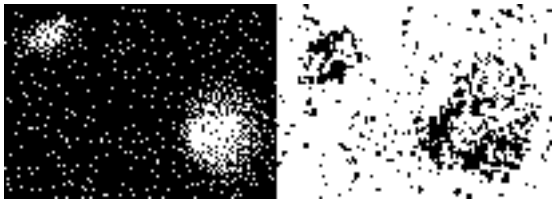


Рис. 2. Изображение исходной сцены (слева) и потоковые векторы сцены с тремя объектами, два из которых наложены друг на друга (справа).

Рис. 3. Результат обработки потоковых данных, указанных на рисунке 2.

Моделирование сцен

В ходе исследования и для подтверждения корректности разрабатываемых методов, применялись сцены, имитирующие в картинной плоскости взаимное движение объектов. Для генерации модельных данных использовалось специальное программное средство генерации видеорядов, позволяющее настраивать для модельных объектов:

— параметры поступательного и вращательного движения;
— параметры изменения размеров изображенных объектов относительно их центров масс;
— параметры яркости.

И для модельного видеоряда в целом:

— число и размеры кадров;
— параметры фона.

Примеры применения

Исследование проводились с помощью отдельного инструментального программного средства, специально предназначенного для обработки информации об оптических потоках. Для этого средства описываемый метод является частью его функциональности.

Рассмотрим пример отработки алгоритма для одного из кадров потока. На рис. 2 показаны точки векторов потока, отфильтрованные пороговым способом. Сюжет сцены представляет собой взаимное перемещение трёх объектов эллипсоидальной формы на сильно зашумлённом фоне, причём рассматриваемый кадр соответствует моменту полного перекрытия одного объекта другим. После обработки описанным выше методом набор векторов потока был на последнем шаге кластеризован на три кластера, показанные на рис. 3 (показаны только выпуклые оболочки кластеров). Кластер 1 соответствует одному объекту, кластеры 2 и 3 вместе — второму, кластер 3 — третьему, спрятанному на кадре за вторым.

Поскольку кластеры формируются на основе вектроного поля потока, являющегося, по сути, оценкой поля скоростей, и, кроме того, изначально задано количество кластеров, значительно превосходящее количество объектов, всегда получается получить набор кластеров, где каждый состоит из совместно движущихся точек. Корректность определения каждой отдельной точки как движущейся проверена статистически и в общем случае (включая ситуации изменения яркости, размеров объектов и пересечения их изображений), в среднем близка к 80%.

Таким образом, путём применения к задаче предложенного метода удаётся получить разделение кадра на изображения объектов исходной сцены, причём уже с учётом возможных пересечений и наложений изображений объектов друг на друга.

Выводы

В результате исследования проблемы сегментации сцен с объектами, заслоняющими друг друга, была установлена принципиальная возможность обнаруживать заслоненные объекты, определять их форму, конфигурацию и траекторию движения.

Указан метод, позволяющий реализовывать подобную сегментацию. Метод реализован в программном приложении, предназначенном для анализа оптических потоков.

Обозначены ближайшие перспективы развития указанного подхода, которые будут реализованы в ближайшее время и помогут решить рассмотренную задачу более полно.

Литература

[1] *Яне Б.* Цифровая обработка изображений — М.: Техносфера, 2007. — 583 С.

[2] *Шапиро Л., Стокман Дж.* BibTitleКомпьютерное зрение — М.: БИНОМ. Лаборатория знаний, 2006. — 752 С.

[3] *Ballard D., Brown C.* Computer vision — Rochester, New York: Department of computer science. University of Rochester, 2006. — 573 P.

[4] *Backer S., Matthews I.* Lucas-Kanade 20 Years On: A Unifying Framework // International Journal of Computer Vision. — 2001. — Pp. 1–30.

[5] *Horn B.K.* Robot vision // MIT Press, Cambridge, MA. — 1986. — Pp. 1–36.

Об алгоритмах сегментации для системы автоматической нотной транскрипции музыкального фольклора

Кальян В. П.
vkalyan@mail.ru
Москва, Вычислительный центр РАН

Приведены содержание результаты эксперимента по настройке правил сегментации на конкретную задачу нотной транскрипции для систем автоматической расшифровки фонограмм народных исполнителей.

Введение

Одной из задач музыкальной информатики является автоматическая нотная транскрипция фольклорных фонограмм. Эта задача до некоторой степени аналогична задаче распознавания речи.

С помощью дискретного преобразования Фурье получают картину спектральной динамики звука, траектории нескольких первых формант, амплитудной огибающей напева. По алгоритмам Терхарда или Гольдштейна строят траектории основного тона. Затем с помощью алгоритмов сегментации и маркировки дают представление мелодии с помощью последовательности нот. На конечной стадии используют автоматический нотатор — нотографический редактор [1].

Первые работы в области автоматической нотной транскрипции появились еще в 70-е годы прошлого столетия.

Известны исследования Мартина Пищальского и Бернарда Галлера из Мичиганского университета [2], Джеймса Мурера, Криса Чейфа и Бернарда Мон-Рено из Станфордского университета [3, 4, 5]. Их работы носили характер исследовательских опытов и преследовали цель отработки простейших алгоритмов нотной транскрипции на примере какого-либо известного напева, сыгранного на одноголосном инструменте, или даже спетого самим исследователем.

Сколько-нибудь масштабных прикладных проектов по автоматизированной нотной расшифровке массивов фонограмм вплоть до настоящего времени опубликовано не было.

Однако логично было бы ожидать, чтобы подобные проекты могли появиться в области музыкальной этнографии.

Так, например, в отечественной музыкальной фольклористике часто ставится задача нотной расшифровки фонограмм народных исполнителей.

В то же время, попыток автоматизированной нотной транскрипции отечественные исследователи не предпринимали. Вероятно здесь дело в том, что разные музыкальные этнографы одну и ту же фольклорную фонограмму могут расшифровать по-разному.

Навыки расшифровки фонограмм фольклорных исполнителей прививаются профессиональным музыкантам еще в музыкальных средних и высших учебных заведениях, и правила нотации у разных исследователей заметно отличаются друг от друга, хотя тенденция к сближению норм фольклористической нотации заметна [6].

Кроме того, нотная интерпретация фольклорного материала может зависеть не только от концепции музыкального фольклориста, но и от задачи расшифровщика, например — дать общую форму движения голоса, или передать характерные черты микроинтонирования исполнителя в данном напеве [7].

С учетом этих особенностей система автоматизированной нотной транскрипции должна быть адаптивной, подстраиваться под фольклористическую концепцию исследователя и конкретную задачу расшифровки.

Настоящая работа посвящена построению алгоритмов сегментации для системы автоматической нотной транскрипции фольклорных фонограмм и исследованию возможностей настройки этих алгоритмов на разные фольклористические задачи.

Описание эксперимента

Для экспериментов был взят массив фольклорных фонограмм сольных исполнителей из Белгородской области. Высота звука P и амплитудная огибающая A фрагмента одного из напевов, как функции времени t, приведены на рис. 1.

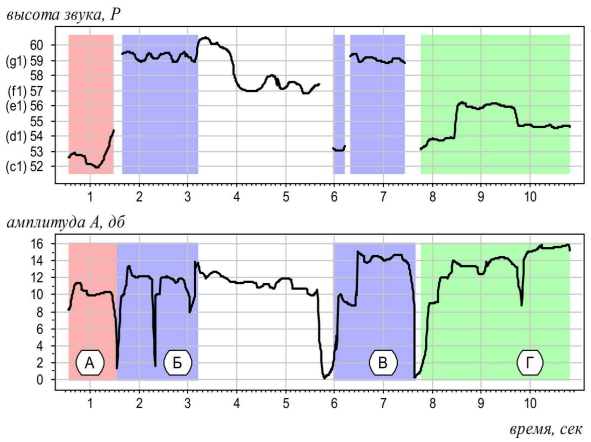


Рис. 1. Траектории высоты и амплитуды звука в напеве как функции времени.

Траектория высоты звука, изображенная на рис. 1, получена с помощью преобразования траектории основного тона из линейной шкалы частот W в логарифмическую шкалу с основанием 2, что соответствует высоте звука P в музыкальном восприятии.

Значение 52 по оси P соответствует высоте ноты до первой октавы. Значение шага между ближайшими целыми значениями (52, 53, 54 и т.д.) соответствует полутону темперированного строя диатоники.

Временные отсчеты функций $A(t_i\ )$ и $P(t_k\ )$ получены с одним и тем же шагом $\delta t = 0.01$ сек так, что для каждого $k = i = 1, \ldots, n$ отсчеты A и P синхронны.

На первом этапе применялся алгоритм поиска локальных максимумов и минимумов в последовательности временных отсчетов функций $P_{\max}(t_k\ )$, $P_{\min}(t_k\ )$, $A_{\max}(t_i\ )$, $A_{\min}(t_i\ )$ для таких отсчетов t_k и t_i, в которых выполнялись простейшие условия «перегиба»:

$$P_{\max}(t_k\ ) = P(t_k), \text{ если } P(t_{k-1}) < P(t_k) > P(t_{k+1});$$
$$P_{\min}(t_k\ ) = P(t_k), \text{ если } P(t_{k-1}) > P(t_k) < P(t_{k+1}).$$

Аналогично,

$$A_{\max}(t_i) = A(t_i), \text{ если } A(t_{i-1}) < A(t_i) > A(t_{i+1});$$
$$A_{\min}(t_i) = A(t_i), \text{ если } A(t_{i-1}) > A(t_i) < A(t_{i+1}).$$

Из найденных $P_{\max}$, $P_{\min}$, $A_{\max}$, $A_{\min}$ были сформированы четыре массива

$$\texttt{MAXP}(P_{\max}, t_{P_{\max}}), \quad \texttt{MINP}(P_{\min}, t_{P_{\min}}),$$
$$\texttt{MAXA}(A_{\max}, t_{A_{\max}}), \quad \texttt{MINA}(A_{\min}, t_{A_{\min}})$$

значений локальных экстремумов и соответствующих им временных значений

$$t_{P_{\max}}(k),\ t_{P_{\min}}(k),\ t_{A_{\max}}(i),\ t_{A_{\min}}(i).$$

Фрагмент напева между первым и вторым максимумом функции $P(t)$ соответствует временному интервалу А на рис. 1. Экспертная оценка этого фрагмента показала, что он соответствует попевке, которая может быть отображена в нотации двумя нотами. Сегментация на основании ближайших пар $A_{\max}$ и $A_{\min}$ выявляет эту последовательность нот, но применение этого алгоритма ко всему напеву дает большой процент ошибок.

Особенно это отчетливо видно на участке голосового вибрато, соответствующего временному интервалу Б на рис. 1. Алгоритм, опирающийся на сегментацию с помощью ближайших пар $A_{\max}$, $A_{\min}$ и $A_{\min}$, $A_{\max}$, разбивает этот фрагмент на 8 сегментов. Экспертная же оценка этого фрагмента выявляет наличие лишь одного сегмента-тона.

Необходимы дополнительные правила, позволяющие отличать плавный переход звучания от одной высоты тона к другой без заметного изменения его интенсивности, и алгоритмы автоматической настройки пороговых значений $P_{\max}$, $P_{\min}$, $A_{\max}$, $A_{\min}$, позволяющих отличать голосовое вибрато от мелизматики и других форм микроинтонирования.

Одной из особенностей исследуемого музыкального материала являлась стабильность звуковысотной организации напева, что дало возможность использовать функцию распределения плотности вероятности N высоты звука $P(t_k\ )$ для построения алгоритмов сегментации и маркировки.

По функции распределения плотности вероятности были построены гистограммы с шагом 0.1 тона, для данного напева; она отображена на рис. 2.

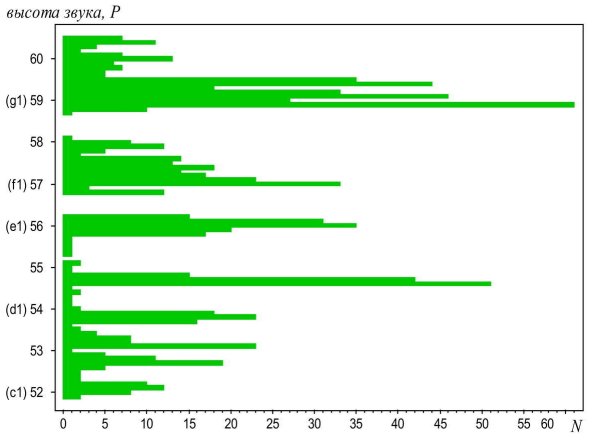


Рис. 2. Гистограмма высоты звука в напеве.

На рис. 2 довольно четко видно разделение звуковысотной шкалы напева на семь зон. Зоны сосредоточены вокруг значений 52, 52.7, 53.6, 54.6, 56.2, 57.7, 59.2.

На основании минимумов функции N вычислялись границы звуковысотных зон, соответствующих ступеням звукоряда в напеве исполнителей.

Из сравнения временных интервалов Б (голосовое вибрато) и В (восходящий скачок с последующим вибрато) на рис. 1 и гистограммы рис. 2. видно, что голосовые вибрато целиком укладываются в 7-ю зону полученного разбиения функции N.

Дальнейшая настройка алгоритма — применение полученных звуковысотных зон и нахождение пороговых значений для разности ближайших пар $A_{\max}(i)$–$A_{\min}(i-1)$ в 1.57 дб в качестве решающего правила позволила выполнить сегментацию всего напева корректно с точки зрения эксперта.

Для ритмической интерпретации был выбран простейший алгоритм вычисления пропорций — метроном 40. В этом случае четверть составляет $\sim$1.3 сек, восьмая $\sim$0.65 сек, шестнадцатая

$\sim$0.33 сек. На нарушение пропорций был применен порог отсечения в 20%.

Нотная расшифровка изображенного на рис. 1 фрагмента напева с помощью описанной группы алгоритмов представлена на рис. 7.

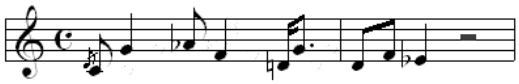


Рис. 3. Нотная расшифровка напева.

Данный метод был применен ко всей выборке напевов и показал надёжность 82%.

Заключение

В докладе описаны эксперименты по настройке правил сегментации на конкретную задачу транскрипции. Изменение параметров пороговых значений для правил сегментации, а также функции распределения вероятности высоты звука позволяет настраивать работу алгоритма на задачу музыкального эксперта. Однако, в данной работе решалась частная задача для музыкального материала со стабильно-высотным интонированием. Для решения более общей задачи с «плавающим звукорядом», или «раскрывающимся ладом» нужны дополнительные исследования музыкального фольклора широкого ареала бытования разных песенных традиций.

Литература

[1] *Кальян В. П.* Музыка, речь и компьютер. М.: ВЦ РАН, 1998. — 40 с.

[2] *Piszczalski M., Galler B.* Automatic Music Transcription // Computer Music Journal. — 1977. — Vol. 1 No. 1, C. 24–31.

[3] *Moorer J.* On The Transcription of Musical Sound by computer // Computer Music Journal. — 1977. — Vol. 1 No. 4, — Pp. 32–38.

[4] *Grey M., Moorer J.* Perceptual Evaluatiion of Synthesized Musical Instrument Tones // Journal Of The Ac. Sosiety Of America. — 1977. — Vol.64 — Pp.434–462.

[5] *Chafe C. Jaffe D.* Source separation and note identification in polyphonic music // Proceedings of the International Conference on Acoustics, Speech, and Signal Processing. — New York. IEEE Press, — 1986.

[6] *Алексеев Э. Е.* Нотная запись народной музыки: теория и практика. М.: Советский композитор, 1990.

[7] *Кальян В. П.* Интонационный анализ народных исполнителей // Методы и модели распознавания речи, Москва: ВЦ РАН, 2007. — С. 48–56.

Геометризованные гистограммы и понимание изображений

Кий К. И.
kikip_46@mail.ru
Москва, ИПМ РАН им. Келдыша

В работе описывается применение метода геометризованных гистограмм к описанию цветных изображений, выделению предметов на цветных изображениях. Разрабатывается техника для распознавания цветных изображений, в частности, для поиска на изображениях предметов с заданной формой и цветовыми характеристиками.

Геометризованная гистограмма цветного (полутонового) изображения была явно введена в [1]. Неявно предшественники этой конструкции уже использовались в предыдущих работах автора [2, 3, 4]. Однако рассматривалась либо геометризованная гистограмма полутоновых изображений, в частности, задаваемых одной из цветовых компонент исходного изображения [3, 4], либо предварительный вариант гистограммы цветного изображения без процедуры принудительного деления цветовых областей [2]. В данной работе мы рассматриваем новый вариант определения с введением «наивных» белого, черного и нескольких оттенков серого цветов, столь успешно используемых в зрении человека, и описываем модернизированные алгоритмы кластеризации. Мы также определяем структурный граф, который связывается с каждым цветным изображением как результат анализа его геометризованной гистограммы, и обсуждаем применения этого графа в прикладных задачах. Несмотря на большое число методов распознавания изображений (главным образом основанных на анализе выделенных контуров) [5], предложенный подход имеет свои преимущества. Он позволяет одновременно работать как с объектами большого, среднего, так и малого размера, в условиях загроможденных сцен с заслонениями, при возможности появления новых объектов в кадре, и обеспечивает режим анализа сцены в реальном времени.

Структурный граф цветовых сгустков цветных изображений

В работе автора [1] для каждого цветного изображения вводится новый геометрический объект — геометризованная гистограмма, которая объединяет статистическое описание изображения с помощью набора значений его характеристик с геометрическим описанием их областей уровня. Геометризованная гистограмма соответствует некоторому разбиению изображения на узкие полосы одинаковой ширины, параллельные горизонтальной (вертикальной) оси Ax прямоугольной системы координат, связанной с изображением. Геометризованная гистограмма определяется тройкой GH, π, B, где GH — расслоенное пространство со слоями — подпространствами пространства конечных интервалов на прямой, $\pi\colon GH \to B$, — отображение проектирования и B — дискретное упорядоченное множество, параметризующее полосы разбиения. Пространство GH есть расслоенное пространство, каждый слой которого есть локальная геометризованная гистограмма соответствующей полосы изображения GH_s, $s \in B$, то есть GH_s есть некоторое подпространство пространства всех конечных интервалов Ax. Каждому интервалу I сопоставлен следующий набор характеристик [1]: $(g_b, r, H_{\min}, H_{\max}, H_{\mathrm{mean}}, S_{\min}, S_{\max}, S_{\mathrm{mean}}, gr_{\min}, gr_{\max}, gr_{\mathrm{mean}}, \mathrm{card}, \mathrm{beg}, \mathrm{end})$, где g_b — фиксированное значение $G/(G+B)$ $(R, G, B$ — стандартные цветовые координаты), r — цветовой диапазон (например, красный, желтый, зеленый, голубой и т. п.), $(H_{\min}, H_{\max}, H_{\mathrm{mean}})$, $(S_{\min}, S_{\max}, S_{\mathrm{mean}})$, $(gr_{\min}, gr_{\max}, gr_{\mathrm{mean}})$ описывают диапазон и среднее значение цветового оттенка, насыщения и полутоновой черно-белой компоненты изображения соответственно, а card, beg, end задают мощность интервала (некоторое целое число) и координаты начала и конца интервала на оси. Напомним, что каждый интервал $I \in GH_s$ принадлежит проекции на ось Ax множества точек полосы с фиксированными значениями g_b, r. В [1] намечена на GH_s процедура кластеризации, в результате которой GH_s разбивается на кластеры, цветовые сгустки. В данной работе также рассматривается усовершенствованная геометризованная гистограмма. В процедуре построения геометризованной гистограммы [1] в дополнение к процедуре разделения цветов аналогичным методом в области малонасыщенных цветов вводятся дополнительные значения основной характеристической функции. На основе значения полутоновой компоненты малонасыщенные точки цветового пространства группируются в подмножества, соответствующие наивным белым, черным и серым цветам, и точкам этих подмножеств присваиваются дополнительные значения характеристической функции.

Опишем процедуру кластеризации подробнее. Во-первых, на пространстве интервалов вводятся меры близости (псевдо-метрики):

$$\begin{aligned} &d_i(I,J) = 1 - \rho_i(I,J),\ i = 1,2,\\ &\rho_1(I,J) = L(I \cap J)/\min(L(I), L(J)),\\ &\rho_2(I,J) = L(I \cap J)/\max(L(I), L(J)), \end{aligned} \qquad (1)$$

где I, J — интервалы, а $L(I), L(J)$ — длины интервалов. Относительно этих мер близости организуется кластеризация.

На первом шаге выбираются семена для выращивания кластеров. Для этого вводится понятие заметности интервала (заметности в данном цветовом диапазоне). Пусть $\mathrm{Dim}\, Ax$ — размерность массива изображения в направлении оси Ax (она равна $\dim X$ или $\dim Y$ в зависимости от того, вертикальная или горизонтальная ось выбрана). Выберем массив размерности $\mathrm{Dim}\, Ax$. Бросаем на него в произвольном порядке интервалы GH_s. В каждой точке выживает точка интервала с максимальной плотностью $\mathrm{card}(I)/L(I)$.

Определение 1. *Для каждого интервала его заметность $V(I)$ равна отношению числа выживших точек к общему числу точек (длине) интервала.*

Так как мы рассматриваем изображения, задаваемые дискретными массивами, все интервалы состоят из конечного числа точек. Такая же процедура может быть проделана внутри интервалов с фиксированным цветовым диапазоном. Мы также можем определить заметность в заданном цветовом диапазоне. Мы начинаем процедуру кластеризации с наиболее заметных интервалов и присоединяем к ним интервалы, которые близки к ним относительно метрик (1) и относительно цветовых характеристик. Для оставшихся интервалов в качестве семян берутся интервалы, наиболее заметные в разных цветовых диапазонах.

Такая организация процедуры кластеризации позволяет выделять даже небольшие окрашенные предметы при достаточно большой ширине полосы на различных фонах. Например, габаритные огни, тормозные сигналы автомобилей и сигналы светофоров, а также цветные метки, нанесенные по маршруту следования робота. На всем множестве цветовых сгустков, построенном для всех локальных гистограмм (всех полос), вводятся дополнительные структуры. Каждый цветовой сгусток характеризуется следующими числовыми значениями: ($H_{\min}$, $H_{\max}$, H_{mean}, $S_{\min}$, $S_{\max}$, S_{mean}, $gr_{\min}$, $gr_{\max}$, gr_{mean}, card, beg, end), где $(H_{\min}, H_{\max}, H_{\mathrm{mean}})$, $(S_{\min}, S_{\max}, S_{\mathrm{mean}})$, $(gr_{\min}, gr_{\max}, gr_{\mathrm{mean}})$ описывают диапазон и среднее значение оттенка, насыщения и полутоновой черно-белой компоненты соответственно, а card, beg, end задают мощность (некоторое целое число, равное сумме мощностей окрашенных интервалов, объединенных в данный цветовой сгусток) и координаты начала и конца интервала, соответствующего цветовому сгустку. Для каждого цветового сгустка с помощью процедуры, которая выполнялась для интервалов геометризованной гистограммы, определяется его заметность. Кроме того, для корректировки результатов кластеризации, изучая

Рис. 1. Цветовые сгустки уличной сцены.

отношение доминирования между сгустками, мы удаляем сгустки, которые не видны на фоне более массивных сгустков, имеющих сходные цветовые характеристики . Оставшиеся сгустки образуют вершины структурного графа STG(CI) заданного цветного изображения CI.

Пример набора цветовых сгустков, порождающих вершины структурного графа цветного изображения уличной сцены, показан на рис. 1.

Изображение формата 320×240 разбито на 24 горизонтальные полосы. Цветовые сгустки накладываются на средние линии соответствующих полос.

Анализ изображения производится на основе структурного графа изображения. Перейдем к определению множества ребер в STG(CI). Определим ребра, соединяющие соседние цветовые сгустки, принадлежащие одной и той же полосе. Пусть имеются два цветовых сгустка с интервалами I_1, I_2. Если эти интервалы не пересекаются, то обозначим через I_3 интервал, лежащий между ними.

Определение 2. *Цветовые сгустки в полосе называются соседними, если $d_1(I_1, I_2) < 1/2$ (пересекающиеся интервалы) или $d_1(I_1 \cup I_3, I_2 \cup I_3) < \varepsilon$ (непересекающиеся интервалы), где ε — некоторая константа.*

Соседние вершины в пределах одной полосы соединяются ребром, если цветовые характеристики соответствующих вершин близки. Опишем процедуру построение связей между цветовыми сгустками, принадлежащими разным, но соседним полосам. Пусть имеются два цветовых сгустка в соседних полосах, и им соответствуют интервалы I_1, I_2.

Определение 3. *Цветовые сгустки в полосе непрерывно продолжают друг друга как геометрические объекты, если $d_1(I_1, I_2) < 1/2$.*

Два цветовых сгустка в соседних полосах соединяются ребром, если их цветовые характеристи-

ки близки. Для отыскания малых предметов (метки на пути движения робота, сигналы светофоров и т. п.) представляют интерес вершины графа, из которых не выходит и в которые не приходит ни одного ребра. Наоборот, для поиска больших объектов необходимо исследовать связные компоненты на структурном графе STG(CI). Объекты достаточно простой формы задаются путями, содержащими по одному сгустку в каждой полосе. Такие объекты, по аналогии с теорией расслоений, мы будем называть сечениями графа STG(CI). Они соответствуют односвязным объектам (без дырок) на изображении. Каждому пути на графе STG(CI) можно поставить в соответствие некоторый геометрический образ, порожденный интервалами цветовых сгустков, которые его образуют.

В [3] описана система понимания изображений, позволяющая находить положение дороги на изображениях дорожных сцен, основанная на сечениях в структурном графе, построенном на основе геометризованных гистограмм $G/(G+B)$. Эта техника хорошо работала на изображениях дорог, окруженных растительностью (проходящих в лесу, по полю). Однако на сложных сценах, содержащих много предметов разного цвета, геометризованная гистограмма, основанная на одной скалярной функции, работала плохо. Как показали многочисленные эксперименты с разнообразными видеопоследовательностями, геометризованные гистограммы, предложенные в настоящей статье, уже имеют необходимые разделяющие свойства.

Практическая реализация

Для реализации предложенных методов сжатого описания и сегментации изображений разработан комплекс программ, написанных на C++ под Windows XP, Vista. Комплекс сопряжён с вводом из видеокамеры с помощью DirectXShow 9. Время обработки одного изображения линейно зависит от числа пикселов и производительности процессора. Имеется также слабая зависимость от числа полос, на которые разбито изображение. При обработке цветных изображений размера 640×480 на процессоре Pentium 4, 3.0 GHz в зависимости от числа полос (в типичных примерах от 16 до 32) скорость обработки порядка 10 fps. Увеличение числа полос незначительно сказывается на времени работы программ. На многопроцессорных персональных компьютерах следующего поколения (без специального распараллеливания обработки, которая возможна) достигается более высокая скорость (порядка 16 fps). Для изображений размера 320×240 скорость обработки в четыре раза выше. Скорость обработки скалярных изображений, например, инфракрасных, выше более чем в шесть раз. Развитый язык содержательного описания результатов обработки позволяет решать многие задачи распознавания при меняющихся условиях окружающей среды. В настоящий момент разрабатывается комплекс программ для обеспечения движения автономного робота в помещениях, включая распознавание ориентиров, основанный на разработанных программных средствах. При этом допускается работа через сеть, включая Интернет (разработаны соответствующие программные средства).

Рис. 2. Цветовые сгустки сцены в помещении.

Примеры работы метода

Проиллюстрируем результаты обработки на конкретном примере. Рис. 2 показывает полутоновую компоненту цветного изображение с наложенными на него цветовыми сгустками (вершинами графа STG(CI)).

Рис. 3 показывает сечения графа STG(CI), нанесенные на него. Вертикальная ось показывает номер цветового сгустка (bunch number), горизонтальная — номер полосы. Вершины графа обозначены квадратными точками, а пути на графе показывают построенные сечения. Рис. 4. показывает геометрическое описание некоторых сечений, наложенное на исходное изображение. С помощью структурного графа хорошо выделяются такие мелкие объекты, как сигналы светофора, габаритные огни и стоп-сигналы автомобилей, цветные электророзетки, огнетушители, электрические щиты в помещениях. На реальных уличных сценах выделяются также части одежды, обуви, открытые участки тела (руки, лица), автомобили, разнообразные вывески и дорожные знаки, газоны и другая растительность. Все это делает перспективным применение разработанных методов в видеонаблюдениях (в частности, в качественном, не пиксельном, анализе движения, которое более устойчиво к помехам).

Выводы

Разработан новый метод представления сцен, удобный для решения задач понимания изображе-

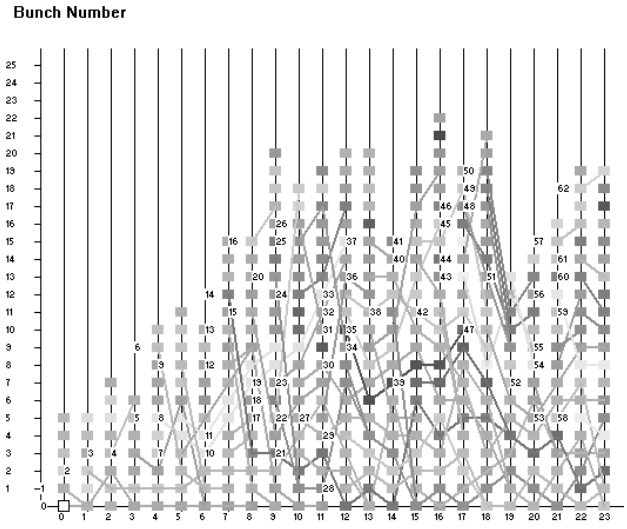


Рис. 3. Сечения, наложенные на граф STG(CI) для сцены рис. 2.

Рис. 4. Некоторые сечения графа STG(CI), наложенные на изображение сцены рис. 2.

ний в реальном времени. Основу представления составляет структурный граф, сопоставляемый любому цветному изображению. Конструкция структурного графа основана на новом понятии геометризованной гистограммы цветного изображения. Многочисленные эксперименты с видеопоследовательностями с использованием комплекса программ, реализующего предложенный метод, показывают, что геометризованная гистограмма и построенный на ее основе структурный граф позволяют давать адекватное описание сложных сцен и разделять различные предметы на изображениях. Построенная техника позволяет выделять реальные объекты в изображениях. В качестве следующих шагов исследования необходимо указать следующие:

1) построение различных систем понимания изображения, основанных на предложенной технике;
2) выделение набора характерных частей изображения на парах и последовательностях кадров для разработки качественного стерео анализа;
3) интеграция описаний, полученных с помощью предложенного метода и стандартных контурных описаний;
4) введение в метод более тонкого анализа полутоновой компоненты;
5) модификация структурного графа с введением неоднозначности и нечеткости, а также интеграция с нейронными сетями при решении задач распознавания.

Литература

[1] *Kiy K. I.* A New Method for Description and Generalized Segmentation of Color Images in Real Time // Int. Conf. on Pattern Recognition and Image Analysis: New Information Technologies, Nizhni Novgorod, 2008. — P. 297–300.

[2] *Kiy K. I., Dickmanns E. D.* A Color Vision System for Real-Time Analysis of Road Scenes // IEEE Intelligent Vehicle'2004 Int. Symp., Parma, Italy, 2004. — P. 54–59.

[3] *Kiy K. I.* An Unsupervised Color Vision System for Driving Unmanned Vehicles // SPIE AeroSense'98 Symp., Enhanced and Synthetic Vision, Proc. of SPIE, Orlando, USA, 1998. — Vol. 3364. — P. 371–382.

[4] *Kiy K. I.* An Implementation of a New Approach to Image Concise Description and Segmentation // IEEE Int. Conf. on Computer Vision and Pattern Recognition, Demo Session, San Juan, USA, 1997. — P. 4.

[5] *Форсайт Д. А., Понс Ж.* Компьютерное зрение. Современный подход. — М.: Вильямс, 2004. — 926 с.

Восстановление трёхмерных изображений по плоским проекциям*

Козлов В. Н.
vnkozlov@mail.ru
МГУ им. М. В. Ломоносова, механико-математический факультет

Плоские и трёхмерные изображения — конечные множества точек в соответственно двухмерном и трёхмерном пространствах. Рассматривается восстановление трёхмерного изображения по плоским проекциям. Поточечное соответствие между проекциями априори не задано. Уровень рассмотрения — теоремный. Имеется компьютерная реализация.

Под изображением (двухмерным) понимается конечное (непустое) множество точек на плоскости. Содержательным обоснованием этому может служить то, что любое реальное (нецветное) изображение можно аппроксимировать изображением из точек, причём в нужной мере можно передать все градации «серого цвета» разной плотностью точек в разных частях изображения. Такое представление не закрывает дорогу и к рассмотрению цветных изображений, поскольку, как известно, цветное изображение можно представить тремя нецветными.

Аналогично, трёхмерное изображение — конечное (непустое) множество точек в трёхмерном евклидовом пространстве.

В уже имеющихся, описанных схемах восстановления точка m проецируется на две плоские сетчатки, проекции её есть точки соответственно m_1 и m_2. Если известно положение этих точек на сетчатках, положение сетчаток относительно друг друга, направления проецирования, то, используя геометрические соображения и построения, можно восстановить положение точки m. Если тело T (трёхмерное изображение) состоит из конечного множества точек, то, восстанавливая положение каждой точки, можно восстановить поточечно всё тело. Проекции S_1 и S_2 тела на две сетчатки несколько разные за счет того, что каждый глаз «видит» тело под своим углом зрения, в своём ракурсе.

Главная проблема в рамках машинного стереозрения — это проблема идентификации соответствующих друг другу точек на двух проекциях. Выше мы говорили об одной точке на каждой из сетчаток. Когда таких точек много, то неясно, какую из них на одной проекции сопоставлять данной точке на другой.

Существует ряд содержательных соображений [1], связанных с попытками приложить такого рода схемы, с одной стороны, к машинному стереозрению, с другой — к объяснению стереоскопического зрения в живых организмах. Эти соображения приводят к необходимости восстанавливать трёхмерное изображение не только по паре S_1 и S_2 его плоских проекций, но и по любой паре $\tilde{S}_1$ и $\tilde{S}_2$, полученной из соответственно S_1 и S_2 аффинными преобразованиями, причем какими именно — не известно. Поточечное соответствие между $\tilde{S}_1$ и $\tilde{S}_2$ тоже априори не известно. Именно в такой постановке рассматривается задача в этой работе.

Рассмотрим тело T и прямую, называемую направлением проекции. Направления проекции назовём разными, если они не параллельны. Проведём через каждую точку тела T прямые, параллельные направлению проекции α и называемые лучами. Полагаем α таким, что на каждом луче находится только одна точка тела. Таких направлений проекции бесконечное множество, не таких — только конечное. Назовём плоскость, пересекающую лучи, плоскостью проекции; изображение, образованное точками пересечения лучей с плоскостью проекции — проекцией тела (на данную плоскость и по данному направлению). Рассматриваем проекции тела T по разным направлениям и на разные плоскости. Взаимнооднозначное соответствие между точками двух изображений назовём их разметкой. Соответствующие друг другу точки обозначаем одной буквой (с разными индексами). Ясно, что описанным выше устанавливается взаимнооднозначное соответствие между точками тела T и точками проекций S_i ($i \in \mathbb{N}$). Если a — точка тела T, то точку проекции S_i, лежащую с ней на одном луче, обозначим через a_i и назовём проекцией точки a. Это устанавливает и взаимнооднозначное соответствие между точками проекций S_i и S_j: соответствующие друг другу точки являются проекциями одной и той же точки тела T. Размеченные изображения S_i и S_j назовем a'-эквивалентными, если можно перевести их одно в другое аффинными преобразованиями так, что совместятся соответствующие друг другу точки (обозначение: $A \approx B$). В противном случае S_i и S_j — a'-разные.

Имея тело T, заданное направление проекции α и меняя плоскости проекции, можно получить некоторое множество $\{S\}$ проекций. Все проекции из $\{S\}$ будут попарно a'-эквивалентными. С другой стороны, тело T — не единственное, проецированием которого можно получить множество $\{S\}$ проекций. Таким будет, например, тело T', полученное заменой каждой точки x тела T, находящейся на луче α_x проецирования, на какую-либо дру-

*Работа выполнена при финансовой поддержке РФФИ, проект № 07-01-00433.

гую точку x' на том же луче. В частном и вырожденном случае все точки тела T' могут находиться и в одной плоскости. Итак, при заданном направлении проекции получить данное множество $\{S\}$ проекций можно проецированием некоторого множества $\{T\}$ тел. Из этого следует, что, имея одну или несколько проекций из множества $\{S\}$, нельзя восстановить тело T. Мало того, нельзя даже распознать, не имеем ли мы дело с вырожденным случаем, когда T — двумерное изображение. Проекции из $\{S\}$ можно интерпретировать как изображения тела T в одном ракурсе. Следовательно, для того чтобы восстановить тело, или даже только определить, не двумерное ли оно, нужно иметь более чем одну проекцию, причем в разных ракурсах (то есть при разных направлениях проекции).

Пусть S_1 и S_2 — проекции тела T по двум разным направлениям. Доказано, что S_1 и S_2 a'-разные тогда и только тогда, когда T — не двухмерное изображение. Это дает возможность по точкам a_1, b_1, c_1, d_1 на S_1 и a_2, b_2, c_2, d_2 на S_2 выяснить, лежит ли точка e в теле T в плоскости треугольника abc.

Далее определяется процедура $\operatorname{Alg} T'$ построения некоторого тела T' по $\tilde{S}_1$ и $\tilde{S}_2$. Пусть $a_1b_1c_1$ и $a_2b_2c_2$ — треугольники, и точка e лежит вне плоскости треугольника abc. Выберем некоторую прямую α' (не параллельную плоскости изображения $\tilde{S}_1$) в качестве направления проекции. Проведем через точки a_1, b_1, c_1, e_1 лучи, параллельные α', и на каждом луче возьмём по точке соответственно a', b', c', e' так, чтобы они не лежали в одной плоскости. Далее показывается, как для произвольной точки d_1 на $\tilde{S}_1$ построить соответствующую ей точку d' тела T'.

Теорема 1. *Если S_1 и S_2 суть a'-разные проекции тела T, $S_1 \approx \tilde{S}_1$, $S_2 \approx \tilde{S}_2$, и тело T' построено по $\tilde{S}_1$ и $\tilde{S}_2$ с использованием $\operatorname{Alg} T'$, то тела T и T' a'-эквивалентны.*

Содержательно теорема состоит в следующем. В процедуре $\operatorname{Alg} T'$ присутствует много моментов, которые могут варьироваться. Можно, например, выбирать разные $\tilde{S}_1$ и $\tilde{S}_2$, по разному брать исходную четвёрку точек на $\tilde{S}_1$ (или на $\tilde{S}_2$) и т. д. Варьируя такие моменты, можно построить некоторое множество $\{T\}$ тел. Теорема, однако, утверждает, что все они a'-эквивалентны телу T и, значит, a'-эквивалентны между собой. Тело посредством процедуры $\operatorname{Alg} T'$ восстанавливается, тем самым, с точностью до аффинных его преобразований. При этом нет необходимости знать расстояние между проекциями S_1 и S_2 (аналог расстояния между сетчатками глаз), направления проекции, тело T' строится по произвольным образом сдвинутым, повернутым, сжатым или растянутым, уменьшенным или увеличенным проекциям тела T.

Процедура $\operatorname{Alg} T'$ и теорема позволяют определить то, что можно назвать приемлемой разметкой (то есть поточечное соответствие между $\tilde{S}_1$ и $\tilde{S}_2$).

Модель реализована в виде программы для компьютера для восстановления трёхмерного изображения по стереофото (стереопаре).

Литература

[1] *Козлов В. Н.* Введение в математическую теорию зрительного восприятия. — М.: Издательство Центра прикладных исследований при механико-математическом факультете МГУ, 2007. — 136 с.

Оценка качества распознавания состояний динамической системы*

Колесникова С. И., Цой Ю. Р.
skolesnikova@yandex.ru
Томский государственный университет систем управления и радиоэлектроники,
Томский политехнический университет

Рассматривается подход к оценке качества процедуры распознавания состояний динамической системы, математическая модель которого имеет достаточно общий характер. Основу данного подхода к оценке качества распознавания составляет заданная информация об оценках вероятностей проявления признака для каждого из состояний. Получены двусторонние оценки для вероятности достижения гарантированного уровня качества процедуры распознавания, при этом использована вероятностная модель пересечения случайным процессом криволинейного уровня.

В данной работе рассматривается вероятностный подход к исследованию возможности оценивания минимальной сложности модели (число признаков и их весовые коэффициенты, число объектов, уровень накопленной статистики, выраженной в частотах встречаемости признаков для каждого из классов) для достижения заданного качества распознавания на предоставленной выборке, объем которой может изменяться динамически.

Общая постановка прикладной задачи

Пусть задан случайный процесс (СП) (X, Y), характеризующий состояние динамической системы (ДС), функционирующего на интервале времени от t_0 до T , где $X(t)$ — вектор переменных состояния системы;

$$Y(t) = f(X(t), \eta) + \xi(t) \qquad (1)$$

— случайная наблюдаемая N_Y-мерная векторная функция; $\xi(t)$, $\eta(t)$ — шумы достаточно общей природы. Относительно СП (X, Y) выдвинуто $I > 1$ альтернативных гипотез $\Omega = \{\Omega_1, \ldots, \Omega_I\}$, составляющих полную группу событий и физически интерпретируемых как классы состояний частично наблюдаемой ДС. Наблюдение величины $Y(t)$ осуществляется в моменты $t_j = t_0 + j\Delta$, $j = 1, \ldots, n$, с шагом дискретизации $\Delta > 0$, по модели данной ДС. Задача состоит в отнесении (в момент t или на некотором фиксированном интервале $[t', t'']$) наблюдаемой реализации $Y(t)$ к состоянию (классу) Ω_i, $i = 1, \ldots, I$.

Сложность практического решения поставленной задачи обусловлена не только наличием возможной нелинейной связи между откликом (выходом) и входным воздействием, но и отклонениями параметров модели от реальных значений (присутствием ошибок систематических и несистематических, неизбежными при моделировании — математическом или имитационном), реальностью времени решения задачи распознавания (диагностики) состояния для последующего (возможного) управления в реальном времени.

Несмотря на большое количество методов интеллектуальной обработки диагностической информации (см., например, обзор в [3]), вопрос об идентификации состояний ДС в такой постановке остается открытым. В связи с этим представляют интерес методы качественного исследования нелинейной ДС. Применение методов теории распознавания образов [1, 2] для решения данной задачи распознавания состояний ДС обосновывается: во-первых, наличием объективной связи между структурой ДС и её состоянием, выражающейся в виде (1); во-вторых, адекватным представлением структуры ДС выбранной моделью, на основе которой возможно получение обучающей выборки; в-третьих, состояния (или подмножества значений $(X(t))$, «восстановленные» на основе обучающей выборки, могут быть соотнесены с реальным поведением ДС.

В работах [3, 4] представлена модель (и реализующий ее комплекс методов и алгоритмов) автоматизированного распознавания состояний ДС — электромеханической системы (ЭМС). Предложена процедура, синтезирующая коллективное итоговое решение на основе нескольких алгоритмов распознавания.

Излагаемая в [4] процедура основана на эквивалентности различия энтропий и условных распределений квантованных значений сигнала (признака) в разных состояниях. Процедура распознавания состояний ДС состоит из 2-х этапов: итогом первого этапа является совокупность информативных (характеристических) признаков в виде оценок условных распределений квантованных значений сигнала в разных состояниях ДС; итогом второго этапа является принятие решения о принадлежности наблюдаемого сигнала одному из состояний ДС на основе определенным образом введенного информационного расстояния (обобщенной меры) между структурным «эталонным» составом сигналов в каждом состоянии и составом, сформированным на основе «новых» наблюдений.

*Работа выполнена при финансовой поддержке РФФИ, проекты №07-01-00452, №09-01-99014.

Проведенные численные эксперименты в работе [4] по распознаванию состояний оцифрованного выходного сигнала в реальном времени показали следующее. При выборе признаков (здесь — целого числа шагов квантования для кодирования сигнала) по обучающей выборке возникает проблема их анализа и учета значимости из большого объема обучающей выборки (порядка 15000 и более, зависящего от интервала дискретизации сигнала $\Delta > 0$).

При проведении численного моделирования и анализа качества распознавания в зависимости от числа уровней квантования было замечено, что, начиная с некоторого значения шага квантования, эффективность распознавания (являя случайный характер) в среднем оставалась на постоянном уровне, т. е. дополнительная «подробность» в знании особенностей сигнала, существенно замедляющая принятие решение о метке состояния в реальном времени (критичном для работы), не влияла на качество распознавания, а иногда и ухудшала его (рис. 1).

Рис. 1. Качество распознавания состояния ДС в зависимости от уровня квантования.

Основные понятия и определения

«Привязка» к классической терминологии задачи распознавания состоит в следующем [1]. Исследуется некоторое множество объектов $O = O_1, \ldots, O_M$ (распознаваемые сигналы). Известно, что O представимо в виде объединения I подмножеств $\Omega = \{\Omega_1, \ldots, \Omega_I\}$, называемых классами (состояниями в контексте данной работы). Предполагается, что объекты из O описываются некоторой системой признаков $Z = Z_1, \ldots, Z_n$ (значений сигналов с метками состояний). Отметим, что для вышеописанной прикладной задачи $n = n(\Delta)$, где $\Delta > 0$ — «динамический» параметр дискретизации сигнала; при $\Delta \to 0$ $n(\Delta) \to \infty$. Данное замечание не ограничивает общности и продиктовано характером прикладной задачи.

Следует уточнить, что дискретизация разбивает сигнал по временной составляющей, а квантование разбивает диапазон значений (амплитуду) сигнала на отрезки равной длины (уровни квантования).

Под надежностью распознавания (качеством работы процедуры распознавания (ПР)) будем понимать способность ПР обеспечить заданный уровень доли правильно распознанных объектов из исследуемой выборки.

Под признаком будем понимать некоторое обобщенное его значение: конъюнкция (по Бонгарду), тестовый набор [1], эталонный элемент, число градаций при квантовании сигнала [4] и т. д.

Заметим (по поводу условия $n \to \infty$), что количество безызбыточных (тупиковых) тестов при числе признаков, равном несколько десятков, может быть весьма большим [5], и нахождение всех безызбыточных тестов становится нереальным даже при современных компьютерных технологиях.

Без ограничения общности будем считать признаки $Z = \{Z_1, \ldots, Z_n\}$ ранжированными каким-либо образом, например, по убыванию весовых коэффициентов признаков (ВКП), вычисленных по какому-либо методу (обзор которых, для разного вида исходной информации, сделан в [5]) и являющихся числовой мерой качества признаков.

Неравенство $|z_j^i - a_j^i| < \delta^i$ будет означать, что признак Z_j (тестовый набор, конъюнкция, эталон, сигнал) корректно распознает объект из i-го класса с точностью δ^i, если расхождение между значениями z_j^i признака Z_j и объекта a_j^i в i-м классе в пределах допустимого. Обозначим через p_{ij} вероятность этого события.

Обозначим через ζ_j случайную величину (СВ), означающую число корректно распознанных объектов (КРО) признаком Z_j, $\zeta_j = \sum\limits_{i=1}^{I} \varkappa_j^i$,

$$\varkappa_j^i = \chi\big[|x_j^i - a_j^i| < \delta^i\big], \quad \chi[x] = \begin{cases} 1, & x = \text{true} \\ 0, & x = \text{false} \end{cases};$$

$\Phi_n = \frac{1}{nm} \sum\limits_{j=1}^{n} w_j \xi_j$ — нормированное, центрированное и взвешенное число КРО по всей совокупности признаков. Здесь $\xi_j = \xi_j - \nu_j$, а величины $\nu_j = \mathsf{E}\zeta_j$, σ_j^2, w_j — математическое ожидание СВ, дисперсия, ВКП j-го признака, соответственно.

Задача распознавания состояния ДС с гарантированной надежностью

Рассмотрим СВ τ — первый момент превышения заданного порогового значения g_0 или требуемого значения качества распознавания (доли корректно распознанных объектов):

$$\tau = \inf\Big\{n \geqslant 1 \colon \sum_{j=1}^{n} \zeta_j \geqslant g_0\Big\}, \tag{2}$$

или

$$\tau = \inf\big\{n \geqslant 1 \colon \Phi_n \geqslant g_n\big\}, \tag{3}$$

где $g_n = nmg_0 - \sum\limits_{j=1}^{n} w_j\nu_j$.

Поставим задачу оценить вероятностное распределение величины τ, или минимально необходимого числа признаков, обеспечивающего заданный уровень надежности распознавания, если известно вероятностное распределение признаков (что естественно не только для решаемой в работах [3, 4] прикладной задачи по распознаванию состояний ДС, но и для приложений, где накоплены соответствующие статистические данные). Знание распределения величины τ позволит оптимизировать процедуру распознавания состояния ДС не только по весовым коэффициентам признаков, но и по интервалам дискретизации и квантования сигнала.

Подход к решению задачи

Для решения поставленной задачи будем использовать вероятностную модель пересечения случайным процессом определенного уровня. Подход, который будет использован для получения оценок вероятностного распределения величины τ (момента остановки случайного процесса), применен А. А. Новиковым [6, 7] для получения асимптотики вероятности непересечения криволинейной границы суммой независимых неодинаково распределенных СВ.

Суть этого подхода заключается во введении новой вероятностной меры, относительно которой момент пересечения криволинейной границы суммой независимых разнораспределенных СВ (2), (3) является моментом пересечения мартингалом некоторого постоянного уровня, для которого уже нетрудно получить непосредственно нужные оценки. Задача, таким образом, сводится к доказательству существования подходящей новой вероятностной меры в данных условиях.

Несмотря на то, что накладываемые ограничения в [6] на поведение неслучайных границ делают невозможным напрямую использовать полученные результаты, тем не менее, будет показано, что используемые рассуждения, по существу, сохраняются и позволяют получить требуемое преобразование и в данной постановке задачи.

Последовательность ξ_j, $j = 1, \ldots, n$, будем считать заданной на вероятностном пространстве (Ω, F, P). Введем последовательность неубывающих σ-алгебр: $F_0 = (\oslash, \Omega)$, $F_n = \sigma\{\xi_1, \ldots, \xi_n\}$. Пусть $\{(a_{j-1}, F_j)\}$ — предсказуемая последовательность, т. е. для каждого j величины (a_{j-1}, F_j) — измеримые [6, 7].

Общая схема получения оценок надежности распознавания

1. Определяется функция $\varphi_j(\lambda)$: $\mathsf{E}\exp(\lambda\xi_j) = \exp(\varphi_j(\lambda))$. Это возможно в силу ограниченности СВ $|\xi_j| \leqslant L_j$, т. к. $|\zeta_j| \leqslant m$. Функция $\varphi_j(\lambda)$ является бесконечно дифференцируемой, и непосредственно из ее определения следует, что

$$\varphi_j'(\lambda) = \mathsf{E}\,\xi_j \exp(\lambda\xi_j - \varphi_j{}'\lambda);$$
$$\varphi_j''(\lambda) = \mathsf{E}\,(\lambda\xi_j - \varphi_j(\lambda))^2 \exp(\lambda\xi_j - \varphi_j(\lambda)).$$

2. На основе функций $\varphi_j(\lambda)$, $j = 1, \ldots, n$, на (Ω, F_{n-1}) вводится вероятностная мера $\widetilde{\mathsf{P}}_n$ по формуле: $\widetilde{\mathsf{P}}_n = \mathsf{E}\chi_A G_n$, где χ_A — индикатор множества , а последовательность $\{G_n\}$ имеет вид: $G_n = \sum\limits_{j=1}^{n} (a_j\xi_j - \varphi_j(a_j))$, где $\{a_j\}$, $j = 1, \ldots, n$, — последовательность ограниченных предсказуемых СВ, подлежащих определению, при этом существенно используется факт принципиальной разрешимости уравнения $\varphi_j(\lambda) = x$ при малых x (см. ниже формулировки и условия лемм). Известно [6, 7], что последовательность $\{G_n, F_n\}$ — мартингал относительно меры P со средним $\mathsf{E}G_n = 1$.

3. Доказывается, что относительно $(F_n, \widetilde{\mathsf{P}}_n)$ исходный процесс Φ_n является мартингалом с нулевым средним, а подвижная граница g_n некоторым постоянным уровнем, для которого выводятся нужные оценки на основании лемм Блэкуэлла и Фридмана, Новикова, Ширяева [6, 7].

4. Осуществляется корректный возврат в исходное пространство (Ω, F, P).

Сформулируем основной результат в виде оценок для надежности распознавания (двусторонних границ) вероятности $R_n = \mathsf{P}(\tau > n)$, имеющих место при выполнении условий, связывающих вероятностные (статистические) характеристики признаков и их ВКП.

Лемма 1. *Функция $\varphi_j'(\lambda)$ не убывает по (λ) и справедлива оценка:*

$$|\varphi_j'(\lambda)| \geqslant |\lambda|\,\sigma_j^2\,\exp\left\{-|\lambda|L_j^{-1}\right\},$$

для любых $|\lambda| \leqslant \Lambda_c = \ln c/2L_j^{-1}$, $c > 1$.

Лемма 2. *При выполнении условия*

$$mg_0 w_j^{-1}/(\Lambda_c\sigma_j^2 L_J^{-1} + \nu_j) \leqslant 1$$

последовательность ограниченных предсказуемых СВ $\{a_j\}$, $j = 1, \ldots, n$, определяется из уравнения:

$$\varphi_j'(a_j) = mg_0 - w_j\nu_j.$$

Теорема 3. *В условиях лемм 1 и 2 имеют место оценки для распределения числа обобщенных признаков (с соответствующими весовыми коэффициентами), обеспечивающих заданную надежность распознавания*

$$R_n \leqslant c_1 \exp\left\{-c_2 g_0^{-2}\sum_{j=1}^{n}\sigma_j^2 + c_3\sum_{j=1}^{n} v_j\right\};$$

$$R_n \geqslant \exp\left\{-c_4\left(\alpha^2 + g_0^{-2}c\sum_{j=1}^{n}\sigma_j^2\right) - c_5\sum_{j=1}^{n} v_j\right\},$$

где $v_j = (g_0 m - \nu_j)/\sigma_j^2$, *величины* $c_i,\ i = 1, \ldots, 5$ *не зависят от* n *и являются известными функциями от величин* α, p, q, c, $c_1 = \left(\frac{5}{4\alpha}\right)^{\frac{1}{2p}}$, $c_2 = \frac{\pi^2}{8p}\left(1 - 3\alpha^{\frac{1}{2}}\right)c^{\frac{1}{2}}$, $c_3 = q - 1$, $c_4 = \frac{\pi^2 p}{8}(1 + 15\alpha^{\frac{3}{4}})$, $c_5 = \frac{q+1}{q-1}$, $\alpha \leqslant \frac{1}{16}$, $\frac{1}{p} + \frac{1}{q} = 1$, $p > 1$.

Доказательство теоремы существенно опирается на результаты, полученные в [6, 7], факт разрешимости уравнения $\varphi_j(\lambda) = x$, имеющий место при выполнении условий леммы 2 и свойств функции $\varphi_j(\lambda)$.

Полученный теоретический результат дает возможность найти минимальное значение числа используемых при распознавании признаков для достижения заданного качества на контрольной динамической выборке, тем самым определить оптимальный уровень квантования сигнала не в ущерб надежности распознавания состояний динамической системы за реальное время.

Заключение

Применимость вероятностных и статистических подходов к распознаванию образов в интеллектуальных системах принятия решений возможна в случае пропусков данных, большой размерности задачи и пр. и никак не связана с тем, в какой шкале измеряются значения оцениваемых признаков.

Рассмотренный вероятностный подход для оценивания качества статистической процедуры состояний динамической системы с целью определения оптимального соотношения величин уровня квантования и дискретизации («число признаков и их весовые коэффициенты — число объектов») может быть также использован при выборе алгоритма распознавания по обучающей выборке статистического характера.

Иллюстративным примером практического применения полученных в данной работе оценок надежности распознавания состояний ДС для предложенных в [3, 4] процедур является их апробирование для распознавание состояний ЭМС с частотно-регулируемым электроприводом переменного тока, являющейся основой нижнего уровня управления большинства современных автоматизированных систем в таких отраслях, как нефтегазовая, горнорудная, металлургическая и др.

В таблице 1 находятся результаты численных экспериментов, проведенных по модели ЭМС — асинхронного двигателя. В 1-й и во 2-й колонках — минимальное число квантов для кодирования сигнала с минимальной амплитудой, указанное экспертом по данным обучающей выборки (Э-оценка) и полученное по формуле (3) (τ), соответственно; в 3-й и 4-й колонках — нижняя (R_1) и верхняя (R_2) оценки вероятности (R_n), соответственно.

Таблица 1. Экспериментальные значения оценок

Э-оценка	τ	R_1	R_2
5	4	0,46	0,96
8	5	0,54	0,89
10	7	0,63	0,91

Результаты численного моделирования в прикладной задаче распознавания состояний ЭМС-объекта (см. таблицу 1) показали непротиворечивость полученных оценок, однако целесообразно дальнейшее исследование этого вопроса, связанное с выяснением свойств оценок и их уточнением.

Предложенные оценки могут быть использованы при организации робастного и адаптивного управления сложными ЭМС в реальном времени.

Литература

[1] *Журавлев Ю. И., Рязанов В. В., Сенько О. В.* Распознавание. Математические методы. Программная система. Практические применения. — М.: Фазис, 2005. — 159 с.

[2] *Загоруйко Н. Г.* Прикладные методы анализа данных и знаний. — Новосибирск: Изд. ИМ СО РАН, 1999. — 273 с.

[3] *Колесникова С. И., Букреев В. Г.* Распознавание состояний динамической системы // Труды XI Международной научно-технической конференции и выставки «Цифровая обработка сигналов и ее применение» DSPA-2009, М.: ИПУ РАН, 2009. — Т. 2. — С. 155–158.

[4] *Колесникова С. И., Букреев В. Г.* Информационный подход к распознаванию состояний динамической системы // Труды X Международной научно-технической конференции «Кибернетика и высокие технологии XXI века» СТ-2009, Воронеж: ВГУ, 2009. — Т. 2. — С. 105–117.

[5] *Колесникова С. И., Янковская А. Е.* Оценка значимости признаков для тестов в интеллектуальных системах // Известия РАН. Теория и системы управления. — 2008. — № 6. — С. 135–148.

[6] *Новиков А. А.* О времени выхода сумм ограниченных случайных величин из криволинейной полосы // Теория вероятностей и ее применения. — 1981. — Т. 26, № 2. — С. 287–301.

[7] *Ширяев А. Н.* Вероятность. — М.: Наука, 1980. — 575 с.

Построение параметрического портрета динамической системы на основе синдромальных представлений*

Котельников И. В.
neymark@pmk.unn.ru
Научно-исследовательский институт прикладной математики и кибернетики
Нижегородского государственного университета им. Н. И. Лобачевского

Приводятся процедуры распознавания образов на основе оптимальных тупиковых нечетких тестов и синдромов для определения множества фазовых портретов, множества аттракторов многомерной многопараметрической динамической системы и областей пространства параметров, на которых они существуют. Выделяются части таких областей с достоверностью результатов, приближающейся к единице.

Содержание доклада является изложением результатов раздела большой темы по исследованию многомерных многопараметрических динамических систем (ДС) методами распознавания образов и статистического моделирования. Выбор объекта и методов исследования не случаен. Исследование ДС классическими методами в настоящее время полностью формализовано только для размерности системы $n \leqslant 2$. При $n = 3$ исследовать возможно, но уже достаточно трудно, и то, только с помощью вычислительной техники. При $n \geqslant 4$ исследовать очень трудно и часто невозможно. Сложившуюся ситуацию Р. Беллман назвал «проклятием размерности». Классические методы А. Пуанкаре и Д. Биркгофа, как отмечал академик А. А. Андронов, «дают нам известное представление о роде и характере движений, но не содержат почти никаких данных для исследования конкретных динамических систем, с которыми мы только и имеем дело». С другой стороны, в распознавании образов мы не встречаем непреодолимых трудностей ни при большом числе признаков объектов, ни при большом числе классов их распознавания. Это, в основном, и определило выбор объекта и методов его исследования [1]. Важно подчеркнуть, что предлагаемый подход является общим формализованным подходом к исследованию многомерных многопараметрических ДС, заданных системой обыкновенных дифференциальных уравнений.

Некоторые понятия, определения, свойства

ДС рассматривается как математическая модель, заданная системой обыкновенных дифференциальных уравнений. Состояние ДС в конкретный момент времени характеризуется положением фазовой точки в пространстве переменных, или фазовом пространстве. С течением времени фазовая точка перемещается в фазовом пространстве по определенной траектории — последовательности своих положений в последовательные моменты времени, являющихся решением дифференциальных уравнений в эти моменты. В силу единственности решения ни сама траектория, ни какие-либо две различных траектории не могут пересекаться в фазовом пространстве ни в одной из своих точек. Для упрощения исследования в предлагаемом подходе рассматриваются лишь такие траектории ДС, которые ведут к устойчивому состоянию равновесия, устойчивому предельному циклу или ограниченной области фазового пространства, в которой фазовая точка совершает хаотические или стохастические движения и никогда из нее не выходит. Области множества таких предельных состояний ДС называются *аттракторами* ДС. Движение фазовой точки не ограничено во времени. Для численного исследования вводится ограничение времени значением T, называемым временем интегрирования. Значение T должно быть достаточным для того, чтобы траектория не только достигла аттрактора ДС, но и внутри области аттрактора была продолжена настолько, чтобы можно было достоверно распознать тип аттрактора рассматриваемой траектории. Для получения траектории необходимо задать значения параметров ДС и начальные условия — положение фазовой точки при нулевом значении времени. Для выбора начальных условий используется не всё бесконечное фазовое пространство, а лишь его часть, ограниченная, например, n-мерным параллелепипедом N. Аналогично, выбор значений параметров производится не из бесконечной области параметров, а из выделенной его части, например, p-мерного параллелепипеда P, где p — число параметров ДС. Конкретные размеры и расположение параллелепипедов N и P задаются на основе знаний предметной области, к которой относится ДС.

При заданных значениях параметров ДС её траектории могут относиться к различным аттракторам в зависимости от начальных условий. Полный набор возможных аттракторов ДС при заданных значениях параметров называется ее *фазовым портретом* при этих параметрах. В пространстве параметров существуют области, задание параметров из которых обязательно приведет к получению одного и того же фазового портрета в фазовом пространстве. Множество таких областей для множе-

*Работа выполнена при финансовой поддержке РФФИ, проект № 08-01-00248.

ства возможных фазовых портретов назовём *параметрическим портретом* ДС. Понятие параметрического портрета представляется новым в исследовании ДС. Мы не нашли нигде ни введения этого понятия, ни какого-либо формализованного способа его построения.

Построение фазовых и параметрического портретов, как было видно, сопровождается рядом ограничений. В силу этого, и фазовые, и параметрический портреты называются *огрублёнными*. Ниже для простоты это определение опускается.

Постановка задачи

Для многомерных многопараметрических ДС, заданных системой обыкновенных дифференциальных уравнений, определить формализованные процедуры распознавания образов и статистического моделирования для получения:

1) множества различных фазовых портретов ДС и вытекающего из него множества различных аттракторов ДС;
2) областей пространства параметров P, соответствующих конкретным фазовым портретам и аттракторам ДС, а также частей этих областей с достоверностью результатов, приближающейся к единице.

Получение множества фазовых портретов и аттракторов динамической системы

Для решения поставленной задачи формируется выборка наборов параметров ДС из заданной области P пространства параметров случайным выбором на основе равномерного распределения. Для каждого набора параметров формируется выборка начальных условий случайным выбором на основе равномерного распределения из заданной области N начальных условий в фазовом пространстве. На основе этой информации определяется фазовый портрет ДС для данного набора параметров путем построения фазовых траекторий для выбранных начальных условий и определения типов соответствующих траекториям аттракторов [2].

Ведётся список СФП полученных фазовых портретов и список полученных аттракторов СА . После получения очередного фазового портрета проверяется его наличие в списке СФП и, если он новый, то вносится в конец списка. Порядковый номер набора параметров, на котором получен фазовый портрет, заменяется в выборке наборов параметров на номер фазового портрета в списке СФП. Аналогичная процедура производится с каждым из полученных в фазовом портрете аттрактором ДС и списком СА аттракторов.

Процедура сравнения предполагает наличие какого-то кодирования аттракторов и фазовых портретов. За основу берется кодирование аттракторов ДС n-мерными кодами по числу фазовых переменных ДС. Для предельного цикла выбирается код из двоек во всех n позициях. Для хаотических движений выбирается аналогичный код из троек. В кодах аттракторов устойчивых состояний равновесия значением 0 выделяются позиции с нулевыми значениями фазовых переменных, значением 1 — конкретные ненулевые значения и значением 9 — множественные ненулевые значения, соответствующие многообразиям устойчивых состояний равновесия. Например, для 4-мерной ДС код 1111 соответствует устойчивому состоянию равновесия со всеми ненулевыми значениями фазовых переменных, код 0110 — устойчивому состоянию равновесия с двумя нулевыми и двумя ненулевыми значениями фазовых переменных, код 0900 соответствует многообразию устойчивых состояний равновесия на оси второй фазовой переменной, а код 0990 — многообразию устойчивых состояний равновесия на плоскости второй и третьей фазовых переменных.

Коды фазовых портретов составляются из кодов соответствующих им аттракторов. Например, код (1111 1111 0910 2222) соответствует фазовому портрету, состоящему из двух различных устойчивых состояний равновесия с ненулевыми значениями фазовых переменных, многообразия устойчивых состояний равновесия на прямой, параллельной оси второй фазовой переменной при фиксированном ненулевом значении третьей фазовой переменной и нулевыми значениями первой и четвертой, и предельного цикла.

Два аттрактора считаются тождественными, если они совпадают по своим кодам. Два фазовых портрета считаются тождественными, если они содержат равное число типов аттракторов, а внутри каждого типа имеют равное число аттракторов с тождественными кодовыми обозначениями. Например, фазовые портреты (2222 1111 1111) и (2222 1111) считаются различными, так как, совпадая по числу типов аттракторов (устойчивый предельный цикл и устойчивое состояние равновесия), они различаются по числу аттракторов устойчивых состояний равновесия. Многообразия устойчивых состояний равновесия при сравнении рассматриваются наравне с аттракторами.

Рассмотренная процедура реализуется на последовательных частях исходной выборки наборов параметров. После реализации очередной части известно число полученных фазовых портретов и аттракторов. Отсутствие увеличения числа фазовых портретов и аттракторов на ряде последовательных частей выборки является признаком достижения процедурой некоторого предельного состояния и возможной причиной для останова процедуры. В остальных случаях момент останова процедуры

выбирает исследователь ДС в зависимости от конкретных стоящих перед ним задач исследования.

Определение областей параметров для фазовых портретов и аттракторов

После окончания процедуры получения фазовых портретов и аттракторов реализованная часть выборки наборов параметров представляет собой готовую обучающую выборку для построения разделяющего решающего правила полученных фазовых портретов в пространстве параметров, поскольку порядковые номера наборов параметров в выборке заменены номерами фазовых портретов (классов) из списка СФП.

Построение решающего правила производится методом оптимальных тупиковых нечётких тестов и синдромов [3], несколько адаптированного к решению рассматриваемой задачи. В результате построения решающего правила получаем покрытие объектов каждого класса выборки, т. е. класса конкретного фазового портрета, набором оптимальных синдромов. Под синдромом понимается в данном случае p-мерный параллелепипед пространства параметров, внутри и на поверхности которого располагаются только точки наборов параметров соответствующего синдрому класса. В нашем случае это будут точки наборов параметров конкретного фазового портрета ДС. Области покрытия могут состоять из большого числа синдромов, часто очень сильно перекрывающихся внутри своего класса. Поскольку формирование выборки производилось на основе равномерного распределения, число точек наборов параметров каждого класса будет пропорционально размерам области параметров класса. Поэтому фазовые портреты с малыми размерами области параметров будут представлены небольшим числом наборов параметров. Однако, предлагаемый подход к исследованию ДС не ставит целью получить с высокой достоверностью всю область параметров конкретных фазовых портретов. С высокой достоверностью предполагается получать лишь часть такой области. Эти части для каждого класса получаются из синдромов покрытия решающего правила с максимальным числом покрываемых наборов параметров.

Области параметров отдельных аттракторов представляют собой объединение областей параметров всех фазовых портретов, которые содержат данный аттрактор. Это означает, что для выбранного таким объединением пространства параметров конкретного аттрактора в фазовом пространстве могут существовать по соседству многие другие аттракторы ДС. Иногда такое соседство может быть опасно для нормальной работы, а иногда и для существования, реальной ДС. Построение рассмотренного параметрического портрета позволяет избежать опасного соседства путём объединения областей параметров только тех фазовых портретов, которые не содержат опасных аттракторов. Это производится довольно простой процедурой на основе полученной обучающей выборки параметров при построении фазовых портретов и аттракторов, и списков фазовых портретов и аттракторов ДС. Более конкретно, реализуется процедура

$$\bigcup_{i=1}^{k} \bigcap_{j=1}^{m} a_{ij}, \qquad (1)$$

где k — число объединяемых областей фазовых портретов, а a_{ij} — порядковые номера аттракторов, образующих фазовый портрет, в списке СА аттракторов. Результатом процедуры (1) является формирование из обучающей выборки параметрического портрета новой обучающей выборки из двух классов. Первый класс содержит наборы параметров фазовых портретов, отобранных процедурой (1) для объединения, и содержащих аттрактор, для которого формируется область. Второй класс объединяет все остальные наборы параметров выборки параметрического портрета. Построение решающего правила по полученной обучающей выборке даёт синдромальное покрытие области параметров рассматриваемого аттрактора ДС.

Формирование областей параметров высокой достоверности

Исходной информацией для построения области параметров высокой достоверности для выбранного фазового портрета является синдром S данного фазового портрета с максимальным числом точек наборов параметров, полученный в решающем правиле в процедуре построения множества фазовых портретов и аттракторов ДС. При малых размерах области параметров рассматриваемого фазового портрета число точек синдрома S может быть очень малым (возможно даже единицей). В этом случае строится искусственный синдром S небольшого размера, покрывающий имеющиеся точки наборов параметров синдрома. На первом этапе необходимо получить синдром S с числом точек $m^* \geqslant 5p$, где p — размерность пространства параметров. По смыслу это значение соответствует минимальному числу объектов обучающей выборки, на которой можно получить решающее правило приемлемой достоверности. Эта задача выполняется реализацией рассмотренной выше процедуры построения множества фазовых портретов и аттракторов, но при выборе точек наборов параметров не из большой области P пространства параметров, а из области S рассматриваемого синдрома. После реализации очередной части исходной выборки наборов параметров на основе полученной

обучающей выборки строится разделяющее решающее правило, и анализируется число точек m наборов параметров в самом результативном по числу точек синдроме S выбранного фазового портрета. При $m < m^*$ производится переход к реализации очередной части исходной выборки параметров. В противном случае — выход из процедуры. В результате мы получили синдром S с числом точек параметров $m \geqslant m^*$. С помощью процедуры (1) при $k = 1$ (рассматривается только один фазовый портрет) получим из основной обучающей выборки новую обучающую выборку из двух классов, первый из которых содержит наборы параметров только рассматриваемого фазового портрета. Запомним эту часть выборки как выборку F. Это подготовительная часть процедуры. Рассмотрим её основную часть.

Зададим значение D^* достоверности результата, с которой мы хотим получить решение. Реализуем процедуру определения фазовых портретов и аттракторов на выборке наборов параметров фиксированной длины m_s, полученной из синдрома S случайным выбором на основе равномерного распределения. Из полученной в результате реализации процедуры обучающей выборки сформируем с помощью процедуры (1) при $k = 1$ (рассматривается только один фазовый портрет) новую обучающую выборку из двух классов, первый из которых содержит только точки наборов параметров рассматриваемого фазового портрета. Подсчитаем значение полученной достоверности $D = m_t/m_s$, где m_t — число наборов параметров, на которых получен нужный нам фазовый портрет. При $D < D^*$ добавим к полученной обучающей выборке двух классов выборку F и построим на этой выборке решающее правило. В результате получим новый синдром S, соответствующий прежнему, но с исключением из него области параметров, соответствующей другим фазовым портретам. Дополнение обучающей выборки двух классов выборкой F производится для того, чтобы закрепить за синдромом S его область параметров с фазовым портретом, для которого определяется область высокой достоверности. Как видно из предыдущего, выборка F не принимает участия в подсчете достоверности D. Для вновь полученного синдрома S сформируем новую выборку F, состоящую из точек наборов параметров рассматриваемого фазового портрета, на которой получен синдром. После этого следует переход в начало основной части процедуры для построения фазовых портретов и аттракторов на новой части точек из m_s наборов параметров, выбранных случайным выбором на основе равномерного распределения из нового синдрома S. При $D \geqslant D^*$ — конец процедуры поиска области параметров с заданной достоверностью. Такой областью будет синдром S, на котором достигнута нужная достоверность.

Аналогичным образом можно определить область параметров высокой достоверности и для конкретного аттрактора ДС. Отличие будет состоять в значении k процедуры (1), где k будет равно числу фазовых портретов, содержащих данный аттрактор.

Выводы

Рассмотренные процедуры распознавания образов и статистического моделирования реализованы в комплексе программ. Полученное программное обеспечение успешно апробировано на очень сложной для исследования четырехмерной математической модели иммунного ответа организма на вторжение инфекции с 14 параметрами. Получены списки фазовых портретов и аттракторов ДС соответственно из 87 фазовых портретов и 20 аттракторов. С заданным уровнем достоверности $D = 0.99$ получены области параметров для ряда фазовых портретов, включая фазовые портреты, представленные в параметрическом портрете лишь одним набором параметров.

Планируется апробация математического обеспечения на математических моделях из других предметных областей. Общность предложенного подхода предполагает успешное решение и этой части исследования.

Литература

[1] *Неймарк Ю. И.* Компьютерная концепция исследования конкретных динамических систем // VII Всероссийская конференция «Нелинейные колебания механических систем», Нижний Новгород: Изд. Нижегородского госуниверситета. — 2005. — С. 17–18.

[2] *Котельников И. В.* Синдромальные процедуры распознавания для исследования фазового пространства конкретных многомерных динамических систем // ММРО–13. М.: МАКС Пресс. — 2007. — С. 146–149.

[3] *Kotel'nikov I. V.* A Sindrome Recognition Method Based on Optimal Irreducible Fuzzy Tests //Pattern Recognition and Image Analysis. — 2001. — V. 11, № 3. — Pp. 553–559.

Многослойное древовидное представление объектов многоканальных изображений*

Ланге М. М., Степанов Д. Ю.
lange_mm@ccas.ru
Москва, Вычислительный Центр РАН

На основе обобщения процедуры представления двумерных тел с полутоновой окраской предложен способ описания объектов многоканальных изображений стопками деревьев. Введена мера различия стопок деревьев. Используя пространство трёхслойных древовидных представлений для описания RGB-объектов, продемонстрирован выигрыш в распознавании лиц, заданных цветными изображениями по сравнению с полутоновыми.

Структурные методы представления данных составляют основу для сокращения вычислительной сложности алгоритмов обработки изображений и, в частности, алгоритмов распознавания образов. К таким методам относятся вейвлет-преобразования [1] и методы древовидной декомпозиции изображений [2, 3]. Для широкого класса объектов, заданных на полутоновых изображениях двумерными телами, в работе [4] предложен метод представления таких объектов деревьями эллиптических примитивов. Этот метод позволил построить быстрый классификатор в пространстве древовидных представлений объектов, который продемонстрировал высокие показатели качества распознавания жестов и подписей с вероятностью ошибки порядка 0,01. Естественное развитие этого подхода состоит в построении древовидных представлений для объектов, заданных несколькими (многоканальными) изображениями. Примерами многоканальных изображений являются полутоновые изображения объектов, снятых в различных ракурсах, а также цветные RGB-изображения. Использование многоканальных данных эквивалентно увеличению размерности пространства представлений и в соответствии с фундаментальной теоремой теории информации [5] должно привести к снижению вероятности ошибки распознавания, выполняемого по схеме декодирования.

Представление двумерных объектов стопками деревьев

Способ представления, предложенный в работе [4], переводит любой двумерный объект, заданный полутоновым изображением и имеющий в плоскости изображения идентифицируемую систему собственных координат, в завершённое бинарное дерево эллиптических примитивов. При достаточно малых размерах пикселя такие представления инвариантны к преобразованиям поворота, смещения и масштабирования объектов в плоскости изображений, а также к изменению уровня яркостной окраски объектов. Пример древовидного представления информативного участка лица, выделенного на полутоновом изображении, показан на рис. 1. В этом представлении дерево содержит девять уровней, образующих пирамиду, в которой каждый уровень $l = 0, 1, \ldots, 8$ содержит 2^l эллиптических примитивов.

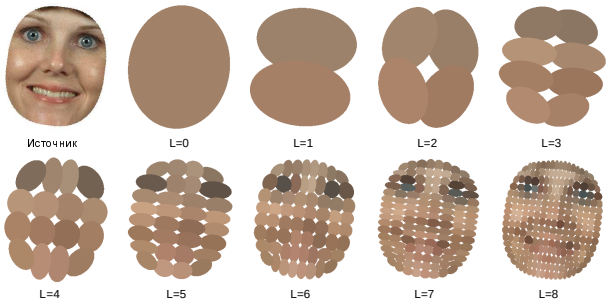


Рис. 1. Древовидное представление лица эллиптическими примитивами.

Формально древовидное представление любого объекта, выделенного на одиночном изображении, задано множеством эллиптических примитивов

$$A = \{a_n, 0 \leqslant n \leqslant n_{\max}\}, \qquad (1)$$

где a_n — примитив, который соответствует в дереве вершине с номером n, находящейся на уровне $l = \lfloor \log_2(n+1) \rfloor$. Нумерация вершин-примитивов производится в соответствии с рекуррентным правилом $n \to (2n+1, 2n+2)$, в котором n-я вершина текущего уровня $l \geqslant 0$ порождает две вершины следующего уровня $l{+}1$. На уровне $l = 0$ корневой примитив имеет номер $n = 0$. Каждый примитив в дереве (1) описывается набором параметров, который включает вектор центра эллипса, единичные векторы его главных осей, радиусы вдоль главных осей и среднее значение яркостной окраски примитива. Указанные параметры задаются в собственных координатах объекта и нормализуются относительно параметров корневого примитива.

Для объектов, заданных N-канальными изображениями, представления вида (1) могут быть построены по каждому каналу независимо. В результате для любого объекта, имеющего на всех N изображениях идентифицируемые системы собствен-

*Работа выполнена при финансовой поддержке РФФИ, проект № 09-01-00573-а.

ных координат, формируется многослойное представление в виде стопки деревьев

$$A^N = (A_1, \ldots, A_N), \tag{2}$$

в которой каждый слой A_i, $i = 1, \ldots, N$, представлен деревом примитивов вида (1). Примерами трёхслойных древовидных представлений являются стопки деревьев, которыми могут быть представлены цветные объекты, заданные RGB-изображениями. Слои в таких стопках являются независимыми древовидными представлениями объекта по R, G и B каналам.

Мера различия стопок деревьев

Для введения меры различия любой пары представлений $(A, \hat{A})$ вида (1) в работе [4] использовано пересечение деревьев $A \cap \hat{A}$, образованное всевозможными парами соответственных примитивов $(a_n \in A, \hat{a}_n \in \hat{A})$ с одинаковыми номерами n. Мера различия пары $(a_n, \hat{a}_n)$ введена на множестве

$$\Omega = \{(a_n, \hat{a}_n) \in (A \cap \hat{A})\},$$

в котором по крайней мере один из примитивов пары $(a_n, \hat{a}_n)$ является концевой вершиной соответствующего дерева. Указанная мера имеет вид

$$D(A, \hat{A}) = \sum_{k=1}^{3} w_k \sum_{(a_n, \hat{a}_n) \in \Omega} d_k(a_n, \hat{a}_n)\, 2^{-\lfloor \log_2(n+1) \rfloor}, \tag{3}$$

где w_1, w_2, w_3 — нормированная тройка параметров $(w_1 + w_2 + w_3 = 1)$, а функции $d_k(a_n, \hat{a}_n) \geqslant 0$ при $k = 1, 2, 3$ соответствуют различиям векторов центров $(k = 1)$, векторов ориентации и радиусов $(k = 2)$ и уровней яркости $(k = 3)$ примитивов a_n и $\hat{a}_n$.

Различие i-х слоёв $A_i \in A^N$ и $\hat{A}_i \in \hat{A}^N$, $i = 1, \ldots, N$, в любой паре стопок деревьев $(A, \hat{A}^N)$ вида (2) определяется функциями $D(A_i, \hat{A}_i)$ вида (3). Поэтому для пары многослойных представлений $(A^N, \hat{A}^N)$ вводится мера различия

$$D^N(A^N, \hat{A}^N) = \sum_{i=1}^{N} \alpha_i D(A_i, \hat{A}_i) \tag{4}$$

с нормализованными параметрами $\alpha_i : \sum_{i=1}^{N} \alpha_i = 1$. Коэффициенты w_k в (3) и коэффициенты α_i в (4) могут быть оценены на этапе обучения из условия минимизации вероятности ошибки распознавания. В частном случае, оценки коэффициентов α_i при $i = 1, \ldots, N$ могут быть получены с использованием нормализованных дисперсий яркостей:

$$\sigma_i^2 = \frac{\sigma^2(A_i)}{\sum_{i=1}^{N} \sigma^2(A_i)}, \tag{5a}$$

$$\hat{\sigma}_i^2 = \frac{\sigma^2(\hat{A}_i)}{\sum_{i=1}^{N} \sigma^2(\hat{A}_i)}, \tag{5b}$$

где $\sigma^2(A_i)$ и $\sigma^2(\hat{A}_i)$ — дисперсии яркостей пикселей объектов, имеющих в i-ых слоях представления $A_i \in A^N$ и $\hat{A}_i \in \hat{A}^N$. С учётом соотношений (5), $\alpha_i = (\sigma_i^2 + \hat{\sigma}_i^2)/2$, $i = 1, \ldots, N$. В случае RGB-изображений возможен выбор одинаковых коэффициентов $\alpha_i = 1/3$, $i = 1, 2, 3$.

Применение к распознаванию лиц по RGB-изображениям

Концепция многослойных древовидных представлений может найти применение для распознавания образов, заданных цветными изображениями, и, в частности, для распознавания лиц. В рамках настоящей работы выполнены эксперименты по выделению лиц на цветных изображениях, построению их представлений в виде стопок деревьев по RGB-каналам и распознаванию лиц в пространстве многослойных древовидных представлений по критерию ближайшего соседа, использующего меру различия вида (4). На этапе предобработки выполняется поиск координат глаз на изображении. Используя расстояние между центрами глаз r_0 строится система декартовых координат (U, V) с учётом симметрии лица относительно оси U и выбора оси V, параллельной линии глаз и смещённой влево вдоль оси U на величину $\beta_0 r_0$ с параметром $\beta_0 > 0$ (рис. 2).

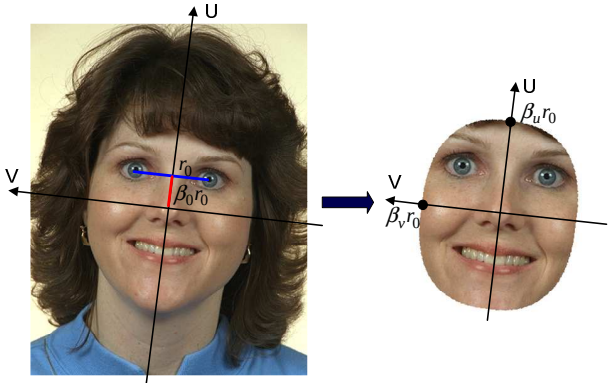


Рис. 2. Выделение информативной части лица на изображении.

В координатах (U, V) информативная часть лица задаётся областью, ограниченной овалоидом

$$\frac{U^m}{(\beta_u r_0)^m} + \frac{V^m}{(\beta_v r_0)^m} \leqslant 1 \tag{6}$$

с радиусами $r_u = \beta_u r_0$ и $r_v = \beta_v r_0$ по осям U и V, и параметром формы $m > 1$. Параметр смещения и параметры радиусов овалоида (6) выбирались равными: $\beta_0 = 0{,}5$, $\beta_u = 0{,}6$ и $\beta_v = 0{,}8$. Экспериментально найден наиболее подходящий параметр формы $m = 2{,}5$. Для объектов, соответствующих участкам лиц, ограниченных на RGB-изображениях фигурами (6), строились представления в виде трёхслойных стопок деревьев вида (2).

Эксперименты проведены с множеством цветных изображений, взятых из базы данных, размещенной на сайте [6]. Исходное множество изображений содержало 108 лиц (18 персон по 6 реализаций), которые были разбиты на две эквивалентные части, соответствующие обучающему и тестовому множествам, по три реализации каждой персоны в обоих множествах. В качестве эталонов использовались представления всех объектов обучающего множества. Распознавание проводилось для объектов тестового множества по критерию ближайшего эталона. Результаты в виде зависимостей доли ошибочных решений (в %) от числа уровней в используемых древовидных представлениях даны на рис. 3 для распознавания по полутоновым (а) и цветным (б) изображениям. В случае (а) использовалась мера вида (3), в случае (б) — мера вида (4) с параметрами: $w_1 = 0{,}1$; $w_2 = 0{,}2$; $w_3 = 0{,}7$. В мере (4) были взяты $\alpha_i = 1/3$ для каналов R, G и B $(i = 1, 2, 3)$.

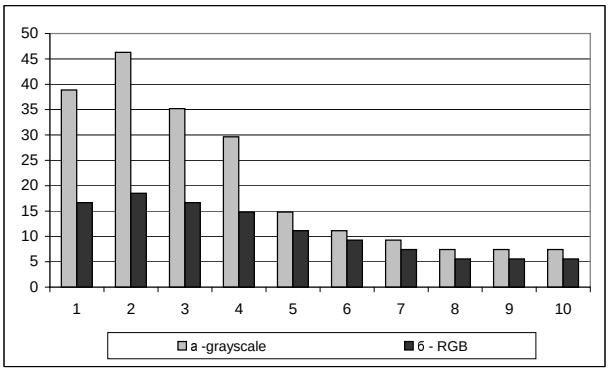


Рис. 3. Экспериментальные зависимости доли ошибок распознавания лиц от уровня древовидных представлений: а — полутоновые изображения; б — RGB-изображения.

Приведённые зависимости демонстрируют снижение доли ошибок при распознавании лиц в пространстве представлений RGB-объектов стопками деревьев по сравнению с пространством древовидных представлений полутоновых объектов. Дополнительное снижение доли ошибочных решений может быть достигнуто за счёт оптимизации процедуры отбора эталонов и введения их сфер влияния.

Выводы

Предложенный способ древовидного описания объектов, заданных многоканальными изображениями, обладает свойством универсальности, что позволяет использовать его для построения иерархически структурированных представлений образов, порождаемых различными источниками. В частности, возможно построение таких представлений для образов, заданных цветными изображениями в стандартах HSI или HSB. Возможно также расширение алфавита признаков примитивов с целью более адекватного описания текстуры образов. Введённая мера различия стопок деревьев допускает обобщение для многослойных пирамидальных представлений. Указанное обобщение может быть использовано для уменьшения вычислительной сложности поисковых процедур в пространстве представлений с многоуровневым разрешением.

Литература

[1] *Gonzalez R., Woods R.* Digital Image Processing. — New Jersey: Prentice Hall, 2002.

[2] *Berretti S., Del Bimbo A.* Multiresolution spatial partitioning for shape representation. // IEEE Proc. of ICPR, 2007. — Vol. 2. — Pp. 775–778.

[3] *Lange M., Ganebnykh S., Lange A.* Moment-based Pattern Representation Using Shape and Grayscale Features. // Lecture Notes in Computer Science — Berlin: Springer, 2007 — Vol. 4477.— Pp. 523–530.

[4] *Ганебных С. Н., Ланге М. М.* Древовидное представление образов для распознавания полутоновых объектов. // Труды Вычислительного центра им. А. А. Дородницына РАН (отдельный выпуск). — М.: ВЦ РАН, 2007. — 32 с.

[5] *Gallager R.* Information Theory and Reliable Communication. — New York: Wiley, 1968.

[6] `face.nist.gov/colorferet/`— The Color FERET Database, USA. — 2003.

Вычислительный алгоритм поиска на изображении прото-объекта

Левашкина А. О., Поршнев С. В.
iconismo@gmail.com
Екатеринбург, УГТУ-УПИ

Предложен алгоритм нахождения прото-объекта с использованием теории восходящего внимания. Описана методика оценки качества алгоритма. Проведено количественное сравнение WK–алгоритма и предложенного в работе алгоритма нахождения прото-объекта, результаты которого свидетельствуют о более высоком качестве результатов последнего.

На основании проведенных в [1] экспериментальных исследований степени субъективности внимания человека был сделан вывод о достаточно высоком уровне согласованности мнений независимых экспертов при выделении на изображениях областей, привлекающих их внимание (областей визуального внимания (ОВВ), прото-объектов, фокусов внимания). В соответствие с теорией визуального внимания (ТВВ) человек распознает изображение и принимает решение о степени похожести изображений именно на основе анализа свойств ОВВ, имеющих однозначную связь с содержанием изображения. В этой связи представляется целесообразным использовать подходы ТВВ в задаче автоматизированного поиска изображений по содержанию. При этом их практическая реализация определяет необходимость разработки алгоритмов автоматизированного поиска на изображениях прото-объектов.

Целью статьи является описание алгоритма нахождения прото-объекта на основе теории восходящего внимания и оценка его работоспособности.

Основные понятия теории визуального внимания

Внимание человека из зрительного поля отбирает сегменты информации, которая далее используется мозгом для более детальной переработки. Согласно [2], *внимание* — это избирательный процесс, позволяющий зрительной системе отделять релевантные внешние раздражители от нерелевантных. Известны два базовых подхода, используемых при объяснении механизмов внимания человека.

1. Подход, основанный на *восходящих процессах* (bottom-up image-based), базируется на том, что распределение внимания полностью определяться свойствами образа (например, неожиданное движение на периферии зрительного поля, отличие цвета образа от фона), при этом решение принимается без учета сознания человека. Зрительная система функционирует по принципу восходящего процесса, когда создание образа становится результатом объединения базовых элементов зрительной системы.

2. Подход, основанный на *нисходящих процессах* (top-down task-dependent), преимущественно базируются на знаниях, ранее полученных наблюдателем, его предшествующем опыте, осмыслении и интерпретации, а также на его ожиданиях.

Процессы, лежащие в основе внимания, могут быть составной частью как восходящих, так и нисходящих процессов. В литературе указывается, что определяющую роль во внимании играют восходящие процессы [2]. Однако сказать, что механизмы внимания как таковые основаны исключительно на восходящих процессах, значит упростить проблему.

В литературе отсутствует однозначный термин для обозначения области изображения, привлекающей внимание человека. В работах [2, 3] используется понятие фокуса внимания. Однако четкое определение этого термина отсутствует.

Далее в нашей работе под *фокусом внимания* будем понимать область изображения, содержащую его элементы, на которых концентрируется внимание человека в данный момент времени.

Отметим, что наряду с термином «фокус внимания» в научной литературе также используется термин «прото-объект» (proto-object) [3]. *Прото-объект* — это область изображения, привлекающая внимание человека и являющаяся составной частью объекта.

В отличие от фокуса внимания, который в большинстве случаев определяется в виде окружности разного радиуса, прото-объект имеет произвольную форму, которая является грубым приближением к наблюдаемому объекту (или объектам).

Известные модели восходящего внимания

Основой различных вычислительных моделей восходящего внимания служит теория интеграции отличительных признаков. Эта теория предполагает, что по всей визуальной сцене параллельно вычисляются только простые визуальные признаки. Внимание объединяет некоторые из вычисленных признаков в более сложное представление, связанное с нахождением объектов на изображении. На основе данного представления выделяется фокус внимания. В большинстве известных математических моделей восходящего внимания (KU-модель [4]; IKN-модель [8]; HKP-модель [9]) фокус внимания задается прямоугольником или

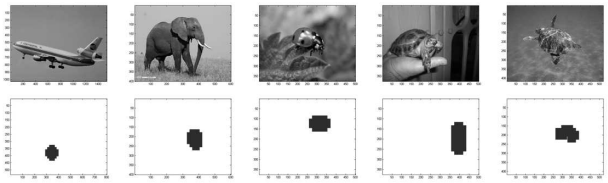


Рис. 1. Примеры прото-объектов, найденный с помощью WK-алгоритма.

окружностью, потому область расположения объекта удается задать только приблизительно.

В WK-модели [5], являющейся дальнейшим развитием IKN-модели, для обозначения области изображения, содержащей элементы, на которых концентрируется внимание человека, используется понятие «прото-объект». Проведенное исследование моделей восходящего внимания показало, что только в WK-модели ставится задача найти область, приблизительно соотвествующую объекту на изображении.

Пусть $I(x)^{\mathrm{RGB}}$ — исходное изображение в пространстве RGB, заданное на поле зрения X, (x — пиксель изображения, $x \subseteq X$).

WK-алгоритм реализуется следующей последовательностью действий:

1. Создание серии изображений с разным разрешением I_p^{RGB}, $p = 1, \ldots, P$, P — количество разрешений.
2. Вычисление карт для интенсивности, цвета и ориентации по каждому из изображений I_p^{RGB}:
   — карта интенсивности M_I^p:
   $M_I^p = \frac{(R^p+G^p+B^p)}{3}$
   — карта красного и зеленого цветов M_{RG}^p:

$$M_{RG}^p = \begin{cases} 0, \max(R^p, G^p, B^p) \leqslant \frac{1}{10}; \\ \frac{R^p - G^p}{\max(R^p, G^p, B^p)}. \end{cases}$$

   — карта синего и желтого цветов M_{BY}^p:

$$M_{BY}^p = \begin{cases} 0, \max(R^p, G^p, B^p) \leqslant \frac{1}{10}; \\ \frac{B^p - \min(R^p - G^p)}{\max(R^p, G^p, B^p)}. \end{cases}$$

   — карта ориентаций M_θ^p: к каждому изображению применяются фильтры для выделения горизонтального, вертикального и диагонального направлений.
3. Объединение карт M_θ^p, M_I^p, M_{RG}^p, M_{BY}^p в одну карту видимости SM (её разрешение совпадает с разрешением исходного изображения $I(x)^{\mathrm{RGB}}$).
4. Выделение области, соответствующей объекту — на карте SM это область с максимальным значением интенсивностей пикселей. Для её нахождения использован нейросетевой алгоритм WTA (Winner Takes All).

WK-алгоритм имеет ряд недостатков:

— находимый прото-объект является лишь грубым приближением к объекту, привлекающему внимание человека;
— оценка алгоритма по базе естественных изображений показала, что с его помощью удалось правильно найти прото-объекты только для 52% изображений.
— на основе визуального анализа найденных прото-объектов (рис. 1) можно сделать вывод, что области, соответствующие прото-объектам, невелики в сравнении с размером изображения, а их границы оказываются весьма грубыми (ступенчатыми).

Алгоритм нахождения прото-объекта

Для устранения недостатков математических моделей восходящего внимания, перечисленных в предыдущем разделе, нами разработан алгоритм нахождения прото-объекта (НПО-алгоритм) на цветных естественных изображениях.

При разработке алгоритма мы исходили из следующих представлений: определяющую роль во внимании играют восходящие процессы, согласно которым создание образа становится результатом объединения базовых элементов, обнаруженных зрительной системой. Очевидно, что область, привлекающая восходящее внимание («бросающаяся» в глаза), должна существенным образом отличаться от окружающих областей по ряду признаков изображения. На естественных изображениях присутствуют разнообразные текстуры, поэтому можно ожидать, что выделить фокус внимания удастся, проведя предварительную обработку изображения текстурным фильтром, и далее локализовать прото-объект, использовав процедуру сегментации изображения с помощью алгоритма JSEG [10]. Выбор данного алгоритма основан на результатах сравнения алгоритмов сегментации изображений [11] с помощью предварительно выбранных критериев для объективной оценки качества сегментации. При этом из всех областей изображения, выделенных в результате сегментации, к прото-объекту естественно относить область, попадающие в фокус внимания.

Пусть $I(x)$ — цветное изображение, заданное на поле зрения X (x — пиксель изображения, $x \subseteq X$).

НПО-алгоритм реализуется следующей последовательностью действий:

1. Создание эскиза изображения $\Psi(\chi)$. Под *эскизом* понимается уменьшенная цветная копия исходного изображения до размера 200 пикселей по большей стороне изображения. Изображение Ψ задано на поле зрения ξ (χ — пиксель изображения, $\chi \subseteq \xi$).
2. Преобразование эскиза в пространство LAB: $\Psi(\chi) \to \Psi(\chi)^{\mathrm{LAB}}$.

3. Нахождение фокуса внимания на каждой компоненте LAB:
   — Сглаживание усредняющим фильтром, область усреднения по умолчанию равна 3×3.
   — Применение разностного текстурного фильтра Local range of image [12]. Значения каждого пикселя χ определяется разностью максимального и минимального значения пикселей в окрестности 3×3 этого пиксела. Получаем изображение $\Psi(\chi)^{\mathrm{t}}$.
   — Пороговая обработка:

$$\Psi(\chi)^{\mathrm{thr}} = \begin{cases} 0, & \text{если } \Psi(\chi)^{\mathrm{t}} < R \max\limits_{i,j} \Psi(\chi)^{\mathrm{t}}; \\ \Psi(\chi)^{\mathrm{t}}, & \text{иначе.} \end{cases}$$

   где $R = 0{,}75$ — значение, определяемое экспериментально.
   — Преобразование Ψ^{thr} в бинарное изображение Ψ^{bin}.
   — Удаление статистически незначимых пикселей χ, идентифицируемых как выбросы. Анализ выбросов выполняется отдельно по осям X и Y. Статистическая значимость выбросов оценивается с помощью критерия Романовского [13]. Пиксели, признанные выбросами хотя бы по одной оси, удаляются.
   — Вычисление вершин прямоугольника, ограничивающего полученное множество белых пикселей. Полученная прямоугольная область характеризует фокус внимания. Изображение F является бинарной маской фокуса внимания, на которой значение пикселя равно 1, если пиксель принадлежит найденной прямоугольной области, иначе 0.

   Результатом выполнения п. 3 является три найденных фокуса внимания: $F(\chi)^L$, $F(\chi)^A$, $F(\chi)^B$.
4. Вычисление обобщенного фокуса внимания F: $F(\chi) = F(\chi)^L + F(\chi)^A + F(\chi)^B$.
5. Удаление пикселов доминантного цвета. Объект на изображении отличается от фона в том числе и цветом (причем фон занимает значительную часть изображения), поэтому целесообразно не относить пикселы доминантного цвета к объекту:

$$\Phi(\chi) = \begin{cases} 0, & \text{если } C_{\min} \leqslant \mathrm{Ind}(\chi) \leqslant C_{\max}; \\ F(\chi), & \text{иначе.} \end{cases}$$

   где $\mathrm{Ind}(\chi)$ — индексированное изображение, $C_{\min}$, $C_{\max}$ — границы диапазона доминантного цвета, вычисляемого следующим образом:
   — преобразование $\Psi(\chi)$ в индексированное изображение $\mathrm{Ind}(\chi)$;
   — построение гистограммы для $\mathrm{Ind}(\chi)$;
   — определение интервала, которому соответствует максимум гистограммы $[C_{\min}, C_{\max}]$;
6. Сегментация изображения $\Psi(\chi)$ алгоритмом JSEG [10]. Намеренно был выбран режим пересегментации (параметр $-l$ устанавливался равным 10).
7. Выбор среди результатов сегментации областей, принадлежащих прото-объекту. Область принадлежит объекту, если в ней присутствует не более 30% пикселов $\chi = 0$ на изображении $\Phi(\chi)$. Найденные области объединяются и образуют прото-объект.
8. Увеличение бинарной маски прото-объекта до размеров исходного изображения $I(x)$.

Иллюстрация получаемых результатов на каждом из этапов выполнения НПО-алгоритма представлена на рис. 2.

Необходимо отметить, что разработанный алгоритм нахождения прото-объекта применим к цветным изображениям, которые не содержат сложных сцен с большим количеством объектов.

Сравнение WK- и НПО-алгоритмов

Для колличественной оценки качества автоматического нахождения прото-объекта были использованы следующие критерии:

— полнота — доля пикселей прото-объекта, входящих в состав объекта, от общего числа пикселей, принадлежащих объекту (который обозначен экспертом).
— точность — доля пикселей прото-объекта, входящих в состав объекта (который обозначен экспертом), от общего числа пикселей, принадлежащих прото-объекту.

Эксперимент №1. Исследование влияния положения объекта на результативность его локализации (на естественных изображениях). Тестирование выполнялось на коллекции из 50 естественных цветных изображениях. На каждом изображении присутствует один хорошо выраженный объект, расположенный не по центру изображения. Значения критериев представлены в таблице 1.

Эксперимент №2. Исследование влияния небольших размеров объекта на результативность

Таблица 1. Результаты сравнения НПО- и WK- алгоритмов (эксперимент №1).

Критерий	НПО-алгоритм	WK-алгоритм
полнота	0.65 ± 0.04	0.47 ± 0.04
точность	0.70 ± 0.04	0.58 ± 0.03

Таблица 2. Результаты сравнения НПО- и WK- алгоритмов (эксперимент №2).

Критерий	НПО-алгоритм	WK-алгоритм
полнота	0.70 ± 0.04	0.53 ± 0.04
точность	0.73 ± 0.04	0.45 ± 0.03

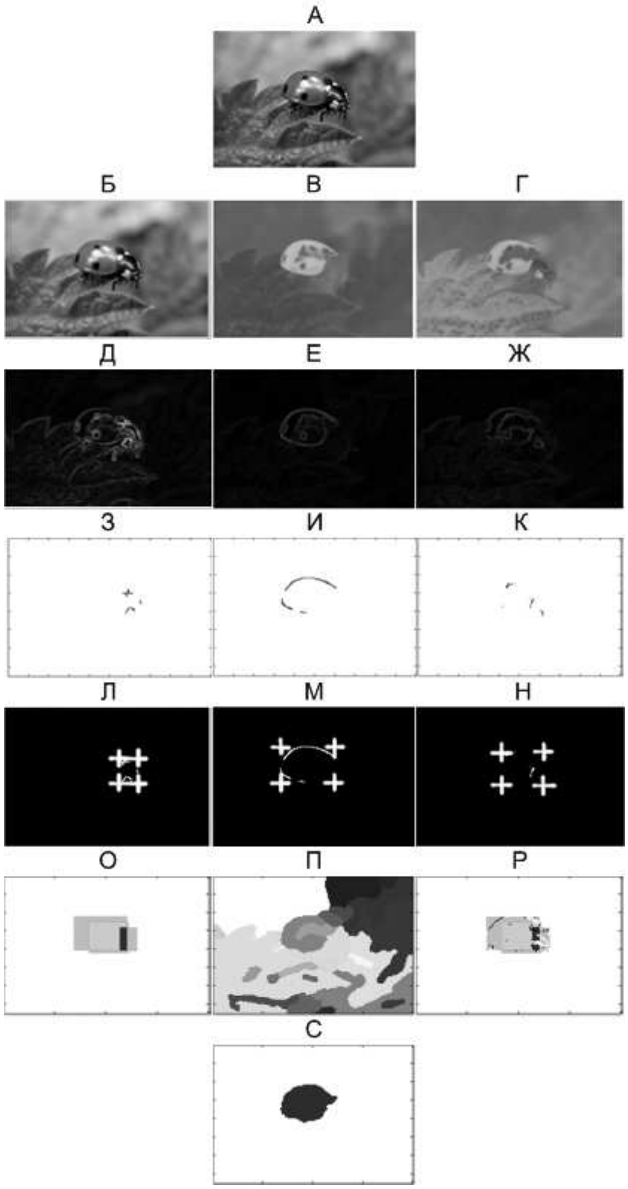


Рис. 2. (А) Исходное цветное изображение, (Б, В, Г) — результат применения усредняющего фильтра к каждой компоненте L, A, B (соотвественно) изображения, (Д, Е, Ж) — результат применения фильтра Local range of image к каждой компоненте, (З, И, К) — результат пороговой обработки и удаления выбросов, (Л, М, Н) — найденные $F(\chi)^L$, $F(\chi)^A$, $F(\chi)^B$, (О) — обобщенный фокус внимания F, (П) — области изображения, выделенные в результате JSEG-сегментации, (Р) — обобщенный фокус внимания после удаление пикселов, соответствующих доминантному цвету, (С) — найденный прото-объект.

его локализации (на естественных изображениях). Тестирование выполнялось на коллекции из 72 естественных изображений, на которых присутствует один выраженный небольшой объект, расположенный не по центру изображения. В процентах от площади всего изображения средняя площадь объекта на изображении составляет 6.2 ± 0.3. Значения критериев представлены в таблице 2.

Из таблиц 1 и 2 видно, что значения всех критериев для НПО-алгоритма выше по сравнению с аналогичными критериями для WK-алгоритма.

Выводы

Предложен вычислительный алгоритм нахождения на изображении прото-объекта, реализующий подходы теории восходящего внимания. Проведено количественное сравнение WK-алгоритма и предложенного НПО-алгоритма, результаты которого свидетельствуют о более высоком качестве работы последнего.

Литература

[1] *Liu T., Sun J., Zheng N., Tang T., Shum H.* Learning to Detect A Salient Object // Computer Vision and Pattern Recognition. — 2007. — Pp. 1–8.

[2] *Шифман Х. Р.* Ощущение и восприятие. — СПб, 2003. — 928 с.

[3] *Moran J., Desimone R.* Selective attention gates visual processing in the extrastriate cortex // Science. — 1985. — No. 229. — Pp. 782–784.

[4] *Koch C., Ullman S.* Shifts in selective visual attention: towards the underlying neural circuitry // Human Neurobiol. — 1985. — No. 4. — Pp. 219–227.

[5] *Walther D.* Interactions of visual attention and object recognition: computational modeling, algorithms, and psychophysics. — PhD thesis, California Institute of Technology. — Pasadena, CA, 2006.

[6] *Walther B., Koch Ch.* Modeling attention to salient proto-objects // Neural Networks. — 2006. — № 19. — С. 1395–1407.

[7] `http://fotki.yandex.ru/` — Яндекс-фотки. Фотохостинг компании «Яндекс». — 2009.

[8] *Itti L., Koch C., Niebur E.* A Model of Saliency-Based Visual Attention for Rapid Scene Analysis // IEEE Transactions on Pattern Analysis and Machine Intelligence. — 1998. — Т. 20, No. 11. — Pp. 1254–1259.

[9] *Harel J., Koch C., Perona P.* Graph-based visual saliency // Proceedings of Neural Information Processing Systems (NIPS). — 2006. — № 3. — С. 23–56.

[10] `http://vision.ece.ucsb.edu/segmentation/jseg/` — JSEG (Segmentation of color-texture regions in images and video). — 2009.

[11] *Левашкина А.О., Поршнев С.В.* Сравнительный анализ супервизорных критериев оценки качества сегментации изображений // Информационные технологии — 2009, No. 5. — Pp. 52–57.

[12] `http://www.mathworks.com/access/helpdesk/help/toolbox/images/index.html?/access/helpdesk/help/toolbox/images/rangefilt.html` — rangefilt (Local range of image). — 2009.

[13] *Третьяк Л. Н.* Обработка результатов наблюдений. — СПб, 2004. — 171 с.

Сравнительный анализ особенностей CBIR-систем

Левашкина А. О., Поршнев С. В.
iconismo@gmail.com
Екатеринбург, УГТУ-УПИ

Изложены результаты сравнительного анализа современных CBIR-систем, полученные на основе собственных исследований, в которых авторы использовали демо-версии следующих CBIR-систем: Img(Anaktisi), MFIRS, CIRES, Tiltomo, INRIA, Retrievr, Alipr, SIMPLIcity, Viper, FS, находящихся в свободном доступе.

В настоящее время задачи оцифровки и хранения больших объемов визуальной информации имеют законченные технические решения, которые вполне соответствуют требованиям пользователей. При существующей необходимости организации доступа к коллекции изображений посредством поиска по текстовой информации, ассоциированной с изображениями, данный подход имеет очевидный недостаток, обусловленный неоднозначностью установления соответствия между визуальным содержанием изображений и его текстовым описанием. В начале 80-х годов для преодоления недостатков поисковых систем на основе текста были начаты разработки методов поиска изображений по содержанию (в зарубежной литературе для обозначения данного подхода используется аббревиатура CBIR – Content-based image retrieval). Известны результаты проведенного в [1] анализа 56 зарубежных CBIR-систем, созданных к началу 2002 г. Однако к настоящему моменту большая часть CBIR-систем, описанная в работе [1], не поддерживается.

Целью статьи является проведение сравнительного анализа демо-версий CBIR-систем, известных к настоящему времени: Img(Anaktisi) [2], MFIRS [3], CIRES [4], Tiltomo [5], INRIA [6], Retrievr [7], Alipr [8], SIMPLIcity [9], Viper [10], FS [11], и находящихся в свободном доступе. Сравнение проводится по следующим критериям:

1) тип изображений в базе поиска;
2) форма запроса к системе поиска;
3) признаки изображений, используемые при поиске;
4) средний процент верно найденных изображений.

Необходимо отметить, что при сравнительном анализе качества поиска CBIR-систем возникает проблема, обусловленная отсутствием возможности использования единой коллекции изображений, поскольку оказалось, что каждая из перечисленных выше CBIR-систем ориентирована на поиск исключительно в собственных базах, содержащих различные по тематике изображения (фотографии, скриншоты телевизионных программ, картины, текстурные образцы, изображения автомобилей, биомедицинские изображения) и др. В этих условиях представляется естественным осуществлять поиск цветных естественных изображений (фотографий) общей тематики в каждой из баз изображений, с которыми работает данная CBIR-систем, и далее вычислять процент верно найденных изображений. Далее проводится сравнительный анализ перечисленных выше критериев сравнения для каждой из рассматриваемых CBIR-систем, представленных в таблице 1.

Таблица 1. Коллекции изображений в CBIR-системах.

№	Название системы	Количество коллекций	Выбранная коллекция
1	MFIRS	4	RGB 1
2	CIRES	1	фотографии
3	Tiltomo	2	«Catchy Colors»
4	INRIA	4	по всем 4м
5	Retrievr(эскиз)	1	Flickr
6	Retrievr(образец)	1	Flickr
7	Alipr	1	фотографии
8	SIMPLIcity	3	фотографии
9	Viper	1	Corel
10	FS	1	Flickr
11	img(Anaktisi)	9	Wang

Тип изображений в базе поиска и форма запроса к системе

Из таблицы 1 видно, что в системах MFIRS, CIRES, Tiltomo, INRIA, Retrievr, Alipr, SIMPLIcity, Viper, FS, Img(Anaktisi)[1] поиск проводится по нескольким коллекциям изображений.

Коллекции естественных изображений или фотографий, использованные для проведения сравнительного анализа в каждой из рассмотренных CBIR-систем, представлены в таблице 1. Примеры изображений одной из использованных в работе коллекций (Flickr) представлены на Рис. 1.

Рис. 1. Примеры изображений из коллекции Flickr.

[1] Особенность коллекций Img(Anaktisi) состоит в том, что в них содержится большое количество дубликатов изображений, что, несомненно, облегчает поиск. Если в качестве запроса задается один из дубликатов, то количество верно найденных изображений велико. Однако, если задать изображение без дубликатов, качество поиска резко снижается.

Таблица 2. Типы-запросов к CBIR-системам.

№	Тип запроса			Система
	Эскиз	Из базы	Польз.	
1	–	+	–	MFIRS
2	–	+	–	CIRES
3	–	+	–	Tiltomo
4	–	+	+	INRIA
5	+	+	+	Retrievr
6	–	+	+	Alipr
7	–	+	+	SIMPLIcity
8	–	+	–	Viper
9	–	+	–	FS
10	–	+	–	img(Anaktisi)

Анализ перечисленных выше CBIR-систем показал, что в них используются следующие способы формирования поисковых запросов:

1) запрос по эскизу (Рис. 2);
2) запрос по образцу, выбираемому из имеющихся в базе изображений;
3) запрос по образцу, заданному пользователем.

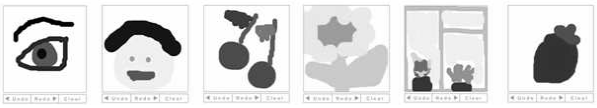


Рис. 2. Примеры запросов эскизов в системе Retrievr

Возможные типы запросов, которые могут быть сформулированы пользователем в каждой из рассматриваемых систем представлены в таблице 2. Из таблицы 2 видно, что в большинстве CBIR-систем запрос задается в виде изображения образца, которое выбирается из имеющихся в базе изображений.

Необходимо отметить, что запрос по эскизу является наиболее неудобным видом запроса. Во-первых, система формирования запроса не имеет достаточного количества инструментов для его реализации. Во-вторых, далеко не все пользователи оказываются способны выразить свой запрос в виде эскиза в форме, пригодной для поиска.

Запрос по изображению-образцу более удобен по сравнению с запросом по эскизу, однако он также имеет ряд недостатков. При задании в качестве запроса целого изображения без указания интересующей пользователя области во многих случаях CBIR-система находит изображения, которые в целом похожи на искомое изображения, однако, зачастую не содержат искомой информации (например, отсутствует искомый объект). Кроме того, возможность задания в качестве запроса одного из изображений, имеющихся в базе изображений, определяет необходимость создания подсистемы быстрого нахождения изображения-запроса, удобной для пользователя. Среди рассмотренных демо-версий подобная подсистема имеется только в CIRES, однако, одновременно в данной подсистеме задано ограниченное число понятий.

В CBIR-системе Eakins можно выделить следующие уровни запросов к CBIR-системам:

— уровень 1: поиск на основе использования некоторого набора признаков изображения (вообще говоря, примитивных);
— уровень 2: поиск на основе одновременного использования признаков изображения и некоторых логических выражений;
— уровень 3: поиск на основе использования абстрактных атрибутов, в том числе высокоуровневых рассуждений о содержании изображений.

Уровни 2 и 3 называются семантическим поиском изображений, а пробел между уровнями 1 и 2, обусловленный различием между ограниченной описательной мощностью низкоуровневых признаков изображений и богатством семантики пользователя называется *семантическим разрывом* [13].

Признаки изображений, используемые при поиске

В ходе проведенного анализа было обнаружено, что признаки, на основе которых выполняется поиск изображений, известны не для всех систем поиска. В доступных описаниях ряда CBIR-систем либо сообщаются только названия признаков, без указания, характеристикой чего они являются, либо указывается, что используются признаки цвета, текстуры, формы и/или пространственного расположения.

Результаты анализа частоты использования признаков в CBIR-системах, показаны в таблице 3 и на рис. 3 (Обозначения: F1 — цвет, F2 — текстура, F3 — форма, F4 — пространственное расположение, ? — информация о признаках отсутствует).

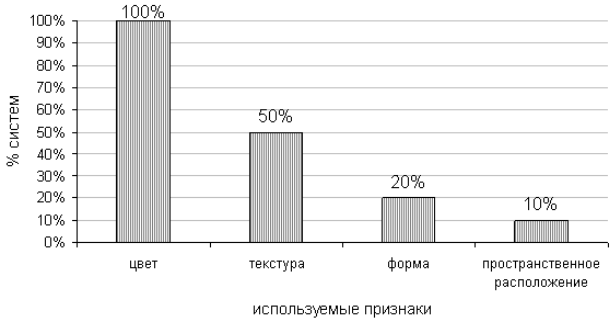


Рис. 3. Признаки, используемые в CBIR-системах.

На сайтах демо-версий систем Alipr, Viper, FS отсутствует информация об используемых признаках. Для системы MFIRS известны только названия признаков. На основании опыта работы с дан-

Таблица 3. Признаки, используемые в CBIR-системах.

№	Система	?	F1	F2	F3	F4
1	MFIRS	×	+			
2	CIRES		+	+		
3	Tiltomo		+	+		
4	INRIA		+	+	+	
5	Retrievr		+			
6	Alipr	×	+			
7	SIMPLIcity		+	+	+	+
8	Viper	×	+			
9	FS	×	+			
10	img(Anaktisi)		+	+		

ными системами можно сделать вывод, что при поиске изображений в наибольшей степени учитываются признаки цвета. Никаких свидетельств того, что используются признаки текстуры, формы и пространственного расположения авторам обнаружить не удалось, поэтому в таблице 3 для этих четырех систем отмечено использование только признаков цвета. Из рис. 3 видно, что наиболее распространенными признаками в системах поиска изображений являются признаки цвета (100%), наименее используемыми — признаки формы (20%) и пространственного расположения (10%).

В системах Img(Anaktisi), MFIRS, INRIA пользователь имеет возможность выбрать из имеющегося набора признак, на основе которого проводится поиск. Возможностью одновременного выбора нескольких признаков обладает только CBIR-система INRIA.

Эмпирическое сравнение CBIR-систем

В информационном поиске часто используется критерий *точности поиска* (precision) — доля истинно релевантных документов в общем числе найденных, а также *точность на уровне n документов* — количество релевантных документов среди первых n выданных документов, деленное на n. Поскольку для большинства рассматриваемых CBIR-систем на первой странице результатов поиска содержится не менее 20 изображений, для оценки качества поиска и информативности первой страницы результатов был выбран критерий точности на уровне первых $n = 20$ найденных изображений. Использование только точности поиска при оценке качества работы системы представляется недостаточным, так как остается неясным, какая доля релевантных документов не была найдена (критерий «полнота», recall). Однако в перечисленных системах поиск ведется по коллекциям, в которых неизвестно количество имеющихся изображений, релевантных тому или иному запросу. Пэтому в указанных условиях в качестве критерия использовалась «точность на уровне n документов».

Для сравнения систем поиска изображений была использована методика, реализующаяся следующей последовательностью действий:

1) выбор CBIR-системы;
2) выбор базы естественных изображений, по которой будет вестись поиск (если такая возможность имеется);
3) выбор запроса к CBIR-системе;
4) вычисление точности на уровне 20 первых найденных изображений;
5) визуальная оценка соответствия запроса и каждого найденного изображения;
6) повторение 50 раз пп. 3–5 для получения статистически значимых результатов;
7) вычисление средней точности на уровне 20.

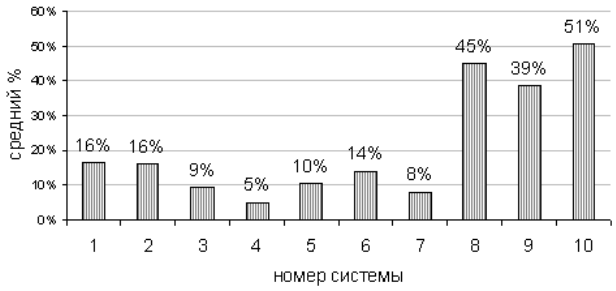


Рис. 4. Результаты сравнения систем поиска изображений.

Данная методика была применена к следующим CBIR-системам: Img(Anaktisi), MFIRS, CIRES, Tiltomo, Retrievr, SIMPLIcity, FS, а также к системам поиска изображений на основе текстовой информации: поиск картинок на Yandex, поиск картинок на Altavista. В том случае, когда система предоставляла пользователю дополнительные настройки, использовались настройки по умолчанию. Результаты сравнения результатов поиска выбранных GBIR-систем представлены в таблице 4 и на рис. 4. Номер на рисунке соответствует номеру системы в таблице 4 (среди систем, перечисленных в таблице 4, отсутствуют INRIA, Alipr, Viper, поскольку в результатах поиска этими системами содержится меньше двадцати изображений, что не позволяет вычислить точность на уровне первых 20 найденных изображений).

Из рис. 4 видно, что:

— для большей части CBIR-систем точность на уровне первых 20-ти найденных изображений не превышает 16%;
— среди CBIR-систем наилучший результат показала система Img(Anaktisi), однако, столь высокий результат поиска сложно считать объективным, поскольку в базе изображений, по которой ведется поиск этой системой, содержится большое количество дубликатов изображений;

Таблица 4. Результаты сравнения систем поиска изображений.

№	Название системы	Точность на уровне 20 (%)
1	MFIRS	16.3 ± 2.7
2	CIRES	16.1 ± 2.5
3	Tiltomo	9.4 ± 2.4
4	Retrievr(эскиз)	4.9 ± 1.2
5	Retrievr(образец)	10.2 ± 2.3
6	SIMPLIcity	13.9 ± 2.7
7	FS	7.4 ± 1.3
8	img(Anaktisi)	45.0 ± 5.5
9	Altavista	38.8 ± 3.4
10	Yandex	50.7 ± 3.6

— CBIR-системы MFIRS и CIRES показали наилучшие результаты среди всех рассмотренных систем, не считая Img(Anaktisi); отметим, что в системе CIRES поиск ведется на основе признаков цвета и текстуры, а для системы MFIRS информация об используемых признаках отсутствует;
— системы поиска изображений на основе текстовой информации (Altavista, Yandex) показали результат выше, чем CBIR-системы;
— среди всех рассмотренных систем самый лучший результат показала система поиска картинок поисковой системы Yandex.

В системах INRIA, Alipr, Viper в результатах поиска содержится меньше двадцати изображений:

— система INRIA: 5 изображений в результатах поиска (точность поиска на уровне 5 найденных изображений составила $10 \pm 3\%$);
— система Alipr: 11 изображений в результатах поиска (точность поиска на уровне 11 найденных изображений составила $12.2 \pm 2.9\%$);
— система Viper: 9 изображений в результатах поиска (точность поиска на уровне 9 найденных изображений составила $15 \pm 3\%$).

Выводы

Проведенный анализ современного состояния современных CBIR-систем, позволяет выделить ряд проблем, возникающих на этапах их разработки и эксплуатации.

1. Проблема формирования запроса: типичным видом запроса является изображение-образец, что менее удобно для пользователя, чем текстовый запрос.
2. Проблема обоснованного выбора признаков изображения и методов их комбинирования, возникающая при поиске в коллекциях разнородного содержания, изображения которых не принадлежат к одной определённой тематике.
3. Проблема ликвидации «семантического разрыва»: наличие указанных проблем приводит к тому, что качество поиска в современных CBIR-систем оказывается недостаточно высоким (для большинства рассмотренных CBIR-систем точность на уровне 20 первых найденных изображений оказалась меньше 16%, в то время как при поиске по текстовым аннотациям в системе Yandex аналогичный показатель оказался равным 51%).

Таким образом, несмотря на достаточно длительный срок, прошедший с начала исследований методов поиска изображений по содержанию и наличие большого числа демо-версий CBIR-систем, разработка новых подходов к решению рассматриваемой задачи остается актуальным.

Литература

[1] Veltkamp R. C., Tanase M. Content based image retrieval systems: A survey — 2002. — http://www.aa-lab.cs.uu.nl/cbirsurvey/

[2] Демо-версия CBIR системы Img(Anaktisi) — 2009 — http://orpheus.ee.duth.gr/anaktisi/

[3] Демо-версия CBIR системы MFIRS — 2009. — http://www.pilevar.com/mfirs/index.php

[4] Демо-версия CBIR системы CIRES — 2009. — http://amazon.ece.utexas.edu/~qasim/cires.htm

[5] Демо-версия CBIR системы Tiltomo — 2009. — http://www.tiltomo.com/

[6] Демо-версия CBIR системы INRIA — 2009. — http://www-rocq.inria.fr/cgi-bin/imedia/circario.cgi/v2std

[7] Демо-версия CBIR системы Retrievr — 2009. — http://labs.systemone.at/retrievr/

[8] Демо-версия CBIR-системы Alipr — 2009. — http://alipr.com/

[9] Демо-версия CBIR-системы SIMPLIcity — 2009. — http://wang14.ist.psu.edu/cgi-bin/zwang/regionsearch_show.cgi

[10] Демо-версия CBIR-системы Viper — 2009. — http://viper.unige.ch/demo/php/demo.php

[11] Демо-версия CBIR-системы FS — 2009. — http://139.82.71.62:8080/fs26/#About

[12] Flickr — online photo management and sharing application — 2009. — http://www.flickr.com/

[13] *Liu Y., Zhang D., Lu G., Ma W.-Y.* A survey of content based image retrieval with high level semantics // Pattern Recognition. — 2007. — No. 40. — Pp. 262–282.

Оценка кривизны методом усреднения локально-интерполяционных оценок*

Лепский А. Е.
lepskiy@mail.ru
Таганрог, Технологический институт Южного федерального университета

В статье рассматривается метод взвешенного усреднения первичных локально-интерполяционных оценок кривизны плоской дискретной вероятностно зашумленной кривой. Исследуются качественные характеристики такой оценки — систематическая и случайная ошибки. Решена задача нахождения оптимальных значений весовых коэффициентов усреднения, при которых минимизируется среднеквадратичная ошибка вычисления оценки кривизны. Предложены алгоритмические процедуры, оптимально корректирующие размер «окна» усреднения.

Введение

Распознавание плоских контурных изображений объектов осуществляется, как правило, путем сравнения контура объекта с эталонными контурами, хранимыми в том или ином виде. Число пикселей контура может достигать нескольких тысяч, причем большинство из них не несет информативной нагрузки. Поэтому стараются выделить небольшое число наиболее информативных точек контура. Такие точки обычно называют контрольными. Контрольные точки контура выделяются путем оценивания кривизны кривой и нахождения локальных максимумов таких оценок. Кривизна — это локальная дифференциальная характеристика кривой, которая показывает насколько данная кривая в точке отличается от прямой. Если Γ — гладкая кривая, то ее кривизна $k(\mathbf{g})$ в точке $\mathbf{g} \in \Gamma$ может быть определена как производная $\theta'_s(\mathbf{g})$ функции наклона $\theta(\mathbf{g})$ (угол между касательной и выбранной осью, например, положительным направлением оси Ox) по длине дуги s.

Реально имеется только дискретная кривая Γ, выделенная тем или иным методом на оцифрованном изображении. Поэтому вместо вычисления кривизны $k(\mathbf{g})$ вычисляют некоторую ее оценку. Простейшую оценку можно получить, если перейти в определении кривизны от производных к конечным разностям. Среди параметров оценки кривизны одним из важнейших является положительный параметр (будем обозначать его через ε), характеризующий величину окрестности $U_\varepsilon(\mathbf{g})$ с центром в точке $\mathbf{g}$, в пределах которой вычисляется оценка, либо в пределах которой наиболее значимы для вычисления оценки точки кривой. Зависимость оценки кривизны от параметра ε будем обозначать через $k_\varepsilon(\mathbf{g})$. Количество и тип других параметров оценки кривизны зависят от способа ее вычисления.

Любая ε-оценка кривизны $k_\varepsilon(\mathbf{g})$ оцифрованной кривой Γ в точке $\mathbf{g}$ характеризуется некоторыми качественными величинами.

Одной из важнейших качественных величин является разность $s_\varepsilon = |k_\varepsilon - k|$, называемая *систематической ошибкой* и характеризующая величину отклонения оценки от точного значения. Систематическая ошибка обусловлена с одной стороны неточностью дискретизации и квантования изображения, а с другой стороны, неточностью метода вычисления оценки. Если рассматривать зависимость систематической ошибки от величины ε окрестности вычисления оценки, то потребуем, чтобы выполнялось условие:

1) для гладкой кривой $\lim\limits_{\varepsilon\to+0} s_\varepsilon(\mathbf{g}) = 0$.

Это свойство означает сходимость оценки кривизны к точному значению.

Кроме того, как правило, изображение, на котором выделяется кривая, является зашумленным. Характер неточности зашумления может быть разным. Ниже будем считать, что неточность имеет аддитивный вероятностный характер. Тогда оценка кривизны будет случайной величиной $K_\varepsilon(\mathbf{g})$, которая качественно характеризуется величиной *смещения* $b_\varepsilon = |\mathsf{E}[K_\varepsilon] - k_\varepsilon|$, где $\mathsf{E}[\cdot]$ — оператор математического ожидания, и величиной *случайной ошибки* (дисперсии) $\sigma^2[K_\varepsilon]$. Потребуем, чтобы для случайной оценки кривизны $K_\varepsilon(\mathbf{g})$ также выполнялись условия:

2) $\lim\limits_{\varepsilon\to\infty} |\mathsf{E}[K_\varepsilon] - k_\varepsilon| = 0$,

3) $\lim\limits_{\varepsilon\to\infty} \sigma^2[K_\varepsilon] = 0$,

Назовём их *условиями устойчивости* оценки кривизны к зашумлению изображения. Если $\sigma^2[K_\varepsilon] = O(\varepsilon^{-\alpha})$, $|\mathsf{E}[K_\varepsilon] - k_\varepsilon| = O(\varepsilon^{-\beta})$, то величина $\min\{\alpha/2, \beta\}$ определяет *порядок устойчивости* оценки кривизны.

Все методы оценивания кривизны и так называемые детекторы углов (в настоящее время известно около 100 таких алгоритмов) реализуют, как правило, два основных подхода к вычислению оценок функции кривизны. В первом подходе сначала с помощью разностного оператора кривизны находят первичные оценки в некоторых точках кривой из ε-окрестности с центром в исследуемой точке $\mathbf{g}$. Затем осуществляется усреднение (сглаживание) полученных первичных оценок. Во вто-

*Работа выполнена при финансовой поддержке РФФИ, проекты № 07-07-00067, № 08-07-00129.

ром подходе осуществляется гладкая аппроксимация дискретной кривой Γ, после чего вычисляется кривизна гладкой аппроксимирующей кривой в исследуемой точке. В работе [6] был проанализирован с точки зрения точности и устойчивости один модельный метод (так называемый *метод геометрического сглаживания*) второго подхода, реализующий неявную аппроксимацию дискретной кривой. Этот метод является бинарным аналогом широко известного детектора Харриса [4]. В работе [7] исследовалась устойчивость к зашумлению некоторых представлений дискретной кривой, полученных в результате агрегирования оценок кривизны, вычисленных методом геометрического сглаживания. В [5] было проведено сравнение с точки зрения точности и устойчивости вычисление оценок кривизны с помощью явной и неявной схем аппроксимации.

Вычисление оценок кривизны только с помощью разностных операторов практические не применяется, поскольку такая оценка будет очень чувствительной к значениям отдельных данных. Для того, чтобы уменьшить эту чувствительность, осуществляют усреднение (сглаживание) полученных в пределах ε-окрестности первичных оценок кривизны. Из всего многообразия процедур усреднения (сглаживания) можно выделить два основных способа. В первом случае вычисляются первичные оценки кривизны в одной точке $\mathbf{g}$, но с разными шагами интерполяции в пределах ε-окрестности. После чего осуществляется усреднение полученных оценок. На такой процедуре усреднения основаны популярные алгоритмы Freeman и Davis [3], Beus и Tiu [1] и другие. В другом способе сглаживаются первичные оценки кривизны, вычисленные в разных точках ε-окрестности. Такую схему сглаживания можно реализовать с помощью интегрального оператора свертки разностных оценок кривизны и некоторого сглаживающего ядра. Применение оператора сглаживания для выделения низкоуровневых признаков на изображении впервые было применено Canny [2] для построения детектора края. Позднее методика Canny была применена для вычисления оценок кривизны.

Для «хороших» оценок и правильно найденных значений параметров все три качественные характеристики-критерии (смещение, систематическая и случайная ошибки) должны быть небольшими. Задача нахождения значения параметра ε, при котором минимизируются несколько критериев, является многокритериальной задачей. Один из путей решения такой задачи — минимизировать «свертку» критериев, например, их сумму или среднеквадратичную ошибку $s_\varepsilon^2 + b_\varepsilon^2 + \sigma^2\,[K_\varepsilon]$.

Итак, для заданного метода вычисления оценки кривизны возникают следующие задачи:

1) найти значения (или оценки) трёх качественных характеристик оценок кривизны;
2) найти оптимальное значение параметра ε, минимизирующего среднеквадратичную ошибку.

Если для дискретной кривой параметр оценки ε характеризует величину окрестности оцифрованного изображения, в пределах которой вычисляется оценка, то будем считать его целочисленным и обозначать через m.

Усреднение первичных оценок кривизны

Предположим, что известны точечные значения $\{(s, y_s)\}_{s=-m}^{m}$ оцифрованной плоской кривой, заданные в «окне» $[-m, m]$. Рассмотрим следующую схему оценивания кривизны в точке $s = 0$.

1) Для всех $l = 1, \ldots, m$ построим интерполяционный многочлен $P_{l,2m}(x; \mathbf{y})$ второго порядка, проходящий через точки (s, y_s), $s = -l, 0, l$. Найдем оценки кривизны $k_{l,m}^{(1)}(\mathbf{y})$ в точке $s = 0$, как кривизны многочленов $P_{l,2m}(x; \mathbf{y})$:

$$k_{l,m}^{(1)}(\mathbf{y}) = \frac{P''_{l,2m}(0; \mathbf{y})}{\left(1 + \left(P'_{l,2m}(0; \mathbf{y})\right)^2\right)^{3/2}}.$$

2) Вычислим α-усредненную оценку кривизны $\tilde{k}_m^{(1)}(\mathbf{y}; \alpha) = \sum\limits_{l=1}^{m} \alpha_l k_{l,m}^{(1)}(\mathbf{y})$, $\alpha = \alpha_1, \ldots, \alpha_m$, где неотрицательные коэффициенты α_l должны удовлетворять условию $\sum\limits_{l=1}^{m} \alpha_l = 1$.

Интерполяционный многочлен $P_{l,2m}(x; \mathbf{y})$ второго порядка, проходящий через точки (s, y_s), $s = -l, 0, l$, можно записать в виде

$$P_{l,2m}(x; \mathbf{y}) = y_0 + \frac{\Delta y_l}{2l} x + \frac{\Delta^2 y_l}{2l^2} x^2,$$

где $\Delta y_l = y_l - y_{-l}$, $\Delta^2 y_l = y_l - 2y_0 + y_{-l}$. Тогда $P'_{l,2m}(0; \mathbf{y}) = \Delta y_l/(2l)$, $P''_{l,2m}(0; \mathbf{y}) = \Delta^2 y_l / l^2$ и

$$k_{l,m}^{(1)}(\mathbf{y}) = \frac{8l\Delta^2 y_l}{(4l^2 + (\Delta y_l)^2)^{3/2}}.$$

Теперь в качестве оценки кривизны в точке $x = 0$ можно взять линейный функционал от вектора оценок $\left(k_{l,m}^{(1)}(\mathbf{y})\right)_{l=1}^{m}$:

$$\tilde{k}_m^{(1)}(\mathbf{y}; \alpha) = \sum_{l=1}^{m} \alpha_l k_{l,m}^{(1)}(\mathbf{y}).$$

Для простоты исследуем следующую упрощенную оценку кривизны $\tilde{k}_m^{(2)}(\mathbf{y})$, полученную методом α-усреднения локально-интерполяционных оценок.

Пусть точечная функция $\{(s, y_s)\}_{s=-m}^{m}$ является четной, то есть $y_{-s} = y_s$ для всех $s = -m, \ldots, m$. Тогда $\Delta y_l = 0,\ l = 1, \ldots, m$. Поэтому

$$k_{l,m}^{(2)}(\mathbf{y}) = P''_{l,2m}(0; \mathbf{y}) = \Delta^2 y_l / l^2;$$
$$\tilde{k}_m^{(2)}(\mathbf{y}; \alpha) = \sum_{l=1}^{m} \alpha_l k_{l,m}^{(2)}(\mathbf{y}) = \sum_{l=1}^{m} \alpha_l \frac{\Delta^2 y_l}{l^2} = 2\sum_{l=1}^{m} \alpha_l \frac{y_l - y_0}{l^2}.$$

Систематическая ошибка усредненной оценки кривизны

Оценим систематическую ошибку оценки кривизны $\tilde{k}_m^{(2)}(\mathbf{y}; \alpha)$, если известны точечные значения кривой класса C^3, заданной явно четной функцией $y(x)$. Тогда точное значение кривизны $k = y''(0)$. Так как

$$y_l = y_0 + 0{,}5\, y''(0)\, l^2 + r_l,$$

где остаток $r_l = \frac{y'''(x_l)}{3!} l^3$, $x_l \in [0, l]$, то $\frac{2}{l^2}(y_l - y_0) = y''(0) + \frac{2}{l^2} r_l$, и для систематической ошибки оценки кривизны $\tilde{k}_m^{(2)}(\mathbf{y})$ имеем

$$s_m = \left|\tilde{k}_m^{(2)}(\mathbf{y}; \alpha) - k\right| = \left|2\sum_{l=1}^{m} \alpha_l \frac{y_l - y_0}{l^2} - y''(0)\right| = \left|\sum_{l=1}^{m} \alpha_l \left(y''(0) + \frac{2r_l}{l^2}\right) - y''(0)\right| = 2\left|\sum_{l=1}^{m} \alpha_l \frac{r_l}{l^2}\right| = \frac{1}{3}\left|\sum_{l=1}^{m} \alpha_l y'''(x_l) l\right|.$$

Поскольку $\min\limits_l y'''(x_l) \sum\limits_{l=1}^{m} \alpha_l l \leqslant \sum\limits_{l=1}^{m} \alpha_l y'''(x_l) l \leqslant \max\limits_l y'''(x_l) \sum\limits_{l=1}^{m} \alpha_l l$, существует $x^* \in [0, m]$ такой, что $\sum\limits_{l=1}^{m} \alpha_l y'''(x_l) l = y'''(x^*) \sum\limits_{l=1}^{m} \alpha_l l$. Следовательно,

$$s_m = s\big(\tilde{k}_m^{(2)}\big) = \frac{1}{3}\big|y'''(x^*)\big| \sum_{l=1}^{m} \alpha_l l. \qquad (1)$$

Из (1) следует, что систематическая ошибка не может быть сколь угодно уменьшена.

Случайная ошибка усредненной оценки кривизны

Предположим, что точечные значения $\mathbf{y}$ кривой подвергнуты аддитивному дискретному некоррелированному стационарному в широком смысле вероятностому зашумлению, получим случайный вектор $\mathbf{Y} = \mathbf{y} + \boldsymbol{\xi}$, $\mathsf{E}[\xi_i] = 0$, $\sigma^2[\xi_i] = \sigma^2$. Тогда оценка кривизны, полученная методом усреднения локально-интерполяционных оценок, будет случайной величиной $\tilde{K}_m^{(2)}$. Поскольку оценка $\tilde{k}_m^{(2)}(\mathbf{y})$ является линейной относительно $\mathbf{y}$, смещение $b\big(\tilde{K}_m^{(1)}\big) = \mathsf{E}\big[\tilde{K}_m^{(2)}\big] - \tilde{k}_m^{(2)} = 0$, а случайная ошибка

$$\sigma^2\big[\tilde{K}_m^{(2)}\big] = \sigma^2\left[2\sum_{l=1}^{m} \alpha_l \frac{Y_l - Y_0}{l^2}\right] = 4\sigma^2\left[\sum_{l=1}^{m} \alpha_l \frac{Y_l}{l^2} - Y_0 \sum_{l=1}^{m} \frac{\alpha_l}{l^2}\right] = 4\sigma^2\left[\sum_{l=1}^{m} \frac{\alpha_l^2}{l^4} + \left(\sum_{l=1}^{m} \frac{\alpha_l}{l^2}\right)^2\right]. \qquad (2)$$

В частности, при равномерном усреднении, то есть при $\alpha_l = 1/m$ для всех $l = 1, \ldots, m$, имеем

$$\sigma^2[\tilde{K}_m^{(2)}] = \frac{4\sigma^2}{m^2}\left(\sum_{l=1}^{m} \frac{1}{l^4} + \left(\sum_{l=1}^{m} \frac{1}{l^2}\right)^2\right) \leqslant \frac{7\pi^4\sigma^2}{45m^2}.$$

Таким образом, при равномерном усреднении случайная ошибка является уменьшаемой при увеличении размера окна m.

Нахождении оптимального размера окна

Случай равномерного усреднения. В случае равномерного усреднения рассмотрим задачу о нахождении оптимального размера окна m, при котором среднеквадратичная ошибка $S(m) = s_m^2\big(\tilde{k}_m^{(2)}\big) + \sigma^2\big[\tilde{K}_m^{(2)}\big]$ будет наименьшей. Можно считать, что $\sigma^2\big[\tilde{K}_m^{(2)}\big] = \frac{7\pi^4}{45(m+1)^2}\sigma^2$. Тогда, исследуя функцию

$$S(m) = \frac{(m+1)^2}{36}\big(y'''(x^*)\big)^2 + \frac{7\pi^4\sigma^2}{45(m+1)^2}$$

с помощью производной, получим, что оптимальное значение m_{opt} размера окна должно удовлетворять условию

$$m_{\text{opt}} = \left[\pi\sqrt{\frac{2{,}4\,\sigma}{|y'''(x^*)|}}\right] - 1. \qquad (3)$$

При этом оптимальном значении размера окна $S(m_{\text{opt}}) = \frac{71}{540}\pi^2\sigma\,|y'''(x^*)|$. Поскольку $y'''(x^*) \approx k'(x^*)$ при $x^* \approx 0$, можно предложить следующую процедуру уточнения размера «окна» m при вычислении оценки кривизны методом усреднения локально-интерполяционных оценок (например, в алгоритме Freeman и Davis [3]) в случае известного уровня зашумления кривой σ. Если $\tilde{k}_{m_i}^{(2)}(\mathbf{g}_i)$ — оценка кривизны в $\mathbf{g}_i$ точке, вычисленная методом усреднения в «окне» размером m_i локально-интерполяционных оценок, то будем вычислять m_{i+1} по формуле

$$m_{i+1} = \left[c\sqrt{\frac{d(\mathbf{g}_i, \mathbf{g}_{i-1})}{\left|\tilde{k}_{m_i}^{(2)}(\mathbf{g}_i) - \tilde{k}_{m_{i-1}}^{(2)}(\mathbf{g}_{i-1})\right|}}\right] - 1,$$

где $c = \pi\sqrt{2{,}4\sigma}$, $d(\mathbf{g}_i, \mathbf{g}_{i-1})$ — расстояние между точками $\mathbf{g}_i$ и $\mathbf{g}_{i-1}$. Кроме того, из формулы (3) следует, что при изменении уровня зашумления σ размер окна (например, в алгоритме Freeman и Davis [3]) необходимо изменять в соответствии с изменением функции $\sqrt{\sigma}$.

Случай произвольного усреднения. В случае произвольного α-усреднения можно поставить задачу о нахождении такого оптимального весового вектора α, при котором среднеквадратичная ошибка $S(\alpha) = s_m^2\big(\tilde{k}_m^{(2)}(\mathbf{y};\alpha)\big) + \sigma^2\big[\tilde{K}_m^{(2)}(\mathbf{y};\alpha)\big]$ будет наименьшей. Размер «окна» m в этом случае можно считать бесконечно большим. Из (1) и (2) имеем $S(\alpha) = 4\sigma^2\tilde{S}(\alpha)$, где

$$\tilde{S}(\alpha) = T^2\left(\sum_{l=1}^{m}\alpha_l l\right)^2 + \sum_{l=1}^{m}\frac{\alpha_l^2}{l^4} + \left(\sum_{l=1}^{m}\frac{\alpha_l}{l^2}\right)^2,$$

$T = \frac{1}{6\sigma}\big|y'''(x^*)\big|$. Функция $\tilde{S}(\alpha)$ является квадратичной и неотрицательной. Поэтому в выпуклой области — симплексе $\alpha_l \geqslant 0$, $l = 1,\ldots,m$, $\sum\limits_{l=1}^{m}\alpha_l = 1$, существует единственная точка минимума функции $\tilde{S}(\alpha)$. Эту точку можно найти с помощью стандартных оптимизационных процедур. Нетрудно видеть, что решение будет зависеть только от величины T, которая характеризует отношение изменения кривизны к уровню шума (отношение сигнал-шум).

В таблице 1 приведены значения ненулевых весовых коэффициентов α_l, $l = 1,\ldots,m$, вычисленных для некоторых значений T.

Таблица 1. Зависимость значений весовых коэффициентов α_l от отношения T сигнал-шум.

T	α_1	α_2	α_3	α_4	α_5	α_6
0,04	0	0	0	0,16	0,41	0,43
0,05	0	0	0,01	0,28	0,53	0,18
0,06	0	0	0,06	0,38	0,56	0
0,1	0	0	0,32	0,68	0	0
0,3	0	0,45	0,55	0	0	0
1	0,23	0,77	0	0	0	0

Из таблицы видно, что чем выше уровень шума по сравнению с изменением кривизны, тем меньше значение T и значимыми для вычисления оценки кривизны должны быть точечные оценки $k_{l,m}^{(1)}(\mathbf{y})$ с большими значениями l. Используя эти весовые коэффициенты, можно оптимизировать локально-интерполяционные методы оценивания кривизны аналогично рассмотренной выше процедуре «уточнения» размера окна m при равномерном усреднении.

Выводы

Таким образом, при некоррелированном стационарном зашумлении оцифрованной кривой оценка кривизны $\tilde{k}_m^{(2)}(\mathbf{y})$, полученная методом усреднения локально-интерполяционных оценок в «окне» $[-m, m]$, имеет:

1) систематическую ошибку

$$s\big(\tilde{k}_m^{(2)}\big) = \tfrac{1}{6}(m+1)\big|y'''(x^*)\big|,$$

которая не может быть сколь угодно уменьшена при изменении размера окна m;

2) случайную ошибку

$$\sigma^2\big[\tilde{K}_m^{(2)}\big] \leqslant \frac{7\pi^4\sigma^2}{(45m^2)},$$

которая может быть сколь угодно уменьшена при увеличении размера окна m;

3) оптимальный размер окна, определяемый формулой (3), (в общем случае, оптимальные весовые коэффициенты) при котором среднеквадратичная ошибка вычисления оценки кривизны будет наименьшей.

Литература

[1] *Beus H. L., Tiu S. S. H.* An improved corner detection algorithm based on chain-coded plane curves // Pattern Recognition, 20, 1987 — P. 291–296.

[2] *Canny J.* A computational approach to edge detection // IEEE Trans. on PAMI, 8(6), 1986 — P. 679–698.

[3] *Freenan H., Davis L. P.* A corner finding algorithm for chain-code curves // IEEE Trans. Computers, 26, 1977 — P. 297–303.

[4] *Harris C., Stephens M. A.* Combined corner and edge detector // Proc. Fourth Alvey Vision Conference, Manchester, UK, 1988. P. 147–151.

[5] *Лепский А. Е., Бачило С. А., Рыбаков О. С.* Анализ двух методов оценивания кривизны дискретной плоской зашумленной кривой // Сб. трудов 3-й междун. конф. «Цифровая обработка сигналов и ее применение», Москва, 2000. С. 12–14.

[6] *Лепский А. Е.* Оценка числовых характеристик случайного веса в одномерной модели зашумления контура плоского изображения // Известия ТРТУ. — 1999. — № 3(13). — С. 197–200.

[7] *Лепский А. Е.* Об устойчивости центра масс векторного представления в одной вероятностной модели зашумления контура изображения // Автоматика и телемеханика. — 2007. — № 1. — С. 82–92.

Оценка кривизны методом аналитического сглаживания локально-интерполяционных оценок*

Лепский А. Е.
lepskiy@mail.ru
Таганрог, Технологический институт Южного федерального университета

В статье рассматривается метод выделения оценок кривизны плоской вероятностно зашумленной кривой на оцифрованном изображении, найденный с помощью усреднения (по Соболеву) локально-интерполяционных оценок кривизны. Такой метод реализован, например, в алгоритме Canny. Исследуются качественные характеристики таких оценок: систематическая ошибка, смещение и случайная ошибка, а также решается задача о нахождении оптимальных параметров этого метода.

Введение

Кривизна, наряду с краем, является одной из важнейших низкоуровневых особенностей изображения. Для плоской гладкой кривой кривизну $k(\boldsymbol{g})$ в точке $\boldsymbol{g}$ определяют как скорость изменения направления касательного вектора при движении точки по кривой, т. е. $k(\boldsymbol{g}) = \theta'_s(\boldsymbol{g})$, где $\theta(\boldsymbol{g})$ — функции наклона (угол между касательной и заданным направлением, например, положительным направлением оси Ox) и производная берется по длине дуги s. Точки, где направление касательного вектора быстро изменяется, являются точками высокой кривизны. Эти точки будут более информативными, чем точки кривой с низкой кривизной в том смысле, что положение именно этих точек на изображении определяет форму объекта.

Реально имеется только дискретная кривая Γ, выделенная тем или иным методом на оцифрованном изображении. Поэтому вместо вычисления кривизны $k(\boldsymbol{g})$ вычисляют некоторую ее оценку $k_{\boldsymbol{\varepsilon}}(\boldsymbol{g})$, где $\boldsymbol{\varepsilon}$ — параметр (или вектор параметров). Вычисление оценки кривизны только с помощью разностных операторов практические не применяется, поскольку такая оценка будет очень чувствительной к значениям отдельных данных. Для того, чтобы уменьшить эту чувствительность, осуществляют сглаживание первичных оценок кривизны, полученных в пределах ε-окрестности рассматриваемой точки. Этот способ, который условно назовем методом аналитического сглаживания, был предложен, по-видимому, Canny [5]. Применительно к вычислению оценки кривизны подход Canny реализуется следующим образом. В качестве оценки $k_\varepsilon(\boldsymbol{g})$ кривизны плоской оцифрованной кривой Γ в точке $\boldsymbol{g}$ используется результат ε-усреднения (сглаживания) самой функции кривизны $k(\boldsymbol{g})$ (или ее оценки, полученной тем или иным методом) с помощью интегрального оператора свертки:

$$k_\varepsilon(\boldsymbol{g}) = (\varphi_\varepsilon * k)(\boldsymbol{g}) = (\varphi_\varepsilon * \theta'_s)(\boldsymbol{g}) = (\varphi'_\varepsilon * \theta)(\boldsymbol{g}),$$

где φ_ε — некоторое ядро усреднения с вещественным параметром ε, $\theta(\boldsymbol{g})$ — угол между касательной и положительным направлением оси Ox, $\theta'_s(\boldsymbol{g})$ — производная функции наклона $\theta(\boldsymbol{g})$ по длине дуги s. В этом методе дифференциальные операции (производные первого порядка) используются только при вычислении функции θ. Оценка кривизны получается численно равной некоторому усреднению оценок функции θ, вычисленный в разных точках кривой.

В этой статье исследуются оценки кривизны, полученные методом аналитического сглаживания, с точки зрения точности этих оценок и устойчивости их к вероятностному зашумлению кривой, которые численно определяются величинами систематической ошибки $s_{\boldsymbol{\varepsilon}} = |k_{\boldsymbol{\varepsilon}} - k|$, смещения $b_{\boldsymbol{\varepsilon}} = \mathsf{E}[K_{\boldsymbol{\varepsilon}}] - k_{\boldsymbol{\varepsilon}}$ и случайной ошибки $\boldsymbol{\sigma}^2[K_{\boldsymbol{\varepsilon}}]$, где $K_{\boldsymbol{\varepsilon}}$ — случайная оценка кривизны вероятностно зашумленной кривой. Вопросы устойчивости и точности некоторых методов вычисления оценок кривизны и различных представлений кривой в случае вероятностного зашумления изображения рассматривались в работе [3]. Устойчивость детерминистских представлений кривой исследовалась в работе [2]. Для устойчивости оценки кривизны к уровню зашумления необходимо выполнение условия $b_{\boldsymbol{\varepsilon}} \to 0$, $\boldsymbol{\sigma}^2[K_{\boldsymbol{\varepsilon}}] \to 0$ при $\varepsilon \to \infty$. С другой стороны, для точной оценки $s_{\boldsymbol{\varepsilon}} \to 0$ при $\varepsilon \to 0$. Поэтому возникает задача нахождения оптимального значения ε, минимизирующего $s_{\boldsymbol{\varepsilon}}$, $b_{\boldsymbol{\varepsilon}}$ и $\boldsymbol{\sigma}^2[K_{\boldsymbol{\varepsilon}}]$. Решению этих задач и посвящена настоящая статья.

ε-усреднение кривизны

Пусть плоская кривая имеет естественную параметризацию с параметром t и $k(t)$ — кривизна этой кривой. Предположим, что $k(t) \in L_p(\mathbb{R})$. Для сглаживания функции кривизны $k(t)$ нам понадобится одномерное ε-усреднение по Соболеву [4]:

$$k_\varepsilon(t) = \frac{1}{\varepsilon} \int\limits_{-\infty}^{\infty} k(\tau)\varphi((t-\tau)/\varepsilon)d\tau,$$

где $\varphi(t)$ — интегрируемая на $\mathbb{R}$ четная неотрицательная функция такая, что $\int_{-\infty}^{\infty} \varphi(t)dt = 1$. Известно, что $k_\varepsilon(t) \to k(t)$ при $\varepsilon \to 0$. Простейшее

*Работа выполнена при финансовой поддержке РФФИ, проекты №07-07-00067, №08-07-00129.

сглаживание можно получить с помощью постоянного финитного ядра

$$\varphi(t)=\begin{cases}1/2, & |t|\leqslant 1,\\ 0, & |t|>1.\end{cases}$$

В этом случае сглаженная функция будет равна

$$k_\varepsilon(t)=\frac{1}{2\varepsilon}\int\limits_{t-\varepsilon}^{t+\varepsilon}k(\tau)d\tau.$$

Более эффективным для сглаживания является свертывание оценок кривизны с гладкими ядрами, например, с ядром Гаусса $\varphi(t)=\left(1/\sqrt{2\pi}\right)e^{-t^2/2}$ или с гладким финитным ядром

$$\varphi(t)=\begin{cases}\frac{1}{\lambda}\exp(\frac{1}{t^2-1}), & |t|\leqslant 1,\\ 0, & |t|>1,\end{cases}$$

где λ — нормирующий множитель. Для простоты всюду ниже будем считать, что ядро является финитным гладким или кусочно-гладким.

Пусть плоская кривая Γ имеет естественную параметризацию $\boldsymbol{w}(t)=x(t)\boldsymbol{i}+y(t)\boldsymbol{j}$, $0\leqslant t\leqslant L$ (L — длина кривой Γ); $\theta(t)$ — кусочно-дифференцируемая функция. Продолжим функцию $\theta(t)$ с отрезка $[0,L]$ на всю числовую прямую, считая, что $\theta(t)$ равна нулю вне отрезка $[0,L]$. Это продолжение также обозначим через $\theta(t)$. Кроме того, будем считать, что кривизна кривой Γ оценивается в такой точке $\boldsymbol{g}=\boldsymbol{w}(t)$, что значения естественного параметра t и параметра усреднения ε удовлетворяют условию $[t-\varepsilon,t+\varepsilon]\subseteq[0,L]$. Тогда сглаживание функции кривизны $k(t)=\theta'(t)$ дает следующее усреднение

$$k_\varepsilon(t)=\theta'*\varphi_\varepsilon=\theta*\varphi'_\varepsilon=\frac{1}{\varepsilon^2}\int\limits_{t-\varepsilon}^{t+\varepsilon}\theta(\tau)\varphi'\left(\frac{t-\tau}{\varepsilon}\right)d\tau.$$

Таким образом, для вычисления усредненного значения кривизны, необходимо произвести свертку функции наклона $\theta(t)$ с производной ядра $\varphi'(t)$. Положительный параметр ε регулирует степень гладкости оценки и величину усредняющего «окна».

Для погрешности усреднения кривизны с помощью финитного ядра справедлива оценка [4]

$$\left|k_\varepsilon(t)-k(t)\right|\leqslant\sup_{|\tau|\leqslant\varepsilon}\left|k(t-\tau)-k(t)\right|\leqslant C_1(\varepsilon,\Gamma)\varepsilon,\quad(1)$$

где $C_1(\varepsilon,\Gamma)=\sup\{k'(\tau)\colon t-\varepsilon\leqslant\tau\leqslant t+\varepsilon\}$, если функция кривизны k дифференцируема в ε-окрестности точки t.

Аналитическое сглаживание первичных оценок кривизны

Пусть $\{\tau_i\}_{i=0}^{n-1}$ — некоторое разбиение отрезка $[t-\varepsilon,t+\varepsilon]\subseteq[0,L]$ (L — длина кривой Γ), $\Delta\tau_i$ — i-й шаг разбиения, $\tilde{k}(\tau_i)$ — дискретная функция локально-интерполяционных оценок кривизны в точках τ_i, $i=0,\ldots,n-1$. Будем считать, что $\tilde{k}(\tau_i)=\frac{\tilde{\theta}(\tau_{i+1})-\tilde{\theta}(\tau_i)}{\Delta\tau_i}$, где $\tilde{\theta}(\tau_i)$ — оценки функции наклона в точках τ_i. Выполним ε-усреднение функции $\tilde{k}$, получим оценку

$$\tilde{k}_\varepsilon(t)=\frac{1}{\varepsilon^2}\sum_{i=0}^{n-1}c_i\tilde{\theta}(\tau_i)\varphi'\left(\frac{t-\tau_i}{\varepsilon}\right),$$

где c_i — множители квадратурной формулы.

Описанный метод был апробирован в ряде алгоритмов. Так, в [6] метод аналитического сглаживания с ядром Гаусса использовался для выделения контрольных точек на модельных контурных изображениях объектов, состоящих из 400-700 точек. Значение ε выбиралось эмпирически из интервала $[4,7]$, а сглаженная функция кривизны при этом содержала около 40 локальных экстремумов.

Систематическая ошибка аналитического сглаживания

Пусть плоская кривая Γ класса C^3 имеет естественную параметризацию $\boldsymbol{w}(t)=x(t)\boldsymbol{i}+y(t)\boldsymbol{j}$. Оценим систематическую ошибку аналитического ε-сглаживания локально-интерполяционных оценок $\tilde{k}$ кривизны k кривой Γ в точке $\boldsymbol{g}=\boldsymbol{w}(t)$. Имеем

$$s_\varepsilon\leqslant\left|\tilde{k}_\varepsilon(t)-k^e_\varepsilon(t)\right|+\left|k^e_\varepsilon(t)-k_\varepsilon(t)\right|+\left|k_\varepsilon(t)-k(t)\right|,\quad(2)$$

где $k^e_\varepsilon(t)=\frac{1}{\varepsilon^2}\sum\limits_{i=0}^{n-1}c_i\theta(\tau_i)\varphi'\left(\frac{t-\tau_i}{\varepsilon}\right)$ — дискретное аналитическое ε-сглаживание точной функции кривизны k, вычисленное по квадратурной формуле численного интегрирования;

$k_\varepsilon(t)=\frac{1}{\varepsilon^2}\int\limits_{t-\varepsilon}^{t+\varepsilon}\theta(\tau)\varphi'\left(\frac{t-\tau}{\varepsilon}\right)d\tau$ — интегральное аналитическое ε-сглаживание точной функции кривизны k;

$\tilde{k}_\varepsilon(t)=\frac{1}{\varepsilon^2}\sum\limits_{i=0}^{n-1}c_i\tilde{\theta}(\tau_i)\varphi'\left(\frac{t-\tau_i}{\varepsilon}\right)$ — аналитическое ε-сглаживание локально-интерполяционных оценок $\tilde{k}$ кривизны k.

Первое слагаемое в неравенстве (2) характеризует погрешность вычисления первичных оценок кривизны, второе — погрешность интегрирования, а третье — погрешность усреднения (1). Оценим первое слагаемое. Имеем

$$\left|\tilde{k}_\varepsilon(t)-k^e_\varepsilon(t)\right|\leqslant\\ \leqslant\frac{1}{\varepsilon^2}\max_i\left|\tilde{\theta}(\tau_i)-\theta(\tau_i)\right|\sum_{i=0}^{n-1}c_i\left|\varphi'\left(\frac{t-\tau_i}{\varepsilon}\right)\right|.$$

Таким образом, первое слагаемое зависит от погрешности вычисления оценки функции наклона $\tilde{\theta}$. Локально-интерполяционная оценка функции наклона $\tilde{\theta}$ может быть вычислена по формуле $\tilde{\theta}(t)=\operatorname{arctg}\left(\Delta y(t)/\Delta x(t)\right)$, где $\Delta x(t)$, $\Delta y(t)$ — конечные разности. На практике конечные разности

Δx и Δy вычисляют с помощью одного из разностных операторов (Робертса, Собеля и др.) [1]. Предположим, что $\Delta x(t) = x(t+\Delta t) - x(t)$, $\Delta y(t) = y(t+\Delta t) - y(t)$. Тогда

$$\Delta x(t) = x'(t) + \tfrac{1}{2}x''(\xi)\Delta t,$$
$$\Delta y(t) = y'(t) + \tfrac{1}{2}y''(\eta)\Delta t,$$

где $\xi, \eta \in (t, t+\Delta t)$. Для разности $\left|\tilde{\theta}(t) - \theta(t)\right|$ имеем

$$\left|\tilde{\theta}(t) - \theta(t)\right| = \left|\operatorname{arctg}\left(\frac{\Delta y(t)}{\Delta x(t)}\right) - \operatorname{arctg}\left(\frac{y'(t)}{x'(t)}\right)\right| \leqslant$$
$$\leqslant \left|\frac{\Delta y(t)}{\Delta x(t)} - \frac{y'(t)}{x'(t)}\right| \leqslant A(\Gamma)\,|\Delta t|\,,$$

где $A(\Gamma)$ — некоторая константа, зависящая от кривой Γ. Тогда

$$\left|\tilde{k}_\varepsilon(t) - k^e_\varepsilon(t)\right| \leqslant \frac{A(\Gamma)}{\varepsilon^2}\max_i\left|\Delta\tau_i\right|\sum_{i=0}^{n-1} c_i\left|\varphi'\left(\frac{t-\tau_i}{\varepsilon}\right)\right|.$$

В частности, для равномерного разбиения $\{\tau_i\}_{i=0}^{n-1}$ отрезка $[t-\varepsilon, t+\varepsilon]$ с шагом $\Delta\tau_i = 2\varepsilon/n$, получим $\left|\tilde{k}_\varepsilon(t) - k^e_\varepsilon(t)\right| \leqslant \frac{2A(\Gamma)}{\varepsilon n}\sum\limits_{i=0}^{n-1} c_i\left|\varphi'\left(1-\frac{2}{n}i\right)\right| = \frac{C_2(\varepsilon,\Gamma)}{\varepsilon n}$, где $C_2(\varepsilon,\Gamma) = 2A(\Gamma)\sup\limits_n\sum\limits_{i=0}^{n-1} c_i\left|\varphi'\left(1-\frac{2}{n}i\right)\right|$.

Второе слагаемое в (2) определяется погрешностью интегрирования и для равномерного разбиения будет равно $\left|k^e_\varepsilon(t) - k_\varepsilon(t)\right| \leqslant C_3(\varepsilon,\Gamma)\left(\frac{\varepsilon}{n}\right)^q$, где q — порядок точности численного интегрирования (обычно, $q \geqslant 2$), $C_3(\varepsilon,\Gamma)$ — некоторая константа.

Оценка третьего слагаемого в (2) определяется неравенством (1). Таким образом, для систематической ошибки кривизны справедлива теорема.

Теорема 1. *Для систематической ошибки аналитического ε-сглаживания первичных оценок кривизны верна оценка*

$$s_\varepsilon \leqslant C_1(\varepsilon,\Gamma)\,\varepsilon + C_2(\varepsilon,\Gamma)\frac{1}{\varepsilon n} + C_3(\varepsilon,\Gamma)\left(\frac{\varepsilon}{n}\right)^q, \quad (3)$$

где q — порядок точности численного интегрирования, n — число точек разбиения, $C_i(\varepsilon,\Gamma)$, $i = 1,2,3$, — ограниченные сверху по ε константы. В частности, если $n = O\left(1/\varepsilon^2\right)$, то $s_\varepsilon \leqslant C(\varepsilon,\Gamma)\,\varepsilon$, где константа $C(\varepsilon,\Gamma)$ ограничена сверху по ε.

Из теоремы 1 следует, что систематическая ошибка аналитического ε-сглаживания первичных оценок кривизны может быть сделана сколь угодно малой с уменьшением ε.

Смещение аналитического сглаживания при сферическом нормальном зашумлении кривой

Пусть Γ — плоская кривая, имеющая естественную параметризацию $\boldsymbol{w}(t) = x(t)\boldsymbol{i} + y(t)\boldsymbol{j}$. Рассмотрим дискретизацию Γ_d этой кривой: $\Gamma_d = \left(\boldsymbol{w}(t_k)\right)_{k=0}^{p-1}$, $\boldsymbol{w}(t_k) = x_k\boldsymbol{i} + y_k\boldsymbol{j}$.

Предположим, что дискретная кривая Γ_d подвергнута аддитивному сферическому нормальному зашумлению $\mathcal{W}_{d,1}(\sigma)$: $\boldsymbol{Y} = \boldsymbol{y} + \boldsymbol{\xi}$, где $\boldsymbol{\xi} \sim N(\boldsymbol{0}, \sigma^2 I)$. Тогда аналитическое ε-сглаживание локально-интерполяционных оценок $\tilde{k}$ кривизны k кривой Γ в точке $\boldsymbol{g} = \boldsymbol{w}(t)$ будет случайной величиной, которую обозначим через K_ε. Оценим величину смещения этой случайной величины.

Предположим, что $\{\tau_i\}_{i=0}^{n-1} \subset \{t_k\}_{k=0}^{p-1}$ — разбиение отрезка $[t-\varepsilon, +\varepsilon] \subseteq [0, L]$ (L — длина кривой Γ), $x(\tau_i) = x_i$, $y(\tau_i) = y_i$, $\Delta x_i = x_{i+1} - x_i$, $\Delta y_i = y_{i+1} - y_i$, $i = 0, \ldots, n-1$. Тогда первичные оценки функции наклона равны

$$\tilde{\theta}_i = \tilde{\theta}(\tau_i) = \begin{cases} \operatorname{arctg}(\Delta y_i/\Delta x_i), & \Delta x_i \neq 0, \\ \pi/2, & \Delta x_i = 0, \end{cases}$$

$i = 0, \ldots, n-1$. В случае вероятностного зашумления кривой для тех индексов $i = 0, \ldots, n-1$, для которых $\Delta x_i \neq 0$, первичные оценки функции наклона будут случайными величинами Θ_i, а ε-сглаживание локально-интерполяционных оценок — случайной величиной $K_\varepsilon = \sum\limits_{i=0}^{n-1} c_i(\varepsilon)\Theta_i$, где $c_i(\varepsilon) = \frac{1}{\varepsilon^2}\varphi'\left(\frac{t-\tau_i}{\varepsilon}\right)$ (см. предыдущий пункт). Справедлива следующая теорема о смещении случайной величины K_ε.

Теорема 2. *Смещение случайной величины K_ε ε-сглаживания локально-интерполяционных оценок при зашумлении $\mathcal{W}_{d,1}(\sigma)$ равно*

$$b(K_\varepsilon) = -2\sigma^2 \sum_{i=0,\ \Delta x_i \neq 0}^{n-1} \frac{c_i(\varepsilon)\sin\tilde{\theta}_i\cos^3\tilde{\theta}_i}{\Delta x_i^2} + R,$$

где

$$|R| \leqslant \frac{2}{3\sqrt{\pi}} \sum_{i=0,\ \Delta x_i \neq 0}^{n-1} \left|c_i(\varepsilon)\right| \left(\frac{\sigma}{|\Delta x_i|}\right)^3$$

Если $\{\tau_i\}_{i=0}^{n-1} \subset (t_k)_{k=0}^{p-1}$ — равномерное разбиение отрезка $[t-\varepsilon, t+\varepsilon]$, $\Delta\tau_i = 2\varepsilon/n$, то $\Delta x_i = x_{i+1} - x_i = (2\varepsilon\operatorname{tg}\alpha_i)/n$, $\operatorname{tg}\alpha_i = (x_{i+1} - x_i)/\Delta\tau_i$ и справедливо следствие.

Следствие 1. *При указанных условиях смещение случайной величины K_ε аналитического ε-сглаживания локально-интерполяционных оценок в случае зашумления $\mathcal{W}_{d,1}(\sigma)$ равно*

$$b(K_\varepsilon) = C_4(\varepsilon,\Gamma)\left(\frac{\sigma n}{\varepsilon}\right)^2 + R, \quad (4)$$

где $|R| \leqslant \tilde{C}_4(\varepsilon,\Gamma)\left(\frac{n\sigma}{\varepsilon}\right)^3$, и константы $C_4(\varepsilon,\Gamma)$, $\tilde{C}_4(\varepsilon,\Gamma)$ ограничены по ε.

Оценка (4) показывает, что смещение случайной величины K_ε можно уменьшить за счет увеличения размера окна ε или за счет уменьшения числа локально-интерполяционных оценок n.

Случайная ошибка аналитического сглаживания при сферическом нормальном зашумлении кривой

Предположим, что дискретная кривая Γ_d подвергнута аддитивному сферическому нормальному зашумлению $\mathcal{W}_{d,1}(\sigma)$ вида $\boldsymbol{Y} = \boldsymbol{y} + \boldsymbol{\xi}$, где $\boldsymbol{\xi} \sim N(\mathbf{0}, \sigma^2 I)$. Справедлива следующая теорема.

Теорема 3. *Случайная ошибка величины K_ε ε-сглаживания локально-интерполяционных оценок при зашумлении $\mathcal{W}_{d,1}(\sigma)$ равна*

$$\boldsymbol{\sigma}^2[K_\varepsilon] = \sigma^2 \sum_{i=0,\ \Delta x_i \neq 0}^{n} \left(b_i^2(\varepsilon) - b_i(\varepsilon) b_{i-1}(\varepsilon) + b_{i-1}^2(\varepsilon)\right) + H,$$

где $b_i(\varepsilon) = c_i(\varepsilon)\cos^2\tilde{\theta}_i / |\Delta x_i|$, $i = 0, \ldots, n-1$, $b_{-1}(\varepsilon) = b_n(\varepsilon) = 0$ и $H = O(\sigma^4)$.

Если $\{\tau_i\}_{i=0}^{n-1} \subset \{t_k\}_{k=0}^{p-1}$ — равномерное разбиение отрезка $[t-\varepsilon, t+\varepsilon]$, $\Delta\tau_i = 2\varepsilon/n$, то $\Delta x_i = x_{i+1} - x_i = (2\varepsilon \operatorname{tg}\alpha_i)/n$, $\operatorname{tg}\alpha_i = (x_{i+1} - x_i)/\Delta\tau_i$ и справедливо следствие.

Следствие 2. *При указанных условиях случайная ошибка величины K_ε аналитического ε-сглаживания локально-интерполяционных оценок в случае зашумления $\mathcal{W}_{d,1}(\sigma)$ равна*

$$\boldsymbol{\sigma}^2[K_\varepsilon] = C_5(\varepsilon, \Gamma)\left(\frac{\sigma n}{\varepsilon}\right)^2 + O\left(\frac{n\sigma}{\varepsilon}\right)^4, \qquad (5)$$

где константа $C_5(\varepsilon, \Gamma)$ ограничена по ε.

Оценка (5) показывает, что случайную ошибку величины K_ε можно уменьшить за счет увеличения размера окна ε или за счет уменьшения числа локально-интерполяционных оценок n.

Оптимальные значения параметров аналитического сглаживания

Поставим задачу о вычислении оптимальных значений параметров метода аналитического сглаживания — размера сглаживающего окна ε и числа первичных оценок кривизны в этом окне n. В качестве функции критерия $\psi(\varepsilon, n)$ будем использовать сумму главных значений верхних оценок систематической ошибки, модуля смещения и среднеквадратичного отклонения. Из (3), (4), (5) имеем

$$\psi(\varepsilon, n) = C_1\varepsilon + \frac{C_2}{\varepsilon n} + C_3\left(\frac{\varepsilon}{n}\right)^q + \sigma^2 C_4\left(\frac{n}{\varepsilon}\right)^2 + \sigma C_5 \frac{n}{\varepsilon}.$$

Константы $C_i > 0$, $i = 1, \ldots, 5$ также зависят от ε и n, но в силу ограниченности констант по ε и n, этой зависимостью пренебрежем. Постоянная $q \geqslant 2$ определяется выбранным методом численного интегрирования. После замены переменных $\varepsilon/n = x$, $\varepsilon n = y$, функция критерия примет вид

$$\tilde{\psi}(x, y) = C_1\sqrt{xy} + \frac{C_2}{\varepsilon y} + C_3 x^q + C_4\sigma^2 x^2 + C_5\sigma x.$$

Составим нормальную систему:

$$\begin{cases} \tilde{\psi}'_x = \frac{C_1}{2}\sqrt{\frac{y}{x}} + qC_3x^{q-1} - \frac{2C_4\sigma^2}{x^3} - \frac{C_5\sigma}{x^2} = 0, \\ \tilde{\psi}'_y = \frac{C_1}{2}\sqrt{\frac{x}{y}} - \frac{C_2}{y^2} = 0. \end{cases} \Rightarrow$$

$$\Rightarrow \begin{cases} \sqrt[3]{\frac{C_1^2 C_2 x^8}{4}} + qC_3x^{q+2} = 2C_4\sigma^2 + C_5\sigma x, \\ y = \left(\frac{2C_2}{C_1}\sqrt{x}\right)^{1/3}. \end{cases}$$

Первое уравнение системы имеет единственный положительный корень. Поэтому функция $\tilde{\psi}(x, y)$ имеет в области $\mathbb{R}^2_+$ единственную точку минимума. Если x и y — положительные решения нормальной системы, то оптимальные значения ε и n найдем по формулам: $\varepsilon = \sqrt{xy}$, $n = \left[\sqrt{y/x}\right]$.

Выводы

При некоррелированном нормальном зашумлении оцифрованной кривой Γ оценка кривизны $\tilde{k}_\varepsilon(t)$, полученная методом аналитического ε-сглаживание, имеет:

1) систематическую ошибку, которая может быть сделана сколь угодно малой с уменьшением ε (оценка (3));
2) смещение и случайную ошибку, которые можно уменьшить за счет увеличения размера окна ε или за счет уменьшения числа локальных оценок n (оценки (4) и (5));
3) оптимальные значения размера окна ε и числа первичных оценок кривизны n, при которых суммарная ошибка вычисления оценки кривизны будет наименьшей.

В качестве преимуществ данного метода отметим его пригодность для обработки полутоновых изображений, возможность его совмещения с процедурой выделения краев. К недостаткам этого метода следует отнести сравнительно большой объем вычислений, требуемый для его реализации.

Литература

[1] *Гонсалес Р., Вудс Р.* Цифровая обработка изображений. — Москва: Техносфера, 2006.

[2] *Каркищенко А. Н., Лепский А. Е.* Об устойчивости центра масс, векторов и дескриптора Фурье векторного представления контура изображения // Автоматика и телемеханика. — 2001. — № 3. — С. 141–151.

[3] *Лепский А. Е.* Об устойчивости центра масс векторного представления в одной вероятностной модели зашумления контура изображения // Автоматика и телемеханика. — 2007. — № 1. — С. 82–92.

[4] *Никольский С. М.* Курс математического анализа, Т.2. — Москва: Наука, 1983.

[5] *Canny J.* A computational approach to edge detection // IEEE Trans. on PAMI, 8(6), 1986.— P. 679–698.

[6] *Liu H. C., Srinath M. D.* Partial shape classification using contour matching in distance transformation // IEEE Trans. on Pattern anal. And Mach. Intel., 11(12), 1990. — P. 1072–1079.

Построение циклических разностных множества Адамара*

Леухин А. Н.
code@marstu.net
Йошкар-Ола, Марийский государственный технический университет

Выполнен обзор известных на сегодняшний день методов построения циклических разностных множеств Адамара. Обсуждается возможность реализации обобщающего метода построения таких множеств.

Введение

Бинарные последовательности широко применяются в современных радиотехнических системах: в качестве моделирующих последовательностей в радарах и сонарах; в качестве ансамблей широкополосных сигнатур в системах беспроводной телефонии с кодовым разделением каналов; в качестве секретных ключей в криптографических симметричных системах; а также в качестве помехоустойчивых кодов, исправляющих ошибки. Наибольший интерес представляют бинарные последовательности с одноуровневой периодической автокорреляционной функцией (ПАКФ) (уровень боковых лепестков равен $a = -1$), построенные на основе циклических разностных множеств типа Адамара.

В 1935 году в работе [1] был сформулирован принцип линейного уплотнения и разделения каналов. В настоящее время широко используется принцип кодового разделения каналов, в основе которого лежит концепция многоканальной широкополосной связи. Оптимальными с точки зрения скорости передачи информации и помехоустойчивости являются ансамбли ортогональных и квазиортогональных сигналов, имеющих широкий равномерный спектр (соответственно одноуровневую ПАКФ с малым уровнем боковых лепестков).

В 1946–1948 годах в фундаментальных работах в области теории информации [2] было показано, что если в качестве ключа в симметричной схеме шифрования использовать случайный ключ, при этом исходный текст и ключ должны иметь одинаковую длину, то криптосистема является стойкой. В начале 1950-х годов появились устройства, выполняющие операцию сложения по модулю 2. Было предложено использовать эти устройства в качестве генераторов ключей для потоков данных в криптографических приложениях. В этих схемах ключ представляет собой бинарную последовательность, которая является псевдослучайной и имеет корреляционные характеристики близкие к белому шуму. Однако вопрос о длине периода таких последовательностей оставался открытым.

*Работа выполнена при финансовой поддержке РФФИ (проект № 09-07-00072-а), фонда HCF (договор № 189), а также в рамках АВЦП «Развития научного потенциала высшей школы» мероприятия 1 (НИР № 1.02.09).

В 1953 году в работе [3] было показано, что для оптимального решения радиолокационных задач разрешения и оценки параметров необходимо использовать шумоподобные коды, роль которых могут играть те же самые псевдослучайные бинарные последовательности с хорошими корреляционными свойствами. Именно при решении локационных задач были получены основные результаты в области синтеза бинарных последовательностей с одноуровневой ПАКФ.

Также в работе [2] были сформулированы основные принципы помехоустойчивого кодирования и передачи информации по каналу связи со сколь угодно малой вероятностью ошибки. В работе [4] была решена задача о потенциальной помехоустойчивости системы. Вскоре появились работы, в которых были рассмотрены вопросы использования бинарных последовательностей с одноуровневой ПАКФ в качестве помехоустойчивых кодов, способных обнаруживать и исправлять ошибки.

Достаточное условие существования бинарного кода длины N с одноуровневой ПАКФ заключается в существовании разностного множества D с параметрами $D(\nu = N, k, \lambda)$. Особое значение для решения прикладных задач имеют разностные множества Адамара. Однако на сегодняшний день не решенной остается задача построения всех разностных множеств такого типа. В работе будет проведен обзор известных методов построения разностных множеств Адамара, а также предложен общий подход к их построению.

Циклические разностные множества Адамара

Матрица Адамара H представляет собой квадратную матрицу порядка n, элементы которой могут принимать только два значения $+1$ и -1, со свойством:

$$H H^{\intercal} = n \cdot I,$$

где $H^{\intercal}$ — транспонированная матрица, I — единичная матрица порядка n, любые две строки с номерами i и j матрицы H образуют взаимноортогональные векторы (т. е. скалярное произведение $\left(H^{\intercal\langle i\rangle}, H^{\intercal\langle j\rangle}\right) = 0;\ i \neq j;\ i, j = 0, \ldots, n-1$).

Порядок матрицы Адамара может принимать лишь значения из множества $\{1, 2, 4t\}$, где t — любое положительное целое число. Однако не для всех значений t удалось построить матрицы

Адамара, наименьший неизвестный случай $4t = 668$.

Известен только один пример матрицы Адамара для $n = 4$, образующий циркулянтную матрицу (строки матрицы образованы циклическими сдвигами),

$$H = \begin{bmatrix} -1 & 1 & 1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 \end{bmatrix}.$$

Однако известно большое множество матриц Адамара, состоящих из циркулянтной подматрицы $(n-1)\times(n-1)$, обрамленной сверху строкой и слева столбцом из элементов, равных $+1$. Например,

$$H = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}.$$

Такие матрицы принято называть циклическими матрицами Адамара. Теория построения циклических матриц Адамара тесно связана с теорией разностных множеств

$$D(\nu, k, \lambda) = \{d_1, d_2, \ldots, d_k\},$$

где d_i — элементы разностного множества, ν, k, λ — параметры разностного множества.

Порядковые номера столбцов (без учета одного столбца) любой строки (кроме нулевой) матрицы Адамара со значением равным $+1$ могут образовывать разностное множество, которое в этом случае принято называть разностным множеством Адамара. Разностное множество Адамара имеет параметры $(\nu, k, \lambda) = (4t-1, 2t-1, t-1)$.

Обзор известных методов синтеза циклических разностных множеств

В работе [5] были открыты линейные рекуррентные последовательности, которые приводят к матрицам Адамара. Впоследствии такие последовательности были названы последовательностями квадратичных вычетов, последовательностями Лежандра или последовательностями Пэли. Длина последовательностей $N = 4t - 1 = p$, где p — простое число, $t \in Z = \{1, 2, 3, \ldots\}$. Систематическое изучение линейных рекурсий над конечными полями (включая поле $GF(2)$) началось в работе [6]. В работе [7] были получены разностные множества Зингера, в частном случае при характеристике конечного поля $p = 2$, множества приводят к бинарным последовательностям максимальной длины (к m-последовательностям длиной $N = 2^m - 1$). Возможно, S. W. Golomb был первым, кто указал на связь между бинарными последовательностями с одноуровневой периодической автокорреляционной функцией и циклическими разностными множествами (ν, k, λ). В частности, в работе [8] он указал на тот факт, что последовательности квадратичных вычетов обладают одноуровневой ПАКФ, а в работе [9] сформулировал задачу нахождения всех бинарных последовательностей с одноуровневыми ПАКФ (или другими словами, задачу определения всех конструкций, которые образуют разностные множества с параметрами $(4t-1, 2t-1, t-1)$). В работе [10] были построены разностные множества Адамара с параметром $\nu = 4b^2 + 27 = p$, где p — любое положительное целое число (последовательности Холла), а в работе [11] — с параметром $\nu = p_1 \ldots p_2$, где $p_2 = p_1 + 2$, p_1 и p_2 — простые числа (последовательности Якоби). В работах [12, 13] были получены GMW-разностные множества, с параметром $\nu = 2^n - 1$ (GMW последовательности четырех типов: GMW последовательности, каскадные GMW последовательности, обобщенные GMW последовательности типа I и обобщенные GMW последовательности типа II).

Только эти бинарные последовательности, образующие циклические разностные множества Адамара (последовательности квадратичных вычетов, последовательности Зингера, последовательности Холла, последовательности Якоби и последовательности GMW), оставались известными вплоть до 1971 года. В работе [14] в результате полного перебора разностных множеств вида $(\nu, k, \lambda) = (127, 63, 31)$ были найдены 6 неэквивалентных примеров, из которых только 3 являлись ранее известными конструкциями циклических разностных множеств. Другие необъясненные примеры бинарных последовательностей появились в работе [15] при полном исследовании циклических разностных множеств вида $(\nu, k, \lambda) = (255, 127, 63)$. Еще большее количество таких примеров появилось в работе [16] при полном исследовании циклических разностных множеств вида $(\nu, k, \lambda) = (511, 255, 127)$, $\nu = 2^9 - 1 = 511$. Таким образом, некоторые из найденных за период с 1971 по 1991 последовательностей размерностей $N = \nu = 2^7 - 1 = 127$, $N = \nu = 2^8 - 1 = 255$ и $N = \nu = 2^9 - 1 = 511$ нельзя было отнести ни к одному из известных на тот момент решений. Следующее открытие новых конструкций циклических разностных множеств Адамара состоялось спустя десятилетие в 2001 году при полном изучении случая $(\nu, k, \lambda) = (1023, 511, 255)$, $N = \nu = 2^{10} - 1 = 1023$ [17].

Интенсивный поиск конструкций для построения найденных, но необъясненных последовательностей с одноуровневой автокорреляционной функцией начался с начала 1990 годов. В период с 1997 года по 2004 год удалось систематизировать и предложить эффективные конструкции для построе-

ния всех необъясненных ранее последовательностей длин $N = 2^m - 1$, полученных для случаев $m = 7, 8, 9, 10$. Были получены следующие результаты в разработке конструкций последовательностей длины $N = 2^m - 1$ с одноуровневой автокорреляционной функцией [18].

а) Периодические 3-term последовательности длины $N = 2^m - 1$ для нечетных $m = 2t + 1$, где $m \geqslant 5$. Отметим, что первоначально, описание 3-term последовательностей появилось в работе [19].
б) Периодические 5-term последовательности длины $N = 2^m - 1$ для $5 \mod 3 \neq 0$, где $m \geqslant 7$.
в) WG-последовательности длины $N = 2^m - 1$, где $m \geqslant 7$, получаемые с помощью WG-преобразований периодических 5-term последовательностей.

В 1998 году в работе [22] было показано, что мономиальные гиперовалы Segre [20] и гиперовалы Glynn [21] первого и второго типа порождают циклические разностные множества Адамара с параметрами ($\nu = 2^n - 1,\ k = 2^{n-1} - 1,\ \lambda = 2^{n-2} - 1$). Вопросы практической реализации гиперовальных последовательностей рассмотрены в работе [23].

Почти в то же самое время, когда велись исследования по построению гиперовальных бинарных последовательностей, в работе [24] было найдено представление WG-последовательностей степенными функциями Касами и моделированием было показано, что такая конструкция имеет одноуровневую ПАКФ для $m \leqslant 23$. Используя это представление, в работе [25] было аналитически доказано, что конструкции, построенные с помощью степенных функций Касами, порождают одноуровневую автокорреляцию в случае нечетных m. Затем в работе [26] было аналитически доказано, что последовательности, построенные с помощью степенных функций Касами, обладают одноуровневой автокорреляционной функцией в случае четных m.

В 2002 году в работе [27] было введено децимационное преобразования Адамара (DHT — decimation-Hadamar transform), с помощью которого удалось сконструировать новые последовательности с одноуровневой ПАКФ.

В 2004 году в работе [28] была выполнена полная классификацию всех известных на тот момент конструкций для циклических разностных множеств Адамара.

Обобщенный метод синтеза циклических разностных множеств Адамара

В работе [29] приводится система уравнений для синтеза фазокодировнных последовательностей с одноуровневой ПАКФ. Задача синтеза сводится к решению системы уравнений:

— для четных N: $K = \frac{1}{2}N - 1,\ n = 1, \ldots, K$,

$$\begin{cases} \cos(\varphi_n) + \cos(\varphi_{N-n}) + \sum\limits_{m=1}^{N-n-1} \cos(\varphi_m - \varphi_{m+n}) + \\ \qquad + \sum\limits_{m=1}^{n-1} \cos(\varphi_m - \varphi_{m+N-1}) = a, \\ \cos(\varphi_K) + \sum\limits_{m=1}^{N-K-1} \cos(\varphi_m - \varphi_{m+K}) = \frac{1}{2}a, \\ \sin(\varphi_n) + \sin(\varphi_{N-n}) + \sum\limits_{m=1}^{N-n-1} \sin(\varphi_m - \varphi_{m+n}) + \\ \qquad + \sum\limits_{m=1}^{n-1} \sin(\varphi_m - \varphi_{m+N-1}) = 0; \end{cases} \tag{1}$$

— для нечетных N: $K = \frac{N-1}{2},\ n = 1, \ldots, K$,

$$\begin{cases} \cos(\varphi_n) + \cos(\varphi_{N-n}) + \sum\limits_{m=1}^{N-n-1} \cos(\varphi_m - \varphi_{m+n}) + \\ \qquad + \sum\limits_{m=1}^{n-1} \cos(\varphi_m - \varphi_{m+N-1}) = a, \\ \sin(\varphi_n) + \sin(\varphi_{N-n}) + \sum\limits_{m=1}^{N-n-1} \sin(\varphi_m - \varphi_{m+n}) + \\ \qquad + \sum\limits_{m=1}^{n-1} \sin(\varphi_m - \varphi_{m+N-1}) = 0. \end{cases} \tag{2}$$

Не ограничивая общности решений, для исключения «повернутых» кодовых комбинаций угол нулевого вектора φ_0 можно положить равным нулю, т. е. $\varphi_0 = 0^o$. Для синтеза бинарных последовательностей, соответствующих разностным множествам Адамара, необходимо найти решения в случае уровня боковых лепестков $a = -1$. Решения системы уравнений (1)–(2) для конкретного значения N позволяют синтезировать все возможные разностные множества Адамара, или дать ответ о невозможности построения таких множеств. Методы решения системы уравнений (1)–(2) рассмотрены в работе [29].

Заключение

На сегодняшний день не существует строго доказательства, что рассмотренными примерами исчерпываются все конструкции циклических разностных множеств Адамара в случае $\nu = 2^m - 1$. Однако отметим, что для случая $m = 10$ новых семейств циклических разностных множеств Адамара найдено не было. Все известные конструкции циклических разностных множеств были получены для случаев $2 \leqslant m \leqslant 9$.

Начиная с размерности $m = 5$ в начале 1950-х годов, в течение примерно каждых следующих 10 лет успешно полностью исследовались разностные множества с новым значение m ($m = 6$ — 1962 год, $m = 7$ — 1971 год, $m = 8$ — 1983 год, $m = 9$ — 1991 год, $m = 6$ — 2001 год). Экстраполируя результаты, можно сделать следующий прогноз: в 2010 ожидается полное исследование циклических разностных множеств Адамара для слу-

чая $m = 11$, а к 2020 году — для случая $m = 12$. В последних работах [30, 31] сообщалось, что новых разностных множеств Адамара найдено не было.

Литература

[1] *Агеев Д. В.* Основы теории линейной селекции // Научно-технический сборник ЛИИС, 1935. — № 10. — С. 8–28.

[2] *Shannon C.* Mathematical theory of communication // BSTJ. — 1948. — Vol. 27, № 3.

[3] *Wodward P. M.* Probability and information theory with applications to radar, N.Y.: Pergamon Press, 1953.

[4] *Котельников В. А.* Теория потенциальной помехоустойчивости. Госэнергоиздат, 1956.

[5] *Paley. R. E. A. C.* On orthogonal matrices // J. Math. Phys. 12 (1933). — P. 311–320.

[6] *Ore O.* Contributions to the theory of finite fields // Trans. Am. Math. Soc. 36 (1934). — P. 243–274.

[7] *Singer J.* A theorem in finite projective geometry and some applications to number theory // Trans. Amer. Math. Soc., 43 (1938). '— P. 377–385.

[8] *Golomb S. W.* Remarks on orthogonal sequences // The Glenn L. Martin Company, Baltimore, MD, July 28, 1954.

[9] *Golomb S. W.* Sequences with randomness properties. Baltimore, Glenn L. Martin Company, 1955.

[10] *Hall M.* Survey of difference sets // Proc. Am. Math. Soc. 7 (1956), pp.975-986.

[11] *Stanton R. G., Sprott D. A.* A family of difference sets. // Canad. J.Math 10 (1958).

[12] *Gordon B., Mill W. H., Welch L. R.* Some new difference sets // Canadian J. Math. 14 (1962). — P. 614–625.

[13] *Scholtz R. A., Welch L. R.* GMW sequences // IEEE Trans. Inform. Theory, 1984. — Vol. IT-30, № 9.

[14] *Baumert L. D.* Cyclic difference sets // Lecture Notes in Mathematics, Vol. 182, Berlin, Springer-Verlag, 1971.

[15] *Cheng U. J.* Exhaustive construction of (255,127,63)-cyclic difference sets // Journal of Combinatorial Theory, Series A 33 (1983). — P. 115–125.

[16] *Dreier R. B., Smith K. W.* Exhaustive determination of (511,255,127) cyclic difference sets, manuscript. Manuscript, 1991.

[17] *Gaal P., Golomb S. W.* Exhaustive determination of (1023,511,255) cyclic difference sets // Mathematics of Computation 70 (2001). — P. 357–366.

[18] *No J. S., Golomb S. W., Gong G., Lee H. K., Gaal P.* A new family of binary pseudo-random sequences having optimal periodic correlation properties and larger linear span // IEEE Trans. Inform. Theory 35 (1998). — P. 371–379.

[19] *Gong G., Gaal P., Golomb S. W.* A suspected infinite class of cyclic Hadamard difference sets // Proceedings of 1997 IEEE Information Theory Workshop, July 6–12, 1997, Longyarbyen, Svalbard, Norway.

[20] *Segre B.* Ovals in finite projective plane // Canadian J. Math. 7 (1955). — P. 414–416.

[21] *Glynn D. G.* Two new sequences of ovals in finite Desarguesian planes of even order // Lecture Notes in Mathematics, Vol. 1036, Berlin, Springer-Verlag, 1983. — P. 217–229.

[22] *Maschietti A.* Difference sets and hyperovals // Designs, Codes and Cryptography 14 (1998). — P. 89–98.

[23] *Chang A. C. , Golomb S. W. , Gong G. , Kumar P. V.* On the linear span of ideal autocorrelation sequences arising from the Segre hyperoval // Sequences and their Applications — Proceedings of SETA'98, Discrete Mathematics and Theoretical Computer Science, London, Springer-Verlag, 1999.

[24] *No J. S., Chung H. B., Yun M. S.* Binary pseudorandom sequences of period 2^m-1 with ideal autocorrelation generated by the polynomial $z^d+(z+1)^d$ // IEEE Trans. Inform. Theory 44 (1998). — P. 1278–1282.

[25] *Dillon J. F.* Multiplicative difference sets via additive characters // Design, Codes and Cryptography, 17 (1999). — P. 225–236.

[26] *Dobbertin H.* Kasami power functions, permutation polynomials and cyclic difference sets // Difference sets, sequences and their correlation properties (Bad Windsheim, 1998), NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci. 542 (1999). — P. 133–158.

[27] *Gong G., Golomb S. W.* The decimation-Hadamard transform of two-level autocorrelation sequences // IEEE Trans. Inform. Theory 48 (2002). — P. 853–865.

[28] *Dillon J. F., Dobbertin H.* New cyclic difference sets with Singer parameters // Finite Fields and Their Applications, 10 (2004). — P. 342–389.

[29] *Leukhin A. N.* Algebraic solution of the synthesis problem for coded sequences // Quantum Electronics. — 2005. — V. 35, № 8. — P. 688–692.

[30] *Golomb S. W., Gong G.* Signal design for good correlation for wireless communication, cryptography, and radar. // Cambridge Univ. Press, 2006.

[31] Sequences, Subsequences, and Consequences // Int. Workshop, SSC 2007, Los Angeles, CA, USA, May 31 – June 2, 2007, Revised Invited Papers, Springer-Verlag 2007.

Регулярный метод синтеза бесконечного множества фазокодированных последовательностей с идеальной периодической автокорреляционной функцией*

Леухин А. Н., Парсаев Н. В., Тюкаев А. Ю., Корнилова Л. Г.
code@marstu.net
Йошкар-Ола, Марийский государственный технический университет

Разработан регулярный метод синтеза бесконечного множества фазокодированных последовательностей с нулевым уровнем боковых лепестков периодической автокорреляционной функции для случаев, когда длина последовательности является квадратом некоторого целого числа или числом, кратным четырем.

Введение

Разработка методов синтеза сложных сигналов, эффективных при решении задач распознавания, разрешения и оценки параметров, наряду с методами их обработки, является важнейшим направлением исследований в области радиотехники. Особый интерес среди синтезируемых кодовых последовательностей представляют фазокодированные последовательности (ФКП). Теория синтеза таких кодов достаточно развита, но далека от своего завершения. На сегодняшний день известно большое множество кодовых последовательностей с хорошими корреляционными свойствами, однако их количество является далеко не полным по сравнению с общих количеством всех возможных кодовых последовательностей заданной длины для фиксированного уровня боковых лепестков периодической автокорреляционной функции (АКФ).

Современное состояние теории синтеза фазокодированных последовательностей заданной длины не позволяет ответить на следующие вопросы:

— всегда ли существуют коды для фиксированного значения уровня боковых лепестков;
— если они существуют, то как определить их возможное число;
— если известно число возможных решений, то как, не конкретизируя метод кодирования, синтезировать сразу все возможные кодовые последовательности.

В первую очередь такое состояние исследуемого вопроса объясняется отсутствием решения фундаментальных проблем дискретной математики, связанных с теорией конечных полей. В работах [5, 6, 7] развивается обобщенный метод синтеза фазокодированных последовательностей, позволяющий получить все известные на сегодняшний день ФКП с одноуровневой АКФ, а также синтезировать большое количество новых ФКП с заданным уровнем боковых лепестков периодической АКФ. В данной работе разработан регулярный метод синтеза, позволяющий получить бесконечное множество ФКП с нулевым уровнем боковых лепестков периодической АКФ с длинами $N = 4z$ и $N = z^2$, где z — любое положительное целое число.

*Работа выполнена при финансовой поддержке по темам НИР в рамках гранта РФФИ №09-07-00072-а, гранта фонда HCF (договор №189), в рамках АВЦП «Развития научного потенциала высшей школы» мероприятия 1 (НИР №1.02.09), а также гос. контракта по программе У.М.Н.И.К. №6538р/9098.

Постановка задачи синтеза ФКП с одноуровневой периодической АКФ

Фазокодированную последовательность $\Gamma = \{\gamma_n\}_{n=0}^{N-1}$ определим на основании выражения:

$$\gamma_n = \exp(i\varphi_n), \quad n = 0, \ldots, N-1, \tag{1}$$

где значение фазы на каждом n-м кодовом интервале определяется из диапазона $\varphi_n \in [0; 2\pi]$, N — количество кодовых элементов в последовательности, i — мнимая единица. Из выражения (1) следует, что модуль каждого кодового элемента ФКП равен 1, т. е.

$$|\gamma_n| = 1, \ n = 0, \ldots, N-1. \tag{2}$$

Для исключения «повернутых» кодовых комбинаций угол нулевого вектора должен быть равен нулю, т. е. $\varphi_0 = 0^\circ$, при этом $\gamma_0 = 1$.

Периодическую АКФ определим из выражения:

$$r_\tau = \sum_{n=0}^{N-1} \gamma_{n+\tau \bmod N} \cdot \gamma_n^*, \quad n = 0, \ldots, N-1, \tag{3}$$

где γ_n^* — комплексно-сопряженный кодовый элемент.

Требуется определить вид кода $\Gamma = \{\gamma_n\}_{n=0}^{N-1}$, чтобы выполнялось условие:

$$r_0 = N, \quad r_1 = r_2 = \cdots = r_{N-1} = a.$$

Значение уровня боковых лепестков a может быть любым вещественным числом из диапазона $a \in [a_{\min}, a_{\max}]$, где верхняя граница диапазона может принимать значение $a_{\max} = N$, а глобальная нижняя граница $a_{\min} \geqslant N/(N-1)$ [7]. На основании выражений (2), (3) задача синтеза ФКП с одноуровневой периодической АКФ при условии $\varphi_0 = 0$ сводится к решению системы уравнений [5]:

— для четных N: $K = \frac{1}{2}N - 1,\ n = 1, \ldots, K$,

$$\begin{cases} \cos(\varphi_n) + \cos(\varphi_{N-n}) + \sum\limits_{m=1}^{N-n-1} \cos(\varphi_m - \varphi_{m+n}) + \\ \qquad + \sum\limits_{m=1}^{n-1} \cos(\varphi_m - \varphi_{m+N-1}) = a, \\ \cos(\varphi_K) + \sum\limits_{m=1}^{N-K-1} \cos(\varphi_m - \varphi_{m+K}) = \frac{1}{2}a, \\ \sin(\varphi_n) + \sin(\varphi_{N-n}) + \sum\limits_{m=1}^{N-n-1} \sin(\varphi_m - \varphi_{m+n}) + \\ \qquad + \sum\limits_{m=1}^{n-1} \sin(\varphi_m - \varphi_{m+N-1}) = 0; \end{cases} \tag{4}$$

— для нечетных N: $K = \frac{N-1}{2},\ n = 1, \ldots, K$,

$$\begin{cases} \cos(\varphi_n) + \cos(\varphi_{N-n}) + \sum\limits_{m=1}^{N-n-1} \cos(\varphi_m - \varphi_{m+n}) + \\ \qquad + \sum\limits_{m=1}^{n-1} \cos(\varphi_m - \varphi_{m+N-1}) = a, \\ \sin(\varphi_n) + \sin(\varphi_{N-n}) + \sum\limits_{m=1}^{N-n-1} \sin(\varphi_m - \varphi_{m+n}) + \\ \qquad + \sum\limits_{m=1}^{n-1} \sin(\varphi_m - \varphi_{m+N-1}) = 0. \end{cases} \tag{5}$$

В такой постановке задачи синтез ФКП с заданным уровнем боковых лепестков периодической АКФ сводится к поиску решений системы уравнений (4), (5), где неизвестными являются углы поворотов элементов кода $\varphi_1, \ldots, \varphi_{N-1}$.

ФКП с нулевым уровнем боковых лепестков периодической АКФ длины, кратной четырем

Анализ системы уравнений (4), (5) показал, что для длины кодовой последовательности, кратной четырем, при нулевом уровне боковых лепестков периодической АКФ существует бесконечное множество решений, задаваемых выражением:

$$\varphi_n = \alpha(n \bmod 2) + \frac{4\pi}{N} \sum_{s=0}^{n-1} \left[\frac{s+1}{2} \right], \tag{6}$$

где $n = 0, \ldots, N-1,\ N = 4z,\ z \in \mathbb{Z}$ — любое целое положительное число, $[x]$ — целая часть числа x, α — произвольное значение фазы, $\alpha \in [0; 2\pi]$.

Коды Чу [4], синтезированные на основе прямой дискретной аппроксимации закона линейной частотной модуляции при длине $N = 4z,\ z \in \mathbb{Z}$ — любое целое положительное число, являются частными случаями кодов, полученных на основании выражения (6).

ФКП с нулевым уровнем боковых лепестков периодической АКФ длины квадратных чисел

Анализ системы уравнений (4), (5) показал, что для длины кодовой последовательности, являющейся квадратным числом, при нулевом уровне боковых лепестков периодической АКФ существует бесконечное множество решений, задаваемых выражением:

$$\varphi_n = \alpha_{n \bmod z} + \frac{2\pi}{z} \left[\frac{n}{z} \right] (n-1), \tag{7}$$

где $n = 0, \ldots, N-1,\ N = z^2,\ z \in \mathbb{Z}$ — любое целое положительное число, $[x]$ — целая часть числа x, $A = \{\alpha_m\}_{m=0}^{z-1}$ — вектор фаз, принимающих произвольные значения $\alpha_0 = 0,\ \alpha_m \in [0; 2\pi]$, $m = 1, \ldots, z-1$.

Частными случаями фазакодированных последовательностей вида (7) с нулевым уровнем боковых лепестков периодической АКФ являются следующие известные коды: коды Фрэнка [1], коды класса p [1], коды Чу [4], а также коды, построенные на основании выражения (6), в том случае, если длина кода $N = 4z = x^2$, где x — положительные целые числа.

Заключение

В работе приводится аналитическое решение задачи синтеза бесконечного множества фазокодированных последовательностей с нулевым уровнем боковых лепестков периодической АКФ длины $N = 4z$ и $N = z^2$, где $z \in \mathbb{Z}$ — любое положительное целое число.

Показано, что коды Фрэнка, коды класса p, коды Чу являются частными случаями синтезированных последовательностей.

Литература

[1] *Варакин Л. Е.* Теория сложных сигналов. — М.: Сов. радио, 1970.

[2] *Свердлик М. Б.* Оптимальные дискретные сигналы. — М.: Сов. радио, 1975.

[3] *Гантмахер В. Е., Быстров Н. Е., Чеботарев Д. В.* Шумоподобные сигналы: анализ, синтез, обработка. — СПб.: Наука итехника, 2005.

[4] *Ипатов В. П.* Широкополосные системы и кодовое разделение сигналов. Принципы и приложения. — М.: Техносфера, 2007.

[5] *Leukhin A. N.* Algebraic solution of the synthesis problem for coded sequences // Quantum Electronics. — 2005. — V. 35, № 8. — P. 688–692.

[6] *А. Н. Леухин, А. Ю. Тюкаев, С. А. Бахтин, Л. Г. Корнилова* Новые фазокодированные последовательности с хорошими корреляционными характеристиками // Электромагнитные волны и электронные системы. — 2007. — № 6. — С. 51–54.

[7] *Леухин А. Н., Парсаев Н. В.* Синтез шумоподобных фазокодированных последовательностей // Ученые записки Казанского государственного университета. Серия физико-математические науки. — 2008. — Т. 150, кн. 2. — С. 38–50.

Ансамбли циклических симплексных последовательностей*

Леухин А. Н.
code@marstu.net
Йошкар-Ола, Марийский государственный технический университет

Рассмотрено решение задачи синтеза циклических симплексных фазокодированных последовательностей (ФКП) с максимально возможными попарными расстояниями между ними. Ансамбли синтезированных последовательностей являются оптимальными для решения задачи распознавания по критерию максимума расстояний между эталонами.

Введение

При решении задач распознавания широко используется критерий минимального расстояния, в соответствии с которым распознаваемый сигнал U будет отнесен к k-му классу, если расстояние между ним и эталоном A_k будет наименьшим среди всех расстояний до других эталонов ансамбля A_m, $m = 0, \ldots, M-1$, где M — объем ансамбля эталонов. В связи с этим важной является задача синтеза ансамбля эталонов, равноудаленных друг от друга на максимально возможное расстояние. Обозначим ансамбль эталонов как:

$$A^{\mathsf{T}} = \begin{bmatrix} A_0 & A_1 & \ldots & A_{M-1} \end{bmatrix}.$$

Пусть каждый m-й эталон ансамбля представляет собой фазокодированную последовательность длины N:

$$A_m = \Gamma^{(m)} = \left\{\gamma_n^{(m)}\right\}_{n=0}^{N-1} = \left\{\exp(i\varphi_n^{(m)})\right\}_{n=0}^{N-1}.$$

В этом случае расстояние между эталонами с номерами m и k можно определить на основании [1]:

$$R_{m,k}^2 = \left\|\Gamma^{(m)}\right\|^2 + \left\|\Gamma^{(k)}\right\|^2 - 2\,\mathrm{Re}\sum_{n=0}^{N-1}\gamma_n^{(m)}\gamma_n^{*(k)}, \quad (1)$$

где $\|\Gamma\|^2 = \sum\limits_{n=0}^{N-1}\gamma_n\gamma_n^*$ — норма эталона, γ_n^* — комплексно-сопряженный отчет.

Из выражения (1) следует, что максимальное расстояние между двумя эталонами будет в том случае, когда величина скалярного произведения двух эталонов будет минимальной, т. е.:

$$R_{m,k}^2 \to \max \quad \text{при } \mathrm{Re}\sum_{n=0}^{N-1}\gamma_n^{(m)}\gamma_n^{*(k)} \to \min. \quad (2)$$

Следуя выражению (2), для решения задачи распознавания широко используются ансамбли ортогональных эталонных ФКП, для которых справедливо условие $\sum\limits_{n=0}^{N-1}\gamma_n^{(m)}\gamma_n^{*(k)} = 0$. В этом случае расстояние между любыми двумя эталонами ансамбля определится, как $R_{m,k}^2 = 2N$. Ансамбли ортогональных последовательностей могут быть сформированы базисными функциями дискретного преобразования Фурье, системой функций Уолша, Адамара и т. д.

Однако, ансамбли симплексных эталонных сигналов будут иметь еще большее (максимально возможное) попарное расстояние между любыми двумя эталонами [2], так как для них справедливо соотношение:

$$\sum_{n=0}^{N-1}\gamma_n^{(m)}\gamma_n^{*(k)} = \frac{N}{1-N}, \quad \text{для всех } N \geqslant 2. \quad (3)$$

В этом случае расстояние между любыми двумя эталонами ансамбля определится как $R_{m,k}^2 = \frac{2N^2}{N-1}$.

Целью данной работы является задача синтеза ансамбля симплексных фазокодированных последовательностей, образующих циркулянтную матрицу вида:

$$A = \begin{pmatrix} \Gamma^{(0)} \\ \Gamma^{(1)} \\ \ldots \\ \Gamma^{(N-1)} \end{pmatrix} = \begin{pmatrix} \gamma_0 & \gamma_1 & \ldots & \gamma_{N-1} \\ \gamma_1 & \gamma_2 & \ldots & \gamma_0 \\ \ldots & \ldots & \ldots & \ldots \\ \gamma_{N-1} & \gamma_0 & \ldots & \gamma_{N-2} \end{pmatrix}. \quad (4)$$

Отметим, что объемы синтезируемых ансамблей являются максимально возможными и равными длине кодовой последовательности $M = N$.

Постановка задачи синтеза ансамблей циклических симплексных последовательностей

Выражение для одноуровневой циклической автокорреляционной функции (АКФ) фазокодированной последовательности $\Gamma = \{\gamma_n\}_{n=0}^{N-1}$ имеет вид:

$$\eta_\tau = \sum_{n=0}^{N-1}\gamma_{n+\tau \bmod (N)}\gamma_n^* = a, \quad \tau = 0, \ldots, N-1.$$

В работах [3, 4] было показано, что минимально возможный уровень отсчетов равномерной корреляционной функции может быть равен:

$$a_{\min} \geqslant \frac{N}{1-N}. \quad (5)$$

* Работа выполнена при финансовой поддержке по темам НИР в рамках гранта РФФИ №09-07-00072-а, гранта фонда HCF (договор №189), в рамках АВЦП «Развития научного потенциала высшей школы» мероприятия 1 (НИР №1.02.09).

Знак $>$ в неравенстве (5) означает, что не для любой размерности N существует ФКП с минимально возможным уровнем боковых лепестков $a_{\min} = \frac{N}{1-N}$. Например, минимальный уровень боковых лепестков $a_{\min} = -5/4 = -1{,}25$ для случая $N = 5$, но $a_{\min} = -1$ для $N = 3$.

Отсчеты m-го эталона в выражении (4) можно представить в виде:

$$\Gamma^{(m)} = \{\gamma_{n+m \bmod (N)}\}_{n=0}^{N-1}, \quad m = 0, \ldots, N-1; \quad (6)$$

тогда попарные скалярные произведения между эталонами с номерами m и k будут представлять собой отсчеты циклической АКФ при соответствующей нумерации:

$$\sum_{n=0}^{N-1} \gamma_n^{(m)} \gamma_n^{*(k)} = \sum_{n=0}^{N-1} \gamma_{n+\tau \bmod (N)} \gamma_n^*. \quad (7)$$

Принимая во внимание выражения (3) и (6), можно сделать вывод о том, что задача синтеза ансамбля циклических симплексных фазокодированных последовательностей сводится к задаче синтеза ФКП с одноуровневой циклической АКФ с минимально возможным уровнем боковых лепестков.

В данной работе будут получены аналитические выражения для ФКП с минимальным уровнем боковых лепестков, и на их основе синтезированы ансамбли циклических симплексных последовательностей.

ФКП с минимальным уровнем АКФ

Для синтеза ФКП с минимальным значением боковых лепестков одноуровневой АКФ можно использовать подход, разработанный в работах [3,4]. Значения фаз кодовых отсчетов в составе ФКП будем записывать в виде вектора:

$$\Psi = \begin{bmatrix} \varphi_0 = 0^\circ & \varphi_1 & \ldots & \varphi_{N-1}, \end{bmatrix} \quad (8)$$

и называть исходным решением задачи синтеза ФКП с одноуровневой АКФ. Нулевой отсчет $\gamma_0 = \exp(i0) = 1$.

Для произвольной длины N может существовать K решений, полученных в результате линейных преобразований некоторого исходного решения:

$$\Psi^{\intercal} = \begin{bmatrix} \Psi^{(0)} & \Psi^{(1)} & \ldots & \Psi^{(K-1)} \end{bmatrix}. \quad (9)$$

На основании исходного фазового вектора (8) в общем случае можно сформировать $K = N$ «автоморфных» решений вида:

$$\varphi_n^{(k)} = \varphi_{n+k \bmod (N)} - \varphi_k, \quad n = 0, \ldots, N-1, \quad k = 0, \ldots, N-1. \quad (10)$$

Также на основании исходного фазового вектора (8) можно сформировать еще $\varphi(N)$ «изоморфных» решений вида:

$$\varphi_n^{(k)} = \varphi_{n\lambda_k \bmod (N)}, \quad n = 0, \ldots, N-1, \quad k = 0, \ldots, \varphi(N)-1. \quad (11)$$

где λ_k — число, взаимно простое с N, $\varphi(N)$ — функция Эйлера от числа N. Кроме того, если существует решение (7), то должно существовать «сопряженное» ему решение вида:

$$\varphi_n^{(k+N)} = \varphi_k - \varphi_{n+k \bmod (N)}.$$

Таким образом, максимальное число возможных кодовых последовательностей (изоморфных, автоморфных и сопряженных решений), полученных на основе некоторой кодовой последовательности, определится как:

$$K = 2\varphi(N)N.$$

Примеры синтеза ансамблей циклических симплексных фазокодированных последовательностей

Приведем примеры результатов синтеза ФКП с минимальным уровнем боковых лепестков циклической АКФ для $N = 2, \ldots, 10$. Ансамбли могут быть представлены в виде циркулянтной матрицы (4), каждая строка которой образует эталон и является циклически сдвинутой исходной фазокодированной последовательностью.

Для $N = 2$, $a_{\min} = -2$ исходное решение является единственным ($K = 1$):

$$\Psi = \begin{bmatrix} 0 & \pi \end{bmatrix}.$$

На его основе может быть получен единственный ансамбль симплексных циклических ФКП:

$$A = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}.$$

Для $N = 3$, $a_{\min} = -1$ на основании единственного исходного решения:

$$\Psi = \begin{bmatrix} 0 & 0 & \pi \end{bmatrix}$$

можно сформировать 2 ансамбля ФКП с почти максимальным расстоянием:

$$A^{(0)} = \begin{bmatrix} 1 & 1 & -1 \\ 1 & -1 & 1 \\ -1 & 1 & 1 \end{bmatrix}, \; A^{(1)} = \begin{bmatrix} 1 & -1 & -1 \\ -1 & -1 & 1 \\ -1 & 1 & -1 \end{bmatrix}.$$

Отметим, что для данной длины ансамблей циклических симплексных ФКП не существует.

Для $N=4$, $a_{\min}=-\frac{4}{3}$ на основании исходного решения:

$$\Psi=\begin{bmatrix}0 & \pi & \pi+\varphi & \varphi\end{bmatrix},$$
$$\varphi=\arccos\left(\tfrac{1}{3}\right),$$

с помощью преобразований (9)–(11) можно получить всего $K=4$ различных решений:

$$\Psi=\begin{bmatrix}0 & \pi & \pi+\varphi & \varphi\\ 0 & \pi & \pi-\varphi & 2\pi-\varphi\\ 0 & \varphi & \pi+\varphi & \pi\\ 0 & 2\pi-\varphi & \pi-\varphi & \pi\end{bmatrix}.$$

Для каждого исходного решения можно сформировать ансамбль симплексных циклических ФКП.

Для $N=5$, $a_{\min}=-\frac{5}{4}$ на основании единственного исходного решения:

$$\Psi=\begin{bmatrix}0 & \varphi & 2\pi-\varphi & 2\pi-\varphi & \varphi\end{bmatrix},$$
$$\varphi=\pi-\arccos\left(\tfrac{1}{4}\right),$$

с помощью преобразований (9)–(11) можно получить всего $K=10$ различных решений. Для каждого исходного решения можно сформировать ансамбль симплексных циклических ФКП.

Для $N=6$, $a_{\min}=-\frac{6}{5}$ существует 2 исходных решения. Для первого — «базисного» исходного решения:

$$\Psi=\begin{bmatrix}0 & \frac{3\pi}{2} & \frac{3\pi}{2}-\varphi & \pi-\varphi & \frac{3\pi}{2}-\varphi & \frac{3\pi}{2}\end{bmatrix},$$
$$\varphi=\arccos\left(-\tfrac{3}{5}\right),$$

с помощью преобразований (9)–(11) можно получить всего K_1=6 различных решений. Для второго:

$$\Psi=\begin{bmatrix}0 & \varphi_1 & \varphi_2 & \varphi_3 & \varphi_4 & \varphi_5\end{bmatrix},$$
$$\varphi_1=\pi,\quad \varphi_2=2\pi-\arccos\left(-\tfrac{2+3\sqrt{6}}{10}\right),$$
$$\varphi_3=\arccos\left(\tfrac{1}{5}\right),\quad \varphi_4=\pi+\arccos\left(\tfrac{1}{5}\right),$$
$$\varphi_5=\arccos\left(\tfrac{2+3\sqrt{6}}{10}\right),$$

с помощью преобразований (9)–(11) можно получить всего $K_2=12$ решений. Общее количество ФКП размерности $N=6$ с минимальным уровнем боковых лепестков $a_{\min}=-\frac{6}{5}$ (и, соответственно, симплексных циклических ансамблей) будет равно $K=18$.

Для $N=7$, $a_{\min}=-1$ на основании исходного решения

$$\Psi=\begin{bmatrix}0 & 0 & 0 & \pi & \pi & 0 & \pi\end{bmatrix}$$

с помощью преобразований (9)–(11) можно получить всего $K=14$ различных решений.

Для $N=8$, $a_{\min}=-\frac{8}{7}$ существует 3 исходных решения. Для первого — «базисного» исходного решения:

$$\varphi_1=\varphi_7=\arccos\left(\tfrac{-2+3\sqrt{2}}{7}\right),$$
$$\varphi_2=\varphi_6=\arccos\left(-\tfrac{1}{7}\right)+\pi,$$
$$\varphi_3=\varphi_5=2\pi-(\varphi_1-\varphi_2),\quad \varphi_4=\arccos\left(-\tfrac{1}{7}\right),$$

с помощью преобразований (9)–(11) можно получить всего $K_1=32$ различных решения. Для второго исходного решения:

$$\varphi_3=2\pi-\arccos\left(\tfrac{2\sqrt{10+\sqrt{2}}\cdot\sqrt{3}(1+2\sqrt{2})+7-7\sqrt{2}}{49}\right),$$
$$\varphi_2=\varphi_3-\arccos\left(\tfrac{1-2\sqrt{2}}{7}\right),\quad \varphi_5=\arccos\left(\tfrac{1+4\sqrt{2}}{7}\right),$$
$$\varphi_1=\pi,\quad \varphi_4=\varphi_5+\pi,\quad \varphi_6=\varphi_3-\pi,\quad \varphi_7=\varphi_2-\pi,$$

с помощью преобразований (9)–(11) можно получить всего $K_1=32$ различных решения. Для третьего исходного решения:

$$\varphi_5=\arccos\left(\tfrac{1+2\sqrt{2}}{7}\right),\quad \varphi_4=\varphi_5+\pi,$$
$$\varphi_1=\varphi_4-\varphi_5+2\pi,\quad \varphi_2=\varphi_4-\varphi_5+\pi,$$
$$\varphi_3=\pi,\quad \varphi_6=\varphi_5+\pi,\quad \varphi_7=\varphi_4+\pi,$$

с помощью преобразований (9)–(11) можно получить всего $K_1=16$ различных решений. Таким образом, общее количество ФКП (циклических симплексных ансамблей) размерности $N=8$ с минимальным уровнем боковых лепестков $a_{\min}=-\frac{8}{7}$ равно $K=80$.

Для $N=9$, $a_{\min}=-\frac{9}{8}$ на основании единственного исходного решения:

$$t=\cos\left(\tfrac{1}{3}\arccos\left(\tfrac{\sqrt{21}}{7}\right)\right),$$
$$z=\frac{\sqrt{1764t^4-420\sqrt{21}t^3-1407t^2+230\sqrt{21}t+529}}{4\sqrt{112t^4-28\sqrt{21}t^3-84t^2+17\sqrt{21}+37}},$$
$$\varphi_1=\pi+\arccos(z),$$
$$\varphi_2=\arccos\left(\frac{\sqrt{84t^2-4\sqrt{21}t+1}}{2\sqrt{28t^4-21t^2-\sqrt{21}t+16}}\right),$$
$$\varphi_3=\arccos\left(-\tfrac{1}{8}\right),\quad \varphi_4=\varphi_3-\varphi_1-\varphi_2,\quad \varphi_5=\varphi_4,$$
$$\varphi_6=\varphi_3,\quad \varphi_7=\varphi_2,\quad \varphi_8=\varphi_1$$

с помощью преобразований (9)–(11) можно получить всего $K=54$ решения.

Для $N=10$, $a_{\min}=-\frac{10}{9}$ существует 6 исходных решений. Для первого исходного — «базисного» решения:

$$\varphi_1=\varphi_9=\varphi_2+\pi,\quad \varphi_2=\varphi_8=\arccos\left(-\tfrac{2}{3}\right),$$
$$\varphi_3=\varphi_7=\pi-\varphi_2,\quad \varphi_4=\varphi_6=2\pi-\varphi_2,\quad \varphi_5=\pi,$$

с помощью преобразований (9)–(11) можно получить всего $K_1=10$ решений. Для второго исходного — «базисного» решения:

$$\varphi_1\approx 149{,}359082^\circ,\quad \varphi_2\approx 171{,}698267^\circ,$$
$$\varphi_3=\varphi_2-\arccos\left(\tfrac{1}{4}\cos(\varphi_1+\varphi_2)+\tfrac{5}{36}\right)+\pi,$$
$$\varphi_4=\varphi_1+\varphi_2-\varphi_3,$$
$$\varphi_5=\varphi_1+\varphi_2,\quad \varphi_6=\varphi_4,$$
$$\varphi_7=\varphi_3,\quad \varphi_8=\varphi_2,\quad \varphi_9=\varphi_1$$

с помощью преобразований (9)–(11) можно получить всего $K_2 = 20$ решений. Для третьего исходного — «базисного» решения:

$$\varphi_1 \approx 149{,}043681^\circ, \quad \varphi_2 \approx 171{,}572709^\circ,$$
$$\varphi_3 = \varphi_2 - \arccos\left(\tfrac{1}{4}\cos(\varphi_1 + \varphi_2) + \tfrac{5}{36}\right) + \pi,$$
$$\varphi_4 = \varphi_1 + \varphi_2 - \varphi_3,$$
$$\varphi_5 = \varphi_1 + \varphi_2, \quad \varphi_6 = \varphi_4,$$
$$\varphi_7 = \varphi_3, \quad \varphi_8 = \varphi_2, \quad \varphi_9 = \varphi_1$$

с помощью преобразований (9)–(11) можно получить всего $K_3 = 20$ решений. Для четвертого исходного решения

$$\varphi_1 \approx 148{,}973278^\circ, \quad \varphi_2 \approx 171{,}544711^\circ,$$
$$\varphi_3 = \varphi_2 - \arccos\left(\tfrac{1}{4}\cos(\varphi_1 + \varphi_2) + \tfrac{5}{36}\right) + \pi,$$
$$\varphi_4 = \varphi_1 + \varphi_2 - \varphi_3,$$
$$\varphi_5 = \varphi_1 + \varphi_2, \quad \varphi_6 = \varphi_4,$$
$$\varphi_7 = \varphi_3, \quad \varphi_8 = \varphi_2, \quad \varphi_9 = \varphi_1$$

с помощью преобразований (9)–(11) можно получить всего $K_4 = 20$ решений. Для пятого исходного решения общего вида

$$\varphi_1 = \pi, \quad \varphi_2 \approx 22{,}03^\circ, \quad \varphi_3 \approx 65{,}352^\circ,$$
$$\varphi_4 \approx 314{,}98^\circ, \quad \varphi_5 \approx 307{,}745^\circ, \quad \varphi_6 = \varphi_5 - \pi,$$
$$\varphi_7 = \varphi_4 - \pi, \quad \varphi_8 = \varphi_3 + \pi, \quad \varphi_9 = \varphi_2 + \pi$$

с помощью преобразований (9)–(11) можно получить всего $K_5 = 40$ решений. Для шестого исходного решения общего вида

$$\varphi_1 = \pi, \varphi_2 \approx 14{,}994^\circ, \quad \varphi_3 \approx 67{,}287^\circ,$$
$$\varphi_4 \approx 159{,}467^\circ, \quad \varphi_5 \approx 126{,}272^\circ, \quad \varphi_6 = \varphi_5 + \pi,$$
$$\varphi_7 = \varphi_4 + \pi, \quad \varphi_8 = \varphi_3 + \pi, \quad \varphi_9 = \varphi_2 + \pi$$

с помощью преобразований (9)–(11) можно получить всего $K_6 = 40$ решений. Общее количество ФКП размерности $N = 10$ с минимальным уровнем боковых лепестков $a_{\min} = -\frac{10}{9}$ равно $K = 150$.

Заключение

Показан подход к решению задачи синтеза ансамблей циклических симплексных фазокодированных последовательностей равноудаленных на максимально возможное расстояние в унитарном пространстве. Показано, что в случае совпадения объёма алфавита M с размерностью кодовой последовательности N ($M = N$) задача сводится к задаче синтеза ФКП с минимальным возможным уровнем боковых лепестков одноуровневой АКФ. Определено значение $a_{\min}$ минимально возможного уровня боковых лепестков одноуровневой АКФ дискретной ФКП. Получены аналитические решения задачи синтеза ФКП, обладающих минимальным уровнем боковых лепестков. На основе каждой такой последовательности может быть сформирован ансамбль симплексных эталонов, оптимальных для решения задачи распознавания. Полученные ансамбли могут быть представлены в виде циркулянтной матрицы, каждая строка которой является эталоном и образовано циклически сдвинутой исходной фазокодированной псоледовательностью. В качестве примера приводятся результаты синтеза всех возможных ФКП с $a_{\min}$ для размерностей $N = 2, \ldots, 10$, на основании которых строятся циклические симплексные ансамбли.

Литература

[1] *Фурман Я. А. и др.* Введение в контурный анализ и его приложения к обработке изображений и сигналов. — М.: ФИЗМАТЛИТ, 2002. — 592 с.

[2] *Golomb S. W., Gong G.* Signal design for good correlation for wireless communication, cryptography, and radar. — Cambridge Univ. Press, 2006.

[3] *Leukhin A. N.* Algebraic solution of the synthesis problem for coded sequences // Quantum Electronics. — 2005. — V. 35, № 8. — Pp. 688–692.

[4] *Леухин А. Н., Парсаев Н. В.* Синтез шумоподобных фазокодированных последовательностей // Учёные записки Казанского государственного университета. Серия физико-математические науки.— 2008. — Т. 150, кн. 2. — С. 38–50.

Аппроксимация энтропии Колмогорова при анализе хаотических процессов на конечных выборках*

Манило Л. А., Немирко А. П.
lmanilo@yandex.ru, apn-bs@yandex.ru
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ»

Рассмотрен метод распознавания хаотических сигналов, заданных конечной выборкой данных. Он основан на использовании приближенной оценки энтропии Колмогорова и предполагает введение специальной коррекции, направленной на снижение влияния ограниченности длины анализируемых фрагментов. На примере анализа биосигналов показана эффективность применения данного подхода к распознаванию сигналов с хаотическими свойствами.

Исследования последних лет свидетельствуют о возрастании интереса к использованию при анализе сигналов различной природы, в том числе и биомедицинских сигналов, методов нелинейной динамики. Хаотические сигналы можно трактовать как процессы, генерируемые режимом динамического хаоса. Это представление позволяет существенно расширить спектр количественных критериев для диагностики состояний живого организма, используя совокупность характеристик для оценки детерминированного хаоса.

Оценка энтропии Колмогорова

Важнейшей характеристикой хаотического движения в фазовом пространстве произвольной размерности считается энтропия Колмогорова (K-энтропия) [1]. Она определяется как средняя скорость потери информации о состоянии динамической системы во времени. Предполагая, что d-мерное фазовое пространство разделено на ячейки размера l^d, а состояние системы $X(t)$ определяется дискретно через интервалы времени τ, можно вычислить K-энтропию

$$K = \lim_{\tau\to 0}\lim_{l\to 0}\lim_{N\to\infty}\frac{1}{N\tau}\sum_{n=0}^{N-1}(K_{n+1}-K_n) = -\lim_{\tau\to 0}\lim_{l\to 0}\lim_{N\to\infty}\frac{1}{N\tau}\sum_{i_0\ldots i_N}P_{i_0\ldots i_N}\ln P_{i_0\ldots i_N},$$

где $K_n = -\sum\limits_{i_0\ldots i_n} P_{i_0\ldots i_n}\ln P_{i_0\ldots i_n}$ — величина, пропорциональная количеству информации, необходимой для определения местоположения системы на заданной траектории $i_0^*\ldots i_n^*$ с точностью l; $P_{i_0\ldots i_n}$ — совместная вероятность того, что траектория $X(t=0)$ находится в ячейке i_0, $X(t=\tau)$ — в ячейке i_1, … , $X(t+n\tau)$ — в ячейке i_n; N — максимальное число оцениваемых точек.

Пределы $l\to 0$ и $N\to\infty$ делают величину K независимой от частного вида разбиения. Величина K считается мерой хаоса: она, как известно, равна нулю для регулярного движения, бесконечна для случайных систем, положительна и постоянна для систем с детерминированным хаосом [1].

В соответствии с теоремой Такенса можно восстановить некоторые свойства аттрактора в фазовом пространстве по временной последовательности одной из составляющих процесса. При этом, в частности, можно найти нижнюю оценку K-энтропии, определяемую в виде

$$K_2 = -\lim_{l\to 0}\lim_{n\to\infty}\frac{1}{n}\ln\frac{C_n(l)}{C_{n+1}(l)} \leqslant K,$$

где $C_n(l)$ — обобщенный корреляционный интеграл, значение которого может быть оценено для последовательности точек бесконечной длины.

Неравенство $K_2 > 0$ означает достаточное условие существования хаоса. Однако требование $N\to\infty$ делает невозможным применение данного критерия для значительного числа практических задач.

Вычисление аппроксимированной энтропии

Требование анализа последовательности отсчетов ограниченной длины N приводит к необходимости нахождения приближенной оценки K-энтропии, названной аппроксимированной энтропией (*Approximate Entropy* — ApEn) [2]. Рассмотрим методику нахождения этой оценки.

Пусть имеется выборка исходных данных $x(1)$, $x(2)$, … , $x(N)$, где N — длина выборки. Задаются значения двух параметров: m — длина анализируемых цепочек, r — величина порога, определяющего размеры ячеек фазового пространства и являющегося одновременно параметром фильтра шумов. Нахождение аппроксимированной энтропии можно представить в виде следующей многошаговой процедуры:

1. Формируют последовательности $X(1), \ldots, X(N-m+1)$, определяемые выражением:

$$X(i) = \big[x(i), x(i+1), \ldots, x(i+m-1)\big],$$

$i = 1, \ldots, N-m+1$.

2. Определяют расстояние между $X(i)$ и $X(j)$:

$$d\big[X(i), X(j)\big] = \max_{k=0,\ldots,(m-1)}\big[|x(i+k)-x(j+k)|\big].$$

*Работа выполнена при финансовой поддержке РФФИ, проекты №07-01-00569, №09-01-00501.

3. Вычисляют $C_r^m(i) = N^m(i)/(N-m+1)$, где $N^m(i)$ — количество пар $\big[X(i), X(j)\big]$, удовлетворяющих условию

$$d\big[X(i), X(j)\big] \leqslant r, \quad j = 1, \ldots, N-m+1.$$

4. Находят усредненное значение натурального логарифма $C_r^m(i)$:

$$\theta^m(r) = \frac{1}{N-m+1} \sum_{i=1}^{N-m+1} \ln C_r^m(i).$$

5. Увеличивая значение m на единицу, повторяют шаги 1 – 4 и находят значения $C_r^{m+1}(i)$, $\theta^{m+1}(r)$.

6. Находят оценку K-энтропии

$$\mathrm{ApEn}(m, r) = \lim_{N\to\infty} \big[\theta^m(r) - \theta^{m+1}(r)\big],$$

которая для выборки данных ограниченной длины принимает вид:

$$\mathrm{ApEn}(m, r, N) = \theta^m(r) - \theta^{m+1}(r).$$

Как видно из последнего выражения, вычисляемая оценка K-энтропии зависит от параметров m, r, N. В работе [3] предложен эффективный алгоритм вычисления этой оценки, а также на ее основе исследованы характеристики ритма сердца при значениях $m = 2$ и r от $0{,}1 \cdot SD_x$ до $0{,}25 \cdot SD_x$, где SD_x — стандартное отклонение исходной выборки данных. Естественно предположить, что более полное представление о свойствах процесса можно получить, анализируя ряд последовательных значений K-энтропии.

Выбор параметров распознавания

Сложность анализа последовательности значений ApEn, $m = 1, 2, 3, \ldots$ для выборки отсчетов конечной длины N связана с проявлением эффекта ложной регулярности процесса. С увеличением параметра m возрастает число одиночных цепочек, что приводит к уменьшению величины приращения энтропии. В результате все большее число редких событий, являющихся по природе своей «неизвестными», идентифицируются как детерминированные. Для исключения ложной регулярности, вносимой одиночными цепочками, предложена корректирующая оценка [4]:

$$\mathrm{ApEn}_{cor}(m) = \mathrm{ApEn}(m) + \mathrm{ApEn}(0)\frac{N_m^{(1)}}{N_{m+1}},$$

где N_{m+1} — число анализируемых цепочек длины $(m+1)$, $N_m^{(1)}$ — число лишь однажды встретившихся цепочек длиной m, $\mathrm{ApEn}(0)$ — значение абсолютной энтропии, вычисленное для исходной последовательности отсчетов.

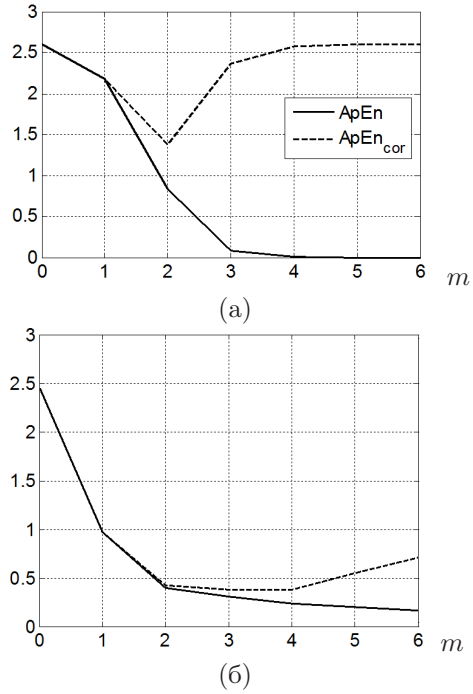


Рис. 1. Примеры оценок $\mathrm{ApEn}(m)$ и $\mathrm{ApEn}_{cor}(m)$ для модельных сигналов: (а) белого шума, (б) отображения Хенона.

В качестве примера на рис. 1 приведены зависимости $\mathrm{ApEn}(m)$ и $\mathrm{ApEn}_{cor}(m)$ для чисто случайного сигнала (а) и детерминированного хаотического процесса (б). На представленных графиках наблюдаются явные отличия полученных зависимостей. Введение коррекции в случае белого шума резко изменяет вид кривой, что позволяет в процессе анализа параметров $\mathrm{ApEn}_{cor}(m)$ выявить различия между стохастическими сигналами и процессами, генерируемыми нелинейными системами.

На основе результатов модельных экспериментов было предложено рассчитывать значения аппроксимированной энтропии при $r = 0{,}15 \cdot SD_x$, $m = 1, \ldots, 6$ и длине фрагментов $N = 300$. При этом нижнюю оценку K-энтропии можно аппроксимировать нижней оценкой $\mathrm{ApEn}_{cor}(m)$.

Анализ свойств аппроксимированной энтропии на выборке модельных сигналов, содержащих детерминированную, стохастическую и хаотическую составляющие, позволил сделать вывод о том, что для оценки степени регулярности изменений, наблюдаемых в дискретной последовательности отсчетов, могут быть использованы следующие параметры:

- значения $\mathrm{ApEn}(m)$ и $\mathrm{ApEn}_{cor}(m)$ при $m = 1, 2, 3$, где вклад одиночных цепочек незначителен;
- относительный минимум $\mathrm{ApEn}_{cor}(m)$: $\mathrm{ME} = \mathrm{ApEn}(0) - \min\limits_{m=1\ldots6}\{\mathrm{ApEn}_{cor}(m)\}$, который аппроксимирует нижнюю границу K-энтропии.

Решение задачи медицинской диагностики

Эти параметры могут эффективно использоваться для распознавания сигналов с хаотическими свойствами, например, при обнаружении фрагментов фибрилляции предсердий (ФП) в ходе слежения за динамикой сердечного ритма. В данной задаче решающие функции строились по оценкам аппроксимированной энтропии с использованием теории линейного дискриминантного анализа (критерий Фишера). Поскольку распознавание ФП предполагает обнаружение этой аритмии на фоне альтернативных нарушений ритма, рассмотрена двухклассовая задача. Первый класс составили сигналы ФП (класс ω_1), во второй класс были объединены реализации нормального ритма и частой экстрасистолией (класс ω_2). Таким образом, класс ω_1 был представлен сигналами с хаотическими свойствами, а ω_2 включал группу сигналов, в которых в случайной картине изменения ритма обнаруживаются скрытые закономерности. Классификация осуществлялась с использованием различного набора признаков, в качестве которых были выбраны параметры $\mathrm{ApEn}(m)$ и $\mathrm{ApEn}_{cor}(m)$.

В ходе экспериментов установлено, что наименьшая ошибка классификации (менее 1%) достигается при $r = 0{,}20 \cdot SD_x$ и использовании следующей совокупности признаков: ApEn(1), ApEn(2), ApEn(3), ME. Все расчеты были выполнены при длине фрагментов $N = 300$, что соответствует 4–5 минутам записи электрокардиосигнала. Кроме того, оценивалась средняя ошибка классификации при анализе более коротких фрагментов сигнала: $N = 150$, 100, 50 отсчетов. Уменьшение длины анализируемых фрагментов привело к возрастанию величины ошибки: 2%, 3,5% и 9,6%, соответственно. Это связано, в первую очередь, с возрастанием числа «неизвестных» цепочек, идентифицируемых как детерминированные.

Полученные результаты сравнивались с результатами обнаружения ФП, использующими альтернативные методы распознавания. Они основаны на проверке гипотезы о гауссовском законе распределения (оценке критерия χ^2), предсказании значений временного ряда авторегрессионной моделью, оценке относительного минимума уточненной условной энтропии. В этих методах используется лишь один информативный признак, являющийся индикатором приступов ФП. Результаты сравнительного анализа показали, что метод, основанный на использовании приближенной оценки K-энтропии, является наиболее надежным и обеспечивает высокое качество распознавания сигналов с хаотическими свойствами.

Выводы

Таким образом, предложен метод распознавания хаотических сигналов на конечных выборках, который основан на анализе параметров приближенной оценки энтропии Колмогорова. Показано, что введение коррекции расширяет число используемых признаков и обеспечивает надежное обнаружение хаотических изменений, порождаемых нелинейными системами.

Литература

[1] *Шустер Г.* Детерминированный хаос: Введение. — М.: Мир, 2007. — 240 с.

[2] *Pincus S. M.* Approximate entropy as a measure of system complexity // Proc. Natl. Acad. Sci. USA. — 1991. — V. 88, Pp. 2297–2301.

[3] Nonlinear Biomedical Signal Processing / Edited by Metin Akay. Volume 2, Dynamic Analysis and Modelling. — New York: IEEE, 2001. — 341 p.

[4] *Манило Л. А., Зозуля Е. П.* Автоматическое распознавание мерцательной аритмии с использованием оценок аппроксимированной энтропии // Информационно-управляющие системы. — 2006. — № 1 (20). — С. 21–27.

Об одном алгоритме распознавания движения на последовательности кадров

Матвеев Д. В.
diman@uniyar.ac.ru
Ярославль, ЯрГУ им. П. Г. Демидова

Предлагается алгоритм обнаружения в потоке зашумлённых кадров движущихся точечных объектов. С этой целью используется теория (r, R)-стратегий [1], предложенная А. Ю. Левиным для обработки точечных изображений. Модифицированный алгоритм позволяет устойчиво и достаточно быстро определять траектории движения при значительном уровне шума.

С бурным развитием компьютерной техники многие задачи по обработке изображений находят новые решения. Работа с большим объёмом данных влечёт за собой жёсткие требования к быстродействию алгоритмов. Большинство существующих алгоритмов обнаружения движения на видеопотоке построены по принципу межкадрового сравнения. Различные модификации данного метода дают неплохие результаты для обнаружения объектов, размеры которых сравнимы с размерами кадра. Построение охранных систем, систем контроля транспортного потока являются наиболее известными приложениями данных идей. В нашем случае рассматривается обнаружение траектории движущегося точечного объекта при наличии шумовой составляющей, визуально неотличающейся от объекта. В таких условиях условиях методы межкадровой разности уже не работают.

Основные факты (r, R)-стратегий

На входе имеем видеопоток (последовательность кадров), полученный с неподвижного регистрирующего источника. Предполагается наличие «белого» шума некоторой интенсивности и возможно равномерное движение точечного объекта. Ставится задача определения траектории движения объекта при различных уровнях шума. Предположим, что завязка (первый шаг) алгоритма выполнена, т. е. известны первые две точки $A_0(x_0, y_0)$, $A_1(x_1, y_1)$ последовательности $A_0, \ldots, A_n$, подозреваемой в том, что она с точностью до ошибок наблюдения является реальной траекторией движущегося объекта. Подтверждение последовательности будем производить следующим образом:

— строим прогноз положения точки на следующем кадре;
— вокруг полученного прогноза рассматриваем окрестность определённого радиуса;
— с помощью некоторого правила будем осуществлять селекцию попавших в окрестность точек и принимать решение о продолжении, либо об обрыве последовательности.

Очевидно, что в окрестность, кроме истинных, могут попадать и шумовые точки, образованные разного рода помехами. Поле помех, как отмечалось выше, будем предполагать пуассоновским с интенсивностью λ и белым в том смысле, что расположение и число помех на различных кадрах независимы. В результате анализа ситуаций в окрестности применяют одно из следующих правил:

1) при наличии в окрестности нескольких точек продолжать последовательности по каждой из них, т. е. допускать размножение последовательностей;
2) выбрать в окрестности одну точку, вероятность принадлежности которой к подтверждаемой последовательности наибольшая, а остальные отбросить как ложные (в нашем случае такой точкой будет ближайшая к центру круговой окрестности).

При отсутствии точек в окрестности в обоих случаях можно осуществлять обрыв последовательности. Это естественно при разумном выборе радиуса окрестности, если вероятность исчезновения сигнала (объекта) пренебрежимо мала. В дальнейшем будем рассматривать именно этот случай. В качестве естественного обобщения правил выбора 1) и 2) введём в рассмотрение следующий класс (r, R)-стратегий [1]. Пусть r, R — две константы, причём $0 < r < R < \infty$. На каждом шаге, начиная с $n = 2$, действуем по следующему правилу:

— если в окрестности радиуса r с центром в прогнозе $\widehat{A}_n$ есть точки, то каждая из них выбирается в качестве A_n;
— если таких точек нет в окрестности радиуса r, но они есть в окрестности радиуса R, то в качестве A_n выбирается ближайшая к $\widehat{A}_n$;
— если точек нет и в окрестности радиуса R, то последовательность обрывается.

В частности, правилу 1) в этих обозначениях отвечает (r, r)-стратегия, а правилу 2) — $(0, R)$-стратегия. Важной особенностью, которая появляется при $r > 0$, является то, что при наличии нескольких близких к $\widehat{A}_n$ точек, все они включаются в число возможных A_n, что должно уменьшить вероятность потери сигнала. Цена, которую приходится за это платить — возможность размножения последовательностей, так как в окрестности с центром $\widehat{A}_n$ могут оказаться несколько точек. Логически возможность неограниченного роста числа последовательностей всегда существует при $r > 0$. Од-

нако вероятностные соображения показывают, что при разумном выборе радиуса r данная реакция — затухающая (с вероятностью 1). При значительном уровне помех подавляющее число начальных пар точек алгоритма будет образовано ложными (шумовыми) точками. Важно, чтобы по мере дальнейшего продолжения эти последовательности, в отличие от реальных траекторий движения наблюдаемых объектов, достаточно быстро обрывались.

Пусть $A_0, \ldots, A_{(n-1)}$ — последовательность из ложных объектов. При использовании (r, R)-стратегии на определённом шаге (на n-м кадре) она может оборваться, а может и породить несколько последовательностей большей длины. Для постоянных r, R получаем однородный во времени ветвящийся марковский процесс размножения и гибели. Ключевую роль для динамики процесса играет параметр a — среднее число последовательностей, рождаемых одной последовательностью за один шаг. В силу наших предположений о поле помех $a = \pi r^2\lambda + \exp(-\pi r^2\lambda) + \exp(-\pi R^2\lambda)$, и при $a < 1$ число ложных последовательностей убывает со скоростью геометрической прогрессии. Но скорость обрыва ложных последовательностей — это не единственное требование к алгоритму, необходимо учесть и стремящуюся к нулю вероятность потери истинных (сигнальных) последовательностей. Как удаётся выяснить, при $\lambda\sigma^2 \ll 1$, где λ — интенсивность пуассоновского поля помех, σ^2 — дисперсия ошибки при измерении координат точек на кадрах видеопотока, удаётся удовлетворить обоим требованиям. Дальнейшие вычисления также показывают, что для (r, r)-стратегии основные параметры выглядят следующим образом: $a = a(r) = \pi r^2\lambda$, $p_n = p_n(r) = \exp\left(-\frac{r^2}{C_n\sigma^2}\right)$, где a — вероятность вырождения шумовых последовательностей, а p_n — вероятность потери сигнала на n-ом шаге, $C_n = \frac{2(n+2)(n+1)}{n(n-1)}$. Отсюда вытекает, что для нашего алгоритма необходимо выполнение условия $0 < r < r_{\max}$, где $r_{\max} = \frac{1}{\sqrt{\pi\lambda}}$, чтобы с вероятностью 1 происходило вырождение шумовых последовательностей.

Особенности реализации

В основе работы алгоритма лежат вышеописанные положения (r, r)-стратегии. На первом шаге мы рассматриваем два первых кадра, на которых выбираются все пары близких точек. Затем на третьем кадре строится прогноз положения точки $\widehat{A}_3$. Прогноз строим, линейно продолжая вектор расстояния между первой парой на такое же расстояние. Вокруг прогноза $\widehat{A}_3$ формируем окрестность радиуса r и все точки, попавшие в окрестность, рассматриваем в качестве A_3. В результате, у нас получились упорядоченные тройки точек, по координатам которых мы и будем вычислять прогноз четвёртой точки. Большинство реальных траекторий движения аппроксимируются многочленами не выше второй степени [2]. Поэтому прогноз будем строить из предположения, что искомая траектория имеет вид

$$x = x(t) = A_1t^2 + B_1t + C_1,$$
$$y = y(t) = A_2t^2 + B_2t + C_2,$$

где A_i, B_i, C_i — константы. Затем рассматриваем окрестность радиуса r и так далее.

При анализе алгоритма выяснился один немаловажный факт, непосредственно влияющий на эффективность. Оказалось, что точки, близко расположенные друг к другу, могут попадать в несколько окрестностей, и, таким образом, многократно порождать ложные последовательности. Чтобы от этого избавиться, предлагается действовать следующим образом: относить точки к ближайшей по евклидовому расстоянию траектории. Параметры траектории рассчитываются по трём предыдущим итерациям. Это значительно снижает количество порождаемых траекторий.

Основные результаты

Модифицированный алгоритм показывает очень хорошие результаты при обработке последовательности сильнозашумленных изображений. Время работы модифицированного алгоритма по сравнению с (r, R)-стратегией сократилось более чем в 10 раз. Верхнюю границу радиуса окрестности возможно уменьшить. Протестирована устойчивая работа метода при $r = \frac{2}{3}r_{\max}$. Тестирование производилось на компьютере с процессором Intel Core2Duo T7200 и объемом оперативной памяти 2ГБ. Обработка изображений размера 500×500 пикселей возможна почти в реальном времени. Уровень шума λ может достигать 0,1. Время работы алгоритма составило 4 секунды для последовательности из 1300 изображений размера 500×500 пикселей. Вопросы обнаружения движения возникают в разных областях, в том числе и в радиолокации. В данном случае рассматривается специфическая модель, где из видимого хаотичного движения точек на последовательности изображений (кадров) удается извлечь информацию о наличии траектории движения и оценке ее параметров. Примером может служить обработка снимков звездного неба для выявления траектории неразличимого глазом движения обьекта при большом количестве ложных, мерцающих и т. п. точек-обьектов.

Литература

[1] *Левин А. Ю., Малков А. Н.* (r, R) — стратегии обнаружения траектории цели в радиолокации // Вычислительные системы и их модели, Ярославль: ЯрГУ,1990. — С. 139–149.

[2] *Гонсалез Р., Вудс Р.* Цифровая обработка изображений— М.: Техносфера, 2005. — 1072 с.

Автоматическая аннотация изображений*

Мельниченко А. С.
melnichenkoalexandra@gmail.com
Таганрог, ТТИ ЮФУ

В предлагаемой работе рассматривается проблема автоматической аннотации изображений. Автоматическая аннотация изображений — это процесс автоматического присвоения системой метаданных в форме заголовка или ключевых слов цифровому изображению на основании только визуальной информации, содержащейся в изображении. Задача вводится и решается в предположении применения её результатов к проблеме поиска в больших коллекциях изображений. Приводится краткий обзор существующих на сегодняшний день направлений исследований в этой области, основные шаги предлагаемого метода решения. Делаются выводы и обозначаются направления дальнейших исследований.

Сегодня с развитием информационных технологий визуальная информация стала занимать не меньшую долю в общем количестве доступной информации, чем текстовая. Появились огромные базы данных изображений, как у отдельного пользователя, так и у различных организаций. Общее же количество изображений в Интернете вообще почти не поддается измерению. В этих условиях очень остро встает вопрос о навигации среди этой массы информации, о быстром и качественном поиске нужных пользователю данных. Но управление большими коллекциями изображений сильно отличается от управления коллекциями числовой или текстовой информации. Основной проблемой, затрудняющей эффективное и однозначное решение проблемы поиска в коллекции изображений, является так называемая проблема «*семантического разрыва*» — отсутствия однозначной связи между низкоуровневыми характеристиками и семантикой изображения. Из оцифрованных изображений система может извлечь набор их числовых характеристик: распределений цветов, текстур, границ. Человек же видит в изображении прежде всего его смысловое содержание, описывая его набором наиболее подходящих этому изображению слов. Проблема семантического разрыва полностью не преодолена в настоящее время, разные подходы пытаются решить её по-разному.

В настоящее время существует два основных подхода к поиску изображений.

Поиск на основе визуального образца (Query–by–Example, QbE). В этом случае в качестве запроса выступает изображение-образец (например, нарисованное от руки, сканированное или низкого качества), а система возвращает все визуально похожие на него изображения, обычно на основе сравнения низкоуровневых характеристик цветов, текстур и форм. Этот подход, таким образом, делает попытку уйти от проблемы семантического разрыва, заставляя пользователя формулировать запрос на языке изображений, а не слов. Поиск по визуальному подобию нашел применение в медицине, в коллекциях медицинских снимков. Такой вид поиска мало применим в Интернете из-за его неудобства для пользователя. Пользователю необходимо иметь образец изображения для поиска, в то время как он не всегда чётко представляет, как именно должно выглядеть искомое изображение в плане распределения цветов и текстур.

Поиск на основе текстового запроса (Query–by–Text, QbT). В этом случае в качестве запроса выступает набор ключевых слов. Поиск по текстовому запросу более удобен для пользователя и обычно применяется там, где более важна семантика искомого изображения. Можно сказать, что поиск по текстовому запросу делает попытку решить проблему семантического разрыва путем «перевода» понятий машинного пространства низкоуровневых понятий в семантическое пространство понятий человека. Для проведения такого поиска все изображения, хранящиеся в базе данных, должны иметь текстовые описания в виде набора ключевых слов (тэгов). Но ручное аннотирование столь больших объемов изображений — слишком трудоёмкий процесс. Поэтому в последнее время большое внимание стало уделяться техникам автоматического получения аннотаций и использования их для целей поиска изображений (*Annotation Based Image Retrieval — ABIR*).

С появлением методов автоматического аннотирования отпала необходимость в ручном аннотировании множества изображений, что позволяет включить в поиск большее количество картинок и улучшить его качество. Поэтому разработка эффективных методов автоматического получения ключевых слов для изображений является очень актуальной задачей для современных поисковых систем.

Обзор методов автоматической аннотации изображений

С осознанием того, что развитие техник машинного зрения пока не позволяет интерпретировать изображение как набор выделенных на изображении объектов и связей между ними, связан тот

*Работа выполнена при финансовой поддержке РФФИ, проекты № 08-07-00129 и № 07-07-00067.

факт, что подавляющее большинство существующих моделей автоматического аннотирования являются вероятностными моделями и нуждаются в обучении на некоторой выборке аннотированных человеком изображений.

Одним из первых исторически и самым простым из вероятностных методов является *Co-occurence Model* [1]. В этом подходе изображение с помощью регулярной сетки делится на равные прямоугольные части, для которых выделяют совместные цветовые и текстурные признаки, и кластеризуют все полученные векторы для всех изображений обучающей выборки. Каждый кластер соотносится со словом. Новое изображение, для которого нужно найти подходящие слова-аннотации, делится на части, для которых находятся ближайшие кластеры и, соответственно, слова.

Следующий по времени возникновения и сложности метод, который использует аналогию машинному переводу в применении к словам и частям изображений — *Translation Model* [2]. Главная идея здесь — составить «таблицу перевода» регионов изображения в ключевые слова. Изображение делится на нерегулярные регионы — «блобы», и вычисляются их низкоуровневые характеристики b: моменты инерции, средний цвет, цветовые вариации, текстурные признаки. Для нахождения «таблицы перевода» производится максимизация вероятности $P(w \mid b)$ условного распределения слов и блобов.

Дальнейшее усовершенствование вероятностных методов — появление моделей, пришедших в эту область из области текстового многоязычного поиска. Решив задачу поиска в двуязычной текстовой коллекции в работе «Cross–Lingual Relevance Model» [5], авторы Лавренко В. и др. перенесли используемую там модель на аннотирование и поиск изображений, назвав модель «*Cross-Media Relevance Model*» [4, 3]. В данном подходе на основании обучающей выборки строится словарь признаков блобов изображений. Строится вероятностная модель, которая позволяет предсказать на основе набора блобов нового изображения вероятность соответствия ему слова из словаря. Главное отличие метода от предыдущих состоит в том, что слова присваиваются всему изображению в целом, а не отдельным блобам. Авторы также представляют несколько моделей для поиска изображений на основе введённой модели релевантности. В работах Лавренко утверждается шестикратное улучшение точности предлагаемых моделей по сравнению с Co-occurence Model методом и двукратное по сравнению с Translation Model.

Еще один вероятностный подход для аннотирования изображений, задействованный в поисковой системе Behold [8], описывает в своей докторской диссертации Алексей Явлинский [6]. Он предлагает *метод непараметрической оценки функции совместного распределения* слов и блобов изображения. Для изображений рассматриваются представления в виде набора векторов действительных значений $x = (x_1, \ldots, x_d)$, $x_i \in \mathbb{R}$ или в виде сигнатур признаков изображения $s = \{(c_1, m_1), \ldots, (c_d, m_d)\}$, где c_i—центры, m_i—массы кластеров.

Существует также группа работ, направленных на аннотирование исключительно глобальных типов сцен (городские виды — природные ландшафты, панорама — макросъемка, съемка в помещении — наружная съемка) и довольно успешно решающих эту задачу. Обзор методов автоматической аннотации можно найти в [7].

Существующие методы можно разделить по типу используемых признаков изображений. Так, в работах, представляющих Co-occurence Model, Translation Model, Cross–Media Relevance Model, используется так называемое региональное представление изображений. В качестве признаков изображений здесь рассматриваются характеристики регионов изображения, выделенные с помощью некоторого алгоритма сегментирования или вручную. Так, последние две упомянутые модели тестировались на одной и той же базе предварительно сегментированных изображений, впервые введённой в [2], поэтому проблема выделения признаков не рассматривалась в соответствующих работах.

Другой способ представления признаков изображений — использование глобальных признаков, соответствующих всему изображению в целом. Такой подход используется Алексеем Явлинским при построении поисковой системы Behold. Сложность использования регионального представления заключается в том, что его построение требует высокоточной и в то же время быстрой сегментации изображения, что в настоящее время является не до конца решенной задачей. Для систем, работающих в реальном времени, выбор признаков является особенно важной задачей. Применяемые признаки должны хорошо описывать и различать изображения, легко вычисляться и занимать немного места в памяти.

Модель релевантности на основе глобальных признаков

Так как задачей, которую мы ставим перед собой в данной работе, является задача автоматического аннотирования изображений для целей поиска изображений в больших коллекциях, хотелось бы создать модель, позволяющую включать в словарь используемых ключевых слов как слова, характеризующие глобальные сцены, так и слова, обозначающие конкретные объекты. В то же время признаки должны достаточно легко и быстро вычисляться. В работах, посвященных глобально-

му аннотированию показано, что глобальные признаки хорошо работают для категоризации типов сцен, в то же время глобальные признаки использовались при построении системы Behold, работающей в Интернете, и показали достаточно хорошие результаты также и при аннотировании различных типов объектов. Поэтому в данной работе также было решено использовать для представления изображений глобальные признаки, которые позволили бы наилучшим образом производить аннотирование изображений различных типов объектов. В качестве модели для присвоения ключевых слов предлагается использование модели, близкой к описанной в работах Лавренко [4, 3], адаптированной для использования предлагаемых глобальных признаков.

Таким образом, предлагается использовать следующие признаки.

Цветовые гистограммы — это простейший признак, выделяющий особенности цветового распределения, полезный для различения сцен. Квантованием трёхмерного цветового пространства RGB на m ячеек по каждой размерности построим трёхмерную гистограмму, которую можно превратить в вектор признаков длиной $m^3 = m \times \times m \times m$. Будем обозначать его как RGB-m^3.

Однородность фона представляет собой одно число, показывающее долю в процентах числа точек изображения, имеющих близкие значения цвета, к общему числу точек изображения. Для вычисления этого числа изображение подвергается сглаживанию с помощью медианного фильтра с окном 5×5, затем для него строится гистограмма RGB-m^3 при $m = 32$, после чего искомое число получается делением наибольшего пика гистограммы на общее количество точек в изображении. Этот признак также может помочь в различении глобальных типов сцен. Обозначим этот тип признаков как Bg-1.

Текстурные признаки. Используются признаки текстуры, включающие важные для человеческого восприятия текстуры характеристики: грубость, контраст, направленность, которые были введены H. Tamura в работе [10]. В ряде работ показано, что текстурный признак Tamura хорошо согласуется с человеческим восприятием текстуры и может быть использован как при различении сцен, так и типов объектов. Для изображения вычисляется вектор признаков из 18 компонент, один из которых отвечает свойству грубости текстуры, один контрасту и остальные направленности. Будем обозначать эти признаки как Tamura-18.

SIFT (scale-invariant feature transform) — метод описания локальных признаков изображений, хорошо зарекомендовавший себя в области распознавания объектов. Впервые этот метод был предложен в [11] для распознавания достаточно большого количества видов объектов. Метод устойчив к изменению освещённости, поворотам, шуму. Построенный на основе SIFT вектор признаков длиной 132 элемента будем называть SIFT-132.

Для построения модели нам необходимо составить «словарь» из используемых признаков. Будем строить словарь признаков размером N, в котором присутствует некоторое количество N_k признаков каждого выбранного типа $k = 1, \ldots, 4$.

Для получения N_1 элементов словаря первого типа необходимо вычислить значения первого типа признаков для всех изображений обучающей выборки, разделить это множество на N_1 групп и взять за элемент словаря центр каждой такой группы. Очевидно, что для такого разделения необходимо применить некий алгоритм кластеризации векторов признаков. Но для обучающей выборки из нескольких тысяч изображений кластеризация такого количества векторов большой размерности, особенно для признаков SIFT-132 и RGB-m^3, является слишком трудоёмкой задачей. Поэтому это разбиение в данной работе производится тремя методами.

Для признаков Bg-1, представляющих собой просто числа, мы получаем разбиение на заданное количество N_2 кластеров с помощью алгоритма *k-means*. Эти кластеры служат для получения N_2 слов словаря признаков.

Для признаков RGB-m^3 при $m = 8$ получаются векторы размерности 512, которые, как и векторы признаков SIFT-132 мы будем кластеризовать *иерархическим методом*, используя в качестве промежуточного уровня вычисление признаков Bg-1. Для этого мы первоначально вычисляем для всех изображений выборки значения признаков Bg-1 и кластеризуем эти числа на некоторое не очень большое количество кластеров N'. Эти кластеры служат первоначальным разбиением для кластеризации признаков SIFT-132 и RGB-m^3. Количество итоговых кластеров признаков SIFT-132 и RGB-m^3 нельзя сказать заранее, так как количество кластеров, на которые разбивается каждый кластер, полученный после кластеризации признаков Bg-1, зависит от его размера. Ограничив максимальное число кластеров второго уровня, получаемых из одного кластера первого уровня неким числом a, можно оценить вклад признаков SIFT-132 в словарь как: $N' \leqslant N_1 \leqslant aN'$, аналогичное справедливо и для RGB-m^3.

Признаки Tamura-18 имеют относительно небольшой размер, поэтому их мы кластеризуем без использования промежуточной кластеризации с помощью алгоритма *k-medoids*, который достаточно эффективно работает для кластеризации многомерных векторов.

Введём обозначения: Q — множество изображений обучающей выборки; $I \in Q$ — изображение из обучающей выборки; $J \notin Q$ — новое неаннотированное изображение не из обучающей выборки, для которого мы хотим построить аннотацию; $W = \{w_1, \ldots, w_n\}$ — словарь ключевых слов, используемых для построения аннотаций.

Получив представление нового изображения J с помощью признаков b_i из словаря признаков $J = \{b_1, \ldots, b_s\}$, $s = 4$, мы хотим выбрать набор ключевых слов $\{w_1, \ldots, w_l\}$, которые наиболее адекватно описывают содержание изображения.

Предположим, что для каждого изображения J имеется вероятностное распределение $P(\cdot \,|\, J)$, назовём его *моделью релевантности* для J. Это распределение содержит все возможные слова и признаки из словарей выборки. Для присвоения изображению l ключевых слов $\{w_1, \ldots, w_l\}$ мы должны знать вероятности слов в этом распределении. Для этого мы должны оценить вероятности $P(w \,|\, J)$ для каждого ключевого слова из словаря:

$$P(w \,|\, I) = P(w \,|\, b_1, \ldots, b_s). \quad (1)$$

Для оценки совместного распределения слова w и признаков $b_1, \ldots, b_s$ на изображении используем обучающую выборку Q. Тогда совместное распределение может быть записано:

$$P(w, b_1, \ldots, b_s) = \sum_{I \in Q} P(I) P(w, b_1, \ldots, b_s \,|\, I). \quad (2)$$

Перепишем предыдущее уравнение, предполагая, что появление признаков $b_1, \ldots, b_s$ на изображении — события взаимно независимые:

$$P(w, b_1, \ldots, b_s) = \sum_{I \in Q} P(I) P(w \,|\, I) \prod_{i=1}^{s} P(b_i \,|\, I). \quad (3)$$

Вероятность появления слова w или признака b на изображении I определятся как:

$$P(w \,|\, I) = (1 - \alpha_I) \frac{N(w, I)}{|I|} + \alpha_I \frac{N(w, Q)}{|Q|}; \quad (4)$$

$$P(b \,|\, I) = (1 - \beta_I) \frac{N(b, I)}{|I|} + \beta_I \frac{N(b, Q)}{|Q|}; \quad (5)$$

где $N(w, I)$ обозначает число раз, которое слово w встречается в аннотации изображения I, $N(w, Q)$ — число раз, которое оно появляется во всех аннотациях обучающей выборки, $|I|$ — общее количество слов и признаков у изображения I, $|Q|$ — общее количество изображений в обучающей выборке. $N(b, I)$ и $N(b, Q)$ обозначают то же самое для признаков соответственно. Параметры сглаживания α_I и β_I выбираются с помощью обучающей выборки. С помощью уравнений (1)–(5) мы можем произвести аннотирование нового изображения J по распределению $P(w \,|\, J)$ выбором требуемого количества ключевых слов с наибольшими вероятностями. Построив аннотации для всех изображений большой коллекции, мы можем производить поиск по запросу из ключевых слов, возвращая в качестве результата изображения с наиболее пересекающимися с запросом аннотациями.

Выводы

Построенная модель, сочетающая использование наиболее дискриминативных признаков, отвечающих восприятию человеком изображений, с вероятностным подходом, хорошо зарекомендовавшим себя в области информационного поиска и, в частности, аннотации изображений, была программно реализована на языке Java с использованием библиотек Apache Lucene и Lire для хранения индексированных изображений. Вычислительный эксперимент проведен на изображениях базы [9]. Оценка полноты и точности полученных результатов входит в план дальнейших исследований.

Литература

[1] *Mori Y., Takahashi H., Oka R.* Image-to-word transformation based on dividing and vector quantizing images with words // Int. Conf. on Multimedia Intelligent Storage and Retrieval Management, 1999 — Pp. 111–115.

[2] *Duygulu P., Barnard K., and de Fretias N., Forsyth D.* Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary // European Conference on Computer Vision, 2002 — Pp. 97–112.

[3] *Jeon J., Lavrenko V., Manmatha R.* Automatic Image Annotation and Retrieval using Cross-Media Relevance Models // Int. Conf. of SIGIR, 2003 — Pp. 120–126.

[4] *Lavrenko V., Manmatha R., Jeon J.* A model for learning the semantics of pictures // 16th Conference on NIPS, 2003 — Pp. 42–46.

[5] *Lavrenko V., Choquette M., Croft W. B.* Cross-lingual relevance models // Int. Conf. of SIGIR on Research and development in information retrieval, 2002 — Pp. 97–112.

[6] *Yavlinsky A.* Image indexing and retrieval using automated annotation // PhD Thesis, University of London, 2007.

[7] *Hanbury A.* A survey of methods for image annotation // Journal of Visual Languages & Computing. — 2008. — Т. 19, № 5. — Pp. 617–627.

[8] `www.behold.cc` — Система для поиска изображений — 2006.

[9] `http://press.liacs.nl/mirflickr` — База изображений mirflickr08 — 2008.

[10] *Tamura H., Mori S., Yamawaki T.* Texture features corresponding to visual perception // IEEE Trans. On Sys. Man, and Cyb. — 1978. — Т. 8, № 6. — Pp. 460–473.

[11] *Lowe D.* Object recognition from local scale-invariant features // Int. Conf. on Computer Vision. — 1999. — Т. 2. — Pp. 1150–1157.

Поиск шаблонов перекрестков на векторной карте городской улично-дорожной сети*

Мехедов И. С.
mehedov@mail.ru
Москва, Вычислительный Центр РАН

В работе рассматривается подход к построению модели улично-дорожной сети векторной карты на основе скелета многоугольной фигуры. Модель дает возможность находить шаблоны перекрестков на карте, выделять области перекрестков и классифицировать перекрестки по конфигурации, а также представлять улично-дорожную сеть во фрагментированной форме (чередование перекрестков и линейных перегонов), служащей основой имитационного моделирования в транспортных геоинформационных системах.

Сегодняшние транспортные проблемы связаны с увеличением интенсивности дорожного движения и ростом спроса на транспортные услуги. Это приводит к дорожным инцидентам, пробкам, проблемам с окружающей средой в условиях современных мегаполисов, вследствие чего растет необходимость грамотного регулирования транспортных потоков и оптимизации городского движения. Одним из способов решения данной задачи является имитационное моделирование, основанное на реальном (или близком к реальному) представлении улично-дорожной сети (УДС). Таким представлением является векторная карта. Для удобства использования в моделях, УДС может быть разбита на множество фрагментов, соответствующих перекресткам и участкам дорог между перекрестками. Такое фрагментированное представление УДС называется базовой моделью улично-дорожной сети (БМДС). Каждый перегон между двумя перекрестками характеризуется, за исключением некоторых случаев, такими постоянными величинами, как: ширина проезжей части, число полос движения, средняя интенсивностью движения, что позволяет сопоставить множество перегонов и множество векторов в некотором пространстве параметров проезжей части. Перекрестки же характеризуются таким параметром, как конфигурация.

Статья посвящена задаче построения БМДС на основе скелетизации многоугольной фигуры и тесно связана с задачей поиска шаблонов перекрестков на карте. Входными данными метода является векторная карта УДС г. Москвы.

Постановка задачи

Рассмотрим всю УДС Москвы как множество многоугольных фигур, любые две из которых могут иметь общий участок границы (рис. 1а). Именно таким образом представлены векторные пространственные данные в информационном ресурсе «Единая государственная картографическая основа г. Москвы» [9]. Необходимо получить другое множество многоугольных фигур, составляющих

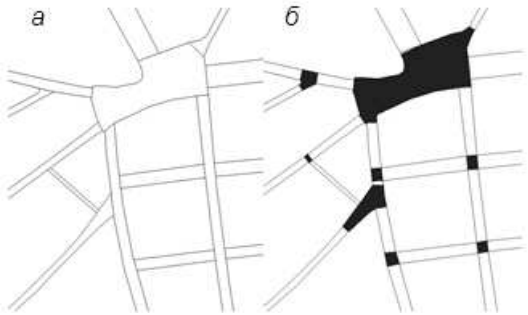


Рис. 1. Фрагмент улично-дорожной сети (слева) и базовой модели дорожной сети (справа). Черным цветом закрашены элементы сопряжения БМДС.

БМДС (рис. 1б), обладающее следующими свойствами:

- каждая многоугольная фигура из множества фигур БМДС соответствует либо перекрестку УДС (элемент сопряжения), либо участку между перекрестками УДС (линейная деталь);
- любые две различные линейные детали не имеют общих точек; любые два различных элемента сопряжения также не имеют общих точек; любой элемент сопряжения имеет общие участки границы как минимум с двумя различными линейными деталями;
- граница многосвязной многоугольной фигуры, являющаяся объединением элементов множества УДС, совпадает с границей многосвязной многоугольной фигуры, являющейся объединением элементов множества БМДС.

Далее необходимо на основе полученной фрагментации УДС классифицировать элементы сопряжения по их конфигурации, а также связать каждую линейную деталь с некоторым вектором значений параметров. Это дает атрибутивное описание БМДС для использования в имитационном моделировании.

Для построения БМДС достаточно найти все элементы сопряжения, в таком случае все оставшиеся многоугольные фигуры будут линейными деталями.

Построение БМДС будем проводить на основе линейной модели УДС (рис. 2). Под линейной мо-

*Работа выполнена при финансовой поддержке РФФИ, проект № 08-01-00670.

делью УДС понимается связный граф, топологически корректно отражающий конфигурацию УДС.

Каждый элемент сопряжения соответствует некоторому подграфу линейной модели. Таким образом, задача выделения элементов сопряжения сводится к задаче поиска шаблонов в линейной модели, где каждый шаблон соответствует определенному типу перекрестка.

Линейная модель УДС

Математической моделью, наиболее точной описывающей топологию многоугольной фигуры, является *скелет многоугольной фигуры*, т.е. множество всех центров максимальных пустых кругов фигуры [2], поэтому в статье рассматривается метод построения линейной модели УДС на основе скелета.

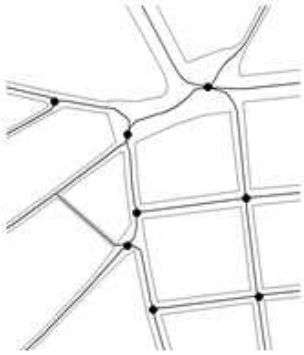


Рис. 2. Линейная модель фрагмента улично-дорожной сети. Точками отмечены вершины графа степени более двух.

Переход к линейной модели от совокупности площадных объектов является известной задачей геоинформатики. В известных работах по этой тематике также используется скелет многоугольной фигуры. Так, в [6, 7] линейная модель УДС строится на основе *спрямленного скелета (straight skeleton)*, подробно описанного в [5]. Недостатками такого скелета является неопределенность *радиальной функции* точки скелета, т. е. величину радиуса пустого круга с центром в этой точке [2]. В [8] используется скелет на основе диаграммы Вороного [2], но он плохо применим к сложным многоугольным фигурам с большим количеством «дыр». Вообще, большинство известных алгоритмов скелетизации хорошо работают с простыми многоугольниками, но либо «боятся» многосвязных многоугольных фигур, либо имеют большую вычислительную сложность.

Существуют два способа построения общего скелета улично-дорожной сети. Первый способ заключается в предварительном объединении всех многоугольных фигур, входящих в УДС, в одну сложную многосвязную фигуру. Координаты вершин такой фигуры точно описываю границу улично-дорожной сети. Вторым способом является скелетизация отдельно каждой многоугольной фигуры с последующим объединением отдельных скелетов в общий граф.

В статье используется первый способ, как наиболее простой, и алгоритм скелетизации, предложенный в [3], применимый к произвольным многоугольным фигурам и имеющий сложность $N \log N$, где N — число вершин фигуры.

Основой линейной модели служит не весь скелет, а его базовая часть, устойчивая к граничным шумам — *базовый скелет* [4].

Сам по себе базовый скелет еще не является моделью УДС. Модель УДС является результатом преобразования скелета к специальному виду. Задача поиска такого преобразования тесно связана с задачей поиска шаблонов перекрестков: с одной стороны, линейная модель порождает классификацию перекрестков по конфигурации, с другой — преобразование скелета к модели выполняется таким образом, чтобы максимально упростить поиск шаблонов.

Классификация перекрестков

Разделим все перекрестки на две большие группы — одноуровневые и многоуровневые (развязки). В группе одноуровневых перекрестков выделим три подгруппы: простые, сложные и комбинированные. К простым перекресткам относятся: Т-образный перекресток (рис. 3а вверху), крестообразный перекресток (рис. 3а внизу), площадь (рис. 3б). К сложным перекресткам относятся: сложный Т-образный (рис. 3в), сложный крестообразный (рис. 3г) и сложная площадь (рис. 3д). Среди комбинированных перекрестков будем по аналогии различать комбинированные Т-образный (рис. 3е), комбинированный крестообразный (рис. 3ж), комбинированную площадь (рис. 3з).

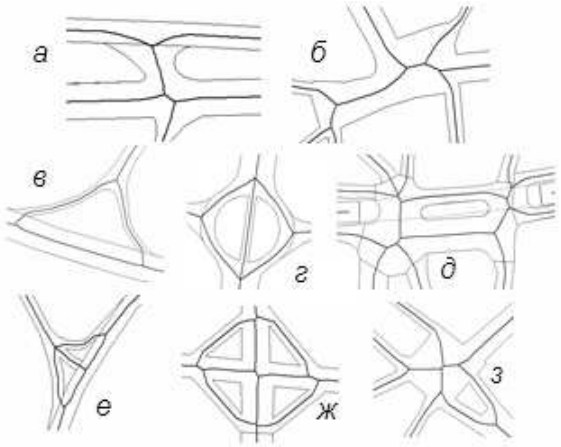


Рис. 3. Типы одноуровневых перекрестков.

Введем три определения.

Определение 1. *Будем говорить, что вершина P модели УДС порождает элемент сопряжения H_P, если H_P соответствует связный подграф модели УДС, содержащий P и не содержащий других вершин степени более двух.*

Определение 2. *Будем говорить, что цикл C модели УДС порождает элемент сопряжения H_C, если H_C соответствует связный подграф модели УДС, содержащий C не содержащий вершин степени более двух, не принадлежащих C, а также не содержащий других циклов, за исключением, быть может, внутри C.*

Определение 3. *Будем говорить, что вершина с циклами P_C модели УДС порождает элемент сопряжения H_{P_C}, если H_{P_C} соответствует связный подграф модели УДС, содержащий P_C, входящую в состав одного или нескольких циклов $C_1, C_2, \ldots, C_N$, и не содержащий вершин степени более двух, не принадлежащих $C_1, C_2, \ldots, C_N$.*

После объединения близких вершин базового скелета (рис. 4), мы получим модель, в которой:

— простые перекрестки порождаются вершинами степени более двух, причем Т-образный перекресток порождается вершиной степени 3, крестообразный — вершиной степени 4, площадь — вершиной степени 5 и более;

— сложные перекрестки порождаются циклами, содержащими по крайней мере две вершины степени более 2, причем сложный Т-образный перекресток порождается циклом, содержащим три вершины степени более 2, и остальные вершины, если они есть, степени 2; сложный крестообразный перекресток — четырьмя вершинами степени более 2; площадь — пятью или более вершинами степени 3, либо содержащим по крайней мере одну вершину степени более 3;

— комбинированные перекрестки порождаются вершинами степени более 3 с циклами, причем комбинированный Т-образный перекресток порождается вершиной степени 3 с циклами, комбинированный крестообразный — вершиной степени 4 с циклами, комбинированная площадь — вершиной степени 5 или более с циклами.

Особо подчеркнем, что любая вершина линейной модели степени более 2 либо порождает перекресток, либо входит состав порождающего элемента. Таким образом, линейные детали БМДС соответствуют *простым цепочкам* линейной модели (т. е. цепям графа, все внутренние вершины которых имеют степень 2, а концевые вершины — степень, отличную от 2).

Таким образом, в линейной модели можно выделить несколько типов неизоморфных подграфов — *шаблонов перекрестков*. Каждый шаблон характеризуется степенью порождающей вершины и (или) наличием и числом циклов.

В следующем разделе подробно опишем такие операции над базовым скелетом, как *объединение близких вершин*, *замена цикла вершиной*, *изъятие цикла*, приводящие к появлению шаблонов перекрестков в линейной модели.

Операции над базовым скелетом, приводящие к шаблонам линейной модели

Объединение близких вершин. Участки УДС, соответствующие перегонам между перекрестками имеют протяженность большую, чем ширину. В области перекрестков же, напротив, ширина УДС сравнима с длиной. На основе этого сформулируем понятие топологической близости вершин для решаемой задачи:

Определение 4. *Две вершины скелета степени более 2 являются топологически близкими в задаче поиска шаблонов перекрестков, если максимальные пустые круги скелета с центрами в этих вершинах, пересекаются.*

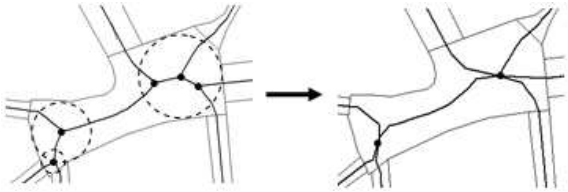


Рис. 4. Объединение близких вершин.

Такое понятие близости вершин, основанное на топологии скелета, формирует *задачу кластеризации вершин скелета на основе радиальной функции* (рис. 4) .

Алгоритм объединения вершин скелета можно найти в [1].

Из-за наличия «дыр» в многоугольной фигуре, описывающей УДС, в модели присутствуют циклы, порождающие сложные перекрестки, и содержащие вершины, порождающие комбинированные перекрестки. Однако, иногда требуется макромодель УДС, в которой любой перекресток соответствует ровно одной вершине модели степени более 2. Для получения такой макромодели УДС, необходимо произвести операции замены цикла вершиной и удаления циклов.

Замена цикла вершиной. Пусть H_C — одноименный сложный перекресток, порожденный циклом C. Заменой цикла на вершину P_C будем считать удалением всех вершин из базового скелета, входящих в C, введение в полученный граф новой вершины в геометрическом центре цикла и соединение ее с соответствующими висячими ребрами графа (рис. 5б). Вершина P_C в данном случае

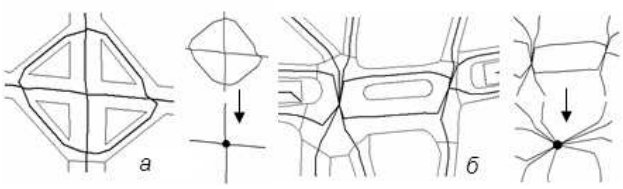


Рис. 5. Изъятие циклов (а) и замена цикла вершиной (б).

порождает простой перекресток макромодели того же типа, что и сложный.

Изъятие циклов. Пусть H_{P_C} — комбинированный перекресток, порожденный вершиной P_C с циклами. Изъятием цикла будем считать удаление всех простых цепочек , не содержащих P_C (рис. 5а). Вершина P_C после такого преобразования порождает простой перекресток макромодели того же типа, что и комбинированный.

Построение БМДС на основе линейной модели

Рассмотрим каждую вершину линейной модели степени более 2. Каждой простой цепочке, выходящей из нее, поставим в соответствие последовательность значений радиальной функции. Та вершина, начиная с которой значения радиальной функции соседних вершин мало отличаются друг от друга, индуцирует границу элемента сопряжения (рис. 6).

Рис. 6. Поиск границ элемента сопряжения, порожденного вершиной. Пунктиром показаны радиусы пустых кругов соседних точек одной из ветвей графа линейной модели.

Заключение

В настоящей работе описана линейная модель дорожной сети на основе скелета многоугольной фигуры. Дана классификация одноуровневых перекрестков на основе топологии линейной модели. Приведен способ построения базовой модели дорожной сети на основе поиска шаблонов перекрестков в линейной модели. Каждый шаблон является подграфом линейном модели определенной структуры, причем все шаблоны попарно неизоморфны.

Линейная модель позволяет проводить более детальную классификацию перекрестков и среди изоморфных шаблонов. Так, на (рис. 7) слева изображен перекресток с круговым движением, а справа — сложный Т-образный перекресток. Перекрестки имеют изоморфные шаблоны типа «сложный Т-образный перекресток», однако разную геометрическую конфигурацию циклов. Эта задача требует более тонкого анализа геометрии скелета и является предметом будущих исследований.

Также в дальнейшем планируется анализ топологии многоуровневых перекрестков (развязок) и генерация соответствующих шаблонов.

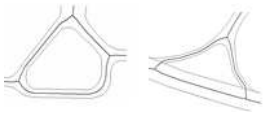


Рис. 7. Перекресток с круговым движением и сложный Т-образный перекресток: сравнение линейных моделей.

Литература

[1] *Домахина Л. Г, Охлопков А.* Изоморфные скелеты растровых изображений // Труды 18-й международной конференции Графикон-2008, Москва. http://www.graphicon.ru/2008/proceedings/Russian/SR4/Paper_3.pdf

[2] *Местецкий Л. М.* Непрерывная морфология бинарных изображений. Фигуры. Скелеты. Циркуляры // М.: ФИЗМАТЛИТ, 2009.

[3] *Местецкий Л. М.* Скелетизация многосвязной многоугольной фигуры на основе дерева смежности ее границы // Сибирский журнал вычислительной математики. — 2006. — Т. 9, № 3. — С. 299–314.

[4] *Метецкий Л. М., Рейер И. А.* Непрерывное скелетное представление изображения с контролируемой точностью // Труды 13-й международной конференции Графикон-2003, Москва. http://www.graphicon.ru/2003/Proceedings/Technical_ru/Reyer.pdf

[5] *Aichholzer O., Aurenhamer F.* Straight Skeletons for General Polygonal Figures in the Plane // Lecture Notes in Computer Science. — Springer-Verlag, 1996. — Vol. 1090. — Pp. 117–126.

[6] *Haunert J.-H., Sester M.* Area Collapse and Road Centerlines based on Straight Skeletons // Geoinformatica. — 2008. — Vol. 12, № 2. — Pp. 169–191.

[7] *Haunert J.-H., Sester M.* Using the Straight Skeleton for Generalization in Multiple Representation Environment // GICA Workshop on Generalization and Multiple Representation, 2004. http://www.ikg.uni-hannover.de/skalen/buendel/PDF/Skeleton.pdf

[8] *Roberts S., Brent Hall G., Boots B.* Street Centerlines Generation With an Approximated Area Voronoi Diagram // In Fisher P. F. (Ed.), Developments in spatial data handling: 11th International Symposium on Spatial Data Handling. — Springer-Verlag, 2007. — Pp. 435-446.

[9] www.egko.ru

Новая технология численного исследования динамических систем методами распознавания образов*

Неймарк Ю. И., Котельников И. В., Теклина Л. Г.
neymark@pmk.unn.ru
Научно-исследовательский институт прикладной математики и кибернетики
Нижегородского государственного университета им. Н. И. Лобачевского

В работе анализируются отличительные особенности новой технологии численного исследования конкретных многомерных и многопараметрических динамических систем методами распознавания образов и представлены основные этапы её алгоритмической реализации.

В работе [1] предложен новый подход к численному исследованию конкретных динамических объектов. Этот подход универсален, он основан на применении методов распознавания образов и статистического моделирования и имеет целью формализацию, а в дальнейшем и автоматизацию исследования прикладных систем, которое в настоящее время требует квалифицированной, длительной и трудоемкой работы даже для систем небольшой размерности и с малым числом параметров, а сравнительно полное исследование сложной многомерной системы практически невозможно. В настоящем докладе излагаются базовые идеи, положенные в основу развития новой технологии численного исследования динамических систем, и представлены главные этапы её алгоритмической реализации.

Базовые принципы новой методики численного исследования конкретных динамических систем

Развитие нового подхода потребовало коренной перестройки сложившихся взглядов на исследование динамических систем и отказа от традиционных методов. В основу новой технологии численного исследования динамических систем легли две базовые идеи.

Первая из них связана с объективно ограниченными возможностями численных методов, что делает невозможным точное численное исследование динамической системы, и неизбежно определённое огрубление в описании структуры её фазового портрета. Поэтому возникла идея о необходимости упрощения задачи исследования, но с получением результатов, достаточных для подавляющего большинства прикладных задач. Такое упрощение связано с *огрублённым* исследованием динамических систем, которое складывается из построения огрублённых фазовых портретов при заданных значениях параметров и огрублённого параметрического портрета исследуемой системы. Огрублённый фазовый портрет строится для заданной ограниченной области фазового пространства и включает в себя описание с заданной степенью статистической достоверности установившихся движений (аттракторов) и некоторых частей их областей притяжения, прилегающих к соответствующим аттракторам. Огрублённый параметрический портрет — характеристика изменения огрублённого фазового портрета в зависимости от значений параметров, упрощённый аналог бифуркационного портрета — строится для заданной ограниченной области в пространстве параметров и состоит из соответствующих различным фазовым портретам областей параметрического пространства, имеющих достаточно простую конфигурацию и отвечающих заданной степени статистической достоверности.

Вторая базовая идея связана с заменой традиционных классических методов исследования достоверным математическим экспериментом. Классические методы исследования в фазовом пространстве динамической системы с ростом его размерности сталкиваются с непреодолимыми трудностями даже при использовании самых современных вычислительных средств, а задача исследования системы в пространстве параметров вообще не имеет разработанных подходов. Что значит — исследовать динамическую систему? Это значит — выяснить, каковы её возможные движения, фазовые портреты, бифуркации. Изменяя начальные условия для построения траекторий при заданных значениях параметров, исследователь изучает движения системы и выясняет её фазовый портрет. Меняя значения параметров, исследователь наблюдает за изменением движений системы и тем самым изучает возможные её бифуркации. Таким образом, в основе исследования лежит эксперимент, состоящий в выборе некоторых значений параметров, в задании начальных условий для проведения эксперимента и в построении для него фазовой траектории. Результат серии таких экспериментов — множество траекторий, отвечающих различным начальным условиям и разным наборам параметров. Эффективным математическим аппаратом для обработки, анализа и отыскания скрытых закономерностей в полученной статистической выборке являются методы распознавания образов, работающие в пространстве большой размерности. Причем очевидно, что построение фазового порт-

*Работа выполнена при финансовой поддержке РФФИ, проект № 08-01-00248.

рета — это распознавание различного вида движений в фазовом пространстве системы на основе обучающей выборки данных, представляющих собой массив отрезков фазовых траекторий (конечные многомерные временные ряды), когда распознаваемыми образами являются установившиеся движения (аттракторы), а построение параметрического портрета динамической системы связано с решением задачи распознавания фазовых портретов в пространстве параметров системы на основе данных о числе и характере аттракторов (типе аттрактора и его локализации в фазовом пространстве) в зависимости от значений параметров. Главная отличительная особенность задач распознавания при исследовании динамических систем состоит в том, что это — задачи распознавания с активным экспериментом, т. е. обучающая выборка формируется исследователем в процессе решения поставленных задач.

Обратимся к алгоритмической реализации новой методики численного исследования конкретных динамических систем.

Основные этапы численного исследования динамических систем методами распознавания образов

Базой для исследования динамических систем методами распознавания являются решающие правила распознавания типа фазовой траектории DR0, построенные на основе анализа обширной статистической выборки, состоявшей из траекторий различных типов для разных и по своей природе, и по размерности фазового пространства динамических систем. Правила используют признаки глобального сжатия и локальной устойчивости движений в фазовом пространстве, они едины для всех систем и представлены в работах [2, 3]. Существующие правила нацелены на распознавание трёх типов фазовых траекторий, а именно: траекторий, стремящихся к состоянию равновесия, траекторий, стремящихся к предельному циклу, и траекторий, принадлежащих хаотическим или стохастическим аттракторам.

Численное исследование конкретной динамической системы распадается на три основных этапа.

Первый этап — предварительный анализ исследуемой модели с целью определения границ областей для исследования в фазовом и параметрическом пространствах системы, а также для выбора начальных данных при планировании эксперимента (длительность, шаг дискретизации, точность счета траектории). Этот этап представляет собой первое знакомство с динамической системой, он наименее формализован и в наибольшей степени опирается на конкретное содержание модели, но, подчеркнем, все получаемые при этом данные необходимы лишь в качестве начальных условий проведения исследования и в процессе решения могут изменяться в зависимости от оценки текущих результатов.

Второй этап исследования связан с построением огрублённых фазовых портретов, характеризующих поведение траекторий в фазовом пространстве системы при некоторых заданных значениях параметров. На этом этапе решаются три основные задачи:

- определение вида и числа устойчивых предельных подмножеств фазового пространства (аттракторов) в исследуемой динамической системе;
- описание и разделение аттракторов в фазовом пространстве;
- выделение областей притяжения для каждого из аттракторов.

Планирование эксперимента на этом этапе включает в себя выбор начальных точек в фазовом пространстве для построения траектории и её перечисленных выше характеристик.

Третий этап — исследование зависимости огрублённого фазового портрета от параметров, и представление результатов такого исследования в виде огрублённого параметрического портрета. Как было отмечено в [1], построение параметрического портрета можно ввести двумя путями: путём построения фазовых портретов и последующего их распознавания в пространстве параметров или путём отыскания всех видов аттракторов в исследуемой системе с последующим решением задачи распознавания аттракторов в пространстве параметров. В последнем случае по областям $O(J_s)$ в пространстве параметров, отвечающим различным видам аттракторов J_s, можно найти и области параметров $Q(R_s)$, соответствующих различным фазовым портретам R_s, в виде

$$Q(R_s) = \bigcap_{k=1}^{m} O(J_{s_k}),$$

если фазовый портрет $R_s = \bigcup\limits_{k=1}^{m} J_{s_k}$. Планирование эксперимента в параметрическом пространстве заключается в выборе значений параметров для построения огрублённого фазового портрета или проверки на наличие в фазовом портрете определенного вида аттракторов.

Какова связь между перечисленными этапами исследования? Первый этап независим, его цель — грубая оценка особенностей проведения эксперимента и получение данных, необходимых для запуска процесса исследования. Третий этап может быть осуществлен только при подключении второго, хотя и в сокращенном его варианте (например, без построения областей притяжения аттракторов). Второй этап исследования — это и необхо-

димая часть третьего, и совершенно независимая часть исследования при описании всех возможных режимов функционирования динамического объекта при определённых значениях параметров.

Весь процесс построения и фазовых, и параметрического портретов можно представить в виде последовательного решения ряда задач планирования и проведения эксперимента, формирования обучающих выборок, анализа, распознавания и классификации данных.

Алгоритмизация процесса численного исследования динамических систем

Как следствие ограничения объёма доклада, приведём упрощённое описание двух основных этапов процесса исследования, но с представлением всех главных решаемых проблем. По этой же причине лишь в одном варианте предложен алгоритм построения параметрического портрета — через отыскание установившихся движений, как наиболее приемлемый для целей прикладного исследования. Отдельно описывается краткая схема решения задач распознавания, как базовая для новой методики исследования динамических систем. Отметим, что все задачи распознавания (классификации) решаются в адаптивном режиме с использованием адаптивных алгоритмов анализа, классификации и распознавания данных на базе оптимальных тупиковых нечётких тестов [4, 5] и универсальной рекуррентной формы метода наименьших квадратов [6]. И, наконец, все результаты исследования должны отвечать заданной степени статистической достоверности $p_0 < 1$ сколь угодно близкой к единице.

Описание алгоритма решения задачи распознавания на основе представленной обучающей выборки (схема I).

1. Анализ обучающей выборки с целью оценки её представительности и достаточности для решения задачи распознавания. При положительном результате переход к п. 2, в противном случае — возврат в основную программу (п. 5, 7, 10 схемы II, п. 6 схемы III) с данными об областях дополнительного исследования (новая область может быть как частью предыдущей, так и выходить за её пределы).

2. Построение решающего правила распознавания.

3. Формирование контрольной выборки путём планирования и проведения эксперимента в области действия решающего правила. Проведение экзамена.

4. Анализ результатов экзамена на соответствие заданной степени статиститческой достоверности. Если соответствие есть, переход к следующей задаче в основной программе, в противном случае выполняется п. 5.

5. Пополнение обучающей выборки по результатам проведенного анализа. Коррекция решающего правила и переход к п. 3.

Описание алгоритма построения огрублённого фазового портрета при заданных значениях параметров (схема II).

1. Планирование и проведение эксперимента в заданной области фазового пространства. Построение фазовой траектории.

2. Анализ траектории с целью формирования информативных признаков, отражающих устойчивость траектории и особенности её поведения при приближении к аттрактору.

3. Определение типа фазовой траектории согласно известным решающим правилам DR0. Определение типа аттрактора, к которому стремится данная траектория, и его локализации в фазовом пространстве. Формирование выборки данных об аттракторах в виде: тип аттрактора; часть траектории из малой окрестности аттрактора TR1; часть траектории, находящаяся вне области аттрактора TR2. Сбор данных о количестве аттракторов и частоте появления новых аттракторов.

4. Анализ выборки на полноту представленных в ней аттракторов в соответствии с заданной степенью статистической достоверности. Если соответствие есть, выполняется п. 5, в противном случае — переход к п. 1, при этом по результатам анализа возможно изменение (уточнение) области планирования и проведения эксперимента.

5. Формирование обучающих выборок по каждому из типов аттракторов в виде множеств TR1 принадлежащих им траекторий. По результатам анализа в п. 1 схемы I возможно расширение выборки за счёт повторения действий 1–3 в указанной области для дополнительного исследования.

6. Определение числа аттракторов одного типа (по каждому из типов) путём решения задачи классификации согласно схеме I.

7. Формирование обучающей выборки для каждого из аттракторов в виде множеств TR1 принадлежащих им траекторий с целью описания аттракторов и разделения их в фазовом пространстве. По результатам анализа в п. 1 схемы I возможно расширение выборки за счет повторения действий 1–3 в указанной области для дополнительного исследования.

8. Описание каждого из аттракторов путём решения задачи распознавания всех выделенных аттракторов в фазовом пространстве системы согласно схеме I.

9. Построение правила для принятия решения о принадлежности произвольной траектории к одному из аттракторов.

10. Формирование обучающей выборки для каждого из аттракторов в виде множеств TR2 принад-

лежащих им траекторий с целью выделения их областей притяжения. По результатам анализа в п. 1 схемы I возможно расширение выборки за счет повторения действий 1–3 в указанной области для дополнительного исследования.

11. Выделение областей притяжения для каждого из аттракторов путём решения задачи разделения множеств траекторий, относящихся к различным аттракторам, в фазовом пространстве системы согласно схеме I.

12. Формирование фазового портрета как совокупности построенных аттракторов и их областей притяжения.

Описание алгоритма построения огрублённого параметрического портрета (схема III).

1. Планирование эксперимента в заданной области пространства параметров.

2. Построение огрублённого фазового портрета согласно схеме II в её сокращенном варианте (п. 1–8).

3. Анализ фазового портрета по данным о типе и количестве аттракторов и их расположении в фазовом пространстве. Кодирование аттракторов с целью автоматизации сравнительного анализа. Формирование выборки данных об аттракторах исследуемой системы в виде: характеристики аттрактора (код) и отвечающих ему значения параметров. Сбор данных о количестве аттракторов и частоте появления новых аттракторов.

4. Анализ выборки на полноту представленных в ней аттракторов для исследуемой области в соответствии с заданной степенью статистической достоверности. Если соответствие есть, выполняется п. 5, в противном случае — переход к п. 1, при этом по результатам анализа возможно изменение (уточнение) области планирования и проведения эксперимента.

5. Формирование первичной обучающей выборки по каждому из выделенных аттракторов в виде множеств отвечающих им наборов параметров (в частности, множество может состоять и из одного объекта).

6. Выбор точек отсчета и направлений в пространстве параметров для проведения эксперимента с целью уточнения границ областей существования для каждого из распознаваемых аттракторов.

7. Расширение обучающей выборки путём проведения для каждой из точек отсчета по всем выбранным направлениям следующих действий:

7.1) планирование эксперимента в пространстве параметров;

7.2) планирование и проведение эксперимента в фазовом пространстве с целью проверки наличия в фазовом портрете исследуемого типа аттрактора;

7.3) при наличии аттрактора обучающая выборка пополняется новым объектом и выполняется п. 7. 1), при отсутствии аттрактора происходит переход к другому направлению или новой точке отсчёта (с повторением для них действий 7. 1)–7. 3).

8. Описание областей в пространстве параметров, отвечающих каждому из выделеных аттракторов, путём решения задачи распознавания установившихся движений в пространстве параметров согласно схеме I.

9. Предварительное определение возможных огрублённых фазовых портретов и областей их существования в пространстве параметров с помощью выделенных аттракторов.

10. Анализ предварительных данных о параметрическом портрете на соответствие заданной степени статистической достоверности путём выполнения действий 3–5 схемы I.

11. Формирование параметрического портрета как совокупности областей параметрического пространства, соответствующих всем выделенным огрублённым фазовым портретам.

Литература

[1] *Neimark Yu. I., Kotel'nikov I. V., Teklina L. G.* Numerical research of the dynamic systems as the pattern recognition problem // Pattern Recognition and Image Analysis: New Information Technologies. Proceedings of the 8-th International Conference (PRIA–9–2008). Nizhni Novgorod, 2008. — Pp. 96–99.

[2] *Котельников И. В.* Синдромальные процедуры распознавания для исследования фазового пространства конкретных многомерных динамических систем // ММРО–13, М.: МАКС Пресс, 2007. — С. 146–149.

[3] *Неймарк Ю. И., Теклина Л. Г.* Анализ фазовых траекторий многомерных динамических систем методами распознавания на основе одномерных временных рядов // ММРО–13, М.: МАКС Пресс, 2007. — С. 191–193.

[4] *Kotel'nikov I. V.* A Syndrome Recognition Method Based on Optimal Irreducible Fuzzy Tests // Pattern Recognition and Image Analysis. — 2001. — V. 11, № 3. — Pp. 553–559.

[5] *Kotel'nikov I. V.* Cluster Analysis of Multidimensional Objects Based on Optimal Irreducible Fuzzy Tests and Syndromes // Pattern Recognition and Image Analysis. — 2004. — V. 14, № 3. — Pp. 361–369.

[6] *Неймарк Ю. И., Теклина Л. Г.* Новые технологии применения метода наименьших квадратов. — Нижний Новгород: Изд. Нижегородского госуниверситета, 2003. — 196 с.

О возможностях изучения хаотических движений в конкретных динамических системах методами распознавания образов и математического моделирования*

Неймарк Ю. И., Таранова Н. Н., Теклина Л. Г.
neymark@pmk.unn.ru
Научно-исследовательский институт прикладной математики и кибернетики
Нижегородского государственного университета им. Н. И. Лобачевского

Для количественного описания результатов вычислительного эксперимента при изучении хаотических и стохастических движений в многомерных динамических системах с большим числом параметров предлагается использовать методы распознавания образов и математического моделирования, которые имеют универсальный характер, легко формализуемы и позволяют преодолеть проблему размерности.

При качественном исследовании динамических систем наряду с основной задачей — выявлением всех видов установившихся движений и возможных бифуркаций в системе — представляет интерес изучение и самих установившихся движений, в особенности хаотических и стохастических аттракторов. Наиболее известными методами исследования хаотических движений являются метод точечных отображений, вычисление показателей Ляпунова, определение размерности и энтропии стохастического аттрактора, а также количественное описание движений с помощью таких статистических характеристик, как распределение вероятностей, корреляционная функция, спектральная плотность. Причем такие исследования для систем большой размерности представляют собой весьма непростую задачу.

В качестве универсального, легко формализуемого и преодолевающего проблему размерности подхода к изучению хаотических и стохастических движений предлагается использовать методы распознавания образов, доказавшие свою эффективность при огрубленном численном исследовании конкретных динамических систем.

1. Распознавание и описание странных аттракторов методами распознавания образов

Огрубленное исследование конкретных многомерных и многопараметрических динамических систем методами распознавания образов и статистического моделирования дает возможность распознать хаотические движения, выделить хаотические и стохастические аттракторы и описать их расположение в фазовом пространстве с помощью набора параллелепипедов, сфер или эллипсоидов, а также отыскать и хотя бы частично с заданной степенью статистической достоверности описать области притяжения выделенных странных аттракторов [1].

Подчеркнем, что распознавание странных аттракторов и дискриминантный анализ по отношению к другим видам установившихся движений (устойчивые состояния равновесия и периодические движения) проводится на основе признаков глобального сжатия и локальной устойчивости (состояния равновесия, циклы) — неустойчивости (хаос) для временных интервалов, ограниченных некоторой величиной $T_{\max}$, то есть, речь идет о распознавании типов установившихся движений в пределах введенных для исследуемой системы понятий малых и больших пространственно-временных величин.

Выделение и описание областей хаотических движений — это именно тот результат, который является основой для дальнейшего изучения хаотических и стохастических движений в каждом из выделенных аттракторов и их эволюции с изменением параметров системы. При этом следует отметить, что все последующие выводы основываются на данных вычислительных экспериментов в предположении, что вычислительный эксперимент по изучению поведения хаотических и стохастических траекторий достаточно адекватно отражает особенности реального эксперимента.

2. Моделирование хаотических движений с помощью однородной марковской цепи

Выделенную область хаотического аттрактора H всегда можно покрыть множеством сфер S_k, $k = 1, \ldots, N$, одного радиуса так, что

$$H \subseteq \bigcup_{k=1}^{N} S_k.$$

Выбираем радиус, а тем самым и количество сфер N так, чтобы при минимальном N с заданной статистической достоверностью p_0 все точки из S_k принадлежали либо выделенному аттрактору, либо его области притяжения. Построенные сферы позволяют смоделировать хаотическое движение как результат вычислительного эксперимента в изучаемом аттракторе, в виде однородной цепи Маркова

*Работа выполнена при финансовой поддержке РФФИ, проект № 08-01-00248.

с конечным числом состояний $A_1, \ldots, A_N$:

$$A_1 = S_1, \;\; A_k = S_k \setminus \left(\bigcup_{i=1}^{k-1} S_i \cap S_k \right), \;\; k = 2, \ldots, N.$$

Главная характеристика марковской цепи — стохастическая матрица P вероятностей перехода p_{ij} из состояния A_i в состояние A_j за один шаг. Такую матрицу, аппроксимирующую результаты численного эксперимента, легко построить для исследуемого аттрактора. По этой матрице можно судить о существовании предельных переходных вероятностей, а тем самым и предельных абсолютных вероятностей нахождения системы в каждом из N состояний после n шагов, когда $n \to \infty$, а также найти эти предельные вероятности и оценить скорость сходимости [2, 3]. Если стационарные вероятности пребывания в каждом из N состояний существуют, то эта важная характеристика странного аттрактора позволяет еще и проанализировать степень его неоднородности.

Дополнительные данные о странном аттракторе можно получить путем отыскания таких статистических характеристик марковской цепи, как среднее (и, соответственно, дисперсия) время возвращения в состояние A_i, $i = 1, \ldots, N$, среднее время обхода всех состояний и др. Все эти характеристики поддаются качественной интерпретации, исходя из конкретного содержания исследуемой системы, и могут быть интересны не только для математиков, но и для специалистов в конкретной области знаний.

3. Исследование хаотических движений с помощью одномерных временных рядов

В результате исследований была установлена возможность изучения движений в фазовом пространстве системы с помощью описания фазовых траекторий $\boldsymbol{x}(t)$ — многомерных временных рядов

$$\boldsymbol{X} = \{\boldsymbol{x}_i = \boldsymbol{x}(t_0 + i\Delta t) \colon i = 0, \ldots, N\},$$

где $\boldsymbol{x}_i = (x_1^i, \ldots, x_n^i)$, одномерными временными рядами

$$\boldsymbol{Y} = \{y_i \colon i = 1, \ldots, N\},$$

представляющими собой изменение со временем расстояний между соседними точками исходного временного ряда, то есть

$$y_i = \sqrt{\sum_{j=1}^{n} (x_j^{i+1} - x_j^i)^2}.$$

Именно такое описание позволило решить задачу распознавания типа фазовых траекторий, а тем самым и выделения хаотических и стохастических аттракторов. Рассматривая временной ряд $\boldsymbol{Y}$ как реализацию случайного процесса, можно построить его количественные характеристики.

3.1. Авторегрессионная модель. Для обнаружения скрытых закономерностей следования значений временного ряда была выбрана модель авторегрессии, во-первых, как доступный и эффективный способ описания и анализа временных рядов и, во-вторых, как аппарат, успешно реализуемый с использованием универсальной рекуррентной формы метода наименьших квадратов (МНК) [4], обладающей широкими адаптивными возможностями.

В качестве наиболее адекватной была выбрана линейная модель скользящей авторегрессии (с анализом последовательности пересекающихся интервалов) с автоматическим выбором порядка авторегрессии и длины анализируемых интервалов для каждого из временных рядов с помощью МНК. Анализ авторегрессионных моделей дал следующие результаты:

1) решение проблемы определения длительности переходного периода до попадания в зону аттрактора, когда при $t > t^*$ характеристики ряда приобретают четкие специфические свойства;
2) выявление закономерного характера изменений в динамике коэффициентов авторегрессии и корней характеристического полинома для траекторий, стремящихся к предельному циклу, а именно: коэффициенты авторегрессии приближаются к некоторым установившимся значениям, а максимальный по модулю корень характеристического уравнения стремится к единице при соответствующем выборе длины анализируемого интервала;
3) для хаотических движений коэффициенты авторегрессии изменяются во времени, а максимальный по модулю корень характеристического уравнения может принимать значения меньшие, равные или большие единицы.

Отыскание закономерностей изменения во времени корней характеристического полинома и (или) коэффициентов авторегрессии — это выявление и описание закономерностей движения в хаотическом или стохастическом аттракторе.

3.2. Спектральный и корреляционный анализ. В дополнение к классическому изучению результатов спектрального и корреляционного анализа временных рядов $\boldsymbol{Y}$ для различного вида хаотических движений, продемонстрировавших, что они являются столь же информативными количественными характеристиками, как и соответствующие данные для исходных траекторий [5], был проведен анализ функций спектральной плотно-

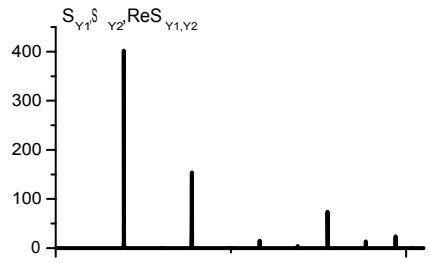


Рис. 1. Спектральные плотности для устойчивого периодического движения.

сти и корреляционных функций для двух временных рядов $\boldsymbol{Y1}$ и $\boldsymbol{Y2}$, соответствующих траекториям $\boldsymbol{x}(t, t^*, \boldsymbol{x^*})$ и возмущенной $\boldsymbol{x}(t, t^*, \boldsymbol{x}^* + \boldsymbol{\delta})$. Если для траекторий, стремящихся к устойчивому предельному циклу, спектральные плотности $S_{\boldsymbol{Y1}}$, $S_{\boldsymbol{Y2}}$ двух рядов и реальная часть взаимного спектра $\operatorname{Re} S_{\boldsymbol{Y1},\boldsymbol{Y2}}$ практически совпали (мнимая часть взаимного спектра $\operatorname{Im} S_{\boldsymbol{Y1},\boldsymbol{Y2}} \approx 0$), то для траекторий, принадлежащих хаотическим и стохастическим аттракторам, каждый ряд имел свои отличные от других спектральные характеристики. Для иллюстрации сказанного на рис. 1 и 2 приведены соответствующие функции для предельного цикла и странного аттрактора из одного фазового портрета динамической системы.

Аналогичная картина наблюдается и для корреляционных функций: совпадение всех характеристик для временных рядов, отвечающих исходной и возмущенной траекториям, принадлежащим устойчивому предельному циклу, и различие — для хаотических движений.

Таким образом, функции взаимного спектра и взаимной корреляции для одномерных временных рядов, соответствующих исходной и возмущенной траекториям, дают возможность ввести количественные оценки степени и характера хаотичности и отслеживать изменение хаотических движений с изменением параметров системы.

3.3. Исследование локальной устойчивости движений в фазовом пространстве. Такое исследование проводится на основе изучения временных рядов $\boldsymbol{Y1}$ и $\boldsymbol{Y2}$, отвечающих, соответственно, траекториям $\boldsymbol{x1}(t) = \boldsymbol{x}(t, t^*, \boldsymbol{x^*})$ и возмущенной $\boldsymbol{x2}(t) = \boldsymbol{x}(t, t^*, \boldsymbol{x}^* + \boldsymbol{\delta})$, а также ряда

$$\boldsymbol{\Delta Y} = \{\Delta y_i \colon i = 0, \ldots, N\},$$

описывающего изменение со временем расстояния между двумя траекториями:

$$\Delta y_i = \sqrt{\sum_{j=1}^{n} (x1_j^i - x2_j^i)^2}.$$

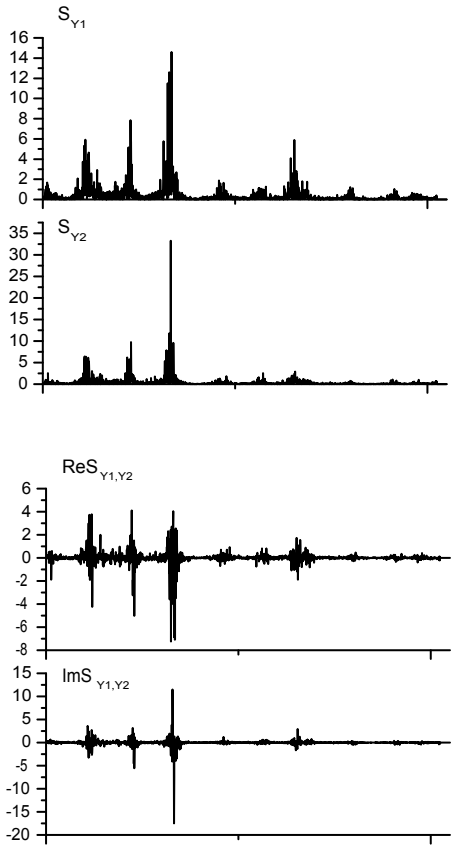


Рис. 2. Спектральные плотности для хаотического движения.

Все статистические характеристики ряда $\boldsymbol{\Delta Y}$ — оценки степени хаотичности движения. О возможности изучения хаотических движений по функциям взаимного спектра и взаимной корреляции для рядов $\boldsymbol{Y1}$ и $\boldsymbol{Y2}$ было сказано выше. Остановимся еще на одном способе оценки «разбегаемости» траекторий.

Если посмотреть на ряды $\boldsymbol{Y1}$ и $\boldsymbol{Y2}$ как на дискретную реализацию параметрически заданной функции $\boldsymbol{Y2}(\boldsymbol{Y1})$, то для устойчивого периодического движения эта функция графически представляется отрезком биссектрисы в первом квадранте координатной плоскости, а для хаотических движений — это кривая, отклоняющаяся от биссектрисы в процессе движения на угол $-\pi/4 \leqslant \alpha \leqslant \pi/4$. Для иллюстрации на рис. 3 приведена такая функция для устойчивого цикла, а на рис. 4 — для некоторых хаотических движений конкретных динамических систем.

«Разбегание» двух траекторий с близкими начальными условиями можно, в частности, описать однородной цепью Маркова с конечным числом s состояний, определяемыми разбиением первого квадранта плоскости $Y1 - Y2$ на $s = 2k - 1$ равных углов, где k-й угол содержит биссектрису. Изуче-

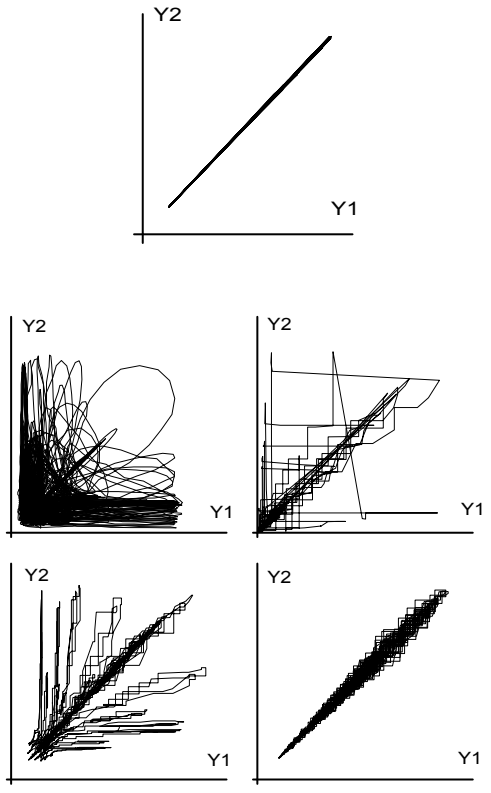


Рис. 3. Примеры характеристик «разбегания» локально неустойчивых траекторий.

ние процесса «разбегания» может быть проведено совершенно аналогично описанному в разделе 2, начиная с построения матрицы вероятностей перехода из одного состояния в другое за один шаг и дальнейшего ее исследования.

Заключение

Проведенные исследования показали, что все предложенные количественные оценки, во-первых, позволяют отличить устойчивые движения от хаотических, и, во-вторых, они различны для разных (по оценкам специалистов) видов хаотических движений, отражая их специфические особенности. Это позволяет использовать их в качестве информативных признаков при изучении хаотических движений, а в дальнейшем и в решении задачи классификации хаоса методами распознавания образов.

Литература

[1] *Neimark Yu. I., Kotel'nikov I. V., Teklina L. G.* Numerical research of the dynamic systems as the pattern recognition problem // Pattern Recognition and Image Analysis: New Information Technologies. Proceedings of the 8th International Conference (PRIA-9-2008). Nizhni Novgorod, 2008. — Pp. 96–99.

[2] *Неймарк Ю. И.* Динамические системы и управляемые процессы. — М.: Наука, 1978. — 336 с.

[3] *Гантмахер Ф. Р.* Теория матриц. — М.: Наука, 1967. — 575 с.

[4] *Неймарк Ю. И., Теклина Л. Г.* Новые технологии применения метода наименьших квадратов. — Нижний Новгород: Изд. Нижегородского госуниверситета, 2003. — 196 с.

[5] *Неймарк Ю. И., Ланда П. С.* Стохастические и хаотические колебания. — М.: Наука, 1987. — 423 с.

Метод синтеза бесконечного множества ансамблей квазиортогональных фазокодированных последовательностей с идеальной периодической автокорреляционной функцией*

Парсаев Н. В., Тюкаев А. Ю., Леухин А. Н.
code@marstu.net
Йошкар-Ола, Марийский государственный технический университет

Выполнен обзор известных на сегодняшний день методов синтеза ансамблей дискретно-кодированных последовательностей с заданными корреляционными и спектральными характеристиками. Разработан алгоритм синтеза бесконечного множества ансамблей квазиортогональных по периодической взаимной корреляционной функции фазокодированных последовательностей с нулевым уровнем боковых лепестков периодической автокорреляционной функции.

Область применения дискретно-кодированных последовательностей с каждым годом расширяется. Например, в системах радиолокации, радионавигации и передачи информации их применяют в качестве модулирующих последовательностей при формировании сложных широкополосных сигналов; в вычислительных системах дискретно-кодированные последовательности применяют в качестве псевдослучайных последовательностей для имитационного моделирования, кибернетической диагностики, встроенного тестового контроля и защиты от несанкционированного доступа; в системах автоматики и телемеханики — для построения самосинхронизирующихся циклических кодов с обнаружением и исправлением ошибок и т. д. [1].

Применяемые в настоящее время сигналы и методы их обработки нацелены, как правило, на минимизацию взаимного влияния сигналов от различных источников. При этом требование минимизации взаимного влияния, т. е. требование максимизации расстояния между сигналами в метрическом пространстве, может вступать в противоречие с другими требованиями, предъявляемым к сигналам, например, минимизации уровня боковых лепестков периодической автокорреляционной функции (АКФ). В современных радиотехнических системах различного назначения наиболее востребованными являются оптимальные по минимаксному критерию ансамбли сложных сигналов, полученных на базе фазокодированных последовательностей (ФКП) с заданными авто- и взаимно корреляционными свойствами [2]. В работе будет проведён обзор наиболее известных на сегодняшний день оптимальных по минимаксному критерию ансамблей ФКП, а также предложен метод формирования бесконечного множества ансамблей квазиортогональных по периодической взаимной корреляционной функции (ВКФ) фазокодированных последовательностей с нулевым уровнем боковых лепестков периодической АКФ.

*Работа выполнена при финансовой поддержке по темам НИР в рамках гранта РФФИ № 09-07-00072-а, гранта фонда HCF (договор № 189), в рамках АВЦП «Развития научного потенциала высшей школы» мероприятия 1 (НИР № 1.02.09), а также гос. контракта по программе У.М.Н.И.К. № 6538р/9098.

Известные оптимальные ансамбли фазокодированных последовательностей

В настоящее время известно чрезвычайно мало оптимальных по минимаксному критерию ансамблей кодовых последовательностей [1, 2], которые достигают теоретическую границу Вэлча, устанавливающую связь между квадратом максимума корреляционного выброса $|R_{\max}|$ и объёмом V ансамбля [3]:

$$R^2_{\max} \geqslant \frac{V-1}{VN-1},$$

где $R_{\max} = \max\{|R_{am}|, |R_{vm}|\}$, $|R_{am}|$ — максимальный по ансамблю боковой лепесток нормированной периодической АКФ, $|R_{vm}|$ — максимум по ансамблю выброса нормированной периодической ВКФ, V — объём ансамбля, N — длина кодовой последовательности.

Рассмотрим некоторые известные ансамбли фазокодированных последовательностей максимально приближенные по своим свойствам к теоретической границе Вэлча.

1. Ансамбли последовательностей квадратичных вычетов [4] нечётной длины N имеют максимум корреляционного выброса $R_{\max} = 1/\sqrt{N}$, при этом объём V ансамбля определяется наименьшим простым числом p_1 в каноническом разложении $N = p_1^{\nu_1} \cdot \ldots \cdot p_r^{\nu_r}$, $1 < p_1 < p_2 < \ldots < p_r$ и равен $V = p_1 - 1$. В случае, когда длина N последовательностей является простым числом, объём ансамбля достигает наибольшего значения равного $N - 1$.

2. Ансамбль частотно-сдвинутых M-последовательностей синтезируется методом посимвольного умножения элементов бинарной M-последовательности длины $N = 2^k - 1$ на дискретные гармоники $\exp\left(i\frac{2\pi k}{N}\right)$, $k = 0, \ldots, N-1$ [5]. Объем такого ансамбля равняется размерности последовательности $V = N$, а максимальный корреляционный выброс — $R_{\max} = \sqrt{N+1}/N \approx 1/\sqrt{N}$, что практически совпадает с границей Вэлча.

3. Ансамбль последовательностей Голда [6] длиной $N = 2^k - 1$ имеет объем $V = N + 2 = 2^k + 1$ и максимальный корреляционный выброс [7]:

$$R_{\max} = \begin{cases} \frac{\sqrt{2(N+1)}+1}{N}, \text{если } k \equiv 1 \mod 2, \\ \frac{2\sqrt{(N+1)}+1}{N}, \text{если } k \equiv 2 \mod 4, \end{cases} \approx$$

$$\approx \begin{cases} \sqrt{\frac{2}{N}}, & \text{если } k \equiv 1 \mod 2, \\ \sqrt{\frac{4}{N}}, & \text{если } k \equiv 2 \mod 4, \end{cases} \quad N \gg 1.$$

4. Ансамбль последовательностей Касами длиной $N = 2^k - 1$, где $k \equiv 0 \mod 2$, имеет объем $V = \sqrt{N+1} = 2^{k/2}$ и максимальный корреляционный выброс [7]:

$$R_{\max} = \frac{\sqrt{N+1}+1}{N} \approx \frac{1}{\sqrt{N}}, \quad N \gg 1.$$

5. Ансамбли последовательностей Камалетдинова строятся на основе двух различным схем [8]. Ансамбль Камалетдинова, построенный по первой схеме, имеет длину последовательностей $N = p(p-1)$, где p — простое число. При этом объём ансамбля равен $V = p + 1 = \frac{\sqrt{4N+1}+3}{2}$, а максимум корреляционного выброса определяется на основании выражения [7]:

$$R_{\max} = \frac{p+3}{N} \approx \frac{1}{\sqrt{N}}, \quad N \gg 1.$$

Ансамбль Камалетдинова, построенный по второй схеме, имеет длину последовательностей $N = p(p+1)$, где p — простое число. При этом объём ансамбля равен $V = p - 1 = \frac{\sqrt{4N+1}-3}{2}$, а максимум корреляционного выброса определяется на основании выражения [7]:

$$R_{\max} = \frac{p+1}{N} \approx \frac{1}{\sqrt{N}}, \quad N \gg 1.$$

Синтез бесконечного множества ансамблей квазиортогональных ФКП с идеальной периодической АКФ

В работе [9] разработан регулярный метод синтеза бесконечного множества фазокодированных последовательностей с нулевым уровнем боковых лепестков периодической АКФ в случае, когда длина последовательностей является квадратом некоторого целого числа. Фазокодированную последовательность определим на основании выражения:

$$\Gamma = \{\gamma_n\}_{n=0}^{N-1} = \{\exp(i\varphi_n)\}_{n=0}^{N-1}, \qquad (1)$$

где модуль каждого кодового элемента $|\gamma_n| = 1$, i — мнимая единица, $N = k^2$, k — целое положительное число, а значение фазы φ_n на каждом n-м кодовом интервале определяется в соответствие с выражением:

$$\varphi_n = \left(\beta_{n \bmod k} + \frac{2\pi}{k}\left[\frac{n}{k}\right](n-1)\right)\lambda,$$

где λ — число, взаимно-простое с N; $[x]$ — целая часть числа x; $B = (\beta_m)_{m=0}^{k-1}$ — вектор фаз, принимающих произвольные вещественные значения из диапазона $\beta_0 = 0$, $\beta_m \in [0; 2\pi]$, $m = 1, \ldots, k-1$.

Исследования показали, что если длина фазокодированных последовательностей вида (1) является квадратом некоторого целого нечётного числа k, то на базе синтезированных последовательностей с идеальной периодической АКФ можно сформировать бесконечное множество ансамблей квазиортогональных ФКП с уровнем модулей отсчётов нормированной периодической ВКФ равным $1/\sqrt{N}$.

Алгоритм синтеза бесконечного множества ансамблей квазиортогональных ФКП с нулевым уровнем боковых лепестков периодической АКФ, в случае $N = k^2$, где $k \equiv 1 \mod 2$, можно представить следующим образом.

1. Определяем систему вычетов по модулю N, взаимно простых с N:

$$\{\lambda_s\}_{s=0}^{\varphi(N)-1},$$

где $\varphi(N)$ — фи-функция Эйлера.

2. Определяем наименьшее простое число p_1 в разложении числа N:

$$N = p_1^{\nu_1} \cdot \ldots \cdot p_r^{\nu_r}, \quad 1 < p_1 < \ldots < p_r.$$

3. Среди всех $C_{\varphi(N)}^{p_1-1}$ сочетаний по $p_1 - 1$ вычетов по модулю N, взаимно простых с N, из $\varphi(N)$ возможных вычетов отбираем w-е сочетания вычетов $\{\lambda_s^{(w)}\}_{s=0}^{p_1-2}$, для которых справедливо условие:

$$\text{НОД}\left(\lambda_q^{(w)} - \lambda_l^{(w)}, N\right) = 1, \quad q \neq l, \quad q, l = 0, \ldots, \varphi(N)-1,$$

где $w = 0, \ldots, W-1$; W — количество формируемых ансамблей; НОД — наибольший общий делитель.

4. w-й ансамбль квазиортогональных ФКП с нулевым уровнем боковых лепестков периодической АКФ будет иметь вид:

$$A^{(w)} = \{\Gamma_j^{(w)}\}_{j=0}^{V-1}, \qquad (2)$$

где $V = p_1 - 1$ — объем ансамбля, а j-я ФКП из ансамбля (2) имеет вид:

$$\Gamma_j^{(w)} = \gamma_{j_{(n)}}^{(w)} = \exp\left(i\left(\beta_{n \bmod k} + \frac{2\pi}{k}\left[\frac{n}{k}\right](n-1)\right)\lambda_j^{(w)}\right),$$

где $j = 0, \ldots, V-1$, $n = 0, \ldots, N-1$.

Заключение

Для случая, когда длина N фазокодированных последовательностей вида (1) является квадратом некоторого целого числа k, где $k \equiv 1 \mod 2$, разработан регулярный метод синтеза бесконечного множества ансамблей квазиортогональных ФКП с нулевым уровнем боковых лепестков периодической АКФ. Объём V каждого формируемого ансамбля определяется наименьшим простым числом p_1 в каноническом разложении числа $N = p_1^{\nu_1} \cdot \ldots \cdot p_r^{\nu_r}$, $1 < p_1 < \ldots < p_r$, и равен $V = p_1 - 1$. При этом для любых двух ФКП Γ_j и Γ_l из такого ансамбля выполняется равенство:

$$|r_\tau| = \begin{cases} 1, & \text{если } j = l \text{ и } \tau = 0, \\ 0, & \text{если } j = l \text{ и } \tau \neq 0, \\ 1/\sqrt{N}, & \text{если } j \neq l, \end{cases}$$

где $|r_\tau|$ — модули отсчётов нормированной периодической ВКФ двух ФКП Γ_j и Γ_l, $\tau = 0, \ldots, N-1$ — значение циклического сдвига. Полученные ансамбли квазиортогональных ФКП с нулевым уровнем боковых лепестков периодической АКФ значительно расширяют класс известных оптимальных по минимаксному критерию ансамблей фазокодированных последовательностей, т. к. имеют максимум корреляционного выброса $R_{\max} = 1/\sqrt{N}$, что совпадает с границей Вэлча.

Литература

[1] *Гантмахер В. Е., Быстров Н. Е., Чеботарёв Д. В.* Шумоподобные сигналы. Анализ, синтез, обработка. — СПб.: Наука и техника, 2005. — 400 с.

[2] *Ипатов В. П.* Периодические дискретные сигналы с оптимальными корреляционными свойствами. — М.: Радио и связь, 1992. — 152 с.

[3] *Левенштейн В. И.* Границы для упаковок в метрических пространствах и некоторые их приложения // Проблемы кибернетики. — М.: Наука, 1983. — Вып. 40. — С. 43–110.

[4] *Леухин А. Н., Тюкаев А. Ю., Бахтин С. А., Парсаев Н. В.* Аналитическое решение задачи синтеза алфавита квазиортогональных фазокодированных последовательностей с дельтовидной автокорреляционной функцией // Электромагнитные волны и электронные системы. — 2009. — № 3. — С. 40–47.

[5] *Goldberg B.-G.* Division multiplexing by frequency shifted biphase modulated M-sequences // IEEE Trans. Aerosp. Electron. Syst. — Vol. 17. — 1981. — Pp. 303–304.

[6] *Gold R.* Optimal binary sequences for spread spectrum multiplexing // IEEE Trans. Inform. Theory. — Vol. 13. — 1967. — Pp. 619–621.

[7] *Ипатов В. П.* Широкополосные системы и кодовое разделение сигналов. Принципы и приложения. — М.: Техносфера, 2007.

[8] *Kamaletdinov B. Zh.* Optimal sets of binary sequences // Problems of Inform. Transmission. — Vol. 32. — 1996. — Pp. 171–175.

[9] *Парсаев Н. В., Леухин А. Н.* Дискретные фазокодированные последовательности с нулевым уровнем боковых лепестков циклической автокорреляционной функции размерности квадратных чисел // Вестник МарГТУ. Серия: Радиотехнические и инфокоммуникационные системы. — 2008. — № 3. — С. 28–35.

Применение методов распознавания образов в системе управления коллекциями графических документов*

Рогов А. А., Рогова К. А., Кириков П. В.
rogov@psu.karelia.ru, ksushar@mail.ru, lispad@gmail.com
Петрозаводский государственный университет

В работе рассказывается об исследованиях по созданию информационной системы для управления коллекцией графических документов петроглифов Северной Фенноскандии. Рассматриваются задачи подготовки документов для организации доступа через WEB, классификации изображений с учётом цветовосприятия и текстурных характеристик, поиска по выбранным комбинациям признаков или эталонам, моделирования зон видимости объектов на местности и определения количества зрителей. Особенностью создаваемой информационной системы является возможность хранения и использования иерархии документов и значений наборов признаков, зависящих от типа документа.

В настоящее время всё большую популярность получают электронные коллекции графических документов. Обычные пользователи создают свои фотоальбомы не только дома, на локальном компьютере, но и в сети Интернет; научные сообщества используют большие объёмы графической информации в своих исследованиях. К сожалению, сейчас нет единой системы, которая позволила бы не только хранить изображения в определенном порядке, но и классифицировать графическую информацию по выделенным параметрам и осуществлять поиск.

Новые компьютерные технологии позволяют создавать информационные системы, которые ранее невозможно было реализовать из-за нехватки вычислительных ресурсов. Заметим, что наиболее требовательными к ресурсам аппаратного обеспечения являются системы, предназначенные для обработки мультимедийной информации, трёхмерной графики, больших объёмов информации, а также системы, в которых реализованы эвристические алгоритмы решения NP-полных задач.

С прогрессом информационных технологий меняются и парадигмы, используемые для создания информационных систем: объектно-ориентированный подход становится не только популярной концепцией программирования, но и находит своё отражение в теории построения баз данных. Объектно-ориентированные базы данных, использующие в качестве данных абстрактные объекты, в некоторых областях позволяют более точно моделировать логическую структуру данных по сравнению с реляционным подходом. При этом современное состояние вычислительной техники позволяет создавать эффективные системы управления объектно-ориентированными базами данных.

Одним из частных случаев объектно-ориентированных баз данных являются базы данных для хранения графических документов, с учётом их внутренних свойств и признаков документов, называемые коллекциями графических документов.

*Работа выполнена при финансовой поддержке РГНФ, проект №08-01-12116в.

Целью данной работы является описание информационной системы, позволяющей создавать, управлять и анализировать коллекции графических документов с учётом графических признаков (цветовосприятия и текстуры изображения).

Общие характеристики системы

Информационная система предназначена для размещения данных на WEB-сервере и обеспечения удобного доступа через сеть Интернет с любого подключенного к сети устройства. Она пригодна для решения большого ряда задач, среди которых можно выделить следующие:

- создание и редактирование иерархической структуры коллекций графических документов;
- хранение различных наборов параметров для каждого графического документа;
- классификация и поиск графических документов по различным комбинациям параметров, а также на основе сходства текстур, цветового восприятия и т. д.;
- описательная статистика коллекции;
- разделение доступа к системе.

Система логически разделена на две части: пользовательскую и административную, и предоставляет простой и удобный интерфейс. Для ввода информации пригодны графические документы в различных форматах, автоматически осуществляется их приведение к единому стандарту, и в автоматизированном режиме предоставляется возможность выделять объекты на них.

Главное отличие предлагаемой системы от существующих систем создания электронных фотоальбомов состоит в возможности приписывать графическому документу набор индивидуальных признаков и осуществлять поиск по выделенной комбинации признаков. На основе признаков автоматически производится классификация объектов коллекции с целью поиска наиболее близких между собой. Кроме того, пользователю предлагается статистическая информация о наличии в коллекции объектов с выделенным набором признаков и анализ выделенных признаков статистическими методами.

Информационная система реализуется на php с использованием web-сервера apache и сервера баз данных mysql. В данный момент прототип информационной системы апробируется на коллекции графических документов петроглифов Северной Фенноскандии.

Типы признаков изображений

Все признаки можно разделить на 2 группы. Первая — признаки, значения которых вводит администратор системы. При этом для вычисления некоторых признаков возможна частичная автоматизация. Во второй группе признаки получаются в автоматическом или авторизованном виде и касаются параметров цветовосприятия и текстур изображения.

Признаки из первой группы описывают каждое изображение по параметрам следующего вида:

- количественные характеристики изображений;
- номинальные переменные с отношением порядка и категоризированные переменные, кроме того, отдельные признаки могут иметь более сложную фасетную структуру [4], которую можно описать с помощью графа;
- текстовые описательные признаки (они не используются при статистическом анализе);
- фрагменты изображений: взаимосвязь и повтор объектов, описываются с помощью ориентированных и неориентированных графов.

Приведём пример фасетной структуры на основе признаков петроглифов Северной Феноскандии. Все изображения делятся на группы, и для каждой группы определяются свои признаки, для каждого значения которого определяются свои признаки и т. д. Примерами признаков для группы лосей/оленей являются следующие: тип головы (удлинённая, укороченная, нормальная), тип шеи (утолщённая, узкая, нормальная), тип рогов (отсутствуют, лося, оленя, ни те, ни другие), тип корпуса (линейный, грузный, массивный) и т. д.

Возможно вычисление некоторых параметров изображения в полуавтоматическом режиме с использованием встроенного функционально программируемого калькулятора. Проиллюстрируем его работу на примере изображений лосей и оленей. Определение некоторых признаков, таких, как толщина шеи, тип головы, тип корпуса и т. д. визуально по рисунку не всегда является точным и часто зависит от эксперта. Для того, чтобы избежать двусмысленности восприятия, предлагается использовать формулы отношения между длинами соответствующих отрезков. На изображении петроглифа лося/оленя выделяют такие отрезки, как длина головы, длина корпуса, длина хвоста, длина туловища. Для того, чтобы задать эти отрезки на рисунке в программе, необходимо мышкой отметить точки начала и конца отрезков. После этого автоматически вычисляются длины этих отрезков. Чтобы определить значение признака, необходимо сравнить длины соответствующих отрезков. В этом случае формулы задаются пользователем. Имеется возможность использовать все арифметические операции, а также операции сравнения. При выполнении всех действий учитывается вычислительная погрешность. С помощью такого калькулятора происходит формализация восприятия изображения, упрощается ввод характеристик, и они становятся более точными.

Во второй группе признаков выделим: характеристики текстуры и цветовосприятия. Одним из стандартных способов представления цветовой характеристики изображения является цветовая гистограмма. Для её построения пространство всех цветов разбивается на подмножества так, чтобы схожие цвета попали в один интервал. Для каждого интервала подсчитывается количество пикселей, чей цвет принадлежит данной области. Для анализа гистограмм используются различные метрики, например, сумма модулей разностей значений элементов гистограмм для каждой цветовой области [1]. Вместо гистограммы можно брать вектор цветовой когеренции [2]. Другим вариантом представления является статистическая модель: рассматривается статистическое распределение различных цветовых каналов. Сравнение распределений является оценкой схожести [3]. Кроме того, можно рассматривать не только одномерные распределения, но и трёхмерные, учитывая все взаимосвязи между каналами (ковариации). Рассматриваются интервалы наиболее часто встречающихся цветов, размеры одноцветных цветовых фрагментов, перевод цветных изображений в бинарные и их анализ. Для анализа текстур одним из применяемых методов является анализ независимых компонент. С его помощью выделяют фильтры, отражающие основные направления текстур для той базы изображений, на основе которой они строятся [1]. Кроме этого, используется спектр фрактальной размерности Реньи [5, 6].

Анализ признаков

С помощью специального модуля корреляционного анализа введённые признаки можно проверить на статистическую независимость с помощью критерия Пирсона. Для этого выделяют признаки, группы объектов и задают уровень значимости. Кроме того, для анализа признаков используются методы описательной статистики.

Классификация и кластеризация

Управление типами документов, а также наборами признаков, общих для всех графических документов коллекции и уникальных для каждого типа, позволяет хранить разнообразные данные

о каждом графическом документе, организовать поиск документов по типу, признаку или набору признаков, задав точное значение или границы варьирования значения каждого признака и точность поиска (количество совпадений признаков).

Для системы создаются модули для статистического анализа документов — количественного анализа, корреляционного анализа связей между группами признаков, а также разбиение изображений на группы методом иерархического кластерного анализа. Признак можно описать как $f\colon X \to D_f$, где D_f — множество допустимых значений признака. Тогда, если заданы признаки $f_1, \ldots, f_n$, то вектор $\big(f_1(x), \ldots, f_n(x)\big)$ называется признаковым описанием графического документа x. В качестве меры расстояния между документом x и эталоном y берётся эвклидово расстояние $R^2(x,y) = \sum\limits_{k=1}^{n} \big(f_k(x) - f_k(y)\big)^2$. Ввиду различия множеств допустимых значений для различных признаков для корректной работы алгоритма выполняется нормировка значений признаков: $\tilde{f}_j(x) = \frac{f_j(x) - \min(f_j)}{\max(f_j) - \min(f_j)}$. В этом случае значение каждого признака будет лежать в пределах $[0;1]$.

Для классификации объектов коллекции на основе признаков применяются методы дискриминантного и кластерного анализов, в частности, метод иерархического кластерного анализа (метод ближайшего соседа), корреляционных плеяд, эталонные алгоритмы. В настоящее время проводятся эксперименты с методом опорных векторов (SVM).

Классификация по цветовому и текстурному анализу может быть осуществлена несколькими методами. Примерами являются поиск по цветовым моментам и цветовым гистограммам [1]. Кроме того, существуют методы поиска нечетких дубликатов [2]. Поиск нечётких дубликатов позволяет предположить, являются ли два объекта частично одинаковыми или нет. Частично одинаковые изображения могут образовывать один кластер. Кроме того, схожесть изображений по степени цветового восприятия может быть осуществлена при сравнении площади белого, определении наличия большого фрагмента определенного цвета и т. д.

Поиск по изображениям

Управление типами документов, а также наборами признаков, общих для всех графических документов коллекции и уникальных для каждого типа, позволяет хранить разнообразные данные о каждом графическом документе, организовать поиск документов по типу, признаку или набору признаков, задав точное значение или границы варьирования значения каждого признака и точность поиска (количество совпадений признаков — для номинальных и категоризированных и интервалы изменения — для количественных переменных). Поиск по изображениям предназначен для поиска изображений, похожих на данное или на его фрагмент. На вход подаётся исследуемое изображение, а на выходе должны появиться изображения из базы данных, наиболее похожие на исходное. Для поиска похожих графических объектов на основе текстур изображений используются методы нейронных сетей и геометрического программирования.

Рассмотрим поиск на примере наскальных изображений Северной Фенноскандии. На сегодняшний день все новейшие материалы по петроглифам представляют собой набор цветных фотографий. Определённую сложность поиска создаёт фактическое отсутствие некоторых частей изображения. Поиск также осложняется тем, что часто невозможно определить, где верх, а где низ изображения. При этом требование, что при поиске необходимо только совпадение контура изображения, позволяет упростить поиск, а значит изображение петроглифа можно рассматривать как бинарное (скале соответствует белый цвет, а петроглифу — черный). В зависимости от выбранных параметров поиска (точность поиска, процент совпадений элементов изображений) будет найдено одно или несколько изображений. Для поиска используются сеть адаптивного резонанса и структурный метод поиска.

В результате поиска пользователю предоставляется доступ к информации о кодовом номере, месторасположении, характерных признаках найденного петроглифа и петроглифах, близких к нему по ранее описанным признакам.

Модуль анализа видимости

Особенным для коллекции петроглифов Фенноскандии стал модуль анализа видимости. Изучая области видимости наскального рисунка, можно ответить на вопрос о целевой аудитории автора, делать выводы о символическом значении данного образа для древних людей. Данные, полученные о видимости, могут быть использованы не только для анализа наскальных рисунков, но полезны специалистам других областей. Например, зная необходимое минимальное число зрителей, можно определять требуемые для этого размеры объекта и его место расположения с учётом ландшафта местности. Подобная информация может быть полезна при архитектурном планировании размеров и позиции нового памятника или монумента.

Видимость объекта характеризуется степенью видимости (отчетливостью видимого объекта) и дальностью видимости, то есть расстоянием, с которого наблюдаемый объект становится различимым глазом. Параметром, определяющим способность субъекта распознать объект, является угол зрения. Для человека с нормальным зрением минимальным углом, при котором объект начинает

быть виден, принимается угол в 10 минут. В свою очередь угол зрения зависит от взаимного расположения в пространстве объекта и субъекта относительно друг друга — от расстояния между ними, наклона наблюдаемого объекта, его высоты, ландшафта местности, контрастности объекта, его освещённости и роста наблюдателя.

Была получена функциональная зависимость угла зрения от остальных параметров (данная модель не учитывает параметр освещённости объекта наблюдения): $\alpha(X) = \arctan\left(\frac{H-h}{X}\right) - \arctan\left(\frac{H-h_1}{X+r}\right)$, где H — высота точки наблюдения (рост человека плюс высота точки расположения), h_1 и h — высота нижнего и верхнего края изображения соответственно, r — расстояние между дальними точками изображения. X — расстояние от точки наблюдения до нижнего края петроглифа. Аналитически выразить функциональную зависимость расстояния от угла и других параметров (для получения максимального расстояния видимости при известном минимальном угле) невозможно. Для нахождения решения применяется численный метод Ньютона: $x \leftarrow x - c(x,u)\big/\frac{dc}{dx}$, где $c(x,u) = \alpha(x) - u$, u — минимальный угол видимости.

Для определения числа зрителей была разработана модель, описывающая распределения параметров, свойственные реальным группам наблюдателей, а также решена задача поиска расположения наблюдателей в зоне видимости объекта, при котором впереди стоящие зрители не закрывают объект от стоящих сзади.

Согласно исследованиям антропологов, рост взрослого человека распределён согласно закону нормального распределения, в то время как рост ребёнка лучше описывается логнормальным распределением. Плотности случайных величин, выражающих рост взрослого и ребёнка, имеют вид $\frac{1}{\sigma_g\sqrt{2\pi}}\exp\left(\frac{(x-\mu_g)^2}{2\sigma_g^2}\right)$ и $\frac{1}{\sigma_c x\sqrt{2\pi}}\exp\left(\frac{(\ln(x)-\mu_c)^2}{2\sigma_c^2}\right)$.

В предложенной модели границей, разделяющей группы взрослых и детей, является пороговое значение роста, превысив которое человек начинает считаться взрослым. Задаются параметры модели: пограничный рост (максимальный для ребёнка cx и минимальный для взрослого gn), минимальный рост ребёнка cn, максимальный рост взрослого gx и процентное соотношение числа взрослых к числу детей cp. На основе этих данных определяются параметры распределений $\mu_g = \frac{1}{2}(gx + gn)$, $\mu_c = \ln\left(\frac{1}{2}(cx + cn)\right)$. С помощью правила «трёх сигм» $\sigma_g = \frac{1}{2}(gx - gn)$, $\mu_c = \ln\left(\frac{1}{6}(cx - cn)\right)$, моделируется набор значений роста для группы человек заданной численности.

Вторая задача относится к NP-полным, поэтому был предложен и реализован в системе эвристический алгоритм. Таким образом, информационная система позволяет, задав параметры группы наблюдателей, наглядно отобразить зону видимости и расположение зрителей.

Выводы

В разработанной информационной системе для создания и управления коллекцией графических документов реализованы модули для проведения классификаций изображений с учётом цветовосприятия и текстурных характеристик и поиска по выбранным комбинациям признаков или эталонам, решении задачи моделирования зон видимости объектов на местности и определения количества зрителей. Особенностью разработанной информационной системы является возможность хранения и использования иерархии документов и значений наборов признаков, зависящих от типа документа. Создаваемое в коллекции дерево для хранения иерархии может иметь неограниченное число уровней и неограниченное число узлов на каждом уровне. Подобный функционал позволяет применять создаваемую коллекцию для хранения каталога запчастей/товаров с иллюстрациями, базы произведений художников, хранимых в музее, и ряде других областей. Апробация решения проходит на коллекции петроглифов Северной Фенноскандии. Для этой коллекции в системе разрабатываются некоторые специфичные модули, применимые с учётом особенностей петроглифов.

В настоящее время, кроме создаваемой с помощью системы коллекции петроглифов Северной Фенноскандии, система апробируется на материалах Карельского государственного краеведческого музея при создании коллекции открыток и коллекции фотографий со строительства Беломорско-Балтийского канала.

Литература

[1] *Васильева Н., Марков И.* Синтез цветовых и текстурных признаков при поиске изображений по содержанию // Труды РОМИП 2007–2008. — СПб.: НУ ЦСИ, 2008. — С. 135–144.

[2] *Кисель Я.* Алгоритм поиска нечётких дубликатов в коллекции изображений // Труды РОМИП 2007–2008. — СПб.: НУ ЦСИ, 2008. — С. 170–173.

[3] *Sticker M., Dimai A.* Color indexing with weak spatial constraints // SPIE Conference, 1996 — Pp. 29–40.

[4] *Обухова О.Л., Гершкович М.М., Бирюкова Т.К., Соловьев И.В., Чочиа А.П.* Открытые электронные коллекции с адаптивным визуальным интерфейсом фасетной навигации // 10-я Всеросс. научн. конф. RCDL-2008, Дубна: ОИЯИ, 2008. — С. 128–138.

[5] *Рогов А. А., Рогова К. А., Спиридонов К. Н., Быстров М. Ю.* Система поиска в электронной коллекции изображений петроглифов Карелии // 10-я Всеросс. конф. RCDL, Дубна: ОИЯИ, 2008. — С. 246–251.

[6] *Спиридонов К. Н.* К вопросу об инварианте графического изображения // Всеросс. конф. ММРО-13, М.: МАКС Пресс, 2007. — С. 393–396.

Обработка многоградационных пространственных изображений с неупорядоченными отсчётами*

Роженцов А. А, Баев А. А., Наумов А. С.
krtmbs@marstu.net
Йошкар-Ола, Марийский государственный технический университет

В работе предложен подход к распознаванию многоградационных изображений пространственных объектов с неупорядоченными отсчетами, задающими поверхность объекта. Методика базируется на формировании вторичного описания пространственного объекта в виде коэффициентов полиномиальной функции гиперкомплексного переменного, проецирующей его отсчеты на поверхность единичной сферы. Показана возможность решения задач оценки параметров преобразований масштабировании и вращения пространственных объектов и задачи распознавания.

Решаемые современными системами технического зрения задачи требуют перехода от плоских сцен к анализу пространственных изображений. Ряд сложившихся к настоящему времени подходов к их обработке базируется на воксельных моделях и связан с трудоемкими процедурами вычисления трехмерного градиента для выделения плоских фрагментов [1]. В работе [2] рассмотрены методы распознавания и сегментации изображений, базирующиеся на локальных признаках объектов в окрестностях задающих их поверхности точек. Рассмотрены три метода, основанные на контекстной информации 3D образа, на сферическом гармоническом анализе, и на анализе спинов изображений [3]. При этом отмечается, что ни один из методов не обладает одновременно инвариантностью ко всем видам геометрических преобразований. Указывается, что было бы идеально, если бы формируемые признаки обладали инвариантностью к преобразованиям вращения и переноса обрабатываемого объекта и были робастными в условиях воздействия шумов и помех. В последнее время получили развитие методы, основанные на представлении изображений пространственных объектов и точечных полей в виде кватернионных сигналов [4], предполагающие переход к формированию описания объекта в виде контура многогранника, но они требуют предварительного упорядочивания отсчетов, задающих поверхность объекта или вершины многогранника. При этом также не учитывается информация о яркости отдельных отметок.

На практике в большинстве случаев изображение точечной отметки занимает более одного элемента разрешения изображения. В ряде случаев для повышения точности локализации отметки в сцене производится даже преднамеренная расфокусировка изображения. При этом точке истинного положения отметки соответствует фрагмент изображения с максимальной яркостью, убывающей по мере удаления от отметки по какому-либо закону, например по экспоненциальному или гауссовому. Исключение из рассмотрения данных о яркости отметок и окружающих их пикселей приводит к потере существенного количества информации и снижению помехоустойчивости, поэтому интерес представляют алгоритмы, сочетающие эффективность контурного анализа и полноту учета яркостной информации. Целью работы является разработка подходов к обработке изображений пространственных объектов, поверхность которых задана неупорядоченными отсчетами с учетом яркости отдельных отметок.

Формирование аналитического описания пространственного объекта

Аппарат, базирующийся на алгебрах Клиффорда и, в частности, на кватернионном анализе, в настоящее время широко применяется в задачах обработки изображений. Например, в работах [5, 6] рассматриваются методы синтеза быстрых алгоритмов дискретных ортогональных преобразований. В работах [7, 8, 9] исследуется возможность применения кватернионного преобразования Фурье к обработке цветных плоских изображений. Работы [10, 11] посвящены вопросам синтеза нейронных сетей с применением алгебр Клиффорда для обработки плоских и пространственных изображений, но вопросы влияния на эффективность их работы помеховых факторов и преобразований вращения и масштабирования не рассмотрены.

Предлагаемый в данной работе подход не требует знания нумерации отметок, задающих форму поверхности объекта, и позволяет выполнять обработку при неизвестных параметрах преобразований вращения и масштабирования. Он основан на формировании вторичного описания формы обрабатываемого пространственного объекта в виде функции гиперкомплексного переменного, проецирующей отсчеты, задающие поверхность объекта, на сферу единичного радиуса. При этом координаты точек поверхности задаются в виде векторных кватернионов.

*Работа выполнена при финансовой поддержке РФФИ, проекты №08-01-00854-а, №07-01-00058-а, №08-01-12000-офи, и по программе «Развитие научного потенциала высшей школы», проект №2.1.2/2204.

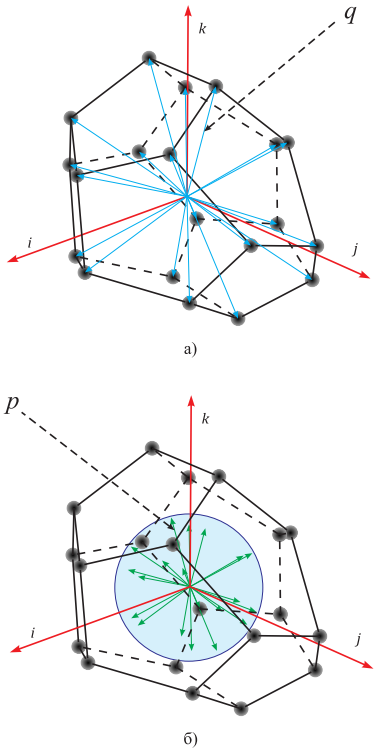


Рис. 1. Пространственный объект (а) и результат проецирования его отсчетов на единичную сферу (б).

Вторичное описание объекта задается в виде коэффициентов полиномиальной проецирующей функции, для получения которых используется метод наименьших квадратов. Он позволяет задавать веса для каждой отметки при вычислении значений ошибок проецирования в качестве которых могут использоваться яркости соответствующих отметок. Если отказаться от процедуры бинаризации изображения, то при расчете параметров проецирующей функции будут учитываться значения яркости не только самой отметки, но и ряда окружающих ее пикселей, благодаря чему увеличится количество информации, используемой при формировании вторичного описания. Это должно привести к повышению качества принимаемых решений. Рассмотрим методику формирования проецирующей функции для многоградационных сцен. Пусть каждый пиксель изображения, учитываемый при формировании вторичного описания, имеет яркость J_n, $n = 0, \ldots, N-1$, где N — количество пикселей, а его координаты задаются векторным кватернионом $q_n = iq_{1,n} + jq_{2,n} + kq_{3,n} = ix_n + jy_n + kz_n$, $n = 0, \ldots, N-1$, где x_n, y_n, z_n — координаты точки в трехмерном пространстве [4, 13], рис. 1 а).

Для отображения отсчетов пространственного объекта на сферу необходимо задать соответствующую функцию, аргументами которой являются кватернионы, описывающие поверхность объекта.

Наиболее простыми и, в тоже время, достаточно универсальными являются полиномиальные отображающие функции гиперкомплексного переменного вида:

$$\sum_{m=0}^{M-1} q_n^m a_m = p_n, \qquad (1)$$

где a_m — коэффициенты полинома, также являющиеся кватернионами, задающие отображение пространственной фигуры на сферу, q_n — кватернионы, соединяющие точки поверхности объекта с началом координат, p_n — кватернионы с единичными модулями, проведенные к поверхности сферы из начала координат в направлении точек поверхности объекта, рис. 1, б).

Коэффициенты полинома a_m могут быть найдены с помощью метода наименьших квадратов. В матричном виде система линейных кватернионных уравнений запишется как $\boldsymbol{qa} = \boldsymbol{p}$, где

$$\boldsymbol{q}_{r,m} = \sum_{n=0}^{N-1} J_n \overline{\boldsymbol{q}_n^r} \boldsymbol{q}_{\boldsymbol{n}}^{\boldsymbol{m}}, \quad \boldsymbol{p}_r = \sum_{n=0}^{N-1} J_n \overline{\boldsymbol{q}_n^r} \boldsymbol{p}_{\boldsymbol{n}},$$

$r, m = 0, \ldots, M-1$. В результате решения системы уравнений будут определены значения коэффициентов a_m, $m = 0, \ldots, M-1$, полиномиальной функции, выполняющей отображение пространственной фигуры на сферу.

Обработка изображений с неизвестными параметрами масштабирования и вращения

При изменении масштаба каждый отсчет q_n умножается на некоторый масштабный множитель μ:

$$q_n^{(\mu)} = \mu q_n.$$

Можно показать, что оценка масштаба в этом случае может быть получена из выражения:

$$\hat{\mu} = \sqrt[m]{\frac{a_m}{a_m^{(\mu)}}}, \qquad (2)$$

где $a_m^{(\mu)}$ — коэффициенты проецирующей функции для объекта с измененным в μ раз масштабом. При вращении объекта каждый его вектор поворачивается на некоторый угол вокруг общей для векторов оси вращения. Аналитически вращение векторов с использованием алгебры кватернионов описывается в виде [4]:

$$q_n^{(b)} = b q_n b^{-1},$$

где b — вращающий кватернион.

Коэффициенты эталонной проецирующей функции a_m связаны с коэффициентами проецируемой функции $a_m^{(b)}$ для повернутого объекта соотношением

$$b^{-1} a_m^{(b)} b = a_m. \qquad (3)$$

Эта взаимосвязь позволяет выполнить оценку параметров вращений на основе анализа коэффициентов проецирующих функций[12].

Распознавание пространственных объектов

При распознавании пространственных объектов на вход распознающего устройства поступают отсчеты, задающие координаты точек поверхности объекта в пространстве. Устройство должно принять обоснованное решение о принадлежности распознаваемого объекта к одному из эталонных классов. В качестве величины, характеризующей меру схожести объектов, может выступать значение скалярного произведения между коэффициентами проецирующего полинома эталонного объекта a_m^* и обрабатываемого объекта a_m:

$$\eta = \sum_{m=0}^{M-1} a_m a_m^*. \qquad (4)$$

Если параметры линейных преобразований неизвестны, то необходимо предварительно выполнить их оценку, произвести обратную коррекцию аналитического описания распознаваемого объекта и вычислить меру схожести в соответствии с (4). На рис. 2 а) приведен тестовый объект, результат его проецирования на сферу, рис. 2 б), и характеристики его распознавания для случаев расчета коэффициентов без учета и с учетом яркости отдельных отметок, рис. 2 в). Во втором случае предполагалось, что вокруг каждой отметок существует «облако» точек, яркость которых уменьшается по мере удаления от центра по гауссовскому закону

$$J(r) = J\exp\bigl(-\alpha r^2\bigr),$$

где J — яркость отметки в центре «облака», r — расстояние от центра «облака», α — декремент затухания. Величина отношения сигнал/шум определялась из соотношения

$$q = \frac{\|Q\|^2}{\sigma_Q^2 N},$$

где σ_Q — величина среднеквадратичного отклонения координатного шума, $Q = \{q_n\}_{n=0}^{N-1}$ — вектор отсчетов, задающих поверхность объекта.

Для имитации влияния яркостного шума выполнялось зашумление значений яркости отметок. Предполагалось, что все отметки в центрах «облаков» имеют одинаковые яркости J, а отношение сигнал/яркостный шум определялось по формуле:

$$q = \frac{J}{\sigma_J}.$$

Как видно из полученных результатов, использование дополнительной яркостной информации

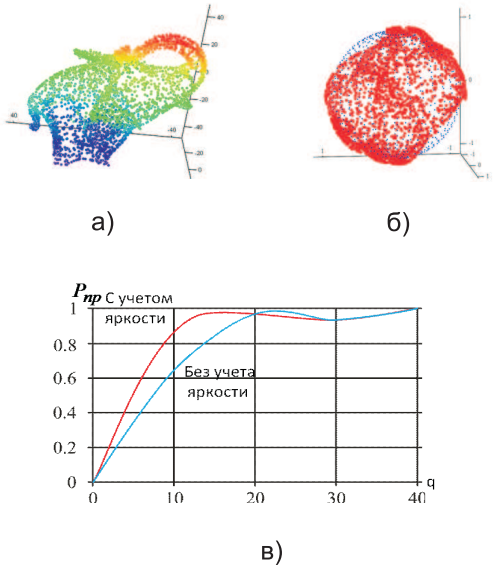


Рис. 2. Изображение эталонного объекта (а), результат его проецирования на сферу (б) и характеристики распознавания (в).

позволяет существенно улучшить характеристики распознавания изображений пространственных объектов. Особенно заметен этот эффект будет при появлении ложных отметок. Поскольку ложная отметка как правило занимает один пиксель, а не группу, то ее вклад в формирование вторичного описания будет пренебрежимо мал, поэтому учет яркостной информации позволяет значительно улучшить характеристики правильного распознавания в условиях возникновения ошибок обнаружения.

Таким образом, учет яркости отдельных отметок при формировании проецирующей функции позволяет значительно улучшить характеристики правильного распознавания пространственных объектов

Заключение

В работе предложен подход к обработке изображений 3D объектов, форма поверхности которых задана неупорядоченными отсчетами с учетом яркости соответствующих элементов и выполнена оценка его эффективности. Учет яркостной информации позволяет снизить в 1,5–2 раза требуемое значение отношения сигнал/шум для надежного распознавания, а также повысить устойчивость системы распознавания к влиянию импульсных помех.

В отличие от описанных в литературе методов, предложенный подход обеспечивает возможность обработки изображений при неизвестных параметрах вращений, масштабирования, переноса, а также в условиях воздействия шумов и помех.

Литература

[1] *Zucker S. W., Hummel R. A.* Three Dimensional Edge Operator // Intell, PAMI-3, 1981. — No. 3. — Pp. 324–331.

[2] *Frome A., Huber D., Kolluri R., Bulow T., Malik J.* Recognizing Objects in Range Data Using Regional Point Descriptors // DARPA E3D program, 2004. — 14 p.

[3] *Johnson A.* Spin-Images: A Representation for 3-D Surface Matching. Carnegie Mellon University — Pittsburgh, Pennsylvania, 1997. — 308 p.

[4] *Фурман Я. А., Кревецкий А. В., Роженцов А. А. и др.* Комплекснозначные и гиперкомплексные системы в задачах обработки многомерных сигналов, под. ред. Я. А. Фурмана. — М.: Наука, 2004. — 456 с.

[5] *Felsberg M., Bulov Th., Sommer G., Chernov V. M.* Fast Algorithms of Hypercomplex Fourier Transforms. In: G.Sommer (Eds) Geometric Computing with Clifford Algebras // Springer Verlag, 2000. — Pp. 231–254.

[6] *Чернов В. М.* Арифметические методы синтеза быстрых алгоритмов дискретных ортогональных преобразований. — М.: Физматлит, 2007.

[7] *Bulow T., Sommer G.* Algebraically Extended Representations of Multi-Dimensional Signals. Computer Science Institute, Christian Albrechts University — Kiel, Gemany, 1997. — 8 p.

[8] *Todd A. Ell, Stephen J. Sangwine.* Hypercomplex Fourier Transforms of Color lmage // IEEE Transactions On Image Processing. — Vol. 16, No. 1, January 2007. — Pp. 22–35.

[9] *Bulow T., Sommer G.* Multi-Dimensional Signal Processing Using an Algebraically Extended Signal Representation. Computer Science Institute, Christian Albrechts University — Kiel, Gemany, 1997. — 16 p.

[10] *Bayro-Corrochano E., Buchholz S., Sommer G.* A new Selforganizing Neural Network using Clifford Algebra. Computer Science Institute, Christian Albrechts University — Kiel, Gemany, 1996. — 5 p.

[11] *Buchholz S., Sommer G.* Quaternionic Spinor MPL. Computer Science Institute, Christian Albrechts University — Kiel, Gemany, 1997. — 6 p.

[12] *Роженцов А. А., Мазанов Е. И., Баев А. А.* Решение проблемы распознавания и оценки параметров 3D изображений при неизвестной нумерации отсчетов их контуров // Доклады 10-й Международной конференции: Цифровая обработка сигналов и ее применение, Москва, 2007. — С. 432–434.

[13] *Арнольд И. В.* Теоретическая арифметика — М.: Учпедгиз, 1938. — 481 с.

Выбор посадочной площадки для беспилотного летательного аппарата*

Рябинин К. Б., Фурман Я. А., Хафизов Р. Г.
krtmbs@marstu.net
Йошкар-Ола, Марийский государственный технический университет

Рассмотрено решение задачи выбора посадочной площадки беспилотного летательного аппарата по результатам обработки точечного поля в виде координат точек поверхности. Предложен метод, состоящий в анализе небольшой группы точек, в пределах которой их можно считать расположенными на плоскости. Нормали к этим плоскостям образуют векторное поле, по которому находится фрагмент с допустимым уровнем шероховатости. Разработан подход, снижающий уровень зашумленности векторного поля.

Одной из проблем эксплуатации беспилотной авиационной системы (БАС) является задача безаварийной посадки в заданном районе. Важность проблемы состоит в том, что в случае выхода из строя спутника, отказа аппаратуры связи и других нештатных ситуаций происходит потеря БАС. При этом пропадает находящаяся на его борту важная информация.

Цель работы заключается в решении задачи выбора в автоматическом режиме посадочной площадки в условиях пересеченной местности. Решение о посадке принимается по результатам анализа степени шероховатости фрагмента подстилающей поверхности.

Подходы к решению задачи выбора посадочной площадки

Для оцифровки изображений 3D подстилающей поверхности дальномерно-угловой датчик фиксирует на поверхности точки и определяет их трехмерные координаты.

После оцифровки получается генеральное множество точек (ГМТ) поверхности $F = \{f(n)\}_{n=0}^{s-1}$ мощности s, которое также будем называть точечным полем (рис. 1а). Из-за неизбежных ошибок измерения расстояний и угловых координат, а также технологических трудностей обеспечения регулярного режима оцифровки, точечное поле имеет нерегулярный характер (рис. 1б). Результаты исследований по обработке 3D точечных полей с целью распознавания задаваемых ими изображений объектов описаны в работах [1, 2].

Используемые подходы поясним с помощью иллюстраций. На рис. 2а и 2б показаны исходная сцена и точечное поле. На рис. 2в изображены результаты планиметрии, в соответствии с которой группе точек ставится в соответствие плоский сегмент W_n, ориентация которого задается нормалью $\boldsymbol{r}_n$. Множество сегментов W_n и нормалей $\boldsymbol{r}_n$, $n = 0, \ldots, s-1$, к ним, представляют собой вторичное описание анализируемой поверхности. Далее

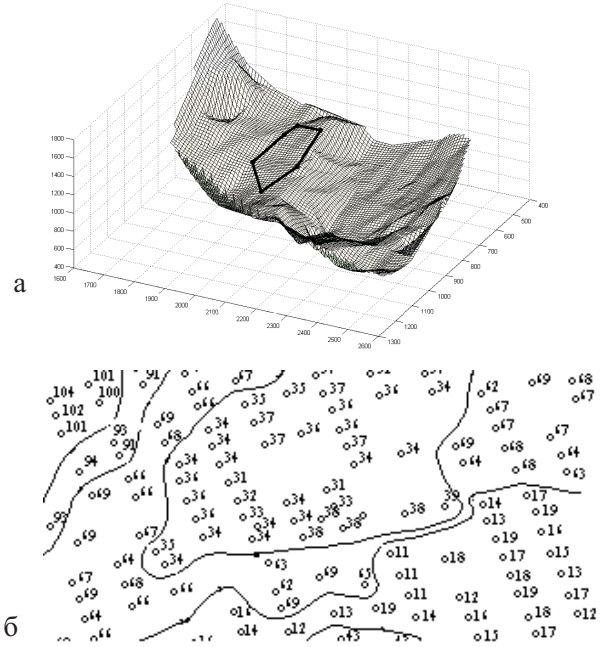


Рис. 1. а) точечное поле 3D поверхности, б) фрагмент поля с нерегулярными отметками.

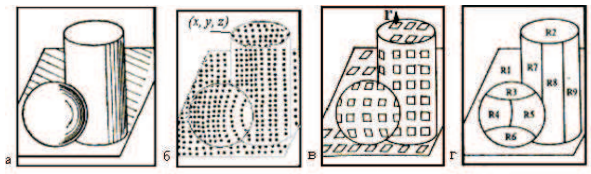


Рис. 2. Описание 3D поверхности, основанное на представлении в виде плоских участков [2].

все сегменты, ориентации которых в пределах заданного порога с определенной точностью совпадают, объединяются в элементарные области (рис.2г). Для выполнения процедуры планиметрии поверхности использован аппарат трехмерного градиента. Кроме высокой трудоемкости, вызванной необходимостью анализировать для одной точки поля содержание трех окон с 27 элементами в каждом, описанный подход имеет такой системный недостаток, как пониженная помехоустойчивость, вызванная дифференциальным характером вычисления градиента. В результате планиметрии форми-

*Работа выполнена при финансовой поддержке РФФИ, проект № 07-01-00058; а также гранта молодым ученым МарГТУ, проект № 951.

руется векторное поле $R = \{\boldsymbol{r}_n\}_{n=0}^{h-1}$, где $\boldsymbol{r}_n$ — нормаль к плоскости n-го сегмента, h — количество полученных сегментов.

Описанный в работах [1, 2] подход соответствует принципам обработки сигнала и может быть принят в качестве базового для решения поставленной задачи. При этом желательно сократить временные затраты на его реализацию. Кроме того, необходимо повысить устойчивость системы к шумам. В связи с этим для получения векторного поля нормалей R вместо метода градиента используется аппарат скалярных (внутренних) произведений векторов в кватернионном представлении, что и обеспечивает более высокую помехозащищенность.

Информативность скалярного произведения векторов, заданных в виде кватернионов

Скалярное произведение нормированных векторных сигналов $\boldsymbol{q} = \{q(n)\}_{n=0}^{s-1}$ и $\boldsymbol{p} = \{p(n)\}_{n=0}^{s-1}$, заданных в линейном действительном пространстве, вводится в виде косинуса угла между ними: $v_E = (\boldsymbol{q}, \boldsymbol{p}) = \sum\limits_{n=0}^{s-1} q(n)p(n) = \cos\varphi$. В задачах, связанных с принятием решений, величина v_E, задающая значение расстояния $R^2 = 2(1 - \cos\varphi)$ между ними, используется в качестве меры с их схожести. Поскольку часто угол φ поворота одного векторного сигнала относительно другого не является информативным параметром, то мера схожести такой пары сигналов должна оставаться равной единице, но при значениях φ, не кратных π, величина $v_E < 1$. Поэтому в подобных случаях, например, при обнаружении сигналов со случайной фазой, перед принятием решения переходят к операциям с модулями, вычисленными в виде суммы квадратурных составляющих [3].

При переходе в унитарное пространстве C скалярное произведение векторов $\boldsymbol{q} = q_1 + iq_2$ и $\boldsymbol{p} = p_1 + ip_2$ становится равным [4]:

$$v_c = v_{c1} + iv_{c2} = v_E - i(q_1p_2 - q_2p_2). \qquad (1)$$

Видно, что v_c информативнее, чем v_E, так как включает его в качестве составной части. Это позволяет однозначно найти угол φ. Переходя к операциям в 3D пространстве, зададим векторы $\boldsymbol{q}$ и $\boldsymbol{p}$ в кватернионном виде: $\boldsymbol{q} = q_1i + q_2j + q_3k$ и $\boldsymbol{p} = p_1i + p_2j + p_3k$. Учитывая тесную связь комплексных чисел с кватернионами, используем представление (1) и свойство линейности скалярного произведения для получения скалярного произведения:

$$v_H = (\boldsymbol{q}, \boldsymbol{p})_H = v_E + \mathrm{hypc}\, v_H = \cos\varphi - \boldsymbol{r}\sin\varphi. \quad (2)$$

Бивектор $\mathrm{hypc}\, v_H$ в виде гиперкомплексной части скалярного произведения (2) равен ориентированной площади $(-\sin\varphi)$ параллелограмма, построенного на векторах $\boldsymbol{q}$ и $\boldsymbol{p}$, умноженной на вектор нормали $\boldsymbol{r}$, $|\boldsymbol{r}| = 1$, к собственной плоскости Ω этих векторов. Видно, что с точностью до знака выражение (2) равно клиффордову произведению векторов $\boldsymbol{q}$ и $\boldsymbol{p}$. Выражение (2) содержит всю необходимую информацию для совмещения вектора $\boldsymbol{q}$ с вектором $\boldsymbol{p}$: угол φ определяется значениями $\sin\varphi$ и $\cos\varphi$, а ось вращения задается нормалью $\boldsymbol{r}$. Плоскостью вращения служит собственная плоскость Ω.

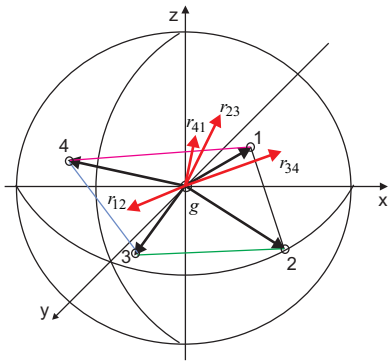


Рис. 3. Триангуляции поверхности, заданной МБТ.

Планиметрия 3D поверхности

В основу планиметрии положен принцип анализа подмножества точек, ближайших к данной текущей точке (полюсу). Это подмножество условно будем называть множеством ближайших точек (МБТ). Для условий небольшой мощности МБТ и сравнительно высокой и равномерной концентрации точек генерального множества $F = \{f(n)\}_{n=0}^{s-1}$ можно сформулировать следующий принцип МБТ: *участок 3D поверхности, на которой расположены точки МБТ, имеет плоскую форму.* Задача анализа конкретного МБТ с полюсом g_n в текущей точке $f(n)$, $n = 0, \ldots, s-1$, заключается в построении плоского сегмента W_n, содержащего точку полюса и проекции остальных точек подмножества. Точность аппроксимации точек МБТ сегментом W_n характеризуется средним значением расстояния точек МБТ от сегмента W_n и СКО этого расстояния.

Случайность расположения в пространстве точек МБТ является причиной задания им не плоской, а многогранной сложной поверхности. Она получается в результате центрированной триангуляции, выполненной из точки его полюса g (рис.3). Принимая парциальные нормали к граням полученной поверхности в качестве независимых нормально распределенных случайных величин, получим, что оценкой максимального правдоподобия нормали к сегменту W_n служит среднее арифметическое парциальных нормалей. Парциальная нормаль $\boldsymbol{r}_{lt}$ к собственной плоскости радиус-векторов $\boldsymbol{g}_l$ и $\boldsymbol{g}_t$, исходящих из полюса g к точкам l и t МБТ, равна нормированной гиперкомплекс-

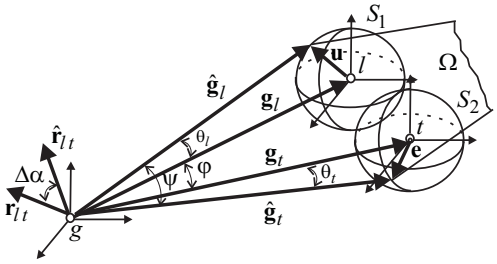


Рис. 4. Векторная диаграмма формирования зашумленной парциальной нормали $\hat{\boldsymbol{r}}_{l,t}$.

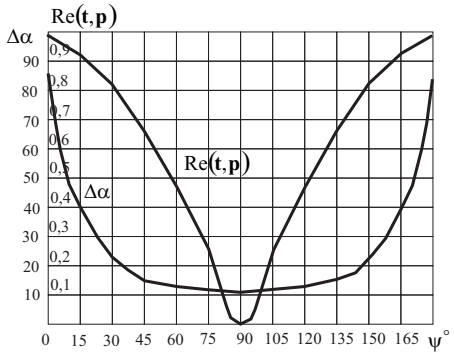


Рис. 5. Зависимости величин ошибки $\Delta\alpha$ парциальной нормали и реальной части скалярного произведения от угла между зашумленными радиус-векторами.

ной части скалярного произведения этих векторов: $\boldsymbol{r}_{l,t} = \mathrm{hypc}(\boldsymbol{g}_l, \boldsymbol{g}_t)/|\,\mathrm{hypc}(\boldsymbol{g}_l, \boldsymbol{g}_t)|$. Обеспечение помехоустойчивости нормалей к сегментам W_n, $n = 0, \ldots, s-1$, является одной из проблем при решении задачи выбора посадочной площадки БАС.

При наличии шумов, как видно из диаграммы на рис. 4, радиус-векторы $\boldsymbol{g}_l$ и $\boldsymbol{g}_t$ складываются с реализациями шумовых векторов $\boldsymbol{l}$ и $\boldsymbol{u}$. Сферы S_1 и S_2 выделяют область интенсивных шумов. Из диаграммы видно, что случайные значения углов θ_l и θ_t приводят к *изменению* ориентации собственной плоскости Ω, и, следовательно, к возникновению углового рассогласования $\Delta\alpha$ между нормалями к зашумленной и незашумленной собственным плоскостям векторов. В тех случаях, когда радиус-векторы попадают в область пересечения сфер, возникают аномальные ошибки, связанные с изменением знака нормали. По мере увеличения угла φ между радиус-векторами растет значение $\sin\psi$ в гиперкомплексной части скалярного произведения $(\hat{\boldsymbol{g}}_l, \hat{\boldsymbol{g}}_t)$, что увеличивает величину нормирующего множителя и, как следствие, ослабляет действие шума. При изменении угла ψ от 90°до 180°действие шума усиливается. На рис. 5 приведены зависимости угловой ошибки $\Delta\alpha$ и реальной части скалярного произведения от величины угла ψ для кватерниона $\boldsymbol{p} = 0{,}4364i - 0{,}2182j + 0{,}8728k$. Второй кватернион $\hat{\boldsymbol{p}}$ был получен добавлением к $\boldsymbol{p}$ кватерниона ошибки $\boldsymbol{l}$ таким образом, что отношение энергии кватернионов $\boldsymbol{p}$ и $\boldsymbol{l}$ была равно 6,25. В процессе поворота кватерниона $\hat{\boldsymbol{p}}$ вокруг фиксированной оси вычислялась ошибка в виде угла $\Delta\alpha$ между направлением этой оси и нормалью к собственной плоскости векторов $\boldsymbol{p}$ и $\hat{\boldsymbol{p}}$. Из графиков видно, что минимальное значение ошибки $\Delta\alpha$ достигается при $\psi = 90°$, причем этому же значению соответствует минимум реальной части скалярного произведения. Таким образом, при формировании парциальной нормали МБТ с минимальным уровнем зашумленности необходимо учитывать обе части скалярного произведения.

После вычисления всех парциальных нормалей МБТ путем их усреднения получается нормаль сегмента, определяется уравнение плоскости, в которой он расположен, и вычисляются проекции точек

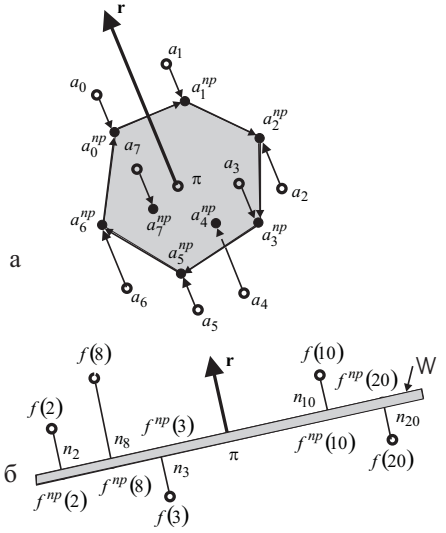


Рис. 6. Последние этапы анализа МБТ: а) контур сегмента, б) расположение точек МБТ относительно построенного сегмента.

на данную плоскость. На рис.6 показаны контур построенного сегмента и расположение точек МБТ относительно сегмента. В результате описанных выше операций, основанных на принципе МБТ, формируется векторное поле нормалей к этим сегментам. Модуль и аргумент каждой нормали характеризуют положение в пространстве каждого сегмента и степень шероховатости его реальной поверхности.

Выделение фрагмента поверхности с заданными свойствами

На данном этапе по сформированному векторному полю $R = \{r_n\}_{n=0}^{h-1}$ производится выбор участков квазиплоской формы. Основной операцией для этого служит выделение кластеров, состоящих из векторов с близкими аргументами [5]. После получения таких кластеров необходимо разрешить плоские фрагменты поверхности, обладающими близкими нормалями, но изолированные

друг от друга. Введем понятие инцидентности точек ГМТ F. Точки множества инцидентны по отношению к точке $f(n)$, если они расположены внутри окрестности C, включающей точку $f(n)$: $R_{nm} \leqslant \rho$, $m,\ n \in C,\ m \neq n$, где R_{nm} — расстояние между точками $f(n)$ и $f(m)$, ρ — радиус окрестности. Две произвольные точки множества *глобально коммутативны*, если хотя бы одна из соединяющих их траекторий полностью состоит из инцидентных точек множества F. По данному определению любые две точки ГМТ глобально коммутативны. Пусть теперь точки $f(n)$ и $f(m)$ обладают одинаковыми свойствами E. Эти точки назовем *локально коммутативными*, если хотя бы одна из соединяющих их траекторий T_r полностью состоит из инцидентных точек со свойством E. Отсюда следует, что точки двух подмножеств, обладающих свойством E, например, коллинеарностью связанных с ними векторов, но являющихся локально некоммутативными, задают два фрагмента поверхности, разделенные точками, не обладающими свойствами E.

Каждый выделенный фрагмент задается своей нормалью, значение которой получено в виде оценки максимального правдоподобия по выборке нормалей сегментов, входящих в кластер, и соответствующей этой нормали плоскостью. Аналогично, как это делалось при анализе МБТ, в этой плоскости формируется контур области, внутри которого расположены проекции соответствующих точек ГМТ. Далее в соответствии с заданными требованиями к посадочной площадке — ее площади, форме ориентации и степени шероховатости, производится сортировка выделенных квазиплоских фрагментов поверхности и выбор, по весовому критерию качества, одного из них в качестве наилучшего. Дисперсия расстояний точек ГМТ до плоского фрагмента служит характеристикой его шероховатости. Кроме того, оцениваются максимально допустимые препятствия вдоль выбранной траектории посадки. На рис. 1а приведено в виде контура изображение плоского участка поверхности, выделенного в соответствии с описанной выше методикой.

Заключение

В докладе рассмотрены теоретические подходы к решению актуальной прикладной задачи выбора на подстилающей поверхности посадочной площадки для БАС. Исходная модель поверхности представлена полем зашумленных декартовых координат ее выборочно взятых ее точек. В основу решения задачи положен описанный в работах [1, 2] алгоритм, модифицированный с целью снижения влияния шумов, высокий уровень которых связан с использованием процедуры вычисления 3D градиента, представляющей по своей сути объемный фильтр высоких частот. В докладе предлагается на базе точечного поля формировать слабозашумленное векторное поле, адекватно отражающее свойства анализируемой поверхности. Это позволяет выделять плоские фрагменты заданной формы, размеров, и с допустимой степени шероховатости.

В теоретическом плане в данной работе решены следующие новые задачи.

1. Для характеристики формы 3D поверхности предложен и реализован подход, основанный на анализе сферической окрестности из небольшого количества близко расположенных точек (метод МБТ). Для принятых условий предполагается, что МБТ задает квазиплоскость. Метод МБТ дает возможность корректно поставить в соответствие точечному полю поле плоских сегментов.

2. Проанализированы причины, вызывающие искажение сформированного векторного поля, и предложен алгоритм, ослабляющий действие шума.

В практическом плане для специалистов по обработке 3D изображений представляют интерес алгоритмы, основанные на анализе МБТ, как допускающие глубокое распараллеливание, существенно облегчающие возможность реализации режима реального времени. Кроме того, при решении всех аспектов задачи выбора площадки была применена математическая конструкция в виде скалярного (внутреннего) произведения векторов в кватернионном представлении. Она задает все характеристики пары 3D векторов: угол между ними, параллелограмм, построенный на этих векторах, нормаль к плоскости параллелограмма и саму плоскость, а также является коннектором — оператором для совмещения векторов [6]. Высокая степень универсальности такого инструментария делает актуальной задачу создания микросхемы частного применения, позволяющей получить результат за один машинный такт, что также повышает возможность достижения режима работы в реальном времени.

Литература

[1] *Zucker S. W., Hummel R. A.* A Three Dimensional Edqe OperatorIEEE Trans. Patter Anal. Mach. Intell, PAMI, 1981. — Vol. 3, No. 3., Pp. 324–331.

[2] *Shirai Y.* Three Dimensional Computer Vision, in Computer Vision and Sensor — Based Robots (G. G. Dodd and L. Rossol, eds), Plenum N.Y., 1979.

[3] *Бакулев П. А.* Радиотехнические системы — М.: Радиотехника, 2005.

[4] *Ефимов Н. В., Розендорн Э. Р.* Линейная алгебра и многомерная геометрия — М.: Наука, 1974.

[5] *Фурман Я. А.* Сегментация и описание трехмерных структур на базе кватернионных моделей // Наукоемкие технологии, 2007. — № 9. Т. 8. — С. 37–49.

[6] *Фурман Я. А., Рябинин К. Б.* Нахождение параметров вращения пространственного группового точечного объекта по результатам его фильтрации // Радиотехника и электроника, 2008. — Т. 53. № 1. — С. 78–89.

Комбинирование ациклических графов соседства в задаче распознавания марковских случайных полей*

Савенков Д. С., Двоенко С. Д., Шанг Д. В.
denissavenkov@gmail.com, dsd@uic.tula.ru, dvietsang@gmail.com
Тула, Тульский государственный университет

Рассматривается задача распознавания объектов, образующих взаимосвязанный массив, представленный как двухкомпонентное случайное марковское поле скрытых классов объектов и их наблюдаемых признаков. Алгоритм распознавания опирается на набор ациклических графов соседства элементов массива данных для снижения потерь при древовидной аппроксимации исходного графа соседства. Разработан алгоритм обучения для определения весов в линейной комбинации ациклических графов соседства. Оценка качества обучения выполнена методом скользящего контроля.

Введение

В теории распознавания образов основное предположение заключается в независимости объектов, предъявленных для распознавания. Но во многих случаях необходимо принимать скоординированные решения о классах объектов.

В задачах обработки такое множество объектов часто понимается как взаимосвязанный массив, элементы которого естественным образом упорядочены, например, вдоль оси времени или пространственной координаты. Элементы такого массива естественно понимать как «соседние», «смежные», «упорядоченные», выразив это свойство соответствующим графом соседства.

Скрытые марковские модели доказали свою эффективность для массивов линейно упорядоченных объектов с цепочечными графами соседства. Но их применение для массивов с произвольными графами соседства элементов оказалось проблематичным [1]. Таким образом, в общем случае эффективное решение возможно лишь для частных видов задач, например [2].

В [3, 4] была предложена модель марковского случайного поля с древовидным графом соседства элементов взаимосвязанного массива и построен эффективный алгоритм распознавания.

Но корректная редукция графа соседства произвольного вида обычно требует в каждом случае разработки специального алгоритма, не уступающего по сложности собственно алгоритму распознавания [5].

Поэтому было предложено [6, 7, 8] заменить исходный граф соседства элементов (например, решетку для растрового изображения) набором ациклических графов соседства. В этом случае редуцированное множество взаимосвязей между элементами массива в древовидном графе компенсируется расширением самого множества ациклических графов разных конфигураций, каждый из которых покрывает все множество вершин.

*Работа выполнена при финансовой поддержке РФФИ, проекты №08-01-12023, 08-01-99003, 09-07-00394.

Эффективный базовый алгоритм распознавания с древовидным графом соседства

Пусть T является массивом взаимосвязанных элементов $t \in T$, который представлен как двухкомпонентное случайное поле (X, Y). Это поле состоит из скрытой компоненты $X = (x_t, t \in T)$, подлежащей восстановлению, где $x_t \in \mathcal{X}$ — классы $\mathcal{X} = \{1, \ldots, m\}$ элементов массива $t \in T$ и наблюдаемой компоненты $Y = (\boldsymbol{y}_t, t \in T)$.

Взаимосвязи между элементами $t \in T$ представлены графом соседства G с ненаправленными ребрами без петель. Как показано в [3, 4], мы остаемся в рамках классической задачи распознавания на этапе обучения при условии, что наблюдения условно независимы относительно реализации скрытого случайного поля классов объектов

$$\psi_t(\boldsymbol{y}_t \,|\, X) = \psi_t(\boldsymbol{y}_t \,|\, x_t).$$

Задача заключается в восстановлении скрытого случайного поля классов объектов X по реализации поля Y.

Пусть скрытое случайное поле классов X восстанавливается по реализации Y на основе байесовского решающего правила

$$\hat{X} = (\hat{x}_t, t \in T), \quad \hat{x}_t = \arg\max_{x_t \in \mathcal{X}} p_t(x_t \,|\, Y).$$

Результатом этапа независимого обучения являются апостериорные маргинальные распределения $p_t(x_t \,|\, \boldsymbol{y}_t)$. Процедура распознавания находит апостериорные маргинальные распределения $p_t(x_t \,|\, Y)$, $t \in T$. Как показано в [3], априорное поле классов X является односторонним марковским

$$q_t(x_t \,|\, X_{(t)}) = q_t(x_t \,|\, x_r),$$

где $X_{(t)}$ является полем без элемента x_t, а вершина t является потомком вершины r относительно дерева G. Апостериорное случайное поле X остается марковским с тем же графом соседства G:

$$p_t(x_t \,|\, X_{(t)}, Y) = p_t(x_t \,|\, x_r, Y_t^+),$$

где Y_t^+ является поддеревом с корнем $\boldsymbol{y}_t$ в дереве, которое представляет поле Y.

Скрытое поле X при определенных условиях распознается за два прохода по дереву G [4].

Сначала вычисляются апостериорные маргинальные распределения $p_t(x_t \,|\, Y_t^+)$ при восходящем просмотре от терминальных вершин, для которых принимается

$$p_t(x_t \,|\, Y_t^+) = p_t(x_t \,|\, \boldsymbol{y}_t).$$

Восходящий просмотр заканчивается в корне дерева, где

$$p_t(x_t \,|\, Y_t^+) = p_t(x_t \,|\, Y).$$

При нисходящем просмотре из корня вычисляются апостериорные маргинальные распределения $p_t(x_t \,|\, Y)$, на основе которых определяются элементы скрытого поля X.

Назовем такую процедуру базовым алгоритмом распознавания

Алгоритм распознавания с комбинированием ациклических графов соседства

Произвольный граф G соседства элементов взаимосвязанного массива не может быть заменен древовидным графом без потери фундаментального свойства исходного графа представлять полную информацию о каждом элементе $t \in T$ относительно других элементов. Например, решетка является графом соседства элементов в растровых изображениях и не является ациклическим графом.

Восстановим скрытое случайное поле X на основе линейной комбинации апостериорных маргинальных распределений классов объектов относительно заранее заданных ациклических графов соседства.

Например, для растровых текстурных изображений с протяженными областями однородности удобно использовать графы, некоторые из которых показаны на рис. 1.

Возьмем граф G_k из заданного набора и применим базовый алгоритм распознавания, описанный выше. Далее, для другого графа соседства снова применим базовый алгоритм распознавания, и т. д. Применение такого алгоритма в отдельности для каждого ациклического графа из заранее заданного набора сформирует на основе начальных распределений $p_t(x_t \,|\, \boldsymbol{y}_t)$ множество распределений

$$p_t^k(x_t \,|\, Y), \quad t \in T, \quad k = 1, \ldots, K,$$

и соответствующих решений о классах $\hat{x}_t^k, \ t \in T$, независимо для каждого графа, где K — число ациклических графов.

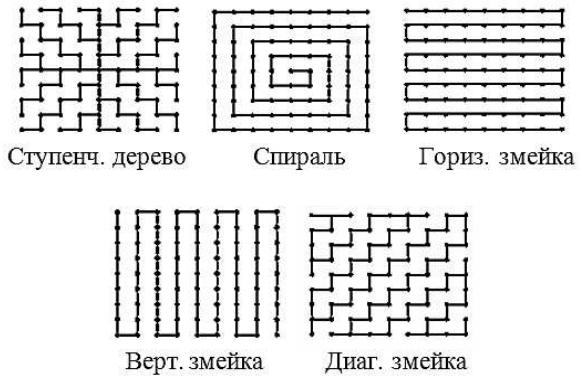


Рис. 1. Ациклические графы соседства элементов взаимосвязанного массива.

Чтобы получить окончательные решения о классах $\hat{x}_t, \ t \in T$, вычислим апостериорное маргинальное распределение $p_t(x_t \,|\, Y)$ в каждом элементе $t \in T$ как взвешенную сумму апостериорных распределений:

$$p_t(x_t \,|\, Y) = \sum_{k=1}^{K} w_k p_t^k(x_t \,|\, Y), \quad \sum_{k=1}^{K} w_k = 1,$$

где $t \in T$, $w_k > 0$.

Назовем это комбинированием. Выполним повторение комбинирования. В этом случае возьмем только что найденные апостериорные распределения $p_t(x_t \,|\, Y)$, $t \in T$ в качестве исходных для каждого ациклического графа соседства в отдельности и вновь повторим комбинирование. Будем его выполнять до тех пор, пока результат не перестанет изменяться. Отметим, что комбинирование с равными весами ациклических графов является частным случаем взвешенного комбинирования.

Назовем такую процедуру итерационным алгоритмом распознавания с комбинированием ациклических графов соседства элементов взаимосвязанного массива.

Обучение распознаванию на основе ациклических графов соседства

Очевидно, что каждый ациклический граф соседства отражает лишь некоторое подмножество взаимосвязей элементов исходного массива данных. Поэтому возникает задача подбора весов графов соседства для конкретного типа массивов данных, предъявленных для распознавания. Рассмотрим ее как задачу обучения распознаванию на основе ациклических графов соседства.

В данной работе предложена процедура определения весов при комбинировании ациклических графов соседства на основе известного алгоритма Гаусса-Зайделя покоординатного спуска.

Будем считать аналогом покоординатного варьирования изменение веса графа G_k в диапазоне от 0 до 1. На начальном шаге распределение весов

всех K графов изменяется от $\frac{1}{K-1}, \ldots, 0, \ldots, \frac{1}{K-1}$, когда граф G_k полностью исключен, а веса остальных $K-1$ графов одинаковы, до распределения $0, \ldots, 1, \ldots, 0$, когда применяется только граф G_k, а остальные графы исключены.

Шаг заканчивается после варьирования весов всех графов и выбора того графа, варьирование веса которого обеспечило минимальное число ошибок распознавания на обучающем множестве массивов.

Пусть теперь на очередном шаге варьируется вес графа G_k в диапазоне $0 \leqslant p \leqslant 1$. Нормированный вес данного графа в линейной комбинации имеет значение $w_k = p$. Остальные графы имеют определенные к данному шагу постоянные веса q_i, $i = 1, \ldots, K$, $i \neq k$, где их сумма также постоянна: $Q = \sum\limits_{i=1}^{K} q_i$, $i \neq k$.

Нормированные веса остальных графов меняются в диапазоне от $w_i = q_i/Q$ до $w_i = 0$, где на интервале варьирования их нормированные веса принимают значения $w_i = q_i(1-p)/Q$.

Отметим, что исходная точка с весами $p = q_k$, q_i, $i = 1, \ldots, K$, $i \neq k$ находится внутри интервала варьирования, где $w_j = q_j$, $j = 1, \ldots, K$.

Результат каждого пробного варьирования проверяется однократным комбинированием уже вычисленных маргинальных распределений $p_t^k(x_t \,|\, Y)$, $t \in T$, $k = 1, \ldots, K$, с подсчетом числа ошибок распознавания для решений $\hat{x}_t$, $t \in T$.

Очередной шаг заканчивается после варьирования весов всех графов и выбора нового веса того графа, для которого было получено минимальное число ошибок распознавания.

Оценка качества обучения методом скользящего контроля

Очевидно, что таким способом подбора весов можно сразу же получить и наилучшую оценку скрытой компоненты X конкретного массива данных. Но в теории распознавания образов это означает подгон результата к особенностям этого массива, то есть переобучение. Это, вообще говоря, может не позволить получить приемлемый результат при распознавании другого массива. Как известно, метод скользящего контроля позволяет оценить обобщающую способность алгоритма обучения.

Пусть дано множество массивов, для которых известны значения скрытой компоненты X. Разделим это множество на n групп, где каждая группа может состоять просто из одного массива. Будем поочередно исключать очередную группу и выполнять подбор весов на оставшихся группах. После того, как веса подобраны, протестируем результат на группе, не участвовавшей в обучении, оценив среднюю ошибку распознавания. В итоге, оценим общую среднюю ошибку распознавания.

Рис. 2. Классы, указанные учителем для двух текстур на шести растровых изображениях размером 201×201.

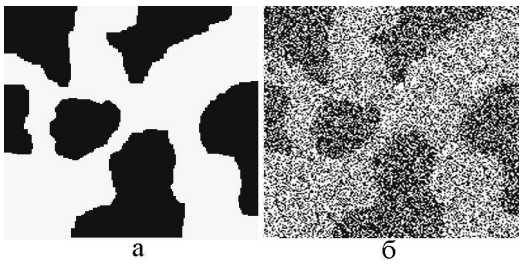


Рис. 3. (а) классы учителя; (б) независимое распознавание — 13727 ошибок (33,98%); цвета смешаны пропорционально вероятностям классов.

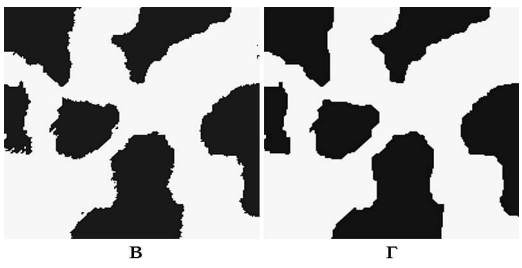


Рис. 4. (в) равные веса графов — 822 ошибки(2,03%); (г) веса (0,190; 0,015; 0,300; 0,322; 0,173) подобраны на других изображениях — 616 ошибок (1,52%); цвета смешаны пропорционально вероятностям классов.

Экспериментальные результаты

На рис. 2 представлено обучающее множество растровых изображений размером 201×201 для распознавания текстур двух классов.

Текстуры классов получены как реализации двумерных нормально распределенных случайных величин с одинаковыми дисперсиями и немного различающимися средними значений зеленого и красного цвета. На рис. 3 для первого изображения, участвовавшего в скользящем контроле, показаны классы текстур (а) и результат независимого распознавания (б). На рис. 4 для него показаны результат комбинирования с равными весами графов соседства (в) и результат с подбором весов (г).

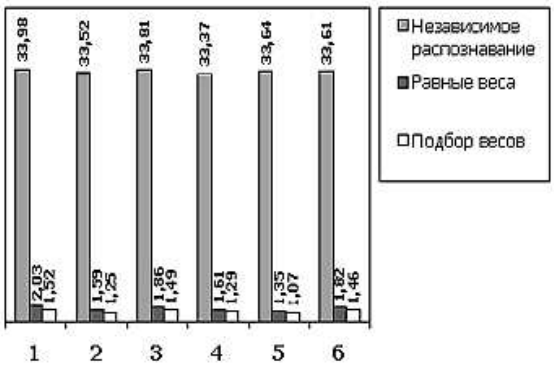


Рис. 5. Среднее значение процента ошибок распознавания текстур двух классов для каждого изображения при скользящем контроле.

На рис. 5 показаны значения ошибок в логарифмической шкале по каждому изображению рис. 2 (слева направо и сверху вниз). В итоге получены средние значения ошибок: 33,65% — при независимом распознавании, 1,71% — при равных весах, 1,35% — при подборе весов.

Заключение

Древовидная аппроксимация графа соседства значительно искажает характер взаимосвязей элементов в массиве данных.

На этапе распознавания принятие решений выполняется на основе взвешенных апостериорных маргинальных распределений, построенных для каждого ациклического графа в отдельности, что позволяет получить лучшее качество распознавания, чем при независимом распознавании и при использовании ациклических графов с равными весами.

Рассмотрен алгоритм обучения распознаванию классов текстур растровых изображений, позволяющий определить веса заранее заданного множества ациклических графов.

Оценка качества обучения выполнена методом скользящего контроля.

Литература

[1] *Li S. Z.* Markov Random Field Modeling in Image Analysis. — London: Springer-Verlag, 2009. — 362 p.

[2] *Schlesinger M. I., Flach B.* Some solvable subclasses of structural recognition problems // Proc. of Czech Patt. Recogn. Workshop, Praha, 2000. — Pp. 55–62.

[3] *Двоенко С. Д., Копылов А. В., Моттль В. В.* Задача распознавания образов в массивах взаимосвязанных объектов. Постановка задачи и основные предположения // Автоматика и телемеханика. — 2004. — № 1. — С. 143–158.

[4] *Двоенко С. Д., Копылов А. В., Моттль В. В.* Задача распознавания образов в массивах взаимосвязанных объектов. Алгоритм распознавания // Автоматика и телемеханика. — 2005. — № 12. — С. 162–176.

[5] *Mottl V. V., Dvoenko S. D., Levyant V. B., Muchnik I. B.* Pattern recognition in spatial data: a new method of seismic explorations for oil and gas in crystalline basement rocks // Proc. 15th ICPR, Barcelona: IEEE CS, 2000. — V. 2. — Pp. 315–318.

[6] *Двоенко С. Д., Савенков Д. С.* Эффективное распознавание взаимосвязанных объектов на основе ациклических марковских моделей // Всеросс. конф. ММРО-13. — Москва: МАКС Пресс, 2007. — С. 302–305.

[7] *Dvoenko S., Savenkov D.* The effective recognition of interrelated objects based on acyclic Markov models // 8th Intern. Conference, PRIA-8-2007, Oct. 8–13, Yoshkar-Ola: RF, 2007. — V. 1. — Pp. 55–57.

[8] *Dvoenko S., Savenkov D.* Selecting of adjacency acyclic graphs for recognition of raster textured images // 9th Intern. Conference, PRIA-9-2008, Sept. 14–20, N. Novgorod: RF, 2008. — V. 2. — Pp. 143–146.

Система верификации владельца карманного компьютера по фотопортрету

Степалина Е. А.
estepalina@mail.ru
Тула, ТулГУ

Целью работы является описание системы верификации владельца карманного компьютера по фотопортрету. В работе приведён пример реализации такой системы для платформы Windows Mobile.

Всё больше конфиденциальной информации хранится на мобильных устройствах, в частности, на карманных персональных компьютерах (КПК). В связи с этим возрастает актуальность защиты информации от несанкционированного доступа на таких устройствах. В целях защиты информации на устройство можно установить систему верификации владельца по биометрическим характеристикам.

Так, например, на ноутбуках успешно применяется верификация по отпечатку пальца владельца устройства, которая, как показывает опыт, даёт около 80% правильных ответов. Для КПК можно реализовать систему верификации по фотопортрету, получаемому со встроенной камеры. При этом не потребуется никаких дополнительных аппаратных элементов, сканеров отпечатков пальцев и т. п.

Целью работы является рассмотрение прототипа такой системы и возможности его реализации для платформы Windows Mobile.

Постановка задачи

Верификация владельца по фотопортрету включает в себя следующие задачи:

— задача обнаружения лица на фотографии (face detection);
— задача распознавания лица (face recognition).

Эти задачи являются частными случаями задачи классификации. Пусть X — множество описаний объектов, Y — конечное множество классов,

$$K^m = \{(x_1, y_1), \ldots, (x_m, y_m)\}$$

— обучающая выборка, в которой $x_i \in X$, $y_i \in Y$, $i = 1, \ldots, m$. Существует неизвестная целевая зависимость — отображение $y^*\colon X \to Y$, значения которой известны только на объектах обучающей выборки. Требуется построить решающую функцию (решающее правило) отнесения произвольного объекта $x \in X$ к некоторому классу $y \in Y$.

В случае задачи обнаружения лица множество описаний объектов — это фотографии, а множество классов содержит два элемента {«фотография содержит лицо», «фотография не содержит лица»}.

В задаче распознавания лиц множество описаний объектов — это лица, выделенные на фотографиях, а множество классов также состоит из двух элементов {«Владелец», «Не владелец»}.

Существует большое число методов решения задач классификации [1]. Для выбора конкретных методов и алгоритмов решения нужно рассмотреть условия, в которых должна работать система верификации.

Выбор мобильной платформы

Мобильная платформа в контексте данной работы — это совокупность программных средств, предназначенных для работы на аппаратных средствах определённой архитектуры. Компания Microsoft — один из лидеров по поставке мобильных платформ и операционных систем для КПК и встраиваемых устройств `CNews.ru`. Она предлагает модульное ядро операционной системы Windows Compact Edition (WinCE), на основе которого можно создавать различные мобильные платформы для конкретных задач. Эти платформы работают в разнообразных встраиваемых устройствах: от промышленных датчиков до «умных» светофоров и MP3-плееров. На основе WinCE созданы платформы Pocket PC, Windows Mobile 5 и 6 для КПК. Разработка систем верификации владельца актуальна для этих платформ.

Особенности разработки программного обеспечения для карманных компьютеров

Можно выделить следующие особенности карманных компьютеров, которые необходимо учитывать при разработке программного обеспечения для них:

— Мощность и объемы оперативной памяти карманных устройств малы для выполнения трудоемких вычислений, связанных с работой алгоритмов распознавания. Особенно это касается процесса обучения алгоритма, который требует значительных вычислительных мощностей. КПК может попросту не хватить памяти для завершения вычислений, не говоря о том, сколько времени потребуется на их выполнение.
— Мобильные устройства весьма ограничены в своих размерах, поэтому размещение в них дополнительных устройств, например сканера отпечатков пальцев, может оказаться затруднительным.
— Существуют удобные инструменты синхронизации систем КПК и настольного компьюте-

ра. Они позволяют передавать файлы между устройствами, организовывать межпроцессное взаимодействие и т. п.

В связи с этим обучение алгоритма распознавания и построение решающего правила можно поручить персональному компьютеру (ПК). После этого решающее правило в виде файла можно передать на КПК и далее на устройстве выполнять процедуру применения готового правила. Таким образом, приложение на КПК получится легким и не потребует значительных ресурсов устройства. Следует отметить, что, если существует возможность переноса кода, написанного для настольных операционных систем, в программы для мобильных устройств, это также упрощает разработку системы.

Структура системы

В соответствии с указанными особенностями удобно организовать систему верификации в виде распределённой системы, состоящей из двух частей: приложения для ПК, и приложения для КПК (рис. 1).

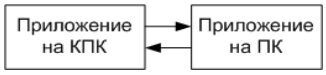


Рис. 1. Структура системы верификации.

Предполагается, что система верификации личности владельца мобильного устройства будет работать следующим образом:

1. *Подготовка к обучению.* Владелец делает несколько фотографий своего лица на камеру устройства. Приложение на КПК детектирует лица на них и передает соответствующие фрагменты фотографий на ПК.
2. *Обучение.* На компьютере запускается программа, которая строит решающее правило для распознавания владельца в виде файла и отправляет его на КПК.
3. *Распознавание.* Для верификации владельцу нужно сделать фотографию своего лица. Приложение на КПК выделит лицо на фотографии и применит для него решающее правило. Если владелец будет «узнан», то доступ к информации, хранящейся на устройстве, будет разрешен.

Каждое из этих действий предполагается реализовать в виде отдельных компонент распределённой системы, см. рис. 2, 3.

Технологии реализации системы

Для программной реализации системы для платформы Windows Mobile существуют все необходимые технологии. Структура системы с точки зрения возможных технологий реализации ее компонентов представлена на рис. 4, 5.

Рис. 2. Структура системы верификации (приложение для КПК).

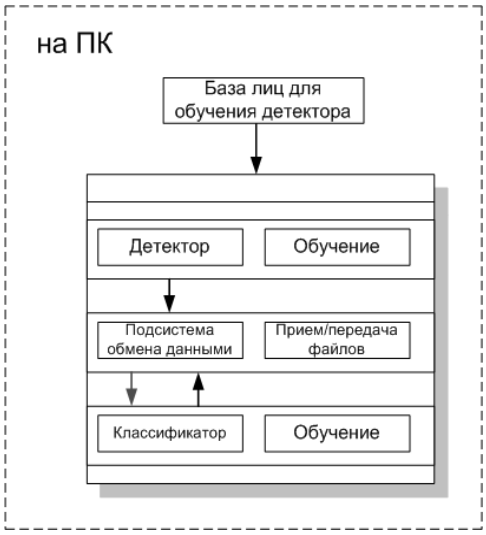


Рис. 3. Структура системы верификации (приложение для ПК).

Для детектирования лица на мобильном устройстве, требуется быстрый и точный алгоритм. Такой алгоритм реализован в библиотеке OpenCV [4], известной разработчикам систем распознавания. Кроме того, библиотека OpenCV после небольших изменений исходного кода успешно компилируется для WinCE-платформ.

Детектор объектов в OpenCV строится в виде xml-описания каскада простых CART классификаторов, усиленных (boosted) по алгоритму Gentle AdaBoost [2, 3]. При этом каждый классификатор работает на haar-like признаках, что обеспечивает высокую скорость детектирования, но не гаранти-

Рис. 4. Технологии реализации компонентов системы на КПК.

Рис. 5. Технологии реализации компонентов системы на ПК.

рует высокой точности [4]. Для достижения максимальной точности детектирования можно применять другие методы, например PCA [1]. Детектор в виде xml-файла с описанием классификаторов удобно передавать на КПК.

Для OpenCV имеются готовые детекторы, построенные на различных базах. Было взято 2 готовых детектора: поставляемый с OpenCV и построенный на базе CBCL [7]. Для проверки их работоспособности в разрабатываемой системе была приготовлена тестовая выборка, содержащая 50 фотографий 10-ти различных персон, снятых камерами 2 и 3 Мрх нескольких мобильных устройств. Для грубого определения эффективности детекторов 50-ти фотографий оказалось достаточно; однако в дальнейшем рассматривается возможность увеличить размер тестовой выборки для более точной оценки. В равном количестве в тестовой выборке присутствовали фотографии: помещение с естественным, с естественным и искусственным освещением, улица при дневном свете (облачно). Тестирование детектора, предлагаемого OpenCV-разработчиками, показало высокий уровень ложных срабатываний, а детектор, обученный на базе OBCL, выделяет недостаточную часть лица, см. рис. 6, 7. Готовые детекторы оказались неподходящими для работы в системе, и возникла необходимость самостоятельного обучения OpenCV.

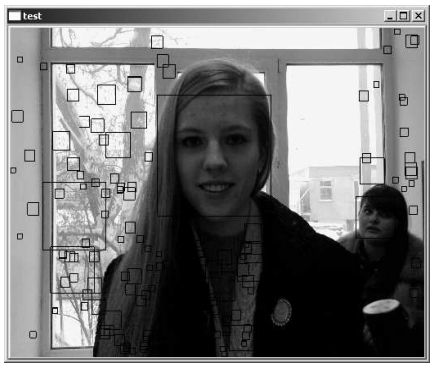


Рис. 6. Детектирование лица детектором OpenCV.

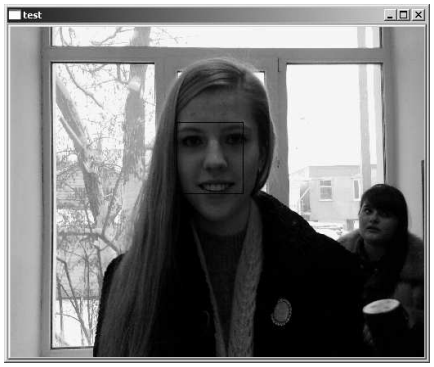


Рис. 7. Детектирование лица детектором CBCL.

Было решено проводить обучение OpenCV на фотографиях из базы FERET [8]. Она содержит 11338 фотографий 994 человек, снятых в период с 1993 по 1996-й год. FERET является крупнейшей базой фотографий лиц, используемой разработчиками систем распознавания по всему миру. В качестве позитивных (содержащих лицо) изображений, было выбрано 2722 фотографии лиц в анфас. Также для обучения использовались 3018 негативных (не содержащих целевого объекта) фотографий [4]. На их основе была подготовлена обучающая выборка черно-белых картинок размером 2020 пикселов. Обучение было выполнено при следующих параметрах: частота ошибочных срабатываний не более $5 \cdot 10^{-6}$, частота правильных ответов $0,995$. Параметры обучения выбирались согласно исследованиям Куранова [3]. Процесс обучения занял около 5 суток, и в результате было построено 16 стадий классификатора. Для тестирования использовалась та же выборка, что и для тестирования готовых детекторов. На тестовой выборке построенный детектор показал приемлемые ре-

Таблица 1. Результаты тестирования детектора, обученного на FERET.

Классифицировано	Тип фотографии	
	Содержащие лицо	Не содержащие лица
Верно	42	3
Ошибочно	8	7
Всего	50	10

зультаты (см. рис. 8), что позволяет использовать его в реальной системе. Результаты тестирования представлены в таблице 1.

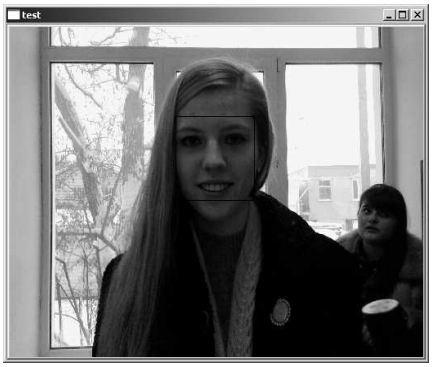


Рис. 8. Детектирование лица детектором, обученным на FERET.

Для решения задачи распознавания владельца можно использовать метод опорных векторов (Support Vector Machine, SVM), также реализованный в OpenCV. Обучение представляет собой подбор оптимальных параметров one-class SVM-модели. Эту модель следует выбрать потому, что обучающая выборка формируется из описаний объектов только одного из двух классов — класса «Владелец». В качестве обучающей выборки можно использовать фрагменты нескольких фотографий, содержащие лицо владельца. Необходимое число фрагментов предполагается выбирать экспериментально, исходя из требуемого качества распознавания. Синхронизация компьютеров для передачи файлов между ними возможна посредством Remote API (RAPI) версии 2 [5]. RAPI предназначен для синхронизации мобильных и настольных Windows-систем. Этот интерфейс содержит все необходимые системные функции для установления соединения, удаленного создания потоков, записи на диск и т. п.

Для создания интерфейса пользователя серверного приложения (на ПК) можно использовать подмножество функций OpenCV-HighGUI, которые позволяют создавать простые оконные приложения для настольных систем. Операционные системы, основанные на WinCE, используют несколько отличный API-интерфейс для работы с окнами, нежели тот, на котором построено HighGUI, поэтому реализовать интерфейс клиентского приложения на HighGUI нельзя. Для этого можно использовать другие библиотеки, например кроссплатформенную wxWidgets [6].

Выводы

Для защиты информации на мобильных устройствах можно использовать систему верификации личности владельца по фотопортрету. Существуют все необходимые средства реализации такой системы для платформы Windows Mobile, на которой работают современные карманные компьютеры. Для снижения вычислительной нагрузки на мобильное устройство трудоёмкие задачи по построению решающих правил для классификации объектов можно выполнять на настольном компьютере, а на мобильном устройстве применять уже готовое решающее правило.

В работе описан прототип системы верификации личности, и приведён пример его технологической реализации для платформы Windows Mobile. По фотографиям лиц из базы FERET обучен детектор, реализованный в библиотеке OpenCV. Этот детектор даёт приемлемые результаты на фотографиях, сделанных на камеру карманного компьютера, то есть пригоден для использования в реальной системе. Требуется обучить выбранную SVM-модель для непосредственной верификации владельца, реализовать синхронизацию приложений между ПК и КПК, разработать пользовательский интерфейс системы.

Литература

[1] *Журавлёв Ю. И., Рязанов В. В., Сенько О. В.* «Распознавание». Математические методы. Программная система. Практические применения. — М.: Фазис, 2006.

[2] *Lienhart R., Maydt J.* An Extended Set of Haar-like Features for Rapid Object Detection. // IEEE ICIP, 2002. — Pp. 900–903.

[3] *Kuranov A., Lienhart R., Pisarevsky V.* An Empirical Analysis of Boosting Algorithms for Rapid Objects With an Extended Set of Haar-like Features. // Intel Technical Report MRL-TR — 2002. — July 02, 01.

[4] `http://note.sonots.com/SciSoftware/haartraining/document.html` — Naotoshi Seo. — 2002.

[5] Microsoft Developer Network. — 2009 — `http://msdn.microsoft.com/en-us/library/ms894770.aspx` —

[6] `www.wxwidgets.org` — wxWidgets. Cross-platform GUI library. — 2009.

[7] `http://cbcl.mit.edu/` — Center for Biological & Computational Learning. — 2009.

[8] `http://face.nist.gov/frvt/feret/feret.htm` — Face Recognition Technology (FERET). — 2009.

Алгоритм векторизации штриховых бинарных изображений

Стержанов М. В., Байдаков И. В.
accept@bk.ru
Минск, Белорусский государственный университет информатики и радиоэлектроники

В данной работе предлагается алгоритм векторизации, основанный на построении промежуточной графовой модели бинарного растра. Предлагаемый метод является дальнейшим развитием идей Ди Зензо, Кропача, Монагана и др. с попыткой улучшения качества за счёт аналитической обработки мест соединений и улучшения производительности.

При сканировании широкоформатных чертежей получается растровое изображение большого размера, что накладывает определенные ограничения на применяемые алгоритмы обработки. В частности, необходима структура данных, которая будет обеспечивать компактное хранение изображения, сохраняя его топологию. Рассмотрим сильные и слабые стороны некоторых методов представления растровой информации. Операция утоньшения позволяет представить объекты на растре линиями единичной ширины. Скелетизированное изображение сохраняет топологию, однако оно чувствительно к шуму, места соединений обрабатываются не всегда корректно. Некоторые алгоритмы скелетизации не сохраняют связность. Алгоритмы выделения контуров можно условно разбить на две группы: отслеживающие и сканирующие. Недостатком контурного препарата является то, что по нему трудно построить топологию исходного изображения. Граф смежности линий является удобным способом представления изображения, состоящего из большого числа горизонтальных или вертикальных отрезков, однако данная структура не хранит информацию о местах соединения, что, безусловно, является серьезным недостатком.

В данной работе предлагается представление каждого объекта изображения в виде планарного нагруженного ориентированного псевдографа, в котором все ребра суть прямолинейные отрезки или дуги плоских кривых, а вершины — точки на плоскости, являющиеся концами отрезков или точками сочленения нескольких отрезков.

Данное представление позволяет обеспечить достаточную степень сжатия информации, и в то же время обрабатывать изображение напрямую в кодированной форме, что позволяет при обработке изображений большого размера обойтись без разбиения на фрагменты с последующей «сшивкой».

Используемая терминология

Под серией будем понимать последовательность пикселей, имеющих одинаковое значение [1]. В зависимости от ориентации серия может быть горизонтальной или вертикальной. Серия однозначно определяется с помощью четырех значений:

- d — направление (вертикальное/горизонтальное);
- pos — номер столбца (строки) матрицы изображения, которому принадлежит серия;
- beg — номер строки (столбца) матрицы изображения, которому принадлежит первый пиксель серии;
- end — номер строки (столбца) матрицы изображения, которому принадлежит последний пиксель серии.

Две серии A и B называются смежными, если выполняются условия (1)–(4):

$$A.d = B.d; \tag{1}$$

$$\bigl|A.\mathrm{pos} - B.\mathrm{pos}\bigr| = 1; \tag{2}$$

$$A.\mathrm{beg} \leqslant B.\mathrm{end} + 1; \tag{3}$$

$$A.\mathrm{end} \geqslant B.\mathrm{beg} - 1. \tag{4}$$

Под путем из серии A к серии B будем понимать последовательность серий $A = A_1, A_2, \ldots, A_n = B$ таких, что A_i является смежной с A_{i+1} для $1 \leqslant i \leqslant n-1$.

Рассмотрим две смежные серии. Они находятся в отношении «родитель–потомок». Серию с меньшим значением pos будем называть родительской. Серию с большим значением pos будем считать дочерней.

Серия будет называться нормальной, если она имеет ровно одного родителя и ровно одного потомка. В противном случае серия является особой.

Под начальной будем понимать серию, не имеющую родителей. Под конечной будем понимать серию, не имеющую потомков. Под серией слияния будем понимать серию, имеющую более одного родителя. Под серией ветвления будем понимать серию, имеющую более одного потомка.

Под полосой будем понимать связное множество нормальных серий. Полной полосой будем называть полосу, которая не является подмножеством другой полосы. Полоса представляет отдельную «ветвь» изображения. Полоса может содержать строго одну серию из столбца изображения, т. е. в полосе отсутствуют случаи ветвления и слияния серий. Под длиной полосы будем понимать количество серий, которые образуют полосу. Под весом полосы будем понимать суммарное число отсчетов ее серий. Если квадрат длины полосы больше ее веса, то полоса закрывается.

Листинг первым ввел понятие линейного скелета [2], который образуется в результате континуального «сжатия» области (без изменения топологии) с выделением подмножества пикселей единичной толщины. Средняя ось формируется центрами дисков максимального радиуса, помещенных внутрь области [3].

Скелетной кривой (СКР) в непрерывном пространстве будет являться либо линейный скелет, либо средняя ось, сохраняющие топологические или геометрические признаки [4]. СКР задается множеством из N целочисленных точек pnt_0, pnt_1, …, pnt_{N-1} и имеет характеристику ширины. На атрибуты СКР задаются следующие ограничения:

$$N \geqslant 3; \tag{5}$$

$$\left|\mathrm{pnt}.x[y]_{i+1} - \mathrm{pnt}.x[y]_i\right| \leqslant 1,\ i = 0, \ldots, N-2; \tag{6}$$

$$\left|\mathrm{pnt}.y[x]_{i+2} - \mathrm{pnt}.y[x]_i\right| \leqslant 4,\ i = 0, \ldots, N-3. \tag{7}$$

Заметим, что условия (6)-(7) являются симметричными относительно координат x и y.

Закрытием полосы является формирование СКР по центральным точкам серий полосы (рис. 1). Серии, образовавшие СКР, удаляются из полосы. СКР представляются только те серии полосы, чьи центральные точки соответствуют условиям (5)–(7). Например, полоса C (рис. 1) не представляется своей центральной точкой, т. к. нарушается условие (7).

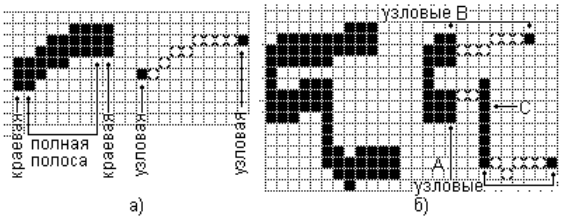


Рис. 1. Пример закрытия полос: а) узловые серии являются краевыми; б) узловые серии A, B, C отличаются от краевых.

Рассмотрим полную полосу S. Серии, являющиеся родительской и дочерней по отношению к первой и последней сериям S соответственно, назовем краевыми. Обозначим краевые серии через E_1 и E_2. Узловыми будем называть серии, к которым прикрепляются СКР при закрытии полосы. Если краевая серия является начальной или конечной, то она является узловой. В этом случае узловая серия «сужается» до своей центральной точки. Пусть краевая серия E_1 будет серией слияния или ветвления. Рассмотрим путь E_1E_2. Тогда узловой будет серия E_{1+n}. В данном случае размеры серии не изменяются. Такой выбор узловых серий сохраняет топологию соединения.

Пусть имеется полная полоса S, состоящая из n серий R. Общее число отсчетов серий полосы S равняется m. После закрытия она представляется кривой C. Вычисляются следующие морфологические свойства. Площадь рассчитывается просто как количество пикселей, соответствующих полосе. Длина вычисляется как евклидово расстояние между центрами узловых серий. Ширина W рассчитывается по формулам:

$$W = \frac{m}{\sum\limits_{i=1}^{n} \sqrt{1 + w_i^2}};$$

$$w_i = \frac{R_i.\mathrm{beg} + R_i.\mathrm{end} - (R_{i-1}.\mathrm{beg} + R_{i-1}.\mathrm{end})}{2}.$$

Построение графовой модели

Каждый столбец матрицы изображения представляется упорядоченным по возрастанию координат списком серий. Выделим связные компоненты (СК) изображения. Два пикселя (B или W) называются связными, если они являются соседями (расстояние между ними равно 1) в выбранной метрике. Связная компонента изображения (СК) — это связное множество пикселей в соответствии с выбранным типом метрики [1]. Мы будем использовать 8-связную метрику. СК можно считать единственной структурной единицей растрового изображения. Благодаря RLE-кодированию возможно использование эффективного алгоритма [5]. Его основная идея заключается в том, что метка СК ассоциируется не с отдельным пикселем, а с серией. Однако для больших изображений размер таблицы эквивалентости является фактором, снижающим производительность. Мы применяем методику «разделяй и властвуй» для нахождения СК изображений большого размера. Основываясь на работе [6], разделим изображение на $N \times N$ частей. Для каждой части применим алгоритм [5], затем выполним процедуру слияния. Затем выполним частичную скелетизацию СК, скан-проход которой заключается в следующем. Изображение сканируется по вертикали, анализируется связность смежных серий и выделяются полосы. Найденные полосы закрываются.

Каждая СК соответствует объекту на изображении и будет представлена ориентированным нагруженным графом. Выполним частичную скелетизацию СК, скан-проход которой заключается в следующем. Изображение сканируется по вертикали, анализируется связность смежных серий и выделяются полосы. Найденные полосы закрываются. После вертикального сканирования изображение поворачивается на 90°, снова выполняется скан-проход, затем изображение поворачивается в исходное положение (рис. 2).

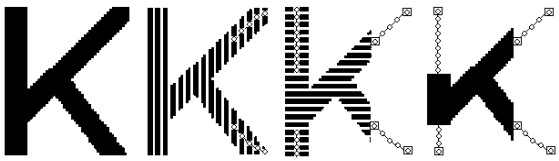


Рис. 2. Пример работы алгоритма частичной скелетизации.

В результате двух скан-проходов прямолинейные отрезки СК заменяются скелетными скривыми (СКР). Группы серий, которые не были заменены СКР на процедуре частичной скелетизации, представляют собой области соединений (например, X-, T-, Y-типа). Из области соединения (ОС) исходят СКР, аппроксимирующие относительно прямолинейные участки. Для каждой СКР, исходящей из ОС, получим вектор направления, построенный по ее начальным точкам. Найдем точку пересечения векторов направлений ОС и соединим ее отрезками с начальными точками СКР. Пометим точки растра, через которые проходят эти отрезки. Затем применим параллельный алгоритм утоньшения для ОС, который не будет удалять помеченные пиксели. Таким способом обеспечивается корректная обработка соединений.

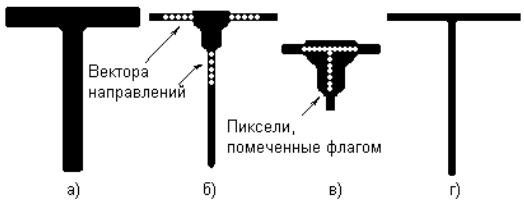


Рис. 3. Обработка соединений: а) исходное изображение; б) результат частичной скелетизации; в) ОС, белым показаны точки соединения; г) результат скелетизации.

На следующем шаге анализом последовательности серий единичной ширины выделим группы серий, удовлетворяющие условиям (5)–(7). Заменим эти серии с помощью СКР. В данном случае СКР будет являться альтернативным способом представления последовательности пикселей единичной толщины. Перейдем к построению графовой модели. По координатам узловых серий (т. е. серий, которые указывают на СКР) сформируем множество вершин V. По точкам СКР построим ребра. Каждая СКР была прикреплена к двум сериям, которые преобразовались в вершины, следовательно, можно найти вершины, из которых исходят ребра.

В результате вышеописанных действий каждая СК изображения представляется нагруженным ориентированным планарным псевдографом, вершинам которого соответствуют концевые и узловые точки отрезков СК, а ребрам — сами отрезки СК, представленные в форме СКР. Независимые подходы к описанию и построению графовых моделей были предложены в [7, 8, 9].

Псевдограф G задается парой $G = (V, E)$, где V — множество вершин, E — мультимножество ребер, каждое из которых соединяет две вершины из V, причем изображения ребер из E на плоскости не пересекаются, поэтому (V, E) представляет собой планарный граф.

Построенная графовая модель обладает важными свойствами. Графовая модель является компактной формой представления СК изображения. Она описывает топологию СК, связи между отрезками СК (ОСК) и позволяет осуществлять эффективное нахождение графических примитивов.

По графовой модели могут быть получены следующие характеристики:

1) отдельной СК: выделение петли на изображении СК; количество ОСК; длина каждого ОСК; общая длина всех ОСК; средняя длина всех ОСК; максимальная и минимальная длина ОСК; средняя элонгация всех ОСК; максимальная и минимальная элонгация ОСК; средняя ширина всех ОСК; максимальная и минимальная ширина ОСК;
2) изображения: количество объектов; количество ОСК всех объектов; суммарное, среднее, максимальное и минимальное значение параметров СК.

Граф может быть преобразован в более компактную форму гиперграфа, гиперребра которого состоят из ребер, соединяющих вершины степени 1 и 2 исходного графа [9] (рис. 4.).

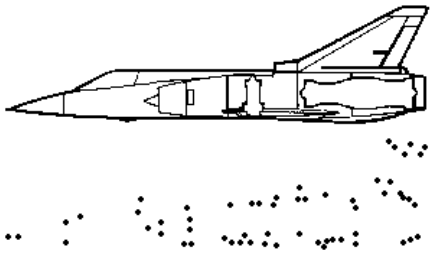


Рис. 4. Исходное изображение и вершины гиперграфа.

Следующим шагом является построение векторной модели на основании имеющегося графа. Из имеющегося псевдографа G получим гиперграф GG. На первом шаге процедуры построения выделим гиперребра графа GG, состоящие из ребер графа G, соединяющих вершины графа G степени один и два. Каждое гиперребро имеет две концевые вершины. Из каждой такой вершины гиперребра GE исходит ноль или более одного ребра E, не принадлежащих гиперребру GE. Рассмотрим два ребра E_1 и E_2, исходящих из вершины степени 3 графа G. Пусть ребра E_1 и E_2 принадлежат гиперребрам GE_1 и GE_2 соответственно. Если реб-

ра E_1 и E_2 являются коллинеарными с некоторой погрешностью (например, 5°), то гиперребра GE_1 и GE_2 объединяются. Введем понятие пути векторизации. Под путем векторизации (ПВ) будем понимать последовательность точек $p_i = (x_i, y_i)$, лежащих на средней линии СК. ПВ $= \{p_0, \ldots, p_n\}$. Если $p_0 = p_n$, то путь является закрытым. Точки ПВ описывают набор отрезков и дуг окружностей (рис. 5).

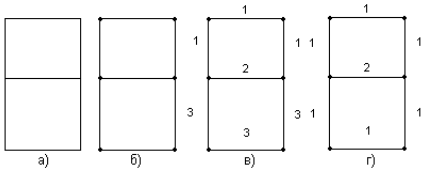


Рис. 5. а) исходная связная компонента; б) полученный граф; в) цифрами показаны гиперребра GE на первом шаге построения г) цифрами показаны ПВ.

Для выделения отрезков применяется метод генерализации Дугласа–Пекера [10]. Пусть имеется путь векторизации ПВ $= \{p_0, \ldots, p_n\}$. Рассмотрим процедуру аппроксимации. Построим отрезок p_0p_n. Пусть p_k — самая удаленная от отрезка p_0p_n вершина. Если расстояние от p_k до отрезка p_0p_n меньше заданного порога, то p_0p_n аппроксимирует последовательность. В противном случае разобьем ПВ на 2 части: $P_1 = \{p_0, \ldots, p_k\}$ и $P2 = \{p_{k+1}, \ldots, p_n\}$. Для каждой части рекурсивно применяется процедура аппроксимации. Введем понятия сегмента. Под сегментом будем понимать прямолинейный отрезок, полученный путем полигональной аппроксимации точек пути векторизации. Пусть имеется путь векторизации, представленный сегментами $S_1, \ldots, S_n$. Рассмотрим пару смежных сегментов S_iS_{i+1}. Из геометрии известно, что три неколлинеарные точки X, Y, Z на плоскости однозначно определяют окружность. Центр окружности C будет находиться на пересечении перпендикуляров, опущенных на середины отрезков XY и YZ. Затем итеративно тестируются смежные сегменты на принадлежность дуге.

Полученные из сегментов отрезки и дуги окружностей получают характеристику ширины ребра графа G, анализом которого они были найдены. В качестве постобработки предлагается стыковка отрезков в местах соединений. Это осуществляется анализом МУТ и МКТ.

Заключение

Эффективный метод представления растрового изображения должен обеспечивать достаточную степень сжатия информации и в то же время позволять обрабатывать изображение напрямую в кодированной форме. Например, векторизация часто осуществляется для изображений большого размера, и разбиение изображения на фрагменты с последующей «сшивкой» является источником ошибок.

Мы предлагаем использование графовой модели как компактного средства хранения и описания структуры растра. Достоинством предлагаемого подхода является простота реализации и достаточно высокое быстродействие. Нами предлагается модификация алгоритма нахождения СК, созданная специально для обработки изображений большого размера. Благодаря кодированию концов серий построение скелета осуществляется быстрее, чем при использовании классических методов попиксельного анализа. Также решается проблема обработки соединений. Недостатком подхода является неполное описание площадных объектов. Алгоритм может быть успешно применен при векторизации планов зданий, технических чертежей.

Литература

[1] *Абламейко С. В., Лагуновский Д. М.* Обработка изображений: технология, методы, применения. Учебное пособие. — Мн: Амалфея, 2000. — 304 с.

[2] *Listing, J.* 1861, Der census raeumlicher complexe oder verallgemeinerungen des euler'schen satzes von den polyeadern., Abhandlungen der Mathematischen Classe der Koeniglichen Gesellschaft der Wissenschaften zu Goettingen

[3] *Blum, H.* 1962, An associative machine for dealing with the visual field and some of its biological implications., in: Biological Prototypes and Synthetic Systems, volume 1, ed.: Bernard,E.E. and Kate,M.R. pp. 244–260.

[4] *Klette G.*, Topologic, Geometric, or Graph-Theoretic Properties of. Skeletal Curves, PhD dissertation, Groningen Univ., 2007 — 122 p.

[5] *Shapiro L. G.* Connected component labeling and adjacency graph construction // Topological algorithms for digital image processing — 1996. — Pp. 1–31.

[6] *Jung-Me P., Looney C. G.., Hui-Chuan C.* Fast Connected Component Labeling Algorithm Using a Divide and Conqure Technique. //CATA 2000 Conference on Computers and Their Applications — 2000, Dec. — Pp. 373–376.

[7] *S. Di Zenzo, L. Cinque, S. Levialdi* Run-based algorithms for binary image analysis and processing. //IEEE Trans. on PAMI. — 1996. — Vol. 18, Issue 1. – Pp. 83–89.

[8] *Костюк Ю. Л., Новиков Ю. Л.* Графовые модели цветных растровых изображений высокого разрешения // Вестник ТГУ — 2002. — № 275, апрель. — С. 153–160.

[9] *Burge M., Kropatsch W. G.* A minimal line property preserving representation of line images //Computing, 1999. — Vol. 62, Issue 4. — Pp. 355–368.

[10] *Douglas D. H., Peucker T. K.* Algorithms for the reduction of the number of points required to represent a digitized line or its caricature // The Canadian Cartographer — 1973. — № 2. — Pp. 112-–122.

Непрерывная классификация дактокарт по особенностям опорных точек изображений отпечатков пальцев*

Ушмаев О. С.
oushmaev@ipiran.ru
Москва, Институт проблем информатики РАН

Значимой проблемой автоматической дактилоскопической идентификации является увеличение производительности за счет различных технологий классификации и снижения объема поиска в больших базах данных. В докладе представлен алгоритм непрерывной классификации дактокарт отпечатков пальцев. Изображению отпечатка пальца ставится в соответствие классификационный вектор небольшой размерности. При идентификации по результатам сравнения классификационных векторов хранимых и предъявленного отпечатка отбирается меньшее подмножество всей хранимой базы. Классификационный вектор строится на основе текстуры вокруг опорных точек отпечатка (центров и дельт). Эксперименты показывают, что за счет использования представленной технологии непрерывной классификации можно значительно увеличить скорость дактилоскопической идентификации.

Введение

Современные автоматизированные дактилоскопические идентификационные системы (АДИС) осуществляют поиск в огромных базах данных, которые могут насчитывать несколько десятков миллионов отпечатков пальцев. Традиционным способом увеличить производительность является разделение всей базы данных на несколько предопределенных классов.

Наиболее распространенным вариантом классификации отпечатков пальцев является деление, предложенное Гальтоном [1] и позднее модифицированное Генри [2]. Все отпечатки разделяются на пять основных классов: дуга, завиток, левая и правая петли и шатровая дуга. В настоящее время существуют более детальные варианты классификации [3]. Значимой проблемой такого деления является сильная внутриклассовая вариация и размытость границ классов [4], что приводит к ошибкам классификации.

В настоящее время существует множество реализаций автоматической классификации отпечатков пальцев по системе Генри. Большинство из них основаны на анализе поля направлений папиллярного узора (рис. 1). В работе [5] представлен способ определения узора на основе относительного расположения опорных точек — центров и дельт. В [6] приведены алгоритмы классификации на основе прослеживания траекторий папиллярных линий. В [7] дано представление поля направлений в виде скрытой марковской модели. В [9] рассматриваются методы кластеризации карт откликов фильтра Габора [8]. Наиболее ранними методами классификации являются синтаксические методы, изложенные в [10, 11]. Преимуществом таких методов ялвяется низкая вычислительная сложность. Наиболее современными и эффективными

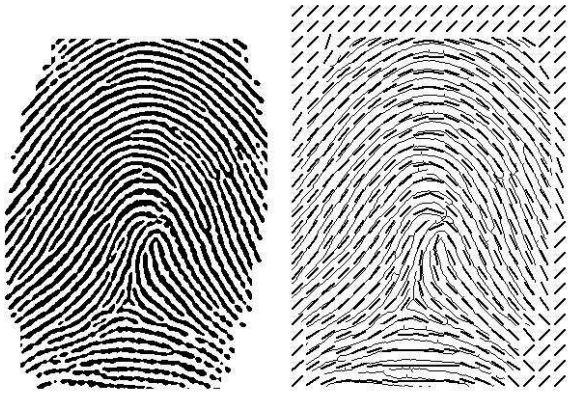


Рис. 1. Поле направлений папиллярных линий отпечатка пальцев.

являются структурные методы [12, 13], которые исследуют «глобальную» топологию поля направлений. Определенный интерес представляет применение нейронных сетей [14]. Однако, несмотря на большое разнообразие методов, ошибки классификации остаются значительными.

Как было отмечено выше, основным назначением классификации является ускорения поиска в большой базе за счет сокращения числа трудоемких операций сравнения отпечатков пальцев. А именно, предъявляемый отпечаток сравнивается только с отпечатками из одного с ним класса. Однако классификация по типам узора имеем один явный недостаток: границы между типами отпечатков очень размыты, поэтому вероятность ошибочной классификации останется достаточно высокой.

Развитием идеи классификации являются многопроходные алгоритмы. На первом шаге быстрый алгоритм просевает базу. На последующих шагах более точные алгоритмы позволяют идентифицировать отпечатки.

Наиболее быстрым алгоритмом сравнения является вычисления евклидового расстояния между короткими векторами признаков. Методы, позво-

*Работа выполнена в рамках НОЦ ИПИ РАН – ВМиК МГУ «Биометрическая информатика» при финансовой поддержке РФФИ, проект № 07-07-00031.

ляющие поставить в соответствие отпечатку пальцев вектор, называются непрерывной классификацией отпечатков пальцев [15]. Чем ближе векторы, тем вероятней, что отпечатки принадлежат одному классу. С точки зрения практического применения отличием непрерывной классификации от классификации по типам узоров является возможность смягчения критерия просева базы отпечатков пальцев за счет выбора порога принятия решения.

Для эффективного использования непрерывной классификации следует стремится к минимально возможной ошибке второго рода (FAR) при нулевой ошибке первого рода (FRR). В крупномасштабных АДИС основная часть времени тратится на сравнение отпечатков, принадлежащих разным людям. Например, при использовании непрерывной классификации с FAR = 0,1 на первом шаге будет в среднем отсеиваться 90% базы, что потенциально соответствует почти десятикратному ускорению.

Задача разработки технологий непрерывной классификации является нетривиальной, поскольку базовые технологии описания и сравнения отпечатков пальцев предполагают работу с двумерными графами (конфигурациями контрольных точек). Для успешного решения задачи непрерывной классификации требуется поиск дополнительной информации в изображениях отпечатков пальцев. В докладе представлен алгоритм непрерывной классификации на основе анализа окрестностей опорных точек отпечатков пальцев, приводятся результаты экспериментов для непрерывной классификации дактокарт.

Алгоритм вычисления классификационного вектора

Если исключить из анализа отпечатка пальца контрольные точки (ветвления и окончания папиллярных линий), то наиболее информативной областью изображения являются окрестности опорных точек — глобальных особенностей папиллярного узора типа дельта и центр. Как показано в [16], опорные точки имеют точное положение и каноническую ориентацию, определенную структурой поля направлений. На рис. 2 приведен пример расположения опорных точек.

Алгоритм вычисления классификационного вектора состоит из следующих шагов:

1. Определение расположения опорных точек отпечатка пальцев.

2. Вычисление канонической ориентации.

3. Извлечение вектора признаков из окрестности опорной точки.

4. Определение весов признаков.

Различные реализации шагов 1 и 2 приведены в множестве статей по автоматической дактилоскопической идентификации. В целях настоящего доклада используется метод [16].

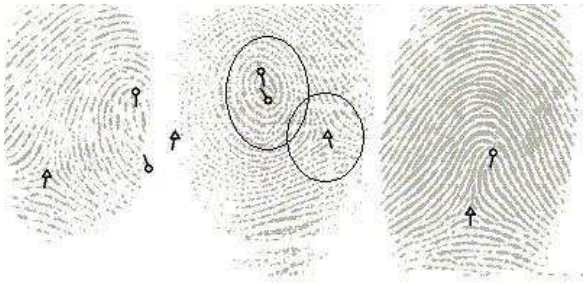


Рис. 2. Положение опорных точек. Выделена информативная зона отпечатка.

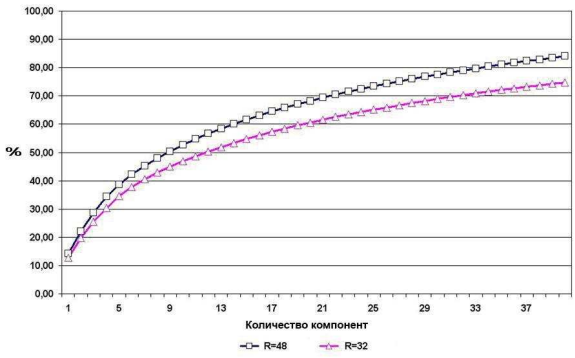


Рис. 3. Накопленная доля дисперсии.

Для извлечения вектора признака проведем анализ изображений вокруг опорных точек методом главных компонент в частотной области. На рис. 3 показана накопленная доля дисперсии в зависимости от числа главных компонент для двух радиусов исследуемой окрестности (32 и 48 точек при разрешении сканера 500ppi). Как видно из рисунка, первые 10 компонент объясняют около 50% общей дисперсии [18], что указывает на высокую информативность полученных компонент.

Для дальнейшего использование классификационный вектор следует нормировать. Как было отмечено выше, целью непрерывной классификации является минимальной ошибка FAR при нулевой ошибке FRR. Поэтому целесообразно выбирать нормировочные коэффициенты методами оптимальной интеграции биометрических технологий [17]. Основной идеей методов [17] является нормировка признаков таким образом, чтобы значения признаков s были равны соотношению $\ln f_{\text{gen}}(s)/\ln f_{\text{imp}}(s)$, где f_{gen} и f_{imp} — плотности распределения признака в «своих» и «чужих» сравнениях (т. е. когда сравниваемые образцы принадлежат одному и разным людям соответственно). С такой нормировкой оптимальным по ошибкам FAR и FRR классификационным правилом является сумма признаков.

На рис. 4 приведены ошибки распознавания FRR и FAR по отдельным компонентам. Ошибки распознавания ожидаемо растут с ростом номера

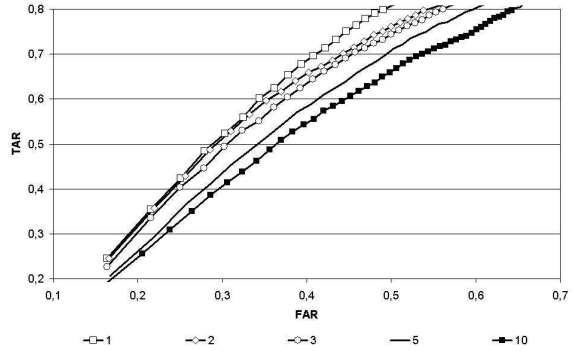


Рис. 4. Ошибки распознавания по отдельными компонентами, TAR = 1 − FRR.

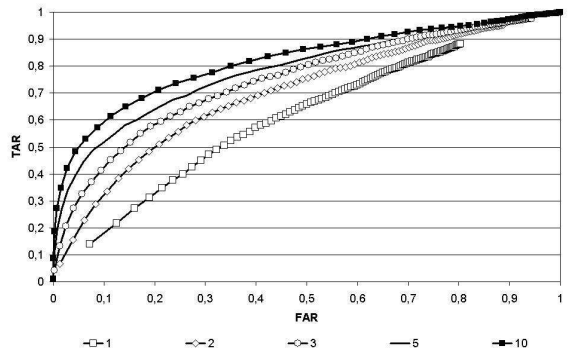


Рис. 5. Ошибки распознавания по совокупности компонент, TAR = 1 − FRR.

компоненты. Ошибки распознавания одновременно по нескольким компонентам приведены на рис. 5. Для оценки параметров метода главных компонент и ошибок распознавания использовались базы изображений отпечатков пальцев FVC2002 [19] (300 человек, по 8 отпечатков пальцев для каждого человека, данные сведены в три массива — два собраны оптическими сканерами, один — ёмкостным).

Эксперименты

Реализация непрерывной классификации в АДИС состоит из двух функций: вычисления классификационного вектора для каждого отпечатка и сравнения классификационных векторов для всей дактокарты. Схематично изложенная в разделе 2 процедура вычисления классификационного вектора для одного отпечатка пальца представлена на рис. 6 (максимальное количество опорных точек каждого типа составляет три).

При использовании нескольких отпечатков ошибки распознавания будут уменьшаться. На рис. 7 представлены ошибки классификации дактокарт в зависимости от числа используемых отпечатков пальцев. Эксперимент проводился следующим образом. Для дактокарт с двумя отпечатками брались откатанные указательные пальцы. Далее последовательно парами добавлялись большие откатанные пальцы, средние, безымянные и мизинцы.

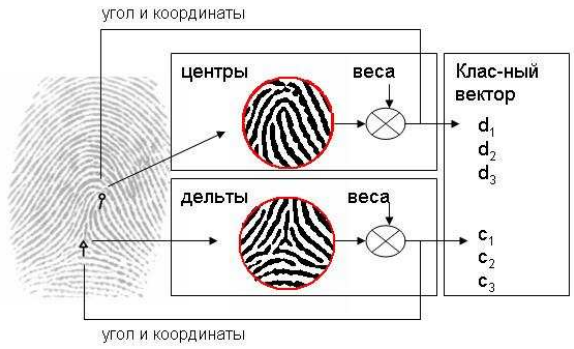


Рис. 6. Алгоритм вычисления классификационного вектора.

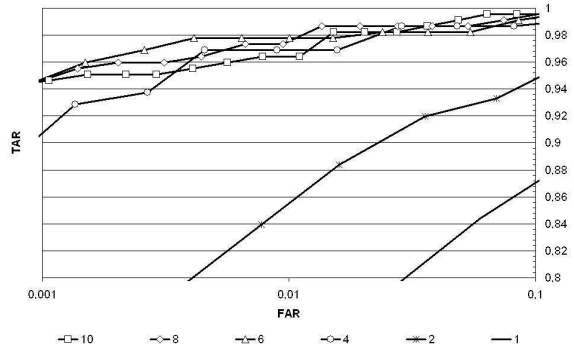


Рис. 7. Ошибки классификации дактокарт в зависимости от размера дактокарты.

На рис. 8 представлен коэффициент ускорения идентификации дактокарт за счет использования представленного метода в зависимости от дополнительной ошибки первого рода FRR. Эксперименты проводились на стандартном тесте NIST SD29 (парные дактокарты для 219 человек) [20]. Сопоставлялось время сравнения дактокарт штатным алгоритмов и рассчитываемое как $t_1 = nt_{\text{basic}}$, и время сравнения с использованием непрерывной классификации $t_2 = nt_{\text{clas}} + np(\theta)t_{\text{basic}}$, где n — число используемых отпечатков, p — доля дактокарт, прошедших отбор по результатам непрерывной классификации, в зависимости от порога θ. Время сравнения штатного алгоритма и непрерывной классификации рассчитывалось из производительности 2000 и 200 000 сравнений в сек на процессоре тактовой частотой 1 ГГц.

Заключение

В докладе представлен метод непрерывной классификации дактокарт на основе анализа окрестностей опорных точек изображений отпечатков пальцев. Реализация разработанного метода позволяет значительно ускорить выполнение опе-

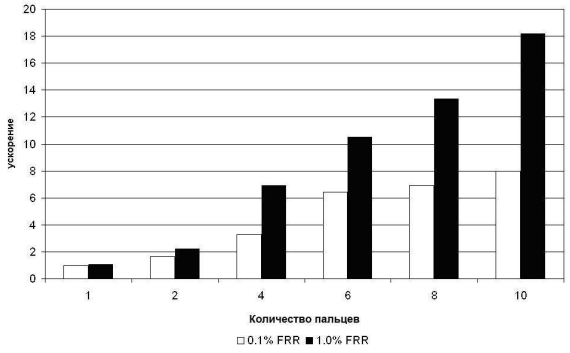


Рис. 8. Ускорение дактилоскопической идентификации в зависимости от размера дактокарты и дополнительной ошибки классификации.

рации идентификации дактокарт без потерь в качестве распознавания.

По сравнению с методами классификации по типам узора предложенный подход имеет следующие отличительные особенности:

1. Недостатком является большая трудоёмкость вычислений.

2. Преимуществом является наличие порога, который позволяет регулировать соотношение «ускорение идентификации — ошибка первого рода».

3. Для принятия решения используется локальная информация вокруг опорных точек, что делает подход более устойчивым к потере части информации об отпечатке пальцев.

Направлением дальнейших исследований является оптимизация выбора классификационных признаков и изучения влияния характеристик исходных изображений.

Литература

[1] *Galton F.* Finger Prints. — New York: Williams Hein & C, 2002. — 216 p.

[2] The Henry classification system. — 2003. — (`www.biometricgroup.com/HenryFingerprintClassification.pdf`).

[3] *Локар Э.* Руководство по криминалистике. — Москва: Юридическое издательство НКЮ СССР, 1941. — 544 с.

[4] *Самищенко С. С.* Атлас необычных папиллярных узоров. — Москва: Юриспруденция, 2001. — 320 с.

[5] *Hong L., Jain A. K.* Classification of fingerprint images // Proceedings of the 15th Scandinavian Conference on Image Analysis, Kangerlussuaq, Greenland. — 1999. — Pp. 435–439.

[6] *Chong M. M. S. et al.* Geometric framework for fingerprint image classification // Pattern Recognition. — 1997. — №. 30. — Pp. 1475–1488.

[7] *Senior A.* A hidden Markov Model Fingerprint Classifier // Proceedings of the 31st Asilomar Conference on Signals, Systems and Computers, Asilomar. — 1997. — Pp. 306–310.

[8] *Daugman J. G.* Complete Discrete 2D Gabor Transforms by Neural Networks for Image Analysis and Compression // IEEE Trans. Acoustics, Speech, and Signal Processing. — 1988. — №. 36. — Pp. 1169–1179.

[9] *Jain A. K., Prabhakar S., Hong L.* A Multichannel Approach to Fingerprint Classification // IEEE Trans. on Pattern Analysis Machine Intelligence. — 1999. — №. 21. — Pp. 348–359.

[10] *Moayer D., Fu K. S.* A Syntactic Approach to Fingerprint Pattern Recognition // Pattern Recognition. — 1975. — №. 7. — Pp. 1–23.

[11] *Rao K., Black K.* Type Classification of Fingerprints: A Syntactic Approach // IEEE Trans. on Pattern Analysis Machine Intelligence. — 1980. — №. 2. — Pp. 223–231.

[12] *Miao D., Maltoni D.* A Structural Approach to Fingerprint Classification // Proceedings of the 23rd International Conference on Pattern Recognition, Vienna. — 1996. — Pp. 120–124.

[13] *Capelli R., Lumini A., D. Miao, D. Maltoni* Fingerprint Classification by Directional Image Partitioning // IEEE Trans. on Pattern Analysis Machine Intelligence, 1999. — №. 21. — Pp. 402–421.

[14] *Kamijo M.* Classifying Fingerprint Images Using Neural Network: Deriving the Classification State // Proceedings of the 3rd International Conference on Neural Network, 1993. — Pp. 1932–1937.

[15] *Lumini A., Maio D., Maltoni D.* Continuous versus exclusive classification for fingerprint retrieval // Pattern Recognition Letters, 1997. — Vol. 18, №.10. — Pp. 1027–1034.

[16] *Bazen A. M., Gerez S. H.* Extraction of Singular Points from Directional Fields of Fingerprints // Mobile Communications in Perspective, University of Twente, Enschede, 2001. — Pp. 41–44.

[17] *Синицын И. Н., Новиков С. О., Ушмаев О. С.* Развитие технологий интеграции биометрической информации // Системы и средства информатики, вып. 14, 2004. — С. 5–36.

[18] *Харман Г.* Современный факторый анализ. — Москва: Статистика, 1973. — 487 с.

[19] The Second Fingerprint Verification Competition // (`bias.csr.unibo.it/fvc2002/`).

[20] NIST SD29, NIST Special Database 29 “Plain and Rolled Images from Paired Fingerprint Cards”.

Трейс-преобразование как источник признаков распознавания*

Федотов Н. Г.
fedotov@pnzgu.ru
Пенза, Пензенский государственный университет

Рассматривается новое геометрическое преобразование (трейс-преобразование), связанное со сканированием изображений по сложным траекториям. Преобразование введено в [1] и исследуется в настоящем докладе. Трейс-преобразование является источником формирования нового класса конструктивных признаков распознавания, характерной особенностью которых является структура в виде композиции трёх функционалов. В библиографии к этой работе и других докладах, представленных на данной конференции, рассмотрены применения предложенного аппарата для решения прикладных задач распознавания образов.

Рассмотрим входную сетчатку распознающей системы, под которой будем понимать сканируемую часть плоскости изображения. В этой части плоскости располагается некоторое изображение, тогда как оставшаяся часть плоскости — фоновая. Таким образом, изображение финитно. Рассмотрим случайную прямую l, которая может пересекать изображение. Предположим, что пересечение прямой l и изображения позволяет нам вычислить некоторое число g, характеризующее их взаимное расположение. Производя серию случайных бросаний прямой l на плоскость, получаем выборку для случайной величины g. Далее, можно определить какую-нибудь эмпирическую характеристику n случайной величины g. В работе [1] рассматривалась реализация описанной процедуры в технических системах, осуществляющих распознавание изображений.

Математическая сторона указанной процедуры интенсивно исследовалась в стохастической геометрии. Было выяснено, что при некоторых условиях характеристика n может иметь явный геометрический смысл. Для нас важно, что, легко реализуясь в технических системах, эта идея может служить исходной точкой для получения новых признаков распознавания образов, как в теоретическом анализе, так и в практической сфере.

В работе [1] приводятся формулы, на основе которых строятся признаки распознавания. Рассматриваются только бинарные изображения (черные фигуры на белом фоне).

1. Рассмотрим изображение в виде кусочно-дифференцируемой кривой, которая может быть границей фигуры. Пусть g — число пересечений этой кривой со случайной прямой l. Тогда математическое ожидание $\mathsf{E}g$ пропорционально длине кривой [2].
2. Рассмотрим изображение в виде выпуклой фигуры. Это может быть выпуклая оболочка некоторой другой фигуры. Пусть g — длина пересечения выпуклой фигуры со случайной прямой l.

Тогда средние величины $\mathsf{E}g^0$, $\mathsf{E}g$, $\mathsf{E}g^2$ пропорциональны соответственно периметру, площади и собственному потенциалу однородного слоя [1].

Приведенные в [1] формулы и их многочисленные аналоги имеют для распознавания образов следующие недостатки:

1) число этих формул ограничено, поскольку ясно выраженных геометрических характеристик не так много, а признаков требуются тысячи и более;
2) формулы применимы только для бинарных изображений.

К достоинствам следует отнести возможности параллельных вычислений (одновременно обрабатывается несколько прямых сразу) и стохастической реализации, последнее позволяет оборвать процесс при достижении нужной точности, кроме того, вычисленные признаки не зависят от движений объектов. Известно, что обычно признаки сильно зависят от поворота и сдвига объекта, в то время как во многих задачах распознавания поворот и сдвиг объектов совершенно неинформативны.

В статье предлагается обобщение приведенного выше подхода с целью преодоления его недостатков и с сохранением достоинств.

Геометрическое трейс-преобразование

Обозначим буквой F финитное изображение. Если дана прямая l, то число g, характеризующее взаимное расположение прямой l и изображения, будем вычислять согласно некоторому правилу T: $g = T(l, F)$; отображение T есть функционал. Для нас желаемым свойством является независимость вычислений от движения объекта, поэтому единственное требование, которое мы накладываем на T, формулируется следующим образом. Пусть изображение претерпело сдвиг и поворот, при этом возникло новое изображение F'. При этом же сдвиге и повороте прямая l перейдет в прямую l', оставаясь, таким образом, «вмороженной» в изображение. Требуется, чтобы $T(l, F) = T(l', F')$. Это равенство должно быть верным для всех прямых и всех допустимых изображений. Такое свойство

*Работа выполнена при финансовой поддержке РФФИ, проект № 09-07-00089.

назовем полной инвариантностью функционала T. Следует отметить, что понятие полной инвариантности весьма сильно расширяет возможности распознавания образов, ибо это не обязательно число пересечений, длина секущей и т. д. Например, если изображение цветное, переменной яркости, то таких функционалов можно найти довольно много. Итак, круг функционалов и обрабатываемых изображений значительно расширен.

Аналогично, как и в стохастической геометрии, определена случайная величина $g = T(l, F)$, распределение которой не зависит от сдвигов и поворотов изображения. Поэтому числовые характеристики этой случайной величины опять могут служить признаками изображений, которые определяются специальными техническими системами. Недостаток нового семейства признаков — первоначальное отсутствие ясного геометрического смысла, и заранее не известна их различающая сила. Однако для распознавания образов это не так важно, ибо решающей все-таки является экспериментальная проверка.

Отметим еще одно свойство вполне инвариантного функционала T (трейс): он не обязательно определяется лишь сечением прямой изображения. Для его вычисления может быть привлечена также и другая информация, например, свойства окрестности этого сечения.

Чтобы понять, что предложенное обобщение в некотором смысле исчерпывает все его возможности, изложим теорию трейс-преобразований. Прямая l, если введены полярные координаты на плоскости, характеризуется расстоянием ρ от начала координат до нее и углом θ (с точностью до 2π) ее направляющего вектора:

$$l = \{(x, y) : x\cos\theta + y\sin\theta = \rho\}, \quad l = l(\theta, \rho),$$

где x, y — декартовы координаты на плоскости.

Таким образом, множество всех направленных прямых, пересекающих круг радиусом R с центром в начале координат («сетчатку»), однозначно параметризуется множеством

$$\Lambda = \{(\theta, \rho) : 0 \leqslant \theta \leqslant \pi,\ -R \leqslant \rho \leqslant R\}$$

при условии, что параметры $(0, \rho)$ и $(\pi, -\rho)$ задают одну прямую. Видно, что множество прямых на сетчатке есть в топологическом смысле не что иное, как лист Мёбиуса. Множество чисел $T(l(\theta, \rho), F)$, зависящее от точки на листе Мёбиуса Λ, есть некоторое преобразование изображения, которое назовем трейс-преобразованием. Если, например, при численном анализе трейс-преобразование представлено матрицей, то будем называть ее трейс-матрицей. Если направить ось 0θ горизонтально, а ось 0ρ — вертикально, то в точке (θ_j, ρ_i) будет расположен элемент матрицы с номером (i, j), то есть значение $T(l(\theta_j, \rho_i), F)$.

Здесь θ_j, ρ_i — некоторые значения равномерных дискретных сеток на указанных осях. Матрица будет 2π-периодична в направлении горизонтальной оси, причем через каждый интервал длины π столбцы ее переворачиваются.

Будем считать дополнительно, что если прямая l не пересекает изображения, то $T(l, F)$ есть заданное число (например,0), или другой фиксированный элемент, если функционал T нечисловой. В этом случае первоначальному изображению F соответствует $T(F)$ — новое изображение (можно трактовать $T(l(\theta, \rho), F)$ как изображение, характеристики которого в точке (θ, ρ)) — его трейс-образ.

К полученному промежуточному изображению (трейс-образу), можно вновь применить трейс-преобразование.

Рассмотрим подробнее вычисление трейс-преобразования.

Пусть F будет некоторой векторной функцией представляющей изображение. Она содержит всю информацию об изображении, яркость, цвет и другие характеристики в каждой точке, поэтому мы можем обозначить её той же буквой, что и изображение F.

Рассмотрим функцию трёх независимых переменных

$$l(\theta, \rho, t) = (\rho\cos\theta - t\sin\theta,\ \rho\sin\theta + t\cos\theta).$$

Это естественное параметрическое представление сканирующей прямой. Параметр t связан с естественной одномерной системой координат на прямой.

Пересечение изображения F прямой l даёт функцию

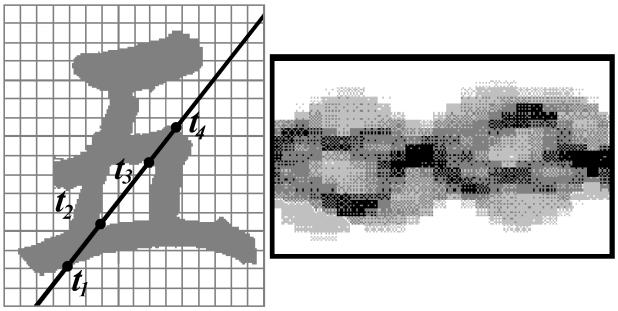
$$f(\theta, \rho, t) = F(l(\theta, \rho, t)).$$

Рассмотрим бинарное изображение китайского иероглифа, состоящее из квадратных пикселов, пересекаемых сканирующей прямой линией l на рисунке 1(а). Данный рисунок демонстрирует получение бинарной функции пересечения $f(\theta, \rho, t)$ действительной переменной t для прямой l. Эта функция $f(\theta, \rho, \bullet)$ равна единице в интервалах пересечения с изображением, так на рисунке 1(а) — это интервалы (t_1, t_2) и (t_3, t_4). В других точках она равна 0:

$$f(\theta, \rho, t) = \begin{cases} 1, & t \in F \cap l \\ 0, & t \notin F \cap l \end{cases}.$$

Следующая идея — вычисление числового значения при использовании этой функции. Пусть T — функционал, применённый к функции $f(\theta, \rho, t)$, где в качестве независимой переменной определим переменную t, таким образом, получим:

$$g(\theta, \rho) = T(l, F) = Tf(\theta, \rho, t) = Tf_{\theta,\rho}. \qquad (1)$$



(а) (б)

Рис. 1. Изображение китайского иероглифа, пересеченное сканирующей прямой линией l (а); Трейс-трансформанта изображения иероглифа (б)

Функционал T назван трейс-функционалом, сам процесс получения функции g под действием трейс-функционала T назван трейс-преобразованием, а функция g — трейс-трансформантой.

Например, пусть $Tf(\theta, \rho, \bullet)$ будет длиной максимального интервала в области определения функции пересечения $f(\theta, \rho, \bullet)$. На рисунке 1 это является значением $t_2 - t_1$.

Можно считать, что каждая сканирующая прямая линия $l \in \Lambda$ должна быть представлена вектором $(-\sin\theta, \cos\theta)^\intercal$ (индекс $^\intercal$ преобразует строку в столбец), поэтому для каждой прямой l существует уникальная пара $(\theta, \rho) \in S^1 \times \mathbb{R}$. Следовательно, множество Λ всех сканирующих прямых на плоскости в топологическом смысле есть цилиндр. В этом случае интерпретируем результат трейс-преобразования как изображение на цилиндре (склеивается левая и правая сторона рисунка с изображением трейс-трансформанты).

В нашем примере функционал T независим от любого сдвига изображения основной функции и при вычислении выражения $Tu(t)$. Также он независим от изменения знака параметра t, то есть $Tu(t) = Tu(-t)$. Это ведет к тому, что мы можем интерпретировать результат трейс-преобразования так, как будто он расположен на листе Мёбиуса. Рисунок с изображением трейс-трансформанты разрежем вдоль вертикальной оси симметрии, правую часть рисунка перевернём вдоль горизонтальной оси симметрии, и склеим левый и правый край рисунка. В нашем представлении этих результатов трейс-преобразований числа для наглядности интерпретируются цветом.

Если выбрать в качестве T функционала суммарную длину пересечения (отрезок t_1–t_2 плюс отрезок t_3–t_4), то в этом частном случае трейс-преобразование совпадет с преобразованием Радона для бинарных изображений. Действительно, пусть пересечение изображения F сканирующей линией l даёт функцию пересечения $f_{\theta,\rho}$. Если интегрировать эту функцию вдоль каждой линии по параметру t, то совокупность интегральных значений яркости для всех линий даёт преобразование Радона. В терминах трейс-преобразования имеем

$$Tf_{\theta,\rho} = \int_{-\infty}^{\infty} f_{\theta,\rho}(t)\, dt.$$

Совокупность $\{Tf_{\theta,\rho}\}$, $\theta \in [0, 2\pi]$, $\rho \in R$ несёт всю информацию об изображении.

Для бинарных изображений, рассматриваемых в вышеприведённом примере, определение суммарной длины пересечений изображений с каждой из сканирующих линий даёт трейс-преобразование эквивалентное преобразованию Радона.

Примеры применения преобразования Радона в качестве трейс-преобразования можно найти в работах [7, 8].

Следует отметить, что при определённом выборе T функционалов трейс-преобразование становится эквивалентным преобразованиям Фурье, Хо, Радона–Хо, но не совпадает с ними.

Трейс-преобразование является эффективным инструментом при изучении движений распознаваемых объектов и их масштабных изменений. Это объясняется тем, что трейс-образ сохраняет информацию о первоначальном объекте, то есть тип трейс-матрицы не изменяется под действием группы движений (поворота, переноса) и гомотетии, но каждое их этих преобразований вносит свою характерную компоненту при формировании трейс-преобразования. Кратко остановимся на том, как меняется изображение $T(l, F)$ при сдвигах и вращениях исходного изображения F. Если первоначальное изображение поворачивается, то его трейс-образ сдвигается по горизонтальной оси.

Если же происходит сдвиг исходного на некоторый вектор, то его трейс-образ претерпевает следующие изменения. Лучше их изложить в терминах трейс-матриц. Столбцы остаются неизменными, на своих местах, но могут сдвигаться вверх или вниз. Вектор сдвига определяют числа a и b такие, что столбец с координатой θ_i сдвигается в вертикальном направлении на $a{\cdot}\cos(\theta_i - b)$. Следует подчеркнуть, что вполне строгим это описание будет лишь в том случае, если трейс-матрицу считать непрерывной, то есть i и j непрерывные параметры.

Обычная евклидова мера $d\theta d\rho$ листа Мёбиуса инвариантна к указанным преобразованиям, поэтому плотность распределения всякой функции, заданной на листе Мёбиуса, в данном случае функции изображения $T(l, F)$ не зависит от указанных преобразований, то есть если изображения F сдвинуто и повернуто до состояния F', то распределение значений функций изображения $T(l, F)$ и $T(l, F')$ одинаковы. Именно поэтому их значения могут трактоваться как случайные функции, не зависящие от движений исходного изображения.

Триплетные признаки распознавания

Рассмотрим формирование триплетных признаков, представляющих последовательную композицию трех функционалов:

$$\Pi(F) = \Theta \circ P \circ T\left(F \circ l\left(\theta, \rho, t\right)\right).$$

Каждый функционал (Θ, P и T) действует на функции одной переменной (θ, ρ и t) соответственно.

Функционал T, соответствующий трейс-преобразованию, подробно рассмотрен выше. В дискретном варианте вычислений результат этого преобразования, или трейс-трансформанта $T\left(F \circ l\left(\theta, \rho, t\right)\right)$, представляет собой матрицу, элементами которой являются, например, значения яркости изображения F на пересечениях со сканирующей линией $l\left(\theta, \rho\right)$. Параметры сканирующей линии θ и ρ определяют позицию этого элемента в матрице. Последующее вычисление признака заключается в последовательной обработке столбцов матрицы с помощью функционала P, а затем в преобразовании полученной периодической функции с помощью функционала Θ в число-признак $\Pi(F)$. Подробное описание триплетных признаков можно найти в работах [3] и [4].

Заключение

Трёхзвенная форма триплетного признака позволяет получить большое число новых конструктивных признаков распознавания, причём в режиме автоматической компьютерной генерации. Обилие признаков даёт возможность расширить круг решаемых задач распознавания, включить в него задачи с большим алфавитом образов: распознавание иероглифов [5], объектов из области нанотехнологий [6], биологических микрообъектов [6], распознавание в задачах технической [9] и медицинской дефектоскопии [8]. Разработанная теория способствует решению смежных задач — поиск фрагмента на изображении, нахождение похожих фрактальных структур, поиск изображений по содержанию, в частности биометрический поиск [7]. На основе трейс-преобразования и триплетных признаков возможна эффективная предварительная обработка изображений, см. [10] и [11, в настоящем сборнике], и анализ текстур [12, в настоящем сборнике].

Триплетные признаки носят универсальный характер и пригодны для распознавания бинарных, тональных и цветных изображений. Благодаря трёхзвенной структуре, возможно получение большого числа (тысяч) триплетных признаков в режиме автоматической компьютерной генерации. Опора на большое число признаков, как показала практика, ведёт к повышению гибкости и интеллектуальности распознающих систем, и увеличению надёжности распознавания.

Литература

[1] *Федотов Н. Г.* Методы стохастической геометрии в распознавании образов. Москва:Радио и связь, 1990. — 144 с.

[2] *Сантало Л.* Интегральная геометрия и геометрические вероятности. Москва:Наука, 1983. — 358 с.

[3] *Fedotov N. G.* The Theory of Image-Recognition Features Based on Stochastic Geometry. // Pattern Recognition and Image Analysis. Advances in Mathematical Theory and Applications.— 1998. — Vol. 8, № 2. — Pp. 264–266.

[4] *Fedotov N. G., Shulga L. A., Moiseev A. V., Kolchugin A. S.* Pattern Recognition Feature and Image Processing Theory on the Basis of Stochastic Geometry // Proc. of the 2nd Int. Conf. on Informatics in Control, Automation and Robotics.— 2005. — Vol. 3. — Pp. 187–192.

[5] *Fedotov N. G., Shulga L. A.* New Theory of Pattern Recognition Feature on the Basis of Stochastic Geometry. // WSCG.— 2000. — Pp. 373–380.

[6] *Федотов Н. Г., Рой А. В.* Анализ биологических микрообъектов с помощью методов стохастической геометрии. // Измерительная техника.— 2004. — № 4. — С. 61–64.

[7] *Федотов Н. Г., Шульга Л. А., Рой А. В.* Интеллектуальная система поиска информации, представленной в виде изображений. // Искусственный интеллект.— 2004. — № 2. — С. 188–192.

[8] *Федотов Н. Г., Кольчугин А. С., Смольякин О. А., Моисеев А. В., Романов С. В.* Формирование признаков распознавания сложноструктурированных изображений на основе стохастической геометрии // Измерительная техника.— 2008. — № 2. — С. 37–45.

[9] *Федотов Н. Г., Никифорова Т. В.* Дефектоскопия сварных соединений на основе методов стохастической геометрии // Машиностроение. Контроль. Диагностика.— 2002. — № 12. — С. 65–68.

[10] *Федотов Н. Г., Шульга Л. А., Моисеев А. В.* Теория признаков распознавания и предварительной обработки изображений на основе стохастической геометрии // Измерительная техника.— 2005. — № 8. — С. 8–13.

[11] *Федотов Н. Г., Романов С. В., Мокшанина Д. А.* Сегментация гистологических изображений. Выделение фолликулов и ядер // Всеросс. конф. ММРО-14. — С. 611–613.

[12] *Федотов Н. Г., Мокшанина Д. А., Романов С. В.* Анализ текстур гистологических изображений // Всеросс. конф. ММРО-14. — С. 608–610.

Концепция группового распознавания образов*

Фурман Я. А.
krtmbs@marstu.net
Йошкар-Ола, Марийский государственный технический университет

Предложена концепция построения системы распознавания сигналов, в соответствии с которой аддитивная смесь сигналов разных классов одновременно подается на ее вход, а реакцией системы является одновременное распознавание сигналов и оценка их весов. Концепция реализуема при образовании базиса эталонными векторами классов алфавита. Исследована помехоустойчивость системы и получены условия равенства ее эффективности с эффективностью системы, распознающей последовательно поступающие сигналы.

В биологии существуют методы распознавания, не получившие, даже в постановочном плане, применения в технических системах. На рис. 1 приведена структура гипотетической системы, в которой звуковые образы от отдельных источников в виде парциальных акустических сигналов $\boldsymbol{q}_n$ суммируются на барабанной перепонке уха, образуя групповой сигнал $\boldsymbol{q}$. Система распознает и оценивает веса c_n, $n = 0, \ldots, M-1$, каждого парциального сигнала. Под распознаванием здесь понимается решение о том, что групповой акустический сигнал в качестве парциальных содержит, например, звуки речи, колокола и автомобиля, а весами являются уровни их громкости. Результаты получают одновременно на выходах, причем номер выхода является номером класса парциального сигнала. Само значение выходной величины служит оценкой $\hat{c}_n$ веса парциального сигнала $\boldsymbol{q}_n$, где c_n, $n = 0, \ldots, M-1$, — вещественные или комплексные числа. Если $c_n = 0$, то принято решение об отсутствии парциального сигнала $\boldsymbol{q}_n$.

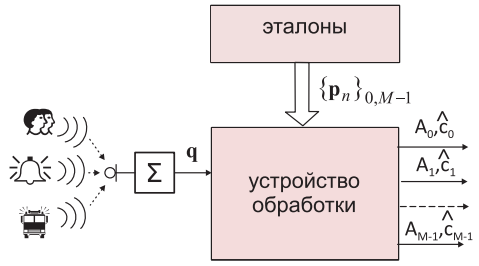


Рис. 1. Структура акустического анализатора.

Концепция группового распознавания, содержание которой отражено в примере, основана на следующих принципах:

1) групповой сигнал $\boldsymbol{q}$ является суммой зашумленных взвешенных парциальных сигналов $\boldsymbol{q}_n$, *одновременно* поступивших на вход;
2) система *одновременно* реагирует на все парциальные сигналы, обеспечивая возможность анализа входной обстановки в темпе поступления информации;
3) групповой сигнал $\boldsymbol{q}$ обрабатывается в *неразрешенном* виде, что создает эффект параллельного распознавания и оценки весов парциальных входных сигналов.

Термин «параллельное распознавание образов» широко используется, но во всех известных автору работах он подразумевает лишь ускорение процесса за счет параллельной работы группы процессоров над частями тем или иным способом декомпозированного алгоритма или сигнала. Как следует из основополагающих публикаций [1, 2, 3], «распознавание представляет собой отнесение исследуемого объекта *одному* из *взаимоисключающих* классов. Это означает, что существует *однозначное* отображение совокупности наблюдений, являющееся конечным множеством $\{\boldsymbol{X}\}$, на множество классов $\{s\} = \{s_1, s_2, \ldots, s_k\}$, количество которых задано» [2]. Таким образом, в существующих системах на вход подается сигнал *только одного класса*, и в этом плане они являются машинами последовательного действия и не отвечают принципам приведенной выше концепции.

Механизмы работы слуховых анализаторов изучены недостаточно хорошо. Учитывая важность решаемых ими задач, ниже с позиции сформулированной концепции рассматривается более простая система распознавания, алгоритм которой основан на теории обработки сигналов; оценивается степень его помехоустойчивости.

Исходные условия

Алфавит классов $A = \{A_n\}_{n=0}^{M-1}$, $\boldsymbol{P} = \{\boldsymbol{p}_n\}_{n=0}^{M-1}$, где $\boldsymbol{p}_n = \{p_n(r)\}_{r=0}^{s-1}$ — единственный эталонный вектор класса A_n, имеющий размерность s. Системы с таким алфавитом применяются для распознавания векторных сигналов в устройствах передачи информации. Эталонный сигнал $\boldsymbol{p}_n$ при распространении по каналу связи случайным образом меняет свои параметры (амплитуду, фазу и др.), зашумляется и принимает вид $\boldsymbol{q}_n = c_n\boldsymbol{p}_n + \boldsymbol{z}_n$, где $\boldsymbol{z}_n = \{z_n(r)\}_{r=0}^{s-1}$ — шумовой вектор. Подобные системы используются для распознавания символьной информации, в навигационных системах и др.

Далее предполагаем, что размерности всех векторов совпадают с числом классов алфавита, т. е. $s = M$. Вообще говоря, данное условие может быть

*Работа выполнена при финансовой поддержке РФФИ, проекты №07-01-0058, №08-01-00854, №08-01-120001-офи.

существенно ослаблено, но пока его выполнение дает возможность корректно ввести условие образования базиса из эталонных векторов классов алфавита A. Если P — квадратная матрица, строками которой служат нормированные векторы $\boldsymbol{p}_n$, $n = 0, \ldots, M-1$, то при выполнении последнего условия детерминант $|\boldsymbol{P}| \neq 0$. Моделью группового сигнала служит линейная комбинация из некоторых d эталонных векторов классов, $1 \leqslant d \leqslant s$:

$$\boldsymbol{q} = \sum_{n=0}^{d-1} \boldsymbol{q}_n = \sum_{n=0}^{d-1} (c_n \cdot \boldsymbol{p}_n + \boldsymbol{z}_n) = \boldsymbol{c} \cdot \boldsymbol{P} + \boldsymbol{z}, \qquad (1)$$

где $\boldsymbol{c} = \{c_n\}_{n=0}^{s-1}$ — вектор весов парциальных сигналов, $\boldsymbol{z} = \sum\limits_{n=0}^{s-1} \boldsymbol{z}_n$.

Реакция последовательной системы на групповой сигнал

Пусть $d = 1$, т. е. распознается один парциальный сигнал $\boldsymbol{q} = c\boldsymbol{p}$. Действие шума пока не учитываем. Сигнал $\boldsymbol{q}$ будет отнесен к классу A_t, если расстояние в признаковом пространстве между векторами $\boldsymbol{q}$ и $\boldsymbol{p}_t$ будет минимальным. Для нормированных векторов решающее правило имеет вид

$$\boldsymbol{q} \in A_t, \quad \text{если} \quad t = \underset{n=0,\ldots,s-1}{\arg\max} (\boldsymbol{q}, \boldsymbol{p}_n). \qquad (2)$$

Как видно из (2), скалярное произведение $(\boldsymbol{q}, \boldsymbol{p}_t)$ является мерой схожести распознаваемого сигнала с эталонным сигналом класса A_t. Здесь и далее $t \in \{0, \ldots, s-1\}$. На основании (2) последовательное устройство представляет собой s-канальную систему. В отдельном n-ом канале находится формирователь скалярного произведения (ФСП) парциального сигнала $\boldsymbol{q}$ и эталонного вектора $\boldsymbol{p}_n$. Решающее устройство принимает решение в пользу класса с номером, равным номеру канала с максимальным значением меры схожести.

Если $\boldsymbol{q} = \boldsymbol{q}_t = c_t \boldsymbol{p}_t$, то

$$(\boldsymbol{q}_t, \boldsymbol{p}_n) = c_t \sum_{r=0}^{s-1} \boldsymbol{p}_t(r)\, \boldsymbol{p}_n(r) = \begin{cases} c_t, & \text{при} \quad n = t; \\ c_t(\boldsymbol{p}_t, \boldsymbol{p}_n), & \text{при} \quad n \neq t. \end{cases} \qquad (3)$$

В силу $(\boldsymbol{p}_t, \boldsymbol{p}_n) < 1$, $t \neq n$, при отсутствии шумов всегда принимается правильное решение.

Пусть теперь значение d в (1) больше единицы. В такой ситуации решающее правило (2) теряет смысл. Тем не менее, определенный интерес представляет рассмотрение сигналов на выходе узлов ФСП. Для произвольного t-го канала с учетом (3) будем иметь:

$$(\boldsymbol{q}, \boldsymbol{p}_t) = c_t + \sum_{\substack{n=0 \\ n \neq t}}^{d-1} c_n(\boldsymbol{p}_n, \boldsymbol{p}_t). \qquad (4)$$

Если парциальный сигнал $\boldsymbol{q}_t = c_t \boldsymbol{p}_t$ не является компонентой вектора $\boldsymbol{q}$, то $c_t = 0$ и выражение (4) примет вид

$$(\boldsymbol{q}, \boldsymbol{p}_t) = \sum_{n=0}^{d-1} c_n(\boldsymbol{p}_n, \boldsymbol{p}_t). \qquad (5)$$

Из (4) и (5) следует, что в режиме обработки последовательным устройством сигнала $\boldsymbol{q}$ на выходе каждого из s каналов всегда присутствует помеховая компонента (5). Поэтому с ростом количества d парциальных сигналов и объема алфавита выходная величина ФСП t-го канала перестает быть мерой схожести сигналов, поскольку определяется в основном помеховой компонентой (5).

Информационно-устойчивая система

Если хотя бы для одной пары эталонных векторов алфавита A имеет место неравенство $(\boldsymbol{p}_m, \boldsymbol{p}_n) \neq 0$, $m \neq n$, то распознающее устройство является информационно избыточной системой. Из правила формирования парциального сигнала (1) следует также, что $(\boldsymbol{q}_m, \boldsymbol{p}_n) \neq 0$.

В информационно-устойчивой системе эталонные векторы алфавита образуют ортонормированный базис, а групповой сигнальный вектор формируется из линейно-преобразованных эталонных векторов алфавита классов. В результате помеховая компонента (5) обнуляется и, в соответствии с (4), выходной сигнал ФСП t-го канала становится равным значению параметра c_t. Поэтому при отсутствии шума парциальный сигнал $\boldsymbol{q}_t$ вызывает ненулевую реакцию *только* t-го канала (рис. 2).

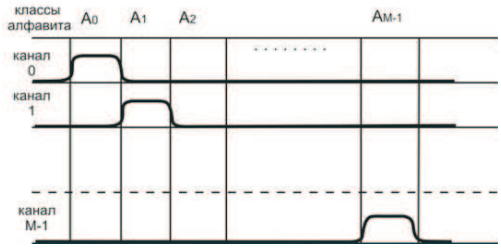


Рис. 2. Каналы информационно-устойчивой системы.

Следовательно, такое устройство успешно работает как при подаче на его вход только одного парциального сигнала, так и группового сигнала (1). При наличии шумов выходной сигнал ФСП t-го канала будет ненулевым даже при отсутствии парциального сигнала $\boldsymbol{q}_t$. Для устранения неопределенности необходимо принять обоснованное решение о характере выходного сигнала. Данную функцию выполняет решающее устройство. Оно состоит из s независимых каналов, каждый из которых содержит обнаружитель и линейный ключ. Обнаружитель t-го канала сравнивает величину $\hat{c}_t$ с порогом u_t. Если $\hat{c}_t \geqslant u_t$, то принимается решение о рас-

познавании сигнала класса A_t, а значение $\hat{c}_t$ через линейный ключ передается на выход. Если же $\hat{c}_t < u_t$, то на выходе канала будет 0, что означает отсутствие парциального сигнала $\boldsymbol{q}_t$. Величина u_t зависит от уровня шума и требуемой надежности правильного распознавания.

Общий случай распознавания

Выше была рассмотрена работа устройства, соответствующего третьему принципу концепции группового распознавания. Ограничение общности принятого подхода состоит в требовании ортонормированности базиса эталонных векторов. Оставим теперь лишь требование образования базиса. Разложим сигнал $\boldsymbol{q}$ в базисе $\boldsymbol{P}$:

$$\boldsymbol{q} = \sum_{t=0}^{s-1} a_t \boldsymbol{p}_t = \boldsymbol{a}\boldsymbol{P} = \boldsymbol{c}\boldsymbol{P}.$$

Здесь $\boldsymbol{a} = \{a_t\}_{t=0}^{s-1}$ — вектор координат (параметров) сигнала $\boldsymbol{q}$ в базисе $\boldsymbol{P}$, который, в общем случае, является косоугольным. Для определения вектора $\boldsymbol{a}$ введем еще один косоугольный базис

$$\boldsymbol{V} = \{\boldsymbol{v}_m\}_{m=0}^{s-1}.$$

Оба базиса биортогональны, т. е.

$$(\boldsymbol{p}_t, \boldsymbol{v}_m) = \begin{cases} 1, & t = m; \\ 0, & t \neq m; \end{cases} \quad t, m = 0, \ldots, s-1. \quad (6)$$

Если теперь скалярно перемножить векторы $\boldsymbol{q} = \boldsymbol{c}\,\boldsymbol{P}$ и $\boldsymbol{v}_m$ и учесть (6), то получим координату a_m в направлении базисного вектора $\boldsymbol{p}_m$:

$$(\boldsymbol{q}, \boldsymbol{v}_m) = \sum_{t=0}^{s-1} a_t(\boldsymbol{p}_t, \boldsymbol{v}_m) = a_m = c_m,$$

где $m = 0, \ldots, s-1$. Умножая обе части равенства $\boldsymbol{q} = \boldsymbol{c}\boldsymbol{P}$ на $\boldsymbol{V} = \boldsymbol{P}^{-1}$ и учитывая (6), получим

$$\boldsymbol{c} = \boldsymbol{q}\boldsymbol{P}^{-1} = \boldsymbol{q}\boldsymbol{V}. \quad (7)$$

Процедура (7) в общем случае при отсутствии флуктуационных шумов решает все задачи, связанные с третьим принципом концепции группового распознавания образов для случая произвольного базиса (рис. 3).

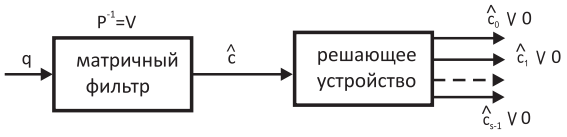


Рис. 3. Структура системы распознавания групповых сигналов.

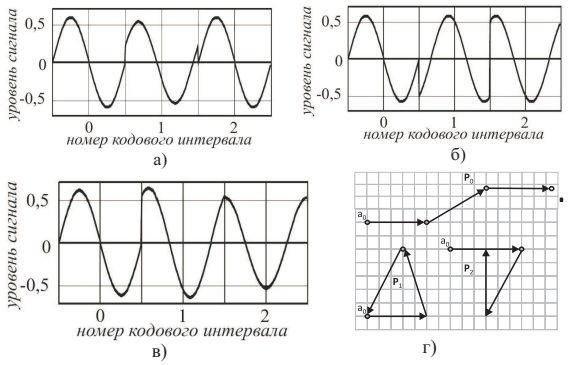


Рис. 4. Фазокодированные сигналы: а), б), в) гармонические колебания в пределах кодовых интервалов классов A_0, A_1 и A_2; г) векторные диаграммы сигналов.

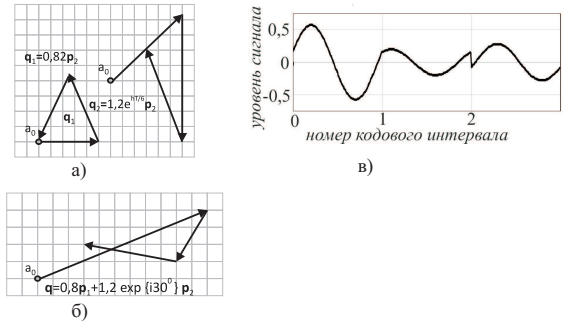


Рис. 5. Групповой сигнал $\boldsymbol{q}$: а) векторные диаграммы парциальных сигналов, б) векторная диаграмма сигнала $\boldsymbol{q}$, в) групповой фазокодированный сигнал.

Пример 1. Алфавит A состоит из классов A_0, A_1 и A_3. Эталонный вектор каждого класса задает фазокодированный сигнал из трех кодовых интервалов (рис. 4):

$$\begin{aligned} \boldsymbol{p}_0 &= \{0{,}59;\ 0{,}5 + 0{,}24\boldsymbol{i}\,;0{,}59\}; \\ \boldsymbol{p}_1 &= \{0{,}58;\ -0{,}29 + 0{,}5\boldsymbol{i};\ -0{,}29 - 0{,}5\boldsymbol{i}\}; \\ \boldsymbol{p}_2 &= \{0{,}60;\ -0{,}30 - 0{,}52\boldsymbol{i};\ -0{,}52\boldsymbol{i}\}. \end{aligned} \quad (8)$$

Строки эталонной квадратной матрицы $\boldsymbol{P}$ получаются из компонент данных векторов, являющихся в общем случае комплексными числами. Определитель матрицы $\boldsymbol{P}$ равен $0{,}964\boldsymbol{i}$. Компонента $p_n(r)$, $n, r = 0, 1, 2$, задает синусоидальное колебание в пределах r-го кодового интервала сигнала n-го класса. Амплитуда колебания равна модулю числа, а начальная фаза — его аргументу. Вектор $\boldsymbol{q}$ группового сигнала образуется в виде линейной комбинации эталонных векторов $\boldsymbol{p}_1$ и $\boldsymbol{p}_2$:

$$\begin{aligned} \boldsymbol{q} &= 0{,}8\boldsymbol{p}_1 + 1{,}2\exp\left(\tfrac{1}{6}\boldsymbol{i}\pi\right)\boldsymbol{p}_2 = \\ &= \{1{,}1 + 0{,}4\boldsymbol{i};\ -0{,}3 - 0{,}3\boldsymbol{i};\ -0{,}5 + 0{,}1\boldsymbol{i}\}. \end{aligned} \quad (9)$$

Из (9) видно, что фазокодированный сигнал $\boldsymbol{p}_1$ масштабируется с весом $c_1 = 0{,}8$, а сигнал $\boldsymbol{p}_2$ — с весом 1,2, и дополнительно поворачивается на угол 30°, т. е. $c_2 = 1{,}2\exp\left(\frac{1}{6}\boldsymbol{i}\pi\right)$, рис. 5. Выходной

сигнал матричного фильтра, полученный при обработке сигнала (9) в соответствии с (7), имеет вид $\boldsymbol{c} = \{0;\ 0{,}8;\ 1{,}1 + 0{,}6\boldsymbol{i}\}$. Отсюда следует, что парциальный сигнал $\boldsymbol{q}_0$ класса A_0 отсутствует, а имеются лишь сигналы классов A_1 и A_2, соответственно с весами $c_1 = 0{,}8$ и $c_2 = 1{,}1 + 0{,}6\boldsymbol{i} = 1{,}2\exp\left(\frac{1}{6}\boldsymbol{i}\pi\right)$.

Анализ матричного фильтра

Одним из существенных отличий описанной системы распознавания является наличие s-канального матричного фильтра, обеспечивающего параллельный режим работы. Он вырабатывает координаты вектора весов $\boldsymbol{c}$ в базисе $\boldsymbol{P}$, играющие роль статистик для принятия решения о классах парциальных сигналов. Обычно в последовательных системах для получения мер схожести используют согласованные фильтры. При наличии широкополосных шумов они являются оптимальными по критерию максимума выходного отношения сигнал/шум [4]. В нашем случае алгоритм формирования матричным фильтром статистик связан с обращением матрицы эталонов (7). В отличие от согласованных фильтров, частотные и временные характеристики которых зависят от свойств только одного эталонного сигнала, характеристики отдельного канала определяются *всеми* s эталонными сигналами. Проведенные исследования показали, что матричный фильтр реализует алгоритм инверсной фильтрации [5]. Были получены выражения для выходного сигнала фильтра и частотного коэффициента передачи. Для оценки помехоустойчивости матричного фильтра его шумовые свойства сравнивались со свойствами согласованного фильтра. Было показано, что t-й канал матричного фильтра проигрывает в D_t раз по значению выходного отношения сигнал/шум, где $D_t = \|\boldsymbol{p}_t\|^2\,\|\boldsymbol{v}\|^2 \geqslant 1$. Поскольку характеристики обнаружителя сигнала на базе согласованного фильтра известны [4], то значение параметра D_t дает возможность построить характеристики правильного распознавания парциального сигнала $\boldsymbol{q}_t$. Из выражения для величины D следует, что эффективности матричного и согласованного фильтров будут совпадать, если эталонная матрица $\boldsymbol{P}$ будет ортогональной.

Выводы

В докладе предложена концепция группового распознавания зашумленной аддитивной смеси векторных сигналов разных классов и рассмотрен пример системы, работающей в соответствии с ней. Каждый векторный сигнал задает отдельный класс, определяемый одним эталонным вектором, представляющим этот сигнал в незашумленном виде, и при стандартных значениях масштаба и угла поворота.

Первый принцип концепции носит технический характер. Его реализация связана с возможностью подачи сигналов на вход системы в темпе их поступления.

Второй принцип выполняется с помощью модифицированного алгоритма умножения вектора на матрицу. По этому алгоритму выходной вектор получается путем суммирования строк матрицы с весом, равным компоненте входного вектора, номер которой равен номеру строки.

Третий принцип реализуется при условии, когда эталонные сигналы классов алфавита образуют базис. Разложение вектора группового сигнала в этом базисе позволяет выделить отдельные (парциальные) сигналы из состава группового сигнала в упорядоченном виде. Данная процедура реализуется матричным фильтром. Он представляет собой совокупность (по числу классов алфавита) ортогональных фильтров, каждый из которых пропускает парциальный сигнал только своего класса, тем самым создавая возможность работы в параллельном режиме. Выход каждого ортогонального фильтра связан с входом обнаружителя, порог срабатывания которого рассчитан на допустимый уровень вероятности ложного решения.

Получены временные и частотные характеристики каналов матричного фильтра и показано, что система распознавания групповых сигналов минимизирует средний риск принятия решения при воздействии белого шума лишь в том случае, когда матрица эталонных векторов классов является ортогональной. Чем больше разница энергий вектор-строки этой матрицы и соответствующего вектора-столбца обратной матрицы, тем хуже помехоустойчивость системы. Это является платой за возможность работать в режиме параллельного распознавания при наличии корреляции между эталонными векторами классов. Условие образования базиса из эталонных векторов классов алфавита возможно лишь при равенстве размерностей этих векторов объему алфавита, т. е. при $M = s$. Если же размерности векторов разные, но не превышают величины M, то эквализация векторов по размерности возможна за счет введения дополнительных нулевых компонент. Данный подход нуждается в дополнительном исследовании.

Литература

[1] *Фукунага А.* Введение в статистическую теорию распознавания образов — М.: Наука, 1979.

[2] *Фомин Я. А., Тарловский Г. Р.* Статистическая теория распознавания образов. М.: Радио и связь, 1986.

[3] *Бакут П. А., Бабуров Э. Ф., Варфоломеев А. М. и др.* Параллельные методы и средства распознавания образов, том 2. — Киев: Наукова думка, 1985.

[4] *Лезин Ю. С.* Оптимальные фильтры и накопители импульсных сигналов — М.: Сов. радио, 1969.

[5] *Василенко Г. И.* Голографическое опознавание образов — М.: Сов. радио, 1977.

Распознавание изображений посредством представлений в различном числе градаций

Харинов М. В., Гальяно Ф. Р.
khar@iias.spb.su, galiano@oogis.ru
Санкт-Петербург, СПИИРАН

Исследуется проблема создания автоматизированной программной системы для выделения и идентификации объектов без использования эталонов и описаний объектов по прототипу. Предлагается способ распознавания объектов по признакам, порождаемым алгоритмами преобразования изображения в новые изображения (признаковые представления), которые нормализуются по встречаемости пикселов независимо от увеличения геометрического масштаба представления и его «изоморфных» преобразований — сдвига, растяжения признаковых осей, упаковки значений пикселов, и других.

В докладе обсуждается способ распознавания изображения на основе его описания как устройства для хранения и передачи информации, которое разрабатывается в рамках объединения нескольких подходов к информации [1, 2, 3].

В нашем подходе, экспериментально обоснованном в задачах стеганографии и целочисленной оценки количества информации, полагается, что изображение обладает аппаратно-независимой цифровой памятью, подобной памяти компьютера [4, 5, 6]. В отличие от памяти современных компьютеров, аппаратно-независимая память состоит не из битов, а из тритов [1], порождается изображением и содержит изображение в инвариантном представлении, вычисляемом независимо от предусмотренных линейных, а также нелинейных преобразований шкалы яркости изображения. Подход формализуется в *физической* модели изображения как запоминающего устройства, в которой каждому пикселу взаимно однозначно сопоставляется последовательность тритов запоминающей ячейки аппаратно-независимой цифровой памяти, и рассматриваются преобразования значений пикселов, а также в *математической* модели, в которой значения пикселов рассматриваются как целостные числа.

В простейшем случае распознавание изображений сводится к обнаружению на них объектов единственного типа. В задаче распознавания от обучаемой программной системы требуется обнаружить на некотором множестве изображений[1] пикселы, относимые к пикселам объекта пользователем, эксплуатирующим обучаемую программную систему. Обучение выполняется посредством последовательного указания пользователем на изображениях пикселов объектов в качестве *контрольных точек*. От результирующего алгоритма требуется, чтобы он не зависел от порядка выбора контрольных точек и упорядочивал пикселы изображения по их близости к пикселам искомых объектов. Если требуется дальнейшая автоматизация распознавания, то задача ограничивается изображениями конкретного типа и сводится к моделированию применяемого пользователем алгоритма выбора контрольных точек.

Формализация решения в рамках сформулированной задачи строится на основе понятия инвариантного представления изображения в заданном числе градаций яркости, выражающимся числом Мерсенна [4]. Целью доклада является описание этого понятия в математической модели изображения, а также обсуждение его применения для распознавания изображений.

Инвариантное представление изображения

Преобразование H изображения u в инвариантное представление Hu относится к преобразованиям улучшения качества изображения посредством выравнивания гистограммы [7] и является «изоморфным» преобразованием.

Определение 1. *Под изоморфным преобразованием понимается преобразование изображения в представление, не меняющееся при упаковке изображения по яркости*[2].

Под *представлением* изображения понимается новое изображение, полученное замещением значений яркости пикселов некоторыми новыми значениями.

Инвариантное представление Hu не зависит от изоморфных преобразований изображения u. Результат двукратного применения преобразования H совпадает с результатом однократного:

$$H^2 = H,$$

т. е. преобразование H относится к *идемпотентным* преобразованиям.

Инвариантное представление Hu строится в «псевдотроичной» системе счисления [4], в которой неотрицательные целые числа раскладываются

[1] При автоматизации редактирования рассматривается одно, а при автоматическом распознавании — все возможные изображения.

[2] *Упаковка* сводится к нумерации по порядку встречающихся на изображении значений яркости и замещению полученными номерами исходных значений яркости пикселов.

по степеням 2, как в двоичной системе счисления, но с коэффициентами 0, 1 и 2, как в троичной системе. Троичность системы позволяет при построении Hu избежать ошибок, возникающих при равноправном выборе. Построение выполняется посредством итеративного разделения исходной шкалы яркости на диапазоны. При этом для бинов гистограммы встречаемости пикселов изображения вычисляют новые яркостные значения, которыми замещают исходные значения яркости пикселов изображения по следующему алгоритму [5, 6]:

Вход: изображение u, гистограмма по яркости;
Выход: инвариантное представление Hu;

1: сопоставить каждому бину гистограммы новое значение яркости, равное нулю;
2: **пока** число различных инвариантных значений не совпадает с числом бинов гистограммы
3: разделить бины гистограммы на последовательности бинов с одинаковыми новыми значениями яркости;
4: поочередно рассматривая последовательности бинов гистограммы, для каждого бина рассматриваемой последовательности новое значение яркости удвоить и
5: **если** сумма бинов в последовательности справа больше суммы бинов слева **то**
6: увеличить на 2;
7: **иначе если** суммы бинов слева и справа совпадают **то**
8: увеличить на 1;
9: заменить яркости пикселов изображения новыми значениями яркости, полученными для соответствующих бинов гистограммы.

Восстановление значений инвариантного представления на промежуточных циклах построения, которые получаются при принудительном завершении алгоритма на той или иной итерации, обеспечивается итеративным применением к инвариантному представлению Hu арифметического преобразования A, см. [4].

Определение 2. *Преобразование A сводится к делению нечетных значений пикселов нацело на* 2 *и удвоению чётных значений пикселов, предварительно поделённых нацело на* 4.

Преобразование инвариантного представления в результате дополнительных циклов выполнения алгоритма описывается линейным преобразованием A^{-1}.

Определение 3. *Преобразование A^{-1} сводится к удвоению каждого значения пиксела с последующим его увеличением на* 1.

Обозначение «A^{-1}» оправдывается тем, что преобразование A^{-1} является правым обратным для A, т. е. AA^{-1} совпадает с тождественным преобразованием. При этом арифметическое преобразование A не выводит из множества инвариантных представлений вида Hu, а предварительное преобразование A^{-1} не влияет на результат последующего преобразования H изображения в инвариантное представление:

$$HAH = AH; \quad HA^{-1} = H.$$

Представление изображения при заданном разрешении

Количество итераций построения инвариантного представления Hu определяет число градаций $2^{r(Hu)+1} - 1$ шкалы инвариантных значений яркости, в которую отображаются значения пикселов изображения u, где r — целочисленная функция.

Определение 4. *Разрешением изображения u по яркости называется число $r(u)$, равное минимальному числу повторений преобразования A, обращающих изображение в нулевое представление из пикселов с нулевыми значениями.*

В модели изображения с аппаратно-независимой памятью параметр разрешения $r(Hu)$, вычисленный для инвариантного представления Hu, определяет число троичных разрядов аппаратно-независимой памяти, занимаемых пикселами и инвариантным представлением в целом, а преобразования A и A^{-1} описывают сдвиг в сторону уменьшения и, соответственно, увеличения разрядов подобно тому, как деление нацело пополам и умножение на 2 описывают сдвиг в разрядах двоичной памяти.

Плоскости тритов аппаратно-независимой памяти $T_n u$, состоящие из тритов одного разряда, вычисляются при последовательных значениях разрешения по яркости для каждой пары представлений изображения u по разностной схеме, совпадающей со схемой вычисления двоичных битовых плоскостей: $T_n u = (H_{n+1} - 2H_n)u$, где $n = 0, 1, 2, \ldots$, а H_n обозначает преобразование изображения u в представление $H_n u$ из n плоскостей тритов, причём нулевое число плоскостей тритов аппаратно-независимой памяти сопоставляется изображению из одинаковых пикселов.

Преобразование H_n выражается через преобразование H, сдвиг посредством A или A^{-1} и параметр разрешения $r(Hu)$ в виде:

$$H_n u = A^{r(Hu)-n} Hu,$$

где, в случае отрицательного значения степени, $A^{r(Hu)-n}$ обозначает преобразование A^{-1}, повторяемое $n - r(Hu)$ раз.

Преобразование H_n является идемпотентным: $H_n^2 = H_n$ и определяет инвариантное представление $H_n u$, не зависящее от изоморфных преобразований изображения. Параметр разрешения

для преобразованного изображения $H_n u$ совпадает с любым наперёд заданным значением $n \equiv r(H_n u)$, которое описывает число троичных разрядов аппаратно-независимой памяти, занимаемых инвариантным представлением после сдвига на соответствующее число разрядов.

Распознавание с обучением

Представление при заданном разрешении по величине значений пикселов оказывается полезным для распознавания на изображении объектов, заданных набором контрольных точек обучающей выборки. Оно позволяет для выделения и распознавания объектов предложить простой способ отбора алгоритмов F преобразования изображения в изображение. Согласно предлагаемому способу преобразования F замещаются композиционными преобразованиями $F_n = H_n F$. Для каждого рассматриваемого преобразования изображения u, которое используется для обучения программной системы, вычисляется максимальное значение параметра разрешения n, при котором пикселы представления $H_n F$ в заданном наборе точек обучающей выборки принимают равные значения:

$$n = \max\colon f_n(x_1) = f_n(x_2) = f_n(x_3) = \cdots \equiv f_n,$$

где $x_1, x_2, x_3, \ldots$ — контрольные точки; $f_n(x_i)$ — значения пикселов представления $F_n u$ в контрольных точках $i = 1, 2, 3, \ldots$; f_n — *признаковое* значение, вычисляемое вместе с параметром разрешения n.

Фактически, обучение по каждому признаковому представлению не зависит от других. Порядок указания контрольных точек не влияет на вычисление результирующих значений n и f_n.

Преобразования вида F_0, для которых устанавливаются нулевые значения $n = 0$, трансформируют изображение в представление из пикселов с одинаковыми нулевыми значениями и исключаются из рассмотрения. Преобразования $F_{n\neq 0}$ с установленными для каждого из них значениями n и f_n порождают *признаковые представления* $F_{n\neq 0} u$, которые продолжают использоваться для обнаружения объектов на обрабатываемых изображениях.

Программная реализация

Под изображением u понимается полутоновое изображение, цветовая компонента полноцветного изображения, или результат преобразования в представление из 64-битных значений пикселов.

На сегодняшний день на языках Java и C++ реализовано алгоритмическое ядро программной системы распознавания, блок-схема которой показана на рис. 1.

Согласно блок-схеме, изображение посредством преобразований F с установленными параметрами

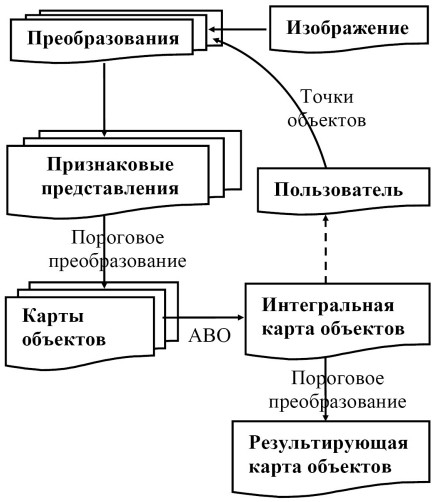


Рис. 1. Блок-схема программной системы.

n и f_n трансформируется по предусмотренным алгоритмам в набор признаковых представлений $F_n u$.

Ввод контрольных точек для каждого преобразования F_n индуцирует модификацию параметров n и f_n, в результате которой указанные параметры, если не сохраняются, то убывают. Преобразования с нулевым значением n отсеиваются для ускорения вычислений. По завершению ввода контрольных точек обработка выполняется при фиксированных параметрах в автоматическом режиме.

Каждое признаковое представление преобразуется в бинарную *карту объектов*, на которой в качестве пикселов обнаруженных объектов помечаются пикселы признакового представления со значениями, равными значению f_n.

По бинарным картам объектов в алгоритмах вычисления оценок (АВО, [8]) строится *интегральная карта объектов*, которая получается суммированием карт объектов с весами, совпадающими с установленными для признаковых представлений значений параметра разрешения. Операции выполняются по обычным правилам сложения матриц и их умножения на число.

Интегральную карту объектов получают конъюнкцией двоичных карт объектов. Экстремальными значениями на интегральной карте помечены пикселы объекта, а пикселы со значениями, близкими к экстремальным, указывают пользователю контрольные точки, при выборе которых для большого количества признаков сохраняется достаточное разрешение. Это позволяет оптимизировать выбор контрольных точек при настройке программной системы на желаемые объекты.

Результирующую карту объектов получают пороговым преобразованием интегральной карты объектов при величине порога, установленной в окрестности экстремального значения.

Для практического использования программной системы важно, что обсуждаемый способ обнаружения объектов позволяет обойтись без составления базы данных эталонов или признаковых описаний объекта по прототипу, вместо которых используются композиции некоторых базовых алгоритмов преобразования изображения в изображение. В качестве обучающей информации запоминаются только настроечные параметры (n и f_n). При этом резко снижается объём обучающей информации и, помимо ускорения обработки, достигается снижение трудоёмкости обучения.

Пример

Рис. 2 демонстрирует обнаружение характерного крестообразного объекта на стандартном изображении «Pentagon»[3].

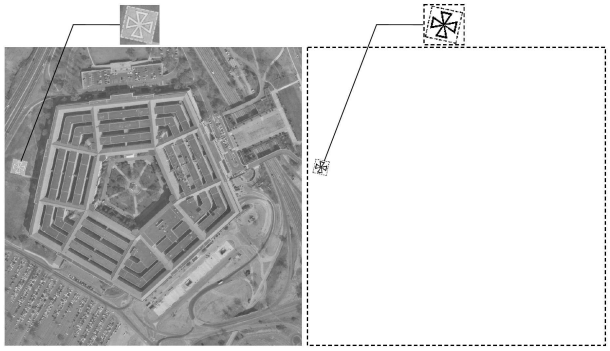


Рис. 2. Распознавание объекта на дистанционном снимке.

На рисунке слева — изображение, справа — результирующая карта объектов. Над изображениями показан объект в увеличенном виде.

Для получения признаковых представлений использованы яркость, количество информации, содержащейся в пикселах изображения, а также геометрические признаки сегментов, содержащих данный пиксел (площадь, периметр сегмента в сегментированном представлении изображения при фиксированном параметре разрешения и др.).

Выделение объекта обеспечивается указанием пользователем на интегральной карте объектов нескольких контрольных точек, занимаемых пикселами с яркостью, близкой к экстремальной. Объект на изображении выделен в виде односвязной крестообразной области, окаймленной прерывистым контуром квадратной формы. Примечательно, что иных пикселов, отождествляемых с пикселами объекта, на всём поле изображения не обнаруживается, а также, что ложные пикселы объекта не обнаруживаются на других изображениях произвольного содержания.

[3] http://sipi.usc.edu/database/

Выводы

Основная идея программной системы распознавания описана в докладе на примере объектов единственного типа. Для распознавания объектов нескольких типов достаточно применить структуру данных двухуровневого распознавания объектов по пересекающимся диапазонам признаков [4].

Особенностью предложенного способа распознавания является отбор преобразований для распознавания заданных объектов посредством приведения признаковых представлений изображения к значениям разрешения, которые устанавливаются в процессе обучения программной системы. Другой особенностью является то, что изображения анализируются программной системой в инвариантной форме представления, которая компенсирует изменчивость входных данных.

Для автоматизации формирования признаковых представлений в рамках модели изображения как запоминающего устройства разрабатываются способы сегментации (упрощения) входного изображения с сохранением его зрительного образа. Текущие результаты по сегментации включены в иллюстративные материалы к докладу и определяют перспективу дальнейших исследований.

Литература

[1] *Брусенцов Н. П.* Вычислительная машина «Сетунь» Московского государственного университета. В кн.: Новые разработки в области вычислительной математики и вычислительной техники. Киев, 1960. — С. 226–234.

[2] *Темников Ф. Е.* Информатика // Известия вузов. Электромеханика. — 1963. — № 11. — С. 1277.

[3] *Юсупов Р. М.* Теоретические основы прикладной кибернетики. Выпуск 1. Элементы теории информации. — Л.: ВИКА им. А. Ф. Можайского, 1973. — 110 с.

[4] *Харинов М. В.* Запоминание и адаптивная обработка информации цифровых изображений / Под ред. Р. М. Юсупова. СПб.: СПГУ, 2006. — 138 с.

[5] Патент РФ № 2329522. Адаптивное встраивание водяных знаков по нескольким каналам / М. В. Харинов, заявители: СПИИРАН, «Самсунг Электроникс Ко., Лтд.» // Бюл. № 20. Опубл. 20. 07. 2008. — 41 с.

[6] Патент РФ № 2331085. Двухкомпонентное встраивание сообщений в изображение / М. В. Харинов, заявители: СПИИРАН, «Самсунг Электроникс Ко., Лтд.» // Бюл. № 22. Опубл. 10. 08. 2008. — 31 с.

[7] *Прэтт У.* Цифровая обработка изображений. — М: Мир, 2007.— Т. 1–2. — 1200 с.

[8] *Журавлёв Ю. И., Никифоров В. В.* Алгоритмы распознавания, основанные на вычислении оценок // Кибернетика. — 1971. — № 3. — С. 1–11.

Распознавание пространственных групповых точечных объектов по их форме и яркости*

Хафизов Р. Г.
KhafizovRG@mail.ru
Йошкар-Ола, МарГТУ

Представлена аналитическая модель пространственного группового точечного объекта, учитывающая не только пространственные координаты его точек, но и их яркость. Исследовано влияние яркостной информации точек пространственного группового точечного объекта на устойчивость его проволочной модели и результаты распознавания.

Введение и постановка задачи

В рассмотренных ранее подходах к упорядочению и распознаванию пространственных групповых точечных объектов (ПГТО) яркость составляющих объект точек предполагалась одинаковой, а отличие объектов друг от друга определялось лишь взаимным расположением этих точек [1, 2, 3]. Такое предположение стало основой для описания ПГТО векторными кватернионами. Реально яркости отдельных точек различны, причем их значения могут различаться в десятки и более раз. Поэтому игнорирование такой информации существенно ухудшает результаты упорядочения точек и распознавания объекта.

Можно, по крайне мере, назвать две области, в которых возникают задачи, связанные с необходимостью использования яркостной информации точек ПГТО. Одной из них является задача идентификации звезды по яркостному портрету окружающих ее звезд в пределах машинного кадра астродатчика, вторая — задача обеспечения радиолокационного наблюдения групповых точечных целей. В этих задачах для принятия решения о классе объекта яркостная информация дополняет информацию о форме и размерах распознаваемого объекта.

Следует отметить, что форма объекта слабо связана с его яркостным портретом, и данный фактор обусловливает одну из особенностей задачи распознавания ПГТО. При решении задачи нумерации граней ассоциированного с ПГТО многогранника неявно использовалось предположение об одинаковой ценности яркостной информации вершин грани и информации о ее форме. Это позволило при нахождении квадрата расстояния между гранями складывать квадрат расстояния, достигаемого за счет разницы форм грани, с квадратом расстояния, получаемого за счет яркостных портретов различных граней. Такой подход является в определенной степени допустимым при решении задачи нумерации граней, относящихся к одному и тому же многограннику. Но при распознавании ПГТО по их проволочным моделям и яркостным портретам для принятия решения о классе ПГТО, целесообразно учитывать различную важность данных двух видов информативных признаков.

Целью данной работы является исследование влияния яркостной информации точек ПГТО на устойчивость его проволочной модели и результаты распознавания в условиях воздействия координатных шумов. Один их подходов к получению кватернионной модели ПГТО, учитывающей как взаимное пространственное расположение, так и яркости каждой из его точек, базируется на интерпретации действительной части полного кватерниона в качестве значения яркости задаваемой им точки. Рассмотрим, как изменится значение меры схожести двух кватернионных сигналов (КТС) при таком подходе.

Общие соотношения для полных кватернионов

В работе [1] показано, что скалярное произведение векторных сигналов, заданных в действительном линейном пространстве $\mathbb{R}$, является составной частью скалярного произведения таких сигналов в кватернионном пространстве $\mathbb{H}$. Дополнительная гиперкомплексная часть обеспечивает более высокую информативность меры схожести объемных изображений, получаемую на основе скалярного произведения в пространстве $\mathbb{H}$ по сравнению с скалярным произведением в пространстве $\mathbb{R}$. Кроме того, представление изображений в кватернионном пространстве $\mathbb{H}$ обладает следующими достоинствами:

— адекватность алгебры кватернионов свойствам трехмерного пространства: операция поворота вектора в этом пространстве является некоммутативной, как и операция умножения кватернионов, реализующая такой поворот;
— операция вращения векторных сигналов в трехмерном пространстве при использовании кватернионов реализуется проще, чем на базе матричных методов;
— условие ортогональности кватернионных сигналов включает в качестве своей основной части условие ортогональности в линейном действительном пространстве $\mathbb{R}$; поэтому для сигна-

*Работа выполнена при финансовой поддержке РФФИ, проекты №08-01-12000 офи, №07-01-00058а, №08-01-00854а.

лов, ортогональных в пространстве $\mathbb{R}$, в пространстве $\mathbb{H}$, существует много дополнительных ненулевых градаций схожести.

Будем полагать, что чисто векторные кватернионы $q(n) = iq_1(n) + jq_2(n) + kq_3(n)$ задают проволочную модель ПГТО. Проволочная модель объекта представляет собой замкнутый полигональный контур. Представленные в кватернионном виде элементарные векторы контура в определенной последовательности проходят без разветвления через все точки вершин. Располагая проволочной моделью изображения объекта, можно аналитическим путем выполнять его различные преобразования, осуществлять фильтрацию и формировать меры схожести между объектами разных классов, оценивать параметры и распознавать изображения объектов. Преобразуем его в полный кватернион, введя вещественную часть $q_0(n)$, равную яркости J_n данной точки. Тогда кватернионный сигнал $\mathbf{Q}^f = \{q^f(n)\}_{n=0}^{s-1}$, где $q^f(n) = q_0(n) + iq_1(n) + jq_2(n) + kq_3(n)$ — полный кватернион, а s — количество точек ПГТО, будет содержать информацию о пространственном положении и яркости всех точек ПГТО. Очевидно, что

$$\mathbf{Q}^f = \mathbf{Q}_0 + \mathbf{Q},$$

где $\mathbf{Q}_0 = \{q_0(n)\}_{n=0}^{s-1}$ — это КТС, состоящий только из вещественных чисел, а $\mathbf{Q} = \{q(n)\}_{n=0}^{s-1}$ — КТС, состоящий из векторных кватернионов. КТС $\mathbf{Q}^f$ далее будем называть полным, а $\mathbf{Q}$ — векторным КТС. Два полных КТС равны друг другу при условии равенства между собой их кватернионов с одинаковыми значениями n: $\mathbf{Q}^f = \mathbf{P}^f$ при $q^f(n) = p^f(n)$, $n = 0, \ldots, s-1$.

Полный КТС $\mathbf{P}^f$, задающий ПГТО, полученный масштабированием исходного ПГТО в μ раз, поворотом на угол 2ψ вокруг оси с направляющим вектором $\mathbf{r}$ и увеличением яркости всех точек в a раз, имеет вид: $\mathbf{P}^f = a\mathbf{Q}_0 + \mu b\mathbf{Q}b^{-1}$, где $b = \cos\psi + \mathbf{r}\sin\psi$ — вращающий кватернион. Для скалярного произведения двух полных кватернионов можно записать

$$(q^f, p^f) = (q, p) + q_0p_0 - q_0p + p_0q.$$

Расстояние между полными кватернионными сигналами

Для оценки влияния яркостей точек на устойчивость его проволочной модели и на эффективность распознавания ПГТО необходимо исследовать влияние этих яркостей на меру схожести формы двух граней ассоциированного с ПГТО выпуклого многогранника и на меру схожести двух полных многогранников. При этом уровень координатных шумов не должен меняться.

Ранее в качестве меры схожести двух кватернионов или двух КТС выбиралось значение их скалярного произведения. Такое решение будет корректным лишь в том случае, когда операции выполняются над нормированными векторами. В условиях принятой модели сигналов в виде полных кватернионов, операция их нормирования из-за учета значений яркости точек приводит к уменьшению его гиперкомплексной части и, как следствие, к искажению положения точки в пространстве. Поэтому мера схожести сигналов должна базироваться не на значении скалярного произведения, а на величине расстояния между векторами, задающими две точки:

$$R^{2,f} = (p_0 - q_0) + R^2.$$

Видно, что при условии $p_0 \neq q_0$ учет яркостей точек, задаваемых кватернионами q и p, приводит к росту расстояния между ними на величину $\Delta R^2 = R^{2,f} - R^2$. Такая же тенденция сохраняется и при учете яркостей всех точек КТС $\mathbf{Q}^f = \{q^f(n)\}_{n=0}^{s-1}$ и $\mathbf{P}^f = \{p^f(n)\}_{n=0}^{s-1}$. При этом

$$\Delta R^2 = R^{2,f} - R^2 = \sum_{n=0}^{s-1}\big(q_0(n) - p_0(n)\big)^2.$$

Отсюда видно, что учет яркостей точек двух ПГТО при условии различия яркостей хотя бы одной пары их точек всегда приводит к росту расстояния между объектами. Таким образом, на базе аппарата полных кватернионов создается возможность принятия решения о принадлежности пространственно расположенных объектов к разным классам не только по различию форм их изображений, но и по различию их яркостных портретов. В этом плане следует ожидать, что учет значений яркостей точек ПГТО снизит вероятность ложных решений при их распознавании и повысит помехоустойчивость получаемой проволочной модели объекта.

Проволочная модель ПГТО представляет собой заданный векторными кватернионами замкнутый пространственный контур, проходящий единственным образом через вершины ассоциированного с ПГТО выпуклого многогранника [2]. Построение проволочной модели содержит этап упорядочения (нумерации) граней многогранника и этап упорядочения (нумерации) его вершин. Наиболее критичным к действию координатных помех, случайным образом смещающих вершины многогранника, является первый этап.

Так, последующая грань назначается из других смежных граней. В качестве последующей грани выбирается грань, максимально удаленная по расстоянию от предыдущей грани, т. е. эти грани имеют наибольшую несхожесть форм и энергий.

В целом ряде случаев наблюдается ситуация, когда все смежные грани очень похожи между собой по этим признакам. В этом случае получение проволочной модели ПГТО, задаваемого правильным многогранником невозможно. Ситуация позитивно меняется, когда точки ПГТО характеризуются различными значениями своих яркостей.

Рассмотрим с этих позиций алгоритм выбора очередной грани G_1, правильного тетраэдра при известной грани G_0.

Заданы:

1) кватернионный код $\mathbf{\Gamma}_0$ контура грани G_0 и кватернионные коды $\mathbf{\Gamma}_1$, $\mathbf{\Gamma}_2$ и $\mathbf{\Gamma}_3$ граней тетраэдра, смежных с гранью G_0;

2) значения яркостей J_0, J_1, J_2 и J_3 точек ПГТО, причем $J_0 \neq J_1 \neq J_2 \neq J_3$.

Необходимо: из смежных с G_0 граней одну по критерию максимума расстояния до контура $\mathbf{\Gamma}_0$ назначить гранью G_1.

Шаг 1. Фильтрация контуров $\mathbf{\Gamma}_1$, $\mathbf{\Gamma}_2$ и $\mathbf{\Gamma}_3$ кватернионным фильтром, согласованным с контуром $\mathbf{\Gamma}_0$. По результатам фильтрации выбирается такой порядок следования ребер в гранях, задаваемых контурами $\mathbf{\Gamma}_1$, $\mathbf{\Gamma}_2$ и $\mathbf{\Gamma}_3$, при которых максимизируются реальные части выходных сигналов фильтра. Новые контуры граней обозначим через $\mathbf{\Gamma}'_1$, $\mathbf{\Gamma}'_2$ и $\mathbf{\Gamma}'_3$. Переход от контуров $\mathbf{\Gamma}_1$, $\mathbf{\Gamma}_2$ и $\mathbf{\Gamma}_3$ к контурам $\mathbf{\Gamma}'_1$, $\mathbf{\Gamma}'_2$ и $\mathbf{\Gamma}'_3$ устраняет различие в расстояниях смежных с G_0 граней, вызванное несовпадающим с $\mathbf{\Gamma}_0$ порядком следования ребер. В результате минимизируется расстояние каждого из контуров $\mathbf{\Gamma}'_1$, $\mathbf{\Gamma}'_2$ и $\mathbf{\Gamma}'_3$ и с контуром $\mathbf{\Gamma}_0$ грани G_0.

Шаг 2. Определяем расстояния между гранью G_0 и каждой их трех смежных с ней граней. Поскольку многогранник является правильным тетраэдром, то расстояния $R_{0,1}$, $R_{0,2}$ и $R_{0,3}$ будут одинаковыми, т. е. $R_{0,1} = R_{0,2} = R_{0,3}$. Поэтому любая смежная с G_0 грань может быть выбрана в качестве очередной грани G_1.

Шаг 3. Для достижения однозначного выбора грани G_1 правильного тетраэдра переходим к описанию граней его контуров полными кватернионами. Новые контуры обозначим как $\mathbf{\Gamma}_0^f$, $\mathbf{\Gamma}_1^f$, $\mathbf{\Gamma}_2^f$ и $\mathbf{\Gamma}_3^f$. Определяем приращения квадрата расстояния ΔR_{01}^2, ΔR_{02}^2 и ΔR_{03}^2 между каждым из контуров $\mathbf{\Gamma}_1^f$, $\mathbf{\Gamma}_2^f$ и $\mathbf{\Gamma}_3^f$ с контуром $\mathbf{\Gamma}_0^f$ грани G_1. В результате получаем меры различия $R_{01}^2 + \Delta R_{01}^2$; $R_{02}^2 + \Delta R_{02}^2$; $R_{03}^2 + \Delta R_{03}^2$ каждой смежной грани тетраэдра с гранью G_0. Поскольку $R_{01}^2 = R_{02}^2 = R_{03}^2$, а яркости точек ПГТО отличаются друг от друга, то полученные меры различия смежных граней по отношению к грани G_0 также будут разными. В результате в качестве грани G_1 можно обоснованно выбрать смежную с ней грань с максимальным значением расстояния.

В качестве примера, рассмотрим объект в виде правильного тетраэдра с вершинами $a = (0; 0; 0)$, $b = (5; 2{,}89; 8{,}16)$, $c = (10; 0; 0)$ и $d = (5; 8{,}66; 0)$. Грань abd обозначена как G_0 и ее контур в кватернионном представлении имеет вид:

$$\mathbf{\Gamma}_0 = \{b-a;\ d-b;\ a-d\} = \\ = \{5i + 2{,}9j + 8{,}16k;\ 5{,}76j - 8{,}16k;\ -5i + 8{,}66k\}.$$

В качестве грани G_1 необходимо выбрать одну из трех граней abc, acd и abd, имеющих с гранью G_0 общее ребро.

Шаг 1. Выбор грани G_1 из трех смежных с G_0 граней по критерию максимума расстояния до грани G_0. Чтобы порядок следования ребер в контурах $\mathbf{\Gamma}_1 = \{a; b; c\}$, $\mathbf{\Gamma}_2 = \{b; c; d\}$ и $\mathbf{\Gamma}_3 = \{a; d; c\}$ не влиял на величину расстояния до грани G_0, нужно выбрать такую последовательность ребер в контурах $\mathbf{\Gamma}_1$, $\mathbf{\Gamma}_2$ и $\mathbf{\Gamma}_3$, при которых расстояния от каждого из них до контура $\mathbf{\Gamma}_0$ было минимальным. В этом случае полученное значение расстояния будет мерой несхожести контуров, связанной лишь различием задаваемых ими форм и размеров.

Такая задача решается путем фильтрации каждого из контуров $\mathbf{\Gamma}_1$, $\mathbf{\Gamma}_2$ и $\mathbf{\Gamma}_3$ фильтром, согласованным с контуром $\mathbf{\Gamma}_0$. В данном случае такой фильтр в ответ на поданный на его вход контур, например, $\mathbf{\Gamma}_1$ будет вырабатывать три значения скалярного произведения контуров $\mathbf{\Gamma}_1$ и $\mathbf{\Gamma}_0$, причем первое значение будет соответствовать контуру $\mathbf{\Gamma}_1$ в виде bca, второе значение — в виде cad, третье — в виде abc. Минимальное расстояние между контурами $\mathbf{\Gamma}_1$ и $\mathbf{\Gamma}_0$ будет соответствовать такой последовательности ребер a, b и c, при котором достигается максимум реальной части полученного скалярного произведения. Определим эти последовательности ребер в контурах $\mathbf{\Gamma}_1$, $\mathbf{\Gamma}_2$ и $\mathbf{\Gamma}_3$. Аналитические представления этих контуров имеют вид:

$$\mathbf{\Gamma}_1 = \{a;\ b;\ c\} = \{b-a;\ c-b;\ a-c\} = \\ = \{5i + 2{,}9j + 8{,}16k;\ 5i - 2{,}9j - 8{,}16k;\ -10i\};$$
$$\mathbf{\Gamma}_2 = \{b;\ c;\ d\} = \{c-b;\ d-c;\ b-d\} = \\ = \{5i - 2{,}9j - 8{,}16k;\ -5i + 8{,}66k;\ -5{,}77j + 8{,}16k\};$$
$$\mathbf{\Gamma}_3 = \{a;\ d;\ c\} = \{d-a;\ c-b;\ a-c\} = \\ = \{5i + 8{,}66j;\ 5i - 8{,}66j, -10i\}.$$

По результатам согласованной с контуром $\mathbf{\Gamma}_0$ фильтрации этих контуров максимум реальной части выходного сигнала достигается соответственно в следующих отсчетах:

— для контура $\mathbf{\Gamma}_1$: $199{,}9 + 70{,}67i + 40{,}8j - 57{,}8k$,
— для контура $\mathbf{\Gamma}_2$: $199{,}9 - 0{,}16i + 81{,}16j - 115{,}5k$,
— для контура $\mathbf{\Gamma}_3$: $199{,}9 - 70{,}7i - 112j + 0{,}2k$.

При этом следование ребер в этих контурах происходило в следующем порядке:

$$\begin{aligned}
\mathbf{\Gamma}'_1 &= \mathbf{\Gamma}_1 = \{a;\, b;\, c\} = \{b-a;\, c-b;\, a-c\} = \\
&= \{5i + 2{,}9j + 8{,}16k;\, 5i - 2{,}9j - 8{,}16k;\, -10i\};\\
\mathbf{\Gamma}'_2 &= \{c;\, d;\, b\} = \{a-b;\, b-c;\, d-b\} = \\
&= \{5i - 8{,}66j;\, -5i + 2{,}89j + 8{,}16k;\, 5{,}77j - 8{,}16k\};\\
\mathbf{\Gamma}'_3 &= \mathbf{\Gamma}_3 = \{a,d,c\} = \{d-a, c-b, a-c\} = \\
&= \{5i + 8{,}66j;\, 5i - 8{,}66j, -10i\}.
\end{aligned}$$

Из полученных значений выходных сигналов согласованного фильтра видно, что все они имеют одинаковые значения своих реальных частей, равное 199,9. Так как многогранник является правильным, то нормы контуров всех граней будут тоже одинаковыми. С учетом длин ребер, равных 10, получим, что расстояния всех граней до грани G_0 будут иметь одно и то же значение

$$R^2 = 300 + 300 - 2 \cdot 199{,}9 = 200{,}2.$$

Из данного результата следует, что в правильном тетраэдре на основании критерия максимума расстояния смежных граней, заданных векторными кватернионами, до грани G_0, отсутствует возможность нумерации грани.

Шаг 2. Для нумерации граней правильного тетраэдра воспользуемся информацией о различной яркости его вершин. Примем следующие значения этих яркостей: $J_a = 3,\ J_b = 6,\ J_c = 12,\ J_d = 24$. Тогда полные кватернионные сигналы, задающие грани тетраэдра, будут иметь вид:

$$\begin{aligned}
\mathbf{\Gamma}_0^f &= \{3 + 5i + 2{,}9j + 8{,}16k;\\
&\qquad 18 + 5{,}76j - 8{,}16k;\, -21 - 5i - 8{,}66j\};\\
\mathbf{\Gamma}_1^f &= \{3 + 5i + 2{,}9j + 8{,}16k;\\
&\qquad 6 + 5i - 2{,}9j - 8{,}16k;\, -9 - 5i\};\\
\mathbf{\Gamma}_2^f &= \{-6 - 5i + 2{,}89j + 8{,}16k;\\
&\qquad 18 + 5{,}77j - 8{,}16k;\, -12 + 5i - 8{,}66j\};\\
\mathbf{\Gamma}_3^f &= \{3 + 10i;\, 12 - 5i + 8{,}66j;\, 24 - 5i - 8{,}66j\}.
\end{aligned}$$

Учет яркости точек ПГТО приводит к росту квадрата расстояния между контуром $\mathbf{\Gamma}_0^f$ грани G_0 и контурами $\mathbf{\Gamma}_1^f$, $\mathbf{\Gamma}_2^f$ и $\mathbf{\Gamma}_3^f$ соответственно на величину 288, 362 и 2261. Расстояния станут соответственно равны 22,09, 19,2 и 47,5. Таким образом, максимальное расстояние будет между контурами $\mathbf{\Gamma}_0^f$ и $\mathbf{\Gamma}_3^f$ и поэтому грань, задаваемая контуром $\mathbf{\Gamma}_3^f$, будет гранью G_1.

В данном случае учет яркостей точек ПГТО позволяет осуществить нумерацию граней ассоциированного с ПГТО правильного тетраэдра. В такой ситуации решить подобную задачу, учитывая лишь разницу форм и размеров граней правильного тетраэдра невозможно. Но и в других ситуациях, когда ассоциированные с ПГТО многогранники являются неправильными телами, часто расстояние между гранями, измеренные без учета яркостей точек, незначительно отличаются друг от друга. В результате сформированная проволочная модель ПГТО обладает низкой помехоустойчивостью. Поэтому различная яркость точек ПГТО служит полезным информативным признаком, дающим возможность увеличить помехоустойчивость проволочной модели объекта при слабом различии граней многогранника.

Выводы

В данной работе исследовано влияния яркостной информации точек ПГТО на устойчивость его проволочной модели и результаты распознавания в условиях воздействия координатных шумов. Разработана аналитическая модель ПГТО, учитывающая не только пространственные координаты его точек, но и их яркость. Основой модели является полный кватернион, векторная часть которого задает положение точки в пространстве, а вещественная — уровень излучаемой ею энергии — яркость, цвет и др. Показано, что использование яркостного портрета ПГТО повышает устойчивость проволочной модели ПГТО для случаев, когда грани ассоциированного многогранника слабо отличаются друг от друга по форме. Кроме того, яркостный портрет ПГТО дает дополнительную информацию о сходстве/различии распознаваемого и эталонных ПГТО, что увеличивает эффективность распознавания.

Литература

[1] *Фурман Я. А., Хафизов Д. Г.* Распознавание групповых точечных объектов в трехмерном пространстве // Автометрия. — 2003. — № 1.

[2] *Фурман Я. А., Рябинин К. Б., Красильников М. И.* Проволочная модель пространственного группового точечного объекта // Автометрия. — 2008. — № 3.

[3] *Фурман Я. А.* Сегментация и описание трехмерных структур на базе кватернионных моделей // Наукоемкие технологии. — 2007. — № 9.

Сравнение эффективности дискретных вейвлетов малого порядка*

Хашин С. И.
khash2@mail.ru
Иваново, Ивановский Университет

В настоящей работе предлагается способ выбора оптимального вейвлета, минимизирующий длину высокочастотной компоненты. Найдены примеры биортогональных вейвлетов, которые почти втрое эффективнее вейвлета Хаара.

Проблема выбора дискретного вейвлета обсуждается во многих работах, см., например, [4, 9, 13]. Однако даваемые рекомендации значительно разнятся, предлагают ориентироваться на свойства гладкости, на количество нулевых моментов, на длину носителя и др. Но в любом случае не дается убедительная мотивировка выбора. Предлагается в каждом случае выбрать вейвлет, «наиболее подходящий к решаемой задаче», но без явных критериев.

Каждая компонента RGB-цвета (или YUV-цвета) для изображения размером $mx \times my$ задается матрицей размера $mx \times my$. Будем рассматривать эту матрицу как вектор из $\mathbb{R}$-пространства L размерности $mx \times my$ с обычной евклидовой метрикой.

Одна из основных идей дискретного вейвлет анализа [1, 3, 4, 9] заключается в разложении пространства L в прямую сумму двух подпространств $L = L' \oplus L''$ — «низкочастотной» и «высокочастотной» компонент.

Помимо различных «биологических» объяснений пользы от такого разложения [2, 5], существует и более формальное. Если взять «случайное» разложение, то можно ожидать, что длина проекции вектора на каждое из подпространств составит примерно $1/\sqrt{2}$ от длины исходного вектора, то есть примерно 71%. Однако для реальных изображений длина высокочастотной компоненты не просто меньше, а меньше на порядок. Это и позволяет подтвердить вывод, что большая часть информации содержится в низкочастотной компоненте.

При применении вейвлетов для сжатия изображений желательно выбрать такой вейвлет, для которого длина высокочастотной компоненты была бы как можно меньше. Например, для стандартного тестового файла `lena.bmp` размером 512×512 точек, длина проекции на высокочастотную компоненту Хааровского разложения составляет лишь 4,281% от длины исходного вектора.

Однако, указанное число больше зависит не от вейвлета, а от выбранного изображения. Более стабильным оказывается отношение длины проекции на заданное подпространство к длине проекции на высокочастотную компоненту разложения Хаара. Это отношение и можно использовать как оценку качества вейвлета. Например, для разложения, соответствующему вейвлету Добеши D4, среднее значение эффективности равно 90,4%, а среднеквадратичное отклонение — 6,5%.

В работах [6, 7, 8] рассматриваются другие подходы к оценке качества вейвлетов, однако предлагаемый метод является значительно более точным и надежным.

В работе [10] рассматривается метод оценки качества ортонормальных базисов, основанных на аналогичных идеях.

В работе [12] предлагается оценка качества вейвлетов по длине высокочастотной компоненты (правда, лишь для ортогональных вейвлетов). Дальнейшего развития идея не получила, вероятно, из-за того, что не удалось сформулировать более стабильную версию оценки.

Учитывая, что одно из важнейших приложений вейвлетов — сжатие изображений, иногда предлагается оценивать качество вейвлета через размер соответствующего сжатого файла. Но математически исследовать эту зависимость слишком сложно, она сильно зависит от деталей реализации сжатия, никак не связанных с вейвлетами. Поэтому больших продвижений в этом направлении на сегодняшний день не отмечается.

Обозначения

Пусть L — пространство действительнозначных функций на множестве целых чисел $(0, \ldots, n-1)$. В случае необходимости будем доопределять функции на все множество целых чисел, полагая их равными 0 за пределами этого интервала. На пространстве L будем рассматривать обычное скалярное произведение и евклидову норму. Через R_k обозначим оператор сдвига на k:

$$R_k(f)(x) = f(x-k). \tag{1}$$

Определение. Пусть $f(x)$ — функция из L. Обозначим через $\rho_2(f)$ подпространство в L, порожденное всеми сдвигами функции f на четное число:

$$\rho_2(f) = \{R_{2k} f, k \in \mathbb{Z}\}. \tag{2}$$

Определение. Пару функций $(a(x), b(x))$ из L таких, что $L = \rho_2(a) \oplus \rho_2(b)$, назовём *вейвлетом*.

*Работа выполнена при финансовой поддержке РФФИ, проект №07-07-00178.

Замечание. При заданной функции a пространство $\rho_2(b)$ определено однозначно. Функцию b можно взять, например, такой: $b(i) = (-1)^i a(n-i)$. Таким образом, по существу, весь вейвлет определяется уже функцией a. Будем называть ее *порождающей* функцией вейвлета.

Примеры. Для вейвлета Хаара порождающая функция равна $(1, -1)$ (все остальные значения нулевые).

Определение. Пусть $a(x), b(x)$ — пара функций из L, удовлетворяющая условиям:

1) $(a, a) = (b, b) = 1$;
2) $(a, R_{2k}(a)) = 0,\ \forall k \in \mathbb{Z},\ k \neq 0$;
3) $(b, R_{2k}(b)) = 0,\ \forall k \in \mathbb{Z},\ k \neq 0$;
4) $(a, R_{2k}(b)) = 0,\ \forall k \in \mathbb{Z}$.

Эти условия означают, что система функций $W = (R_{2k}a, R_{2k}b)$, полученных из пары (a, b) сдвигом на все четные числа, образует ортонормальный базис в пространстве L. В этом случае будем называть базис W *ортогональным вейвлет-базисом*, а саму пару функций (a, b) — ортогональным вейвлетом.

Если $a(x)$ — порождающая функция неортогонального вейвлета, для нее условия (2) могут не выполняться. Поэтому величины

$$\{(a, R_2(a)),\ (a, R_4(a)),\ \dots\} \tag{3}$$

могут служить мерой отклонения вейвлета, порожденного функцией, от ортогональности. Эти числа будем называть *дефектом*.

Эффективность вейвлета и функции

Определения. Пусть $a(x)$ — функция из L такая, что размерность подпространства $\rho_2(a)$ равна $n/2$. Для произвольного вектора v длины n через $P_a(v)$ обозначим проекцию вектора v на подпространство $\rho_2(a)$. Если M — матрица ширины n, то через $P_a(M)$ обозначим матрицу, строки которой являются проекциями на подпространство $\rho_2(a)$ строк исходной матрицы M.

Через $\Phi(M, a)$ обозначим отношение длин векторов $P_a(M)$ и $P_h(M)$, где h — функция Хаара $h = (1, -1)$. Будем называть $\Phi(M, a)$ *эффективностью* функции a на матрице M.

Эффективностью $\Phi(a)$ функции a будем называть среднеквадратичное значение величины $\Phi(M, a)$ на наборе матриц, соответствующим достаточно представительному набору фотореалистичных изображений. Будем измерять эффективность в процентах (чем меньше, тем лучше).

Пример 1. Согласно определению, эффективность функции Хаара $h = (1, -1)$ равна в точности 100%.

Пример 2. Эффективность вейвлета Добеши D4 на наборе из 45 высококачественных изображений получается следующая (в процентах):

81,42	89,56	97,77	95,32	95,71	75,78	81,90	87,30
86,35	70,60	86,31	77,25	86,48	87,51	74,62	86,12
85,30	90,54	90,02	86,24	95,22	92,00	94,70	94,66
96,35	95,15	97,05	95,06	92,22	93,99	90,01	93,18
96,48	95,49	93,52	96,56	93,35	93,16	89,46	96,06
		93,37	95,53	94,10	95,15	95,69.	

Среднее значение равно 90,44%, а среднеквадратичное отклонение — 6,47%. Этот пример демонстрирует устойчивость введенного параметра.

Оптимальные ортогональные вейвлеты

С точки зрения сжатия изображений следует выбирать вейвлет с наименьшим значением параметра. Во всех рассматриваемых случаях наилучшее качество достигалось в случае, когда нулевой момент порождающей функции $\sum a(i)$ равен 0. Это условие можно сразу считать необходимым для выбора оптимального вейвлета.

При нахождении вейвлетов Добеши ([4]) свободные параметры выбираются так, чтобы обнулить следующие моменты функции:

$$\sum ia(i), \quad \sum i^2 a(i),\ \dots.$$

Однако, обнуление высших моментов не приводит к улучшению качества вейвлета. Более того, детальные численные эксперименты показали, что увеличение длины вейвлета не приводит к улучшению качества. Например, при длине 4 качество наилучшего вейвлета равно 81% (чем меньше, тем лучше), причем оптимальный вейвлет очень близок в вейвлету Добеши D4. При длине 6 качество оптимального вейвлета равно 74.8%, при длине 12 — 69.3%, и дальнейшее увеличение длины ортогонального вейвлета никак не способствует улучшению качества (проверки производились вплоть до длины 40).

Другими словами, ни при какой длине ортогонального вейвлета не удается подобрать его коэффициенты так, чтобы длина высокочастотной компоненты была заметно меньше 70% от длины Хааровской компоненты.

Кроме того, следует учесть, что качество вейвлета определяется не математически, а экспериментально — на примере некоторого, достаточно представительного набора изображений. Поэтому вряд ли можно надеяться получить точность более одного процента. А среди ортогональных вейвлетов с носителем заданной длины существует очень большая область, в которой качество весьма близко к наилучшему. Поэтому процесс нахождения оптимального ортогонального вейвлета крайне неустойчив.

Оптимальные биортогональные вейвлеты

В случае биортогональных вейвлетов ситуация существенно лучше. Рассмотрим два тесто-

вых изображения: `lena.bmp` — стандартный тестовый файл размером 512×512, и `res.bmp` — файл размером 512×5768, вырезки из разных высококачественных изображений, склеенные в один файл по высоте. Для каждого из них были найдены оптимальные вейвлеты. Их эффективность показана в таблице.

Длина	Эффективность на `lena.bmp` (%)	Эффективность на `res.bmp` (%)
2	100,000	100,000
3	77,212	89,395
4	69,190	79,488
5	59,312	67,158
6	50,813	56,736
7	47,511	54,404
8	45,808	52,083
9	42,700	48,608
10	40,284	44,994
11	38,904	44,194
12	38,501	43,449
13	36,458	41,518
13	35,162	39,541
15	34,522	39,179
16	34,619	38,730

Другими словами, биортогональный вейвлет длины 16 можно выбрать так, чтобы длина высокочастотной компоненты была почти в 3 раза меньше длины Хааровской компоненты.

Насколько устойчив этот показатель? Для ответа на этот вопрос был взят набор из 118 высококачественных изображений, для каждого из них измерено качество оптимального вейвлета, найденного на предыдущем этапе. Результат показан в следующей таблице.

Длина	Средняя эффективность на 118 изображениях (%)	Среднеквадратичное отклонение
2	100,00	0,00
3	89,36	5,68
4	80,24	9,25
5	67,57	6,99
6	57,44	5,66
7	55,00	6,50
8	52,80	7,26
9	49,32	6,29
10	45,85	5,66
11	44,92	5,86
12	44,24	6,15
13	42,26	5,63
13	40,41	5,71
15	40,15	5,80
16	39,84	5,77

Коэффициенты оптимального вейвлета длины 4 (эффективность 79,488%):

$$\begin{matrix} 0{,}28427718 & -0{,}63033235 \\ 0{,}65364725 & -0{,}30759209 \end{matrix} \tag{4}$$

Его дефект равен 0,38.

Коэффициенты оптимального вейвлета длины 8 (эффективность 52,083%):

$$\begin{matrix} 0{,}10019483 & -0{,}30143907 & 0{,}37739102 \\ -0{,}47695892 & 0{,}49804116 & -0{,}41037433 \\ 0{,}31872135 & -0{,}10557604 & \end{matrix} \tag{5}$$

Его дефект равен (0,767, 0,344, 0,0638).

Коэффициенты оптимального вейвлета длины 16 (эффективность 38,730%):

$$\begin{matrix} 0{,}03978018 & -0{,}13798418 & 0{,}17211180 \\ -0{,}23737729 & 0{,}26391534 & -0{,}31290316 \\ 0{,}30681224 & -0{,}33707334 & 0{,}36269543 \\ -0{,}34031110 & 0{,}32069307 & -0{,}26426394 \\ 0{,}23285584 & -0{,}16274193 & 0{,}12736702 \\ -0{,}03357600 & & \end{matrix} \tag{6}$$

Его дефект равен (0,927, 0,748, 0,518, 0,298, 0,137, 0,0428, 0,0053).

Дефект полученных вейвлетов довольно велик (первый коэффициент близок к 1), то есть полученные вейвлеты весьма далеки от ортогональности. Поэтому для эффективного использования полученных вейвлетов, по всей видимости, придется все-таки переходить к ортогональному базису. Это существенно повысит требуемый объем вычислений, но без этого потеряется вся найденная эффективность.

Судя по всему, при переходе к вейвлетам с еще более длинным носителем их эффективность можно еще несколько улучшить. Некоторые вычисления показывают, что для вейвлетов длины 32 можно ожидать качества порядка 30%, то есть они будут более чем втрое превосходить вейвлет Хаара.

Выводы

1. Предложена математическая оценка качества вейвлета с точки зрения проблемы сжатия изображений (чем меньше параметр, тем лучше). Продемонстрирована устойчивость предложенной оценки на реальных изображениях.

2. Экспериментально исследованы оптимальные ортогональные вейвлеты. Оказалось, что увеличение длины носителя от 6 до 20 и даже до 40 не дает практически никакого повышения качества вейвлета, его не удается снизить ниже 69%.

3. Экспериментально исследованы оптимальные биортогональные вейвлеты. Оказалось, что при длины носителя 6 качество наилучшего вейвлета равно 56%, при длине 8 — 52%, при длине 16 —

38,7%. Таким образом, имеется практический интерес в использовании оптимальных вейвлетов большой длины для сжатия графики и видео (см. например, [11]).

Литература

[1] *Ватолин Д. С.* Алгоритмы сжатия изображений. — М.: МГУ, 1999. — 80 с.

[2] *Ватолин Д., Ратушняк А., Смирнов М.* Методы сжатия данных. — М.: Диалог-МИФИ, 2003. — 382 с.

[3] *Воробьев В. И., Грибунин В. Г.* Теория и практика вейвлет-преобразования. — С.-Петербург: ВУС, 1999.

[4] *Добеши И.* Десять лекций по вейвлетам. — М.: РХД, 2001.

[5] *Дремин И. М., Иванов О. В., Нечитайло В. А.* Вейвлеты и их использование // УФН. — 2001. — Т. 171, № 5. — С. 465–501.

[6] *Злобин А. С.* Эффективность базисных функций вейвлет преобразований // Доклады 48 конференции МФТИ, Секции радиотехники и защиты информации, М. МФТИ, 2005.

[7] *Умняшкин С. В., Кочетков М. Е.* Анализ эффективности использования дискретных ортогональных преобразований для цифрового кодирования коррелированных данных // Известия вузов. Электроника. — 1998. — № 6. — С. 79–84.

[8] *Умняшкин С. В.* Эффективность применения ортогональных преобразований для кодирования дискретных сигналов с точки зрения корреляционной теории // Интеллектуальные системы в производстве. — 2003. — № 1. — С. 100–123.

[9] *Фрейзер М.* Введение в вейвлеты в свете линейной алгебры. — М.: Бином, 2008. — 488 с.

[10] *Хашин С. И.* Оптимальный ортонормальный базис в компьютерной графике // Вестник ИвГУ, Иваново. — 2008. — Вып. 2. — С. 122–127.

[11] *Хашин С. И.* Применение методов распознавания образов для сжатия видеоинформации // Докл. всеросс. конф. ММРО-13. — М.: МАКС Пресс, 2007. — С. 420-424.

[12] *Karel J. M. H. et al.* Optimal discrete wavelet design for cardiac signal processing // Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference, Shanghai, China, September 1-4, 2005.

[13] *Salomon D* Data Compression. The Complete Reference (3d Edition) // New York: Springer-Verlag, 2004. — 920 p.

Исследование зависимости СКО редискретизации цифровых сигналов от величины апертуры окна интерполяции

Чичагов А. В.
mail2chi@yandex.ru
Москва, Вычислительный центр РАН

В работе приводятся результаты исследования зависимости набора средне-квадратических ошибок редискретизации тестовых цифровых сигналов от значений параметров процедуры редискретизации, в частности, величины апертуры окна интерполяции.

В настоящее время в цифровых устройствах, таких, например, как цифровые осцилографы, сейсмографы, томографы и др. для реконструкции или визуализации цифровых сигналов (изображений, полей) используют различные алгоритмы и программы редискретизации цифровых сигналов.

Современные цифро-аналогововые преобразователи сигнала выполняют кусочно-постоянную интерполяцию между выборочными значениями сигнала, поэтому при выборе частоты дискретизации сигнала, близкой к удвоенной частоте верхней границы фурье-спектра сигнала, восстановленный аналоговый сигнал обычно имеет достаточно большую погрешность.

Существенно уменьшить погрешность восстановления аналогового сигнала можно как путем увеличения частоты дискретизации сигнала при выполнении аналого-цифрового преобразования, так и с помощью редискретизации исходной выборки сигнала к выборке с большей частотой дискретизации и затем выполнения высокоскоростного цифро-аналогового преобразования.

Второй способ часто является более предпочтительным, так как для хранения сигнала требуется меньший объем памяти. Однако в этом случае необходимо использовать довольно точные методы редискретизации цифровых сигналов.

В работе [1] были рассмотрены некоторые методы повышения частоты дискретизации и проанализированы ошибки редискретизации сигнала в цифровых осциллографах реального времени. В данной работе рассматривается общий метрологический подход к проблеме оценки точности алгоритмов и программ редискретизации цифровых сигналов на примере двух алгоритмов редискретизации сигналов, построенных на основе классических методов интерполяции Лагранжа и Уиттакера.

Постановка задачи и идея решения

Требуется оценить точность редискретизации цифрового сигнала при заданных значениях параметров процедуры редискретизации. Компоненты вектора параметров процедуры редискретизации цифрового сигнала $\boldsymbol{p} = (C_T, M_a, K_D)$ представляют:

- тип интерполяционной кривой (кодовое значение или имя метода интерполяции) C_T = = «Lagrange» | «Whittaker»;
- величину апертуры окна интерполяции M_a;
- коэффициент редискретизации цифрового сигнала $K_D = F_D^{(d)}/F_D^{(s)}$;

где $F_D^{(s)} = 1/T_D^{(s)}$ — частота дискретизации исходного сигнала (источника), $F_D^{(d)} = 1/T_D^{(d)}$ — частота дискретизации выходного (редискретизованного) сигнала.

Мультиинтерполяционное преобразование или, иначе, «передискретизацию» цифрового сигнала можно описать следующим выражением:

$$x'[n] = \sum_{m=-M}^{M} x\big[\tilde{i}(n)-m\big] f\Big(nT_D^{(d)} - \big(\tilde{i}(n)-m\big)T_D^{(s)}\Big),$$

где $x'[n]$ — элемент выходной выборки цифрового сигнала, а $\tilde{i}(n) = (nF_D^{(s)})/F_D^{(d)}$ представляет индекс элемента исходной выборки $\{x[i]\}$ цифрового сигнала, который наиболее близок к «физическому» моменту времени, соответствующему текущему индексу n элемента выходной выборки, $f(nT_D^{(d)} - iT_D^{(s)})$ — весовая функция («ядро») или функция вклада узла $x[i]$ исходной выборки в текущее значение $x'[n]$ выходной выборки, M — полуширина апертуры окна интерполяции.

В работе использовалась процедура редискретизации [2], реализующая данное преобразование:

Вход: $\boldsymbol{p} = (C_T, M_a, K_D)$, $\boldsymbol{x} = \{x[i]\}$;
Выход: $\boldsymbol{x'} = \{x'[n]\}$;
1: инициализация: $n := 0$; $i := 0$;
2: **пока** $(i < \boldsymbol{x}.length)$ **повторять**
3: $\quad x'[n] := \varphi(n, i)$;
4: $\quad n := n + 1$;
5: $\quad i := (n * F_D^{(s)})/F_D^{(d)}$;

Здесь переменные «i» и «n» можно интерпретировать как «счетчик принятых элементов» и «счетчик отправленных элементов» соответствующих выборок данных, а $\varphi(n, i)$ — обозначает некоторую интерполяционную функцию, в частности,

функцию Лагранжа

$$\varphi(n,i)=\sum_{m=-M}^{M}x[i+m]\prod_{\substack{m'=-M\\m'\neq m}}^{M}\frac{nT_D^{(d)}-(i+m')T_D^{(s)}}{(m-m')T_D^{(s)}}$$

или «усеченную» функцию Уиттакера

$$\varphi(n,i)=\sum_{m=-M}^{M}x[i+m]\frac{\sin\pi F_D^{(s)}\left(nT_D^{(d)}-(i+m)T_D^{(s)}\right)}{\pi F_D^{(s)}\left(nT_D^{(d)}-(i+m)T_D^{(s)}\right)}.$$

Легко показать, что «ядро» интерполяционной формулы Уиттакера представляет аппроксимацию «ядра» интерполяционной формулы Лагранжа для равномерной неограниченной выборки измерений. Действительно,

$$\prod_{\substack{m'=-M\\m'\neq m}}^{M}\frac{(s-m')}{(m-m')}=\prod_{\substack{m'=-M\\m'\neq m}}^{M}\frac{(s-m)+(m-m')}{(m-m')}=$$

$$=\prod_{\substack{m'=-M\\m'\neq m}}^{M}\left\{1+\frac{(s-m)}{(m-m')}\right\}\to\prod_{\substack{k=-\infty\\k\neq 0}}^{\infty}\left\{1+\frac{(s-m)}{k}\right\}=$$

$$=\prod_{k=1}^{\infty}\left\{1+\left(\frac{(s-m)}{k}\right)^2\right\}=\frac{\sin\pi(s-m)}{\pi(s-m)}.$$

Заметим, что указанные интерполяционные формулы, при неограниченном увеличении апертуры окна интерполяции, асимптотически соответствуют известной формуле теоремы Котельникова о дискретизации сигналов.

В настоящей работе предполагается, что с точки зрения оценки математической погрешности процедура «восстановления непрерывного сигнала» соответствует процедуре «редискретизации цифрового сигнала» при достаточно большой величине коэффициента редискретизации цифрового сигнала.

Идея алгоритма вычисления оценки точности редискретизации цифрового сигнала состоит в следующем. Разделим ожидаемый диапазон частот исходного цифрового сигнала $\left[0,\frac{1}{2}F_D^{(\mathrm{src})}\right]$ на L_0 узких частотных полос (интервалов) и оценим точность редискретизации цифрового сигнала в каждой частотной полосе. Пусть $\bar{\zeta}_s^2(\boldsymbol{p})$ — средне-квадратическая ошибка редискретизации гармоники полосы s, где $0\leqslant s<L_0$, тогда ошибку редискретизации произвольного цифрового сигнала $\bar{\zeta}_x^2(\boldsymbol{p})$ можно оценить по формуле

$$\bar{\zeta}_x^2(\boldsymbol{p})=\frac{\sum_{s=0}^{L_0-1}\bar{\zeta}_s^2(\boldsymbol{p})E_s^{(x)}}{\sum_{s=0}^{L_0-1}E_s^{(x)}},$$

где $\boldsymbol{p}=(C_T,M_a,K_D)$ — вектор параметров процедуры редискретизации, $E_s^{(x)}$ — выборка сглаженного квадрата модуля фурье-спектра исходного цифрового сигнала.

Таким образом, при известном спектре сигнала пользователь может оценить точность редискретизации цифрового сигнала, если ему предоставить набор оценок средне-квадратических ошибок редискретизации полосовых гармоник $\{\bar{\zeta}_s^2(\boldsymbol{p})\}$ для заданных значений вектора параметров $\boldsymbol{p}$ процедуры редискретизации цифрового сигнала. Передача информации может быть реализована с помощью оперативного доступа либо к серверу базы данных СКО ошибок редискретизации полосовых гармоник, либо к серверу контрольно-измерительного приложения (КИП), вычисляющего требуемые оценки.

План вычислительного эксперимента

Сценарий вычислительного эксперимента по оценке средне-квадратических ошибок редискретизации полосовых гармоник для заданных значений параметров $\boldsymbol{p}$ процедуры редискретизации цифрового сигнала представим в виде последовательности следующих действий:

1. Задание значений параметров процедуры редискретизации,
2. Задание значений параметров процедуры синтеза набора исходных тестовых сигналов (гармоник),
3. Вычисление набора ожидаемых (т. е. вычисленных «аналитически» или без помощи исследуемой процедуры редискретизации) тестовых сигналов, соответствующего заданным значениям параметров набора исходных тестовых сигналов и параметров процедуры редискретизации,
4. Вычисление набора фактических (вычисленных с помощью исследуемой процедуры редискретизации) тестовых сигналов, соответствующего заданному набору исходных тестовых сигналов и заданному набору значений параметров процедуры редискретизации,
5. Вычисление набора величин погрешностей редискретизации для рассматриваемых наборов ожидаемых и фактических тестовых сигналов.

Определим нормализованную средне-квадратическую ошибку (СКО) редискретизации тестового сигнала $\bar{\zeta}_s^2(\boldsymbol{p})$ в следующем виде

$$\bar{\zeta}_s^2(\boldsymbol{p})=\frac{1}{\sigma_s^2(T_1-T_0)}\sum_{t=T_0}^{T_1-1}\left|z_s(t)-\tilde{z}_s(t)\right|^2,$$

где $z_s(t)$ — ожидаемый цифровой сигнал (аналитическая кривая), $\tilde{z}_s(t)$ — фактический (редискретизированный) цифровой сигнал (мультиинтерполяционная кривая), T_0 и T_1 — соответственно начальная и конечная границы фрагментов цифровых сигналов, по которым оценивается СКО редискретизации, а $\sigma_s^2=\frac{1}{T_1-T_0}\sum\limits_{t=T_0}^{T_1-1}\left|z_s(t)\right|^2$ — нормi

ровочный коэффициент, в качестве которого выберем оценку дисперсии ожидаемого цифрового сигнала на рассматриваемом фрагменте, выбранном так, чтобы исключить влияние краевых эффектов.

Формально можно считать, что $z_s(t) = x_s(t) + iy_s(t)$ представляет комплексную величину, хотя это не имеет принципиального значения. Тестовый комплект сигналов выберем так, чтобы тестовые сигналы имели одинаковую амплитуду при одинаковой норме для всех допустимых значений s, включая $s = 0$.

В качестве тестовых сигналов выберем семейство тригонометрических функций. Набор цифровых сигналов источника или, иначе, набор исходных тестовых выборок данных представим в виде семейства временных рядов

$$x_s^{(\mathrm{src})}[i] = A_0 \cos\bigl(\pi(s/L_0)i\bigr);$$
$$y_s^{(\mathrm{src})}[i] = A_0 \sin\bigl(\pi(s/L_0)i\bigr);$$

где $i = [F_D^{(s)}t]$ — целочисленный сдвиг (смещение) элемента выборки данных, а $s = (2F/F_D^{(s)})L_0$ — нормализованная частота цифрового сигнала источника (гармоники). Выборка целочисленных значений $s \in 0, 1, \ldots, L_0{-}1$, где L_0 — размер выборки, представляет эквидистантный набор или спектр цифровых сигналов источника в полосе частот $\bigl[0, \frac{1}{2}F_D^{(s)}\bigr]$. Частота дискретизации источника $F_D^{(s)}$ выбирается равной некоторому терминальному значению, например, 2000 Гц.

Компоненты вектора параметров процедуры синтеза набора исходных тестовых сигналов $\boldsymbol{q} = (L_0, N_0, A_0)$ имеют следующую интерпретацию:

- параметр L_0 представляет размерность набора исходных тестовых цифровых сигналов («количество» гармоник),
- параметр N_0 представляет объем выборки исходного тестового цифрового сигнала («длина» гармоники),
- параметр A_0 представляет «размах» исходного тестового цифрового сигнала («амплитуда» гармоники).

Опцию выбора величины амплитуды гармоники A_0 пользователь может использовать для детального исследования или экспресс-оценки влияния эффекта квантования амплитуды сигнала на точность редискретизации, а опции выбора «количества» гармоник L_0 и объема выборки N_0 — для экспресс-оценок СКО редискретизации цифрового сигнала.

Набор ожидаемых (expected) тестовых цифровых сигналов, соответствующий заданным значениям параметров набора исходных тестовых сигналов и параметров процедуры редискретизации, запишем в виде

$$x_s^{(\mathrm{exp})}[i] = A_0 \cos\bigl(\pi K_D^{-1}(s/L_0)i\bigr)$$

и

$$y_s^{(\mathrm{exp})}[i] = A_0 \sin\bigl(\pi K_D^{-1}(s/L_0)i\bigr),$$

где $K_D = F_D^{(d)}/F_D^{(s)}$ — коэффициент редискретизации цифрового сигнала. Набор фактических (actual) тестовых цифровых сигналов представим в виде

$$\boldsymbol{x}_s^{(\mathrm{act})} = \mathrm{SRC}\bigl(\boldsymbol{p}, \boldsymbol{x}_s^{(\mathrm{src})}\bigr)$$

и

$$\boldsymbol{y}_s^{(\mathrm{act})} = \mathrm{SRC}\bigl(\boldsymbol{p}, \boldsymbol{y}_s^{(\mathrm{src})}\bigr),$$

где SRC(...) обозначает линейное преобразование исходного цифрового сигнала или, иначе, используемую процедуру редискретизациии.

Отчет с результатами вычислительного эксперимента должен быть представлен пользователю в удобном виде и содержать:

- набор значений параметров процедуры редискретизации $\boldsymbol{p} = (C_T, M_a, K_D)$,
- набор значений параметров процедуры синтеза набора исходных тестовых сигналов $\boldsymbol{q} = (L_0, N_0, A_0)$,
- таблицу значений нормализованных среднеквадратических ошибок редискретизации тестовых цифровых сигналов $\bar{\zeta}_s^2(\boldsymbol{p})$, где $s \in 0, 1, \ldots, L_0{-}1$.

Используя отчет о вычислительном эксперименте и зная спектр мощности цифрового сигнала, пользователь может оценить точность редискретизации цифрового сигнала для заданных значений параметров процедуры редискретизации.

Практика проведения вычислительных экспериментов, однако, показывает, что вместо рассмотренной выше таблицы значений нормализованных средне-квадратических ошибок редискретизации тестовых цифровых сигналов $\bar{\zeta}_s^2(\boldsymbol{p})$, где $s \in 0, 1, \ldots, L_0{-}1$, удобнее рассматривать таблицу логарифмов среднеквадратических ошибок или, точнее, использовать децибельную логарифмическую шкалу $10 \log \bar{\zeta}_s^2(\boldsymbol{p})$ (Дб) для представления величин СКО редискретизации тестовых цифровых сигналов внешним пользователям.

Анализ результатов вычислительного эксперимента

Цель рассматриваемого вычислительного эксперимента состоит в изучении зависимости среднеквадратической ошибки редискретизации набора тестовых цифровых сигналов от величины апертуры окна интерполяции. Формально, данный вычислительный эксперимент представляет логически связанную серию реализаций планов вычислительных экспериментов, описанных выше.

При проведении данного вычислительного эксперимента был получен большой экспериментальный материал, однако ввиду наличия полиграфических и других специфических ограничений, на

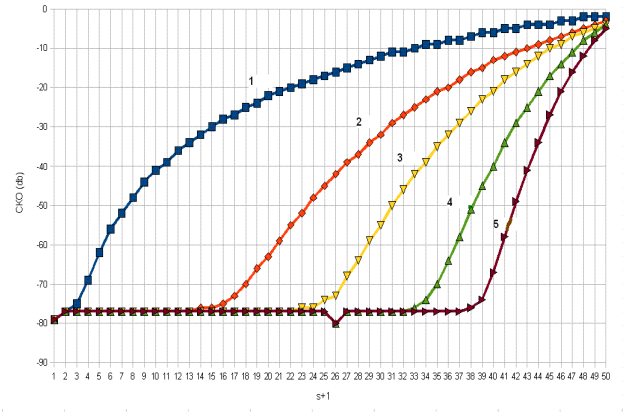


Рис. 1. Зависимость СКО редискретизации по формуле Лагранжа от нормализованной частоты для различных величин апертуры окна интерполяции. Кривые $1, \ldots, 5$ соответствуют апертуре $M_a = 3, 10, 20, 50, 100$, при $K_D = 25$, $L_0 = 50$, $N_0 = 1000$, $A_0 = 10000$.

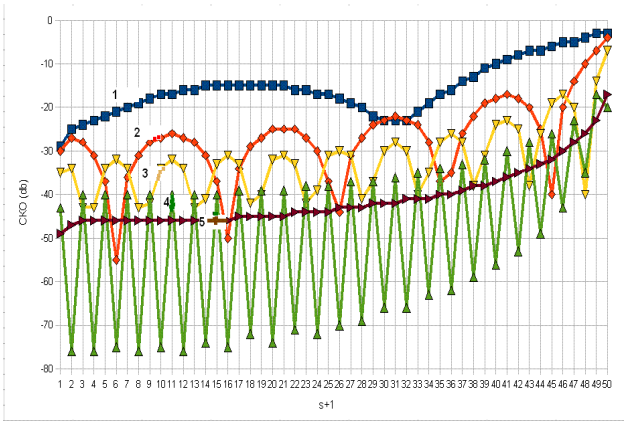


Рис. 2. Зависимость СКО редискретизации по формуле Уиттакера от нормализованной частоты для различных величин апертуры окна интерполяции. Кривые $1, \ldots, 5$ соответствуют апертуре $M_a = 3, 10, 20, 50, 100$, при $K_D = 25$, $L_0 = 50$, $N_0 = 1000$, $A_0 = 10000$.

иллюстрациях, приводимых в настоящей работе, представлена только небольшая часть существующего экспериментального материала.

На рис. 1 показаны графики зависимости нормализованной средне-квадратической ошибки редискретизации по формуле Лагранжа для набора тестовых цифровых сигналов (гармоник) при различных значениях апертуры окна интерполяции. Легко заметить, что редискретизация по формуле Лагранжа является более точной для низкочастотных гармоник, чем для высокочастотных.

Для сравнения, на рис. 2 показаны графики зависимости нормализованной средне-квадратической ошибки редискретизации по формуле Уиттакера для набора тестовых цифровых сигналов (гармоник) при тех же значениях апертуры окна интерполяции.

Можно отметить, что в отличие от редискретизации по формуле Лагранжа, СКО редискретизации по формуле Уиттакера является квазипериодической функцией нормализованной частоты гармоники s, причем величина периода этой функции зависит от значения апертуры окна интерполяции.

Причиной такого характера поведения СКО редискретизации по формуле Уиттакера является, по мнению автора, явление Гиббса, которое обычно проявляется при «обрезании» импульсной характеристики линейного фильтра или, в данном случае, ядра интерполяционной функции Уиттакера. Отсутствие «колебаний» на графике СКО редискретизации, который идентифицирует кривая (5) рис. 2, обусловлено стробоскопическим эффектом, который «проявился» в данной серии измерений.

Выводы

Получены кривые зависимости спектра СКО редискретизации тестовых цифровых сигналов (гармоник) от величины апертуры окна интерполяции, которые показаны на рис. 1 и рис. 2. Визуальное парное сравнение полученных зависимостей показывает, что при выборе частоты дискретизации сигнала более чем в три-четыре раза выше максимальной частоты спектра сигнала, процедура редискретизации цифрового сигнала, построенная на основе интерполяционной формулы Лагранжа, при одинаковых значениях апертуры окна интерполяции (лежащих в исследованном интервале $[1, 100]$), обычно оказывается значительно (в среднем приблизительно на 20 ± 10 Дб) точнее процедуры редискретизации цифрового сигнала, построенной на основе интерполяционной формулы Уиттакера.

Литература

[1] *Вьюхин В. Н.* Дискретизация в эквивалентном времени и интерполяция в цифровой осцилографии реального времени // Автометрия. — 2008. — №3. — С. 26–32.

[2] *Чичагов А. В.* Конвертор частоты дискретизации цифрового сигнала // Модели и методы распознавания речи. — М.: ВЦ РАН им. А. А. Дородницына, 2008. — С. 24–46.

Решение задачи декомпозиции сигналов заданной формы методами теории измерительно-вычислительных систем*

Чуличков А. И., Демин Д. С.
achulichkov@gmail.com
Москва, Физический факультет МГУ

В задаче декомпозиции считается, что регистрируемый сигнал можно представить в виде суммы сдвинутых относительно друг друга сигналов заданной формы с неизвестными амплитудами и величинами сдвигов. Результат регистрации сигнала сопровождается аддитивным шумом. Предлагается метод, позволяющий оценить число слагаемых, их амплитуду и положение на оси абсцисс.

Постановка задачи декомпозиции

Рассмотрим ситуацию, в которой наблюдаемый сигнал является линейной комбинацией сигналов заданного вида («элементарных сигналов»), каждый из которых сдвинут по оси абсцисс на неизвестную величину. Задачей декомпозиции назовем задачу определения числа слагаемых линейной комбинации, их коэффициентов (амплитуд элементарных сигналов) и параметров сдвигов. Основой решения являются методы теории измерительно-вычислительных систем [1–4]. Считая, что сигналы наблюдаются в дискретные моменты времени $0 = t_1 < \ldots < t_n$, будем считать, что сигнал является вектором n-мерного евклидова пространства $\mathbb{R}^n$, заданный своими координатами — значениями сигнала в заданные моменты времени. Формально запишем схему регистрации сигнала $\xi = (\xi(t_1), \ldots, \xi(t_n)) \in \mathbb{R}^n$, в виде

$$\xi(t_i) = \sum_{j=1}^{k} c_j^{(k)} \varphi(t_i - t_{0j}) + \nu(t_i), \quad i = 1, \ldots, n, \quad (1)$$

где $\varphi(\cdot)$ — известный элементарный сигнал, $c_j^{(k)}$ — априори неизвестные вещественные коэффициенты, t_{0j} — априори неизвестные времена задержки j-го элементарного сигнала, $j = 1, \ldots, k$, $\nu(\cdot)$ — шум, моделирующий погрешности измерения. Сигналы, формирующиеся в соответствие с данной моделью, возникают в различных физических задачах, связанных с зондированием слоистых структур (например, при акустическом зондировании атмосферы), задачах спектрометрии и т. д. Для получения информации о структуре исследуемого объекта необходимо, в первую очередь, оценить число элементарных сигналов k и времена задержки t_{0j}, $j = 1, \ldots, k$, несущие в случае зондирования атмосферы информацию о числе слоистых структур и их пространственном положении, соответственно.

В векторном виде схема (1) запишется в виде

$$\boldsymbol{\xi} = \Phi \boldsymbol{c}^{(k)} + \boldsymbol{\nu}, \quad (2)$$

где $\Phi\colon \mathbb{R}^k \to \mathbb{R}^n$ — линейный оператор, задаваемый матрицей $\Phi_{ij} = \varphi(t_i - t_{0j})$, $i = 1, \ldots, n$, $j = 1, \ldots, k$, векторы $\boldsymbol{\xi}$ и $\boldsymbol{\nu}$ принадлежат пространству $\mathbb{R}^n$, $\boldsymbol{c}^{(k)} \in \mathbb{R}^k$, $k < n$.

В работе предлагаются методы оценки параметров k, $\boldsymbol{c}^{(k)} \in \mathbb{R}^k$, $\boldsymbol{\tau} = (t_{01}, \ldots, t_{0k}) \in \mathbb{R}^k$ по наблюдению (2) при известной функции $\varphi(\cdot)$, если шум $\boldsymbol{\nu}$ имеет нулевое математическое ожидание $\mathsf{E}\boldsymbol{\nu} = \mathbf{0}$ и корреляционную матрицу $\sigma^2 I$.

Заметим, что сформулированная модель является нелинейной, т. к. в (2) неизвестными является как матрица $\Phi = \Phi(\boldsymbol{\tau})$, так и вектор $\boldsymbol{c}^{(k)}$.

При заданном числе слагаемых k задачу декомпозиции рассмотрим как задачу поиска наиболее точных оценок значений коэффициентов $\boldsymbol{c}^{(k)}$ и максимально надежных параметров $\boldsymbol{\tau}$. При фиксированных значениях параметров k и $\boldsymbol{\tau}$ задачу оценивания коэффициентов $\boldsymbol{c}^{(k)}$ поставим как задачу наилучшего линейного оценивания:

$$\mathsf{E}\big\|R\boldsymbol{\xi} - \hat{\boldsymbol{c}}^{(k)}\big\|^2 = \inf_{\boldsymbol{c}^{(k)} \in \mathbb{R}^k} \Big\{ \mathsf{E}\big\|R\boldsymbol{\xi} - \boldsymbol{c}^{(k)}\big\|^2 \;\Big|\; R\colon \mathbb{R}^n \to \mathbb{R}^k \Big\}.$$

При этом надежность коэффициентов в соответствии с [1, 2] будем определять величиной

$$\alpha_k(\boldsymbol{\xi}) = \inf\Big\{ \big\|\boldsymbol{\xi} - \Phi(\boldsymbol{\tau})\boldsymbol{c}^{(k)}\big\|^2 \;\Big|\; \boldsymbol{c}^{(k)}, \boldsymbol{\tau} \in \mathbb{R}^k \Big\}. \quad (3)$$

Такой метод оценки носит название метода максимальной надежности [1]. Погрешность оценок метода максимальной надежности вычислены в [5].

Сформулируем принцип максимальной надежности для оценки числа слагаемых k. Как показано в [6], надежность $\alpha_k(\boldsymbol{\xi})$ определяет распределение возможности на множестве $k = 1, 2, \ldots$: вариант этого распределения можно записать как функцию $\mu_0(\alpha_k(\boldsymbol{\xi}))$, где $\mu_0(\cdot)$ — монотонно невозрастающая функция, определенная на множестве $[0, \infty)$ и такая, что $\mu(0) = 1$, $\mu(+\infty) = 0$. С ростом k, очевидно, функционал $\alpha_k(\boldsymbol{\xi})$ не убывает, а значит, надежность растет. Это означает, что в задаче (3) вектор $\Phi\boldsymbol{c}^{(k)}$ все точнее приближает вектор $\boldsymbol{\xi}$. Если эта точность превысит точность измерения вектора $\boldsymbol{\xi}$, то аппроксимация $\Phi\boldsymbol{c}^{(k)}$ будет приближать не только «полезный сигнал» $\Phi\boldsymbol{c}^{(k)}$, но и погрешность (шум) $\boldsymbol{\nu}$. Принцип оценки значения k состоит в том, чтобы выбрать такое k, надежность которого достаточно высока, т. е. отличие $\Phi\boldsymbol{c}^{(k)}$ от $\boldsymbol{\xi}$ можно объяснить шумом $\boldsymbol{\nu}$.

*Работа поддержана грантом РФФИ № 08-07-00120.

Задача декомпозиции при фиксированном k

Запишем схему измерений (2) в виде

$$\xi_i = \int_0^T \varphi(t_i - t)c(t) + \nu(t_i), \quad i = 1, \ldots, n, \qquad (4)$$

где сигнал $c(t)$, $t \in [0, T]$, представляет собой линейную комбинацию дельта-функций $\delta(t - t_{0j})$, $j = 1, \ldots, k$. Переходя к дискретной схеме измерений, зададим набор значений $0 = t_{01} < \ldots < t_{0N} = T$ и перепишем (4) в виде

$$\boldsymbol{\xi} = \Psi \boldsymbol{c}^{(N)} + \boldsymbol{\nu}, \qquad (5)$$

где $\boldsymbol{c}^{(N)} \in \mathbb{R}^N$ — вектор евклидова пространства $\mathbb{R}^N$, $N \geqslant k$, $\{\boldsymbol{\xi}, \Psi\boldsymbol{c}^{(N)}, \boldsymbol{\nu}\} \subset \mathbb{R}^n$, линейный оператор $\Psi\colon \mathbb{R}^N \to \mathbb{R}^n$ задан своей матрицей $\Psi_{ij} = \varphi(t_i - t_{0j})$, $i = 1, \ldots, n$, $j = 1, \ldots, N$, $N \geqslant k$. Для того, чтобы схема (5) была эквивалентна (2), положим, что вектор $\boldsymbol{c}^{(N)}$ содержится в множестве $C_{k,e}$, элементы которого имеют не более k отличных от нуля координат, при этом параметр e указывает номера нулевых координат следующим образом: $e = e_1 \ldots e_N$ есть двоичное число, такое, что если $e_j = 0$, то j-я координата c_j вектора $\boldsymbol{c}^{(N)} \in C_{k,e}$ равна нулю, если же $e_j = 1$, то $c_j^{(N)} \neq 0$. При этом условии ненулевые координаты вектора $\boldsymbol{c}^{(N)}$ задают значение коэффициентов при сигналах $\varphi(\cdot)$ в (1), а номер ненулевой координаты — положение элементарного сигнала $\varphi(\cdot)$ (значение «сдвига» элементарного сигнала по оси времени). Множества $C_{k,e}$ при каждом k и e замкнуты в $\mathbb{R}^N$.

Для оценки координат вектора $\boldsymbol{c}^{(N)} \in C_{k,e}$ воспользуемся теоремой о нелинейном уточнении оценки [1, 3], согласно которой если $\tilde{\boldsymbol{c}}$ — некоторый случайный вектор $\mathbb{R}^N$, оценивающий вектор $\boldsymbol{c}^{(N)}$, и известно, что $\boldsymbol{c}^{(N)}$ принадлежит некоторому замкнутому множеству $C \subset \mathbb{R}^N$, то оценка, равная вектору из C, ближайшему к $\tilde{\boldsymbol{c}}$, обладает среднеквадратичной погрешностью, не большей, чем $\tilde{\boldsymbol{c}}$. Если множество невыпукло, то в качестве уточнения оценки $\tilde{\boldsymbol{c}}$ можно взять любой из ближайших к $\tilde{\boldsymbol{c}}$ векторов множества C.

Зафиксируем k и e и построим наилучшую линейную оценку $\hat{\tilde{\boldsymbol{c}}}$ вектора $\boldsymbol{c}^{(N)}$, считая, что $\boldsymbol{c}^{(N)}$ — любой вектор из $\mathbb{R}^N$. Эта оценка является решением задачи

$$\mathsf{E}\big\|R_0\boldsymbol{\xi} - \hat{\tilde{\boldsymbol{c}}}\big\|^2 = \inf_{\boldsymbol{c}^{(N)} \in \mathbb{R}^n} \Big\{ \mathsf{E}\big\|R\boldsymbol{\xi} - \boldsymbol{c}^{(N)}\big\|^2 \;\Big|\; R\colon \mathbb{R}^n \to \mathbb{R}^N \Big\}, \qquad (6)$$

и, как показано в [1, 2], равна $\hat{\tilde{\boldsymbol{c}}} = \Psi^-\boldsymbol{\xi}$, где $\Psi^-\colon \mathbb{R}^n \to \mathbb{R}^N$ — линейный оператор, псевдообратный $\Psi\colon \mathbb{R}^N \to \mathbb{R}^n$. Погрешность этой оценки равна $\mathsf{E}\big\|R_0\boldsymbol{\xi} - \hat{\tilde{\boldsymbol{c}}}\big\|^2 = \sigma^2 \operatorname{tr} R_0 R_0^*$. Здесь $\operatorname{tr} Q$ — след оператора Q.

Нелинейное уточнение $\hat{\boldsymbol{c}}^{(N)}$ оценки $R_0\boldsymbol{\xi}$ при фиксированном e получим, приравняв нулю значения координат вектора оценки $R_0\boldsymbol{\xi}$, соответствующих нулевым разрядам двоичного параметра e. Остальным координатам вектора $\hat{\boldsymbol{c}}^{(N)}$ припишем значения, равные значениям координат вектора $R_0\boldsymbol{\xi}$.

Выбор параметра e при фиксированном k осуществим, указав множество $C_{k,e}$, ближайшее в $\mathbb{R}^N$ к $R_0\boldsymbol{\xi}$. Для этого упорядочим координаты векторов $R_0\boldsymbol{\xi}$ и $\hat{\boldsymbol{c}}^{(N)}$ по невозрастанию абсолютных значений координат вектора $R_0\boldsymbol{\xi}$ и присвоим первым k координатам вектора $\hat{\boldsymbol{c}}^{(N)}$ значения, равные значениям соответствующих координат вектора $R_0\boldsymbol{\xi}$. Значения оставшихся $N - k$ координат также положим равными нулю. Если выбор ненулевых координат неоднозначен, в качестве оценки можно выбрать любой из таких векторов. Полученный вектор $\hat{\boldsymbol{c}}^{(N)}$ будем считать оценкой вектора $\boldsymbol{c}^{(N)}$ при фиксированном k.

Выбор параметра k

Для оценивания k из принципа максимальной надежности будем использовать статистику

$$\alpha_k(\boldsymbol{\xi}) = \big\|\boldsymbol{\xi} - \Psi\hat{\boldsymbol{c}}^{(N)}\big\|^2.$$

Вычисляя $\alpha_k(\boldsymbol{\xi})$ для $k = 1, 2, \ldots$, добьемся такого ее значения, которое правдоподобно описывается шумом.

Решение задачи декомпозиции при слабо обусловленной матрице Ψ

Оценка $R_0\boldsymbol{\xi}$ является несмещенной оценкой вектора $\boldsymbol{c}^{(N)}$ по измерению (5). Она сопровождается погрешностью $\sigma^2 \operatorname{tr} R_0 R_0^*$, которая в случае слабой обусловленности матрицы Ψ может оказаться неприемлемо большой, такой, что шум $R_0\boldsymbol{\nu}$ будет больше «полезного сигнала» $R_0\Psi\boldsymbol{c}^{(N)}$. Такая оценка окажется малоэффективной для уточнения координат вектора $\boldsymbol{c}^{(N)}$.

Уменьшить погрешность оценки вектора $\boldsymbol{c}^{(N)}$ можно, отказавшись от условия несмещенности. Для этого будем использовать два подхода.

В первом подходе линейную оценку $R\boldsymbol{\xi}$ вектора $\boldsymbol{c}^{(N)}$ будем искать, решая задачу на минимум [1, 2]

$$\inf\Big\{ \big\|R\Psi - I\big\|_2^2 \;\Big|\; \mathsf{E}\big\|R\boldsymbol{\nu}\big\|^2 \leqslant \varepsilon \Big\}.$$

Если $R \in \mathbb{R}_n \to \mathbb{R}^N$ — ее решение, то вектор $R\boldsymbol{\xi}$ интерпретируется как выходной сигнал прибора $R\Psi$, ближайшего к идеальному I при заданном уровне шума на выходе прибора $R\Psi$, если на вход его подан сигнал $\boldsymbol{c}^{(N)}$.

В результате такого подхода существенно подавляется шум, однако вместо набора импульсов, поступающего на вход Ψ, на выходе прибора $R\Psi$

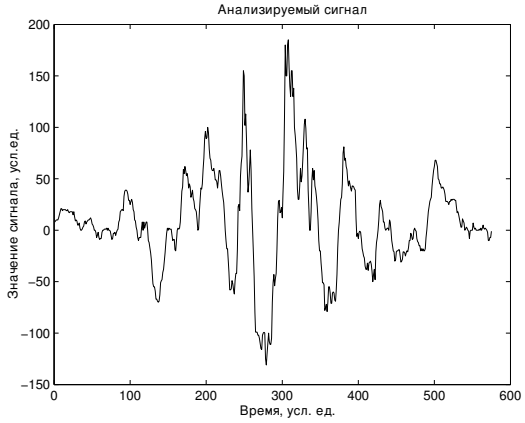


Рис. 1. График анализируемого экспериментально полученного сигнала.

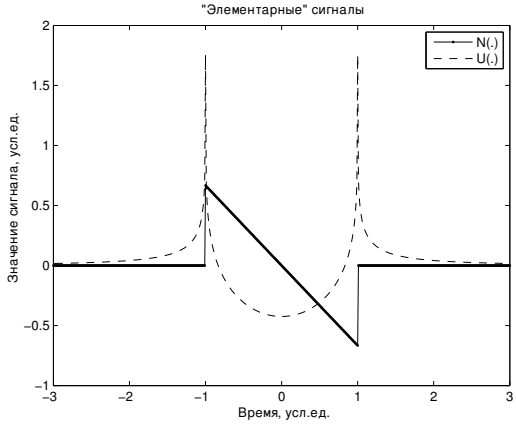


Рис. 2. Зависимость «элементарных» сигналов $N(\cdot)$ и $U(\cdot)$ от времени.

появятся «размытые» импульсы. Иными словами, разрешающая способность прибора $R\Psi$ оказывается конечной, однако ее можно оценить, построив графики отклика прибора $R\Psi$ на входные δ-образные сигналы.

Другой подход состоит в отказе от оценивания всего вектора $\boldsymbol{c}^{(N)}$, и оценивании лишь той его проекции, которая в наименьшей степени поражена шумом [1]. В этом случае оценка проекции дается линейным оператором R_s, и его действие на вектор $\boldsymbol{\xi}$ можно интерпретировать как выходной сигнал прибора $R_s\Psi$, измеряющему проекцию $\boldsymbol{c}^{(N)}$ на линейное подпространство $\mathbb{R}^N$, в наименьшей степени пораженное шумом.

Наконец, можно построить стохастическую модель входного сигнала $\boldsymbol{c}^{(N)}$, и для его оценивания из измерения (5) воспользоваться методом наилучшей в среднем квадратичном линейной аппроксимации случайного вектора $\boldsymbol{c}^{(N)}$ по наблюдаемому вектору $\boldsymbol{\xi}$.

Вычислительный эксперимент

В вычислительном эксперименте проводилась декомпозиция синтезированного сигнала и экспериментально полученного сигнала (рис. 1) в соответствие со следующей моделью регистрации:

$$\xi(t) = \sum_{j=1}^{k} a_j^{(k)} N(t - t_{0j}) + b_j^{(k)} U(t - t_{0j}) + \nu(t). \quad (7)$$

Характерный вид взаимно ортогональных «элементарных» сигналов $N(\cdot)$ и $U(\cdot)$ приведен на рис. 2.

Синтезированный сигнал представлял собой сумму четырех элементарных сигналов $U(x-2) + N(x-5) - (U(x-8) + N(x-8))$, искаженных нормально распределенным шумом с нулевым математическим ожиданием и стандартным отклонением 0,1 (рис. 3).

Решим задачу линейного оценивания (6) для некоторого $k = \tilde{k}$ с использованием подхода,

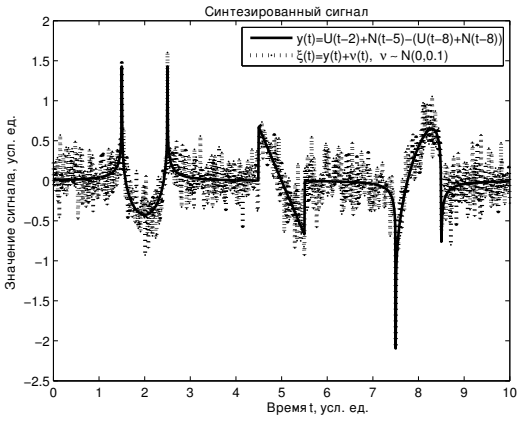


Рис. 3. График анализируемого в вычислительном эксперименте синтезированного сигнала.

при котором оценивается только наименее пораженная шумом проекция вектора $\boldsymbol{c}^{(N)}$. Известно, что оператор ортогонального проецирования на линейную оболочку m сингулярных векторов (Π_m) оператора Ψ, соответствующих его m максимальным собственным значениям, является проектором на подпространство, наименее пораженное шумом, из всех подпространств размерности m пространства $\mathbb{R}^n$. Данное подпространство является пространством максимальной размерности, таким, что погрешность оценивания вектора $\tilde{\boldsymbol{c}}$ вектором $\Pi_m\Psi^-\boldsymbol{\xi}$, например, не превышает заданного порога. Для построения этой проекции воспользуемся методом SVD-разложения: если

$$\Psi = USV^+,\ S = \operatorname{diag}(s_1, \ldots, s_n),\ s_1 \geqslant \ldots \geqslant s_n,$$

то, отбросив меньшие сингулярные значения, получим

$$\Psi_m = US_mV^+,\ S = \operatorname{diag}(s_1, \ldots, s_m, \underbrace{0, \ldots, 0}_{n-m}),$$

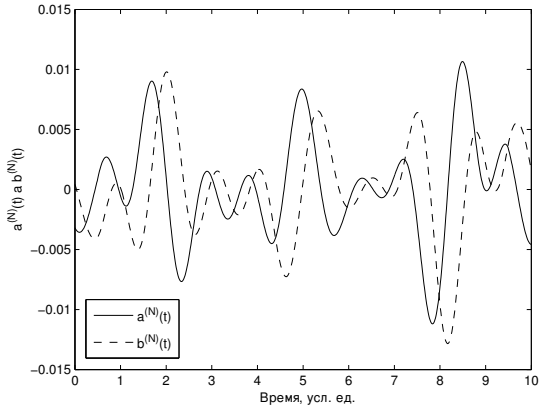


Рис. 4. Линейная оценка вектора коэффициентов $\tilde{\boldsymbol{c}} = (\tilde{\boldsymbol{a}}, \tilde{\boldsymbol{b}})$ для синтезированного сигнала, вычисленная с помощью SVD-разложения при $m = 17$.

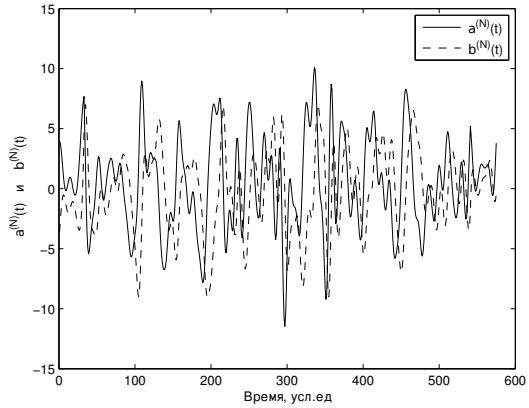


Рис. 5. Линейная оценка вектора коэффициентов $\tilde{\boldsymbol{c}} = (\tilde{\boldsymbol{a}}, \tilde{\boldsymbol{b}})$ для экспериментального сигнала, вычисленная с помощью SVD-разложения при $m = 75$.

и $\tilde{\boldsymbol{c}} = \Pi_m \Psi^- \boldsymbol{\xi} = \Psi_m^- \Psi_m \Psi^- \boldsymbol{\xi} =$

$$= V S_m^- U^+ U S_m V^+ V S^{-1} U^+ \boldsymbol{\xi} =$$

$$= V S_m^- S_m S^{-1} U^+ \boldsymbol{\xi} = V S_m^- U^+ \boldsymbol{\xi},$$

где $S^- = \operatorname{diag}(s_1^{-1}, \ldots, s_m^{-1}, \underbrace{0, \ldots, 0}_{n-m})$.

Линейные оценки для синтезированного и экспериментального сигналов приведены, соответственно, на рис. 4 ($k = 7$) и 5 ($k = 35$).

Далее построим оценки $\hat{\boldsymbol{c}}^{(N)} = (\hat{\boldsymbol{a}}^{(N)}, \hat{\boldsymbol{b}}^{(N)})$, см. рис. 6, 7, 8, 9; $\tilde{k} = 35$.

Если оказывается, что значение статистики $\alpha_{\tilde{k}}(\boldsymbol{\xi})$ правдоподобно описывается шумом (менее заданного порога), следовательно, выбранное k является искомой оценкой размерности исходной линейной комбинации (7), t_{0j}, $j : e_j \neq 0$ являются оценкой нелинейного параметра максимальной надежности, $\hat{\boldsymbol{c}}^{(N)}$ — наиболее точной оценкой векторов коэффициентов. Если же оказывается, что значение статистики $\alpha_{\tilde{k}}(\boldsymbol{\xi})$ более заданного порога (45 для синтезированного сигнала и $2 \cdot 10^5$ для экспериментального), то $\tilde{k}$ увеличивается и эксперимент повторяется.

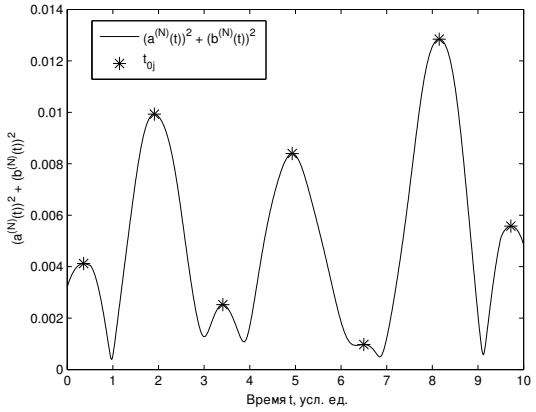


Рис. 6. Принцип выбора t_{0j}. Синтезированный сигнал.

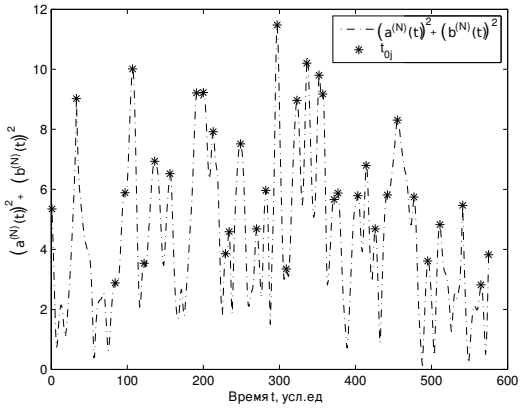


Рис. 7. Принцип выбора t_{0j}. Экспериментальный сигнал.

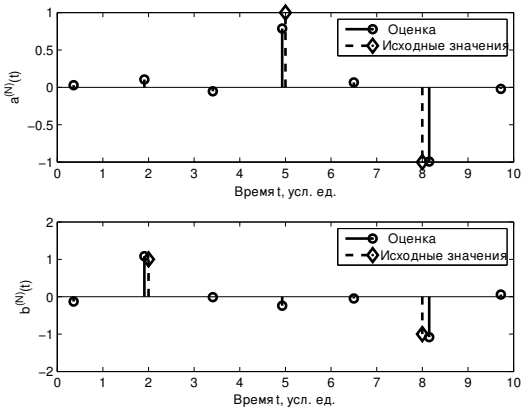


Рис. 8. Нелинейное уточнение вектора коэффициентов $\hat{\boldsymbol{c}}^{(N)} = (\hat{\boldsymbol{a}}^{(N)}, \hat{\boldsymbol{b}}^{(N)})$. Синтезированный сигнал.

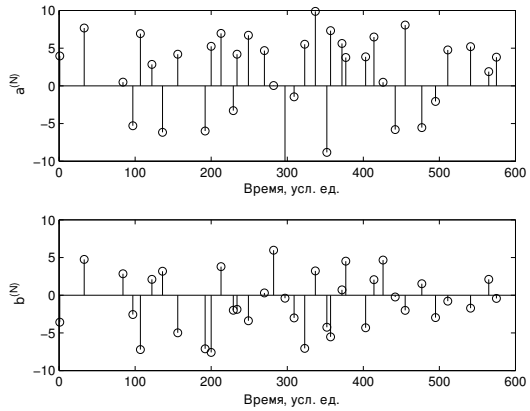


Рис. 9. Нелинейное уточнение вектора коэффициентов $\hat{\boldsymbol{c}}^{(N)} = (\hat{\boldsymbol{a}}^{(N)}, \hat{\boldsymbol{b}}^{(N)})$. Экспериментальный сигнал.

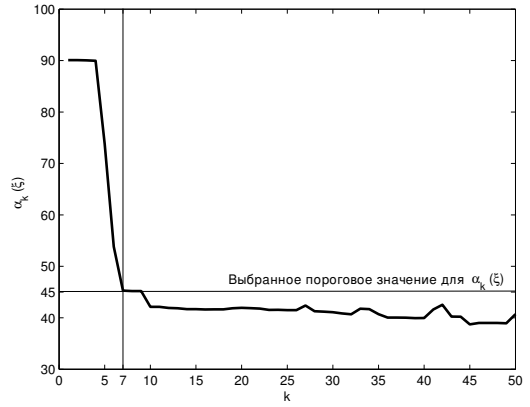


Рис. 10. Зависимость статистики $\alpha_{\tilde{k}}(\boldsymbol{\xi})$ от размерности линейной комбинации $\tilde{k}$. Синтезированный сигнал.

Как видно из рис. 8, результат оценивания величины коэффициентов $\boldsymbol{c}^{(k)}$ и временных задержек $\boldsymbol{\tau}$ близки к исходным значениям, использованным при синтезе сигналов. Несовпадение размерности линейной комбинации в результате оценивания связано с выбором порога на статистику $\alpha_{\tilde{k}}(\boldsymbol{\xi})$. При анализе синтезированного сигнала порог был выбран, равным 45 исходя из того, что он отделяет области, где эта статистика принимает существенно различающиеся значения. Резкое уменьшение величины статистики $\alpha_{\tilde{k}}(\boldsymbol{\xi})$ в окрестности аргумента $k = 7$ позволяет сделать вывод о том, что в случае выбора величины порога 45, а размерности $k = 7$, «полезная составляющая» анализируемого сигнала оказывается полностью учтена, а влияние шума еще не становится существенным.

Немонотонность зависимости $\alpha_{\tilde{k}}(\boldsymbol{\xi})$ от $\tilde{k}$, приведенная на рис. 10, 11, объясняется ошибками, возникающими при приближенном решении задачи нелинейной оптимизации.

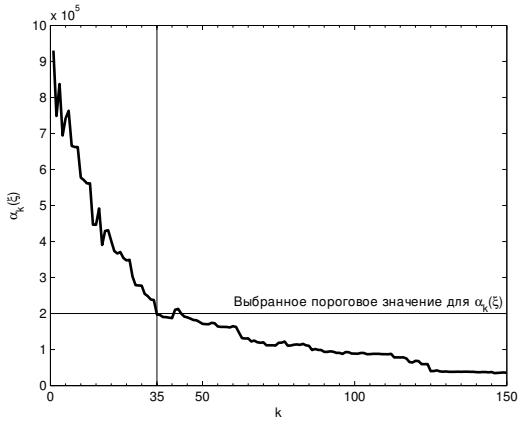


Рис. 11. Зависимость статистики $\alpha_{\tilde{k}}(\boldsymbol{\xi})$ от размерности линейной комбинации $\tilde{k}$. Экспериментальный сигнал.

Выводы

Предложенный в работе метод декомпозиции сигналов, представимых в виде линейной комбинации шума с известными корреляционными свойствами и сигналов заданной формы с неизвестными амплитудами и величинами временных сдвигов, позволяет оценить число слагаемых, их амплитуду и положение на оси абсцисс.

Эффективность метода продемонстрирована в вычислительном эксперименте по акустическому зондированию атмосферы для декомпозиции синтезированного и реального сигналов.

Литература

[1] *Пытьев Ю. П.* Методы математического моделирования измерительно-вычислительных систем. — М.: Физматлит, 2002.

[2] *Чуличков А. И.* Основы теории измерительно-вычислительных систем. Стохастические линейные измерительно-вычислительные системы — Тамбов: Издательство Тамбовского государственного технического университета, 2000. — 140 с.

[3] *Пытьев Ю. П.* К теории нелинейных измерительно-вычислительных систем // Математическое моделирование — 1992. — Т. 4, № 2.

[4] *Пытьев Ю. П.* Математические методы интерпретации эксперимента — М.: Высшая школа, 1989.

[5] *Пытьев Ю. П., Сухорукова Г. В., Чуличков А. И.* Задачи дистанционного зондирования: математическое моделирование, анализ и интерпретация результатов // Математическое моделирование, 1994. — Т. 6, № 11.

[6] *Пытьев Ю. П.* Возможность как альтернатива вероятности. Математические и эмпирические основы, применения. — М.: Физматлит, 2007.

Морфологический подход к вейвлет-анализу сигналов*

Чуличков А. И., Демин Д. С., Цыбульская Н. Д.
achulichkov@gmail.com
Москва, Физический факультет МГУ

Предлагается подход к анализу особенностей формы сигналов, использующий локальность и масштабируемость вейвлет-анализа и инвариантность к нелинейным искажениям сигнала, связанным с (неизвестными) особенностями условий их формирования. Приводятся примеры использования морфологических методов анализа для решения задач классификации сигналов и оценки параметров их формы.

Вейвлет-анализ является высокоэффективным способом решения целого ряда задач, связанных с выделением амплитудных и частотных особенностей сигналов.

С другой стороны, хорошим инструментом решения задач узнавания, классификации объектов, а также оценивания параметров объекта по регистрации поступающих от него сигналов (изображений) являются методы морфологического анализа [1, 2, 3]. Морфологические методы предназначены для анализа изображений, полученных при неконтролируемых и неизвестных условиях их формирования. Для этого рассматривается класс преобразований изображения, моделирующий вариации условий их формирования, и максимальный инвариант этих преобразований называется формой изображения. Конструктивно форму можно построить, указав метод сравнения изображений по форме: считается, что форма g не сложнее, чем форма f, если найдется преобразование F из заданного класса $\mathbf{F}$, такое, что $F*f=g$. Изображение g не сложнее по форме, чем f, тогда и только тогда, когда найдутся такие условия формирования изображения объекта (сцены), изображенного на f, при которых его изображением будет g. На практике класс $\mathbf{F}$ выбирают так, чтобы множество V_f всех изображений, форма которых не сложнее f, было выпукло и замкнуто в линейном метрическом пространстве всех изображений, тогда искомым инвариантом является проектор на это множество. В терминах проектора решаются многие задачи анализа формы сигнала. Формой сигнала в морфологическом анализе называют как множество V_f, так и проектор на V_f.

Данная работа предлагает совместное применение описанных выше подходов, а также содержит описание применения полученных методов к решению двух задач: оценивания параметров формы сигнала и классификации сигналов. При этом для решения первой задачи применяется совмещение двух подходов. Для второй — их последовательное применение.

Принципы построения морфологических вейвлетов

В настоящей работе сформулирован подход, объединяющий достоинства вейвлет- и морфологических методов анализа сигналов (изображений).

Пусть сигнал (изображение) рассматривается как элемент $L^2(X)$, где область задания сигнала (поле зрения изображения) X является подмножеством евклидова пространства R^n размерности n. Рассмотрим множество преобразований масштаба и сдвига носителя $H\in X$ базового вейвлета $\psi(\cdot)$, таких, что результат $H_{x_0,\lambda}$ любого из этих преобразований целиком содержится в X. Пусть, кроме того, задан класс преобразований $\mathbf{F}$, моделирующий изменение условий формирования сигнала. Тогда множество

$$V_{\psi,\lambda,x_0}=\{f(\cdot)\,|\,f(x)=F*\psi(\lambda(x-x_0)),F\in\mathbf{F}\},\quad V_{\psi,\lambda,x_0}\subset L^2(H_{x_0,\lambda})$$

есть множество фрагментов сигнала (изображения), форма которых не сложнее, чем форма базового вейвлета.

Обозначим P_{ψ,λ,x_0} оператор проецирования на множество V_{ψ,λ,x_0} в пространстве всех изображений. Тогда мерой близости фрагмента предъявленного сигнала $\xi\in L^2(X)$ на подмножестве $H_{x_0,\lambda}\subset X$, полученном из H преобразованием масштаба λ и сдвига на $x_0\in R^n$, по форме к вейвлету $\psi(\lambda\cdot-x_0)$, определится значением дроби

$$\Delta_\psi(\lambda,x_0)=\frac{\|P_0\xi-\xi\|^2}{\|\xi-P_{\psi,\lambda,x_0}\xi\|^2},\qquad(1)$$

где P_0 — проектор на множество сигналов, равных константе почти всюду на H_{x_0}, [1, 2, 3]. Чем больше значение $\Delta_\psi(\lambda,x_0)$, тем ближе фрагмент рассматриваемого сигнала ξ по форме к вейвлету $\psi(\lambda\cdot-x_0)$. Вычисляя $\Delta_\psi(\lambda,x_0)$ для всех λ и x_0, при которых подмножество $H_{x_0,\lambda}$ целиком содержится в X, найдем те значения, которые соответствуют максимальному сходству по форме выделенного фрагмента сигнала с заданным вейвлетом.

Морфологические вейвлеты были успешно применены для решения задач выделения, классификации сигналов, содержащих информацию об определенном классе физических явлений, и задач оценивания их параметров.

*Работа выполнена при финансовой поддержке РФФИ, проект № 08-07-00120.

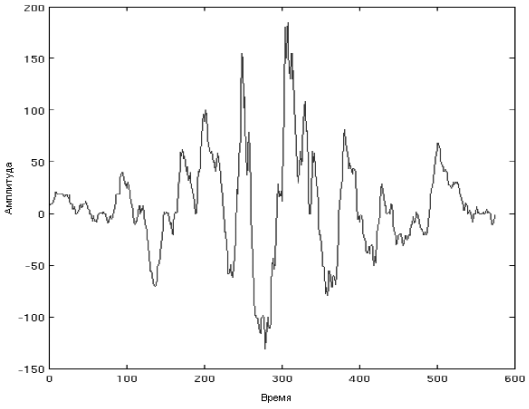


Рис. 1. Результат регистрации сигнала, прошедшего через канал связи и искаженный шумом.

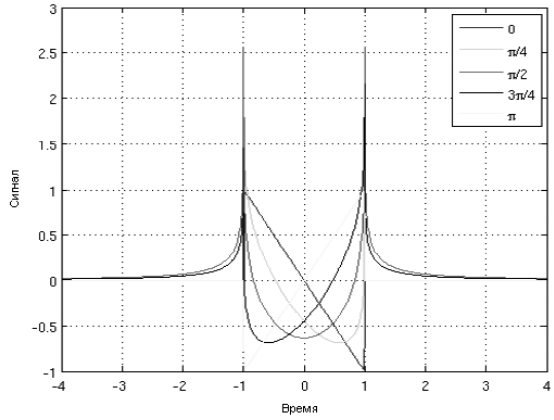


Рис. 2. Результат преобразования $F(1,0,\varphi)$ N-образного импульса для различных φ.

Вейвлет-анализ при оценивании параметров формы сигнала

Пусть исходный сигнал имеет вид N-образного импульса:

$$g(t,\tau_0) = cf(t/\tau_0), f(t) = \begin{cases} -t, & |t| < 1, \\ 0, & \text{иначе}\ . \end{cases}$$

Значение параметра τ_0, равное половине длины N-образного импульса, неизвестно. Сигнал $g(\cdot,\tau_0)$ будем рассматривать на конечном носителе H.

Распространяясь по каналу связи, исходный импульс претерпевает искажения: на выходе канала он имеет вид $q((t-t_0)/\tau_0)$, $t \in [0,T]$. Процесс регистрации сопровождается аддитивным гауссовым некоррелированным шумом $\nu(t)$ с нулевым математическим ожиданием. Результат регистрации — сигнал $q((t-t_0)/\tau_0)+\nu(t)$, $t \in [0,T]$ — изображен на рис. 1. Требуется оценить сдвиг t_0 импульса по оси времени и параметр масштаба τ_0.

Искажения сигнала при распространении по каналу связи опишем следующим образом: будем считать, что $q(t) = F(a,b,\varphi) * g(t)$, $t \in H$, где функция $F(a,b,\varphi)$ преобразует любой сигнал вида $f(t) = \sum\limits_{k=1}^{\infty} C_k \sin(k\omega_0 t)$ в сигнал

$$F(a,b,\varphi) * f(t) = a\sum_{k=1}^{\infty} C_k \sin(k\omega_0 t + \varphi) + b,$$

где a, b, φ являются неизвестными параметрами. На рис. 2 приведен график сигнала, полученного преобразованием $F(1,0,\varphi)$ из N-образного импульса при различных значениях φ.

На рис. 3 приведен график функции двух переменных $F(1,0,\varphi) * g(t)$.

Множеством фрагментов сигнала, форма которых не сложнее, чем форма, задаваемая базовым

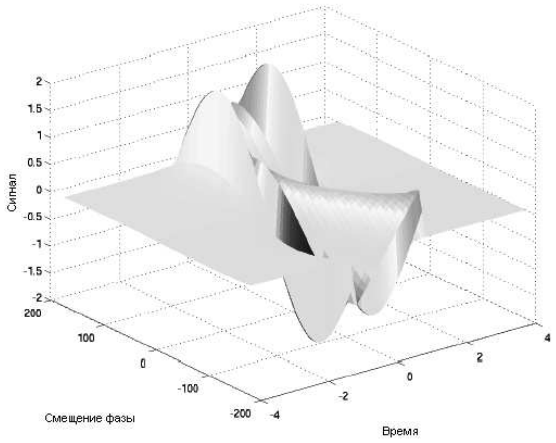


Рис. 3. График $F(1,0,\varphi) * g(t)$ как функции времени и фазы.

вейвлетом, сжатым в τ^{-1} раз и сдвинутым на t_0, в этом случае является

$$V_{g,\tau^{-1},t_0} = \Big\{F(a,b,\varphi) * g\Big(\frac{t-t_0}{\tau_0}\Big)\Big\},$$

где константы a ,b, φ определяют неизвестные условия регистрации сигнала, а константы t_0 и τ^{-1} — сдвиг и масштаб базового вейвлета соответственно.

Схема регистрации сигнала принимает следующий вид:

$$\xi(t) = q((t-t_0)/\tau) + \nu(t),\ q(t) \in V_f,$$

где сдвиг t_0 и масштабный множитель τ^{-1} неизвестны.

Сжимая множество H в τ^{-1} раз и сдвигая его по интервалу $[0,T]$ так, чтобы результат этого преобразования целиком содержался в $[0,T]$, определим фрагмент сигнала ξ на полученном подмножестве, и вычислим проекцию этого фрагмента

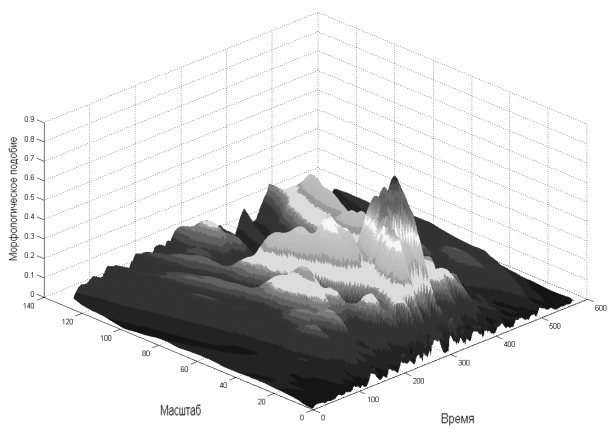


Рис. 4. Сходство по форме фрагмента сигнала с заданным вейвлетом как функция сдвига и масштаба.

на V_{g,τ^{-1},t_0}. Заметим, что в данном случае этот проектор является линейным оператором и легко вычисляется.

Локальное сходство выбранного фрагмента по форме с заданным вейвлетом определялось по формуле (1). График зависимости $\Delta_g(\tau^{-1}, t_0)$ от положения t_0 подвижного фрагмента и масштабного множителя τ представлен на рис. 4. Максимальное значение морфологического сходства сигнала и элементарных функций позволяет оценить искомые параметры.

Классификация сигналов по форме их вейвлет-спектра

Другим способом объединения морфологического и вейвлет- анализа является последовательное вычисление вейвлет-спектра сигнала и его исследование морфологическими методами.

Этот подход был применен при решении задачи классификации акустических сигналов, возникающих в результате разрядов в изоляции высоковольтного оборудования. Считалось, что все акустические сигналы можно разделить на четыре класса. Первые три из них соответствовали различным видам электрических разрядов, а четвертый — шумам, не связанным с разрядами. Цель работы состояла в отделении сигналов первых трех классов, связанных с предаварийными ситуациями, от посторонних виброзвуков (сигналов четвертого класса), не связанных с опасностью аварии.

Для решения поставленной задачи сначала вычислялся вейвлет-спектр сигнала (рис. 5).

Далее на основании анализа полученных изображений определялись формы изображений вейвлет-спектров сигналов трех классов, соответствующих разрядам:

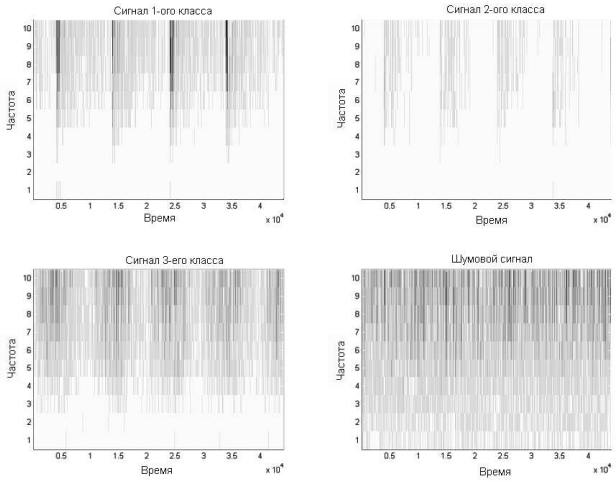


Рис. 5. Изображения вейвлет-спектра сигналов четырех классов.

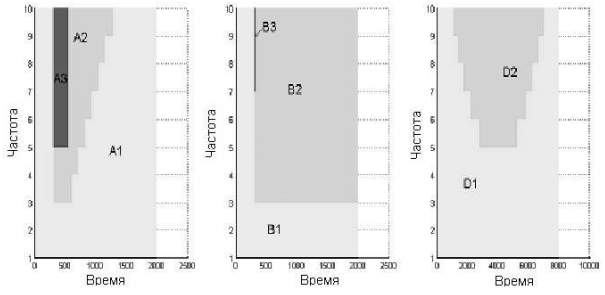


Рис. 6. Разбиения поля зрения, задающие формы изображений вейвлет-спектров сигналов из трех классов.

$$V_1 = \left\{ f^{(1)}(\mathbf{x}) = \sum_{j=1}^{3} c_j \chi_j(\mathbf{x}),\ c_1 \leqslant c_2 \leqslant c_3 \right\};$$

$$V_2 = \left\{ f^{(2)}(\mathbf{x}) = \sum_{j=1}^{3} b_j \psi_j(\mathbf{x}),\ b_1 \leqslant b_2 \leqslant b_3 \right\};$$

$$V_3 = \left\{ f^{(3)}(\mathbf{x}) = \sum_{j=1}^{2} a_j \varphi_j(\mathbf{x}),\ a_1 \leqslant a_2 \right\}.$$

В данном случае формы фрагментов изображений вейвлет-спекторв определяются разбиением соответствующего подмножества поля зрения на области одинаковой яркости. Эти множества для каждого из трех типов сигналов представлены на рис. 6.

Проекции предъявленного фрагмента изображения g вейвлет-спектра сигнала на форму V_i определялись как решения задач наилучшего приближения:

$$\|g - P_i g\| = \min\{\|g - z\| \colon z \in V_i\}, \quad i = 1, 2, 3.$$

Так как шумовое изображение не обладает регулярной структурой, считалось, что оно являет-

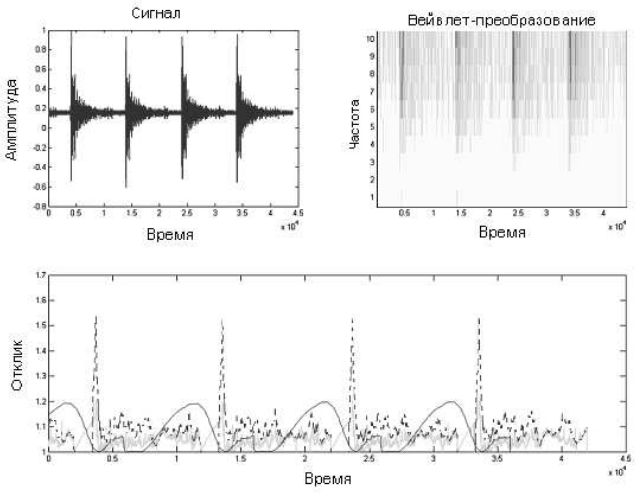


Рис. 7. Результат работы классификатора.

ся изображением ровного поля зрения, искаженным белым шумом. Мерой отличия изображения от изображения однородного поля зрения является величина

$$\|g - P_0 g\| = \min\{\|g - z\| \colon z = \text{const}\},\ \ i = 1, 2, 3.$$

Дробь

$$\Delta_i(g) = \frac{\|g - P_0 g\|}{\|g - P_i g\|}$$

принимает тем большие значения, чем ближе форма предъявленного изображения к форме изображений из множества V_i по сравнению с формой шумового изображения. Если значение $\Delta_i(g)$ хотя бы для одного из классов с номерами $i = 1, 2, 3$ превышает заданный порог, сигнал относится к тому классу, для которого $\Delta_i(g)$ максимально, $i = 1, 2, 3$. В противном случае принимается решение о шумовом характере предъявленного сигнала.

На рис. 7 приведен результат распознавания сигнала: исходный сигнал (вверху слева), изображение его вейвлет-спектра (вверху справа) и отклик классификатора (значение дроби $\Delta_i(g)$ для каждого из трех классов $i = 1, 2, 3$) как функция сдвига фрагмента по оси абсцисс (времени). Предъявленный сигнал соответствует классу, выделенному пунктирной линией.

Выводы

Методы, объединяющие вейвлет- и морфологические методы анализа сигналов, позволили решить задачи автоматической классификации сигнала и оценки параметров формы сигнала.

При решении задачи оценивания параметров сигнала построен морфологический вейвлет, позволяющий найти в анализируемом сигнале интересующий исследователя участок и оценить его характеристики.

Построенный классификатор позволяет отличить как сигнал, являющийся предвестником аварии, от посторонних виброзвуков, не связанных с опасностью аварии, так и конкретные предаварийные сигналы между собой.

Литература

[1] *Pyt'ev Yu. P.* Morphological Image Analysis. // Pattern Recognition and Image Analysis. — 1993. — V. 3, № 1. — P. 19–28.

[2] *Pyt'ev Yu. P.* The Morphology of Color (Multispectral) Images. // Pattern Recognition and Image Analysis. — 1997. — V. 7, № 4. — P. 467–473.

[3] *Pyt'ev Yu. P.* Methods for Morphological Analysis of Color Images. // Pattern Recognition and Image Analysis. — 1998. — V. 8, № 4. — P. 517–531.

Представление результатов распознавания речи*

Чучупал В. Я.
chuchu@ccas.ru
Москва, ВЦ РАН

Передача нескольких возможных альтернативных вариантов позволяет существенно улучшить эффективность применения систем распознавания речи. В работе приведен эффективный в вычислительном смысле алгоритм формирования сетевой структуры для представления промежуточного результата автоматического распознавания речи. Заметного ухудшения вероятности правильного распознавания при этом не наблюдается, а объем требуемых вычислений дает возможность использовать предложенные алгоритмы в системах реального времени.

Представление промежуточного результата распознавания речи

Корректность автоматического распознавания речи на практике довольно часто далека от желаемой. Поэтому для повышения вероятности правильного распознавания речи используется любая дополнительная информация, например, лексические или синтаксические ограничения, которые свойственны конкретному приложению.

Представление результата распознавания высказывания в виде одной, наиболее вероятной последовательности слов, так называемый 1-й лучший вариант, не является хорошим способом представления промежуточного результата распознавания, т. к. ограничивает возможность повторного анализа альтернатив и принятия лучшего решения с использованием дополнительных ограничений, предоставляемых приложением. Более предпочтительным оказывается вариант передачи результата распознавания в виде набора n-лучших (обычно $n \gg 10$) возможных последовательностей слов.

Возможность предоставления результата в виде списка n-лучших гипотез предусмотрена основными существующими стандартами на интерфейс с системами распознавания речи, такими как SAPI (Microsoft Speech Application Program Interface) или MRCP (Multimedia Resource Control Protocol).

Наиболее очевидный способ представления n лучших в виде вектора из n гипотез практически приемлем только для небольших значений $n \leqslant 10$. Он неэкономен с точки зрения использования памяти и вычислительных ресурсов процессора.

Поскольку, как правило, большинство гипотез в списке n-лучших отличаются друг от друга в нескольких словах, существует очевидная возможность для более компактного описания промежуточного результата распознавания высказывания в виде сетей, где ребра (например) соответствуют произнесению слов, а вершины — межсловным переходам. Такие структуры, графы слов (Word Graph [1]), решетки слов (Word Lattice [3]) или сети конфузных слов (Word Confusion Network [4]), обеспечивают сохранение точности при относительно небольших требованиях к памяти. Перебор вариантов на такой сетевой структуре также предполагает умеренный объем вычислений.

Представление словаря

При распознавании речевой сигнал моделируется как последовательность произнесения слов заданного словаря. Произнесение слова рассматривается как последовательность произнесений более мелких единиц, обычно фонов (контекстно-зависимых реализаций фонем), для которых строятся акустические модели. Последовательность фонов, которые соответствуют акустической реализации слова, определяется путем задания произносительной транскрипции слова. Набор произносительных транскрипций для всех слов словаря называется произносительным словарем.

Для словарей объемом свыше 1000 слов линейная организация произносительного словаря не является удовлетворительной, так как значительное число слов имеет одинаковые приставки, корни и т. п. Можно ожидать существенного сокращения требуемых ресурсов памяти и процессора, если уменьшить объем дублирования. Например, использовать древовидные структуры. В этом случае слова, имеющие одинаковые начальные фоны в произносительной транскрипции, имеют общий корень и ветвь до того момента, пока не встретятся разные фоны (префиксное дерево). Дуги такого дерева соответствуют фонам, каждая последовательность дуг от корня дерева до любого его листа соответствует некоторому слову словаря (точнее, его произносительной транскрипции). Такую структуру называют лексическая или произносительная сеть [6].

Поскольку высказывание — это последовательность из нескольких слов, при поиске произносительный лексикон делается циклическим — листья деревьев (концы слов) соединяются ребрами с вершинами.

Современные алгоритмы распознавания речи можно рассматривать как поиск на деревьях. Стоимость перехода по ребру дерева определяется правдоподобием соответствующего ребру слова слова-

*Работа выполнена при финансовой поддержке РФФИ, проект № 07-01-00657а.

ря. Для оценки этого правдоподобия используется процедура Витерби, которая также определяет и оптимальный путь, то есть наиболее вероятную последовательность слов. По определению алгоритма в каждой вершине хранится величина правдоподобия для наиболее вероятного пути в эту вершину и указатель на предшествующую вершину, с помощью которого можно затем восстановить оптимальный путь, то есть получить 1-лучшую.

Алгоритм гарантированного определения N лучших гипотез получается, если в каждой вершине дерева запоминать правдоподобия N наиболее вероятных путей и хранить, соответственно, N обратных указателей. По окончании вычислений перебор по всем гипотезам для всех концов слов даст точный список N лучших гипотез. Вычислительная нагрузка при использовании такого подхода увеличивается минимум в N раз, что делает алгоритм при $N > 4$ малопригодным для применения в реальных системах. В следующей таблице приведен коэффициент реального времени $\varkappa$ (отношение времени распознавания к длительности высказывания) для процессора Intel Pentium 3.3 ГГц при $n = 1, \ldots, 4$. При уменьшении n в 4 раза время распознавания уменьшилось в 7 раз.

n	4	3	2	1
$\varkappa$	0,364	0,236	0,138	0,049

Идея субоптимального (не гарантирующего построение правильного списка N лучших) алгоритма Word Dependent n-Best была предложена в [2]. Можно существенно сократить объем вычислений ценой незначительного снижения вероятности правильного обнаружения N лучших гипотез. Для этого в вершинах графа нужно хранить минимальную информацию об L, $L \ll N$ наиболее вероятных путях. Информация о каждом пути включает только его правдоподобие и последнее предшествующее данному слово. Практическая применимость алгоритма обусловлена экспериментально тем, что граница между произнесениями двух слов обычно не зависит от других предшествующих слов и, следовательно, есть смысл принимать во внимание только «короткую» историю, или одно предшествующее слово. При этом увеличение ошибки (по сравнению с точным алгоритмом N лучших) составляет около 1% для словарей размером до 1000 слов при $L = 2$.

Использование приведенного выше субоптимального алгоритма приводит к необходимости построения алгоритмов генерации N лучших гипотез, выбора компактного сетевого способа записи N лучших гипотез и построении оптимального алгоритма перебора на полученной сети. Соответствующие алгоритмы приведены ниже. При этом в качестве представления промежуточного результата распознавания использован граф слов.

Алгоритм формирования графа слов

Входными данными являются результаты выполнения алгоритма поиска [2]:

- наиболее правдоподобная последовательность слов, которая определяет минимальную стоимость $h^\star$ и момент окончания T_{end};
- h_n — оценка стоимости n-лучшего пути;
- список гипотез о произнесенных словах L_{words}, который состоит из упорядоченных по времени завершения записей о возможных произнесениях слов.

Каждая запись о слове $w \in L_{\text{words}}$ включает следующую информацию:

- w.id — идентификатор слова;
- w.start — время начала слова;
- w.end — время конца слова;
- w.cost — стоимость наилучшего пути до слова включительно;
- w.pred — указатель на наиболее вероятное предшествующее слово.

В приведенном описании алгоритмов использованы следующие обозначения для структур данных: L_{open} — список открытых вершин графа слов, L_{close} — список закрытых вершин, L_{start} — множество начальных вершин, L_{end} — множество конечных вершин (то есть концов слов). Запись w.tail обозначает стоимость пути от вершины конца слова w до целевой вершины.

Выходом алгоритма является граф слов WG: список вершин и направленных дуг. Выделены списки начальных L_{start} и конечных L_{end} вершин. Каждый путь от начальной до конечной вершины соответствует некоторой гипотезе о произнесенной последовательности слов. Через V_w обозначим вершину, которая соответствует некоторому слову w, $\overrightarrow{V_{\bar{w}}V_w}$ — дуга от $V_{\bar{w}}$ к V_w. Через $h(V)$ обозначим стоимость попадания в вершину V. Создание вершины V_w, которая соответствует слову w, означает также присваивание этой вершине вышеперечисленных атрибутов слова.

Граф слов, вычисленный в соответствии с алгоритмом, не является оптимальным по памяти: одна и та же гипотеза о произнесении слова может иметь несколько вершин в графе, однако каждая вершина имеет только одну последующую, что облегчает последующий поиск. После окончания работы алгоритма для каждой вершины графа V_w вычислена функция стоимости оптимального пути к целевой вершине, который проходит через данную вершину $h(V_w)$.

Построенный граф дает возможность выполнить повторно поиск оптимального пути или оптимальной последовательности слов, используя в качестве стоимости показатель, отличный от величины правдоподобия данных для оптимального пути.

Алгоритм 1. Формирование графа слов.

$L_{\text{open}} := \{\varnothing\}$, выберем: $h_{\text{thr}} := h_n$;
для всех $w \in L_{\text{words}}$
  если $(w.\text{cost} > h_{\text{thr}}) \vee (w.\text{end} \approx T_{\text{end}})$ **то**
    создать V_w: $V_w \in WG$;
    $w.\text{tail} := 0$; $h(V_w) := w.\text{cost}$;
    $L_{\text{open}} := V_w$; $L_{\text{end}} := V_w$;
пока $L_{\text{open}} \neq \varnothing$
  выберем $V_w := \arg\min_{V \in L_{\text{open}}} h(V)$;
  $L_{\text{close}} := V_w$; $\acute{w} := w.\text{pred}$;
  если $\acute{w} := \varnothing$ **то**
    $L_{\text{start}} := V_{\acute{w}}$;
  иначе
    $\acute{w}.\text{tail} := w.\text{tail} + (w.\text{cost} - \acute{w}.\text{cost})$;
    если $\acute{w}.\text{cost} + \acute{w}.\text{tail} < h_{\text{thr}}$ **то**
      создать $V_{\acute{w}}$: $V_{\acute{w}} \in WG$;
      $h(V_{\acute{w}}) := \acute{w}.\text{cost} + \acute{w}.\text{tail}$;
      $L_{\text{open}} := V_{\acute{w}}$; $WG := \overrightarrow{V_{\acute{w}} V_w}$;
      для всех $\bar{w} \in L_{\text{words}}$
        если $\bar{w}.\text{end} \approx \acute{w}.\text{end}$ **то**
          если $\bar{w}.\text{cost} + \acute{w}.\text{tail} < h_{\text{thr}}$ **то**
            Создать $V_{\bar{w}}$: $V_{\bar{w}} \in WG$;
            $h(V_{\bar{w}}) := \bar{w}.\text{cost} + \acute{w}.\text{tail}$;
            Присвоить $V_{\bar{w}}$ атрибуты слова $\bar{w}$;
            $L_{\text{open}} := V_{\bar{w}}$; $WG := \overrightarrow{V_{\bar{w}} V_w}$;

Алгоритм 2. Поиск на графе слов.

$L_{\text{open}} := L_{\text{start}}$;
если $V_w \in L_{\text{open}}$ **то** $g(V_w) := g(w)$;
пока $L_{\text{open}} \neq \varnothing$
  $V_w := \arg\min_{V \in L_{\text{open}}} g(V)$; $L_{\text{close}} := V_w$;
  если $V_w \in L_{\text{end}}$ **то**
    $i := 1$; $V_i := V_u$;
    пока $V_i \neq L_{\text{start}}$
      $V_{i+1} := \Psi(V_i)$; $i := i + 1$;
    последовательность $\{V_i\}$ — лучшая;
    остановка;
  иначе
    пока $\exists V_u : \overrightarrow{V_w V_u} \in WG$
      $g(V_u) := g(V_w) + g(u)$;
      $\hat{h}(V_u) := g(V_u) + (h(V_u) - u.\text{cost})$;
      если $V_u \in L_{\text{close}}$ **то**
        $L_{\text{open}} := V_u$;
        $h^\star(V_u) := \hat{h}(V_u)$;
      иначе если $\hat{h}(V_u) < h^\star(V_u)$ **то**
        $h^\star(V_u) := \hat{h}(V_u)$;
      создать указатель: $\Psi(V_u) := V_w$;

В частности, таким образом может быть найдена оптимальная последовательность слов по принципу Sentence N-best [8], использованы более сложные схемы оценок правдоподобия, например, специально построенных акустических счетов, модели языка, межсловные (crossword) акустические модели или взвешенные по краям оценки правдоподобия [6].

Пусть $g(V)$ обозначает такую «улучшенную» функцию стоимости пути из начала графа в вершину V. Тогда алгоритм пересчета на графе слов может быть получен простым обобщением (число начальных вершин может быть равно размеру словаря) оптимального алгоритма перебора $A^\star$. При этом функция h используется в оценочной стоимости пути $h^\star$ от данной вершины к целевой: $h^\star(V_w) = g(V_w) + w.\text{tail}$.

Алгоритм поиска на графе слов

При записи промежуточного результата распознавания в виде графа слов представление в виде списков лучших гипотез не требуется, и для выбора наиболее правдоподобной гипотезы о речевом высказывании требуется выполнить поиск на графе слов в соответствии с новой функцией стоимости $g(w)$. Предполагается, что алгоритм вычисления стоимости $g(w)$ задан. Алгоритм поиска имеет следующий вид.

Заключение

В докладе рассмотрена проблема представления промежуточных результатов автоматического распознавания речи. В качестве экономичного и вычислительно эффективного способа представления результатов использованы графы слов. Приведены алгоритмы формирования графа слов и пересчета результатов распознавания речи на графе.

Литература

[1] *Ney H.* WordGraphs: An efficient interface between continuous speech recognition and language understanding // IEEE Int.Conf on Acoustics, Speech and Signal Processing, ICASSP, 1993— Pp. 119–122.

[2] *Schwartz R., Nguyen L., Makhoul J.* Mupliple Path Search Strategies // Kluwer Academic Pulisher, 1996— Pp. 423–456.

[3] *Young S.* The HTK Book. Cambridge University, 1997.

[4] *Mangu L., Brill E., Stolcke A.* Finding consensus in speech recognition: word error minimization and other applications on confusion networks // Computer Speech and Language, 14, 2000. — Pp. 373–400.

[5] *Hakkani-Tur D. Bechet F. Riccardi G. Tur G.* Beyond ASR 1-best: Using word confusion networks in spoken language understanding // Computer Speech and Language, 20, 2006. — Pp. 495–514.

[6] *Demuynck K. Duchateau J., Compernolle D., Wambacq P.* An efficient search space representation for large vocabulary continuous speech recognition // Speech Communication. 30, 2000. — Pp. 37–53.

[7] *Нильсон Н.* Искусственный интеллект. Методы поиска решений // М: Мир, 1973.

[8] *Young S., Bloothooft G.* Corpus-based methods in language and speech processing // Kluwer Academic Publishers, 1997.

Прикладные задачи и системы интеллектуального анализа данных

Код раздела: AP (Applied Problems)

- Дистанционное зондирование и ГИС.
- Медицинские приложения анализа сигналов.
- Биомедицинские приложения анализа изображений.
- Приложения в области промышленности и техники.
- Прогнозирование свойств химических соединений.
- Анализ генетических последовательностей и белков.
- Социально-экономические приложения.
- Анализ и понимание текста.
- Информационный поиск.

Разделение малонаполненных классов методом скользящего контроля*

Барчуков М. А., Двоенко С. Д.
barchukovmaxim@rambler.ru, dsd@uic.tula.ru
Тула, Тульский государственный университет

Рассмотрена задача распознавания малонаполненных классов аминокислотных последовательностей, заданных значениями взаимных близостей. Для ее решения модифицирован известный алгоритм обучения для представления оптимальной разделяющей гиперплоскости в случае, когда исходное пространство признаков не задано. Показано, что результат обучения эффективно улучшается при сдвиге разделяющей гиперплоскости в направлении малонаполненного класса.

Введение

Методология анализа данных традиционно основана на изучении свойств матрицы данных — таблицы, строки которой содержат результаты измерений характеристик изучаемого явления в каждом акте измерения (объекты), столбцы которой содержат результаты измерений каждой из характеристик изучаемого явления (вариационные ряды как признаки). В рамках естественного геометрического подхода объекты представляют собой векторы в абстрактном многомерном метрическом пространстве, образованном признаками.

В современных условиях значительно расширились способы представления информации об объекте исследования. В частности, стало обычным явлением, что данные сразу же представлены в виде матрицы парных сравнений элементов анализируемого множества. Результатом сравнения в количественных шкалах может быть неотрицательная численная величина различия или неотрицательная величина противоположной характеристики — сходства элементов множества.

В частности, сегодня общепринятым является определение похожести последовательностей аминокислот, образующих полимерные цепи молекул белков, на основе парного выравнивания (элаймента), выполняемого, например, программой Fasta [1]. Следует отметить, что во многих актуальных задачах обработки невозможность получения традиционной матрицы данных часто объективно обусловлена [2].

Если матрица сходства положительно полуопределена, то ее можно рассматривать как матрицу скалярных произведений в некотором неизвестном нам метрическом (евклидовом) пространстве, размерность которого не превышает числа элементов множества. В этом случае, как известно, такая матрица сходства (скалярных произведений) может быть преобразована в матрицу различий (расстояний) и наоборот. Таким образом, соответствующая матрица различий может пониматься как матрица расстояний в том же неизвестном нам пространстве. Если элементами множества являются признаки, то в качестве матрицы их сходства достаточно взять матрицу модулей или квадратов взвешенных скалярных произведений (корреляций) признаков.

Распознавание объектов на основе их взаимного сходства

Пусть относительно начала координат ω_0 вычислена матрица $C(N,N)$ взаимных скалярных произведений $(\omega_i \circ \omega_j)$ объектов $\omega_i, \omega_j \in \Omega$, где N — число объектов.

Разделимость классов Ω_1 и Ω_2 означает, что, по меньшей мере, выпуклые оболочки множеств объектов из разных классов не пересекаются. Если выпуклые оболочки разделяемых множеств не соприкасаются, то в «зазоре» между ними можно построить более одной разделяющей гиперплоскости. В отсутствие иной априорной информации обычно выбирают гиперплоскость, наиболее удаленную от выпуклых оболочек разделяемых множеств. Такая оптимальная разделяющая гиперплоскость обеспечивает наименьшее число ошибок распознавания объектов, не участвовавших в обучении.

Для объектов $\omega_i, \omega_j \in \Omega$ обучающего множества Ω, представленных взаимными скалярными произведениями, представление решающего правила в неизвестном нам метрическом пространстве имеет вид

$$(\omega_a \circ \omega) + a_0 = c_{a\omega} + a_0, \qquad (1)$$

где ω_a — «направляющий» объект, a_0 — его «смещение» относительно начала координат ω_0, ω — новый объект. Если (1) положительно, то $\omega \in \Omega_1$, иначе $\omega \in \Omega_2$.

Пусть объекты $\omega_i \in \Omega$ обучающего множества $\Omega_1 \cup \Omega_2$ представлены своими нормализованными взаимными близостями $s(\omega_i, \omega_j) \geqslant 0$, $\omega_j \in \Omega$, где близость объекта с самим собой $s(\omega_i, \omega_i) = 1$.

Применим алгоритм Б. Н. Козинца, модифицированный для взаимных скалярных произведений объектов [3], чтобы построить правило (1) для распознавания новых объектов ω. В итоге необходимо вычислить величину

$$(\omega_a \circ \omega) + a_0 = s(\omega^+, \omega) - s(\omega^-, \omega)$$

*Работа выполнена при финансовой поддержке РФФИ, проекты №08-01-12023, №08-01-99003, №09-07-00394.

для распознавания нового объекта ω, где элементы $\omega^+ \in \Omega_1$ и $\omega^- \in \Omega_2$ найдены на этапе обучения алгоритмом Козинца.

Алгоритм Козинца обладает следующим замечательным свойством: между элементами $\omega^+ \in \Omega_1$ и $\omega^- \in \Omega_2$ выпуклых оболочек разделяемых множеств Ω_1 и Ω_2 гарантированно не содержится ни одного элемента обучающей совокупности $\Omega_1 \cup \Omega_2$. Таким образом, если $d(\omega^+, \omega^-)$ — длина интервала между объектами ω^+ и ω^-, то интервал длины $(1-\varepsilon)\, d(\omega^+, \omega^-)$ на «оси» направляющего объекта ω_a посередине между элементами ω^+ и ω^-, где $0 < \varepsilon < 1$ — достаточно малая наперед заданная величина, не содержит ни одного объекта.

Такой интервал ограничен двумя «точками» на «оси» направляющего объекта ω_a:

$$a_0' = -\big(s(\omega^+, \omega^+) - s(\omega^+, \omega^-)\big) + \varepsilon d(\omega^+, \omega^-)/2,$$
$$a_0'' = -\big(s(\omega^+, \omega^-) - s(\omega^-, \omega^-)\big) - \varepsilon d(\omega^+, \omega^-)/2,$$

где первая точка представлена смещением a_0' от начала координат ω_0 в сторону класса Ω_1, а вторая точка представлена противоположным смещением a_0'' от начала координат ω_0 в сторону класса с Ω_2.

Пусть n_1 — число объектов первого класса («свой» класс), n_2 — число объектов второго класса («чужой» класс), n_{11} — число правильно распознанных объектов первого класса, n_{22} — число правильно распознанных объектов второго класса, n_{12} — число своих объектов, неправильно распознанных как чужие объекты, n_{21} — число чужих объектов, неправильно распознанных как свои объекты. Соответствующие выборочные оценки вероятностей имеют вид:

$p_{11} = n_{11}/n_1 = n_{11}/(n_{11} + n_{12})$ — вероятность правильного распознавания своих объектов (сенситивность);

$p_{12} = n_{12}/n_1 = n_{12}/(n_{11} + n_{12})$ — вероятность неправильного распознавания своих объектов (негативная ошибка);

$p_{21} = n_{21}/n_2 = n_{21}/(n_{21} + n_{22})$ — вероятность, что чужие объекты распознаны как свои (позитивная ошибка);

$p_{22} = n_{22}/n_2 = n_{22}/(n_{21} + n_{22})$ — вероятность правильного распознавания чужих объектов (специфичность).

Таким образом, при скользящем контроле возникают ошибки двух типов: когда свои объекты распознаются как чужие, и наоборот.

В результате, ошибки этих двух типов возникают тогда, когда распознаваемый объект ω оказывается единственным в «промежутке» между выпуклыми оболочками разделяемых множеств, но с «неправильной» стороны от разделяющей гиперплоскости. Если, тем не менее, такой объект может быть отделен от чужого класса, то ошибку легко устранить, «сместив» гиперплоскость к множеству Ω_1 на величину $a_0' + \eta$ или к множеству Ω_2 на величину $a_0'' - \eta$, где $0 < \eta < 1$ — достаточно малая наперед заданная величина порога разделимости.

Задача разделения малонаполненных классов аминокислотных последовательностей

Одно из основных положений молекулярной биологии заключается в том, что первичная структура белка, то есть его последовательность аминокислотных остатков, содержит полный объем информации, которая однозначно определяет его пространственную структуру.

Как правило, конфигурации белковых макромолекул оказываются похожими в больших группах эволюционно близких белков. В результате оказывается, что множество существенно различающихся пространственных структур белков значительно меньше множества всех известных белков. Вследствие этого, проблема выявления пространственной структуры белковых макромолекул оказывается задачей распознавания.

Рассматриваемое в данной работе множество белков было отобрано др. Сан Хо-Кимом из Национальной лаборатории Лоуренса в Беркли. Это множество содержит 420 белковых последовательностей, образующих 51 класс, которые были выбраны из базы данных SCOP [4] по принципу наименьшей похожести. Порог наименьшей похожести белков друг на друга в каждом из семейств по результатам парного выравнивания протеиновых цепочек программой Fasta составил 27%. В результате были получены малонаполненные классы по сравнению с общим размером множества отобранных белков. Например, три наименьших класса состояли только из трех белков, а три класса наибольшего размера содержали 38, 31 и 20 белков

Из-за того, что исходная матрица близостей $S(N, N)$ оказалась отрицательно определена и имела пять небольших отрицательных собственных чисел, все значения близостей $s(\omega_i, \omega_j)$ были нормализованы к величине $s_{ij}/\sqrt{s_{ii}s_{jj}}$ и возведены в квадрат. После такого преобразования матрица близостей оказалась положительно полуопределена. Значения всех ее собственных чисел оказались расположены в диапазоне от $\lambda_1 = 15{,}752$ до $\lambda_{418} = 0{,}336$, где $\lambda_{419} = \lambda_{420} = 0$.

Таким образом, предполагалось, что объекты $\omega_i,\ i = 1, \ldots, N$ были представлены своими скалярными произведениями $(\omega_i \circ \omega_j)$ с остальными объектами $\omega_j,\ j = 1, \ldots, N$ в неизвестном метрическом пространстве размерности не выше 420.

Ранее были проведены эксперименты, в которых каждый объект ω_i, представленный своими близостями $s_{ij},\ j = 1, \ldots, N$, рассматривался как вектор в 420-мерном пространстве, названном «проекционным». Необходимость такого пред-

ставления была обусловлена тем, что оказывалось возможным использовать все известные алгоритмы кластер-анализа и распознавания без их модификации [2].

Предварительный анализ данной коллекции белков, помещенных в проекционное пространство, показал плохую разделимость классов в целом. В результате, в экспериментах по скользящему контролю «один класс против всех» плохая разделимость была подтверждена: только 14 классов из 51 были распознаны с ошибкой 19,4% (75 из 93 объектов). В таблице 1 показаны размеры и названия этих классов.

В итоге, общая доля правильно распознаваемых объектов оказалась не выше 22% (93 из 420). В таблице 2 показаны доли в процентах правильного распознавания и использованные алгоритмы. Результаты были получены путем построения решающего правила по максимуму апостериорной вероятности (Posteriori Probability, PP) и по методу ближайших соседей (Nearest Neighbor, NN) [2].

Распознавание классов аминокислотных последовательностей

В данной работе считалось, что объекты были представлены взаимными близостями в неизвестном метрическом пространстве, в отличие от предыдущего представления векторами в известном проекционном пространстве высокой размерности. В данном случае было необходимо модифицировать алгоритмы распознавания и кластер-анализа, как показано в [3, 5].

Легко увидеть, что отдельный класс белков (первый, т. е. «свой» класс обучения) содержит небольшое число объектов по сравнению с остальными классами (второй, т. е. «чужой» класс). Поэтому в экспериментах «один против всех» методом скользящего контроля «форма» отдельного малонаполненного класса обучения в гипотетическом пространстве оказывается нестабильной по сравнению с формой другого класса обучения, образованного всеми остальными объектами из оставшихся 50 классов белков. Эта нестабильность и определяет низкое качество распознавания малонаполненного класса в скользящем контроле (высокий уровень ошибки).

Эксперименты показали, что все ошибки скользящего контроля оказались только одного типа, когда объект из своего класса распознается как чужой. Поэтому необходимо изменить смещение на новое для улучшения результата распознавания. В итоге, результат распознавания оказался значительно лучше. В таблице 3 показан результат распознавания 14 классов, которые ранее достаточно хорошо распознавались в проекционном пространстве. В данной работе было достигнуто значение 97.4% доли правильного распознавания при сколь-

Таблица 1. Хорошо выделяемые классы.

Класс	Название	Размер
1	Globin	12
2	Cytochrome C	7
6	EF Hand	13
7	Cyclin	4
8	Cytochrome P450	5
11	Cupredoxins	9
14	Crystallins/protein S/yeast killer toxin	5
21	Acid proteases	5
23	Lipocalins	6
25	Barrel-sandwich hybrid	6
39	Periplasmic binding protein I	7
47	N-terminal nucleophile aminohydrolases	4
49	C-type lectin	6
50	Protein kinases (PK), catalytic core	4

Таблица 2. Распознавание хорошо выделяемых классов.

Класс	% правильных	Алгоритм
1	92% (11)	PP
2	100% (7)	1NN
6	85% (11)	3NN
7	75% (3)	1NN
8	80% (4)	PP
11	78% (7)	3NN
14	60% (3)	PP
21	60% (3)	3NN, PP
23	50% (3)	PP
25	67% (4)	3NN, PP
39	86% (6)	3NN
47	100% (4)	1NN
49	83% (5)	1NN
50	100% (4)	3NN

Таблица 3. Повышение качества распознавания хорошо выделяемых классов.

Класс	Размер	% правильных
1	12	100% (12)
2	7	100% (7)
6	13	92% (12)
7	4	100% (4)
8	5	100% (5)
11	9	100% (9)
14	5	100% (5)
21	5	100% (5)
23	6	83% (5)
25	6	100% (6)
39	7	100% (7)
47	4	100% (4)
49	6	100% (6)
50	4	100% (7)

Таблица 4. Плохо выделяемые классы.

Класс	Название	Размер	% правильных
10	Common fold of difteria toxin/ trnscription factors/ cytochrome	5	20% (1)
12	C2 domain	3	33% (1)
33	Thioredoxin fold	5	20% (1)
36	S-adenosyl-L-methionine-dependent methyl-transferases	5	20% (1)
41	Lysozyme	4	25% (1)
13	Viral coat and capsid proteins	15	47% (7)
17	OB-fold	17	12% (2)
24	Double-stranded beta-helix	6	33% (3)
44	Cystatin	7	43% (3)

Таблица 5. Распознавание 28 классов.

Класс	Размер	% правильных
3	8	75% (6)
4	8	87% (7)
5	11	100% (11)
9	31	71% (22)
15	4	75% (3)
16	8	50% (4)
18	5	60% (3)
19	4	75% (3)
20	6	67% (4)
22	7	57% (4)
26	38	100% (38)
27	9	56% (5)
28	4	50% (2)
29	14	79% (11)
30	3	67% (2)
31	9	56% (5)
32	9	89% (8)
34	9	56% (5)
35	3	67% (2)
37	12	100% (12)
38	5	60% (3)
40	7	57% (4)
42	4	75% (3)
43	8	75% (6)
45	20	55% (11)
46	7	100% (7)
48	4	100% (4)
51	3	67% (2)

зящем контроле (91 объект из 93), выполненного алгоритмом Козинца.

Доля правильно распознаваемых объектов для всех классов достигла 73% (307 из 420).

Невозможность получения более высокого результата была обусловлена множеством из 9 классов с низким уровнем (меньше 50%) доли правильно распознаваемых объектов (табл. 4). Эксперименты показали, что было невозможно улучшить долю правильного распознавания объектов для классов с номерами 10, 12, 33, 36 и 41 при сдвиге разделяющей гиперплоскости как было предложено выше. Для другой группы классов с номерами 13, 17, 24 и 44 этот показатель был улучшен незначительно. Повышение качества распознавания для остальных 28 классов показано в таблице5.

Заключение

Применение проекционного пространства оказывается удобным из-за возможности непосредственного использования известных алгоритмов распознавания без их модификации. Но при таком подходе свойства пространства оказываются неизвестными, а его высокая размерность делает весьма проблематичным применение метода скользящего контроля.

Представление объектов взаимными близостями и их погружение в «традиционное» метрическое пространство позволило улучшить результат распознавания, а отсутствие необходимости восстановления самого пространства позволило резко уменьшить трудоемкость скользящего контроля. Но для этого потребовалось модифицировать применяемый алгоритм распознавания.

Литература

[1] *Pearson W. R.* Flexible Sequence Similarity Searching with the FASTA3 Program Package // Methods Mol. Biol. — 2000. — № 132. — P. 185–219.

[2] *Mottl V. V., Dvoenko S. D., Seredin O. S., Kulikowski C. A., Muchnik I. B.* Featureless Pattern Recognition in an Imaginary Hilbert Space and Its Application to Protein Fold Classification // 2nd International Workshop on Machine Learning and Data Mining in Pattern Recognition (MLDM), Leipzig: Springer, 2001. — P. 322–336.

[3] *Двоенко С. Д.* Распознавание элементов множества, представленных взаимными расстояниями и близостями // Всеросс. конф. ММРО-14, 2009 (в данном сборнике). — С. 112–115

[4] *Dubchak I., Muchnik I., Mayor C., Dralyuk I., Kim S.-H.* Recognition of a Protein Fold in the Context of the SCOP Classification // Proteins: Structure, Function, and Genetics — 1999. — № 35. — P. 401–407.

[5] *Двоенко С. Д.* Кластеризация элементов множества на основе взаимных расстояний и близостей // Всеросс. конф. ММРО-13., М.: МАКС Пресс, 2007. — С. 114–117.

Алгоритм множественного трекинга лабораторных животных*

Ветров Д. П., Кропотов Д. А.
vetrovd@yandex.ru, dkropotov@yandex.ru
Москва, ВМиК МГУ, Вычислительный Центр РАН

В работе рассматривается задача сопровождения (трекинга) нескольких лабораторных мышей, находящихся в клетке со встроенной системой видеонаблюдения. Предложены алгоритмы отделения изображений мышей от фона, выделения отдельных особей и их последующей идентификации. Также в работе предложен специальный способ тестирования алгоритма множественного сопровождения.

Введение

Одним из основных инструментов в современных когнитивных исследованиях являются методы анализа поведения человека или животного. Построение описания поведения вручную является очень трудоемким процессом, поэтому все большую популярность получают системы автоматического анализа поведения. Такие системы, как правило, включают в себя модуль видеонаблюдения, модуль сегментации видеосигнала (выделения в нем элементарных структурных единиц поведения — поведенческих актов) и модуль поиска закономерностей в поведении. Первичным элементом любой системы автоматического анализа поведения является модуль видеонаблюдения, позволяющий осуществлять сопровождение наблюдаемого объекта (трекинг) и, возможно, оценивать ряд его характеристик, например, геометрическую форму, скорость и т. п. Наибольшее распространение такие системы получили для наблюдения за лабораторными мышами.

Серьезным ограничением современных систем видеонаблюдения за животными является их неспособность осуществлять наблюдение и идентификацию одновременно для нескольких животных, находящихся в клетке. Это сильно ограничивает область их применения, т.к. не позволяет анализировать социальное поведение. Кроме того, большую часть времени вне экспериментов животное проводит в т.н. домашней клетке, содержащей обычно 3–5 мышей. Для изучения изменений поведения животного в течение долгих интервалов времени (недели, месяцы) желательно осуществлять видеонаблюдение за ним и в домашней клетке. Современные технологии позволяют это сделать аппаратно, но надежных методов для множественного трекинга и идентификации мышей в клетке до сих пор не создано. Это связано в первую очередь со сложностями при разделении мышей, которые любят сбиваться в кучку и находиться в тесном телесном контакте, а также с персонификацией найденных особей. Существующие аналоги обычно либо используют вид сбоку [3], либо сильно ограничивают возможные перекрытия тел мышей [6].

*Работа выполнена при финансовой поддержке РФФИ, проект № 08-01-00405.

В настоящей работе для решения первой проблемы предложено использовать EM-подобную процедуру разделения смеси распределений. При этом E-шаг отличается от классического подсчета вероятности принадлежности к компоненте смеси в том смысле, что одна и та же точка на изображении может принадлежать двум и более мышам (например, когда одна мышь забралась на спину другой). Для идентификации предложен алгоритм, основанный на определении близости к положению мыши на предыдущем кадре.

Вычитание фона

Входными данными для рассматриваемых алгоритмов являлись видеоролики с разрешением 500×550 пикселей, содержащих запись движения одиночной мыши в клетке в течение 15 минут. На основе этих записей был создан виртуальный видеоролик, позволяющий отлаживать систему множественного видеотрекинга.

Для отслеживания животных внутри сцены используется традиционный метод видеослежения, основанный на «вычитании фона» [4]. Изображение было переведено в серую шкалу, и для каждого пикселя оценивалась медиана интенсивности серого по N кадрам видеосъемки сцены при том же положении камеры и освещении, что и в процессе эксперимента, но без животных. Далее для кадров с животными пиксели, интенсивность которых составляла менее 60% от соответствующего ему медианного значения, помечались как точки объекта. Найденные пиксели объекта фильтровались с помощью методов математической морфологии [7]. Это позволило отсечь одиночные выбросы, помехи и отдельные мелкие объекты, например, насекомых, попавших в кадр. Значение порога было выбрано так, чтобы к объекту не причислялись тени, отбрасываемые мышами, и их отражения от стен клетки. В результате вычитания фона выделялась маска, содержащая изображения мышей.

Разделение особей

В данной работе предполагалось, что общее количество мышей, находящихся в клетке, известно. В том случае, если в результате выделения масок животных на видеокадре количество компонент связности равнялось количеству мышей, на-

Рис. 1. Пример работы системы трекинга одной мыши в клетке. Белым показан выделенный контур мыши. Также отмечены найденные ключевые точки: звездочкой — центр масс, плюсиком — нос, крестиком — хвост.

ходящихся в клетке, то задача сводилась к определению контуров и характерных точек (носа, точки крепления хвоста), аналогичному случаю наблюдения за одной мышью [1]. Пример работы алгоритма определения контура и характерных точек показан на рис. 1.

Также для каждого контура находился эллипс, наилучшим образом его приближающий. Для этого для каждой маски мыши M_k рассчитывалась матрица

$$A_k^{-1} = \frac{\sum\limits_{\boldsymbol{x}_i \in M_k} (\boldsymbol{x}_i - \mathbb{E}_k\boldsymbol{x})(\boldsymbol{x}_i - \mathbb{E}_k\boldsymbol{x})^\intercal}{\sum\limits_{\boldsymbol{x}_i \in M_k} 1},$$

где

$$\mathbb{E}_k\boldsymbol{x} = \frac{\sum\limits_{\boldsymbol{x}_i \in M_k} \boldsymbol{x}_i}{\sum\limits_{\boldsymbol{x}_i \in M_k} 1},$$

а $\boldsymbol{x}_i$ — пиксели кадра. Уравнение эллипса, наилучшим образом приближающего контур мыши, имело вид

$$(\boldsymbol{x} - \mathbb{E}_k\boldsymbol{x})^T A_k (\boldsymbol{x} - \mathbb{E}_k\boldsymbol{x}) = \frac{1}{4}. \qquad (1)$$

Если количество компонент связности оказывалось меньше количества мышей, то анализировались эллипсы, вписанные в контур каждой мыши на предыдущем кадре. Для каждой компоненты связности определялись те особи, контуры которых оказывались внутри компоненты связности. Это определялось путем подсчета ближайшей компоненты связности к эллипсам с предыдущего кадра.

Далее для каждой компоненты связности, содержащей более одной мыши, производилось разделение мышей с помощью процедуры, представляющей собой модификацию EM-алгоритма разделения смеси распределений [5]. Обозначим через C_k — внутренность эллипса (1). Алгоритм разделения особей представляет собой двухшаговую итеративную процедуру. На первом шаге (аналог E-шага EM-алгоритма) осуществляется расчет переменных $\gamma(z_{ik})$, определяющих принадлежность данной точки маски $\boldsymbol{x}_i$ к k-ой особи. Заметим, что в отличие от классического EM-алгоритма, сумма принадлежностей к различным особям не обязана равняться единице, т.к. одна и та же точка может принадлежать двум и более животным одновременно, например, из-за того, что одно животное влезло на спину другому:

$$\gamma(z_{ik}) = \begin{cases} 1, & \boldsymbol{x}_i \in C_k \text{ или } k = \arg\min\limits_j \rho_M(\boldsymbol{x}_i, \mathbb{E}_j\boldsymbol{x}); \\ 0, & \text{иначе}; \end{cases}$$

где $\rho_M(\boldsymbol{x}_i, \mathbb{E}_j\boldsymbol{x})$ — расстояние Махаланобиса, задаваемое матрицей A_j.

На втором шаге (аналог M-шага EM-алгоритма) осуществляется уточнение эллипсов. Обозначим через α_k угол наклона большей полуоси эллипса (направления, соответствующего носу[1]), взятого с предыдущего кадра. Для уточнения матрицы A_k сначала центрируем систему координат и выполняем преобразование поворота на угол $-\alpha_k$:

$$\boldsymbol{y}_i = R_{-\alpha_k}(\boldsymbol{x}_i - \mu_k),$$

где

$$\mu_k = \frac{\sum\limits_i \gamma(z_{ik})\boldsymbol{x}_i}{\sum\limits_i \gamma(z_{ik})} = \frac{\sum\limits_i \gamma(z_{ik})\boldsymbol{x}_i}{n_k}.$$

Затем осуществляется поиск наиболее вероятной матрицы A_k путем максимизации апостериорной плотности на матрицу A'_k, представляющей собой матрицу A_k, повернутую на угол $-\alpha_k$. Функция правдоподобия на A'_k принимает следующий вид:

$$p(Y|A'_k) = \left(\sqrt{\det A'_k}\right)^{n_k} \exp\left(-\frac{1}{2}\sum_i \boldsymbol{y}_i^\intercal A'_k \boldsymbol{y}_i\right).$$

Для того, чтобы учесть априорные представления о характерной форме эллипса, приближающего контур мыши, введем априорное распределение Уишарта на множестве матриц

$$p(A'_k) = \\ = B(V,\nu)\big(\det A'_k\big)^{\frac{\nu-3}{2}} \exp\big(-\tfrac{1}{2}\operatorname{tr}(V^{-1}A'_k)\big) \sim \\ \sim \mathcal{W}(A'_k|V,\nu), \quad (2)$$

[1] Об идентификации точек носа и крепления хвоста см., например, работу [1]

где $B(V,\nu)$ — (несущественная для нас) нормировочная константа. Вопрос выбора параметров распределения Уишарта будет рассмотрен в следующем разделе. В связи с тем, что распределение Уишарта является сопряженным для нормального распределения с матрицей A'_k, то апостериорное распределение может быть вычислено аналитически. Оно также будет представлять собой распределение Уишарта. Выражение для максимума апостериорной плотности для A'_k имеет вид

$$A'_k = (n_k + \nu - 3)\left(\sum_i \gamma(z_{ik})\, \boldsymbol{y}_i^{\mathsf{T}} \boldsymbol{y}_i + V^{-1}\right)^{-1}.$$

Чтобы получить матрицу A_k, осуществляем обратный поворот системы координат на угол α_k

$$A_k = R_{\alpha_k} A'_k R_{\alpha_k}^{\mathsf{T}}.$$

Описанные два шага повторяются итерационно до сходимости. В качестве начального приближения для каждой особи, чей контур входит в данную компоненту связности, используются матрицы и центры эллипсов с предыдущего кадра.

Настройка параметров априорного распределения

Введение априорного распределения позволяет отсекать заведомо нереалистичные эллипсы, например, слишком длинные или слишком узкие, которые не могут описывать контур реальной мыши. Распределение Уишарта было выбрано в качестве априорного в том числе и потому, что оно позволяет получить выражение для максимума апостериорной плотности в явном виде при использовании гауссовского распределения в качестве функции правдоподобия (т. е. является сопряженным к последнему). Для набора априорной информации мы воспользовались несколькими видеороликами с одиночной мышью в клетке, снятыми в тех же условиях. Для каждого кадра решалась задача вписывания эллипса в контур мыши. Затем полученные эллипсы поворачивались так, чтобы их первый собственный вектор (т. е. собственный вектор, отвечающий большему собственному значению) был параллелен оси абсцисс. Обозначим через $\mathbb{E}A$ среднее арифметическое получившихся матриц. Известно, что математическое ожидание распределения Уишарта выражается формулой $\mathbb{E}A = \nu V$, отсюда

$$V = \frac{\mathbb{E}A}{\nu}.$$

Для оценки параметра ν воспользуемся информацией об изменении длины большей полуоси эллипса r_1. Легко показать, что если неотрицательно определенная матрица имеет распределение Уишарта, то ее собственные значения имеют гамма-распределение. Пусть r_1 имеет гамма-распределение с параметрами a и b, т. е. $r_1 \sim \mathcal{G}(r_1|a,b)$. Оценим их значения, используя известные соотношения между ними и первыми двумя моментами распределения

$$\mathbb{E}r_1 = \frac{a}{b}, \quad \mathbb{D}r_1 = \frac{a}{b^2}.$$

С другой стороны, известно, что распределение Уишарта является многомерным обобщением гамма-распределения, причем параметру a соответствует значение $\frac{1}{2}\nu$ [2]. Приняв предположение о близости характеристик разброса длины большей полуоси и характеристики изменчивости эллипса, в качестве значения ν примем величину

$$\nu = 2a = 2\frac{(\mathbb{E}r_1)^2}{\mathbb{D}r_1}.$$

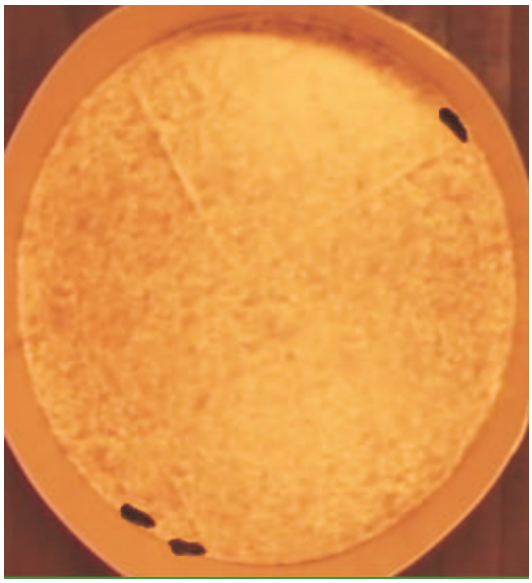


Рис. 2. Пример кадра из виртуального ролика, содержащего наложенные изображения трех мышей.

Построение тестового видеоролика и результаты экспериментов

Для оценки качества разработаного алгоритма видеотрекинга был разработан специальный виртуальный видеоролик, содержащий изображения многих мышей, информацию об их форме и идентификаторы особей. Данный видеоролик был получен путем обработки видеоклипа, содержащего запись поведения одной мыши в течение сравнительно долгого промежутка времени. Этот клип был разбит на p равных частей. Из каждого кадра был выделен контур мыши, а затем был получен новый ролик путем наложения p контуров, взятых с соответствующих кадров каждой из частей, на статическую модель фона. Кадр из получившегося ролика показан на рис. 2. Пример работы ал-

Рис. 3. Пример кадра из виртуального ролика с разделением особей с помощью предложенной EM-подобной процедуры. Белыми эллипсами показаны спрогнозированные формы мышей, плюсами отмечены точки носа.

горитма разделения особей показан на рисунке 3. Алгоритм успешно отработал на видеороликах, содержащих 3 мыши в течение 15 минут, не сделав при этом ни одного сбоя сопровождения или идентификации. Примеры его работы будут показаны в ходе доклада.

Литература

[1] *Конушин А. С., Ветров Д. П., Воронин П. А., Синдеев М. С., Ломакина-Румянцева Е. И., Кропотов Д. А., Зарайская И. Ю., Анохин К. В.* Система видеонаблюдения за поведением лабораторных животных с автоматической сегментацией на поведенческие акты // Труды межд. конф. Графикон-2008, М.: МАКС Пресс, 2008. — С. 199–206.

[2] *Bishop, C.* Pattern Recognition and Machine Learning, Springer, 2006.

[3] *Branson K., Belongie S.* Tracking multiple mouse contours (without too many samples) // Proc. of the 2005 IEEE Comp. Soc. Conf. on Computer Vision and Pattern Recognition (CVPR'05) 2005, — Vol. 1, Pp. 1039–1046.

[4] *Cheung S., Kamath C.* Robust techniques for background subtraction in urban traffic video // Visual Communications and Image Processing, 2004, Pp. 881–892.

[5] *Dempster A., Laird N., Rubin D.* Maximum likelihood from incomplete data via the EM algorithm // Journal of the Royal Statistical Society, 1977, B 39, 1, 1–38.

[6] *Goncalves W. N., Monteiro J. B. O., de Andrade Silva J., Machado B. B., Pistori H., Odakura V.* Multiple mice tracking using a combination of particle filter and k-means // Proc. of XX Brazilian Symp. on Computer Graphics and Image Processing, 2007, Pp. 173–178.

[7] *Serra J.* Image Analysis and Mathematical Morphology. New York: Academic, 1982.

«Полигон» — распределённая система для эмпирического анализа задач и алгоритмов классификации*

Воронцов К. В., Ивахненко А. А, Инякин А. С., Лисица А. В., Минаев П. Ю.
vokov@forecsys.ru, lisitsa@forecsys.ru
Москва, Вычислительный Центр РАН, ЗАО «Форексис»

Система «Полигон» предназначена для массового выполнения типовых экспериментов по тестированию алгоритмов классификации на модельных и реальных данных. В отличие от существующих систем такого типа, «Полигон» является Интернет-ресурсом, имеет централизованное хранилище задач и результатов тестирования, распределённую наращиваемую пользователями сеть вычислительных серверов и расширенную визуальную методику тестирования, основанную на $t\times q$-кратном скользящем контроле. Ресурс ориентирован на специалистов по анализу данных, прикладных экспертов, разработчиков алгоритмов, научных работников, учащихся и преподавателей вузов.

На конференции ММРО-13 было анонсировано начало работ по созданию распределённой системы тестирования алгоритмов классификации на задачах со стандартным представлением исходных данных в виде матрицы «объекты–признаки» [2]. В настоящее время прототип системы доступен по адресу `http://poligon.MachineLearning.ru`.

«Полигон» автоматизирует типовое эмпирическое исследование по сравнению качества заданного набора алгоритмов на заданном наборе задач. Проведение такого рода сравнительных экспериментов считается обязательным как при разработке новых методов классификации, так и при решении практических задач классификации.

В настоящее время широко известные системы анализа данных Matlab, R, STATISTICA, SAS, SPSS и др. имеют мощные библиотеки алгоритмов классификации. Свободно доступные системы для решения задач машинного обучения (RapidMiner, WEKA) укомплектованы наборами реальных задач и процедурами скользящего контроля для проведения упомянутых выше типовых экспериментов. Тем не менее, создание новой системы имеет смысл в силу ряда причин.

1. Перечисленные системы не гарантируют воспроизводимость и верифицируемость результатов тестирования. Все они устанавливаются локально на компьютере пользователя, оставляя ему возможность по-своему реализовать методику тестирования, модифицировать как исходные данные, так и сами алгоритмы. Как следствие, эмпирические результаты, представленные в разных публикациях, оказываются несопоставимыми, даже если тестирование проводилось на одних и тех задачах, как правило, из репозитория UCI [3].

2. В системах с открытым кодом R, RapidMiner, WEKA невозможно широкое использование коммерческих алгоритмов. В то же время, создатели алгоритмов могут быть заинтересованы в предоставлении их для пробного использования, не предполагающего извлечения коммерческой выгоды, и при этом максимально простого для пользователя. «Полигон» предоставляет удобную площадку для такого использования, при этом алгоритмы остаются на серверах правообладателя, и он имеет возможность ограничивать доступ к ним.

3. В «Полигоне» нет ограничений на язык программирования, на котором реализуются алгоритмы. RapidMiner и WEKA допускают только Java, что снижает скорость выполнения алгоритмов и затрудняет перенос в систему готовых алгоритмов, написанных на других языках.

4. В настоящее время стандартной методикой тестирования считается $t\times q$-кратный скользящий контроль [7]. В минимальном варианте для каждой пары $\langle$алгоритм, задача$\rangle$ вычисляется один скалярный критерий — средняя частота ошибок на контроле, как правило, с доверительным интервалом. В «Полигоне» реализована расширенная методика тестирования, включающая ряд широко известных методов, дающих более детальное и наглядное представление о качестве классификации. Эти методы обобщены и унифицированы таким образом, чтобы каждый из них, по возможности, позволял анализировать качество отдельно по классам, сопоставлять обучение и контроль и вычислять доверительные интервалы для всех оцениваемых величин.

Цель данного сообщения — показать возможности реализованной в «Полигоне» методики тестирования и привлечь научное сообщество к активному использованию и пополнению системы алгоритмами и задачами.

Класс решаемых задач

С точки зрения «Полигона» алгоритм классификации — это функция, принимающая на входе:

- обучающую выборку в виде матрицы «объекты–признаки» $(x_{ij})_{\ell\times n}$, где x_{ij} — значение j-го признака на i-м обучающем объекте x_i;
- вектор ответов $(y_i)_{i=1}^{\ell}$, соответствующих объектам x_i, где $y_i \in Y$, Y — множество классов;

*Работа поддержана РФФИ (проекты №07-07-00372, №08-07-00422, №07-07-00181), программой ОМН РАН «Алгебраические и комбинаторные методы математической кибернетики и информационные системы нового поколения».

— матрицу потерь $(C_{yy'})_{|Y|\times|Y|}$, где $C_{yy'}$ — штраф за отнесение объекта класса y к классу y';
— вектор информации $(I_j)_{j=1}^n$ о типах признаков;
— тестовую выборку в виде матрицы «объекты–признаки» $(x'_{ij})_{k\times n}$, где x'_{ij} — значение j-го признака на i-м тестовом объекте x'_i;

и выдающая на выходе:

— вектор ответов $(y'_i)_{i=1}^k$ на тестовой выборке;
— матрицу оценок $(p'_{iy})_{k\times|Y|}$ принадлежности каждого тестового объекта каждому из классов.

Оценки принадлежности могут быть апостериорными вероятностями (в байесовских классификаторах), или просто значениями дискриминантных функций классов. Если алгоритм не вычисляет вещественных оценок принадлежности, то полагается $p'_{iy} = [y'_i = y]$.

Методика тестирования

Процедура скользящего контроля. Производится N разбиений выборки $X = \{x_1, \ldots, x_L\} = X_n^\ell \sqcup X_n^k$ на обучающую подвыборку длины ℓ и контрольную длины $k = L - \ell$, где $n = 1, \ldots, N$ — номер разбиения. Обозначим через $a_n \colon X \to Y$ *функцию классификации*, получаемую при n-м разбиении в результате обучения по выборке X_n^ℓ.

Оценка скользящего контроля для произвольной функции от разбиения $\xi(n)$ определяется как среднее $\hat{\mathsf{E}}\xi = \frac{1}{N}\sum\limits_{n=1}^{N}\xi(n)$. Разбиения строятся по стандартной методике $t\times q$-fold cross-validation [7, 8]: генерируется t случайных разбиений выборки X^L на q блоков примерно равной длины и пропорциональными долями классов, и каждый блок поочерёдно становится контрольной выборкой. Таким образом, $N = tq$ и $k = \frac{L}{q}$, с точностью до округления. Каждый объект $x_i \in X^L$ выступает t раз в роли контрольного и $(t-1)q$ раз в роли обучающего. При достаточно больших t это позволяет строить доверительные интервалы для случайной величины $\xi(n)$ с помощью порядковых статистик, не делая никаких дополнительных предположений о виде распределения ξ.

Далее рассматриваются методы оценивания качества, реализованные в системе «Полигон», и приводятся примеры интерпретации результатов. Для иллюстрации взят алгоритм SVM и медицинская задача `Liver_Disorders` из репозитория UCI — разделение пациентов с нарушением работы печени (145 объектов класса 1) и здоровых людей (200 объектов класса 2). Число признаков равно 6.

Средняя частота ошибок на обучении $\nu^\ell = \hat{\mathsf{E}}\nu(a_n, X_n^\ell)$ и на контроле $\nu^k = \hat{\mathsf{E}}\nu(a_n, X_n^k)$ используется в эмпирических исследованиях чаще всего. В данной задаче $\nu^\ell = 22\pm5\%$, $\nu^k = 27\pm11\%$. Поверхностный анализ на таких оценках, как правило, и завершается. Однако уже простое разделение частоты по классам

класс 1, частота ошибок (обуч./конт.): 36%/43%;
класс 2, частота ошибок (обуч./конт.): 12%/18%;

выявляет важную особенность задачи: распознать больного намного труднее, чем здорового.

В «Полигоне» наряду с доверительными интервалами строятся графики эмпирических распределений частот ошибок на обучении и на контроле, а также *переобученности* $\delta_n = \nu(a_n, X_n^k) - \nu(a_n, X_n^\ell)$. Примеры таких графиков можно найти в [1].

Анализ вариации и смещения (bias-variance) [8]. Вводится функция *среднего предсказания* $\tilde{y}(x)$ — это класс, к которому объект $x_i \in X$ относится большинством функций a_n, $n = 1, \ldots, N$. *Смещение* на объекте x_i определяется как $B(x_i) = \big[\tilde{y}(x_i) \neq y_i\big]$. Соответственно, объекты выборки X разделяются на *смещённые* ($B(x_i) = 1$) и *несмещённые* ($B(x_i) = 0$). Смещённость означает, что объект плохо описывается данной моделью классификации (под моделью понимается параметрическое семейство алгоритмов).

Вариация $V(x_i)$ на объекте x_i определяется как доля разбиений n, при которых $a_n(x_i) \neq \tilde{y}(x_i)$. Эта величина характеризует изменчивость ответов на данном объекте по отношению к составу обучающей выборки. Неустойчивы, как правило, классификации объектов, находящихся вблизи границы классов. Поэтому число объектов с наибольшими вариациями (близкими к $\frac{1}{2}$) характеризует толщину пограничного слоя между классами. Чем тоньше этот слой, тем «правильней» модель классификации подобрана под задачу.

Заметим, что выявление пограничных объектов в многомерных пространствах признаков является нетривиальной задачей. «Полигон» решает эту задачу универсально, независимо от природы задачи и конструкции алгоритма; при этом список «пограничных» объектов может быть выдан пользователю в явном виде.

Суммирование смещений и вариаций по объектам даёт характеристики, называемые *смещением* и *вариацией выборки*. В сумме они составляют среднюю частоту ошибок классификации. Большая величина смещения говорит о том, что модель классификации, возможно, выбрана неудачно, и для решения данной задачи следует искать другой алгоритм. Большая величина вариации говорит о том, что результат обучения слишком сильно зависит от состава выборки; в таких случаях качество классификации можно улучшить, оставаясь в рамках той же модели классификации, путём регуляризации или композиции алгоритмов.

Объекты, которые оказываются смещёнными относительно большого числа различных алгоритмов, можно считать шумовыми выбросами. Таким

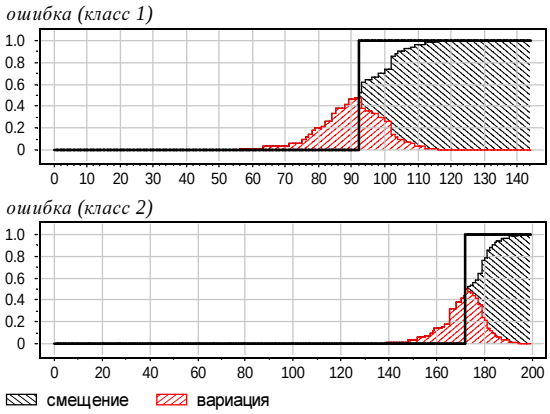


Рис. 1. Анализ вариации и смещения, по классам.

образом, «Полигон», при одновременном использовании большого числа разнообразных алгоритмов, позволяет идентифицировать выбросы более объективно, не привязываясь к какой-либо конкретной модели классификации.

«Полигон» позволяет анализировать разложение ошибки на смещение и вариацию отдельно по классам, рис. 1. По оси абсцисс отложены объекты, отсортированные по возрастанию ошибки на них. По оси ординат отложены значения вариации и ошибки на объектах. На несмещённых объектах ошибка состоит только из вариации, на смещённых — из разности смещения и вариации [4]. Видно, что у класса 1 (больные) пограничная зона больши́х вариаций намного шире, чем у второго класса. Опять-таки, это означает, что первый класс отделить труднее.

Кривая ошибок (ROC-кривая, receiver operating characteristic) используется для представления результатов классификации в тех случаях, когда соотношение цены ошибок I и II рода заранее неизвестно [6]. Предполагается, что имеется два класса: «положительный» и «отрицательный». По оси X откладывается доля *ошибочных положительных* классификаций, по оси Y — доля *правильных положительных* классификаций. Собственно, кривая получается в результате варьирования порога θ в функции классификации вида $a(x) = \text{sign}(f(x) - \theta)$, где $f(x)$ — вещественная дискриминантная функция.

Для многоклассовой задачи в роли положительного выступает каждый из классов по очереди, все остальные объединяются в один отрицательный класс; в результате строится серия из $|Y|$ ROC-кривых. Сопоставление ROC-кривых разных классов позволяет судить о том, какие алгоритмы, и при каких соотношениях цены ошибок, более целесообразно применять.

Чем больше *площадь под кривой* (AUC, area under curve), тем выше качество классификации. Критерий AUC является оценкой качества классификации, не зависящей от выбора соотношения цены ошибок. Различие между ROC-кривыми на обучении и контроле (рис. 2) позволяет судить о величине переобучения. Возможны ситуации, когда переобучение существенно различается в левой-нижней и в правой-верхей ветвях кривой; в таких случаях можно давать рекомендации об использовании данного алгоритма классификации только при определённых соотношениях цены ошибок.

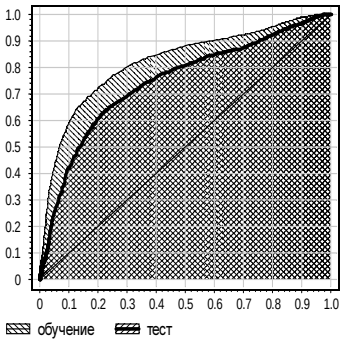


Рис. 2. Кривая ошибок.

Распределение отступов. Понятие *отступа* (margin) $m(x_i)$ объекта x_i определено для алгоритмов, формирующих оценки принадлежности классам, $m(x_i) = p_{iy_i} - \max\limits_{y \neq y_i} p_{iy}$. В зависимости от значения $m(x_i)$ объекты разделяются на четыре типа: *шумовые* ($m \ll 0$), *пограничные* ($m \approx 0$), *неинформативные* ($m > 0$), *эталонные* ($m \gg 0$). В прикладных задачах такая типизация объектов имеет, как правило, самостоятельную ценность.

«Полигон» строит распределения отступов [5], усредняя их по разбиениям n. Отдельно строятся распределения отступов только по обучающим и только по контрольным объектам, что позволяет оценивать величину переобучения и число пограничных объектов в «зоне неуверенной классификации», рис. 3. Также строятся распределения отдельно по классам. По оси абсцисс откладываются объекты, упорядоченные по возрастанию среднего отступа; по оси ординат — значения отступов.

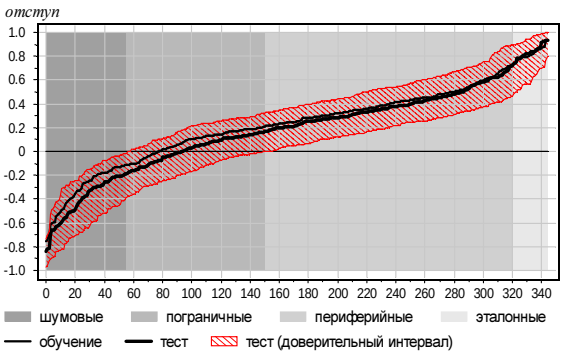


Рис. 3. Распределение отступов.

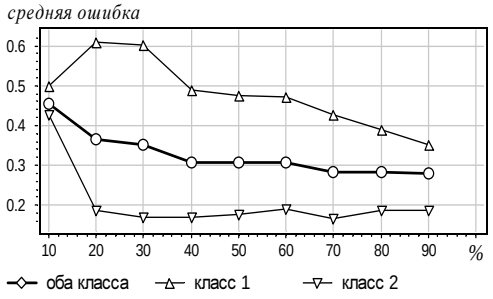


Рис. 4. Кривая обучения.

Кривая обучения (learning curve) в стандартном варианте — это зависимость средней частоты ошибок на контроле от длины обучающей выборки. «Полигон» дополняет стандартную методику построением кривых обучения отдельно по каждому классу с наложением их на одном графике, рис. 4. Здесь для класса 2 приемлемое качество классификации достигается, начиная с длины обучения 20% и далее сохраняется на постоянном уровне. Для класса 1 средняя ошибка снижается медленно, не успевая «выйти на насыщение». Отсюда можно сделать вывод, что для повышения качества классификации на объектах класса 1 (больные) необходимо существенно увеличить число обучающих объектов, причём только класса 1.

Архитектура системы

Распределённая система «Полигон» состоит из *Центрального Сервера* (ЦС), который хранит репозиторий задач, результаты тестирования, индивидуальные настройки пользователей и отчетов, и *Вычислительных Серверов* (ВС), обеспечивающих работу алгоритмов. ВС принимают от центрального сервера задания на решение задач классификации и возвращают результаты работы алгоритмов. Функции ВС может выполнять любой компьютер в сети Интернет, на котором установлена специальная программа — *Менеджер ВС*, осуществляющая запуск алгоритмов и обмен данными с ЦС. Эта программа предоставляется разработчиками «Полигона». На одном вычислительном сервере может работать несколько алгоритмов. Любой зарегистрированный пользователь может установить Менеджер ВС на свой компьютер, реализовать один или несколько алгоритмов и объявить свой ВС в системе «Полигон».

Начальный набор задач формируется из репозитория UCI [3] и других общедоступных источников данных. Пользователи имеют возможность загружать в систему свои задачи и устанавливать на них права доступа.

Один раз вычисленные результаты тестирования сохраняются в центральной базе данных «Полигона», и при повторном запросе выдаются без обращения к алгоритмам. Во большинстве случаев это позволяет получать отчёты очень быстро. Когда алгоритм обновляется, его сохранённые результаты стираются.

Добавление алгоритмов

Для разработчиков новых алгоритмов под платформой *.NET* предоставляется библиотека классов, содержащая основные структуры данных и базовый класс алгоритма, требующий написания двух функций — обучение и контроль. На сайте проекта имеются подробно документированные примеры реализации алгоритмов. Разработчикам алгоритмов на других языках программирования предоставляются специальные «обёртки» для реализации алгоритмов в виде *dll* (как *.NET*, так и *native*), исполняемых *exe*-файлов, web-сервисов, *matlab*-функций, и т. п. В частности, адаптация уже существующего алгоритма заключается в настройке или доработке одной из возможных «обёрток», исходный код которых находится в открытом доступе и подробно документирован.

Выводы

Методика тестирования, реализованная в системе «Полигон», существенно расширяет и унифицирует известные методы анализа качества классификации, основанные на скользящем контроле. Она позволяет не только констатировать тот или иной уровень ошибок, но и выявлять их причины, идентифицировать шумовые и пограничные объекты (независимо от типа алгоритма), целенаправленно подбирать алгоритм под конкретную задачу.

Литература

[1] *Ботов П. В.* Точные оценки вероятности переобучения для монотонных и унимодальных семейств алгоритмов // Всеросс. конф. ММРО-14 — М.: МАКС Пресс, 2009 — С. 7–10 (в настоящем сборнике).

[2] *Воронцов К. В., Инякин А. С., Лисица А. В.* Система эмпирического измерения качества алгоритмов классификации // Всеросс. конф. ММРО-13. — М.: МАКС Пресс, 2007. — С. 577–581.

[3] *Asuncion A., Newman D. J.* UCI Machine Learning Repository // University of California, Irvine. — 2007. `www.ics.uci.edu/~mlearn/MLRepository.html`.

[4] *Domingos P.* A unified bias-variance decomposition and its applications: ICML'17. — 2000. — Pp. 231–238.

[5] *Garg A., Roth D.* Margin distribution and learning algorithms: ICML'03. Washington, DC USA. — 2003. — Pp. 210–217.

[6] *Hand D., Till R.* A simple generalization of area under the ROC curve for multiple class classification problems // Machine Learning, 45. — 2001. — Pp. 171–186.

[7] *Langley P.* Crafting papers on machine learning // In Proceedings of ICML'2000. Pp.1207–1216.

[8] *Webb G. I.* MultiBoosting: A technique for combining boosting and wagging // Machine Learning. — 2000. — Vol. 40, No. 2. — Pp. 159–196.

Прогнозирование оттока абонентов телекоммуникационной компании как задача обучения распознаванию образов*

Гуз И. С., Татарчук А. И., Фрей А. И.
iguz@forecsys.ru, tatarchuk@forecsys.ru, frey@forecsys.ru
Москва, ЗАО «Форексис»

Рассматривается задача прогнозирования оттока абонентов телекоммуникационной компании, предлагается методология её решения и приводятся результаты экспериментального исследования на примере данных конкурса Teradata Center из Duke University.

Телекоммуникационная индустрия является одной из наиболее динамически развивающихся областей современной экономики. Условия жесточайшей конкуренции проявляются в ежегодном оттоке до 25% абонентов каждой компании. Вследствие этого традиционные проблемы организации взаимодействия с клиентами здесь ощущаются особо остро. Удержать абонента обходится компании, как правило, в 4–5 раз дешевле, чем привлечь нового, тогда как вернуть ушедшего абонента будет стоить уже в 50–100 раз дороже. Именно поэтому от решения задачи выявления абонентов, которые только собираются отказаться от услуг компании, напрямую зависят финансовые показатели каждого участника телекоммуникационного рынка.

Прогнозирование оттока абонентов

На отток абонентов может влиять огромное количество различных факторов: неудовлетворительное качество услуг связи, наличие более выгодных предложений со стороны других операторов, индивидуальные особенности абонентов и многие другие. При этом прогнозирование оттока было бы практически невозможным, если бы оно опиралось только на анализ причин, вызывающих отток, поскольку выявить все потенциальные причины оттока крайне затруднительно. Поэтому широкое распространение получили методы машинного обучения, позволяющие оценивать лояльность абонентов к компании в будущем по огромному количеству показателей их активности в прошлом, лишь косвенно отражающих истинные причины оттока.

Основные показатели активности абонентов обычно представлены в виде временных рядов, отражающих фактическое использование абонентом услуг мобильного оператора с различной детализацией. На основании этих рядов по специальным методикам, индивидуальным для каждого оператора, определяется факт ухода абонента, т. е. момент времени, с которого можно считать этого клиента потерянным для компании. Если условия методики выполняются для некоторого абонента на всем интервале достоверного ухода **C** (рис. 1), то по истечении этого времени абонент считается ушедшим на интервале **B**, если на интервале **A** он еще был активным. Таким образом, на некотором интервале времени **B** в прошлом всех абонентов можно разделить на «активных» и «уходящих». В то же время естественным представляется желание определять уход абонентов заблаговременно, когда еще возможно эффективно использовать маркетинговые воздействия по их удержанию. Это означает, что факт ухода на интервале **B** необходимо определять по состоянию активного, еще только «уходящего» абонента не менее чем за **A** и не более чем за **A** + **B** дней. Таким образом, задача прогнозирования оттока сводится к задаче классификации множества различных состояний абонентов на начальный момент времени интервала **A** на два класса — «активных» и «уходящих».

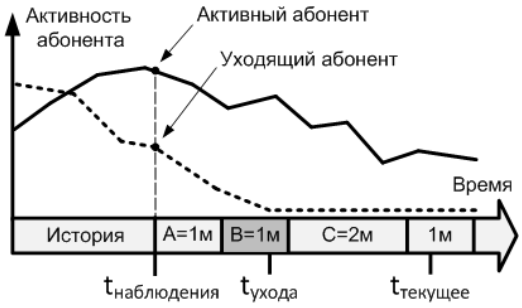


Рис. 1. Разделение абонентов на «активных» и «уходящих» (ориентировочные длины интервалов: 1м — 1 месяц, 2м — 2 месяца).

По доступным временным рядам формируется признаковое описание $F = \{f_1, \ldots, f_n\}$, $f_j \colon \mathbb{X} \to \mathbb{F}_j$ абонентов $x \in \mathbb{X}$, которое наиболее полно отражает их состояние на момент наблюдения, а также характеризует потенциальные причины оттока. Признаки могут измеряться в номинальных, порядковых, относительных или абсолютных шкалах. Количество анализируемых признаков n на практике варьируется от нескольких сотен до нескольких тысяч. Целевой признак принимает только два значения $y(x) \in \mathbb{Y} = \{0, 1\}$, соответствующих классам «активный» и «уходящий».

На начало интервала **A** (рис. 1) формируется обучающая выборка $(X, Y) = \{(x_i, y_i)\}_{i=1}^{N} \subset \mathbb{X} \times \mathbb{Y}$, в которую входят «активные» $y_i = 0$ и «уходящие»

*Работа поддержана РФФИ (проекты №08-07-00422, №07-07-00181, №08-01-12022-офи).

$y_i = 1$ абоненты. Для этих абонентов формируется признаковое описание в виде матрицы абонент-признак $\big[f_j(x_i)\big]_{i=1}^{N}{}_{j=1}^{n}$. Размер обучающей выборки N может достигать сотен тысяч, что накладывает дополнительные ограничения на средства хранения и обработки данных, а также на вычислительную сложность алгоритмов классификации.

Задача обучения состоит в построении алгоритма классификации — функции $\hat{y}\colon \mathbb{F}_1 \times \ldots \times \mathbb{F}_n \to \mathbb{Y}$, дающей прогноз ухода абонента по его признаковому описанию, сформированному на любой заданный момент времени.

Наряду с прогнозом ухода алгоритм должен оценивать вероятность ухода $\hat{p}(y \,|\, x),\ y \in \mathbb{Y}$. Эти вероятности используются при выделении сегмента наиболее склонных к оттоку абонентов для проведения маркетинговых воздействий по удержанию.

Сокращение признакового описания

В задаче прогнозирования оттока абонентов особенно остро встаёт проблема отбора и преобразования признаков. Использование всех признаков вычислительно проблематично, может приводит к переобучению и ухудшению качества прогнозов из-за наличия неинформативных и неточно измеренных признаков. Кроме того, наборы информативных признаков могут существенно отличаться для разных сегментов абонентов.

Методы отбора признаков принято разделять на три основные группы [1]. *Встроенные* (embedded) методы являются неотъемлемой частью алгоритмов обучения классификации. *Методы-оболочки* (wrappers) основаны на сокращенном эвристическом переборе подмножеств признаков и оценивании качества классификации на каждом подмножестве. *Методы фильтрации* (filters) оценивают полезность признаков ещё до обучения модели классификации. Основным достоинством последних является их вычислительная эффективность по сравнению с методами первых двух групп. В задачах классификации с большими объемами данных применение фильтрации практически неизбежно.

Одномерный отбор (univariate selection) основан на индивидуальном оценивании каждого признака $f \in F = \{f_1, \ldots, f_n\}$ по некоторому критерию информативности $J\colon F \to \mathbb{R}$. Все признаки ранжируются по убыванию значений критерия информативности $J(f_{j_1}) \geqslant \ldots \geqslant J(f_{j_n})$, а оптимальное подмножество однозначно определяется порогом минимальной информативности. На практике в качестве критерия $J\colon F \to \mathbb{R}$ применяются либо коэффициент корреляции с целевым признаком, либо различные статистические тесты, такие как критерий однородности Колмогорова-Смирнова или критерий согласия Пирсона [1, 2].

Направленный отбор (incremental selection), в отличие от одномерного, учитывает взаимозависимости признаков и реализует «жадную» стратегию выбора оптимального подмножества признаков $\tilde{F} \subseteq F = \{f_1, \ldots, f_n\}$. В простейшем случае организуется процесс поочерёдного добавления $\tilde{F}^{(k+1)} = \tilde{F}^{(k)} \cap f$ или удаления $\tilde{F}^{(k+1)} = \tilde{F}^{(k)} \setminus f$ признаков по критерию $J(\tilde{F}^{(k+1)}) \to \max\limits_{f \in F}$. Отличие разных методов заключается в выборе критерия информативности $J(\tilde{F})$ подмножеств признаков. Как правило, критерий основан на вычислении различных статистик, оценивающих разделимость классов, либо сразу по всему подмножеству признаков, например, критерий Фишера, либо по результатам попарного сравнения добавляемого признака с уже отобранными признаками [3].

Для сильно зависимых признаков методы фильтрации становится мало эффективными, поскольку такие признаки будут практически неразличимы по оценкам информативности. Подобная ситуация характерна для признаков, вычисленных по одним и тем же или сильно зависимым характеристикам абонентов. Однако простое игнорирование «дублирующих» признаков может привести к потере информации, существенной для дальнейшего анализа специфических сегментов абонентов.

Синтез признаков (Feature Extraction) — это формирование новых признаков как функций от исходных $f_j(x) = g_j(f_1(x), \ldots, f_n(x)),\ j = n, \ldots, n+n'$. Выбор функций g_j, возлагается на экспертов в предметной области. В задаче прогнозирования оттока абонентов имеет смысл брать отношения признаков $g(x) = f_j(x)/f_{j'}(x)$, полученных для одной и той же характеристики на разных интервалах определения.

В результате синтеза размерность признакового описания катастрофически увеличивается, что не позволяет использовать даже достаточно эффективную процедуру направленного отбора. В этом случае представляется разумным исключать из дальнейшего анализа заведомо малоинформативные признаки сначала по результатам одномерного анализа и только затем одним из методов направленного отбора.

Логические методы классификации

Одним из существенных требований, предъявляемым к моделям классификации абонентов является их интерпретируемость экспертами-маркетологами. Поэтому наибольшее распространение в мировой практике решения данной задачи получили логические алгоритмы — решающие деревья, решающие списки, взвешенное голосование логических закономерностей или бинаризованных исходных признаков, веса которых определяются, как правило, методом логистической регрессии.

Логическая закономерность $\varphi_y\colon X \to \{0, 1\}$ — это предикат, который выделяет $(\varphi_y(x) = 1)$ достаточно много объектов $x \in X$ класса $y \in \mathbb{Y}$ и прак-

тически не выделяет объекты других классов. Логические закономерности чаще всего ищут в виде конъюнкций $\varphi_y(x) = \bigwedge\limits_{j\in\omega} \beta_j(f_j(x))$ элементарных логических условий (термов) $\beta_j(f_j(x)) \in \{0,1\}$ над небольшим числом признаков. Обычно мощность набора признаков $|\omega|$ не превышает 3–7, иначе конъюнкция утрачивает интерпретируемость.

Для поиска наиболее информативных конъюнкций применяются эвристические методы отбора признаков из класса методов-оболочек (wrappers) и различные критерии информативности [4].

При наличии пропусков в данных часть термов в конъюнкции могут быть не определены, тогда считается, что правило не выделяет данный объект, $\varphi_y(x) = 0$. Таким образом, в логических алгоритмах легко обходится проблема пропусков как на этапе обучения, так и на этапе классификации новых объектов.

После того, как найдено достаточное количество закономерностей $\varphi^k_{y_k}(x)$, $y_k \in \mathbb{Y}$, $k = 1,\ldots,K$, выделяющих каждая свою часть объектов, они объединяются в композицию.

Решающий список (decision list). При классификации объекта $x \in X$ закономерности $\varphi^k_{y_k}(x)$, $k = 1,\ldots,m$ применяются последовательно до тех пор, пока одна из них не выделит объект, $\varphi^k_{y_k}(x) = 1$, и тогда x будет отнесен к классу y_k.

Оценка вероятности $\hat{p}(y_k \,|\, x)$ вычисляется как доля правильно классифицированных объектов обучающей выборки, покрытых этим правилом после удаления объектов, классифицированных предыдущими правилами. Объём данных позволяет достаточно надежно вычислять несмещенные оценки вероятностей по отложенной (контрольной) части обучающей выборки.

Взвешенное голосование (weighted voting) закономерностей представляет собой функцию вида $\hat{y}(x) = \arg\max\limits_{y\in\mathbb{Y}} \sum\limits_{k=1}^{K_y} \alpha^k_y \varphi^k_y(x)$, где K_y — число закономерностей класса y. Для настройки весов α^k_y используется модификация алгоритма бустинга [6]. Оценка вероятности $\hat{p}(y \,|\, x)$ вычисляется с помощью логистической функции:

$$\hat{p}(0 \,|\, x) = 1 - \hat{p}(1 \,|\, x) = \frac{1}{1 + \exp(aw(x) + b)},$$

где $w(x) = \sum\limits_{k=1}^{K_1} \alpha^k_1 \varphi^k_1(x) - \sum\limits_{k=1}^{K_0} \alpha^k_0 \varphi^k_0(x)$; параметры a и b подбираются путём калибровки Платта [5].

Логистическая регрессия (logistic regression) является другим подходом к настройке весов α^k_y в модели взвешенного голосования. В отличие от бустинга, где каждая закономерность $\varphi^k_y(x)$ и её вес α^k_y настраиваются поочерёдно так, чтобы компенсировать ошибки предыдущих, в логистической регрессии сначала строятся все закономерности, затем определяются их веса. Для этого используется итерационный метод наименьших квадратов (IRLS) на основе алгоритма Ньютона-Рафсона. Оценки вероятностей вычисляется с помощью аналогичной калибровки.

Оценивание качества прогнозов

Доля ошибок классификации (error rate) является стандартным способом оценивания качества алгоритмов классификации. Для задачи прогнозировании оттока характерна значительная несбалансированность классов (уходящих в 30–100 раз меньше, чем активных в каждый месяц). При этом наивный алгоритм классификации, относящий всех абонентов к активным, будет приводить к обманчиво невысокой доле ошибок. Кроме того, доля ошибок классификации никак не характеризует качество оценок вероятности $\hat{p}(y \,|\, x)$. Поэтому стандартный способ оценивания качества обучения практически не используется в задаче прогнозирования оттока абонентов.

Качество на критическом сегменте (top-decile lift)

определяет качество классификации 10% абонентов тестовой выборки с наибольшими оценками вероятности принадлежать к классу уходящих $\hat{p}(1 \,|\, x)$, поскольку именно эти абоненты представляют наибольший интерес для проведения маркетинговых кампаний по удержанию [7].

Показатель вычисляется как отношение доли $\hat{\pi}_{10\%}$ фактически ушедших абонентов на этом сегменте и доли $\hat{\pi}$ фактически ушедших абонентов на всей тестовой выборке

$$\mathrm{TDL} = \frac{\hat{\pi}_{10\%}}{\hat{\pi}}. \tag{1}$$

Чем больше величина TDL, тем выше качество прогнозирования на критическом сегменте.

Коэффициент Джини (Gini coefficient) оценивает качество классификации всей тестовой выборки, а не только критического сегмента [7].

Для абонентов тестовой выборки $(X', Y') = \{(x'_i, y'_i)\}_{i=1}^M \subset \mathbb{X} \times \mathbb{Y}$ вычисляется доля π_l абонентов с большей вероятностью ухода

$$\pi_l = \frac{1}{M} \sum_{i=1}^{M} \big[\hat{p}(1 \,|\, x'_i) > \hat{p}(1 \,|\, x'_l)\big],$$

а также доля фактически ушедших из них:

$$\pi^c_l = \frac{1}{M_c} \sum_{i=1}^{M} \big[\hat{p}(1 \,|\, x_i) > \hat{p}(1 \,|\, x_l)\big]\big[y_i = 1\big],$$

где M_c — число ушедших на тестовой выборке. Коэффициент Джини определяется выражением

$$\mathrm{GC} = \frac{2}{M} \sum_{i=1}^{M} (\pi^c_i - \pi_i) \leqslant 1 - \frac{M_c}{M}. \tag{2}$$

Таблица 1. Качество на критическом сегменте.

TDL Current/Future	Взвешенное голосование	Решающий список	Логистическая регрессия
Исходные признаки	1.71/1.84	1.44/1.54	0.91/1.14
Лучшие из исходных	1.85/1.92	1.54/1.79	1.53/1.58
Исходные и синтезир-е признаки	0.83/1.10	1.01/1.28	0.83/1.08
Лучшие из исходных и синтезир-х	2.34/2.13	2.41/**2.27**	**2.53**/2.24

Таблица 2. Коэффициент Джини.

Gini Current/Future	Взвешенное голосование	Решающий список	Логистическая регрессия
Исходные признаки	0.205/0.215	0.163/0.169	0.147/0.148
Лучшие из исходных	0.206/0.227	0.181/0.204	0.197/0.210
Исходные и синтезир-е признаки	0.196/0.201	0.158/0.167	0.156/0.144
Лучшие из исходных и синтезир-х	0.248/0.234	0.243/0.241	**0.29/0.292**

Экспериментальное исследование

Предложенный подход тестировался на данных конкурса Teradata Center из Duke University [8].

Данные состоят из трех непересекающихся выборок: Calibration (100 000 абонентов), Current (51 306 абонентов) и Future (100 462 абонентов). Число признаков $n = 170$. Обучение производилось по выборке Calibration, по остальным двум выборкам вычислялись оценки качества. Прогнозирование оттока на выборке Future является более сложной задачей, поскольку эта выборка сформирована на полгода позже, чем Current.

Обучение моделей проводилось по четырем наборам признаков. Первый содержит все исходные 170 признаков. Второй содержит только 15 «лучших» признаков по результатам фильтрации из 170 исходных признаков. Третий состоит из 170 исходных и 35 синтезированных признаков — попарных отношений сильно коррелированных исходных признаков. Четвертый набор сформирован в результате фильтрации 15 «лучших» из всех исходных и синтезированных признаков.

В таблице 1 приведено качество моделей на критическом сегменте (1). В каждой ячейке таблицы указаны показатели качества TDL на тестовых выборках Current/Future.

Выбор 15 «лучших» из исходных признаков повышает качество прогнозирования на тестовых выборках. Добавление к исходным признакам синтезированных приводит к переобучению и ухудшает качество всех моделей как на Current, так и на Future. Однако совместное применение предварительной фильтрации и синтеза новых признаков даёт наилучший результат. На выборке Current лидирует логистическая регрессия с TDL = 2.53, но более устойчивой во времени оказалась модель решающего списка, которая на выборке Future даёт TDL = 2.27.

В таблице 2 приведены значения коэффициента Джини (2). Здесь логистическая регрессия демонстрирует недосягаемое для других алгоритмов качество и устойчивость прогноза.

Тестовая выборка Current содержит 51 306 абонентов оператора мобильной связи, из которых 1.8% или 924 абонента, перестанут пользоваться услугами оператора в следующем месяце. Если случайно выбрать 10% абонентов выборки, то среди них будет отобрано около 92 уходящих абонентов. Логистическая модель, показавшая наилучший результат (Табл. 1), выделила среди 10% наиболее склонных к оттоку абонентов 233 из 924 абонентов фактически ушедших в следующем месяце.

Для оператора с миллионной абонентской базой, при ежемесячном оттоке около 1.8% абонентов, внедрение и использование средств по автоматическому мониторингу оттока абонентов позволит ежемесячно целенаправленно воздействовать на 2 754 уходящих абонентов больше, чем при прямой маркетинговой компании.

Литература

[1] *Duch W.* Filter methods // Feature extraction, foundations and applications 2006 — Pp. 89–118.

[2] *Biesiada J., Duch W.* Feature selection for high-dimensional data: a Kolmogorov-Smirnov correlation-based filter // 4th International Conference on Computer Recognition Systems 2005 — Pp. 95–104.

[3] *Pękalska E. et al.* Pairwise selection of features and prototypes // Advances in Soft Computing 2005 — Pp. 271–278.

[4] *Furnkranz J., Flach P. A.* Roc 'n' rule learning-towards a better understanding of covering algorithms // Machine Learning. — 2005. — Vol. 58, No. 1. — Pp. 39–77.

[5] *Niculescu-Mizil A., Caruana R.* Predicting good probabilities with supervised learning // 22nd Intern. Conf. on Machine Learning. — 2005. — Pp. 625–632.

[6] *Cohen W. W., Singer Y.* A simple, fast and effective rule learner // 16th National Conference on Artificial Intelligence. — 1999. — Pp. 335–342.

[7] *Hand D.* Construction and assessment of classification rules // Journal of Classification 2000 — Pp. 355–356.

[8] `http://www.salford-systems.com/churn.php` — The Duke/NCR Teradata Churn Modeling Tournament.

Сравнительный анализ применения нечетких дескрипторов при решении задачи «структура–активность» для выборки гликозидов*

Девятьяров Д. А., Кумсков М. И., Апрышко Г. Н., Носеевич Ф. М., Прохоров Е. И., Перевозников А. В., Пермяков Е. А.

kumskov@mail.ru

Москва, МГУ им. М. В. Ломоносова, мехмат, кафедра вычислительной математики;
Москва, Российский онкологический научный центр им. Н. Н. Блохина;
Москва, Институт органической химии им. Н. Д. Зелинского РАН

Проведены вычислительные эксперименты для задачи «структура–активность» с использованием дескрипторов, построенных на основе нечетких функций принадлежности. Были сформированы четкие и два типа нечетких дескрипторов для выборки гликозидов. Проведен анализ сформированных матриц «молекула-признак» рядом различных методов машинного обучения. Сравнение результатов прогноза для четких и нечетких дескрипторов показало, что при обработке матриц методом ANFIS нечеткие дескрипторы дают значительно лучшее качество прогноза, чем их четкие аналоги.

Данная работа посвящена анализу применения дескрипторов, основанных на нечетких функциях принадлежности, при решении задачи «структура–активность» [2]. Детальная постановка задачи приведена в [4]. В предыдущих работах для решения задачи были применены структурные трехмерные дескрипторы — пары и тройки особых точек, определенных на триангулированной молекулярной поверхности химического соединения. Был использован структурный символьный спектр молекулярного графа, представляющий собой число повторений молекулярных фрагментов в молекулярном графе путем полного перечисления всех пар, троек, четверок особых точек (ОТ) [2].

Однако описание с помощью подобных структурных дескрипторов имеет ряд недостатков, приведенных ниже.

1. Дескрипторы формируются на основе разбиения интервала значений расстояний между особыми точками (ОТ) на молекулярной поверхности, в результате чего описание с помощью структурных дескрипторов в значительной степени зависит от выбора данного разбиения. При этом неясно, каким образом возможна оптимизация выбора разбиения, так как значения дескрипторов не связаны непрерывно с выбором параметров — точек разбиений.
2. Значения структурных дескрипторов «разрывны» относительно параметров молекулярной поверхности: при непрерывном изменении координат ОТ значения дескрипторов не зависят непрерывно от этих аргументов и могут меняться только скачкообразно.
3. При моделировании биологической активности задача «структура–активность» осложняется тем, что молекулы могут менять свою конформацию (пространственную укладку) в пространстве. В результате при изменении конформации даже незначительное изменение взаимного расположения ОТ может привести к значительному изменению значений дескрипторов. Таким образом, классифицирующая функция, построенная на основе структурных дескрипторов, может работать ошибочно на относительно гибких молекулах.

В [5] было предложено решение вышеописанных проблем с помощью использования подхода нечеткой логики и применения нечетких дескрипторов.

В данной работе данный подход был применен к действительным данным — выборке гликозидов, протестированных на противоопухолевую активность.

Полученные матрицы «молекула–признак» были проанализированы с помощью различных методов машинного обучения. Подобные вычислительные эксперименты были проведены для двух видов нечетких дескрипторов, а также классических «четких» дескрипторов. Это позволило провести сравнительный анализ применения нечетких дескрипторов к базе химических соединений.

Этап формирования дескрипторов

Были сформированы 24 матрицы «молекула–признак», соответствующие разным наборам следующих параметров формирования дескрипторов: тип функции принадлежности; способ разбиения интервала электростатического заряда; количество интервалов разбиения расстояния между ОТ; количество интервалов разбиения расстояния между ОТ и парами ОТ.

В данной работе нас интересует зависимость качества прогноза от первого параметра — типа функции принадлежности, на основе которой формируются значения дескрипторов. Для определения дескрипторов необходимо задать нечеткие множества с функциями принадлежности $g_i(x)$, $0 \leqslant g_i(x) \leqslant 1,\ \ j = 1, \ldots, N$ на отрезке возможных

*Работа выполнена при финансовой поддержке РФФИ, проект № 07-07-00282.

расстояний между ОТ $x \in [0; d_{\max}]$, где $d_{\max}$ — максимальное значение из расстояний между ОТ для всех элементов обучающей выборки. Пример приведен на рис. 1.

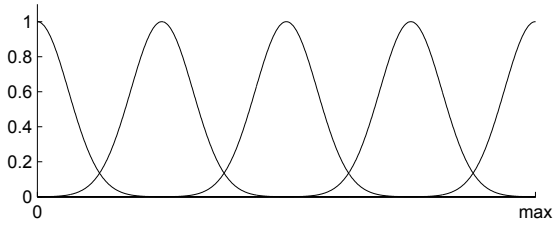


Рис. 1. Пример задания нечетких функций принадлежности расстояний.

При заданных функциях принадлежности можно определить значение нечеткого дескриптора для данного молекулярного графа. Для определения дескриптора, соответствующего нечеткому структурному фрагменту 2-ого порядка $\{L_i, L_j, G_k\}$, где L_i, L_j — метки ОТ, G_k — нечеткий интервал с функцией принадлежности $g_k(x)$, необходимо перечислить все неупорядоченные пары ОТ, встречающиеся в молекулярном графе. Для каждой такой пары определим степень сходства пары и структурного фрагмента $\{L_i, L_j, G_k\}$, равной $g_k(d)$, где d — расстояние между ОТ, если ОТ имеют метки L_i и L_j, и 0 в противном случае. Значение дескриптора определяется как сумма всех значений степени сходства данного структурного фрагмента и молекулярных фрагментов (пар ОТ), присутствующих в молекулярном графе конформации. Для определения значения дескриптора, соответствующего структурному фрагменту 3-его порядка $\{\{L_i, L_j, G_k\}, L_m, G_n\}$, где $\{L_i, L_j, G_k\}$ — нечеткий структурный фрагмент 2-ого порядка, L_m — метка третьей ОТ, G_n — нечеткий интервал с функцией принадлежности $g_n(x)$, необходимо перечислить все неупорядоченные тройки ОТ, встречающиеся в молекулярном графе. Для каждой такой тройки необходимо проверить, можно ли разбить ее на две такие группы F_1 и F_2 (состоящие из одной и двух особых точек соответственно), так, что степень сходства F_2 и $\{L_i, L_j, d_k\}$ положительна и $F_1 = L_m$. Если такое разбиение невозможно, то полагаем, что степень сходства тройки ОТ молекулярного графа и $\{\{L_i, L_j, G_k\}, L_m, G_n\}$ равна 0. В противном случае вычислим расстояние $d(F_1, F_2)$ между $F_1 = L_m$ и F_2 (под расстоянием здесь понимается наименьшее, наибольшее или среднее из всех расстояний между F_1 и каждой из особых точек F_2). Окончательно, положим степень сходства тройки ОТ молекулярного графа и $\{\{L_i, L_j, G_k\}, L_m, G_n\}$ равной произведению $g_k(d)g_n(d(F_1, F_2))$. Как и в случае дескрипторов 2-го порядка, значение «нечеткого» дескриптора 3-го порядка определяется как сумма всех значений степени сходства данного структурного фрагмента и молекулярных фрагментов (троек ОТ), присутствующих в молекулярном графе конформации. Аналогичным образом формируются значения «нечетких» дескрипторов более высокого порядка.

Ниже приведены функции принадлежности, примененные в данной работе: тип 1 соответствует четким дескрипторам, типы 2 и 3 — двум видам нечетких дескрипторов.

Четкие функции принадлежности. В случае четких дескрипторов функции принадлежности имеют следующий вид:

$$g_i(x) = \begin{cases} 1, & \text{если } d_{i-1} < x \leqslant d_i, \\ 0, & \text{в противном случае,} \end{cases}$$

где $d_0 = 0$, d_1, …, $d_N = d_{\max}$ — точки разбиения.

Кусочно-линейные трапециевидные функции принадлежности имеют следующий вид:

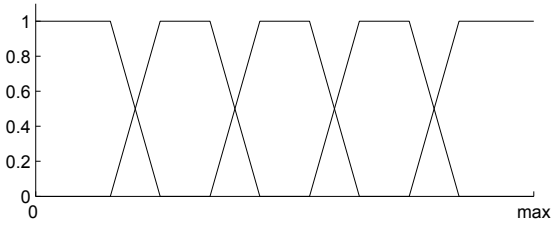


Рис. 2. Нечеткие трапециевидные функции принадлежности расстояний.

Кусочно-линейные треугольные функции принадлежности представляют собой вырожденный случай трапециевидных и наиболее отличны от четких функций принадлежности:

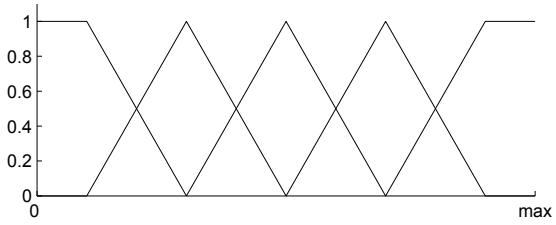


Рис. 3. Нечеткие треугольные функции принадлежности расстояний.

Число полученных дескрипторов варьировалось от 2100 до 3255 в зависимости от параметров формирования матрицы. Были сформированы матрицы для различного количества точек разбиения интервала расстояний между ОТ и интервала расстояний между ОТ и парой ОТ. Точки разбиения были выбраны равномерно на $[0; d_{\max}]$.

Этап анализа матрицы

Полученные 24 матрицы были обработаны различными методами машинного обучения, и для каждого метода было проведено сравнение качества предсказаний для четких и нечетких дескрипторов. Все методы применены в режиме скользящего контроля (leave-one-out cross validation), так как размер выборки не позволяет разделение на обучающую и тестовую выборку. Таким образом, качество прогноза вычислялось как R^2_{cv} [4].

Были реализованы следующие методы.

МГУА. Были реализованы деревья решений с применением метода группового учета аргументов (МГУА) [1]. В качестве опорных функций использовались линейные комбинации дескрипторов. Предварительно для получения деревьев решения был применен иерархический агломеративный метод кластерного анализа с евклидовой метрикой на отобранном множестве дескрипторов и центроидальным методом объединения кластеров. Множество дескрипторов для определения метрики кластерного анализа было отобрано с помощью МГУА, примененного ко всех выборке. Молекулы, не вошедшие в кластеры, были отнесены к выбросам. Применение подобного алгоритма к задаче «структура–активность» изложено в [3].

МГУА на конъюнкциях / дизъюнкциях. Действительные значения дескрипторов были преобразованы в бинарные данные. Полученное множество было разбито на кластеры агломеративным методом кластерного анализа. К бинарным матрицам «молекула– признак» для каждого кластера был применен МГУА, использующий в качестве опорных функций конъюнкции ряда дескрипторов и отдельно от них дизъюнкции ряда дескрипторов. Когда добавление очередного дескриптора не давало улучшения прогноза, алгоритм МГУА останавливался.

Предварительная бинаризация каждого столбца также проходила с помощью иерархического кластерного анализа. Значения дескрипторов в каждом столбце разбивались на заданное число кластеров так, что сумма числа элементов в двух самых больших была больше (а разность между ними — меньше) некоторого процента от общего количества элементов. Элементы остальных кластеров были распределены по наименьшему евклидову расстоянию до центров выделенных кластеров.

Оптимальные типы метрики и меры сходства для кластерного анализа были подобраны отдельно.

ANFIS на главных компонентах. С помощью SVD-разложения (singular value decomposition), были выделены главные компоненты матрицы «молекула–признак». Далее был применена система нечёткого логического вывода ANFIS (Adaptive Network-based Fuzzy Inference System) [6] к определенному числу первых (по модулю собственных значений) главных компонент. Оптимальные параметры, а также количество задействованных главных компонент были подобраны. В частности, рассматривалось такое количество главных компонент, что при добавлении очередной компоненты качества прогноза на скользящем контроле R^2_{cv} не улучшалось.

МГУА-kNN. Применен алгоритм МГУА [1] с использованием метода k ближайших соседей (kNN). В качестве опорных функций использован следующий вариант метода kNN: для нового объекта предсказываем класс большего числа объектов, находящихся на расстоянии, не превышающем радиус облака всех точек $R = \min\limits_{Y\in M}\max\limits_{X\in M} d(X,Y)$, где d — евклидово расстояние, заданное на векторе определенных дескрипторов, M — множество всех объектов выборки. Итерации по добавлению нового дескриптора прекращались, когда при добавлении очередного дескриптора качество переставало улучшаться.

Данный алгоритм может не выдать прогноз для молекулы в следующих случаях:

— молекула опознается как выброс (в случае, если у молекулы на расстоянии радиуса графа нет других молекул в метрике, основанной на дескрипторах, отобранных для метода ближайших соседей);
— метод kNN не выдает однозначный класс объекта, так как среди ближайших соседей 2 класса имеют одинаковое максимальное число представителей.

Численные результаты и их анализ

В таблицах 1 и 2 приведены результаты применения ANFIS на главных компонентах и МГУА-kNN к выборке гликозидов. Показаны результаты только двух из вышеописанных методов из-за ограничений объема работы. В двух таблицах продемонстрированы разные показатели качества. Однако способ их вычисления незначительно отличается друг от друга, так как МГУА-kNN выдавал не более 2 неоднозначных предсказаний.

Сравнение всех 4 методов обработки матрицы показало, что метод МГУА-kNN выдает лучшее качество прогноза. Однако при этом не прослеживалось улучшений при применении нечетких дескрипторов. Возможно, это происходило по той причине, что обычные дескрипторы без использования преимуществ нечеткой логики уже выдавали высокое качество прогноза.

Преимущества применения нечетких дескрипторов были зафиксированы при обработке матриц методом ANFIS. Нечеткие дескрипторы в данном случае работают заметно лучше, чем класси-

Таблица 1. Результаты применения ANFIS на главных компонентах к выборке гликозидов (коэффициенты качества прогноза на скользящем контроле).

	Четкие дескрипторы	Нечеткие трапециевидные дескрипторы	Нечеткие треугольные дескрипторы
1	61,8%	67,1%	75,0%
2	67,1%	68,4%	69,7%
3	73,7%	69,7%	72,4%
4	65,8%	67,1%	67,1%
5	63,2%	71,1%	64,5%
6	71,1%	72,4%	76,3%
7	65,8%	69,7%	68,4%
8	68,4%	65,8%	69,7%

Таблица 2. Результаты применения МГУА-kNN к выборке гликозидов (доля верных предсказаний среди однозначных предсказаний на скользящем контроле).

	Четкие дескрипторы	Нечеткие трапециевидные дескрипторы	Нечеткие треугольные дескрипторы
1	92,1%	90,8%	89,5%
2	89,5%	90,8%	86,8%
3	89,5%	92,1%	93,4%
4	92,1%	90,8%	90,8%
5	92,1%	90,8%	89,5%
6	88,2%	94,7%	94,7%
7	93,4%	96,1%	96,1%
8	96,1%	92,1%	94,7%

ческие дескрипторы: фактически при всех комбинациях параметров построения дескрипторов значение функционала качества на нечетких дескрипторах превышало значение функционала качества на четких дескрипторах. При применении других методов улучшения прогнозирующей способности при переходе к нечетким дескрипторам не наблюдалось: четкие и нечеткие дескрипторы давали лучшие прогнозы относительно друг друга без особой закономерности. При этом среди примененных методов машинного обучения не нашлось такого, который давал очевидно лучшие предсказания на четких дескрипторах.

Выводы

В работе приведены результаты применения нечетких дескрипторов, предложенных в ранних работах авторов. Сравнение прогностической способности сформированных моделей для матриц четких и нечетких дескрипторов показало зависимость от метода машинного обучения, примененного при анализе матрицы «молекула–признак». Зафиксировано, что при обработке матриц методом ANFIS качество прогноза было значительно выше для нечетких дескрипторов. Однако подобное явление не наблюдалось при применении других методов. Сравнение показателей качества для разных методов машинного обучения показало, что применение МГУА-kNN приводит к значительно лучшему прогнозу.

Предлагаются следующие дальнейшие исследования в данной области.

1. Использовать другие наборы функции принадлежности, в частности, гладкие функции, такие, как Гауссова, и рассмотреть другие способы нахождения точек разбиения и количество точек разбиения; проанализировать результаты на новых типах дескрипторов.
2. В случае, если метод в целом дает плохие результаты прогноза (например, МГУА), следует, модифицировать метод, возможно, адаптировав его к конкретной задаче, и проводить эксперименты уже модифицированным методомю
3. Удалить выбросы при анализе матрицы. Дело в том, что один или два выброса, которые могли оказаться в выборке в результате неправильных результатов тестирования на активность, могут портить всю модель, в результате чего качество прогноза значительно ухудшается.

Литература

[1] *Ивахненко А. Г., Зайченко Ю. П., Дмитриев В. Д.* Принятие решений на основе самоорганизации — М.: Сов. Радио, 1976. — 220 с.

[2] *Кохов В. А.* Метод количественного определения сходства графов на основе структурных спектров // Известия РАН, Техническая Кибернетика. — 1994. — No. 5. — С. 143–159.

[3] *Кумсков М. И., Митюшев Д. Ф.* Применение метода группового учета аргументов для построения коллективных оценок свойств органических соединений на основе индуктивного перебора их «структурных спектров» // Проблемы управления и информатики. — 1996. — No. 4. — С. 127–149.

[4] *Прохоров Е. И., Перевозников А. В., Воропаев И. Д., Кумсков М. И., Пономарёва Л. А.* Поиск представления молекул и методы прогнозирования активности в задаче «структура–свойство» // Всеросс. конф. ММРО-14. — М.: МАКС Пресс, 2009. — С. 589–591.

[5] *Devetyarov D. A., Zaharov A. M., Kumskov M. I., Ponomareva L. A.* Fuzzy logic application for construction of 3D descriptors of molecules in QSAR problem // 8th Intern. Conf. «Pattern Recognition and Image Analysis: New Information Technologies», Yoshkar-Ola. — 2007. — Vol. 2. — Pp. 249–252.

[6] *Roger Jang J.-S.* ANFIS: Adaptive-Network-Based Fuzzy Inference Systems // IEEE Transactions on Systems, Man, and Cybernetics. — May 1993. — Vol. 23, No. 03. — Pp. 665–685.

Разработка автоматизированной технологии распознавания трехмерных дефектов в композитных конструкциях по тепловизионным изображениям*

Димитриенко Ю. И., Краснов И. К., Николаев А. А.
Dimit@serv.bmstu.ru, A.A.Nikolaev@yandex.ru
Москва, Московский государственный технический университет им. Н. Э. Баумана

В статье рассматриваются разработанные автоматизированная технология и программно-математический комплекс «TSHCSS3D» распознавания трехмерных дефектов и их форм в изотропных и композитных конструкциях по изображениям, полученным тепловым неразрушающим контролем.

В настоящее время существует достаточное количество методов неразрушающего контроля (НК), на основе которых созданы и создаются технические системы, реализующие диагностирование — распознавание дефектов (определение наличия, расположения и геометрических параметров). В таких системах распознавание дефектов производится либо оператором «на глаз», либо экспертными системами машинного зрения. При распознавании дефектов в различных диагностируемых объектах оператору по изображению с дефектами (зашумленному и часто имеющему недостаточную к фону контрастность, четкость) бывает затруднительно принять правильное решение. Системы машинного зрения, основанные на традиционном подходе обработки каждой точки изображения и имеющие более низкую зрительную эффективность по сравнению с человеком, «захлебываются», обрабатывая большие объемы данных, и выдают, часто неоднозначные, результаты. При такой ситуации хорошим результатом диагностирования — распознавания считается определение наличия дефектов и их плоскостных геометрических характеристик. Вопрос об определении трехмерной формы дефектов по данным, полученным быстрыми недорогими методами НК, не стоит. Для композитных элементов конструкций ситуация с определением трехмерной формы дефектов осложняется неоднородностью структуры и свойств материала. В разработанной авторами статьи автоматизированной технологии распознавания трехмерных дефектов в композитных конструкциях использован наиболее перспективный (а также: быстрый, простой, дешевый и безопасный) активный тепловой НК (ТНК). Предлагаемая технология основывается на обработке входных тепловизионных изображений методами теории контурного анализа, многократном численном моделировании процесса ТНК, решении задач построения трехмерных геометрических моделей (ГМ) и распознавания форм трехмерных дефектов.

Постановка общей задачи распознавания трехмерных дефектов

Задача распознавания трехмерных дефектов состоит из двух этапов:

1) построение трехмерных ГМ дефектов на фоне ГМ объекта контроля;
2) автоматическая классификации полученной трехмерной ГМ (определение типа дефекта).

Исходными данными для распознавания трехмерных дефектов являются:

— тепловизионное изображение (достаточно одного), представляющее собой матрицу, элементами которой являются значения температуры на поверхности объекта контроля;
— информация об условиях проведения ТНК;
— данные о геометрических параметрах объекта контроля (трехмерная геометрическая модель);
— информация о тепловых эффективных характеристиках объекта контроля.

По этим данным необходимо определить наличие или отсутствие дефектов, построить трехмерную ГМ и определить тип (форму) дефекта.

Задача первого этапа является задачей распознавания изображений [1], и решается на основе двух разработанных методик: методики определения плоскостных геометрических параметров дефектов и методики предварительного распознавания формы трехмерных дефектов.

Задача второго этапа решается на основе разработанной методики распознавания формы трехмерных дефектов.

Определение плоскостных геометрических характеристик дефектов

Данная методика основывается на решении задач подавления шумов (использованы медианные фильтры) и задач теории контурного анализа. Из одного исходного зашумленного тепловизионного матричного изображения производится выделение системы контуров $\{\Gamma_{(j)}\}$, $\Gamma_{(j)} = \{\gamma_{(j)}(n)\}_{j=0}^{k-1}$ (реализованы алгоритмы «жука» и Розенфельда). Данная система контуров образует контурный каркас и непрерывный контурный «скелет» из геометрических центров контуров, указывающий на коли-

*Работа выполнена при финансовой поддержке РФФИ, проекты № 07-08-00574-а и № 09-08-00323-а.

чество предполагаемых дефектов (при этом автоматически решается задача сегментации дефектов в плане) и изменение формы дефектов по толщине объекта контроля. Далее по всей структуре контурного каркаса производится вычисление площадей и распознавание зашумленных контуров [2, 3] соседних уровней (строятся контурные согласованные фильтры [2]). Распознавание контуров соседних уровней контурного каркаса позволяет из всего множества системы контуров выделить по одному контуру для каждого предполагаемого дефекта. Каждый такой контур несет данные о геометрическом центре дефекта, ориентации, приблизительной форме (т. е. об искомых плоскостных геометрических характеристиках).

Предварительное распознавание формы трехмерных дефектов

Решение задачи предварительного распознавания формы трехмерных дефектов в композитных элементах конструкций основано на многократном численном моделировании тепловых процессов в объекте контроля при наложении граничных условий ТНК. При этом накладывается ограничение на распознавание только одного дефекта. Масштабируемая информация о плоскостных геометрических характеристиках дефектов и информация о толщине объекта контроля образуют трехмерное пространство пробных дефектов, где осями являются масштабы контура дефекта $|\mu|$, толщины z и глубины h залегания от поверхности. Для данного пространства задача распознавания формы ставится как задача минимизации функционала относительной ошибки $\bar{H}(D_m(|\mu_k|, z_l, h_q)) \to \to \min$. Относительная ошибка пробного дефекта вычисляется как $\bar{H}_{D_m} = \sum\limits_i \sum\limits_j \left(1 - \frac{\theta^{ij}_{NMI}}{\theta^{ij}_{TI}}\right)^2$, где θ^{ij}_{NMI} — значение элемента матрицы температурного поля, полученного численным моделированием ТНК, θ^{ij}_{TI} — значение элемента матрицы тепловизионного изображения. Матрица θ^{ij}_{NMI} вычисляется решением трехмерной задачи нестационарной теплопроводности (методом конечных элементов (КЭ)) с полной системой уравнений вида

$$
\begin{cases}
\rho(M)c_v(M)\frac{\partial\theta(M,t)}{\partial t} = \nabla\cdot\big(\underline{\Lambda}(M)\cdot\vec{\nabla}\theta(M,t)\big), \\
\qquad\qquad M\in\Omega,\ t\in(t_0,t_{\max}); \\
\theta(M,t)\big|_{t=0} = \theta_0(M), \qquad M\in\Omega\cup\Sigma; \\
\theta(M,t)\big|_{M\in S_d} = \theta_d(M,t), \quad t\geqslant t_0, \\
-\boldsymbol{n}\cdot\underline{\Lambda}(M)\cdot\vec{\nabla}\theta(M,t)\big|_{M\in S_p} = p(M,t)\big|_{M\in S_p},\ t\geqslant t_0, \\
-\boldsymbol{n}\cdot\underline{\Lambda}(M)\cdot\vec{\nabla}\theta(M,t)\big|_{M\in S_c} = \\
\qquad = \alpha(M,t)\big(\theta(M,t)-\theta_c(M)\big)\big|_{M\in S_c},\ t\geqslant t_0, \\
-\boldsymbol{n}\cdot\underline{\Lambda}(M)\cdot\vec{\nabla}\theta(M,t)\big|_{M\in S_{\mathrm{Def}}} = \\
\quad = \sigma_0\varepsilon\big(\theta^4(M,t)-\theta^4(M_1,t)\big)\big|_{M,M_1\in S_{\mathrm{Def}}},\ t\geqslant t_0,
\end{cases}
$$

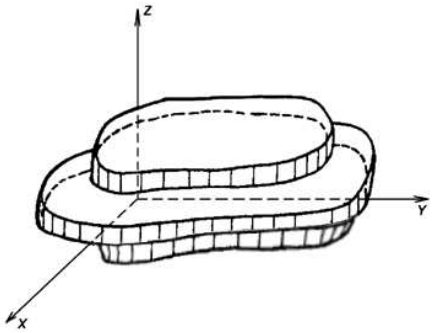


Рис. 1. Набор криволинейных цилиндров, описывающий трехмерную модель дефекта.

где ρ — плотность; M, M_1 — материальные точки области Ω; c_v — теплоемкость; $\underline{\Lambda}$ — симметричный тензор теплопроводности; $\vec{\nabla}\theta$ — градиент температуры; θ_0 — температура по всему телу в момент времени t_0 (начальные условия); θ_d — температура на поверхности тела (граничное условие первого рода); p — тепловая нагрузка; α — коэффициент теплоотдачи; θ_c — температура среды, в которую происходит конвективный теплообмен; σ_0 — постоянная Стефана–Больцмана; ε — поглощательная способность поверхности дефекта; S_d, S_p, S_c — участки поверхности тела с граничными условиями постоянной температуры, теплового потока и конвективного теплообмена; S_{Def} — внутренняя поверхность дефекта.

Минимизация функционала относительной ошибки $\bar{H}(D_m) \to \min$ производится применением метода Хука-Дживса, при этом в качестве начального приближения задается пробный дефект с минимальными параметрами $|\mu_k|, z_l, h_q$.

Результатом предварительного распознавания формы трехмерных дефектов является «оптимальный» пробный дефект, дающий при численном моделировании ТНК температурное поле поверхности контроля, максимально близкое к снятому тепловизионному полю. Распознанный при этом дефект описывается областью, образуемой набором конечных элементов (тетраэдрами). Форма дефекта D_i описывается набором непересекающихся криволинейных цилиндров $\{\Phi^i_j\}$, образуя при этом трехмерную проволочную ГМ дефекта (рис. 1).

Задача распознавания формы трехмерных дефектов

Получение трехмерной ГМ дефекта позволяет произвести непосредственное распознавание формы дефекта. Под распознаванием формы дефекта в разработанной автоматизированной технологии прежде всего понимается определение типа (класса) и подтипа (подкласса) дефекта (например, таких классов: внутренняя раковина, внутреннее расслоение, внутренняя трещина, поверхностная трещина; именно тип дефекта и его пространственные

характеристики являются исходными данными для дальнейшего решения задач определения текущей опасности и прогнозирования долговечности исследуемой конструкции).

Распознавание формы дефектов производится следующим образом. Ставится задача распознавания объекта (дефекта) D_i с ГМ $\{\Phi_j^i\}$ из множества классов (типов) числом N, которая решается следующим образом. Основание каждого криволинейного цилиндра Φ_j^i, входящего в ГМ $\{\Phi_j^i\}$, сравнивается с набором оснований эталонных моделей (производится распознавание контуров [2,3] на основе теории контурного анализа (строятся контурные согласованные фильтры)) и определяется наиболее вероятная ГМ. Далее производится вычисление отношения высот криволинейных цилиндров к ширинам оснований. На основе полученных данных производится автоматическое отнесение дефекта к одному из подтипов (типов).

Программно-математический комплекс «TSHCSS3D», пример работы

Рассмотренная в данной статье концепция автоматизированной технологии распознавания трехмерных дефектов легла в основу разработанного в 2008 г. программно-математического комплекса «TSHCSS3D». Данный программно-математический комплекс (ПМК) полностью основывается на собственных разработках авторов и реализован в «MS Visual Studio 2005». В качестве исходной информации в ПМК используются файлы с тепловизионным изображением и с трехмерной геометрией объекта контроля. Разработан удобный диалоговый интерфейс, через который вводятся остальные параметры задачи распознавания дефектов.

В 2008–2009 гг. ПМК «TSHCSS3D» прошел стадии отладки и тестирования, в том числе и на изображениях реальных композитных оболочек с дефектами, полученных на ведущих предприятиях в области производства и диагностики композитных материалов (ОАО «Центральный НИИ Специального машиностроения» и ОАО «Технологический институт ВЕМО»). Ниже приведен пример распознавания дефекта в многослойной композитной пластине.

Диагностируется композитная пластина (материал стеклопластик, размеры: 250×70×9 мм) с искусственно созданным внутренним дефектом в форме сплюснутого цилиндра (радиус 12,5 мм, глубина залегания 2,5 мм, толщина 0,3 мм; образец создан в лаборатории НК ОАО «ЦНИИ СМ» и является характерным для демонстрации ТНК). Тепловизором получено изображение температурного поля поверхности при ТНК (рис. 2). Применив к данному изображению программный модуль ПМК «TSHCSS3D», реализующий методику определения плоскостных геометрических характеристик дефектов, получаем следующую систему контуров (рис. 3) и предварительную форму дефекта в плане (окружность). Применив методику предварительного распознавания формы трехмерных дефектов к имеющейся КЭ модели части диагностируемой пластины (рис. 4), получаем для поля пробных дефектов (рис. 5) следующие решения H_D (рис. 6, 7) и «оптимальный» пробный дефект, имеющий следующий вид (рис. 8). Применив к найденной дефектной области методики распознавания формы трехмерных дефектов, позволяет отнести данный дефект к внутренним расслоениям со спюснутой цилиндрической формой (окончательные размеры: радиус 12,97 мм, толщина 0,5 мм;

Рис. 2. Температурное поле поверхности пластины с дефектами, полученное с тепловизора.

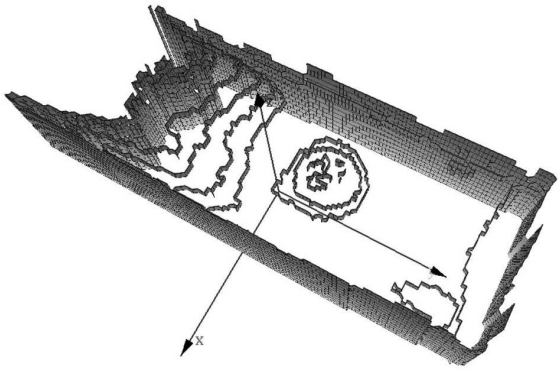


Рис. 3. Система контуров для исследуемой области с дефектом.

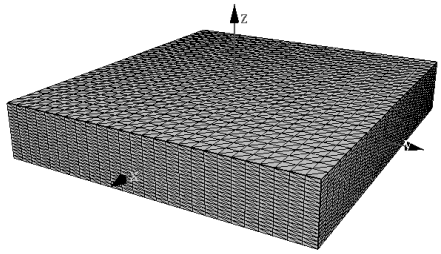


Рис. 4. Расчетная КЭ сетка (78624 КЭ).

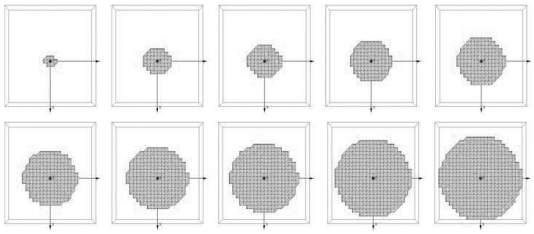


Рис. 5. Поле пробных дефектов (в плане).

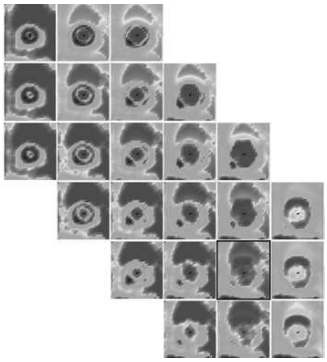


Рис. 6. Расчитанные решения H_D для поля пробных дефектов толщиной 0,5 мм.

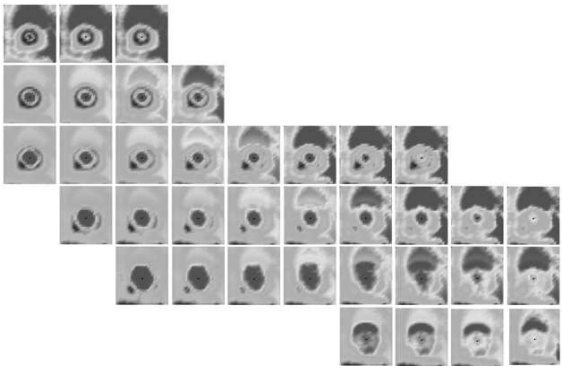


Рис. 7. Расчитанные решения H_D для поля пробных дефектов толщиной 1 мм.

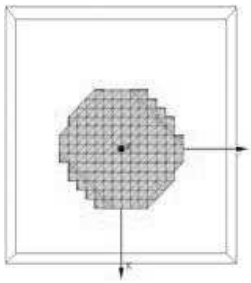


Рис. 8. Распознанный «оптимальный» пробный дефект (толщина 0,5 мм, описывающий радиус 14,5 мм и глубина залегания 2,5 мм).

глубина залегания 2,5 мм). Произведем оценку вычислительной эффективности разработанной технологии применительно к данному примеру. Рассмотрим наиболее времязатратный этап — методику предварительного распознавания формы дефектов. Общее количество пробных дефектов для построенной КЭ сетки равно 1360. «Оптимальный» дефект получен решением для 60 пробных дефектов, что составляет 4,4% от общего количества пробных дефектов и является хорошим показателем вычислительной эффективности разработанной методики при данном уровне точности.

Выводы

Впервые созданная на основе теории распознавания образов и теории механики сплошной среды и не имеющая аналогов автоматизированная технология распознавания трехмерных дефектов в композитных конструкциях по тепловым изображениям позволяет получать трехмерную форму и тип дефектов с высокой точностью, что подтверждается испытаниями на реальных композитных образцах.

Дальнейшая разработка данной автоматизированной технологии распознавания трехмерных дефектов будет проводиться путем добавления к существующим задачам задачи сегментации (разрешения) дефектов.

Литература

[1] *Журавлев Ю. И., Гуревич И. Б.* Распознавание образов и распознавание изображений // Распознавание, классификация, прогноз. — 1989. — Т. 2, № 5. — С. 5–73.

[2] *Фурман Я. А.* Введение в контурный анализ. Приложения к обработке изображений и сигналов. — М.: Физматлит, 2003. — 592 с.

[3] *Николаев А. А.* Распознавание неоднородностей, определение их геометрических характеристик и построение 3D геометрических моделей в задачах неразрушающего контроля // Всеросс. конф. ММРО-13. — М.: МАКС Пресс, 2007.— С. 506–508.

Система обработки эндоскопических изображений, реализующая возможность количественных измерений линейных размеров*

Дулькин Л. М., Салахутдинов В. К., Алёхин А. И., Дорошенко Д.
vsalakhutdinov@gmail.com
Москва, Центральная клиническая больница РАН; НИИСИ РАН

Рассмотрено решение задачи количественного измерения линейных размеров объектов эндоскопических медицинских исследований. Показано, что предложенные методы и средства позволяют увеличить точность измерений размеров более чем на порядок и довести ее до величин, приемлемых для адекватной диагностики.

В настоящее время основным, а в ряде случаев и единственным методом неинвазивной визуализации внутренних органов и диагностики широкого спектра заболеваний является эндоскопия [1]. В обследуемую полость организма вводится малогабаритная телевизионная система, с помощью которой обеспечивается визуализация и телевизионная трансляция изображения обследуемых объектов [2]. При этом для выбора оптимальной стратегии лечения крайне важно не только обнаружить сам факт патологических изменений, но и адекватно оценить их масштаб. Как правило [3], при малых размерах патологий оказывается достаточно терапевтического лечения, лечение патологий средних размерах требует эндоскопического лечения, а при крупных размерах необходимо полостное вмешательство.

Проблема заключается в том, что в современных эндоскопических системах не измеряется расстояние до объекта, поэтому с помощью традиционно используемых в эндоскопии методов и средств невозможно соотнести видимые размеры объектов на изображении с их истинными размерами. Ошибка в оценке расстояния до объекта может привести к неверной оценке размера патологии и, как следствие, к выбору неверного метода лечения. Практика показывает, что погрешность количественных измерений размеров с помощью традиционно используемых в эндоскопии методов и средств в ряде случаев (в частности, при урологических обследованиях) в 3–5 раз выше, чем это требуется для адекватной выработки стратегии лечения [4].

В работе представлен подход к решению этой проблемы, основанный на масштабировании с помощью тестового изображения известных размеров. На изображение объекта проецируется световое пятно известного размера. Размеры проекции используются для определения масштабного коэффициента анализируемых патологий. Этот подход известен и нашел широкое применение [5, 6].

При практической реализации этого подхода возникают трудности, связанные с тем, что оптическая система эндоскопа имеет большие аберрации [7]. Кроме того, в биологических тканях происходит рассеяние света, которое приводит к размытию пятна и искажению масштабирования.

Предложена система построения и обработки изображений, позволяющая оценивать истинные размеры патологий по их эндоскопическим изображениям. В разработанной системе реализованы возможности оценки и устранения искажений размеров, вызванных рассеянием света.

Световолокно расположено в инструментальном канале эндоскопа так, что оптическая ось его коллиматора, решающего задачу преобразования расходящегося светового пучка в пучок с постоянным сечением, коллинеарна оптической оси телевизионной системы эндоскопа. С помощью персонального компьютера реализуется управление светодиодным источником света осветителя и лазером на входе световолокна, а также обработка изображения телевизионной камеры эндоскопа.

Для измерения размера объект помещается в районе центра изображения на мониторе. По сигналу измерения:

1. Регистрируется в оперативной памяти и отображается на мониторе в виде статического кадра текущее изображение, состоящее из расположенного вблизи центра измеряемого объекта.

2. По переднему фронту импульса кадровой синхронизации телевизионной камеры эндоскопа выключается светодиодный источник подсветки и включается лазерный источник на входе световолокна.

3. Регистрируется в оперативной памяти следующий телевизионный кадр, состоящий из изображения проекции на измеряемый объект коллимированного светового пучка лазера.

4. В результате обработки полученного на шаге 3 изображения количественно (в пикселях) измеряется сечение лазерного пучка.

5. На статическом изображении измеряемого объекта вручную отмечается линия, размер которой подлежит измерению. Длина этой линии в миллиметрах определяется как ее размер в пикселях, деленный на размер (в пикселях) сечения лазерного пучка, умноженный на заранее известное сечение лазерного пучка в миллиметрах.

*Работа поддержана грантами РФФИ №09-07-00309-а, №09-07-00444-а и №08-07-12089-офи.

Нетривиальность задачи обусловлена тем, что, в зависимости от характеристик тканей измеряемого объекта, лазерный пучок может проникать (а может и не проникать) в них на значительную глубину. Это может приводить к появлению на изображении ареола, интенсивность которого сравнима с интенсивностью основного пучка и, как следствие, к значительным погрешностям измерений. Медицинская практика показывает, что погрешность измерений, обусловленная рассеянием, может при работе с реальными изображениями в медицинской эндокопии доходить до 70

С целью повышения точности измерений лазерный пучок на выходе коллиматора формировался в виде кольца. При этом в процессе измерения размеров проекции лазерного пучка на измеряемый объект (шаг 4) измеренное значение интенсивности в центре использовалось как порог. Клинические исследования показали, что устранение влияния диффузного рассеяние в тканях позволяет снизить погрешность измерений примерно в три раза, что достаточно приемлемо в практической медицине.

Были проведены клинические исследования на лабораторных животных, которые показали, что предложенные методы и средства позволяют увеличить точность измерений размеров более чем на порядок, что значительно снижает долю неверных решений о выборе характера лечения.

Литература

[1] *Cotton P. B., Williams C. B.* Practical Gastrointestinal Endoscopy: The Fundamentals. Blackwell Science. — 2003. — 213 pp.

[2] *McPhee S. J., Papadakis M. A.* Current Medical Diagnosis and Treatment 2009. — McGraw Hill. 2009. — 1728 p.

[3] *Соколов В. А. и др.* Эндоскопическая тактика при трудноудаляемых конкрементах холедоха // Сб. 9-ый московский международный конгресс по эндоскопической хирургии (6–8 апреля 2005г.). — М. — 2005. — С. 353–355.

[4] *Дулькин Л. М., Салахутдинов В. К. и др.* Method of longitudinal stereoscopy for 3D visualization of endoscopic images // In proc. of Int. Symp. Topical Problems of Biophotonics-2007. — Pp. 1–30.

[5] *Хорн Б.* Зрение роботов. — М.: Мир, 1989. — 358 с.

[6] *Курикша А. А., Жулина Ю. В.* Обработка искаженных оптических изображений // Радиотехника и электроника. — 2000. — Т. 66, № 3. — С. 23–40.

[7] *Борн М., Вольф Э.* Основы оптики. — М.: Наука, 1970. — 856 с.

[8] *Дулькин Л. М. и др.* Классификации категорий сложности диагностической и лечебной эндоскопической ретроградной панкреатохолангиографии и степени риска развития осложнений // 6-ой Московский международный конгресс по эндоскопической хирургии. — М. — 2002.

Разработка реконструктивного метода обработки хронометрических данных*

Каримов М. Г., Магомедов М. А., Магомедов М. Г., Шамилова М. М.
karmaggas@mail.ru, mmagomedoff@gmail.com
Махачкала, Дагестанский Государственный Университет

В настоящей работе, используя математическое моделирование и вычислительный эксперимент, продемонстрирована эффективность модифицированного метода реконструкции функции по проекционным данным хронометрических измерений.

Задача восстановления функции распределения физической величины по различным проекциям возникает во многих областях науки и техники. Одно из научных направлений, занимающихся подобными задачами, известно как томография [1]. Томография — хорошо развитая область науки, представляющая большой практический интерес [2]. В данной работе она используется для восстановления истинной картины по реально измеряемым физическим образам, в том числе, по измеряемым в сканирующей туннельной микроскопии образам — сканам. Основная задача реконструктивной томографии, независимо от области приложения, сводится к проблеме восстановления неизвестной двумерной функции распределения $f(x,y)$ по её линейным интегралам — проекциям, известным как преобразование Радона, для конечного числа направлений. Преобразование Радона функции $f(x,y)$ вдоль прямой, заданной определённым углом φ, образованным с осью x, и расстоянием s от начала координат, производится по следующей формуле:

$$R(s,\varphi) = \iint\limits_{-\infty}^{\infty} f(x,y)\delta\big(x\cos\varphi + y\sin\varphi - s\big)dxdy. \quad (1)$$

Оно может быть видоизменено с учётом фильтрующего свойства δ-функции Дирака, а также переходя от системы координат $\{x,y\}$ к вращающейся системе координат $\{s,\tau\}$ с общим центром вращения. Используя формулы перехода:

$$\begin{cases} s = x\cos\varphi + y\sin\varphi; \\ \tau = -x\sin\varphi + y\cos\varphi; \end{cases}$$

преобразование Радона (1) может быть записано следующим образом:

$$R(s,\varphi) = \int\limits_{-\infty}^{\infty} f(s\cos\varphi - \tau\sin\varphi, s\sin\varphi + \tau\cos\varphi)d\tau.$$

Большинство существующих классических методов восстановления функции по проекциям основываются на применении обобщённой проекционной теоремы с использованием преобразования Фурье, которая устанавливает связь между Фурье-образами функции $f(x,y)$ и его преобразованием Радона $R(s,\varphi)$, и заключается в том, что одномерный Фурье-образ проекции при фиксированном угле φ есть сечение двумерного Фурье-образа функции, то есть

$$\hat{R}(\omega,\varphi) = \hat{f}(\omega\cos\varphi, \omega\sin\varphi), \quad (2)$$

где $\hat{R}(\omega,\varphi)$ — Фурье-образ преобразования Радона $R(s,\varphi)$ по переменной s, а $\hat{f}(\omega\cos\varphi,\omega\sin\varphi)$ — сечение Фурье-образа $\hat{f}(\omega_1,\omega_2)$. Используя (2), можно получить выражение для искомой функции:

$$f(x,y) = \int\limits_0^{\pi} d\varphi \int\limits_{-\infty}^{\infty} |\omega|\hat{R}(\omega,\varphi)\, e^{i2\pi\omega(x\cos\varphi + y\sin\varphi)}d\omega.$$

По сути это выражение есть решение интегрального уравнения Радона (1). Но в нашем случае представляет практический интерес получение решения для реальных физических задач измерения, в том числе и для хронометрических измерений с конечной точностью.

В настоящей работе проводится обсуждение задачи томографии, когда известны и доступны хронометрические данные («грубые изображения»). Предложен «хронотомографический» алгоритм восстановления искомой функции, известный как одноэтапный метод свертки с последующим обратным проецированием. Сделана попытка использовать очевидные и хорошо изученные преимущества классической радоновской томографии для решения задачи восстановления функции по хронометрически измеряемым с определённой точностью проекциям [3].

Математически, процесс хронометрических измерений физической величины можно формализовать и описать как интегральное преобразование функции $f(x,y)$ следующим образом:

$$R(s,\tau,\varphi) = \int\limits_{-\infty}^{\infty} f(s\cos\varphi - \tau'\sin\varphi, s\sin\varphi + \tau'\cos\varphi)\times \\ \times\, h(\tau-\tau')d\tau'. \quad (3)$$

Задача заключается в том, чтобы найти неизвестную функцию $f(x,y)$ по ее образам $R(s,\tau,\varphi)$,

*Работа выполнена при финансовой поддержке грантов РФФИ № 08-01-00802а и № 09-01-96508.

т. е. по ее измеряемым «грубым» изображениям. Функция $h(\tau)$ есть аппаратная функция хронометрического прибора, моделируемая функцией Гаусса:

$$h(\tau) = \frac{1}{\sigma\sqrt{\pi}} e^{\frac{-\tau^2}{\sigma^2}},$$

где $\sigma^2 = \frac{\Delta\tau^2}{\ln 2}$, а параметр $\Delta\tau$ характеризует точность измерений прибора. При $h(\tau) \equiv 1$ интегральное преобразование (3) есть обычное преобразование Радона, и в этом случае функция $f(x, y)$ восстанавливается единственным образом в классе суммируемых функций [1, 2].

Для случая $h(\tau) \neq 1$ интегральному преобразованию (3) с известной левой частью $R(s, \tau, \varphi)$ могут удовлетворять много различных функций $f(x, y)$. Вопрос нахождения функции $f(x, y)$ усложняется еще и тем, что функция $R(s, \tau, \varphi)$ реально бывает задана для конечного числа значений φ. Задачи такого типа возникают в приложениях, где требуется реконструкция (восстановление) функций, например, в томографии, по конечному числу направлений φ, по которым проводится сканирование. Поэтому всегда стоит вопрос о точности восстановления функции $f(x, y)$.

Учитывая, что $R(s, \tau, \varphi)$ — трехмерная функция, т. е. функция трех переменных, в то время как $f(x, y)$ есть двумерная функция, очевидно, что эта «избыточность» может быть использована разными путями для получения решения данной обратной задачи интегральной геометрии для гораздо худших условий «неполноты» данных, чем для традиционной задачи Радона.

Изучение реконструктивных возможностей данной задачи предполагает изучение свойств преобразования (3) для различных аппаратных функций, с целью выработки эффективных алгоритмов восстановления $f(x, y)$ по заданной функции $R(s, \tau, \varphi)$ и оценке точности алгоритмов, для внедрения в диагностическую технологию. В приложениях $f(x, y)$ является неизвестной функцией распределения физической величины, и задача заключается в нахождении такого алгоритма, который дает решение $f_0(x, y)$ уравнения (3), которое наиболее близко к искомому $f(x, y)$.

Метод восстановления функции $f(x, y)$ основывается на обратном проецировании предварительно фильтрованных проекций $I(s, \tau, \varphi)$:

$$f(x, y) = \pi\sigma^2 \int_0^{\pi} I(s, \tau, \varphi) d\varphi, \qquad (4)$$

где

$$I(s, \tau, \varphi) = \int_{-\infty}^{\infty} R(s', \tau, \varphi) h(s - s') Q(s - s') ds. \qquad (5)$$

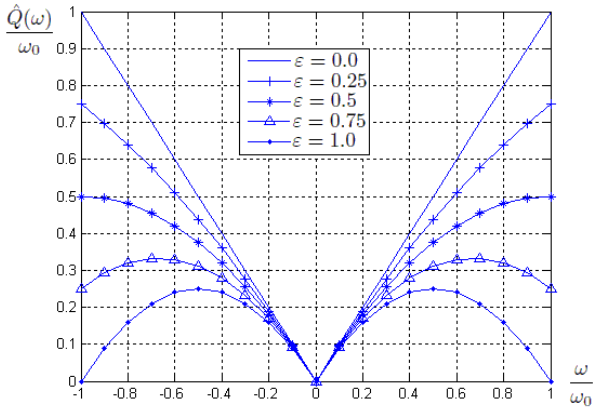


Рис. 1. Графики функции $\frac{\hat{Q}(\omega)}{\omega_0}$ при различных ε.

В последнем выражении (5) $Q(s)$ представляет собой функцию реконструкции со спектром $\hat{Q}(\omega) = |\omega|\left(1 - \varepsilon\frac{|\omega|}{\omega_0}\right)$. Параметр $\varepsilon \in [0, 1]$, как параметр регуляризации, предназначен в том числе для подавления высокочастотных артефактов, и в каждом конкретном случае ε подбирается с учётом спектральных особенностей изображения. Параметр ω_0 соответствует максимальной частоте спектра функции $f(x, y)$, используется для подбора шага выборки — дискретизации Δs и удовлетворяет условию Найквиста $\Delta s \leqslant \frac{1}{2\omega_0}$. При $\varepsilon = 0$, $\hat{Q}(\omega) = |\omega|$, что соответствует так называемому фильтру низких частот. На рис. 1 представлены графики приведённых спектров реконструирующей функции $\hat{Q}(\omega)$ для различных ε. Функция реконструкции $Q(s)$ определяется по формуле:

$$Q(s) = \int_{-\omega_0}^{\omega_0} |\omega|\left(1 - \varepsilon\frac{|\omega|}{\omega_0}\right) e^{2\pi i\omega s} d\omega. \qquad (6)$$

В процессе вычислений использовалась функция реконструкции, полученная по формуле (6) и имеющая удобный для табулирования вид:

$$Q(s) = \begin{cases} \left(1 - \frac{2}{3}\varepsilon\right), & s = 0; \\ 2\left(1 - \varepsilon\left(1 - \frac{2}{(2b)^2}\right)\right)\operatorname{sinc}(2b) + \\ \quad + (2\varepsilon - 1)\operatorname{sinc}^2(b) - \frac{4\varepsilon}{(2b)^2}, & s \neq 0; \end{cases} \qquad (7)$$

где $b = \pi\omega_0 s$, а $\operatorname{sinc}(x) = \frac{\sin(x)}{x}$.

Для демонстрации основных возможностей метода реконструкции изображений, предложенного в данной работе, проводилась серия вычислительных экспериментов над модельными изображениями размером 256×256. Сначала по данным изображениям формировались проекции $R(s, \tau, \varphi)$, для некоторого числа проекций n_φ. После чего каждая полученная проекция фильтровалась по схеме (5), используя функцию реконструкции (7) для

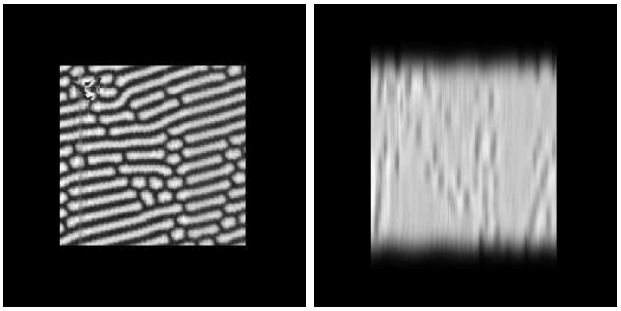


Рис. 2. Слева изображение $f(x,y)$, справа измеряемое изображение — скан $R(s,\tau,\varphi)$ для $\varphi = 0$ и ошибки измерения $\Delta\tau = 0{,}05$.

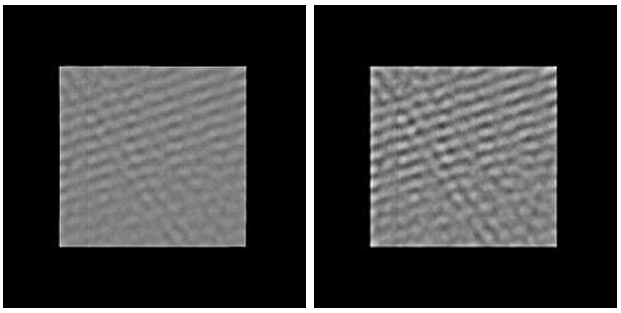


Рис. 3. Реконструкция изображений по радоновской томографии $R(s,\varphi)$ для 10 проекций ($n_\varphi = 10$). Слева: для $\varepsilon = 0$. Справа: для $\varepsilon = 1$.

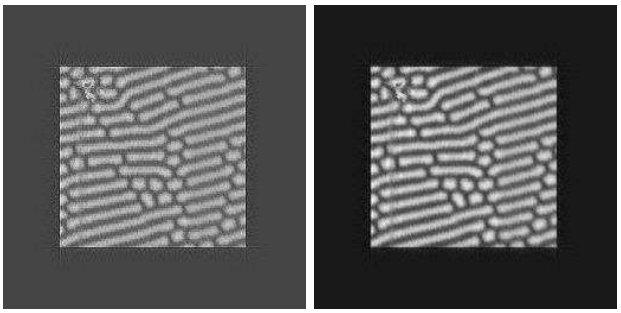


Рис. 4. Восстановленное изображение методом хронотомографии по 10 проекциям ($n_\varphi = 10$). Слева: $\varepsilon = 0$, параметр подобия $\Pi = 94{,}74$. Справа: $\varepsilon = 1$, $\Pi = 99{,}04$.

различных $\varepsilon \in [0,1]$, и над полученными проекциями проводилась операция обратного проецирования по формуле (4), для формирования реконструированного изображения. Полученный таким образом результат сравнивался с исходным изображением $f(x,y)$, используя параметр подобия $\Pi = \Pi(\varepsilon, n_\varphi)$, определяемого как:

$$\Pi(\varepsilon, n_\varphi) = 100\left[1 - \frac{1}{NM}\sum_{x,y}\bigl(f(x,y) - f_{\varepsilon,n_\varphi}(x,y)\bigr)^2\right],$$

где $f(x,y)$ — исходное изображение, $f_{\varepsilon,n_\varphi}(x,y)$ — восстановленное изображение по n_φ ракурсам для заданного ε, а N и M — соотвественно высота и ширина изображений. На рис. 2 показаны $f(x,y)$ и соотвествующий результат измерения данной функции для $\varphi = 0$. На рис. 3 представлены результаты реконструкции изображения радоновской томографии для числа проекций $n_\varphi = 10$, для параметров $\varepsilon = 0$ и $\varepsilon = 1$. На рис. 4 показаны результаты реконструкции изображения методом хронотомографии для числа проекций $n_\varphi = 10$ и для параметров $\varepsilon = 0$ и $\varepsilon = 1$.

Качественное сравнение результатов процесса восстановления функции, как показывают рис. 2 (справа) и рис. 4, указывает на то, что хронотомографический метод в обоих случаях предпочтительнее, чем хронометрический. С другой стороны анализ качества изображений приведенных на рис. 3 и рис. 4, позволяет утверждать, что хронотомографический метод восстанавливает лучше и эффективнее метода радоновской томографии когда число ракурсов ограниченно. Детальная количественная оценка эффективности хронотомографического метода и сравнение его с другими современными методами является предметом отдельных и более детальных исследований, которые проводятся и будут опубликованы.

Литература

[1] *Radon J.* Uber die Bestimmung von Funktionen durch ihre Interalwerte langs gewisser Mannigfaltigkeiten // Berichte Sachsische Akademie der Wissenschaften. — 1917. — v. 69, — p. 262-267

[2] *Наттерер Ф.* Математические аспекты компьютерной томографии. — М.: Мир, 1990.

[3] *Каримов М. Г.* Стохастическая корреляционная томография // ЖЭТФ. — 2000. — Т. 117, № 4. — С. 673–681.

Вычислительные методы обработки и интерпретации многоспектральных и гиперспектральных аэрокосмических изображений*

***Козодеров В. В.*[1], *Дмитриев Е. В.*[2], *Егоров В. Д.*[2]**
vkozod@mes.msu.ru
[1]Москва, МГУ им. М. В. Ломоносова, [2]Москва, Институт вычислительной математики РАН

Рассматривается задача классификации и восстановления параметров растительного покрова по данным аэрокосмических измерений. Предлагаются оригинальные процедуры, основанные на методах вычислительной математики, которые могут быть использованы при создании комплексного информационно-математического обеспечения космических систем дистанционного зондирования (ДЗ). Обработку данных ДЗ составляют два основных этапа: распознавание наблюдаемых объектов по их спектральным образам и оценка параметров, характеризующих состояние этих объектов. В основе процедур распознавания лежат логические правила принятия решений о принадлежности текущего элемента разрешения к тому или иному классу, использующие характерные спектральные признаки соответствующих объектов. Процедуры оценки параметров состояния растительного покрова реализуются на основе решения прямой задачи формирования интенсивности уходящего излучения, регистрируемого аппаратурой ДЗ, и обратной задачи восстановления указанных параметров.

Современные системы аэрокосмического дистанционного зондирования (ДЗ) позволяют получать данные измерений различного пространственного разрешения в форме многоспектральных и гиперспектральных изображений. Традиционные подходы к использованию географически привязанных данных ДЗ состоят в построении так называемых географических информационных систем (ГИС), которые объединяют базы данных различного назначения (рельеф, почвенный покров и др.). В основе ГИС-технологий лежит анализ пространственного распределения регистрируемых данных с точки зрения классификации наблюдаемых объектов природно-техногенной сферы. Программное обеспечение ГИС разрабатывается, в основном, зарубежными фирмами и содержит ряд вычислительных процедур обработки данных ДЗ, считающихся на сегодняшний день стандартными.

Среди этих процедур можно выделить: синтезирование данных различных спектральных каналов, которое производится с целью пространственной привязки различных объектов; кластерный анализ — разделение множества регистрируемых данных на классы без процесса обучения расчётного классификатора; анализ главных компонент — разложение регистрируемых данных как случайных функций для уменьшения их размерности. Основу практических приложений при реализации технологий ГИС с использованием данных ДЗ составляют стандартные преобразования изображений (географическая привязка, выделение характерных контуров, классификация объектов и др.), модели многофакторной регрессии и концепция «вегетационных индексов».

Несмотря на существенные достижения последних лет, методы обработки аэрокосмических изображений остались фактически неизменными. Используемые процедуры, основанные только на эмпирических методах исследования, служат в первую очередь для визуального дешифрирования обрабатываемых данных, и при получении количественных оценок они, как правило, не дают представления о реальной точности решения возникающих прикладных задач.

Предлагаемые методы и подходы

На протяжении последних лет мы развивали методы восстановления параметров растительного покрова по данным аэрокосмических измерений. В наших работах мы придерживались постановки данной задачи как обратной задачи переноса излучения в системе «земная поверхность — атмосфера». Созданные на данный момент процедуры включают в себя этапы распознавания образов наблюдаемых объектов по их многоспектральным (гиперспектральным) изображениям и обращения функционала интенсивности отражённого излучения, область значений которого соответствует данным ДЗ.

Поскольку оценка параметров, характеризующих состояние объектов класса «растительность», производится для каждого элемента разрешения обрабатываемых изображений, разрабатываемые процедуры идеальны для распараллеливания, что даёт возможность максимально эффективного использования современных многопроцессорных вычислительных систем. Разработанные программы реализованы с использованием технологии MPI. К настоящему моменту они успешно применялись для проведения массовой обработки данных ETM+ на многопроцессорных системах Межведомственного суперкомпьютерного центра РАН и Института вычислительной математики РАН.

*Работа выполнена при финансовой поддержке РФФИ, проекты №08-07-13515-офи_ц и №09-05-00171-а.

Определение контуров и классификация лесных массивов, лугов, болот, водоёмов, вместе с объектами энергетики, промышленности, сельского хозяйства, населенными пунктами и дорожно-транспортной сети возможно благодаря тому, что соответствующие им характеристики собственного и отражённого электромагнитного излучения определённым образом упорядочены. Благодаря пространственной упорядоченности интенсивности излучения можно распознать объекты по их характерной форме. Однако в наших подходах используется спектральная упорядоченность, которая позволяет классифицировать каждый отдельный пиксель аэрокосмического изображения.

Спектральные характеристики разнородных поверхностей, вообще говоря, не совпадают. Для построения алгоритма распознавания определяется функция отклика, характерная для заданного типа объектов. Общая схема интерпретации как многоспектральных, так и гиперспектральных аэрокосмических изображений, включает следующие основные этапы: классификация изображений и сегментация объектов с близкими спектральными свойствами; выделение типичных объектов со специфическими характеристиками регистрируемого излучения; построение моделей формирования внутренней структуры отдельных элементов разрешения; привлечение дополнительной информации по интерпретации объектов смешанного типа.

Число измерительных каналов аппаратуры многоспектрального зондирования обычно не превышает десяти, а в случае гиперспектрального зондирования используются данные нескольких сотен спектральных каналов. В системах гиперспектрального ДЗ разрешение по спектру достигает величин порядка одного нанометра. При высоком спектральном разрешении в исходных данных проявляются отдельные линии поглощения излучения геологическими минералами, почвенными образованиями, искусственными материалами (крыши зданий, асфальто-бетонные покрытия) и атмосферной средой (молекулярный кислород, озон, водяной пар). Для таких высокомолекулярных соединений, как хлорофилл — основной пигмент фитоэлементов (листья и хвоя деревьев), заметны целые полосы поглощения. Таким образом, использование гиперспектральных данных открывает новые возможности для обнаружения объектов со специфическими свойствами по данным аэрокосмического ДЗ на основе анализа тонкой структуры обрабатываемых спектров, что создаёт предпосылки для создания более точной технологии диагностики состояния наблюдаемых объектов.

Новые возможности использования всего многообразия измерительных каналов аппаратуры аэрокосмического ДЗ требуют развития математических моделей формирования полей уходящего излучения и адекватных вычислительных процедур анализа и интерпретации получаемых данных. В отличие от традиционных методов, основанных на использовании относительных градаций аэрокосмических изображений, в предлагаемых нами подходах реализована возможность объединения данных моделирования и мониторинга. Таким образом, с использованием методов вычислительной математики удаётся осуществить переход от исходных измерительных данных к параметрам состояния, с которыми имеют дело пользователи соответствующей информационной продукции.

Для лесных экосистем одним из таких параметров является объём зелёной фитомассы растительности, который, с одной стороны, увязывается эмпирическими соотношениями с общим объёмом биомассы древесины, а с другой стороны, — с содержанием углерода (основным параметром климатических моделей). Преимущества предлагаемых подходов перед существующими аналогами состоят в возможности получения этого и других количественных показателей экологического состояния выбранных регионов . Вместо различных преобразований исходных яркостных образов, изменчивых от одной обрабатываемой сцены к другой, каждый элемент разрешения представляется в терминах указанных параметров состояния, инвариантных относительно условий солнечного освещения и углов визирования выбранных объектов.

В предлагаемой нами методике решаются прямые задачи расчёта спектральных интенсивностей уходящего излучения, регистрируемых аппаратурой аэрокосмического ДЗ, при его взаимодействии с природными средами (земная поверхность и атмосфера), а также обратные задачи восстановления параметров состояния наблюдаемых объектов по данным получаемых измерений.

Прямая задача заключается в построении расчётной базы спектральных образов наблюдаемых объектов. Основной функционал интенсивности уходящего излучения оказывается зависящим от множества параметров, таких как тип лесной растительности, тип межкроновой растительности, условия затенения фитоэлементов при их освещении прямым солнечным излучением и диффузным рассеянным излучением, приходящим со всех участков небесной сферы, оптическая толщина атмосферы и др. В реальности функционал представляет собой интеграл свертки суммарного падающего излучения с весовой функцией чувствительности используемой аппаратуры ДЗ в пределах телесного угла визирования наблюдаемого объекта при конкретных значениях зенитного угла визирования соответствующего объекта и разности азимутов визирования и Солнца.

Для почвенно-растительного покрова возникает необходимость включения в соответствующую

расчётную схему особенностей взаимодействия падающего солнечного излучения с отдельными фитоэлементами. При разработке моделей взаимодействия учитываются спектральные отражательные способности фитоэлементов и условия их затенения при заданных зенитных углах Солнца. Результаты решения прямой задачи можно представить себе как процесс создания некой «книги», каждая «страница» которой описывается координатами «плотность полога леса — ажурность крон деревьев» для соответствующих типов лесных экосистем. В этих же координатах отображаются и значения объёма фитомассы классифицируемых типов растительного покрова. Для каждой «страницы» такой «книги» рассчитываются спектральные интенсивности регистрируемого излучения.

Для решения обратной задачи по восстановлению параметров, характеризующих состояние растительности (в том числе объёма фитомассы), для каждого элемента многоспектрального (гиперспектрального) спутникового изображения осуществляется поиск наилучшего соответствия между измеренными значениями спектральных интенсивностей излучения и их значениями, полученными в результате модельных расчётов. В предлагаемом нами подходе реализуется проверка всех возможных вариантов решения обратной задачи путём сравнения измеряемых данных и исходной базы расчётных данных. Число вариантов определяется довольно большим количеством различных параметров, что обуславливает необходимость программной оптимизации поиска.

Поскольку полученное таким образом решение может быть неединственным, то из полученного ансамбля решений выбирается наиболее вероятное. В расчётной процедуре компьютерного поиска решений рассматриваемой обратной задачи можно изменять разрешение сетки в координатах «плотность лесного полога — ажурность крон деревьев». Для грубого разрешения точность решения будет невелика при малых затратах компьютерного времени. Если же уменьшить область поиска решения, то компьютерное время расчёта будет большим, но при этом возрастёт и точность.

Выходную продукцию для каждого элемента обрабатываемых изображений составляют: тип объекта; прозрачность атмосферы; объём зелёной фитомассы растительности (типичные значения от нуля до приблизительно 35 Т/Га); оценки ошибки воспроизведения объёма зелёной фитомассы растительности (соответствуют точности решения обратной задачи восстановления этой величины); тип растительности (для лесной растительности выделяется 11 классов породного состава: от полностью лиственных до полностью хвойных пород); тип межкроновой лесной растительности; изрезанность верхней границы лесного полога, сомкнутость полога и ажурность крон деревьев.

Численные эксперименты

Приложения созданного алгоритмического и программного обеспечения отрабатывались в применении к данным аппаратуры MODIS/Moderate-resolution Imaging Spectroradiometer («Видеоспектрорадиометр среднего разрешения») среднего пространственного разрешения и аппаратуры ETM+/Enhanced Thematic Mapper («Усовершенствованный тематический картограф») спутника Landsat-7 высокого пространственного разрешения. В дополнение к этому проводилась обработка данных, полученных в процессе лётных испытаний двух типов отечественных гиперспектрометров разработки НПО «Лептон», г. Зеленоград (около 200 спектральных каналов, пространственное разрешение около 2 м с высоты 1 км), для выбранного тестового региона. Полученные результаты опубликованы в работах [1, 2, 3].

На рис. 1 приведён один из результатов обработки изображения ETM+ на дату съёмки 02.09.1999 соответствующей восточной части г. Тверь и ближайших окрестностей.

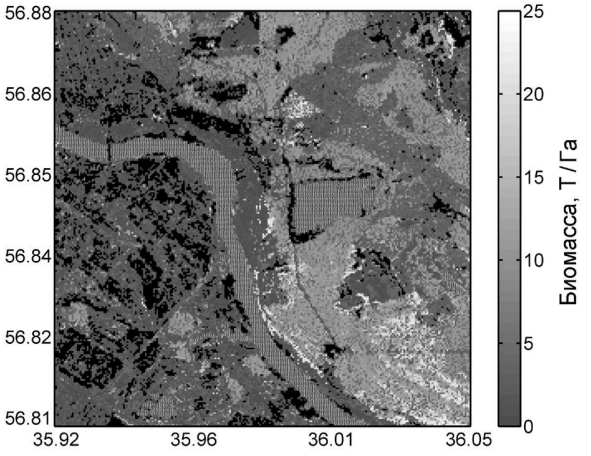


Рис. 1. Восстановление биомассы зелёных фитоэлементов лесной растительности.

Изначально при обработке мультиспектральных изображений выделялись следующие классы: растительность, водоёмы, облачность, открытые почвогрунты и антропогенные объекты. Нестандартная схема распознавания объектов на рассматриваемых многоспектральных изображениях включает использование всех 6 задействованных каналов дистанционного зондирования. Поскольку рис. 1 представлен в градациях серого, почвогрунты и антропогенные объекты были обозначены единым черным цветом, водные объекты выделены сетчатой текстурой, а восстановленные значения биомассы зелёных фитоэлементов

лесной растительности изображены в градациях от тёмно-серого (минимальные значения) к белому (максимальные значения). На горизонтальной и вертикальной осях представлены приблизительные значения долготы и широты в градусах.

Валидация данного изображения может быть проведена на качественном уровне с использованием известной системы Google Earth. Можно видеть, что наряду с крупными водными объектами, такими как р. Волга и песчаный карьер, воспроизводятся более мелкие объекты, соответствующие городским отстойникам и портовой зоне. Ложного распознавания водных объектов не выявлено. Также хорошо распознаются открытые грунты (например, песчаные берега карьера) и промышленные зоны. Легко можно видеть, что наибольшие значения биомассы соответствуют лесным массивам. Наряду с крупными загородными лесными угодьями, можно увидеть лесопарковые зоны внутри г. Тверь, например, Бобачевскую рощу.

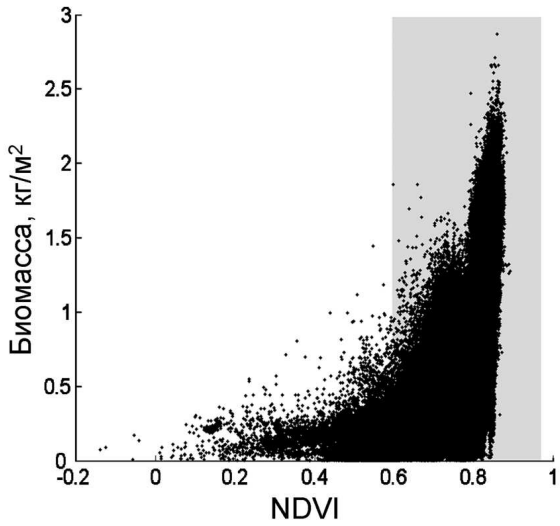


Рис. 2. Зелёная биомасса лесной растительности как функция NDVI.

В традиционных подходах пользователи ориентируются на многочисленные комбинации вегетационных индексов. Наиболее известный — нормализованный разностный вегетационный индекс NDVI [4]. На сегодняшний день считается, что NDVI может успешно применяться при значениях более 0,3. Однако наши результаты показывают, что описание состояния растительности, которое нередко проводится с помощью данных по NDVI, представляется затруднительным. На рис. 2 выделены значения элементов разрешения, соответствующие достаточно большим значениям вегетационного индекса (NDVI>0,6).

Можно видеть, что значениям этого индекса, близким к максимальным, соответствуют самые разные значения объёма фитомассы от нуля до почти 30 Т/Га. Более того, около 30 % точек соответствует значениям биомассы менее 5 Т/Га, а около 15 % точек соответствуют биомассам более 15 Т/Га. Данный результат говорит о том, что индекс NDVI фактически не чувствителен к огромным различиям в содержании биомассы соответствующих растительных сообществ. Таким образом, мы видим необходимость дальнейшего развития предлагаемого нами нового подхода к обработке многоспектральных и гиперспектральных аэрокосмических изображений.

Выводы

Предложена методика обработки и интерпретации многоспектральных и гиперспектральных аэрокосмических изображений, которая включает решение задач распознавания объектов природно-техногенной сферы и количественной оценки состояния различных типов растительности (лесной, болотно-луговой, сельскохозяйственной и др.). Разработано программное обеспечение, предназначенное для обработки абсолютно калиброванных данных дистанционного аэрокосмического зондирования и ориентированое на использование высокопроизводительных многопроцессорных вычислительных систем. Видимые перспективы развития лежат в адаптации разработанного программного обеспечения для обработки данных гиперспектральных аэрокосмических измерений и разработке критериев определения оптимального набора измерительных каналов сканирующих спутниковых радиометров и гиперспектрометров.

Литература

[1] *Козодеров В. В., Кондранин Т. В., Косолапов В. С., Головко В. А., Дмитриев Е. В.* Восстановление объёма фитомассы и других параметров состояния почвенно-растительного покрова по результатам обработки многоспектральных спутниковых изображений // Исследование Земли из космоса. — 2007. — № 1. — С. 57–65.

[2] *Козодеров В. В., Кондранин Т. В., Дмитриев Е. В., Егоров В. Д., Борзяк В. В.* Инновационная технология обработки многоспектральных космических изображений земной поверхности // Исследование Земли из космоса. — 2008. — № 1. — С. 56–72.

[3] *Козодеров В. В., Кондранин Т. В., Казанцев О. Ю., Бобылев В. И., Щербаков М. В., Борзяк В. В., Дмитриев Е. В., Егоров В. Д., Каменцев В. П., Беляков А. Ю., Логинов С. Б.* Обработка и интерпретация данных гиперспектральных аэрокосмических измерений для дистанционной диагностики природно-техногенных объектов // Исследование Земли из космоса. — 2009. — № 2. — С. 50–61.

[4] *Rouse J. W., Haas R. H., Schell J. A., Deering D. W.* Monitoring vegetation systems in the great plains with ERTS // Third ERTS Symposium, NASA SP-351. — 1973. — V. 1. — Pp. 309–317.

Прикладные технологии распознавания количественных характеристик растительности по цифровым многоспектральным и гиперспектральным аэрокосмическим изображениям*

***Кондранин Т. В.*[1], *Козодеров В. В.*[2], *Дмитриев Е. В.*[3], *Егоров В. Д.*[3], *Борзяк В. В.*[2]**

kondr@kondr.rector.mipt.ru

[1]Московский физико-технический институт,

[2]Московский государственный университет им. М. В. Ломоносова

[3]Москва, Институт вычислительной математики РАН

Предлагаются новые подходы к решению задач классификации объектов природно-техногенной сферы и восстановления параметров растительного покрова, с использованием математических методов распознавания образов и регуляризации некорректных обратных задач. Обсуждаются проблемы и возможности технологической реализации предлагаемых методов для создания автоматизированной системы обработки и интерпретации данных аэрокосмического мониторинга.

Достижения в исследованиях Земли из космоса ассоциируются в настоящее время с метеорологическими, экологическими и другими приложениями материалов космических съемок, представленных в форме цифровых многоспектральных и гиперспектральных изображений. Число каналов многоспектрального дистанционного зондирования (ДЗ) обычно не превышает десяти, а в случае гиперспектрального ДЗ используются сотни спектральных каналов.

В существующих приложениях в одних случаях используются обзорные изображения невысокого пространственного разрешения за разные даты съемок, в других — детальные изображения высокого пространственного разрешения для конкретных территорий. Большим спросом пользуются результаты компьютерного отображения материалов космических съемок в близких к реальным цветах. Так, например, в поисковой системе Google Earth в доступной форме на картографической основе (с разной детализацией) можно найти цветные изображения большинства регионов и многих объектов на земном шаре. В настоящее время рынок насыщен разнообразными космическими снимками территорий и конкретных природных или техногенных объектов. Развитие технических возможностей аппаратуры дистанционного зондирования стимулирует процессы усовершенствования существующих и создания новых методов и технологий обработки и анализа изображений.

Тенденцией становится конкуренция космических держав и частных корпораций по совершенствованию аппаратуры ДЗ в различных областях спектра с миниатюризацией измерительных средств (уменьшение массогабаритных характеристик аппаратуры, энергопотребления и т.д.). В этой связи наряду с тяжелыми космическими аппаратами (КА), оснащенными современными многофункциональными измерительными комплексами ДЗ, развивается рынок малых КА, ориентированных на решение специализированных, прикладных задач с использованием данных ДЗ.

Для оснащения таких КА разрабатываются все более совершенные измерительные средства типа гиперспектрометров со многими десятками и даже сотнями спектральных каналов, одновременно обеспечивающие получение изображений высокого пространственного разрешения. Возможности использования всей совокупности спектральных каналов таких аппаратурных комплексов ДЗ для решения прикладных задач далеко не очевидны. С одной стороны, понятно, что с помощью таких средств должно обеспечиваться повышение точности распознавания образов природных объектов по соответствующим изображениям. Но, в то же время, резкое возрастание потоков данных измерений неизбежно затрагивает приложения методов вычислительной математики при автоматизации процесса обработки получаемых изображений. В этом случае модели классификации наблюдаемых объектов и атмосферной коррекции становятся важнейшей составной частью процесса обработки данных ДЗ. Таким образом, возникает необходимость создания нового алгоритмического и программного обеспечения.

Большинство существующих приложений по аэрокосмическим методам использования данных ДЗ привязано к созданию географических информационных систем (ГИС), когда результаты обработки данных проводятся специалистами-интерпретаторами на основе готового программного обеспечения, поставляемого зарубежными фирмами. В этом программном обеспечении заложены стандартные операции по обработке данных, их отображению в определенной проекции карты, в виде цветокодирования результатов обработки и других усовершенствований. Реальные возмож-

*Работа выполнена при финансовой поддержке РФФИ, проекты № 08-07-00284-а, № 08-07-13515-офи_ц, № 09-05-00171-а и в рамках проекта по аналитической ведомственной целевой программе «Развитие научного потенциала высшей школы (2009–2010 годы)» на 2009 год.

ности построения моделей преобразования исходных данных в параметры состояния, других математических преобразований, которые требуются пользователям соответствующей информационной продукции, в имеющемся программном обеспечении не заложены.

В работах [1, 2, 3, 4, 5] впервые показаны приложения новых оригинальных методов, разработанных алгоритмов и специализированного программного обеспечения распознавания образов объектов природно-техногенной сферы и количественной оценки их состояния. В [1] рассматривается информационное обеспечение решения задач оценки параметров состояния природно-техногенных объектов по данным космического и локального мониторинга регионов. Примеры приложений новых подходов для получения информационной продукции использования данных ДЗ продемонстрированы в [2] при обработке многоспектральных изображений аппаратуры MODIS/Moderate-Resolution Imaging Spectroradiometer (видеоспектрорадиометр среднего разрешения) спутника Terra. Прикладные аспекты технологии решения указанных задач при обработке данных аппаратуры ETM+/Enhanced Thematic Mapper (усовершенствованный тематический картограф) спутника Landsat-7 высокого пространственного разрешения показаны в [3, 4]. Соответствующие приложения по обработке данных летных испытаний гиперспектрометрической аппаратуры даны в [5].

Основные этапы обработки многоспектральных и гиперспектральных изображений

Обработка данных ДЗ начинается с их так называемой «нормализации»: согласованная калибровка, приведение к единой форме и т.д. Следующим этапом является условно названная межотраслевой обработка данных, одинаковая для разных потребителей: географическая привязка, трансформирование изображений в определенную проекцию карты, радиометрическая и другие виды коррекции. Эта обработка имеет целью устранить искажения, которым подвергается отраженное от того или иного объекта излучение. Завершающим этапом, разным для различных потребителей, является так называемая тематическая обработка, которую условно можно разбить на 3 этапа.

Во-первых, на изображениях выделяются объекты исследований (например, на основе предварительного физико-географического районирования выбранной территории, на основе карты-нарезки сельскохозяйственных полей, карты землепользования, дорожно-транспортной сети и т.д.). На этом этапе тематической обработки в интерактивном режиме взаимодействия специалиста-интерпретатора с системой обработки данных проводится оконтуривание так называемых генетически однородных подмножеств, возможно использование метода главных компонент, чтобы выделить наиболее связные области в общей картине пространственной изменчивости наблюдаемых объектов. Преобразования являются основной исходной предпосылкой «визуального дешифрирования». Эта терминология наиболее часто применяется при развитии приложений аэрокосмических снимков в географических исследованиях.

Во-вторых, проводится автоматизированное распознавание образов объектов аэрокосмического мониторинга для выделенных генетически однородных подмножеств. Распознавание включает в себя этапы классификации, т. е. построение необходимого алфавита классов, и идентификации, т. е. опознавание этих классов на обрабатываемых изображениях. Распознавание может проводиться как «с обучением» (с использованием обучающей выборки по текущей информации с тестовых участков изображения или по априорной информации из банка спектральных образов), так и «без обучения» на основе кластерного анализа изображений, т. е. знания условий группирования объектов в классы. В результате данного этапа обработки на изображениях опознаются конкретные объекты, для которых далее проводится оценка их состояния.

В-третьих, для конкретных объектов по текущему математическому описанию их свойств, информации из банка данных и используемых модельных описаний взаимодействия излучения с природными средами делаются выводы о принадлежности текущего описания к определенным классам состояния соответствующих объектов. В конечном итоге для каждого элемента обрабатываемого изображения определяется принадлежность объекта к выделенным классам и восстанавливаются количественные параметры, характеризующие состояние этих объектов. Для почвенно-растительного покрова одним из таких параметров является объем зеленой фитомассы и связанный с ним общий объем биомассы древесной и другой растительности.

Распознавание образов объектов по их изображениям

Общая методология распознавания образов объектов по их многоспектральным и/или гиперспектральным изображениям основывается на нескольких принципах: принцип перечисления числа классов; принцип общности свойств объектов; принцип кластеризации.

Задание класса перечислением образов, входящих в его состав, предполагает реализацию процесса автоматического распознавания образов посредством сравнения с некоторым эталоном, для которого этот образ точно известен. Множество обра-

зов, принадлежащих одному классу, запоминается системой распознавания. При предъявлении системе новых образов она последовательно сравнивает каждый из образов с другими, хранящимися в ее памяти. В соответствии с заданными признаками система относит новый образ к наиболее близкому из имеющихся эталонов. Задание класса с помощью свойств, общих для всех входящих в его состав подобразов, предусматривает реализацию процесса автоматического распознавания путем выделения подобных признаков. Основное допущение здесь — образы, принадлежащие одному и тому же классу, обладают рядом общих свойств или признаков, отражающих подобие таких образов. Рассмотрение принципа общности свойств оказывается связанным с необходимостью развития методов выбора признаков, являющихся в некотором смысле оптимальными.

Когда образы некоторого класса представляют собой векторы, компонентами которых являются действительные числа, этот класс можно рассматривать как кластер и выделять только его свойства в пространстве образов кластера. Кластер — это область группирования определенных объектов по заданным их признакам. Построение систем распознавания, основанных на реализации данного принципа, определяется взаимным пространственным расположением отдельных кластеров.

Для реализации перечисленных основных принципов построения автоматических систем распознавания образов существуют три основных типа методологии [6, 7]: эвристическая, математическая и синтаксическая.

За основу эвристического подхода взяты интуиция и знания оператора-исследователя. Обычно системы, построенные такими методами, включают набор специфических процедур, разработанных применительно к конкретным задачам распознавания. Многочисленные приложения аэрокосмических методов в географических исследованиях характеризуют этот подход. Эти приложения развиваются исключительно исходя из опыта специалиста-интерпретатора при заранее заданном программном обеспечении обработки изображений.

В основу математического подхода положены правила распознавания, которые формулируются в рамках определенного математического формализма с помощью принципов общности свойств объектов и их кластеризации. Используются детерминированные и статистические методы построения таких систем распознавания. Этот подход допускает значительное расширение процедур обработки данных путем подключения новых модельных представлений к формированию полей уходящего излучения в соответствии с требованиями количественной оценки состояния объектов аэрокосмического ДЗ.

Синтаксические методы применяются для построения специфических систем распознавания с использованием лингвистической общности свойств объектов.

Расстояния при математическом описании образов играют ключевую роль в процессе обработки информации, заключенной в образе. Для многоспектральных изображений цифровые матрицы данных спектральной интенсивности зарегистрированного излучения представляются в виде отдельных элементов разрешения (строки $i = 1, \ldots, I$; столбцы $l = 1, \ldots, L$) для каждого из дискретных $k = 1, \ldots, K$ измерительных каналов: каждый элемент отображается K-мерным вектором образов (например, $K = 6$ для данных аппаратуры ETM+). В случае гиперспектральных изображений вводится понятие гиперкуба данных спектральной интенсивности зарегистрированного излучения, в котором наряду с цифровыми матрицами по пространственным координатам для каждого элемента разрешения (строки–столбцы i, l) представлены практически непрерывные данные по третьей координате — длине волны излучения ($K \sim 200$).

Если следовать рассматриваемой схеме описания образов наблюдаемых объектов, то их поэлементное распознавание осуществляется на основе выбранной меры близости предъявляемых векторов спектральных образов соответствующих объектов (значений интенсивности регистрируемого излучения) некоторым «эталонным образам». Под эталонами при этом понимаются некоторые объекты, для которых известно точно, к какому классу образов они принадлежат. В частности, минимум евклидова расстояния в пространстве образов между текущими значениями регистрируемых спектров J_k (каналы $k = 1, \ldots, K$ — общее число каналов) и спектров, относящихся к «эталонным классам» $J_k^{(n)}$ (номера эталонов $n = 1, 2, \ldots, N$ — общее число таких эталонов) может служить эффективной информационной мерой разделения изображений на классы.

В соответствии с используемыми расчетными процедурами решается задача классификации объектов на обрабатываемых изображениях (облака, водоемы, почвогрунты, разные типы растительности и др.). Следующий этап — решение задачи поэлементного восстановления количественных параметров состояния объектов (объем зеленой фитомассы разных типов экосистем, породный состав лесной растительности, тип межкроновой растительности и др.) на основе обращения основного функционала расклассифицированных на первом этапе обработки многоспектральных или гиперспектральных данных.

Оценка параметров состояния почвенно-растительного покрова

Технологические особенности реализации предлагаемого способа получения новой информационной продукции по количественной оценке состояния почвенно-растительного покрова осуществляются в два этапа обработки многоспектральных и гиперспектральных изображений. Сначала производится распознавание наблюдаемых объектов по пороговым уровням регистрируемой интенсивности излучения с использованием характерных мод гистограмм, другим характерным признакам математического описания образов. Затем для элементов разрешения, относящихся к классу «растительность» осуществляется преобразование данных многоспектрального (гиперспектрального) зондирования в параметры состояния различного типа экосистем (лесные, болотные, луговые, сельскохозяйственные и др.). Одним из таких параметров является объем зеленой фитомассы растительности, который в результате предлагаемых преобразований восстанавливается для каждого элемента разрешения класса «растительность» наряду с определением типа растительности и типа подстилающей поверхности для этих элементов. Поэлементное восстановления данного параметра — суть расчетной процедуры обращения основного функционала интенсивностей уходящего излучения, регистрируемых соответствующей аппаратурой ДЗ.

В применении к лесной растительности при обращении функционала используется специальная программа минимизации расстояния между пересечениями изолиний интенсивности уходящего излучения для разных каналов и значениями объема фитомассы растительного покрова в координатной плоскости «плотность лесного полога — ажурность крон деревьев». Для других типов растительности используется аналогичная процедура поиска решений обратной задачи в указанной координатной плоскости. Элементы предлагаемой технологии обработки изображений по данным абсолютно калиброванной аппаратуры ДЗ ориентированы на использование высокопроизводительных многопроцессорных вычислительных систем при решении прикладных задач аэрокосмического мониторинга земной поверхности. Вместе с тем, исследуются возможности повышения расчетной эффективности разрабатываемого программного обеспечения при его реализации на персональных компьютерах.

Выводы

Основу предлагаемой технологии распознавания образов природно-техногенных объектов и количественной оценки состояния этих объектов по данным аэрокосмического ДЗ составляют модели описания взаимодействия оптического излучения с природными средами и методы вычислительной математики при нахождении информационной меры близости описаний текущих элементов изображений к описаниям некоторых «эталонных» объектов. Методы, алгоритмы и расчетные программы реализации технологии объединяют решение прямых задач формирования и трансформации излучения в системе «подстилающая поверхность — атмосфера — приемная аппаратура ДЗ», а также постановку обратных задач восстановления параметров состояния наблюдаемых объектов по данным ДЗ для каждого элемента разрешения соответствующей аппаратуры. При этом используются достижения в области распознавания образов и анализа сцен, компьютерных и геоинформационных технологий.

Литература

[1] *Кондранин Т. В., Козодеров В. В., Топчиев А. Г., Головко В. А., Косолапов В. С.* Информационное обеспечение задач оценки состояния природно-техногенной сферы с использованием космического и локального мониторинга // Сб. «Современные проблемы дистанционного зондирования Земли из космоса», вып. 3, т. 1. Москва: Изд-во ООО «Азбука-2000», 2006. — С. 185–191.

[2] *Козодеров В. В., Кондранин Т. В., Косолапов В. С., Головко В. А., Дмитриев Е. В.* Восстановление объема фитомассы и других параметров состояния почвенно-растительного покрова по результатам обработки многоспектральных спутниковых изображений // Исследование Земли из космоса. — 2007. — № 1. — С. 57–65.

[3] *Кондранин Т. В., Козодеров В. В., Казанцев О. Ю., Бобылев В. И. и др.* Технология оценки состояния природно-техногенной сферы по данным аэрокосмического мониторинга // Сб. «Современные проблемы дистанционного зондирования Земли из космоса», вып. 5, т. 2. Москва: Изд-во ООО «Азбука-2000», 2008. — С. 512–522.

[4] *Козодеров В. В., Кондранин Т. В., Дмитриев Е. В., Егоров В. Д., Борзяк В. В.* Инновационная технология обработки многоспектральных космических изображений земной поверхности // Исследование Земли из космоса. — 2008. — № 1. — С. 56–72.

[5] *Козодеров В. В., Кондранин Т. В., Казанцев О. Ю., Бобылев В. И. и др.* Обработка и интерпретация данных гиперспектральных аэрокосмических измерений для дистанционной диагностики природно-техногенных объектов // Исследование Земли из космоса. — 2009. — № 2. — С. 50–61.

[6] *Дуда Р., Харт П.* Распознавание образов и анализ сцен. — Москва: Мир, 1976. — 282 с.

[7] *Ту Дж., Гонсалес Р.* Принципы распознавания образов. — Москва: Мир, 1978. — 412 с.

Частотный анализ данных магнитной энцефалографии в аудиторном эксперименте*

Корнилина Е. Д., Махортых С. А., Семечкин Р. А.
ekornilina@gmail.com, makh@impb.ru, ras@impb.psn.ru
Москва, МГУ ВМК; Пущино, ИМПБ РАН

Целью настоящей работы является создание методов, алгоритмов и программ анализа и классификации экспериментальных данных магнитной энцефалографии для решения задачи картирования функциональных областей головного мозга. Методика применялась к реальным данным. В результате дополнительной процедуры выделения полезных частотных составляющих сигнала, была решена задача локализации двух токовых источников биомагнитной активности в височных долях.

В настоящее время актуальной является задача картирования головного мозга — выделение функций различных областей и локализация участков, отвечающих за те или иные действия, совершаемые испытуемым. Для решения этой задачи предлагаются различные неинвазивные подходы, основными из которых являются электроэнцефалография (ЭЭГ) и магнитная энцефалография (МЭГ). Обработка данных МЭГ позволяет совершать более точную, чем при использовании результатов ЭЭГ, трёхмерную локализацию источников нейронной активности. В связи со слабостью магнитных полей, возникающих на поверхности головы, возникает потребность в использовании высокочувствительного оборудования, что существенно повышает стоимость эксперимента. При проведении пространственно-временного анализа необходима предварительная фильтрация шумовой составляющей сигнала. Чтобы подготовить данные для решения обратной задачи — нахождения источника активности магнитного поля — был произведён частотный анализ с помощью спектрального преобразования Фурье и вейвлет-преобразования. Результаты обработки экспериментальных данных МЭГ при подаче звукового сигнала повышают точность локализации источников биомагнитных сигналов, обнаруживая повышение нейронной активности в областях правой и левой височных долей мозга.

Постановка задачи

Многоканальная магнитная энцефалография (МЭГ) — высокотехнологичный метод получения информации о функционировании различных областей головного мозга. Несмотря на возникающие технические проблемы и высокую стоимость экспериментального оборудования, использование данных МЭГ является перспективным направлением биомедицины. Основное преимущество заключается в том, что по сравнению с электрическим, магнитное поле испытывает значительно меньшие искажения на внутричерепных неоднородностях и покрывающих тканях, что существенно повышает точность локализации источников. В качестве исходных данных использовались результаты эксперимента, проведённого в медицинской школе Нью-Йоркского Университета. На поверхность головы испытуемого надевалось 148-канальное измерительное устройство, покрывающее примерно две трети поверхности сферы, с которого снимались показания с частотой 500 Гц. Изменения величины магнитного поля наблюдались при подаче звукового сигнала, который представлял собой щелчки, подаваемые с частотой 7 Гц в один из слуховых каналов.

В ходе выполнения работы решались следующие задачи:

1) разработка методов классификации пространственной и временной составляющих магнитных полей головного мозга человека в норме и при патологиях;
2) изучение магнитных полей головного мозга здорового человека при подаче аудиторного стимула;
3) повышение точности локализации источников биомагнитных сигналов по данным МЭГ при аудиторной стимуляции испытуемого.

Анализ пространственной структуры сигнала

Для изучения пространственной структуры исходных данных было произведено разложение в базисе сферических гармоник [1]:

$$f(\theta,\varphi)=\sum_{n=0}^{N}\sum_{k=0}^{n} a_{nk}p_n^k\cos(k\varphi)+b_{nk}p_n^k\sin(k\varphi),$$

где p_n^k — присоединённые полиномы Лежандра.

Ниже приводится выражение для коэффициентов a_{nk}:

$$a_{nk}\int_{-1}^{1}\int_{0}^{2\pi}(p_n^k)^2\cos^2(k\varphi)\,d\mu\,d\varphi=$$
$$=\int_{-1}^{1}\int_{0}^{2\pi}f(\mu,\varphi)p_n^k\cos(k\varphi)\,d\mu\,d\varphi.$$

*Работа выполнена при финансовой поддержке РФФИ, проекты №08-07-00353, №07-01-00564 и №08-01-12030-офи.

Таким образом, было получено признаковое описание в виде набора коэффициентов для каждого момента времени в любой точке пространства. При решении задачи аппроксимации в качестве подзадачи возникла задача экстраполяции на область, не занятую датчиками (они располагались преимущественно в затылочной части). Была предложена процедура пошагового приближения значений функции:

1) инициализация: в области, не занятой датчиками, значение магнитной индукции считать равным среднему значению магнитной индукции на границе;
2) коррекция: используя значения магнитной индукции в области, не занятой датчиками, из предыдущего шага и реальные данные для остальных точек, вычислить новые коэффициенты разложения;
3) обновление: получить новые значения магнитной индукции по найденным коэффициентам разложения в области, не занятой датчиками;
4) вернуться к шагу 2

Для аппроксимации использовался итеративный алгоритм:

1) получение коэффициентов разложения по начальному приближению для области без датчиков;
2) экстраполяция по полученному разложению на неизвестную область;
3) использование линейного приближения по трем точкам для поиска значения подынтегральной функции;
4) получение новых значений коэффициентов;
5) восстановление по коэффициентам значения в искомой точке;

Для правильного выбора трех приближающих значений была построена триангуляция точек, соответствующих положениям датчиков с помощью программы, разработанной И. В. Поповым, И. В. Седых [2]. Некоторые треугольники (преимущественно в «нижней» части, относящейся к челюсти) первоначального разбиения не отвечали критерию Делоне (внутрь окружности, описанной вокруг любого построенного треугольника, не должна попадать ни одна из заданных точек триангуляции) [3], поэтому в целях повышения точности приближения было решено добавить 10 точек в эту область, задав в каждой из них значение, равное среднему арифметическому значений на границе. Это заметно улучшило результаты. Использовались два критерия корректности результатов — точность и гладкость. Точность — это величина, характеризующая аппроксимацию в точках, в которых значение магнитной индукции было известно изначально:

$$G = 1 - \frac{\sum_{i=1}^{N} \left(B_i - B_i^0\right)^2}{\sum_{i=1}^{N} \left(B_i^0\right)^2},$$

где B_i^0 — исходные данные, B_i — полученные значения.

Гладкость оценивалась визуально. Для этого строилась аппроксимация магнитного поля по всей поверхности сферы и производилась визуализация распределения — каждая точка поверхности располагалась таким образом, что ее радиус-вектор в сферических координатах пропорционален величине магнитной индукции в ней. Очевидно, при глубине разложения 1 (т. е. аппроксимации с помощью одного коэффициента) значение магнитной индукции в каждой точке пространства будет постоянным, т. е., исходя из наших построений, должна получиться сферическая поверхность. С увеличением глубины разложения точность приближения магнитного поля поверхности головы повышается, на наблюдаемой поверхности образуются выступы и углубления. Полученные поверхности вполне удовлетворяют условию гладкости.

В целях повышения точности вычислений было исследовано два подхода к вычислению коэффициентов разложения:

1) вычисление сразу всех коэффициентов на некоторой глубине n;
2) постепенное увеличение глубины разложения, от единицы до n, с шагом step $= 1, 2, 3$.

Результаты измерений показали, что при увеличении глубины разложения с шагом 3 точность оказывается выше, чем в других подходах.

Некоторый аналог описанного итерационного метода экстраполяции можно встретить в монографии [4], в главе, посвящённой восстановлению неизвестной фазовой функции в методе рентгеноструктурного анализа. Отмечается сходство с решением фазовой проблемы методом простой итерации. При этом подчёркивается отсутствие строгого доказательства сходимости подобных процедур, однако практические задачи демонстрируют работоспособность данного метода. Это подтверждается и в результатах настоящего исследования.

Следует отметить, что вычислительная сложность задачи связана с большими объёмами данных (один эксперимент — около 500 Мб) и сложностью реализации оптимизационной процедуры в многомерном пространстве параметров, поэтому при вычислении коэффициентов для каждого момента времени использовалась несколько упрощённая версия интегрирования. Поверхность сферы покрывалась равномерной сеткой, а само значение интеграла принималось равным сумме всех значений в узлах сетки, умноженное на площадь поверхности и поделённое на число узлов. Оказалось, что

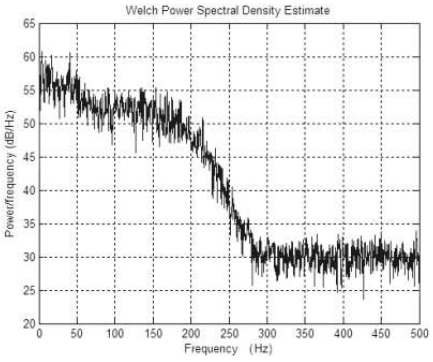


Рис. 1. Спектрограмма данных МЭГ.

подобная процедура приводит к довольно высокой точности уже при количестве узлов, равном 300–400, а дальнейшее увеличение их числа не сильно улучшает ситуацию. При этом время работы алгоритма на вычисление набора коэффициентов в рассматриваемый момент сократилось в 10–60 раз.

Локализация источника вызванной активности

Таким образом, по данным МЭГ для каждого источника были получены коэффициенты разложения в базисе сферических функций для каждого из моментов времени, соответствующих проведённым измерениям. Тем самым попутно решалась задача фильтрации шума и получалось полное признаковое описание магнитного поля на поверхности головы, значительно упрощающее дальнейшее исследование.

Перед распознаванием типа активности сигнала, производилось выделение наиболее информативных признаков сигнала с помощью процедур, описанных в [5].

После этого сигнал характеризовался уже всего тремя признаками, что позволило провести спектральный анализ и вейвлет-анализ временной изменчивости информативных признаков и подготовить данные для того, чтобы в дальнейшем можно было с помощью программы MRIAN (Пущино, ИМПБ) решить обратную задачу МЭГ.

Для локализации частотных характеристик сигнала по времени использовалось оконное Фурье-преобразование, при этом его базисные функции имели постоянное разрешение по времени и частоте.

Для коэффициентов разложения методом Уэлча были получены значения спектральной плотности для различных моментов времени. С помощью спектрограммы (рис. 1) были найдены частоты, на которых наблюдались максимумы амплитуды. По каждой из этих частот был создан фильтр и восстановлены значения магнитной индукции. Обратная задача решалась уже для новых значений.

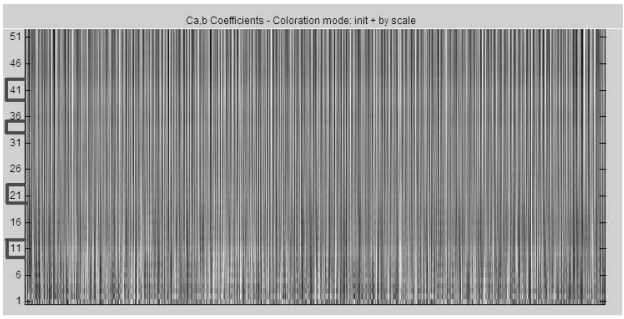


Рис. 2. Коэффициенты вейвлет-разложения в базисе Хаара.

вейвлет-анализ [6] позволяет находить как частотный состав сигнала, так и его локализацию во времени. Тем самым появляется возможность анализировать эволюцию частотных характеристик, а при получении спектральной информации на каждой из частот, дополнительно использовать сведения о прошлом и будущем в описании сигнала.

В качестве базиса вейвлет-преобразования при получении коэффициентов непрерывного одномерного вейвлет-разложения был использован вейвлет Хаара с глубиной разложения равной 5. Для полученных коэффициентов были выделены интервалы частот, которым соответствуют максимумы в спектре (рис. 2). На выделенной частоте было извлечено 10 пиковых моментов времени и для каждого из них произведена локализация источника.

Результаты

При спектральном анализе были выделены три основных частоты — 10, 20 и 30 Гц, активность на частотах, больших 70 Гц считалась шумом. В результате вейвлет-преобразования было выделено четыре диапазона частот, в каждом из которых было выбрано по одной основной частоте — 11, 21, 34 и 40 Гц.

После выделения локальных максимумов на каждой из частот, были получены новые данные пространственной структуры МЭГ для анализа средствами программы MRIAN.

При решении обратной задачи с одним диполем на каждой из частот, как для случая спектрального анализа, так и для случая вейвлет-преобразования были получены следующие результаты. На первых двух частотах источник сигнала локализовался в височных областях. На рис. 3, 4 показаны различные сечения поверхности головы (венечное и осевое) с указанием положения диполя — белый кружок и величины его момента — белый отрезок.

Локализация диполя на частотах, превышающих 30 Гц, была неустойчивой, эту активность можно отнести к так называемому гамма-ритму [7]. Многие нейрофизиологи рассматривают колебания выше 30 Гц как высокочастотный шум и отфильтровывают их при анализе, считая эти частоты на-

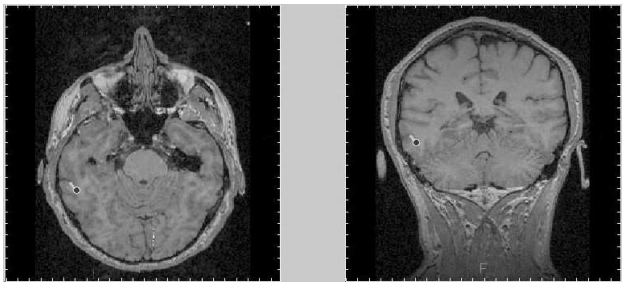


Рис. 3. Локализация источника в левой височной доле.

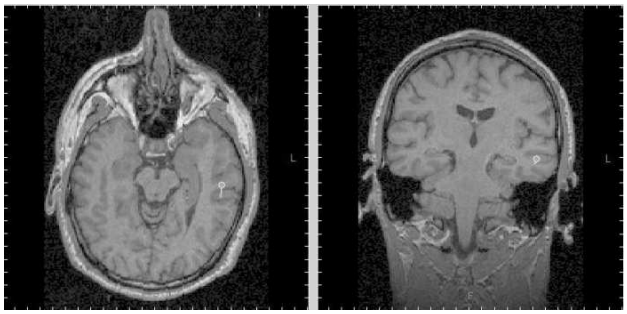


Рис. 4. Локализация источника в правой височной доле.

водками от потенциалов мышц головы и шеи. Показано, что во многих случаях за гамма-волны принимают электромиографическую активность или активность, порождённую миниатюрными движениями глаз [7]. Корректная регистрация гамма-ритма возможна лишь после проведения одновременной записи МЭГ и миограммы, и сопоставления этих данных.

Стоит отметить, что в ходе вейвлет-анализа на каждой из частот выделялось несколько пиковых точек. Это было сделано с целью исключения «выбросов» функции — реакции на непредусмотренные внешние раздражители и т. п. Устойчивая локализация источников биомагнитной активности в височных долях при аудиторной стимуляции испытуемого подтверждает корректность предлагаемой методики анализа данных МЭГ.

Выводы

Разработаны эффективные методы решения задач классификации данных магнитной энцефалографии. Использование таких методов позволило существенно сократить объем обрабатываемых данных и повысить точность вычислений. Данный подход существенно повышает точность решения обратной задачи и позволяет достигать приемлемой точности локализации источников сигнала. Адекватность разработанного метода была подтверждена анализом данных МЭГ при предъявлении звукового раздражителя. В этих условиях выявлена локализация источников повышенной биомагнитной активности в проекционных слуховых зонах в височной коре. Обнаружено, что источники вызванной биомагнитной активности в контрлатеральной и ипсилатеральной подаваемому сигналу областях различаются по амплитудно-частотным характеристикам. Метод может использоваться также в задачах комплексной диагностики различного рода нейрофизиологических заболеваний, в частности, он был использован для случая слуховых галлюцинаций, которые могут возникать как самостоятельное заболевание (tinnitus), так и сопровождать течение некоторых болезней (паркинсонизм).

Литература

[1] *Джеффрис Г., Свирлс Б.* Методы математической физики.— М.: Мир, вып. 3, Т. 3. — 345с.

[2] *Popov I. V., Sedykh I. V.* Russ. J. Numer. Anal. Math Modelling. — Vol. 22, No. 6. — Pp. 591–600.

[3] *Скворцов А. В.* Триангуляция Делоне и ее применение — Томск: Изд-во Томского уни-та, 2002. — 128 с.

[4] *Лунин В. Ю.* Определение пространственной структуры биологических макромолекул Компьютеры и суперкомпьютеры в биологии — М.: Институт компьютерных технологий, 2002, — С. 327–349.

[5] *Дергузов А. В., Махортых С. А.* Распознавание патологической активности в записях магнитных энцефалограмм при болезни Паркинсона // Электронный журнал «Исследовано в России», 2005. — № 149. — С. 1562–1573.
zhurnal.ape.relarn.ru/articles/2005/149.pdf

[6] *Астафьева Н. М.* вейвлет-анализ: основы теории и примеры применения // Успехи физических наук, 1996. — Т. 166., № 11.

[7] *Whitham E. M., Pope K. J., Fitzgibbon S. P.* Scalp electrical recording during paralysis: quantitative evidence that EEG frequencies above 20 Hz are contaminated by EMG // Clin Neurophysiol 118 (8): 1877-88. DOI: 10. 1016 / j. clinph. 2007. 04. 027. PMID 17574912.

[8] *Устинин М. Н., Махортых С. А., Молчанов А. М. и др.* Задачи анализа данных магнитной энцефалографии Компьютеры и суперкомпьютеры в биологии. — М.: Институт компьютерных технологий, 2002, — С. 327–349.

[9] *Махортых С. А., Семечкин Р. А.* Спектральные методы анализа магнитных энцефалограмм. Динамика неоднородных систем // Труды Института системного анализа РАН, 2008. — Т. 33 (вып. 12), — С. 236–250.

[10] *Press W. H., Teukolsky S. A., Vetterling W. T., Flannery B. P.* Numerical Recipes in C. The Art of Scientific Computing. Cambridge University Press, 1992.

Алгоритмы поиска и классификации изображений линейных объектов на космоснимках*

Кандоба И. Н., Костоусов В. Б., Костоусов К. В., Перевалов Д. С.
vkost@imm.uran.ru
Екатеринбург, Институт математики и механики УрО РАН

В работе рассматривается три алгоритма автоматического анализа космоснимков, составляющие вычислительное ядро экспериментального программного комплекса ДЕКОС. Это алгоритм поиска контуров линейных объектов с помощью приближенного решения задачи глобальной оптимизации, алгоритм автоматической классификации линейных объектов с помощью локальной параболической модели и алгоритм распространения результатов классификации. Приводятся примеры работы, обсуждаются результаты тестирования.

В настоящее время геоинформационные системы (ГИС) и результаты их работы используются во многих сферах человеческой деятельности. Одним из основных источников входных данных ГИС являются цифровые космические снимки земной поверхности высокого разрешения. В то же время, многие задачи анализа таких снимков являются весьма трудоёмкими и решаются сейчас в ручном или полуавтоматическом режиме человеком-оператором. Одна из таких задач — поиск и классификация линейных топографических объектов типа дорожной и речной сети [1].

В универсальных ГИС ArcGIS [2] и Панорама [3], используемых в промышленности, средства для поиска линейных объектов разработаны слабо. С другой стороны, сложность настройки и качество получаемого результата в специализированных системах автоматического анализа ALFIE [4] и GeoAIDA [5] не позволяют применять их в промышленности в достаточной мере. Поэтому построение специализированных систем автоматического поиска линейных объектов остается трудной и актуальной задачей [1].

В работе описывается несколько алгоритмов автоматического анализа линейных объектов. Реализация этих алгоритмов составляет вычислительное ядро программного комплекса ДЕКОС (ДЕшифрирование КОСмоснимков) [6].

Проблема автоматического анализа изображений линейных объектов

Проблема автоматического анализа изображений линейных объектов на космических снимках решается уже несколько десятилетий [1]. Она состоит из двух взаимосвязанных задач: поиска контуров линейных объектов и классификации объектов с целью отнесения их к известным классам топографических объектов (например, автодороги, лесные тропы, рельсовые пути, реки, ручьи, каналы).

*Работа выполнена при финансовой поддержке РФФИ, проект №09-01-00523 и фундаментальной программы Президиума РАН №29 «Математическая теория оптимального управления», проект П(29)7-2.

Поиск контуров. Задача поиска контуров может формулироваться в полуавтоматической, либо автоматической постановке. В полуавтоматическом случае это задача трекинга дороги, где пользователь указывает ключевые точки, через которые алгоритм должен провести искомые контуры. Мы будем рассматривать более сложный, автоматический случай, когда пользователь задает на входном изображении лишь область поиска, в которой алгоритм должен построить дорожную и речную сети [1].

Классификация. Задача классификации топографических объектов состоит в отнесении некоторого участка изображения к одному из классов. В рассматриваемом случае такой участок является окрестностью контура интересующего линейного объекта.

Алгоритм поиска контуров линейных объектов

Качественный алгоритм автоматического поиска контуров линейных объектов должен содержать по крайней мере три составляющих: построение первоначального решения путем поиска экстремума некоторого функционала, зависящего от изображения; учёт контекста для исправления ошибок поиска; уточнение контуров найденных объектов [1].

Мы будем рассматривать только задачу оптимизации. Учёт контекста является бурно изучаемым направлением, но практически значимых результатов очень немного. А уточнение контуров можно считать проработанной задачей, которую решают методами адаптивных контуров [1].

Мы рассматриваем функционал, который связывает яркости пикселов изображения с моделью искомых объектов в виде протяженных гладких кривых с возможными пересечениями. В настоящее время такой способ задания функционала представляется одним из самых перспективных в силу универсальности, которую можно заложить в модель линейных объектов. Его основным недостатком является вычислительная трудоемкость, так как поиск экстремума обычно осуществляется методами, основанными на алгоритмах типа ими-

тации отжига [7]. Они позволяют добиться относительно хорошего качества результата, но время вычислений является неприемлемо большим при обработке изображений в режиме реального времени. Другие известные способы, которые обладают более высокой вычислительной эффективностью, основаны на предобработке изображения с целью поиска точек, соответствующих центру или краю изображения линейного объекта, а затем построении по ним контуров искомых объектов [8, 9]. Но они зачастую являются эвристическими, и требуют тщательной настройки управляющих параметров.

Для обеспечения достаточного качества решения задачи и приемлемых временных показателей работы алгоритма, в программном комплексе ДЕКОС реализован алгоритм, который за приемлемое время находит приближенное решение оптимизационной задачи. Минимизируемый функционал имеет следующий вид:

$$F(\boldsymbol{S}) = \sum_{x,y} L\big(\, x,\, y,\, \chi_{\boldsymbol{S}}(x,\, y),\, \boldsymbol{S}\,\big) + \sum_{i} G\big(\, r_i,\, \boldsymbol{S}\,\big),$$

где $\boldsymbol{S}$ — некоторый набор ломаных, состоящих из отрезков $\{r_i\}_i$; $\chi_{\boldsymbol{S}}(x,\ y)$ — характеристическая функция носителя $\bigcup_i [r_i]$ этих ломаных на пиксельной сетке.

Функция L описывает локальную яркостную модель линейного объекта. Она вычисляется по значениям яркости пикселов изображения в окрестности пиксела $(x,\ y)$. Если $\chi_{\boldsymbol{S}}(x,\ y) = 1$, то есть пиксел $(x,\ y)$ принадлежит носителю $\bigcup_i [r_i]$, то значение L является оценкой того, что $(x,\ y)$ лежит на изображении некоторого линейного объекта, с направлением, определяемым по $\{r_i\}_i$. Если $\chi_{\boldsymbol{S}}(x,\ y) = 0$, то значение L является оценкой того, что $(x,\ y)$ не лежит на изображении никакого линейного объекта. Оценивание производится на основе локальной модели линейного объекта, согласно которой пикселы вдоль объекта обладают близкими значениями яркости, поэтому в качестве оценки берётся дисперсия этих значений.

Функция G является штрафной функцией, формализующей глобальные геометрические свойства искомой сети линейных объектов: множество $\{r_i\}_i$ должно задавать протяженные гладкие кривые, с возможными пересечениями. Поэтому штрафуются короткие кривые и точки излома.

Предлагаемый алгоритм ищет приближенное решение $\{r_i\}_i$ задачи минимизации $\arg\min F(\boldsymbol{S})$ в два этапа. На первом этапе с помощью «жадного» алгоритма минимизируется L. А именно, входное изображение разбивается на квадраты. В каждом квадрате ищется отрезок, соединяющий две стороны квадрата, с минимальным значением L на этом отрезке. На втором этапе из найденных отрезков

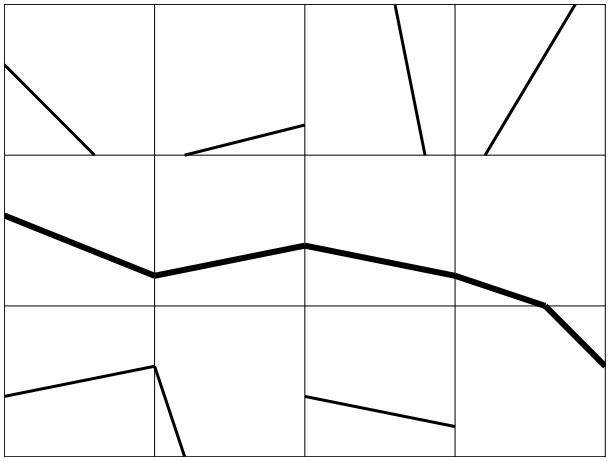


Рис. 1. Схематичное изображение процесса поиска линейных объектов. Жирной линией показана найденная протяженная ломаная без изломов.

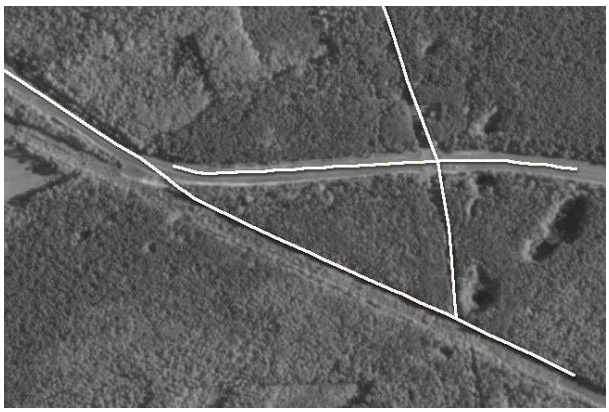


Рис. 2. Результат поиска линейных объектов. Найденные объекты изображены белыми линиями.

с помощью «жадного» алгоритма поиском в глубину конструируется набор протяженных ломаных без изломов для минимизации G, см. рис. 1 и 2.

Время работы алгоритма вполне приемлемо для практического применения. Качество результатов его работы является не очень высоким, что связано с упрощением исходной задачи минимизации. В дальнейшем предполагается повысить качество работы за счет комбинации методов градиентного спуска и имитации отжига.

Алгоритмы классификации линейных объектов

Алгоритмы классификации топографических объектов обычно основываются на использовании базы знаний о свойствах изображений объектов и пространственных связях между объектах [4, 5, 10].

В комплексе ДЕКОС используется два алгоритма. Первый алгоритм осуществляет классификацию объекта по его изображению, используя информацию из окрестности его контура. Этот ал-

горитм применяется для первоначальной разметки найденных контуров.

Второй алгоритм использует уже имеющиеся результаты классификации для распространения классификации на необработанные контуры. Он применяется после исправления ошибок, которые были сделаны алгоритмом классификации по изображению, а также для ускорения ручной классификации.

Алгоритм классификации по изображению. Чтобы осуществить классификацию, для точек объекта строится локальная модель линейного объекта, описанная ниже. С её помощью вычисляется вектор цветовых и геометрических признаков $\boldsymbol{V} = (V_1, \ldots, V_n)$. Затем для каждого из K классов вычисляется величина

$$C_k(\boldsymbol{V}) = \sum_{i=1}^{n} w_{ki} V_i, \quad k = 1, \ldots, K.$$

Здесь w_{ki} — веса, которые предварительно находятся с помощью статистического анализа базы данных примеров объектов каждого класса.

Если для $k^* = \arg\max\limits_k C_k(\boldsymbol{V})$ значение $C_{k^*}(\boldsymbol{V})$ не менее некоторого порога T, задаваемого пользователем, то k^* объявляется результатом классификации. В противном случае, объект считается не классифицированным.

Опишем подробней локальную модель линейного объекта и вычисляемые с её помощью признаки. Удобно предположить, что точки линейного объекта на изображении представляют собой гладкую кривую с толщиной. Поэтому будем считать, что для каждой точки объекта можно подобрать такую систему координат с центром в этой точке, что для некоторой окрестности этой точки все пикселы (x', y') изображения объекта в окрестности аппроксимируются фигурой, которую можно назвать параболой с толщиной $W = 2w + 1$, что можно записать неравенствами

$$-r \leqslant x' \leqslant r, \quad ax'^2 - w \leqslant y' \leqslant ax'^2 + w, \qquad (1)$$

где $w \geqslant 0$ — переменный параметр, характеризующий толщину параболы, $r = 50$ — параметр, характеризующий длину параболы.

Для нахождения системы координат и параметров a и w предлагается использовать результат поиска максимума функционала, который характеризует перепад яркости в точках границы изображения объекта. Для определения этого функционала обозначим через $m_{a,w}$ и $\sigma_{a,w}$ среднее значение и среднеквадратичное отклонение набора яркостей изображения в пикселах

$$Z_{a,w} = \big\{(x', y') \colon x' = -r, \ldots, r,\ y' = ax'^2 + w\big\}.$$

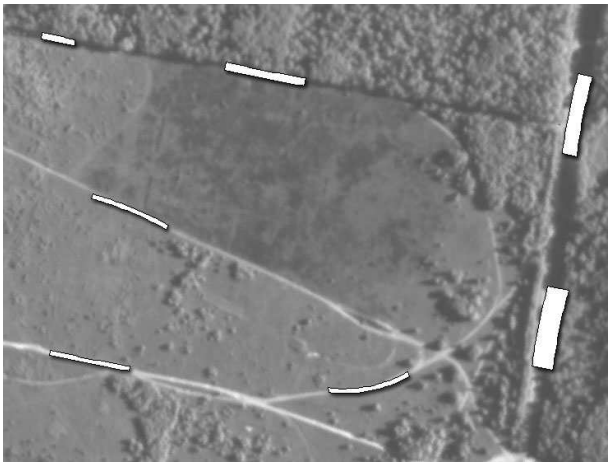


Рис. 3. Примеры вычисления локальных моделей линейных объектов. Найденные фигуры выделены белым цветом.

Тогда функционал можно определить как

$$Q(a, w) = \frac{|m_{a,w} - m_{a,w+1}|}{\sigma_{a,w}} + \frac{|m_{a,-w} - m_{a,-w-1}|}{\sigma_{a,-w}}.$$

На рис. 3 показаны примеры вычисления локальных моделей нескольких линейных объектов. Теперь можно определить признаки, из которых составляется вектор $\boldsymbol{V}$:

1) локальная яркость объекта, равная средней яркости пикселов (1);
2) контраст, равный сумме абсолютных значений перепадов яркости пикселов $(x', y' \pm w)$ и $(x', y' \pm (w + 1))$ для $x' = -r, \ldots, r$;
3) однородность яркости профиля, равная среднему значению $\sigma_{a,i}$ для $|i| \leqslant w$;
4) яркость вне объекта, равная среднему значению $m_{a,i}$ для $|i| = w + 1, \ldots, 2w$;
5) локальная ширина объекта $W = 2w + 1$;
6) локальная кривизна объекта, находимая по a;
7) уверенность, что это линейный объект, равная $Q(a, w)$.

Алгоритм распространения классификации. Работа алгоритма распространения классификации заключается в выборе из множества неклассифицированных объектов Ψ таких объектов, свойства изображений которых были бы похожи на свойства некоторого наперед заданного объекта ψ_0. После нахождения данным объектам устанавливается класс, совпадающий с классом объекта ψ_0, см. рис. 4.

Алгоритм состоит из нескольких шагов. Вначале контур каждого объекта $\psi \in \Psi \bigcup \{\psi_0\}$ разбивается на фрагменты длиной $l = 20$ пикселей. Так как контур мог быть найден в результате автоматического анализа и потому может проходить не точно через центр изображения объекта, производится уточнение положения каждого фрагмента.

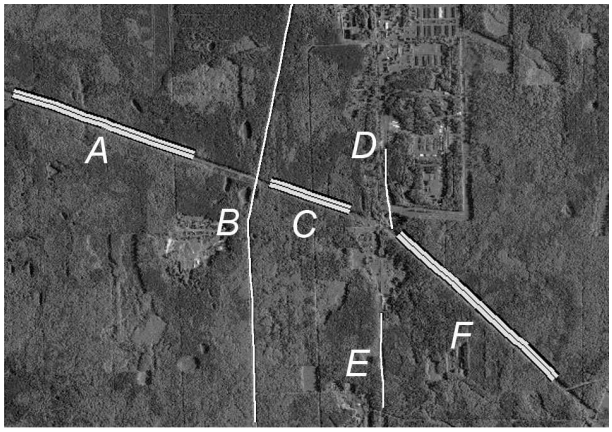


Рис. 4. Пример работы алгоритма распространения классификации. В качестве эталонного объекта ψ_0 взят A. Алгоритм правильно распространил классификацию на объекты C и F.

Для этого находятся параметры модели, аналогичной описанной в предыдущем пункте, но с переменным центром координат и, используя значение найденной ширины, осуществляется смещение вдоль нормали к центральной линии объекта.

Затем для каждого фрагмента считается вектор поперечного профиля средней яркости шириной $6w + 1$:

$$\boldsymbol{P} = (P_{-3w}, \ldots, P_{3w})^{\intercal},$$

где P_i есть средняя яркость пикселов изображения из множества пикселов

$$\big\{(x', y') \colon x' = 0, \ldots, l,\ y' = i\big\},$$

рассматриваемых в такой локальной системе координат, что концы фрагмента имеют координаты $(0, 0)$ и $(l, 0)$.

Введем в качестве меры различия между профилями $\boldsymbol{P_1}$ и $\boldsymbol{P_2}$ величину

$$\mathrm{diff}(\boldsymbol{P_1}, \boldsymbol{P_2}) = 1 - \rho(\boldsymbol{P_1}, \boldsymbol{P_2}),$$

где ρ — нормированный коэффициент корреляции.

Среди усреднённых профилей для кривой ψ_0 ищется такой профиль $\boldsymbol{P}$, для которого радиус наименьшего шара в смысле значений diff, содержащего $T_{\boldsymbol{P}} = 60\%$ остальных профилей, был бы минимален. Обозначим этот радиус $R_{\boldsymbol{P}}$.

После этого для каждой из кривых $\psi \in \Psi$ считается число тех профилей, для которых расстояние до $\boldsymbol{P}$ не превосходит $\alpha R_{\boldsymbol{P}}$, где α — порог чувствительности, задаваемый пользователем. Если для кривой ψ процент таких усреднённых профилей от общего числа усреднённых профилей больше $T_m = 40\%$, то объект ψ считается похожим на эталонный объект ψ_0, и ему устанавливается тот же класс, что и ψ_0.

Выводы

Представленные в работе алгоритмы тестировались на ряде полутоновых космических снимков. Результаты показали, что алгоритмы выдают достаточно много пропусков объектов и отказов в классификации. В то же время, они дают относительно мало ложных срабатываний и неверной классификации. Требуемая правка результатов их работы сравнительно небольшая, и составляет в среднем 20% от числа найденных объектов. Это позволяет утверждать, что алгоритмы могут быть применены в промышленном производстве для первоначальной разметки электронных карт.

Для улучшения качества работы алгоритмов предполагается использование контекстной информации, которая позволит учитывать не только свойства самого объекта, но и наличие и относительное расположение других соседних объектов.

Литература

[1] *Mena J. B.* State of the art on automatic road extraction for GIS update: a novel classification // Pattern Recognition Letters. — 2003. — Vol. 24. — Pp. 3037–3058.

[2] ArcGIS, `http://www.esri.com`.

[3] ГИС Панорама, `http://www.gisinfo.ru`.

[4] *Priestnall G., Hatcher M. J., Morton R. D., Wallace S. J., Ley R. G.* A Framework for Automated Extraction and Classification of Linear Networks // Photogrammetric Engineering & Remote Sensing. — 2004. — Vol. 70, №. 12. — Pp. 1373–1382.

[5] *Bückner J., Pahl M., Stahlhut O., Liedtke C.-E.* A Knowledge-Based System for Context Dependent Evaluation of Remote Sensing Data // 24th DAGM Symposium, 2002. — Pp. 58–65.

[6] *Ефимов С. А., Кандоба И. Н., Костоусов В. Б., Перевалов Д. С., Скрипнюк В. В.* Система автоматизированного топографического дешифрирования изображений земной поверхности // Геодезия и картография. — 2008. — № 5, — С. 34–40.

[7] *Lacoste C., Descombes X., Zerubia J.* Road network extraction in remote sensing by a Markov object process // ICIP. — 2003. — Vol. 3, — Pp. 1017–1020.

[8] *Geraud T., Mouret J.* Fast road network extraction in satellite images using mathematical morphology and Markov random fields // EURASIP J. Appl. Signal Process. — 2004. Vol. 1. — Pp. 2503–2514.

[9] *Lee H. Y., Park W., Lee H.-K., Kim T.-G.* Towards Knowledge-Based Extraction of Roads from 1m-resolution Satellite Images // Proc. IEEE Southwest Symposium on Image Analysis and Interpretation, 2000. — Pp. 171–176.

[10] *John A. Richards, Xiuping Jia* Remote Sensing Digital Image Analysis: An Introduction, 4th Edition — Berlin: Springer, 2006. — 476 p.

Метод анализа коротких отрезков временных рядов*

Котов Ю. Б., Гурьева В. М.
kotsem@voxnet.ru
Москва, Институт Прикладной математики РАН

Алгоритм анализа временных рядов, представленных короткими отрезками, разработан в ходе решения проблемы классификации гестозов у беременных по коротким (до суток) отрезкам записи артериального давления и пульса [1]. Обнаружены отрезки относительной стабильности гемодинамики длительностью до нескольких часов, позволяющие оценить состояние системы управления кровообращением.

Гестозом принято называть заболевание беременных, проявляющееся в увеличении артериального давления, появлении белка в моче и отеков [2, 3]. В отсутствие лечения оно способно прогрессировать, приводя к эклампсии — осложнению, связанному с нарушением сознания, судорогами и возможной гибелью пациентки.

Описание сигнала

В последнее время в медицинскую практику входит автоматическое измерение артериального давления и пульса с помощью автономного прибора (монитора), укрепленного на теле больного и выполняющего периодические измерения в заранее заданные моменты времени [4]. Прибор сохраняет результаты измерений во внутренней памяти и сообщает компьютеру врача во время сеанса связи.

Пример записи наблюдений [5], произведенной самим монитором для больной $N = 295$, приведен на рис. 1. Поскольку значения измеряемых величин (каждая в своих единицах измерения) имеют практически совпадающие численные пределы, то шкалы использованы общие. В дальнейшем будем рассматривать безразмерные величины, численно равные этим значениям. Сама форма диаграммы наглядно иллюстрирует дискретность измерений во времени. По оси абсцисс отложено время наблюдения в часах с 14 часов 25 марта по 14 часов 26 марта 2002 года. По оси ординат отложены значения частоты пульса HR (уд/мин) и значения давлений: систолического SYS (верхнего) и диастоличесого DIA (нижнего) в мм рт. ст.

Момент очередного измерения совпадает с вертикальным отрезком прямой, верхний конец которого отвечает систолическому давлению, а нижний — диастолическому. Значения частоты пульса обозначены точками при тех же абсциссах, что и отрезки. Для наглядности точки частоты пульса последовательно соединены отрезками прямых, образующими ломаную в нижней части рисунка.

Обращает на себя внимание изменчивость давлений и частоты пульса. Это естественное состояние системы кровообращения живого организма, постоянно подстраивающейся к потребностям всех остальных систем. Общий вид диаграммы соответствует модели медленных (многочасовых) процессов в организме, обусловленных его состоянием, и сравнительно быстрых реакций на кратковременные возмущения. Медленные процессы соответствуют длительно протекающему болезненному состоянию организма. Процесс можно представить в виде обычного графика (рис. 2).

По оси абсцисс отложено «непрерывное» время в часах, по оси ординат — безразмерные значения переменных процесса.

Особенности данных

Традиционные методы анализа временных рядов молчаливо используют предположение о достаточной их длине [6]. Это позволяет пользоваться преобразованиями Фурье для выявления скрытой периодичности, автокорреляционными методами для выявления локальной связи значений, усреднением для подавления случайного шума. Импульсы помех большой амплитуды при этом можно просто вырезать, почти не искажая содержательного сигнала.

В данном случае такой «щедрый» подход может привести к потере существенной части данных. Приходится укладываться в жесткий регламент, заданный самим течением болезни — все измерения должны быть выполнены за сутки, хотя суточный ритм организма диктует выделение времени на сон, еду, лечебные процедуры и на физическую активность. Эти виды жизнедеятельности вносят свои изменения в регистрируемые величины. Реально суточное наблюдение дает от 40 до 70 точек, не очень равномерно заполняющих ин-

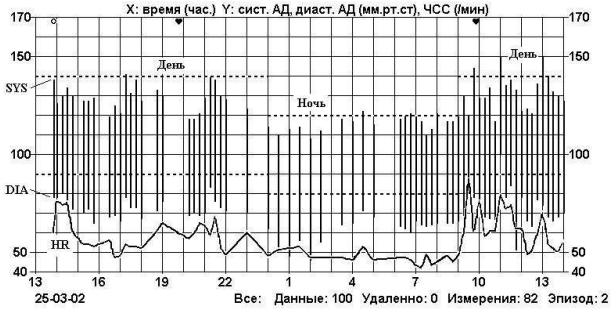


Рис. 1. Диаграмма измерений ($N = 295$).

*Работа выполнена при финансовой поддержке РФФИ, проект № 07-01-00376.

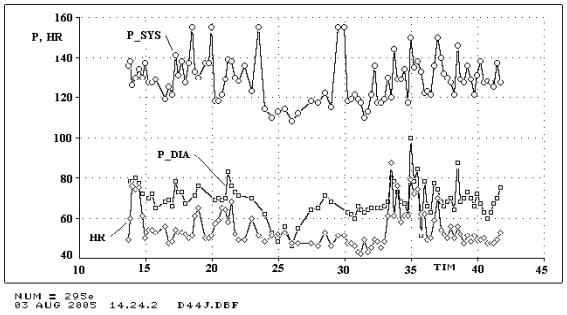


Рис. 2. Графики процесса ($N = 295$).

тервал наблюдения, поэтому дорого каждое успешное измерение (не забракованное аппаратом).

Постановка задачи

Задача состоит в том, чтобы обнаружить признаки возможного опасного развития событий в ближайшие дни. Оговоримся, что в данной работе ни одного случая эклампсии не зарегистрировано, их удалось блокировать вовремя, в начале опасного развития процесса, благодаря информации, сообщенной алгоритмом диагностики.

С точки зрения лечащего врача гестозы делятся на несколько вариантов: отвечающие на лечение (класс G), плохо поддающиеся лечению (класс B) и упорные (практически не поддающиеся лечению, класс E).

У больных с упорным гестозом уровень диастолического давления несколько выше, а частота пульса несколько ниже [1, 5], чем у благополучных больных. Данная работа посвящена приданию точного смысла этим наблюдениям.

На графиках (рис. 2) есть достаточно длинные интервалы, в течение которых взаимное расположение кривых сохраняется. Можно предположить, что эти интервалы отвечают кратковременным стабильным состояниям гемодинамики, коррелирующим с состоянием организма, от которого и зависит возможность угрожающего ухудшения.

Выделение медленного компонента

При малом количестве наблюдений и единичных больших выбросах локальное усреднение не гарантирует корректного представления малоамплитудного медленного компонента. Поэтому мы остановились на медианном сглаживании [7]. В этом подходе для представления значения функции в центральной точке отрезка (по абсциссе) используется медиана распределения значений ординаты на отрезке. Медиана гораздо менее, чем среднее, чувствительна к отдельным выбросам, но хорошо представляет медленный компонент процесса.

На сглаженной записи процесса видно, что медленные компоненты процессов изменяются согласованно. В ночные часы (23–28) наблюдается снижение давлений, постепенно восстанавливающихся лишь к 33-му часу (9 ч. утра). Все время наблюдения сглаженная кривая диастолического давления (SDI) находится выше сглаженной кривой частоты сердечных сокращений (SHR). Для описания взаимного расположения этих кривых воспользуемся разностью SDH = SDI − SHR (рис. 3).

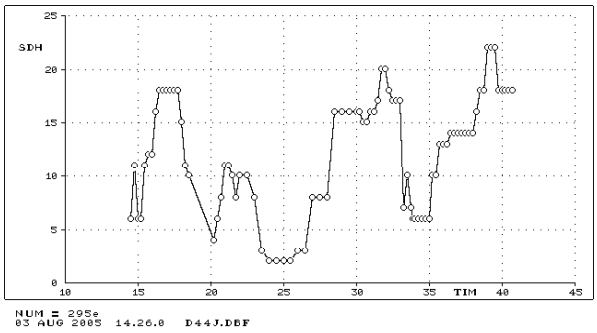


Рис. 3. Разность SDH для $N = 295$.

В этом примере все значения SDH остаются положительными. В другом крайнем случае (пациент $N = 427$) кривая оказывается целиком в отрицательной области.

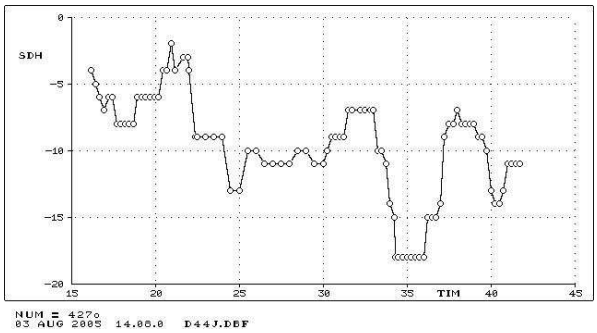


Рис. 4. Разность SDH для $N = 427$.

Наряду с такими вполне четкими вариантами кривых, в массиве наблюдений есть некоторое количество кривых, переходящих в течение суток из отрицательной области в положительную или наоборот. По-видимому, гемодинамика беременной бывает настолько подвижна, что использовать всю кривую целиком для классификации тяжести заболевания нельзя. Рассмотрим сегменты этой кривой по отдельности.

Интервалы относительной стабильности SDH встречаются на обеих кривых. Для первой кривой (рис. 3) можно указать интервалы 16–18, 23–26, 28–31 и 36–38 час. Для второй кривой (рис. 4) — интервалы 18–21, 22–24, 26–28, 31–33 и 34–37 час. Каждый из них можно охарактеризовать разбросом ординаты, ее медианой, моментом начала и продолжительностью по оси абсцисс.

Настройка модели

Материал для анализа: 273 суточных записи давления и частоты пульса у беременных, для которых хотя бы один врач в ходе обследования или клинического ведения пациента предположил возможность развития гестоза. Из них по итогу наблюдения: упорные гестозы (E) составили 18 случаев, гестозы, плохо поддающиеся лечению, (B) — 9 случаев, гестозы средней тяжести (G) — 44, сочетанные (с другими заболеваниями) гестозы (M) — 25, артериальная гипертензия (H) — 75, благополучные больные (O) — 102. Для получения решающего правила использовали сравнение самой тяжелой пары (B, E) с благополучными (O).

Распределения значений сглаженной разности SDH (медиана * и межквартильный интервал -) приведены на рис. 5. Вдоль оси абсцисс отложены безразмерные значения SDH, а строки диаграммы соответствуют классам тяжести заболевания.

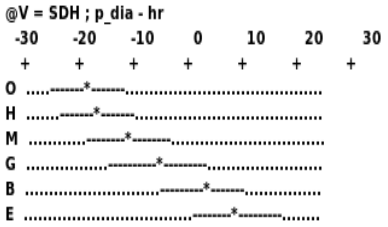


Рис. 5. Распределения SDH.

Виден сдвиг распределений вправо при переходе к более тяжелым классам. Из диаграммы видно, что использование только статистических характеристик давления не решает задачу различения благополучных пациентов и упорного гестоза.

Рассмотрим предлагаемый алгоритм.

Первым этапом обработки записи является медианное сглаживание.

Следующий шаг — определение параметров интервалов времени со стабильными значениями SDH продолжительностью не менее DX часов и принадлежащих определенному периоду суточного наблюдения.

Интервалы длиной не менее 2.5 час встречаются более, чем у 90% пациентов. Для построения диагностического правила минимальная длительность интервала была выбрана DX = 2 час.

Плотность стабильных интервалов варьируется в течение суток. Их меньше в дневное беспокойное время и больше в часы отдыха больных, когда пациент, может быть, и не спит, но предпочитает лежать, а не ходить.

Основная доля моментов начала интервалов приходится на время с 20 до 32 часов (8 час утра). Это время включим в условия правила диагностики нарушений гемодинамики.

Параметры алгоритма используют свойства массива данных, присущие исследованному классу беременных с подозрением на гестоз. Для исследований больных с другими заболеваниями следует определять константы метода заново.

Граница разделения классов по SDH на класс O (благополучные) и класс BE (с опасностью развития тяжелого гестоза) лежит в интервале $(-1.99, 0)$. При этом разделение по количеству интервалов, удовлетворяющих условиям классификации, можно представить таблицей сопряженности (таб. 1). Наличие 185 правильных интервалов у 102 больных говорит о том, что некоторые больные имеют по несколько правильных интервалов на протяжении записи в критическом интервале времени суток.

Таблица 1. Соответствие ответов правила классам.

классы	$<= -1.99$	$>= 0$	
O	185	13	$\chi^2 = 156$
BE	5	48	p_F=1.4.10$^-$33

Если в сглаженной суточной записи артериального давления и частоты пульса беременной в период от 20 до 32 часов имеются участки стабильных значений SDH длиной не менее 2 часов каждый, то уровни SDH на этих участках свидетельствуют о принадлежности записи к классу с опасностью развития упорного гестоза, если все значения SDH на этих участках положительны. Если же эти значения все отрицательны и лежат ниже -2, то опасности упорного гестоза не предвидится.

Выводы

В результате сравнения динамики артериального давления и пульса получено правило формального выделения класса больных с нарушенной гемодинамикой. Поскольку заметных отклонений механических характеристик системы кровообращения врачи не обнаружили, видимо, речь может идти о нарушении работы регулирующих механизмов системы кровообращения у изученной категории пациентов.

Использование полученного правила позволило своевременно начать усиленное лечение больных, предрасположенных к критическому ухудшению состояния.

Внешне записи похожи на динамику системы регулирования, обладающей несколькими устойчивыми состояниями под воздействием внешних возмущений. Для эволюции динамических систем медленные процессы, обычно связываемые с воздействием второго уровня управляющей системы, оказываются более важны, нежели быстрые флюктуации [8, 9].

Литература

[1] *Гурьева В. М., Котов Ю. Б., Логутова Л. С., Петрухин В. А.* Способ диагностики гестоза тяжелой степени у беременных // Патент на изобретение № 2215469.

[2] *Савельева Г. М.* Современные подходы к диагностике, профилактике и лечению гестоза // Методические указания. — 1999. — № 99/80 М. — С. 28.

[3] *Савельева Г. М.* Патогенетическое обоснование терапии и профилактики гестозов // Вестник Российской ассоциации акушеров-гинекологов. — 1998. № 2. — С. 21–26.

[4] *Ольбинская Л. И., Мартынов А. И., Хапаев Б. А.* Мониторирование артериального давления в кардиологии. — М.: Издательский дом «Русский врач», 1998. — 51 с.

[5] *Гурьева В. М., Логутова Л. С., Котов Ю. Б., Петрухин В. А.* Суточный мониторинг артериального давления и частоты сердечных сокращений при диагностике гестоза // Российский вестник акушера-гинеколога. — 2003. — № 1. — С. 4–9.

[6] *Брандт З.* Анализ данных. Статистические и вычислительные методы для научных работников и инженеров. — М.: Мир, ООО «Издательство АСТ», 2007. — 686 с.

[7] *Тюрин Ю. Н., Макаров А. А.* Анализ данных на компьютере. — М.: ИНФРА-М, 2003. — 544 с.

[8] *Гельфанд И. М., Цетлин М. Л.* Принцип нелокального поиска в задачах автоматической оптимизации // ДАН СССР. — 1961. — Т. 137, № 2. — С. 295–298.

[9] *Капица С. П., Курдюмов С. П., Малинецкий Г. Г.* Синергетика и прогнозы будущего. — М.: Наука, 1997. — 285 с.

О прогнозировании спроса на периоды календарных праздников*

Красоткина О. В., Каневский Д. Ю.
krasotkina@forecsys.ru
Москва, ЗАО «Форексис»

Внедрение систем автоматизированного прогнозирования покупательского спроса в современных торговых сетях является важнейшим фактором, обеспечивающим конкурентоспособность торгового предприятия. В данной статье предлагается методика прогнозирования покупательского спроса на периоды календарных праздников, позволяющая прогнозировать спрос в праздник для товаров с любой длиной истории, обладающих неустойчивым спросом, принадлежащих любым отраслям потребительского рынка. Показано, что предложенная методика прогнозирования позволяет существенно сократить потери от неточных прогнозов спроса за счет оптимизации товарных запасов в магазинах.

Введение

Праздничные дни всегда сопровождаются высокой покупательской активностью, и торговые сети готовятся к ним с особой тщательностью. Ключевым фактором планирования становятся точные прогнозы спроса на период праздников. Проблема прогнозирования объёмов продаж (sales forecast) является частным случаем задачи прогнозирования временных рядов и, естественно, в данной области уже накоплен большой арсенал методов [1, 2]. Однако частная задача прогнозирования продаж на периоды календарных праздников, когда существенным образом меняется характер спроса, в литературе исследована недостаточно. Стандартным методом решения этой задачи является введение в модель временного ряда мультипликативной или аддитивной сезонной составляющей [1]. Естественно, для этого требуется иметь достаточно длинную историю временного ряда (хотя бы несколько лет), что в практических приложениях выполняется далеко не всегда. Например, стремительная эволюция рынка мобильных телефонов и бытовой техники ограничивает реальный срок жизни этих товаров всего несколькими месяцами.

Задача прогнозирования продаж осложняется ещё и тем, что целью, как правило, является прогнозирование спроса на заданный товар в заданном магазине. Временные ряды такой детализации сильно зашумлены, что не дает возможности достаточно точно оценивать параметры сезонной составляющей временного ряда.

В данной работе предлагается методика прогнозирования на периоды праздников, основанная на объединении сведений о праздничных продажах целых групп товаров. Данная методика применима для прогнозирования любых товаров, в том числе обладающих короткой историей и неустойчивым спросом; совместима с любыми алгоритмами прогнозирования; не требует сложной настройки и обладает низкой вычислительной сложностью.

*Работа выполнена при поддержке РФФИ, проекты №07-07-00181 и №08-01-12022-офи.

Задача прогнозирования спроса на периоды календарных праздников

Рассмотрим множество временных рядов $\mathcal{R} = \{R_i\}_{i=1}^n$, где $R_i = (r_t^i)_{t=1}^N$ — временной ряд продаж i-го товара. Задача прогнозирования состоит, во-первых, в подборе алгоритма прогнозирования A, во-вторых, определения вектора его параметров $\boldsymbol{\mu}$, и, наконец, определения прогнозного значения спроса в день, отстоящий от дня конца истории N на τ отсчетов:

$$\hat{r}_{N+\tau}^i = A(R, \mathcal{U}, \boldsymbol{\mu}, \tau), \qquad (1)$$

где $\mathcal{U} = \{U_j\}_{j=1}^k$ — совокупность внешних сигналов $U_j = (u_t^j)_{t=1}^N$, которые могут влиять на прогнозируемую переменную. Выбор алгоритма и его параметров мы оставляем за рамками данной статьи, так как эти вопросы достаточно подробно рассмотрены в литературе [1–4]. Будем предполагать, что и алгоритм, и его параметры зафиксированы, и на основании них уже получен прогноз (1). Такой прогноз мы будем называть *регулярным*. Целью данной статьи является разработка методики быстрой корректировки регулярного прогноза в предположении, что день, соответствующий моменту времени $N + \tau$ является праздничным. Так как в реальных автоматизированных системах число одновременно прогнозируемых временных рядов n имеет порядки 10^4–10^8, то такая корректировка не должна занимать много времени.

Каждый анализируемый временной ряд имеет привязку к календарю — каждой точке ряда r_t^i поставлена в соответствие дата d_t. Под *календарным праздником* понимается слитный интервал времени, привязанный к определённым датам и повторяющийся из года в год. Все праздники делятся на типы (Новый год, Рождество, Восьмое марта и т. д.), которые известны заранее.

В основе предлагаемого алгоритма прогнозирования лежит следующая гипотеза: будем считать, что *во время праздника спрос увеличивается в k раз, где величина k имеет характерные значения для данного товара*. Данная гипотеза принимается для большинства товаров, за исключением тех,

которые вне праздника практически не продаются (например, ёлочные игрушки). В данной методологии предлагается выделять такие товары на предварительном этапе и использовать для их прогнозирования сезонные методы.

На практике длина временного рядов может оказаться недостаточной для вычисления коэффициента k, поэтому в данной работе предлагается строить эмпирическое распределение величины k, объединяя коэффициенты увеличения спроса по группам товаров.

Объединение прогнозируемых рядов в группы задается иерархическим *классификатором*. Примерами классификаторов являются номенклатурные перечни (товарные классификаторы), а также группировки по магазинам и далее по регионам (региональные классификаторы). Дадим строгое определение понятия классификатора. Рассмотрим множество $\mathcal{S} = \{S_j \subseteq \mathcal{R}\}$, составленное из всех подмножеств множества $\mathcal{R}$, включая $\mathcal{R}$ и $\varnothing$. *Классификатор* $C \subset (\mathcal{S} \times \mathcal{S})$ устанавливает на множестве $\mathcal{R}$ отношение родства и представляет собой множество упорядоченных пар вида $(\mathcal{S}_i, \mathcal{S}_j)$, где первый элемент является родительским по отношению ко второму. Классификатор удобно изображать в виде графа, который, как правило, является деревом (рис. 1). Множество узлов самого нижнего уровня иерархии представляет собой множество временных рядов продаж $\mathcal{R}$.

Для преодоления неустойчивости рядов продаж предлагается оценивать значение коэффициента увеличения спроса не по временному ряду продаж одного товара, а по *профилю* — совокупному ряду продаж группы товаров определенного уровня иерархии. Выберем некий ряд R_i. Зададимся в каждом из имеющихся классификаторов C_j, $j = 1, \ldots, m$ уровнем иерархии g_j. Совокупность уровней иерархии обозначим $\boldsymbol{g} = (g_j)_{j=1}^m$. Уровень иерархии g_j в каждом классификаторе C_j задает для ряда R_i множество его соседей $\mathcal{R}(g_j)$ в данном классификаторе по группе заданного уровня иерархии g_j. Множество, являющееся пересечением множеств соседей ряда во всех классификаторах, обозначим $\mathcal{R}(\boldsymbol{g}) = \bigcap_{j=1}^{m} \mathcal{R}(g_j)$. *Профилем* товара $p(\boldsymbol{g})$ для заданных уровней иерархии $\boldsymbol{g}$ называется временной ряд, полученный как среднее в каждый момент времени значение по всем рядам множества $\mathcal{R}(\boldsymbol{g})$, по которым в данный момент не было пропусков. Пару «профиль — тип праздника» будем далее называть *прецедентом*.

Алгоритм прогнозирования спроса в дни, принадлежащие праздничному интервалу, состоит из следующих этапов.

1. Определение множества временных рядов, которые будут использоваться для вычисления распределения коэффициента увеличения спроса в праздник.
2. Вычисление профилей.
3. Определение для каждого профиля характера потребительской активности в данный праздник.
4. Вычисление коэффициента усиления спроса для каждого прецедента.
5. Агрегирование полученных коэффициентов усиления в распределение.
6. Корректировка регулярного прогноза с учетом полученного распределения.

Разделение прецедентов на классы по характеру потребительской активности. Каждый прецедент по характеру потребительской активности будем относить к одному из трех классов: строго праздничный, праздничный и непраздничный. Товары первого класса почти не продаются вне праздников. Третий класс характеризуется отсутствием праздничной активности покупателей. Для определения типа прецедента будем использовать следующую процедуру. Сформируем два множества отсчетов времени T_h (множество моментов времени праздника) и T_w (множество моментов времени предыстории). Во множество T_h включим дни, принадлежащие интервалам праздника данного типа и не содержащие маркетинговых акций, других праздников и дефицита. Во множество T_w включим дни, не принадлежащие празднику, не содержащие маркетинговых акций, праздников, дефицита, и отстоящие от интервалов праздника данного типа не более чем на τ отсчетов. Сформируем выборки праздничных P_h и непраздничных P_w значений профиля, соответствующих множествам T_h и T_w. Посчитаем на полученных выборках набор статистик, среди которых количество элементов и средние значения в каждой выборке, отношение средних, статистика Лемана-Розенблатта однородности двух выборок [5]. Далее на полученных статистиках определяется набор предикатных правил. Например, таким правилом может быть превышение статистикой Лемана-Розенблатта критического значения для данного уровня значимости. На базе сформированных предикатных правил строится бинарное решающее дерево, относящее прецедент к одной из перечисленных выше групп. Далее для праздничных прецедентов вычисляется коэффициент увеличения спроса, для непраздничных прецедентов коэффициент увеличения спроса устанавливается равным 1. Для строго праздничных товаров используется сезонная методика прогнозирования, выходящая за рамки данной статьи.

Вычисление коэффициента увеличения спроса в праздник. Для вычисления коэффициента увеличения спроса k для праздничных преце-

дентов используются выборки значений праздничных $P_h = (p_i^h)_{i=1}^l$ и непраздничных $P_w = (p_j^w)_{j=1}^s$ значений профиля, полученные на предыдущем этапе. В предлагаемой методологии реализовано два способа вычисления коэффициента усиления k.

Первый способ выбора коэффициента усиления основан на статистическом критерии однородности выборок Лемана-Розенблатта [5]. Коэффициент праздничного спроса подбирается по сетке с целью минимизации статистики критерия T:

$$\hat{k} = \arg\min_k T(kP_h, P_w).$$

Второй способ основан на вычислении коэффициента увеличения спроса на интервале праздника как результата решения задачи наименьших квадратов, построенной по двум выборкам следующим образом:

$$\hat{k} = \arg\min_k \sum_{i=1}^{l}\sum_{j=1}^{s}\big(f(p_i^h) - f(kp_j^w)\big)^2,$$

где f — монотонная функция, в качестве которой могут быть взяты линейная или логарифмическая зависимость.

Агрегирование выборок коэффициента увеличения спроса в праздник. Агрегирование выборок коэффициентов k для подгрупп или товаров внутри группы более высокого уровня иерархии происходит следующим образом. Если среди агрегируемых профилей больше заданного порога составляют строго праздничные профили, то прецедент считается строго праздничным. Иначе происходит объединение коэффициентов усиления праздничных и непраздничных профилей. Агрегирование происходит простым объединением выборок коэффициента усиления, выбранных для агрегирования подгрупп. Характер агрегирования определяется двумя параметрами — уровнем профиля и уровнем агрегирования в каждом из используемых классификаторов.

Уровень агрегирования g_a определяет уровень предка временного ряда в дереве классификатора, потомки которого будут использоваться для формирования выборки коэффициента усиления.

Уровень профиля g_p в классификаторе определяет, профили какого уровня иерархии будут формировать распределение коэффициента усиления.

Временной ряд, уровень профиля и уровень агрегирования определяют подмножество профилей, участвующих в формировании выборки коэффициента усиления. На рис. 1 сплошным контуром обозначено множество профилей, участвующих в построении коэффициента увеличения спроса для товара i_1, где уровень уровня агрегирования $g_a = 2$, а уровень профиля $g_p = 1$. На том же рисунке пунктирным контуром обозначено множество профилей для товара i_2 и $g_a = 2$, $g_p = 1$.

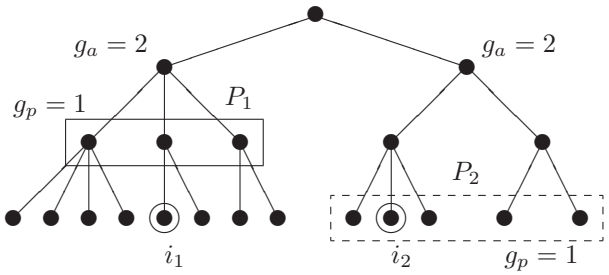


Рис. 1. Пример агрегирования профилей по уровню в классификаторе.

Оценка устойчивости распределения коэффициента усиления спроса. В выборке коэффициента усиления, полученной на предыдущем этапе, могут содержаться выбросы, которые могут негативно отразиться на качестве прогноза. Если для полученной в предыдущем пункте выборке подтверждается гипотеза нормальности, то для удаления выбросов используется тест Граббса (Grubb's test). В противном случае для фильтрации выбросов используются диаграммы «ящик с усами» (box & wishkers plot) [6].

Процедура прогнозирования. Допустим, на предыдущих этапах получена выборка значений коэффициента увеличения спроса $K = \{k_i\}_{i=1}^l$ для данного товара в данном магазине в день прогноза и выборка значений регулярного прогноза $R_{N+\tau} = \{\hat{r}_{N+\tau}^{(j)}\}_{j=1}^v$. Если регулярный прогноз является точечным, то выборка состоит из одного значения, $v = 1$. Если в качестве регулярного прогноза выступает распределение, то выборка достаточного объема генерируется из данного распределения. На основании выборок коэффициента усиления и регулярного прогноза формируется выборка значений скорректированного прогноза на день, принадлежащий интервалу праздника $R_{N+\tau,h} = \{\hat{r}_{N+\tau,h}^c\}_{c=1}^{lv} = \{k_i \cdot \hat{r}_{N+\tau}^{(j)}\}_{i=1\,j=1}^{l\quad v}$. Точечный прогноз получается применением принципа минимизации среднего риска

$$\hat{r}_{N+\tau,h}^* = \arg\min_{c=1,\ldots,lv} \sum_{c'=1}^{lv} L\big(\hat{r}_{N+\tau,h}^c, \hat{r}_{N+\tau,h}^{c'}\big), \qquad (2)$$

где $L(x, y)$ — использующаяся в данной задаче функция потерь. $L(x, y)$ может являться одной из стандартных функций ошибки прогноза, как, например, $(x - y)^2$ или $|x - y|$.Однако в прикладных задачах завышение и занижение прогноза относительно реального значения спроса могут штрафоваться по-разному. Например, при планировании объема закупок товара для магазина завышенный прогноз означает замораживание средств и увеличение складских расходов, а заниженный прогноз — потерю потенциальной прибыли. Это приводит к необходимости использования несимметричных функций потерь [3, 4], требующих в общем

случае явного построения распределения прогноза и численного решения задачи (2).

Зачастую в реальных ситуациях необходимо прогнозировать суммарный объем продаж на некоторый интервал времени, например на неделю или месяц. При этом на интервале прогнозирования могут встречаться подынтервалы, отвечающие различным праздникам. В этом случае интервал прогнозирования разбивается на совокупность элементарных подынтервалов, каждый из которых либо полностью принадлежит интервалу одного из праздников, либо не содержит праздников вовсе. Затем должен быть выполнен прогноз отдельно на каждый элементарный подынтервал, с выбором алгоритма в зависимости от типа праздника. Прогноз на полный интервал получается суммированием прогнозов на элементарных интервалах.

При регулярном прогнозировании дни, принадлежащие праздникам, как правило, не участвуют в вычислении прогнозов. При этом, особенно в случае длинных праздников, теряется существенная информация о продажах. Чтобы устранить этот недостаток, временной ряд продаж очищается от влияния праздника. Для этого значения продаж в праздничные дни делятся на среднее по выборке значение коэффициента увеличения спроса.

Эксперименты

Предложенная методика проверялась на данных розничных продаж компаний, представляющих различные сектора потребительского рынка: электроники, продовольственных товаров, косметической продукции. Для всех вариантов данных использовались три вида функции ошибки:

1) линейная $L_1 = \frac{1}{n}\sum\limits_{i=1}^{n}\frac{\sum_{t=1}^{N}|p_t^i-\hat{p}_t^i|m_t^i}{\sum_{t=1}^{N}m_t^i}$;
2) квадратичная $L_2 = \frac{1}{n}\sum\limits_{i=1}^{n}\frac{\sum_{t=1}^{N}(p_t^i-\hat{p}_t^i)^2 m_t^i}{\sum_{t=1}^{N}m_t^i}$;
3) несимметричная L_{ns}, учитывающая экономические факторы потерь от завышения и занижения прогноза [3, 4].

Для каждого варианта входных данных и функции потерь на интервале обучения подбирались уровни профиля и агрегирования товарного и регионального классификаторов, а также функция для вычисления коэффициента праздничного спроса. Затем алгоритм тестировался на контрольном интервале, в который вошли все праздничные дни за один год. Для всех вариантов входных данных удалось добиться существенного улучшения качества прогнозирования на период календарных праздников по сравнению с регулярным алгоритмом прогнозирования. В таблице 1 приведены значения относительного уменьшения ошибки прогнозирования с помощью предложенной методики по сравнению с регулярным прогнозированием.

Таблица 1. Относительное уменьшение ошибки прогнозирования в период календарных праздников.

Сектор рынка	Функция потерь	Улучшение качества прогнозирования, %
Электроника	L_1	28,33
	L_2	37,59
Парфюмерия	L_1	27,66
	L_2	42,98
	L_{ns}	45,49
Продукты	L_1	15,49
	L_2	31,42
	L_{ns}	45,49

Заключение

В работе представлена методология прогнозирования продаж на период календарных праздников, учитывающая дополнительную информацию о потребительском спросе для групп товаров, представленную в виде набора иерархических классификаторов. Это свойство обеспечивает возможность прогнозирования праздничного спроса для неустойчиво продающихся товаров и товаров с короткой историей (менее года). В экспериментах на реальных данных показано, что предложенная методика позволяет добиться существенного улучшения качества прогнозирования. Дальнейшее улучшение может быть достигнуто за счет привлечения более сложных моделей покупательской активности в праздники, автоматического подбора параметров описанных процедур, автоматического определения периодов праздничной активности, анализа близости распределений коэффициентов усиления спроса для «соседних» групп классификатора.

Литература

[1] *Лукашин Ю. П.* Адаптивные методы краткосрочного прогнозирования временных рядов // Москва: Финансы и статистика, 2003, — 412 с.

[2] *Diebold F. X., Gunther T. A., Tay A. S.* Evaluating Density Forecasts // International Economic Review, 39, 863-883., 1998, (`www.ssc.upenn.edu/~fdiebold/papers/paper16/paper16.pdf`).

[3] *Arminger G., Schneider C.* Assymetric loss function for evaluating quality of forecastin time series for goods management systems: // Tech Rep.22.199, SFB 475, Univerdity of Dortmund, 1999.

[4] *Баринова О. В.* Об одном методе прогнозирования временных рядов с несимметричнфм функционалом потерь // ММРО-12, Москва: Макс-Пресс, 2005. — С. 25–28.

[5] *Большев Л. Н., Смирнов Н. В.* Таблицы математической статистики. — Москва: Наука, 1983. — 415 с.

[6] *Tukey J.* Exploratory Data Analysis. — Addison-Wesley, Reading, MA. 1977.

Локализация границ разномасштабных клеточных структур на основе вейвлет-анализа*

Кревецкий А. В., Ипатов Ю. А.
itinf@marstu.net
Йошкар-Ола, Марийский государственный технический университет

Решена задача автоматизации анализа изображений препаратов поперечных сечений стволов растений. Разработан алгоритм обнаружения границ клеточных структур с отличающимися размерами клеток на основе вейвлет-преобразования. Найдены характеристики качества принимаемых решений полученного алгоритма.

Одной из актуальных проблем наземной лесной таксации является высокая стоимость, временные затраты и, в определенной мере, субъективный характер проводимых измерений [1].

Решение проблемы возможно за счет автоматизации измерений на основе внедрения аппаратно-программного инструментария. Сегодня многие задачи лесной таксации автоматизированы, однако они относятся к области аэрокосмических наблюдений [2]. Вопросы создания инструментария для наземной лесной таксации в силу разнообразия и сложности регистрируемых изображений остаются нерешенными.

В настоящей работе предлагается метод автоматизации актуальной задачи наземной лесной таксации, состоящей в обнаружении и измерении положений границ ранней и поздней древесины на поперечных сечениях стволов растений. Метод основан на вейвлет-анализе цифровых высокодетальных изображений клеточных препаратов и пространственной локализации особенностей вейвлет-спектров. В работе также исследуются результаты работы реализации предлагаемого алгоритма в виде законченной программной модели в среде MATLAB [3].

Постановка задачи

На рис. 1 приведен объект исследования в виде изображения препарата микроспила древесины, полученного с помощью микроскопа при сорокакратном увеличении. Анализ текстурных признаков [4] изображений данного типа показывает, что на них можно выделить две разнотекстурные области, которые классифицируют как ранняя и поздняя древесины. Характерным элементом этих текстур является клетки растений с отличающимся размером. На основе анализа данного изображения $I(x, y)$, где x и y — пространственные координаты, необходимо обнаружить и определить пространственные координаты границ текстур. Осложняет решение задачи высокая и близкая по значению дисперсия яркости в обеих текстурах, неизвестная угловая ориентация изображения φ, неравномерность освещенности препарата и недетерминированный размер элементов каждой текстуры.

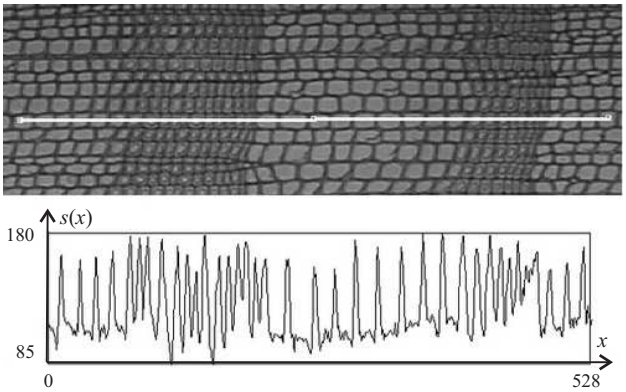


Рис. 1. Распределение яркости вдоль заданной линии на изображении препарата микроспила.

Таким образом, одним из основных требований к синтезируемому алгоритму служит высокая устойчивость к перечисленным факторам при сохранении достаточной для практики точности локализации границ.

Модель изображения препарата

Исследования статистических, геометрических и корреляционных свойств изображений данного класса, но для меньшей детальности, были проведены авторами настоящей работы ранее [2,5] для синтеза оптимального алгоритма локализации текстурных границ, однако изучение тонкой структуры распределений яркости для высокодетальных изображений проводится впервые. На рис. 1 приведена строка изображения $I(x, y)$ микроспила с высоким разрешением, а также построен график яркостных отсчетов $s(x, y{=}\,\mathrm{const})$ этой строки. Результаты спектрального анализа фрагментов подобных строк для разных текстур показывают, что их моды амплитудных спектров-Фурье смещены, но сами спектры существенно перекрываются. Это затрудняет использование Фурье-анализа для сегментации указанных областей. Трехмерное представление отсчетов яркости $I(x, y)$ приведено на рис. 2, что также визуально подтверждает разночастотный характер яркостного распределения для двух клеточных структур.

*Работа выполнена при финансовой поддержке РФФИ, проект № 07-01-00058.

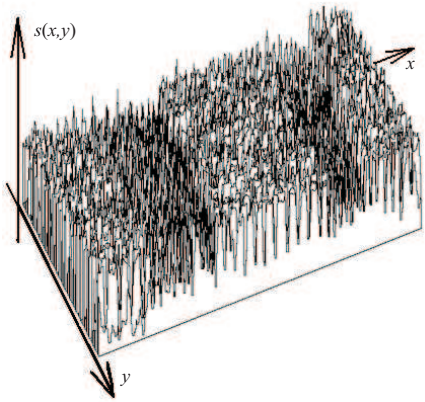


Рис. 2. Оценка отсчетов яркости фона для полученного снимка.

Текстуры ранней древесины имеют периодическое изменение яркостных импульсов, соответствующее более крупным размерам клеток, с частотой $\omega_R(x) = \omega_1 + \Delta\omega_1$, где ω_1 — среднее значение частоты для областей данного вида, а $\Delta\omega_1$ — случайная составляющая частоты, распределенная по случайному закону с нулевым математическим ожиданием и дисперсией σ_1^2. Аналогично, текстуры поздней древесины имеют периодическое изменение яркостных импульсов, соответствующее более мелким размерам клеток, с частотой $\omega_P(x) = \omega_2 + \Delta\omega_2$, где ω_2 — среднее значение частоты для областей данного вида, а $\Delta\omega_2$ — случайная составляющая частоты, распределенная по случайному закону с нулевым математическим ожиданием и дисперсией σ_2^2. При этом $\omega_1 < \omega_2$. Заметно и отличие скважности яркостных импульсов: большее значение скважности имеют области ранней древесины. Классический Фурье-анализ таких сигналов не дает практически значимых результатов по разделению спектров разных текстур и определению их границ в связи с тем, что базисными функциями разложения здесь служат незатухающие гармоники. Как известно из литературы [3, 7, 8, 10], наиболее подходит для локализации особенностей процессов, например, в виде радиотехнических сигналов [8, 9], разномасштабный анализ на основе вейвлет-разложения.

Алгоритм обнаружения границ квазипериодических текстур

В основе синтезируемого алгоритма лежит прямое вейвлет-преобразование сигнала $s_i(x) = s(x, y_i)$, где i — номер строки в кадре [7]:

$$C_i(a,b) = \int_0^X s_i(x) a^{-1/2} \Psi\left(\frac{x-b}{a}\right) dx, \qquad (1)$$

где a и b — параметры вейвлета $\Psi(*)$, задающие масштабирование и смещение по оси x соответственно.

Анализ литературы по вейвлет-разложениям показывает, что на сегодняшний день не существует универсальных и строгих критериев выбора данных функций [8, 9]. Поэтому для решения поставленной в настоящей работе задачи использовался выбор наилучшего из 12 наиболее известных и включенных в пакет MATLAB вейвлетов (Добеши, Хаара, Гаусса и других). В качестве критерия отбора использован максимум вероятности правильного обнаружения границ между текстурами.

В результате в качества базового для алгоритма обнаружения границ текстур был выбран вейвлет Хаара (рис. 3).

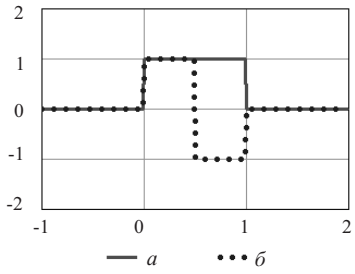


Рис. 3. Вейвлет Хаара: а — масштабирующая функция, б — материнский вейвлет.

Алгоритм обнаружения границ квазипериодических текстур сводиться к выполнению следующих шагов:

1) компенсация неравномерности освещения и неизвестной угловой ориентации исследуемых типов изображений. Этот этап был подробно рассмотрен авторами в работе [6];

2) отбор k строк $s_i(x)$, $i = 1, \ldots, k$ с дисперсией яркости σ_i^2, превышающей среднюю дисперсию яркости кадра σ^2. Этот отбор необходим, так как некоторые строки попадают в область между рядами клеток, где нарушается принятая модель распределения яркостных импульсов;

3) применение прямого вейвлет-преобразования (1) к каждому одномерному сигналу $s_i(x)$, $i = 1, \ldots, k$. На рис. 4 представлены значения коэффициентов вейвлет-разложения в плоскости (a, b).

Светлым участкам спектрограммы соответствует большее значение коэффициентов $C_i(a, b)$. Внизу спектрограммы расположены коэффициенты с малыми значениями параметра a. Они дают детальную картину закономерностей в $s_i(x)$. Сверху — коэффициенты с большими значениями a, дающие огрубленную картину закономерностей в $s_i(x)$. Чистым гармоническим сигналам соответствуют яркие горизонтальные полосы, где модуль вейвлет-коэффициента велик. Локальным особенно-

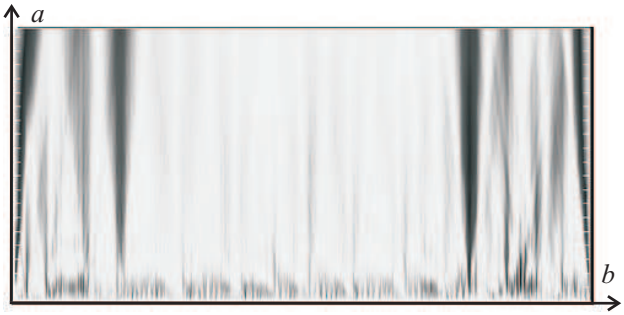


Рис. 4. Вейвлет-спектрограмма сигнала.

стям (нарушениям закономерности) отвечают темные вертикальные полосы, выходящие из точки, где находится особенность [9];

4) усреднение значения коэффициентов вейвлет-преобразования вдоль оси масштаба спектрограммы:

$$Q_i(b) = \frac{1}{M}\sum_{m=1}^{M} C_i(m\Delta a, b), \qquad (2)$$

где Δa — шаг вейвлет-преобразования по масштабу. При этом сохраняются несглаженными особенности, сохраняющиеся и в детальном и огрубленном образе сигнала [10].

На рис. 5 приведена усредненная вейвлет-спектрограмма (2) для одной строки изображения. Здесь локальные максимумы соответствуют границам текстур.

5) усреднение значений коэффициентов по всему изображению для k строк, отобранных на втором шаге:

$$Q(b) = \frac{1}{k}\sum_{i=1}^{k} Q_i(b). \qquad (3)$$

При этом учитываются все локальные особенности каждой строки изображения наблюдаемого кадра. На рис. 6 приведена усредненная вейвлет-спектрограмма по отобранным строкам.

6) пространственная локализация экстремумов усредненной спектрограммы по равенству нулю первой производной (первой разности) и отрицательному значению второй производной (второй разности) [11]. Пример индикации результатов обнаружения границ текстур приведен на рис. 7.

Характеристики работы алгоритма

Для оценки эффективности функционирования программной реализации алгоритма необходимо знать статистические характеристики его работы на исследуемом классе изображений.

Проведена серия экспериментов для нахождения средних ошибок первого и второго рода при обнаружении границ текстур. В результате получено: вероятность ошибки первого рода составила

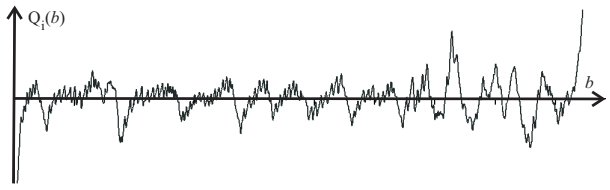


Рис. 5. Усредненная вейвлет-спектрограмма для одной строки изображения.

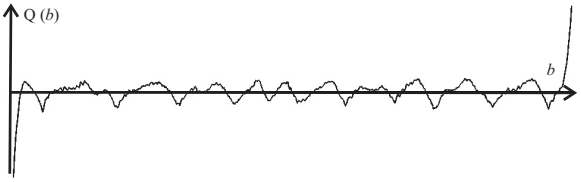


Рис. 6. Усредненная вейвлет-спектрограмма по всем отобранным строкам кадра.

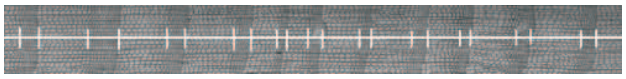


Рис. 7. Результат обнаружения границ разнотекстурных переходов.

$$F = \frac{n_1}{N} = 0{,}153,$$

где n_1 — количество ситуаций, соответствующих случаю, когда текстурная граница обнаружена, но, по мнению эксперта, её нет на изображении, N — экспертная оценка количества границ текстур в кадре; вероятность ошибки второго рода

$$1 - D = \frac{n_2}{N} = 0{,}212,$$

где n_2 — количество ситуаций, соответствующих случаю, когда граница есть, но она не обнаружена алгоритмом.

Выводы

Созданный алгоритм и его программная реализация в среде MATLAB позволяют автоматизировать процесс нахождения разнотекстурных границ на изображениях микроспилов древесины и могут служить готовым инструментом при решении исследовательских и инженерных задач наземной лесной таксации.

Алгоритм обеспечивает среднюю вероятность 0,79 правильного обнаружения границ для рассмотренных классов объектов при вероятности ошибки первого рода 0,153 и при этом обладает высокой вычислительной эффективностью (порядка 237 млн. операций типа сложение/умножение для черно-белого изображения 2392×269 (255 градаций серого). Это на современном процессоре Intel Core 2 Duo составляет менее 1 секунды.

Литература

[1] *Уголев Б. Н.* Древесиноведение с основами лесного товароведения. — М.: МГУЛ, 2001. — 340 с.

[2] *Ипатов Ю. А., Кревецкий А. В.* Сегментация цветных аэрофотоснимков на основе алгоритма селективного обучения // Вестник МарГТУ. Радиотехнические и инфокоммуникационные системы, Йошкар-Ола: МарГТУ, 2008. — № 2. — С. 22–26.

[3] *Смоленцев Н. К.* Основы теории вейвлетов. Вейвлеты в MATLAB. — М.: ДМК Пресс, 2005. — 304 с.

[4] *R. M. Haralick, K. Shanmugam, I. Dinstein* Textural features for image classification // IEEE Trans. Syst. Man Cybern, 1973. — v. 3. — P. 610–621.

[5] *Кревецкий А. В., Ипатов Ю. А.* Обнаружение и измерение параметров протяженных текстурных переходов на изображениях дендрохронологических срезов // VIII Междунар. конф. «Оптико-электронные приборы и устройства в системах распознавания образов, обработки изображений и символьной информации (Распознавание–2008)», Курск: КурГТУ, 2008. — С. 186–187.

[6] *Кревецкий А. В., Ипатов Ю. А.* Автоматическая коррекция пространственных искажений в изображениях разнотекстурных областей // Всерос. конф. «Информационные технологии в профессиональной деятельности и научной работе», Йошкар-Ола: МарГТУ, 2008. — Ч. 2. — С. 82–84.

[7] *Блаттер К.* Вейвлет-анализ. Основы теории. — М.: Техносфера, 2004. — 280 с.

[8] *Малла С.* Вейвлеты в обработке сигналов. — М.: Мир, 2005. — 671 с.

[9] *Дьяконов В.* Matlab. Обработка сигналов и изображений. Специальный справочник. — СПб.: Питер, 2002. — 608 с.

[10] *Короновский А. А., Храмов А. Е.* Непрерывный вейвлетный анализ и его приложения. — М.: ФИЗМАТЛИТ, 2003. — 176 с.

[11] *Корн Г., Корн Т.* Справочник по математике для научных работников и инженеров. — М.: Наука, 1984. — 832 с.

Задача распознавания статистических таблиц*

Кудинов П. Ю.
pkudinov@gmail.com
Москва, Вычислительный центр РАН

Рассматривается задача организации поиска статистических данных, представленных в сети Интернет. Обсуждается специфика исходных данных, методы распознавания статистических таблиц, построение полуавтоматической системы распознавания, её концепция и требования к интерфейсу эксперта.

Одним из основных направлений деятельности экспертов в области экономики является анализ статистических показателей. В настоящее время сбором статистических данных в России занимаются государственные и коммерческие структуры. Ежегодно результатом их труда являются десятки тысяч таблиц, представленных как в бумажном, так и в разнообразных электронных форматах — HTML, Microsoft Excel, Microsoft Word, PDF. Сканирование текста и большая доля ручной работы ведет к большому числу ошибок, поэтому исходные данные являются разнородными, неточными и противоречивыми. Такими же особенностями обладают данные, размещенные в сети Интернет. В последствии планируется рассматривать Интернет как основной источник табличных данных.

Многообразие форматов и представления таблиц, их содержания и структуры делает задачи экспертного анализа статистических данных трудоемкими, требующими большой доли ручной работы, состоящей в поиске и отборе нужных данных по всем таблицам.

В связи с этим актуальной является задача автоматического выделения статистических показателей из таблиц, их приведение к стандартному виду и сохранение в едином хранилище данных. Эта задача решается в контексте распознавания таблиц и состоит в построении описания каждого значения статистического показателя по исходной таблице.

Аналогичная задача решалась в [1] с целью автоматического распознавания автомобильных объявлений, представленных в виде таблиц и размещенных в сети Интернет. Предлагаемое решение состоит в полном описании содержания исходных данных и ручном составлении «выделяющей онтологии» (Extracting ontology), позволяющей в автоматическом режиме обрабатывать автомобильные объявления.

В [4] предложен подход, основанный на эмпирических методах, пригодных для распознавания таблиц фиксированной структуры из фиксированной прикладной области. Имеется обучающая выборка таблиц, на основе которой происходит построение исходной базы известных значений ячеек

Распределение численности занятых в экономике регионов Российской Федерации по возрастным группам в 2000 г. (в процентах)

	Всего	в том числе в возрасте, лет						Средний возраст, лет
		до 20	20-29	30-39	40-49	50-59	60-72	
Российская Федерация	**100**	**2,0**	**21,5**	**27,2**	**30,3**	**14,1**	**5,0**	**39,3**
Центральный федеральный округ	**100**	**1,6**	**20,0**	**26,6**	**30,1**	**15,7**	**6,1**	**40,1**
Белгородская область	100	1,8	19,8	28,4	30,5	12,6	6,9	39,9
Брянская область	100	1,9	22,8	27,7	30,7	12,3	4,6	38,8
Владимирская область	100	2,2	21,0	26,5	30,6	14,4	5,2	39,4
Воронежская								

Рис. 1. Пример исходой таблицы простой структуры.

с текстом. При распознавании каждой новой таблицы происходит поиск значений в базе. Пополнения базы в процессе обучения не предполагается. Несмотря на то, что система показывает неплохие результаты (ошибка на уровне 90%) на таблицах из той же совокупности, откуда была взята обучающая выборка, данный метод не применим для таблиц, содержащих статистические данные.

В настоящей работе предлагается концепция полуавтоматической системы распознавания статистических таблиц и организации поиска статистических данных.

Исходные таблицы

Под исходными таблицами понимаются таблицы, преобразованные в формат HTML и очищенные от различных HTML-тегов стилевого оформления. Такие таблицы обладают определенной структурой, включающей в себя: заголовок таблицы, обычно находящийся за её пределами; блоки данных, содержащих действительные числа; блоки атрибутов (ключей). Ячейки таблицы могут быть объединены по вертикали или горизонтали. На рис. 1 представлен пример исходной таблицы простой структуры с одним блоком данных (выделен светло-серым) и двумя блоками атрибутов (выделены тёмно-серым).

Построение описания числового значения показателя в такой таблице состоит в восстановлении названия показателя, имеющего данное значение. Например, значение 19,8, находящееся в таблице на рис. 1 в 5-й строке, 4-м столбце, будет иметь следующее описание: «распределение численности занятых в экономике регионов Российской Федера-

*Работа выполнена при финансовой поддержке РГНФ (проект № 08-02-12104в) и РФФИ (проекты № 08-07-00305, № 08-01-12022-офи).

Вывоз основных товаров длительного пользования промышленными предприятиями и организациями оптовой торговли из регионов Российской Федерации

	1995	1999	2000		1995	1999	2000
Легковые автомобили , шт.				**Уральский федеральный округ**			
Центральный федеральный округ				Свердловская область	2756	221	6
Костромская область		40	-	Челябинская область	57	-	-
Московская область		10312	14419	**Сибирский федеральный округ**			
г. Москва		5664	1956	Республика Хакасия	63	-	-
Южный федеральный округ				Алтайский край	9667	-	-
Ростовская область		1186	267	Красноярский край	123	18062	12175
Приволжский федеральный округ				Новосибирская область	3520	727	72
Республика Татарстан		18872	22273	Омская область	3055	-	-
Удмуртская Республика		1951	7512	Томская область	78	-	-

Рис. 2. Пример таблицы с несколькими заголовочными блоками и блоками данных.

ции по возрастным группам, Белгородская область, в возрасте 20–29 лет, в 2000 г., в процентах».

Однако далеко не все таблицы имеют простую структуру, так как комбинирование блоков атрибутов и значений позволяет создавать таблицы произвольной структуры. Например, на рис. 2 представлен пример более сложной таблицы, состоящей из двух блоков данных (светло-серый цвет), и пяти блоков атрибутов (тёмно-серый цвет). Следовательно, при построении системы распознавания таблиц необходимо учитывать всё разнообразие допустимых блочных структур.

Исходные таблицы сложны не только разнообразием структуры, но и произвольным текстом атрибутов. В них обнаруживаются грамматические ошибки, слияние слов, замены русских букв на совпадающие по начертанию английские, используются сокращения и синонимы, текст может быть ошибочно разделен на несколько ячеек или наоборот, несколько ячеек объединены в одну.

Задача распознавания

Задача распознавания исходных статистических таблиц заключается в автоматизации построения описания каждого числа, найденного в таблице, и состоит из двух этапов: построение описания для числа и приведение к нормальной форме, пригодной для хранения в базе данных. Нормальная форма включает в себя следующие компоненты: список атрибутов из базы данных, единицу измерения и период времени, к которому привязано значение показателя. Важно отметить, что единицу измерения и период времени нельзя включить в список атрибутов, так как в этом случае затруднятся операции конвертирования единиц измерения и агрегации данных по времени, необходимые эксперту прикладной области.

Метод решения

Решение задачи подразумевает наличие некоторого стандартного набора значений атрибутов, которые используются в процессе нормализации. В качестве исходной базы известных атрибутов используется база элементов общероссийских классификаторов [2]. Общероссийские классификаторы представляют собой иерархические структуры, каждый элемент которых имеет свой код в определенном формате. В качестве примеров классификаторов можно привести Общероссийский классификатор объектов административно-территориального деления (ОКАТО), Общероссийский классификатор информации о населении (ОКИН), Общероссийский классификатор единиц измерения (ОКЕИ).

Основными проблемами при построения автоматической системы распознавания являются, с одной стороны, произвольный текст в ячейках таблицы, а с другой — изменяющиеся классификаторы, в которых редактируются названия элементов и их состав. Следовательно, возникают задачи поиска наиболее близкого из известных атрибутов к данному атрибуту и корректного пополнения базы известных атрибутов. В связи со спецификой исходных данных, разработка полностью автоматической системы, основанной на традиционных методах машинного обучения, не представляется возможной, так как классические методы предполагают наличие представительной обучающей выборки.

Для решения задачи применяется концепция динамического обучения (on-line learning). В соответствии с этим подходом, система работает в полуавтоматическом режиме, принимая на входе последовательность (поток) таблиц. К каждой таблице применяется алгоритм, выдающий эксперту результаты распознавания. Эксперт их анализирует и корректирует. Все скорректированные экспертом решения становятся отрицательными прецедентами и пополняют обучающую выборку. Затем система перестраивает свои модели и переходит к распознаванию следующей таблицы. Важным требованием к динамическому алгоритму является то, что после каждой подстройки моделей результат классификации на всех ранее просмотренных объектах измениться не должен.

Процедура распознавания состоит из семи этапов, представленных на рис. 3, из которых лишь один производится экспертом вручную.

1. Выделение названия таблицы. Название таблицы является атрибутом и содержится в заголовке, расположенном за пределами таблицы. Помимо названия таблицы, заголовок может содержать единицу измерения показателей, дату или год, к которому относятся данные.

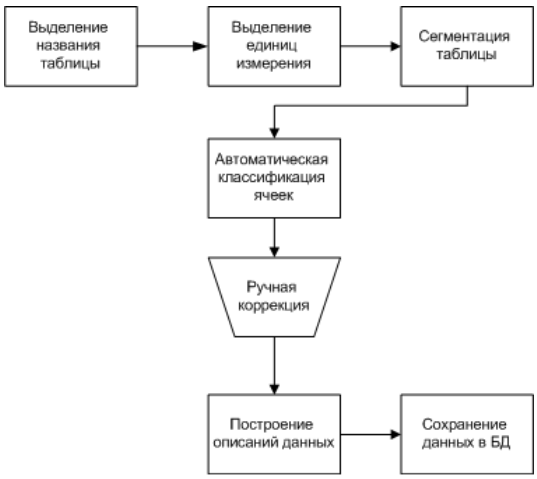


Рис. 3. Общая схема работы системы.

При выделении названия сначала выделяется заголовок таблицы. При этом предполагается, что он находится в тексте, расположенном до или после таблицы. Для выделения заголовка производится анализ абзацев и предложений в исходном HTML коде и поиск в них известных атрибутов из базы данных. Если заголовок не был найден, происходит выделение нового заголовка в соответствии с известными шаблонами положений заголовков в тексте. В этом случае эксперту предлагается убедиться в том, что заголовок был выделен правильно.

Затем из заголовка исключаются единицы измерения и дата. Оставшийся текст считается названием таблицы. После этого выделенное название добавляется в таблицу в виде первой строки с одной ячейкой и становится объектом распознавания наравне со всеми другими атрибутами.

2. Выделение единиц измерения. Для корректной обработки статистических данных экспертам прикладной области необходимо понимать единицы измерения значений показателей. Во многих таблицах единицы измерения содержатся в названиях таблиц или тексте ячеек. В этом случае несложно определить единицу измерения для каждой ячейки. Однако в отдельных случаях единицы измерения необходимо выделять на основе анализа текста ячеек. Это необходимо для корректного поиска по базе атрибутов, так как в ней хранятся значения, не содержащие единиц измерения.

Для выделения единиц измерения используется база известных системе единиц измерения, а также известные ей шаблоны расположения единиц в тексте. Например, в заголовках таблицы единицы измерения часто написаны в скобках (см. рис. 1). В тексте ячеек единица измерения находится в конце предложения и отделена запятой (на рис. 1 «, лет», на рис. 2 «, шт»). Однако нельзя рассчитывать на то, что база единиц измерения будет неизменной (единицей может быть любое существительное), и не будут найдены новые шаблоны в необработанных таблицах. Система выделяет все известные ей единицы измерения и сообщает эксперту о ячейках, для которых не удалось установить единицу. Каждый такой случай должен рассматриваться экспертом на этапе ручной корректировки.

3. Сегментация таблицы. В существующих статистических таблицах значения показателей представлены действительными числами в отдельных ячейках. Также имеются ячейки, содержащие атрибуты. Блоки данных и блоки атрибутов имеют прямоугольную форму. Задача сегментации таблицы состоит в их выделении. Решение этой задачи разделяется на два этапа.

Сначала для каждой ячейки таблицы вычисляется оценка степени её принадлежности к классу значений, который содержит текстовое представление всех действительных чисел. Значение оценки p_{ij} для ячейки с индексами (i, j) вычисляется с помощью решающего списка, построенного по обучающей выборке, состоящей из таких признаков, как отношение длины текста с числами к длине всего текста, положение чисел (в начале, в конце, в середине), наличие в соседних строках / столбцах повторов строковой части записи, и т.д. Таким образом, значение 0 будет соответствовать ячейкам, содержащим атрибуты, а значение 1 — ячейкам, содержащим значения показателей.

После вычисления оценок для каждой ячейки имеем матрицу $P = (p_{ij}) \in \mathbb{R}^{m \times n}$, $p_{ij} \in [0, 1]$, в которой каждый элемент p_{ij} соответствует ячейке исходной таблицы с индексами (i, j). Далее необходимо покрыть матрицу P минимальным числом блоков типа 0 или 1 таким образом, чтобы суммарная ошибка в каждом блоке была минимальной. Суммарная ошибка в блоке типа $t \in \{0, 1\}$ вычисляется по формуле $E = \sum\limits_{(i,j)\in\Omega} |t - p_{ij}|$, где Ω — множество индексов ячеек, содержащихся в блоке.

Результаты сегментации используются на последующих этапах обработки таблицы.

4. Классификация атрибутов. База данных с известными атрибутами представляет собой иерархическую структуру, на первом уровне которой находятся названия различных классификаторов. Классификация атрибутов, расположенных в сегментированных блоках атрибутов, заключается в установлении классификатора каждого атрибута и вычислении оценки степени принадлежности к данному классификатору. Это реализуется с помощью методов нечеткого поиска, основанных на расстоянии Левенштейна [3], вычисляющего число вставок, замен и удалений символов для приведения одной строки к другой. Таким образом, для каждого атрибута c_{ij}, находящего-

ся в ячейке с индексами (i, j), ищется ближайший известный атрибут $\overline{a}$ в базе данных, и выдается оценка $e_{ij} = 1 - \mathcal{L}(c_{ij}, \overline{a})/\max(|c_{ij}|, |\overline{a}|)$, где $\mathcal{L}$ — расстояние Левенштейна. Найденный атрибут войдет в нормальную форму названия показателя.

5. Ручная коррекция. Правильно распознанные ячейки, со значением e_{ij} выше некоторого порога, который задается экспертно, обрабатываются без вмешательства пользователя. Тем не менее, многообразие таблиц и их содержания не позволяет рассчитывать на автоматическую обработку всех ячеек. После этапа классификации необходимо выделить ячейки, обработка которых требует вмешательства эксперта. Деятельность эксперта в основном заключается в редактировании текста ячеек, добавлении новых атрибутов и синонимов атрибутов, исправлении неверных классификаций ячеек и неверных привязок ячеек-значений к ячейкам-атрибутам.

Например, в классификаторе ОКАТО содержится элемент «Город Москва столица Российской Федерации город федерального значения» и нет элемента «г. Москва», который встречается во всех таблицах, содержащих информацию о регионах России. Очевидно, что такого рода ситуации неразрешимы не только автоматической системой, но и неподготовленным пользователем. Таким образом, важной задачей является обеспечение возможности добавления синонимов в базу атрибутов, а также выбор основного атрибута, который будет использоваться в нормальной форме названия показателя, среди множества синонимичных атрибутов. В приведенном примере целесообразно использовать «г. Москва», так как это значение встречается в большинстве таблиц и является общеупотребительным.

Интерфейс эксперта представляет собой веб-приложение и обладает следующими функциями:

1) создание новых классификаторов и их пополнение;
2) корректировка классификации ячеек (изменение похожего атрибута). Такие случаи должны рассматриваться отдельно разработчиками системы;
3) добавление синонимов атрибутов;
4) редактирование текста ячеек;
5) выделение единиц измерения и дат;
6) редактирование названия таблицы;
7) перезапуск обработки данной таблицы.

Важно, чтобы все действия, совершаемые экспертом, сохранялись и анализировались системой, с целью их дальнейшего использования при обработке следующих таблиц. Это позволяет рассчитывать на то, что с каждой обработанной таблицей доля времени, затраченного на ручную корректировку, будет уменьшаться.

6. Построение названий показателей. После определения класса каждой заголовочной ячейки можно приступать к построению названий показателей. Названия строятся для всех ячеек из всех блоков данных, выделенных в результате сегментации. В направлении от ячейки (i, j), для которой строится описание, просматриваются все заголовочные ячейки в j-м столбце и i-й строке. При этом среди всех заголовочных ячеек данного столбца или строки выбираются ячейки из разных классов. К построенному названию добавляется единица измерения данного показателя и период времени, к которому он относится. Образованная нормальная форма названия показателя добавляется вместе с его значением в базу данных распознанных значений.

7. Сохранения данных в БД. Результатом обработки каждой таблицы является список значений показателей с описанием в нормализованной форме. Каждый показатель со своим описанием сохраняется в базу данных, ориентированную на быстрый поиск показателя по тексту входящих в него атрибутов и построение пространственно-временных рядов, необходимых для пользователя системы.

Заключение

Сформулирована проблема распознавания статистических таблиц и описана концепция системы распознавания, основанной на методах динамического обучения. Выделены основные этапы обработки таблиц, только один из которых требует вмешательства эксперта; описаны возможности интерфейса, посредством которого происходит обучающий диалог между экспертом и системой. Результатом работы системы являются значения статистических показателей и стандартизированные названия, позволяющие организовать эффективный поиск по базе показателей.

Литература

[1] *David W. Embley, Cui Tao* Automating the Extraction of Data from HTML Tables with Unknown Structure // Brigham Young University, Provo, Utah 84602, U.S.A.

[2] *Постановление Правительства РФ от 10 ноября 2003 г. № 677* Об общероссийских классификаторах технико-экономической и социальной информации в социально-экономической области.

[3] *Левенштейн В. И.* Двоичные коды с исправлением выпадений, вставок и замещений символов // Доклады АН СССР, 1965, 163.4: 845–848.

[4] *Ashwin Tengli, Yiming Yang and Nian Li Ma* Learning Table Extraction from Examples //School of Computer Science, Carnegie Mellon University, Pittsburgh, PA-15213.

Автоматическая сегментация поведения лабораторных животных на основе выделяемых контуров*

Ломакина-Румянцева Е. И., Ветров Д. П., Кропотов Д. А.
lr2kate@gmail.com
МГУ им. М. В Ломоносова, ВМК; Вычислительный Центр РАН

Предлагается использовать при сегментации поведения животного на поведенческие акты информацию о контуре животного в каждый момент времени. Сегментация осуществляется с помощью алгоритма на основе скрытых марковских моделей с использованием априорного распределения длины сегмента.

Введение

Необходимость создания высокопроизводительных и экономически эффективных методов поведенческого фенотипирования (скрининга) лабораторных мышей привела к появлению автоматических домашних клеток, предоставляющих исследователям возможность оказывать на мышей различные стимулирующие воздействия, и оборудованных системами видеонаблюдения [2]. Однако это привело к взрывному росту сложности и временных затрат на анализ данных. Современные методы анализа поведения, например выделение поведенческих шаблонов и стереотипии [1], требуют разделения траекторий движения на отдельные поведенческие акты. Эта задача в настоящее время может быть выполнена только с привлечением опытного специалиста в области поведения животных. Существующие системы видеонаблюдения за поведением животных позволяют определять некоторые акты с помощью простейших эвристических метрик, например, сравнивая длину мыши с порогом, что обеспечивает крайне низкую точность распознавания. Автоматические методы сегментации траекторий на данный момент позволяют выделять только периоды двигательной активности и неподвижности, требуют тщательной настройки параметров, что существенно ограничивает их применимость на практике [3].

Использование метода скрытых марковских моделей для автоматической сегментации поведения лабораторных животных показало многообещающие результаты [4]. В данной работе для сегментации используется признаковое пространство, расширенное с помощью информации о контуре животного в каждый момент времени.

Признаковое пространство

Многие системы видеотрекинга позволяют выделять только координаты точки, соответствующей центру масс животного. На основе этих данных рассчитываются такие признаки как скорость, ускорение, дисперсия скорости и ускорения, кривизна движения. При дополнительном выделении координат носа и точки прикрепления хвоста становится возможным рассчитать также «вытянутость» животного, угол поворота головы, изменение этого угла и т. п.

Все эти признаки, безусловно, использовались при экспериментах, однако было решено расширить признаковое пространство для улучшения результатов. Система видеотрекинга, используемая в данной работе, позволяет выделять в каждый момент времени не только координаты уже упомянутых трёх точек, но и контур животного. Выделение контура осуществляется на основе моделей активной формы [5]. Для получения обучающего набора все контуры центрируются и поворачиваются вокруг центра таким образом, чтобы точка носа лежала на оси x. Затем для каждого контура через одинаковые интервалы берётся N точек. В соответствующий каждому контуру вектор размерности $2N$ сначала записываются координаты x всех точек, а затем координаты y. На рис. 1 приведён пример нахождения контура животного со взятыми на нём $2N$ точками. К получившемуся набору применяется метод главных компонент [6], т. е. вычисляются характерные изменения контура.

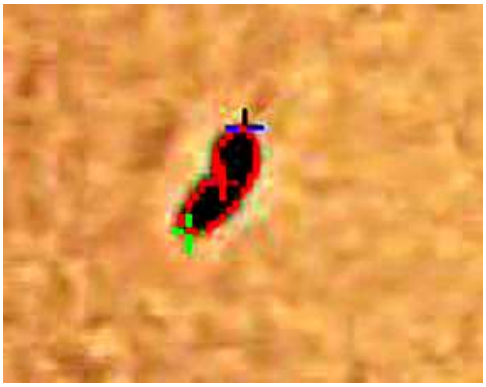


Рис. 1. Пример выделения контура животного.

Обозначим за $x_i \in \mathbb{R}^{2N}$ контур животного в момент времени i, а за $\tilde{x}^k \in \mathbb{R}^{2N}$ — контур k-го животного из обучающего набора. Тогда под $\bar{x} = \frac{1}{M}\sum_{k=1}^{M} \tilde{x}_k$ будем понимать математическое ожидание контура, то есть среднестатистический контур, а главные компоненты набора будем обозначать $y_i \in \mathbb{R}^{2N}$. Тогда дополнительные l признаков в момент вре-

*Работа выполнена при финансовой поддержке РФФИ, проект № 08-01-00405.

мени i будут вычисляться следующим образом:

$$f_i^k = (x_i - \bar{x})^\intercal y_k, \text{ где } k = 1, \ldots, l.$$

По результатам экспериментов было решено взять $l = 5$, так как выборочная дисперсия существенно падает после пятой главной компоненты. На рис. 2 и 3 приведены примеры изменений контура относительно среднестатистического контура, соответствующих первой и второй главным компонентам.

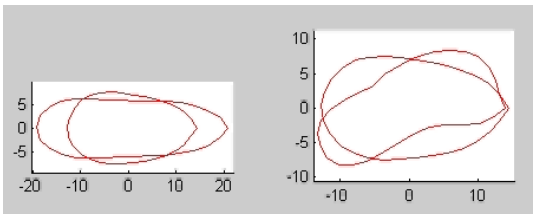


Рис. 2. Первая и вторая главные компоненты.

Метод сегментации

В данной работе для сегментации траектории использовался подход, основанный на скрытых марковских моделях [7]. Скрытые марковские модели являются примером вероятностной модели для обработки последовательностей событий и часто используются для анализа и сегментации сигналов. Предполагается, что мышь в каждый момент времени находится в одном из состояний поведения, которые характеризуются вектором признаков, вычисляемых по траектории, полученной с помощью системы видеотрекинга. Каждое такое состояние трактуется как ненаблюдаемое (скрытое) состояние марковского процесса. Параметры процесса оцениваются по выборке, составленной из траекторий, размеченных экспертом. Помимо этого, учитывается априорное распределение длины сегмента, в течение которого мышь находится в одном состоянии.

В имевшихся выборках экспертами было выделено 9 различных поведенческих актов, часть из которых была сгруппирована в обобщённые состояния. Основанием для группировки являлась частота встречаемости состояния. Окончательно множества состояний выглядели следующим образом: `Run` (бег, ходьба), `Turns` (повороты головы и тела), `Rears` (стойки на задних лапах), `Quiet` (состояние покоя), `Groom` (умывание).

Сегментация новой траектории осуществляется вычислением наиболее вероятной последовательности фаз, основанной на признаках, рассчитанных для каждого момента времени.

Описание алгоритма

Обозначим через $\bar{x}(t) \in \mathbb{R}^d$ наблюдаемый вектор признаков, вычисляемый по траектории мыши для каждого момента времени $t = 1, \ldots, T$. Пусть $z(t) \in \{z_1, \ldots, z_k\}$ — обобщённое состояние (фаза) мыши в момент времени t.

Необходимо найти вектор

$$\big(z^*(1), \ldots, z^*(T)\big) = \\ = \underset{(z(1),\ldots,z(T))}{\arg\max}\ p\big(\bar{x}(1), \ldots, \bar{x}(T),\ z(1), \ldots, z(T)\big).$$

Пусть наблюдаемая траектория $\boldsymbol{x}(t)$ разбита на S сегментов, соответствующих состояниям $z^1, \ldots, z^S \in \mathbb{Z}$. Обозначим через t_i время начала каждого сегмента, $t_0 = 1$, $t_S = T$. Таким образом, на участке от t_{i-1} до $t_i - 1$ обобщённое состояние для всех элементов последовательности равно z^i. Пусть, кроме того, известно некоторое априорное распределение $p_{Y_j}(\tau)$ длины сегмента τ для каждого состояния Y_j.

Тогда вероятность разбиения выглядит следующим образом:

$$p\big(\bar{x}(1), \ldots, \bar{x}(T), z(1), \ldots, z(T)\big) = \\ = p\left(z^1\right) \prod_{i=1}^{S-1} p_{z^i}\left(t_i - 1 - t_{i-1}\right) \prod_{t=t_{i-1}}^{t_i - 1} p\left(\bar{x}(t) \mid z^i\right) \times \\ \times \prod_{i=2}^{S} p\left(z^i \mid z^{i-1}\right) \prod_{t=t_{S-1}}^{t_S} p\left(\bar{x}(t) \mid z^S\right) \left(\sum_{\tau=t_S - t_{S-1}}^{+\infty} p_{z^S}(\tau) \right).$$

Последний множитель учитывает, что последний сегмент может продолжаться сколь угодно долго вне пределов нашего измерения.

Для оценки плотности вероятности $p\left(\bar{x}(t) \mid z(t)\right)$ для каждой фазы $\{z_1, \ldots, z_k\}$ воспользуемся следующим методом. Приведём сначала набор признаков к некоррелированному виду с помощью метода главных компонент. Обозначим преобразованные признаки $g(t) = Q\bar{x}(t)$, где $Q^\intercal = Q^{-1}$ — ортогональная матрица перехода к новому базису, а $\mathsf{E}\bar{g}(t)\bar{g}(t)^\intercal = \operatorname{diag}(\lambda_1^2, \ldots, \lambda_d^2)$. Теперь для каждой фазы построим одномерную оценку плотности значений преобразованных признаков $\hat{p}\left(g^i(t) \mid z(t) = z_j\right)$, $i = 1, \ldots, d$, $j = 1, \ldots, k$. Для этого гистограмму распределения обучающей выборки для каждого состояния и каждой главной компоненты приблизим смесью из пяти нормальных распределений с помощью EM-алгоритма [8]. Гистограмма распределения значений $g^1(t)$ и соответствующая ей аппроксимация пятью гауссианами для фазы `Groom` изображена на рис. 4. Совместная плотность распределения признаков при данной фазе оценивалась как произведение одномерных оценок плотностей распределения $g^i(t)$:

$$\hat{p}\big(\bar{x}(t) \mid z(t)\big) = \hat{p}\big(\bar{g}(t) \mid z(t)\big) = \prod_{i=1}^{d} \hat{p}\big(g^i(t) \mid z(t)\big).$$

Вероятность перехода из фазы z_i в фазу z_j и априорная вероятность каждого состояния легко

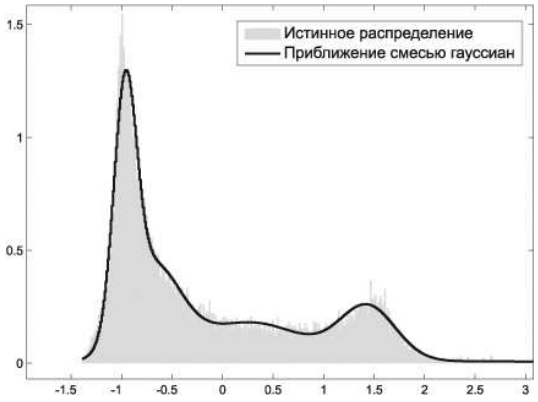


Рис. 3. Гистограмма распределения значений $g^1(t)$ и соответствующая ей аппроксимация для фазы `Groom`.

оцениваются следующим образом:

$$\hat{p}\big(z\,(t)=z_j \,|\, z\,(t-1)=z_i\big) = \frac{\big|\big\{t\colon z(t)=z_j,\, z(t-1)=z_i\big\}\big|}{\big|\big\{t\colon z(t-1)=z_i\big\}\big|},$$

$$\hat{p}\big(z\,(1)=z_i\big) = \frac{\big|\big\{t\colon z(t)=z_i\big\}\big|}{T}.$$

В классическом методе скрытых марковских моделей предполагается, что вероятность длины сегмента τ задаётся следующим образом:

$$p_{Y_j}(\tau) = \big(1-p\,(Y_j|Y_j)\big)p\,(Y_j|Y_j)^{\tau-1}, \quad \tau > 0.$$

В данной работе под априорной вероятностью длины сегмента τ будем понимать следующее

$$p_{Y_j}(\tau) = \begin{cases} 0, & \text{если } \tau < k_0, \\ \big(1-p\,(Y_j|Y_j)\big)p\,(Y_j|Y_j)^{\tau-k_0}, & \text{если } \tau \geqslant k_0, \end{cases}$$

где k_0 — минимально допустимая длина сегмента. Также для проведения процедуры сегментации нам потребуется знать значение величины $\sum\limits_{\tau=t}^{+\infty} p_{Y_j}(\tau)$. Можно показать, что:

$$\sum_{\tau=t}^{+\infty} p_{Y_j}(\tau) = \begin{cases} 1, & \text{если } t < k_0, \\ p\,(Y_j|Y_j)^{\tau-k_0}, & \text{если } t \geqslant k_0. \end{cases}$$

Здесь, как и в предыдущей формуле, k_0 — минимально допустимая длина сегмента. Построение оптимальной сегментации сводится к максимизации вероятности разбиения.

Для дальнейших построений введём функцию Беллмана $V_t(Y_j)$ для каждого момента времени t и состояния Y_j как вероятность наилучшей сегментации от момента времени t_0 до момента времени t, заканчивающейся в состоянии Y_j, следующим образом:

$$V_t(Y_j) = \max\Big\{f\,(Y_j)\,, \max_{Y_i \neq Y_j} g\,(Y_j)\Big\}.$$

Здесь $f\,(Y_j)$ — вероятность наилучшей сегментации от момента времени t_0 до момента времени $t-1$, заканчивающейся в состоянии Y_j, при сохранении состояния Y_j в момент времени t; а $g\,(Y_j)$ — вероятность наилучшей сегментации от момента времени t_0 до момента времени $t-1$, заканчивающейся в состоянии Y_i, с переходом в состояние Y_j в момент времени t. Можно показать, что:

$$f\,(Y_j) = V_{t-1}\,(Y_j) + \log p\,(\bar{x}\,(t)\,|Y_j) + \\ + \log \sum_{\tau=t-t(Y_j)}^{+\infty} p_{Y_j(\tau)} - \log \sum_{\tau=t-t(Y_j)-1}^{+\infty} p_{Y_j(\tau)};$$

$$g\,(Y_j) = V_{t-1}\,(Y_j) + \log p\,(\bar{x}\,(t)\,|Y_j) + \log p\,(Y_j|Y_i) + \\ + \log p_{Y_j}\,(t-t\,(Y_j)-1) - \log \sum_{\tau=t-t(Y_j)-1}^{+\infty} p_{Y_j(\tau)}.$$

Здесь $t\,(Y_j)$ обозначает начало сегмента, в который входит момент времени $t-1$, для каждого состояния Y_j. Обозначим через $S_t\,(Y_j)$ предшествующую точку оптимальной сегментации.

$$S_t\,(Y_j) = \begin{cases} Y_j, & \text{если } f\,(Y_j) > \max\limits_{Y_i \neq Y_j} g\,(Y_j)\,, \\ \arg\max\limits_{Y_i \neq Y_j} g\,(Y_j)\,, & \text{иначе.} \end{cases}$$

Тогда можно последовательно вычислить значения функции Беллмана и функции $S_t\,(Y_j)$ для всех моментов времени $1 \leqslant t \leqslant T$. Выполняя обратный проход, получаем оптимальную разметку траектории

$$(z^*\,(T)\,, z^*\,(T-1)\,, \ldots, z^*\,(1)) = \\ = \big(\arg\max_{Y_j} V_T\,(Y_j)\,, S_T\,(z^*\,(T))\,, \ldots, S_2\,(z^*\,(2))\big).$$

Эксперименты и будущая работа

Предложенная система была протестирована на 13 видеозаписях изучающего поведения мышей полёвок в эксперименте «открытое поле», общее время записи — 325 минут. Из них 150 минут были использованы как обучающая выборка, остальные — как контрольная. Результаты автоматической сегментации были сопоставлены с сегментацией, выполненной вручную, см. таблицу 1, где в каждой ячейке указано соответствующее число кадров.

Таблица 1. Матрица точности распознавания фаз поведения мыши.

Реальный класс: Класс. как:	`Groom`	`Quiet`	`Run`	`Turns`
`Groom`	5683	7120	22	599
`Quiet`	2097	100859	95	283
`Run`	0	21	9890	796
`Turns`	850	7281	356	10382

Общий процент ситуаций совпадения экспертной разметки и результата работы алгоритма составил 87.8%. Лучше всего распознаётся фаза `Run`. Также, благодаря выделению контуров, стабильно распознаётся фаза `Turns`. Существуют некоторые трудности с распознаванием фаз `Quiet` и `Groom`, что связано с визуальной схожестью этих поведенческих актов.

В дальнейшем планируется поставить ряд экспериментов с обучением без учителя для выявления участков стационарного поведения, не связанных с поведенческими актами, которые выделяются экспертами.

Литература

[1] *Magnus S. Magnusson* Discovering hidden time patterns in behavior: T-patterns and their detection // Behavior Research Methods, Instruments & Computers. — 2000. — Т. 32, № 1. — С. 93–110.

[2] *Spruijt B. M., DeVisser L.* Advanced behavioral screening: automated home cage ethology // Drug Discovery Today: Technologies — 2006. — Vol. 3, № 2. — Pp. 231–237.

[3] *Cherepov A. B., Mukhina T. V., Anokhin K. V.* Automatic segmentation of mouse behavior during video tracking in home cages // 5th Int. Conf. on Methods and Techniques in Behavioral Research «Measuring Behavior 2005», Wageningen, 2005.

[4] *Konushin A., Kropotov D., Vetrov D., Lomakina-Rumyantseva E., Zarayskaya I., Anokhin K., Voronin P., Sindeyev M., Kutuzova V.* Система видеонаблюдения за поведением лабораторных животных с автоматической сегментацией на поведенческие акты // Proceedings of GraphiCon'2008, Moscow, 2008 — Pp. 199–205.

[5] *Voronin P., Konushin A.* Video tracking laboratory rodents using active shape models // 9th Int. Conf. on Patterns Recognition and Image Analysis: New Information Technologies «PRIA-9-2008», Nizhni Novgorod, 2008. — Pp. 299–302.

[6] *Jolliffe I. T.* Principal Component Analysis, Series: Springer Series in Statistics, 2nd ed. — Springer, 2002.

[7] *Elliot R. J., Aggoun L., Moore J. B.* Hidden Markov Models: Estimation and Control. — Springer, 1995.

[8] *Dempster A., Laird N., Rubin D.* Maximum likelihood from incomplete data via the em algorithm // Journal of the Royal Statistical Society. — 1977. — Т. 39, № 1. — С. 1–38.

Визуализация исследовательской активности организаций с использованием таксономии предметной области

Миркин Б. Г., Насименто С., Мониш-Перейра Л.
mirkin@dcs.bbk.ac.uk
Лондон, Биркбек Колледж;
Москва, Высшая школа экономики

В ситуации, когда существует многоуровневая таксономия предметной области, такая как Классификация тематических единиц информатики международной Ассоциации вычислительных машин (ABM-классификация), деятельность исследовательской организации может быть отображена на эту таксономию для целей анализа и планирования. Эффективность такого отображения зависит от уровня обобщения. Мы предлагаем метод, включающий два этапа обобщения: один через выявление кластеров тематических единиц, второй — через оптимальную «постановку» кластеров в структуре таксономии. Исходные данные формируются в виде матриц связи между предметными единицами на основе информации о том, какие темы развиваются отдельными исследователями. Эта работа поддержана Португальским Фондом науки и техники и опирается на собранные нами данные о некоторых департаментах информатики в Португалии и Великобритании.

Введение: ABM-классификация и ее использование

Классификация тематических единиц информатики международной Ассоциации вычислительных машин (ABM-классификация) [1] делит информатику на 11 категорий первого уровня таких как «хардвер», «софтвер», «данные», «теория вычислений», «математика вычислений», «информационные системы», «вычислительные методы», «приложения». Эти категории подразделяются на 81 более мелких предметов второго уровня; из них только 59 не сводятся к тривиальным рубрикам типа «разное». Например, категория «вычислительные методы» включает такие темы как «символическое и алгебраическое манипулирование», «искусственный интеллект», «компьютерная графика», «распознавание образов». Распознавание образов, в свою очередь, делится на единицы третьего уровня: «модели», «кластер-анализ» и пр.

Такие таксономии обычно используются для аннотации и поиска документов или публикаций в различных коллекциях, как это делается для ABM-классификации на веб-портале Ассоциации вычислительных машин [1]. Некоторые другие применения:

1) стандарт для автоматически выявляемых онтологий [2];
2) определение семантического сходства при информационном поиске [3] или электронном обучении [4];
3) средство ассоциации между потребностями пользователей программного обеспечения и исследователей, создающих новое обеспечение [5].

Мы предлагаем еще одно направление использования ABM-классификации — для анализа направлений исследований, хотя, конечно, наш метод может быть использован и в других предметных областях. Следует отметить, что существующие практические системы анализа и оценки исследовательской деятельности ориентированы, прежде всего, на анализ индивидуальной активности (см., например, систему RAE в Великобритании [8]), тогда как здесь акцент делается на интегральном представлении организации как целого, что может быть полезно для таких задач как

1) обзор научной деятельности организации;
2) позиционирование организации в ABM-классификации.
3) обзор разработок в стране или иной территориальной единице, с возможностью количественной оценки того, насколько достаточен или наоборот избыточен уровень усилий в том или ином направлении.
4) анализ направлений, не вписывающихся в таксономию, что потенциально может вести к накоплению качества и новым точкам роста или иным неожиданным продвижениям;
5) планирование инвестиций и структурных изменений.

Метод кластер–постановка

Данная работа включает следующие элементы:

1) разработка электронного интерфейса для того, чтобы каждый член организации мог самостоятельно отобрать тематические единицы, относящиеся к его научным разработкам, и оценить степень интенсивности работы по каждой из них (в принципе, такого рода информация может быть получена на основе анализа документов в интернете, однако это может применяться только в ситуациях, когда все работы хорошо представлены такими документами);
2) метод вычисления сходства между тематическими единицами;
3) метод для отыскания, возможно пересекающихся, кластеров тематических единиц (экстенсиональное обобщение);

4) метод оптимальной постановки кластера тематических единиц в АВМ-классификации (интенсиональное обобщение).

Опишем вкратце последние два из них.

Пересекающиеся кластеры

Исходная информация — матрица сходства $A = (a_{ij})$ между тематическими единицами $i, j \in I$, соответствующими висячим вершинам таксономии. Мы используем подход восстановления данных, представленный для случая четких кластеров в [11], а для нечетких — в [10]. Здесь ограничимся только случаем четких кластеров, которые отыскиваются по одному так, чтобы максимизировать отношения Рэлея

$$g(S) = \frac{s^{\intercal} A s}{s^{\intercal} s} = a(S)|S|. \qquad (1)$$

где $s = (s_i)_{i \in I}$ — бинарный индикатор кластера S ($s_i = 1$ если $i \in S$ и $s_i = 0$ в противном случае), $a(S)$ — среднее сходство a_{ij} внутри S, $|S|$ — число тематических единиц в S.

Квадрат критерия равен доле квадратичного разброса данных, учитываемой кластером S [11].

Критерий локально максимизируется алгоритмом последовательного отбора объектов в/из S, начиная с произвольного $i \in I$, ADDI-S [11]. Критерием присоединения или удаления единицы j к S является результат сравнения среднего сходства j и S с адаптивным порогом $\pi = a(S)/2$, выражающего прирост критерия 1. Начиная с каждого $i \in I$, ADDI-S порождает пересекающиеся или даже совпадающие субоптимальные кластеры, которые фильтруются затем так, чтобы образовать базис дизъюнктивной кластерной модели матрицы A.

Оптимальная постановка

Рассмотрим типичный случай, представленный на рис. 1: кластер тематических единиц (черные висячие вершины) данной организации не совсем вписывается в структуру таксономии, так как распределен между тремя ее кустами. Ясно, что ему соответствует более общая категория — но ее нет в таксономии. В варианте (А) кластер представлен двумя более общими категориями — ценой введения нескольких пробелов в них и даже одного выброса — в средний куст, не охватываемый выбранными категориями. В варианте (В) кластер поднят еще выше, до категории, охватывающей все три куста, но со значительно увеличившимся количеством пробелов. Оба варианта — компромиссные, не точно учитывающие реальный кластер, что напоминает задачу о постановке музыкального голоса.

Для формализации этой ситуации введем понятие категории-направления — вершины таксономического дерева, принимаемой в качестве обобщенной характеристики кластера. С каждой категорией-направлением автоматически связывается число пробелов — не членов кластера среди ее детей. Будем штрафовать каждую категорию-направление величиной p, каждый пробел — величиной q, и каждый выброс (элемент кластера, не покрываемый категориями-направлениями) — величиной r.

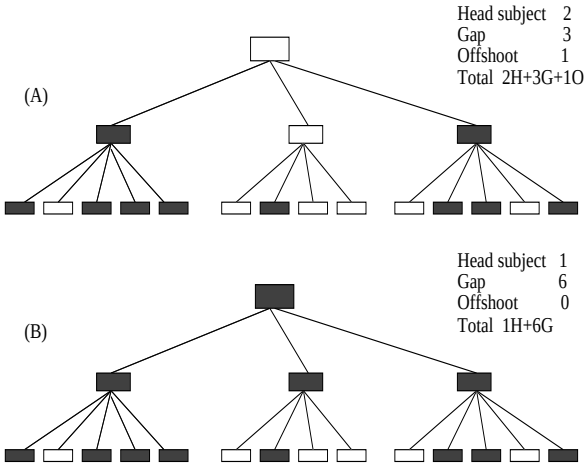


Рис. 1. Варианты постановок кластера в иерархии: постановка (В) более экономна, чем (А), если пробелы значительно дешевле, чем категории-направления.

Задача об оптимальной постановке: выбрать такие категории-направления, которые минимизируют суммарную величину штрафа.

Решение задачи может быть получено с помощью рекурсивного алгоритма вычисления оптимальной постановки в родительской вершине дерева таксономии по оптимальным постановкам в детях, как это сделано в [9], где для другой прикладной области подобная задача решена в условиях бинарного дерева и при отсутствии выбросов. Особенностью этого алгоритма является необходимость проведения вычислений для каждого из двух различных предположений о природе родительской вершины:

1) категория-направление унаследована ею от своего родителя;
2) категория-направление не унаследована, но появилась именно в родительской вершине.

Пример применения

Изложенный подход был применён к данным о 49 членах Департамента информатики Нового университета Лиссабона (Португалия). Для простоты используются только показатели сходства между вершинами второго, а не третьего, уровня АВМ-классификации; их оказалось 26 из 59. С помощью алгоритма ADDI-S получено 6 значимых кластеров $C_1, \ldots, C_6$ (здесь и далее используются обозначения АВМ-классификации [1]):

C_1: вклад 27.08%, интенсивность 2.17,
4 элемента: D3, F1, F3, F4;

C_2: вклад 17.34%, интенсивность 0.52,
12 элементов: C2, D1, D2, D3, D4, F3, F4, H2, H3, H5, I2, I6;
C_3: вклад 5.13%, интенсивность 1.33,
3 элемента: C1, C2, C3;
C_4: вклад 4.42%, интенсивность 0.36,
9 элементов: F4, G1, H2, I2, I3, I4, I5, I6, I7;
C_5: вклад 4.03%, интенсивность 0.65,
5 элементов: E1, F2, H2, H3, H4;
C_6: вклад 4.00%, интенсивность 0.64,
5 элементов: C4, D1, D2, D4, K6.

Оптимальная постановка полученных кластеров, в основном, использует соответствующие категории-направления:

F: теория вычислений (C_1);
C: организация вычислительных систем (C_3);
I: вычислительные методы (C_4);
H: информационные системы (C_5);
D: программное обеспечение (C_6).

Единственное исключение — кластер C_2, представляемый двумя категориями-направлениями: D и H. Это противоречие структуре таксономии, по-видимому, объясняется тем, что в последние годы тема «Инженерия программного обеспечения», охватывающая эти два направления, и имеющая третий ранг в ABM-классификации, превратилась в дисциплину первого ранга — что следовало бы учесть в новой структуре. Тот факт, что этот кластер имеет выбросы во все остальные направления, развиваемые в департаменте, может интерпретироваться как определенная его центральность, обеспечивающая его единство.

Заключение

Данный метод может рассматриваться как метод профилирования организаций, в котором обобщение производится для обеих сторон процесса — экстенсиональной и интенсиональной. Принципиальным является то, что вся работа ведется только в терминах таксономии.

Очевидно, профилирование организации в терминах категорий-направлений может быть сделано более информативным, если специально выделять те из них, которые оказались успешными по тем или иным параметрам (внедрение, цитирование и пр.).

Потенциально данный метод мог бы стать полезным инструментом интеграции и визуализации в анализе деятельности научных и других организаций.

Литература

[1] The ACM Computing Classification System, 1998, `www.acm.org/class/1998/ccs98.html`.

[2] *Thorne C., Zhu J., Uren V.* Extracting domain ontologies with CORDER, Tech. Report kmi-05-14 // Open University, 2005, Pp. 1–15.

[3] *Miralaei S., Ghorbani A.* Category-based similarity algorithm for semantic similarity in multi-agent information sharing systems // IEEE/WIC/ACM International Conference on Intelligent Agent Technology, 2005, Pp. 242–245.

[4] *Yang L., Ball M., Bhavsar V., Boley H.* Weighted partonomy-taxonomy trees with local similarity measures for semantic buyer-seller match-making // Journal of Business and Technology. Atlantic Academic Press, 2005, 1 (1), Pp. 42–52.

[5] *Feather M., Menzies T., Connelly J.* Matching software practitioner needs to researcher activities // Proc. of the 10th Asia-Pacific Software Engineering Conference (APSEC'03), IEEE, 2003, 6.

[6] *Weiss S. M., Indurkhya N., Zhang T., Damerau F. J.* Text mining: predictive methods for analyzing unstructured information, Springer Verlag, 2005, 237 p.

[7] *Middleton S., Shadbolt N., Roure D.* Ontological user representing in recommender systems // ACM Trans. on Inform. Systems, 2004, 22 (1), Pp. 54–88.

[8] RAE2008: Research Assessment Exercise, 2007, `www.rae.ac.uk`.

[9] *Mirkin B., Fenner T., Galperin M., Koonin E.* Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes // BMC Evolutionary Biology, 2003, 3:2.

[10] *Nascimento S., Mirkin B., Moura-Pires F.* Modeling proportional membership in fuzzy clustering // IEEE Trans. on Fuzzy Systems, 2003, 11 (2), Pp.173–186.

[11] *Mirkin B.* Clustering for Data Mining: A Data Recovery Approach. Chapman & Hall/ CRC Press, 2005, 276 p.

Морфология и синтаксис в задаче семантической кластеризации*

Михайлов Д. В., Емельянов Г. М.
Dmitry.Mikhaylov@novsu.ru
Великий Новгород, ГОУ ВПО НовГУ им. Ярослава Мудрого

Рассматривается задача семантической кластеризации текстов Естественного Языка (ЕЯ). На основе теории Анализа Формальных Понятий предложен подход к выработке качественных оценок моделей морфологии и синтаксиса как инструментальных средств выделения объектов и признаков.

Одна из центральных задач понимания текста — выделение класса Семантической Эквивалентности (СЭ). В общих чертах установить факт СЭ означает доказать идентичность ролей сходных понятий относительно сходных ситуаций.

Наиболее близка данной идее обработка текстов на основе коммуникативной грамматики. Хорошим примером является поисковая система Exactus [5].

Тем не менее, существуют задачи сравнения смысла, отличные от традиционного для поисковых систем взаимодействия «запрос–ответ».

Примером является тестовое задание открытой формы в системе контроля знаний [3]. Необходимо не столько отобразить ответ на предметную область, сколько оценить близость ответу, «правильному» с точки зрения разработчика теста. Анализ взаимной близости ответов здесь требует учета лексико-функциональной синонимии, в частности — расщепленных значений и конверсивов [3].

Актуальная *глобальная задача*, которой посвящена настоящая работа — автоматизация накопления знаний о взаимодействии семантики, синтаксиса и морфологии, необходимых для установления СЭ, непосредственно по ЕЯ-текстам.

Постановка проблемы

Сформулируем задачу СЭ следующим образом.

Пусть G есть множество ЕЯ-текстов. По результатам синтаксического разбора каждого $T_i \in G$ требуется выявить:

- множество $V(T_i)$ *ситуаций*, описываемых T_i;
- множество $M(T_i)$ *объектов* и/или *понятий*, значимых в ситуациях из множества $V(T_i)$;
- тернарное отношение $I \subseteq G \times M \times V$, ставящее в соответствие каждому $m \in M,\ M = \bigcup_i M(T_i)$, ту ситуацию $v \in V,\ V = \bigcup_i V(T_i)$, в которой он фигурирует относительно T_i.

Множества M и V выделяются на основе *синтаксических контекстов существительных* — последовательностей соподчиненных слов вида

$$S_{ki} = \{v_1, \ldots, v_{n(k,i)}, m_{ki}\}. \qquad (1)$$

Здесь $v_1 \in V(T_i)$ является предикатом (глаголом или словом, производным от него). Существительное m_{ki} обозначает некоторое понятие, значимое в ситуации v_1. Индекс k есть порядковый номер последовательности среди выявленных из T_i. Целочисленное значение $n(k,i)$ равно количеству соподчиненных существительных $\{v_2, \ldots, v_{n(k,i)}, m_{ki}\}$.

Кроме того, для всех $\{v_l, v_{l+1}\} \subset S_{ki}$ существует *синтаксическое отношение* R_q:

$$v_l R_q v_{l+1}, \quad v_{n(k,i)} R_q m_{ki}, \qquad (2)$$

тип q которого определяется предлогом для связи главного слова с зависимым и падежом зависимого.

Транзитивность отношения R_q дает основание утверждать, что $\{v_2, \ldots, m_{ki}\} \subset M(T_i)$. В конечном итоге, тип указанного отношения между v_1 и словом справа от него в (1) определяет роль относительно v_1 для каждого $m \in \{v_2, \ldots, m_{ki}\}$.

На основе I выделяются группы текстов, сходных по встречаемости объектов в одних и тех же ситуациях. Данная задача наиболее естественно решается методами Анализа Формальных Понятий (АФП, [2]). При этом для $A \subseteq G$ и $B \subseteq M \times V$ вводится пара отображений:

$$A' = \big\{(m,v)\colon m \in M,\ v \in V \;\big|\; \forall\, T_i \in A\colon m(T_i) = v\big\};$$
$$B' = \big\{T_i \in G \;\big|\; \forall\,(m,v) \in B\colon m(T_i) = v\big\}.$$

Пара множеств (A, B) таких, что $A' = B$ и $B' = A$, называется *формальным понятием* (ФП).

Тернарному отношению I здесь ставится в соответствие *формальный контекст* $K = (G, M, V, I)$, для которого строится *решетка ФП* $\mathrm{Re}(G, M, V, I)$.

Визуализация Re диаграммой линий [2] позволяет графически отображать группировку текстов.

Тем не менее, актуальной является проблема точности синтаксического анализа как инструмента выделения понятий и их признаков. Известные синтаксические анализаторы реализуют стратегию разбора на основе наиболее вероятных связей [1].

Вместе с тем, часто требуется исследовать природу выявляемых синтаксических связей. При неправильном разборе нужно установить причину использования той или иной стратегии (правила) с учетом особенности отражения ситуации, описываемой анализируемой фразой, в заданном ЕЯ.

Целью настоящей работы *является* разработка модели автоматического выделения и классификации наиболее вероятных синтаксических связей для множества СЭ-фраз.

*Работа выполнена при финансовой поддержке РФФИ, проект № 06-01-00028.

Методы решения

Предлагаемое решение поставленной проблемы основано на закономерностях выражения смысла в заданном ЕЯ его носителем.

Как уже обсуждалось нами ранее [3], языковой опыт человека можно разделить в соответствии с разделением концептуальной картины мира. При этом основополагающим является понятие ситуации употребления ЕЯ как основы его генезиса.

Под *ситуацией употребления ЕЯ* понимают описание нового социального опыта (содержания совместных действий) средствами этого ЕЯ [3].

Указанное описание выполняется в некоторой знаковой системе с целью обобщения и передачи знаний от человека к человеку.

Формально фиксируемый ситуацией S языковой контекст представляется тройкой:

$$S = (O, R, T), \tag{3}$$

где O есть множество объектов-участников S, R — множество отношений между $o \in O$, $T \subset G$ — множество форм языкового описания S.

Предположим, что T состоит из синонимичных фраз, каждая из которых описывает одну и ту же ситуацию действительности (относительно языкового контекста S). Выбор $T_i \in T$ для описания S является равновероятным. В силу произвольности R предположим, что его элементами являются синтаксические отношения вида (2).

При этом все ЕЯ-фразы из T являются строго синонимичными, а

$$O = \bigcup_{T_i \in T} \{M(T_i) \cup V(T_i)\}.$$

Поскольку S есть (по определению) полное и независимое описание контекста, то имеем задачу.

Задача 1. На основе ЕЯ-фраз из T найти R, используя отношения между $o \in O$ в качестве признаков последних относительно (3).

Рассмотрим текст $T_i \in T$ как множество символов. Тогда для любого $T_i \in T$ справедливо:

$$T_i = T_i^C \cup T_i^F,$$

где T_i^C — общая неизменная часть для всех $T_i \in T$, T_i^F — изменяемая часть. На множестве T_i^F выражаются синтагматические зависимости, которые определяют возможность сосуществования словоформ в линейном ряду и задаются с помощью R.

Пусть W_{ij} — буквенный состав слова, j — его порядковый номер в ЕЯ-фразе. Тогда

$$W_{ij} = W_{ij}^C \cup W_{ij}^F, \text{ где} \tag{4}$$

$W_{ij}^C \subset T_i^C$ — неизменная, $W_{ij}^F \subset T_i^F$ — флективная часть, изменяемая при склонении (спряжении).

Таким образом, на основе попарного сравнения W_{ij} различных T_i требуется найти:

1) W_{ij}^C и W_{ij}^F каждого W_{ij} при $\left|W_{ij}^C\right| \to \max$;
2) Отношение R_q, определяющее допустимость сочетания $\left(W_{ij}^F, W_{ik}^F\right)$, $k \neq j$.

Введем в рассмотрение индексное множество J для неизменных частей всех слов, употребленных во всех фразах из T.

Определение 1. *Моделью L линейной структуры предложения $T_i \in T$ назовем последовательность индексов $j \in J$ неизменных частей слов, присутствующих в T_i.*

При этом порядок индексов в L идентичен порядку следования соответствующих слов в T_i. Поэтому $L(T_i)$ позволяет однозначно восстановить ЕЯ-фразу T_i на множестве всех слов для всех фраз из T. И наоборот, для $\forall T_i \in T$ на индексном множестве J можно однозначно построить $L(T_i)$.

Для формирования множества R в (3) необходимо найти совокупность указанных моделей, удовлетворяющих требованиям проективности [4].

Модель L следует считать проективной в содержательном смысле, если все стрелки выявленных синтаксических связей могут быть проведены без пересечений по одну сторону прямой, на которой записана L. Кроме того, если из позиции некоторого индекса выходят несколько стрелок, то эту позицию не должны накрывать стрелки, выходящие из позиций других индексов.

С учетом линейной природы синтагм дополним вышеуказанные требования следующим образом.

Пусть $h(j, L(T_i))$ — позиция индекса j в модели $L(T_i)$. Тогда множество связей относительно $L(T_i)$

$$D : T_i \to \left\{ \Big(h\big(j, L(T_i)\big),\, h\big(k, L(T_i)\big) \Big) :\ j \neq k \right\}.$$

Определение 2. *Связь*

$$d_{qi} = \Big(h\big(j, L(T_i)\big),\, h\big(k, L(T_i)\big) \Big)$$

является допустимой для модели $L(T_i)$, если

$$\exists \{T_l, T_m\} \subset T, \quad l \neq m,$$

такие, что и $L(T_l)$, и $L(T_m)$ содержат в качестве подпоследовательности либо $\{j, k\}$, либо $\{k, j\}$.

При этом пара индексов (j, k) соответствует одной синтагме, а индекс q — типу синтаксического отношения, которое ей соответствует.

Положим, что для всех $T_i \in T$, $i = 1, \ldots, |T|$, все $d_{qi} \in D(T_i)$ удовлетворяют Определению 2.

Определение 3. *Будем считать, что модель $L(T_i)$ проективна относительно R в (3), если*

$$\sum_{q=1}^{|D(T_i)|} \Delta_{qi} \leqslant \left|L(T_i)\right|, \text{ где}$$

$$\Delta_{qi} = \left|h(j, L(T_i)) - h(k, L(T_i))\right|.$$

На основе $\bigcup_i D(T_i)$ формируется граф синтагм (V^J, I^J). Элементами множества вершин V^J этого графа являются множества пар (j,k), $\{j,k\} \subset J$, сгруппированных по некоторому общему для них индексу k. Множества E_1 и E_2, входящие в V^J, будут соединены ребром из I^J, если $\exists\, \{j,k,m\} \subset J$: $(j,k) \in E_1$, $(k,m) \in E_2$ и $j \neq m$.

Анализом $\left(V^J, I^J\right)$ строится дерево-прецедент $\left(V_1^J, I_1^J\right)$ для $\bigcup_i T_i$, $i = 1, \ldots, |T|$. Формально

$$V_1^J = J,\ I_1^J = \left\{(j,k) \colon \exists\, E \in V^J, (j,k) \in E\right\}. \quad (5)$$

При этом индекс $k \in V_1^J$ соответствует корню дерева $\left(V_1^J, I_1^J\right)$, если $\exists\, E_1 \in V^J$, в котором пары индексов сгруппированы по k, $|E_1| > 1$, а k не содержится ни в одной паре индексов для $\forall\, E_2 \in V^J$: $E_1 \neq E_2$.

Содержательно корень соответствует предикатному слову в (1), которое (по определению) обозначает ситуацию. Поскольку исследуемая проблема точности синтаксического анализа характерна для ситуаций с двумя и более участниками, то число дочерних узлов у корня полагается больше одного.

Будем использовать маршруты в дереве (5) для выделения классов отношений из R в (3) согласно сформулированной нами Задаче 1.

Пусть

$$G^F = \left\{f_{ij} \colon f_{ij} = \odot\left(W_{ij}^F\right)\right\},$$
$$I^F = \left\{(f_{ij}, f_{ik}) \colon s(j,k) = \text{true},\ \{j,k\} \subset J\right\}.$$

Здесь $\odot$ есть конкатенация, последовательно выполняемая над символами из W_{ij}^F в (4). Отношение s задается рекурсивно на основе (V^J, I^J):

1) $s(j_1, j_1) = \text{true}$;
2) $s(j_1, j_2) = \text{true}$, если:
   — либо $\exists\, E_1 \in V^J$: $(j_1, j_2) \in E_1$, причем $\exists\, j_3 \in J$, для которого $s(j_2, j_3) = \text{true}$;
   — либо $\exists\left(E_1, E_2\right) \in I^J$: $\exists\, j_3 \in J$, при этом $(j_1, j_3) \in E_1$, $(j_3, j_2) \in E_2$, а $s(j_3, j_2) = \text{true}$.

Введем в рассмотрение формальный контекст:

$$K^F = \left(G^F, M^F, I^F\right), \quad (6)$$

в котором $M^F = G^F$, а $I^F \subseteq G^F \times M^F$.

Модель (6) выделяет классы в R по характеру изменения флективной части зависимого слова в каждом $R_q \in R$ с учетом бинарности последнего.

Рассмотрим задачу поиска флексий для слов в составе расщепленных значений и конверсивов. Будем рассматривать Расщепленное Предикатное Значение (РПЗ) — совокупность вспомогательного глагола (связки) и некоторого существительного, называющего ситуацию. Для РПЗ, как и для конверсивов (слов, обозначающих ситуацию с точки зрения разных ее участников) представления вида (4) не могут быть найдены попарным сравнением буквенного состава слов во всех $T_i \in T$.

Пусть $W_k^P \in T_i$ — последовательность символов слова, для которого не найдено представления (4).

Рассмотрим

$$T_i^{\odot} = \left\{w_{ij} \colon w_{ij} = \odot(W_{ij})\right\}.$$

Положим также, что $\exists\, T_i^P \subset T_i$, определяющее последовательность

$$P_i^{\odot} = \left\{u_k \colon u_k = \odot\left(W_k^P\right), \bigcup_k W_k^P = T_i^P\right\}.$$

Лемма 1. *Последовательность $P_i^{\odot}$ содержит предикатное слово, если $\exists\, \{j, 0, k\} \subset L(T_i)$:*

$$\{w_{ij}, u_1, \ldots, u_p, w_{ik}\} \subset T_i^{\odot},$$

где $\{u_1, \ldots, u_p\} = P_i^{\odot}$, а $p = \left|P_i^{\odot}\right|$.

Лемма 2. *Слово $u_k \in P_i^{\odot}$ входит в состав РПЗ, если $\exists\, T_j \in T$: $L(T_j) \neq L(T_i)$, а $u_k \in P_j^{\odot}$.*

При этом $\nexists\, T_k \in T$, для которого $P_k^{\odot} \subset P_i^{\odot}$, а $L(T_k) \neq L(T_j)$ и $L(T_k) \neq L(T_i)$.

Пусть $P_i^{\odot\prime}$ — последовательность слов, каждое из которых удовлетворяет условию Леммы 2.

Теорема 1. *Для формирования контекста (6) необходимо и достаточно найти множество $T' \subset T$:*

$$T' = \left\{T_i \colon |P_i^{\odot\prime}| \to \max\right\}. \quad (7)$$

Другой критерий отбора $T_i \in T$ основан на минимизации числа слов, не представимых как (4).

Для $u_k \in \bigcup_i P_i^{\odot\prime}$: $T_i \in T'$ представление (4) формируется сравнением буквенного состава со всеми $u_j \in \bigcup_l P_l^{\odot} \colon T_l \in (T \setminus T')$. При этом необходимо, чтобы $2\left|W_k^C\right| > \left|W_k^F\right| + \left|W_j^F\right|$, где $W_k^P = W_k^C \cup W_k^F$, а $W_j^P = W_j^C \cup W_j^F$.

Дерево (5) преобразуется следующим образом с учетом вышесказанного для всех $T_i \in T'$:

1) корень изменяется с $k = 0$ на значение k для $u_k \in P_i^{\odot\prime}$, имеющего максимальную встречаемость в различных $T_i^{\odot}$;
2) левое поддерево остается без изменений;
3) правое поддерево перевешивается на узел j для $u_j \in P_i^{\odot\prime}$ наименьшей встречаемости;
4) для всех $\{u_l, u_m\} \subset P_i^{\odot\prime}$ дочерним будет узел для слова с меньшей встречаемостью.

В итоге основу формирования контекста (6) составляют T_i, которые наиболее полно описывают ситуацию S.

Экспериментальная апробация

На материале результатов теста открытой формы был проведен машинный эксперимент по выделению и классификации синтаксических отношений предложенным в работе методом.

Вопрос теста: «Каковы негативные последствия переобучения при скользящем контроле?»

Таблица 1. Правильные ответы $T_i \in T'$ в (7).

основа	флективная часть + предлог					
заниженн	ость	ости	ость	ости	ость	ости
эмпирическ	ого	ого	ого	ого	ого	ого
риск	а	а	а	а	а	а
нежелательн	ого	ое	ого	ое	ым	ое
переобучени	я	е	я	е	ем	е
явля	естя	—	ется	ется	—	—
следстви	ем	—	—	—	—	—
служ	—	ит	—	—	—	—
причин	—	ой	—	ой	—	—
результат	—	—	ом	—	—	—
связан	—	—	—	—	a:c	—
привод	—	—	—	—	—	ит:к

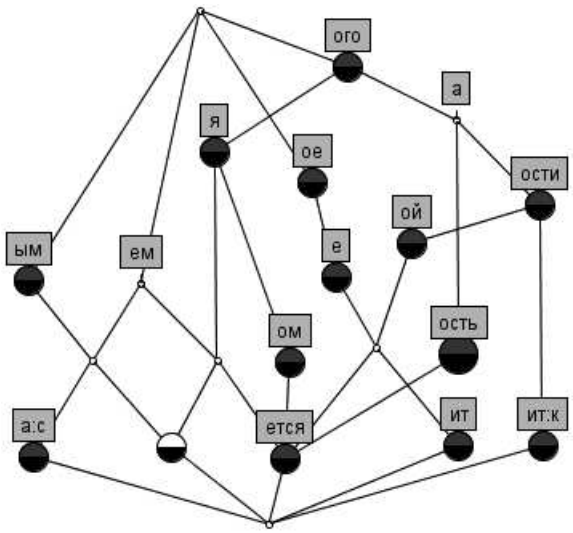


Рис. 1. Синтаксические отношения в решетке ФП.

Было получено двадцать семь вариантов правильного ответа, которые служили исходными данными при формировании контекста (6).

На рис. 1 представлена решетка Re^F для T'. При $P_i^{\odot\prime} \cap P_i^{\odot} \neq \varnothing \;\; \forall\, u_m \in \left(P_i^{\odot} \backslash P_i^{\odot\prime}\right)$ есть предлог и представляется вместе со словом $u_l \in P_i^{\odot\prime}$, стоящим слева от него в $P_i^{\odot}$, см. таблицу 1.

Содержательная интерпретация Re^F может быть получена выделением морфологических классов слов с учетом структуры последовательностей (1) согласно приведенным ниже правилам.

Пусть $\mathcal{L}$ — базис импликаций [2] для K^F из (6).

Правило 4. ФП $(A^F, B^F) : A^F \subseteq G^F, B^F \subseteq M^F$, соответствует предикатному слову в (1), если

$$\exists\,(Pr \to Cs) \in \mathcal{L}\colon |Pr| = 1,\ Pr \subset B^F \text{ и}$$
$$\exists\,(Pr_1 \to Cs_1) \in \mathcal{L}\colon Pr \subset Cs_1.$$

При этом $Pr \cup Cs = B^F,\ \ Pr_1 \cup Cs_1 = B^F$.

Правило 5. ФП (A^F, B^F) соответствует прилагательному для m_{ki} в (1), если $\exists\,(Pr \to Cs) \in \mathcal{L}$, причем $B^F \backslash Cs = \varnothing$.

В противном случае ФП (A^F, B^F) соответствует существительному из $\{v_2, \ldots, m_{ki}\} \subset S_{ki}$.

Синтаксические отношения выделяются анализом наименьшей верхней грани каждой пары ФП в Re^F и образуют классы по сходству характера флексии зависимого слова. Отдельному классу соответствует область в решетке, а наименьшая верхняя грань множества ФП этой области — прецеденту класса.

В примере на рис. 1 классы отношений соответствуют словоизменению прилагательных (нежелательн-ого, эмпирическ-ого) и существительных в составе генитивных конструкций (результат-ом переобучени-я, следстви-ем переобучени-я). Последний в силу транзитивности отношений (2) может включать сочетания существительного (вне генитивных конструкций) с глаголом.

Поскольку основу формирования решетки составляют те ЕЯ-фразы, которые максимально точно описывают ситуацию, а значит и более четко передают смысл, то выявленные отношения будут соответствовать искомым наиболее вероятным синтаксическим связям относительно (3).

Заключение

Предложенная в работе модель позволяет решить две важные задачи, актуальные для семантической кластеризации ЕЯ-текстов.

Во-первых, автоматически выделить лучший способ выражения нужной мысли в заданном ЕЯ, что позволит избежать ошибок синтаксического анализа при использовании его как инструмента формирования объектов и признаков.

Во-вторых, автоматизировать разработку синтаксических стратегий и правил при исследовании случаев применения определенных грамматических конструкций в тематическом корпусе текстов. Качественные оценки формируемых знаний здесь могут быть даны на основе мер схожести решеток по аналогии с мерами схожести для ФП [3].

Литература

[1] `http://www.aot.ru` — 2009.

[2] *Ganter B., Wille B.* Formal Concept Analysis — Mathematical Foundations. — Berlin: Springer-Verlag, 1999. — 284 с.

[3] *Mikhailov D. V., Emelyanov G. M., Stepanova N. A.* Formation and clustering of Russian's nouns's contexts within the frameworks of Splintered Values // 9^{th} Int. Conf. PRIA-9-2008. — Nizhni Novgorod: NNSU, 2008. — Vol. 2. — P. 39–42.

[4] *Кибрик А. Е.* Очерки по общим и прикладным вопросам языкознания. — М.: КомКнига, 2005. — 336 с.

[5] *Осипов Г. С., Тихомиров И. А., Смирнов И. В.* Exactus — система интеллектуального метапоиска в сети Интернет // 10-я конф. КИИ-2006. — М.: Физматлит, 2006. — Т. 3. — С. 859–866.

Математические модели и алгоритмы в задачах атрибуции фольклорных текстов

Москин Н. Д.
moskin@karelia.ru
Петрозаводск, Петрозаводский государственный университет

В данной статье показано, как можно применить теоретико-графовые модели в задачах атрибуции фольклорных текстов и их сравнительного анализа. Также в работе приводится краткое описание информационной системы «Фольклор», в которой реализованы алгоритмы решения подобных задач.

При исследовании коллекций фольклорных текстов часто возникают сложные проблемы, которые трудно решить традиционными методами. К таким проблемам можно отнести жанровую дифференциацию и атрибуцию текстов, обнаружение устойчивых языковых конструкций (мотивов) и их классификацию, реконструкцию текстов.

С 2001 года на кафедре информатики и математического обеспечения Петрозаводского государственного университета ведется работа над исследованием фольклорных коллекций с помощью математических методов и компьютерных технологий. Среди таких коллекций можно выделить корпус бесёдных песен Заонежья XIX – начала XX века [6]. Каждой песне ставится в соответствие набор характеристик: фамилия, имя, отчество автора и собирателя, год и место записи, вид, жанр, тема, темп, движение в танце и др. К настоящему времени ряд характеристик песен оказались утраченными или изначально не были зафиксированы, поэтому возникает задача атрибуции этих характеристик. Другой интересной проблемой является обнаружение в текстах данной коллекции схожих сюжетов и их классификация.

Для решения этих задач мы предлагаем использовать теоретико-графовые модели. В гуманитарных областях знаний графы используются для автоматической обработки текстов, информационного поиска, реферирования и индексирования текстов, автоматического перевода, стилистической диагностики, в задачах атрибуции анонимных текстов и т. д. [4, 7, 8] В фольклористике графы применялись крайне мало, такие работы единичны [1].

Моделирование семантической структуры фольклорных песен с помощью графов

Бесёдная песня, как и другой фольклорный текст, состоит из устойчивых фрагментов (мотивов), которые повторяются в других текстах в разной последовательности, образуя таким образом новые сюжеты. По выражению известного фольклориста Б. Н. Путилова мотив является «узловой категорией художественной организации произведения фольклора».

Содержательную основу мотива можно представить в виде помеченного мультиграфа, в узлах которого находятся основные персонажи песни, животные, явления природы, предметы обихода и т. д. Между объектами устанавливаются связи двух видов: локальные и глобальные, соответствующие синтагматическим и парадигматическим отношениям в тексте (на рисунке они отмечены сплошной линией и пунктиром соответственно). Рассмотрим фрагмент бесёдной песни «Как назябло, навеяло лицо» из сборника В. Д. Лысанова «Досюльная свадьба, песни, игры и танцы в Заонежье Олонецкой губернии» (Петрозаводск, 1916 г.) [3]:

Красна девица во тереме сидит, да
Жемчужное ожерельицо садит; да
Разсыпалось ожерельицо, да
По всему высоку терему. Да
Не собрать, не собрать жемчуга, да
Что ль ни батюшку, ни матушки, да
Что ль ни братцам, ни ясным соколам, да
Ни сестрицам, белым лебедям, да
А собрать соберет жемчужок, да
Разудалый, добрый молодец.

Теоретико-графовая модель данного фрагмента изображена на рис. 1.

Если связать графы мотивов, объединив одинаковые объекты в одну вершину, то подобную структуру можно изобразить в виде единого графа сюжета фольклорной песни. Подробнее ознакомиться с методикой построения моделей и исходными текстами бесёдных песен можно в [5, 6].

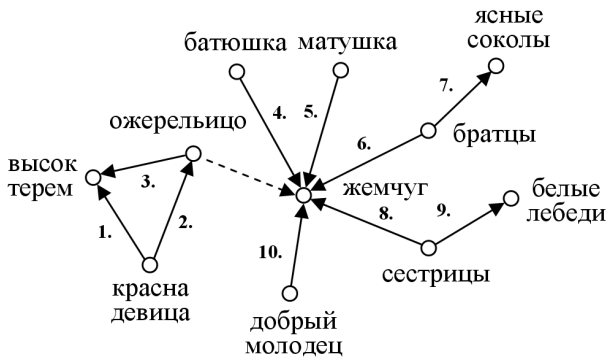


Рис. 1. Граф фрагмента песни «Как назябло, навеяло лицо».

Основные подходы к сравнению и классификации теоретико-графовых моделей

Рассмотрим задачу атрибуции фольклорных текстов в следующей постановке. Пусть задано множество песен с основными характеристиками и множество соответствующих им теоретико-графовых моделей. Граф $G = (V, E, \alpha, \beta, L_V, L_E)$, где $V = \{v_i\}_{i=1}^m$ — множество вершин, $E \subseteq V \times V$ - множество ребер. Функция $\alpha\colon V \to L_V$ задает метки вершинам графа (номера групп), $\beta\colon E \to L_E$ задает тип отношения: локальное или глобальное.

Введем расстояние $d(G_i, G_j)$ на множестве графов, которое удовлетворяет следующим свойствам метрики:

$d(G_i, G_j) \geqslant 0$ (неотрицательность);
$d(G_i, G_j) = 0 \Leftrightarrow G_i \cong G_j$;
$d(G_i, G_j) = d(G_j, G_i)$ (симметричность);
$d(G_i, G_j) \leqslant d(G_i, G_k) + d(G_k, G_j)$
(неравенство треугольника).

При определении расстояния можно использовать числовые характеристики графов: например, параметры связности семантической структуры или показатели распределения связей по вершинам. Подобным образом проводила исследование деревьев зависимости И. П. Севбо при диагностике авторского стиля художественных произведений [7].

Другой подход основан на определении степени неточностей, которые возникают при нахождении изоморфизма графов и подграфов. В зарубежной литературе данное направление получило название «graph matching» [9]. Эти методы нашли свое применение в обработке изображений, химии, молекулярной биологии, дактилоскопии и т. д.

Одним из способов оценки возможных ошибок сравнения является максимальный общий подграф. Максимальным общим подграфом графов G_1 и G_2 называется такой граф $\widehat{G}$, который изоморфен $G_1' \in G_1$ и $G_2' \in G_2$ и содержит максимальное число вершин. Тогда

$$d(G_1, G_2) = 1 - \frac{|\widehat{G}|}{\max\{|G_1|, |G_2|\}},$$

где $|G|$ обозначает число вершин в графе G. В нашем случае максимальный общий подграф можно интерпретировать как общую структуру, образующую сюжет в двух текстах.

Еще один способ количественной оценки сходства графов является мера на основе операций редактирования (вставка, удаление и переименование вершин и ребер). Эта мера является расширением известного правила сравнения строк Вагнера-Фишера. Поскольку на практике часто одни операции являются более значимыми по сравнению с другими, каждой операции ставится в соответствие ее вес $\gamma\colon \Sigma \to R^+$, где Σ — множество операций редактирования. Тогда расстоянием $d(G_1, G_2)$ определяется как последовательность $\sigma \subset \Sigma$ операций редактирования для графов G_1 и G_2, которые преобразуют один граф в другой с минимальным суммарным весом:

$$d(G_1, G_2) = \min\{\gamma(\sigma)\colon G_1 \xrightarrow{\sigma} G_2\}.$$

Эта мера позволяет отразить изменение сюжета при передачи «из уст в уста», при котором некоторые отношения утрачивались или, наоборот, добавлялись в текст, тем самым образуя новые варианты. Чтобы учитывать порядок появления ребер в графе автором была предложена модификация этих метрик [5].

Далее нужно установить, какая метрика (а, следовательно, структурные особенности графа) связана с той или иной характеристикой песни. Для этого строилась матрица попарных расстояний $D = (d_{ij})_{i,j=1}^n$ между графами $G_1, G_2, \ldots, G_n$, которая затем анализировалась при помощи методов многомерного шкалирования пакета Statistica 6.0, а также иерархического алгоритма кластерного анализа, где расстояние между кластерами задается по методу ближнего соседа.

Поскольку фольклорный текст вариативен, т. е. существует несколько вариантов одного и того же текста, записанных разными собирателями или в разных местах, то в этом случае для их представления был использован «средний граф» [10]:

$$\widetilde{G} = \arg\min_{G \in H} \sum_{i=1}^{k} d(G, G_i),$$

где $H = \{G_1, G_2, \ldots, G_k\}$ — множество графов, соответствующих различным вариантам одного текста. Также с помощью среднего графа можно выделить наиболее типичные тексты с заданной характеристикой. Например, типичная «семейная» песня (характеристика «тема») — «Все мужовья до жен добры» из сборника Ф. Студитского [6].

Агрегация графов

При определении расстояния между графами могут возникнуть следующие проблемы:

— если размерность графа велика, то время выполнения алгоритма занимает очень много времени;
— в фольклорном тексте могут присутствовать лишние несущественные отношения (например, при повторах), которые следует отбросить при анализе.

Для решения этих проблем мы предлагаем использовать агрегирующие графы основных потоков связей с небольшим числом вершин и ребер.

Один из методов построения подобных графов предложен в работе [2]. В этом методе накладывается следующее ограничение: разбиение вершин графа $v_i \in V$ осуществляется на непересекающиеся группы, объединение которых дает исходное множество V. Число групп можно определить, например, по числу мотивов песни.

Рассмотрим матрицу смежностей для графа G, $A = (a_{ij})_{i=1}^{m}{}_{j=1}^{m}$ между m вершинами из множества V. Обозначим $R^t = \{R_1, R_2, \ldots, R_t\}$ — разбиение множества V на t непустых непересекающихся классов. Тогда задача построения агрегированного графа состоит в том, чтобы максимизировать функционал:

$$F(R^t, r) = \sum_{k=1}^{t}\sum_{l=1}^{t} r_{kl} \sum_{i\in R_k}\sum_{j\in R_l} (a_{ij} - \alpha),$$

где α — порог значимости показателей связи, который определяется в зависимости от характера исследования, а элементы матрицы $r = (r_{kl})_{k=1}^{t}{}_{l=1}^{t}$ находятся следующим образом:

$$r_{kl} = \begin{cases} 1, & \text{если } \sum\limits_{i\in R_k}\sum\limits_{j\in R_l} (a_{ij} - \alpha) \geqslant 0; \\ 0, & \text{иначе.} \end{cases}$$

Алгоритм, описанный в [2], на первом этапе строит начальное разбиение R^t, а затем на втором этапе производится локальное улучшение начального приближения. При исследовании фольклорных песен пороговое значение было экспериментально установлено равным 0.2 (при таком значении большинство связей попадает в основные потоки). На рисунке 2 приведен пример агрегирующего графа для песни «Девушка в горенке сидела» [6], где число t равно 5.

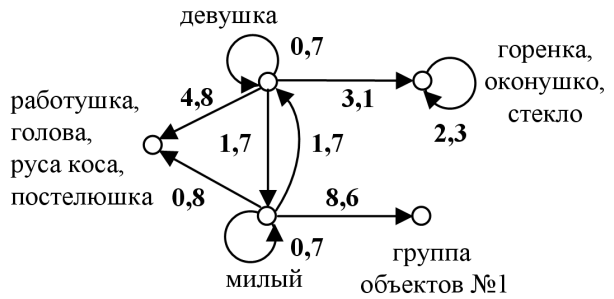


Рис. 2. Граф основных потоков связей для песни «Девушка в горенке сидела».

Здесь группа объектов № 1 — это «поле», «ясен сокол», «белые руки», «желтые кудри», «резвые ноги», «высокий терем», то есть все объекты, связанные с «милым дружком».

Информационная система «Фольклор»

Для ввода, редактирования и анализа фольклорных текстов и их теоретико-графовых моделей необходимо специальное программное обеспечение. С этой целью была разработана информационная система «Фольклор» в среде визуального программирования Delphi 7.0 [5]. Система содержит следующие функциональные модули.

1. Модуль ввода и редактирования фольклорных текстов.
2. Модуль анализа текстов. Включает процедуры графематического, морфологического и контент-анализа текстов, встроенный морфологический словарь.
3. Модуль автоматизированного построения теоретико-графовых моделей. В системе реализована следующая пошаговая процедура:
   шаг 1: выбор параметров построения графа;
   шаг 2: определение объектов в тексте;
   шаг 3: разбиение объектов на группы;
   шаг 4: определение связей в тексте;
   шаг 5: разбиение связей на группы.
   Пользователь может в любой момент скорректировать полученный граф (например, удалить или добавить связи и объекты, изменить их свойства и т. д.).
4. Модуль визуализации графов. Включает методы двухмерной и трехмерной визуализации теоретико-графовых моделей, основанные на физических аналогиях, а также процедуру поуровневого изображения деревьев.
5. Модуль агрегации графов (алгоритмы Мучника и Куперштоха-Трофимова).
6. Модуль сравнения и классификации теоретико-графовых моделей. Содержит модуль определения параметров графов (коэффициент связности, распределение объектов и связей на группы, функциональные веса вершин и др.), модуль вычисления метрик на множестве деревьев, модуль вычисления метрик на основе общих подграфов и операций редактирования, модуль кластерного анализа.

В настоящее время информационная система содержит 562 текста из четырех фольклорных коллекций (песни, былины, духовные стихи и записи о народных святых Нижегородского края). При этом теоретико-графовые модели, хранящиеся в системе, могут быть построены по разным принципам. Например, духовные стихи и былины представлены в виде деревьев зависимости, отражающих их синтаксическую организацию.

В данный момент идет работа над разработкой усовершенствованного формата хранения теоретико-графовых моделей (с использованием технологии XML), вводом в систему новых фольк-

лорных текстов, улучшением метода визуализации графов (с учетом упорядоченности и иерархичности их элементов), разработкой процедуры поиска схожих мотивов, основанной на модификации алгоритма Ульмана поиска изоморфизма подграфу.

Результаты анализа текстов

На основе коллекции бесёдных песен была составлена выборка из 50 текстов, собранных из разных источников [6]. В результате исследования были получены следующие закономерности.

1. Характеристика «темп» песни (частый, быстрый, медленный, протяжный и т. д.) связана с числом вершин и ребер в графе. Например, если в графе песни число вершин $m > 14$ и число ребер $n > 17$, то эта песня с большой вероятностью исполнялась в быстром темпе.
2. Характеристика «тема» (любовная, семейная, хвалебная, шуточная и т. д.) связана с распределением объектов песни на группы. Например, в любовных песнях, чаще чем в остальных, встречаются объекты групп «части человеческого тела», «проявление качеств человека» и «земля и воды», а для семейных песен характерны группы «разные предметы» и «конструкции», почти не встречаются объекты группы «проявление качеств человека».
3. Характеристика «вид» (бесёдная, плясовая, свадебная, бытовая и т. д.) связано с распределением объектов по числу их связей. Например, для песен вида «бесёдная», «свадебная бесёдная» и «плясовая» в среднем значение $\Delta(G)$ (максимальная степень вершины) равно 6.5 и не превышает 10 и т. д.

На основе полученных закономерностей можно либо установить вероятное значение пропущенной характеристики, либо диапазон возможных значений, либо отбросить неподходящие значения. Например, с большой вероятностью можно предположить, что песня «Уж ты Ванюшка-Иван» в записи Е. В. Барсова, где $m = 21$ и $n = 26$, имеет «быстрый» темп исполнения (эта характеристика чаще всего не фиксировалась собирателем). С другой стороны, песня «Право, матушка, мне тошненько» в записи О. Х. Агреневой-Славянской не имеет «семейную» тему, поскольку в ней нет объектов групп «разные предметы» и «конструкции», зато есть объект группы «проявление качеств человека».

В данном исследовании при построении теоретико-графовых моделей и интерпретации результатов принимали участие фольклористы Института языка, литературы и истории Карельского научного Центра Российской Академии Наук (г. Петрозаводск) и сотрудники отдела фольклора музея-заповедника «Кижи».

Выводы

Полученные результаты говорят о том, что теоретико-графовые модели и заданные на них метрики позволяют искать зависимости между особенностями семантической структуры фольклорных песен и их основными характеристиками, что может помочь фольклористам при решении сложных задач атрибуции и сравнения фольклорных текстов. При этом для обеспечения большей надежности результатов необходимо увеличить выборку песен, привлекая материалы из архивов Петрозаводска, Санкт-Петербурга и Москвы.

Данное исследование может быть продолжено апробацией рассмотренных методов на примере коллекций других фольклорных жанров, в задачах реконструкции текстов, для решения вопросов жанровой дифференциации (например, духовных стихов и былин) и т. д.

Литература

[1] Зарубежные исследования по семиотике фольклора: Сб. ст. / Сост. Е. М. Мелетинский, С. Ю. Неклюдов. — М.: Наука, 1985.— 316 с.

[2] *Куперштох В. Л., Трофимов В. А.* Алгоритм анализа структуры матрицы связи // Автоматика и телемеханика. — 1975. — № 11. — С. 170–180.

[3] *Лысанов В. Д.* Досюльная свадьба, песни, игры и танцы в Заонежье Олонецкой губернии. — Петрозаводск: Северная скоропечатня Р. Г. Каца, 1916. — 119 с.

[4] *Марусенко М. А.* Атрибуция анонимных и псевдонимных литературных произведений методами теории распознавания образов. — Л: Издательство ЛГУ, 1990. — 164 с.

[5] *Москин Н. Д.* Теоретико-графовые модели структуры фольклорных текстов, алгоритмы поиска закономерностей и их программная реализация // Дисс. на соиск. уч. ст. к.т.н. — Петрозаводск, 2006. — 121 с.

[6] *Москин Н. Д.* Теоретико-графовые модели структуры фольклорных песен и методы их анализа // Круг идей: Междисциплинарные подходы в исторической информатике. Труды X конференции Ассоциации «История и компьютер». — Москва: МГУ, 2008. — С. 280–300.

[7] *Севбо И. П.* Графическое представление синтаксических структур и стилистическая диагностика. — Киев: Наукова Думка, 1981. — 192 с.

[8] *Скороходько Э. Ф.* Семантические сети и автоматическая обработка текста. — Киев: Наукова Думка, 1983. — 218 с.

[9] *Bunke H.* Graph matching: theoretical foundations, algorithms, and applications // Proc. Vision Interface. — Montreal, 2000. — P. 82–88.

[10] *Jiang X., Munger A., Bunke H.* On median graphs: properties, algorithms, and applications // IEEE Transactions on Pattern Analysis and Machine Intelligence. — 2001. — Vol. 23, № 10. — P. 1144–1151.

Распознавание скрытой периодичности в геномах модельных организмов*

Назипова Н. Н., Теплухина Е. И., Тюльбашева Г. Э., Чалей М. Б.
nnn@impb.psn.ru
Пущино, ИМПБ РАН

Создана вычислительная технология распознавания участков скрытой периодичности в геномных последовательностях, основанная на спектрально-статистическом подходе. Данные, полученные с помощью этой технологии, представлены в базе данных HETEROGENE, которая является достоверным и неизбыточным информационным ресурсом по скрытой периодичности в геномах модельных организмов, имеющих особое значение для генно-инженерных исследований. Первый выпуск базы содержит периодические последовательности из геномов модельных эукариот *Caenorhabditis elegans*, *Saccharomyces cerevisiae* и *Drosophila melanogaster*.

Геном представляется символьной последовательностью, составленной из четырёх видов букв, которые, на первый взгляд, разбросаны в случайном порядке. Однако некоторые участки геномов обладают выраженной периодичностью. Такие периодические участки называются тандемными (последовательно расположенными один за другим) повторами. Тандемный повтор характеризуется длиной (длина образца) и кратностью. Каждая копия образца (и сам образец) независимо подвергаются дополнительному мутированию (заменам, вставкам, удалениям отдельных элементов последовательности, новым тандемным дупликациям внутренних фрагментов и др.). Через определённое время каждая копия повтора сильно дивергирует, точный тандемный повтор становится размытым. Под скрытой периодичностью в нуклеотидных последовательностях понимают размытые тандемные повторы [1], в частности, — нечёткие тандемные повторы, не повреждённые вставками и делециями. Задача состоит в распознавании размытых тандемных повторов при неизвестных длине периода и кратности.

Введение

Большой интерес к распознаванию периодических структурных элементов генома обусловлен важной ролью, которую они играют в жизни. Одним из самых распространённых применений таких распознаваний является диагностика генетических болезней. Более 12-ти неврологических генетических болезней человека связаны с тандемными тринуклеотидными повторами. В нормальной популяции такие повторы разнообразны и относительно коротки. У больных людей они имеют кратность от 5 до 2000 копий в зависимости от локуса, вызывающего заболевание. Наиболее известная форма врожденной олигофрении (fragile-X mental retardation) вызывается появлением многочисленных (около 200) копий триплета CGG в 5′-нетранслируемой области гена FMR1. Болезнь Хантингтона (Huntington's disease) — болезнь мозга, которая вызывается наличием критического количества тандемных копий триплета CAG, кодирующего глутамин, в гене IT-15. Миотоническая дистрофия мышц (myotonic dystrophy) вызывается накоплением экстремального количества (не менее 50 копий) повторов того же самого триплета CAG в гене, который находится на 19-й хромосоме и кодирует мышечный белок протеинкиназу. Редкая форма атрофии (spinal and bulbar muscular atrophy) вызывается чрезмерным тандемным копированием триплета CAG в гене, кодирующем рецептор андрогена AR, у здоровых людей может быть до 36 тандемных копий этого фрагмента ДНК. Наследственная атаксия Фридрейха (Friedreich's ataxia) вызывает поражение нервной системы и сердца и вызывается чрезмерной тандемной дупликацией триплета GAA в гене, кодирующем белок фратаксин (frataxin).

Кроме того, с тандемными повторами связывают такие болезни, как рассеянный склероз [2], шизофрения [3], болезнь Альцгеймера [4], и рак [5]. Для всех этих случаев появление повторов в определённой части генома означает патологию.

Ещё одним применением тандемных повторов является ДНК-типирование (определение иммунологической принадлежности клетки, ткани или организма по результатам анализа состава антигенов) в криминалистике [6]. По районам периодичности можно реконструировать эволюционную историю генома, можно изучать дифференциацию между отдельными индивидуумами и географически или по времени изолированными популяциями. Тандемные повторы могут служить генетическими маркерами (участками ДНК с известной локализацией), используя которые, можно изучать эпидемии инфекционных болезней [7]. Данные полных геномов для тотальной локализации тандемных повторов появились сравнительно недавно, методы поиска периодичностей (тандемных повторов) в полных геномах стали активно развиваться и продолжают развитие в последнее время [8]. Надо отметить, однако, что научный и практиче-

*Работа выполнена при финансовой поддержке РФФИ, проекты №09-07-00455, №08-01-12030, №06-07-89274.

ский интерес к распознаванию скрытой периодичности в геномах организмов способствует развитию методов их поиска и созданию баз данных [9, 10, 11, 12, 13, 14], главным образом, для микро- и минисателлитных последовательностей (первые имеют характерную длину периода от 2 до 6 нуклеотидов, вторые — от 7 до 200 нуклеотидов). Кроме того, имеет место недостаточная достоверность методов поиска скрытой периодичности и большая избыточность представляемых в базах результатов.

При создании базы данных HETEROGENE [15] для поиска скрытой периодичности в геномах был применён спектрально-статистический подход [16, 17, 18], который, благодаря специально разработанному NEP-критерию, позволяет получать достоверные результаты в практических условиях недостаточного статистического материала при поиске микросателлитных, минисателлитных и сателлитных (с длинами периода более 200 нуклеотидов) участков. Общая методика выявления скрытой периодичности напоминала shotgun-стратегию секвенирования геномов [19], когда сначала секвенируют относительно короткие и перекрывающиеся фрагменты и затем производят их компьютерную сборку в более протяжённые участки.

В исходной программе поиска периодичности спектрально-статистический подход [17] реализован в пределах перекрывающихся окон сканирования разной длины. Сканирование каждого генома повторяется многократно с учётом различных уровней дивергенции выявляемых повторов, которые можно характеризовать значением величины $pl \in [0, 1]$, называемой уровнем сохранности буквы.

Для устранения высокой избыточности и коррекции получаемых первичных результатов был создан специальный комплекс программ. Комплекс позволяет там, где это обосновано, объединять найденные в геноме пересекающиеся или смежные участки скрытой периодичности и корректировать уровень сохранности и размер паттерна (длину периода L) окончательно выявляемого участка. Каждый, вновь выявленный путём слияния, участок вновь проходит проверку NEP-критерием. Таким образом, совместная работа программы поиска скрытой периодичности на основе спектрально-статистического подхода и программного комплекса последующей корректировки и фильтрации данных представляет технологию выявления неизбыточной, статистически значимой скрытой периодичности без ограничений на длину выявляемых районов периодичности.

Методы и алгоритмы

Спектрально-статистический подход к поиску скрытой периодичности [17] выделяет спектрально-статистические характеристики нуклеотидной последовательности, чувствительные к периодичности. Для каждого участка производится расчёт двух значений — pl (уровня сохранности паттерна) и H (спектрально-статистического параметра). Из совместного анализа этих характеристик предлагается оценка размера паттерна периодичности (L). Достоверность этой оценки проверяется затем с помощью специального критерия, названного NEP-критерием. Участки, удовлетворяющие NEP-критерию, считаются обладающими свойством скрытой периодичности.

Начальный этап поиска районов скрытой периодичности. Сканирование геномов осуществлялось с помощью перекрывающихся окон длиной от 30 до 3840 нуклеотидов. Каждый геном сканировался шестикратно, с учётом возможных различных уровней дивергенции повторов, которые можно характеризовать величиной pl уровня сохранности буквы ($pl \in [0, 1]$). Поиск скрытой периодичности проводился для значений $0{,}5 \leqslant pl \leqslant 1$. Границы каждого найденного участка варьировались до достижения наиболее высокого значения pl.

Последующая коррекция результатов. Полученные на первом этапе результаты проходят четыре последующих этапа коррекции и фильтрации. Для этого создано программное обеспечение, которое позволяет:

- удалять полные вхождения участков строго по кратности длин периодов L и близости pl;
- расширять в обе стороны уже найденные участки так, чтобы при неизменном pl сохранялось значение H;
- удалять подлежащие объединению короткие фрагменты с близкими значениями уровней сохранности и кратными длинами периодов;
- сливать смежные, пересекающиеся и отстоящие друг от друга на расстояние, не превышающее половины длины периода, участки с близкими значениями pl уровней сохранности и производить проверку новых длинных фрагментов на соответствие NEP-критерию.

По уровню сохранности степень близости участков, подлежащих слиянию, задаётся значением параметра ε (в текущей реализации на разных этапах использовались разные значения ε от 0,02 до 0,05). В случае, когда новый фрагмент, получающийся слиянием нескольких участков, имеет статистически значимую периодичность согласно NEP-критерию, в базу данных попадает именно он, а все входящие в него участки с близкими значениями уровня сохранности удаляются. При этом может оказаться, что объединённый участок после проверки NEP-критерием меняет свои количественные характеристики — длину периода и уровень сохранности. Поэтому каждый из четырёх последующих этапов слияния участков начинается с верифи-

кации слияний, сделанных на предыдущем этапе, а заканчивается новыми слияниями. Таким образом, проверяется целесообразность замены группы пересекающихся участков одним новым. В результате все участки, имеющие кратные длины периода и сравнимые (с точностью до ε) значения уровня сохранности, которые вошли в новый участок, удаляются из базы данных. Если внутри нового объединённого участка имеется участок с кратной длиной периода и существенно большим значением уровня сохранности, то такой участок (один, с самым высоким значением уровня сохранности) остаётся в базе.

Система администрирования базы данных реализована на основе СУБД MySQL. Разработано программное обеспечение управления базой через веб-интерфейс, которое позволяет пополнять и корректировать базу. Организована поисковая система по различным ключевым полям базы с возможностью сортировки данных любого из этих полей. Для каждого участка подключены модули для просмотра графиков спектров значений уровней сохранности и значений спектрально-статистического параметра, модуль визуализации участка периодичности, а также находящийся в свободном доступе модуль просмотра аннотированного фрагмента последовательности, внутри которого лежит участок периодичности, Sequence Viewer 2.1 (`http://www.ncbi.nlm.nih.gov/projects/sviewer`).

Результаты и обсуждение

Создана вычислительная технология выявления районов скрытой периодичности в сильно протяженных геномных последовательностях. Она позволяет получать неизбыточные достоверные данные о тандемных повторах любой длины — начиная с длины периода в 2–3 нуклеотида и до десятков тысяч нуклеотидов, практически без ограничения сверху. Эта технология уникальна благодаря своей универсальности. Она отличается от всех существующих программных разработок в этой области, которые всегда заточены на выявление какого-то конкретного вида тандемных повторов, тем, что одинаково легко распознает минисателлитные, микросателлитные и сателлитные повторы. Тем самым получается полная и не избыточная картина скрытой периодичности целого генома. По данным, полученным с помощью этой технологии, спроектирована и реализована база данных районов скрытой периодичности HETEROGENE [15].

Первая версия базы содержит результаты обработки геномов трёх эукариотических организмов — круглого червя *Caenorhabditis elegans*, пекарских дрожжей *Saccharomyces cerevisiae* и плодовой мушки *Drosophila melanogaster*. Это модельные организмы, взятые из разных биологических видов, биология которых довольно хорошо изучена. Полные геномы этих организмов стали известны за последнее десятилетие, что открыло новые перспективы для их систематических генноинженерных и биоинформационных исследований. Например, зная картину распределения тандемных повторов по отдельным хромосомам, можно делать выводы об эволюционной истории современной структуры хромосом как в рамках одного организма, так и для групп организмов различных видов.

Созданная база данных HETEROGENE может иметь как практическое применение в качестве ресурса генетических маркеров (в основном, микросателлитных последовательностей с длиной периода до 7 нуклеотидов), так и общее биологическое значение. Особенность созданного ресурса состоит в том, что он содержит неизбыточные и достоверные данные (прошедшие дополнительное тестирование специальными критериями) по последовательностям со скрытой периодичностью без ограничений на длину периода — здесь есть и микросателлитные последовательности, и участки с длиной периода в несколько тысяч нуклеотидных пар. Для каждого периодического участка указана наиболее вероятная длина периода, обоснованность выбора которой иллюстрируется приведёнными спектрами значений pl уровней сохранности и значений спектрально-статистического параметра H проявления неоднородностей [17]. Это позволяет для периодических последовательностей каждого генома наиболее точно оценить спектр длин периодов, их кратностей, уровней сохранности и т. д. Также можно оценить и общую долю районов периодичности в геноме каждого организма. Анализ такой информации совместно с аннотированием геномных последовательностей способствует пониманию смысла и созданию моделей возникновения скрытой периодичности в геномах различных биологических организмов.

Литература

[1] *Benson G.* Tandem repeats finder: a program to analyze DNA sequences // Nucl. Acids Res. — 1999. — V. 27, № 2. — Pp. 573–580.

[2] *Guerini F. R. et al.* Interleukin-6 gene alleles affect the risk of Alzheimer's disease and levels of the cytokine in blood and brain // Neurobiol. Aging. — 2003. — V. 61. — Pp. 520–526.

[3] *Licastro F., et al.* Myelin basis protein gene is associated with ms in DR4- and DR5-positive Italians and Russians // Neurology. — 2003. — V. 24. — Pp. 921–926.

[4] *Brzustowicz L. M., et al.* Location of a major susceptibility locus for familial schizophrenia on chromosome 1q21-q22 // Science. — 2000. — V. 288. — Pp. 678–682.

[5] *Sidransky D.* Nucleic acid-based methods for the detection of cancer // Science. — 1997. — V. 278. — Pp. 1054–1058.

[6] *Butler J.* Forensic DNA Typing: Biology and Technology Behind STR Markers. — London: Academic Press, 2003.

[7] *Cummings C. A. and Relman D. A.* Microbial Forensics-cross-examining pathogens // Science. — 2002. — V. 296. — Pp. 1976–1979.

[8] *Krishnan A. and Tang F.* Exhaustive whole-genome tandem repeats search // Bioinformatics. — 2004. — V. 20. — Pp. 2702–2710.

[9] *Mudunuri S. B. and Nagarajaram H. A.* IMEx: Imperfect microsatellite extractor // Bioinformatics.— 2007.— V. 23. — Pp. 1181–1187.

[10] *Bikandi J., et al.* In silico analysis of complete bacterial genomes: PCR, AFLP-PCR, and endonuclease restriction // Bioinformatics. — 2004. — V. 20. — Pp. 798–799.
`http://insilico.ehu.es/microsatellites/`

[11] *Denoeud F., Vergnaud G.* Identification of polymorphic tandem repeats by direct comparison of genome sequence from different bacterial strains: a web-based resource // BMC Bioinformatics. — 2004. — V. 5. — Pp. 4. `http://minisatellites.u-psud.fr/`

[12] *Sreenu V. B., et al.* MICdb — Database of prokaryotic microsatellites // Nucleic Acids Research, 2003. — V. 31. — Pp. 106–108.
`http://210.212.212.7/MIC`

[13] *Shelenkov A. A., et al.* MMsat — a database of potential micro- and minisatellites // Gene. — 2008. — V. 409. — Pp. 53–60.
`http://victoria.biengi.ac.ru/mmsat`

[14] *Boeva V., et al.* Short fuzzy tandem repeats in genomic sequences, identification, and possible role in regulation of gene expression // Bioinformatics. — 2006. — V. 22. — Pp. 676–684.

[15] HETEROGENE database — 2008.
`http://www.jcbi.ru/lp_baze`

[16] *Чалей М. Б., Назипова Н. Н., Кутыркин В. А.* Совместное использование различных критериев проверки однородности для выявления скрытой периодичности в биологических последовательностях // Математическая биология и биоинформатика (электр. журнал). — 2007. — Т. 2. — С. 20–35.
`www.matbio.org/downloads/Chaley2007(2_20).pdf`

[17] *Chaley M., Kutyrkin V.* Model of perfect tandem repeat with random pattern and empirical homogeneity testing poly-criteria for latent periodicity revelation in biological sequences // Mathematical Biosciences. — 2008. — V. 211. — Pp. 186–204.

[18] *Chaley M., Nazipova N., Kutyrkin V.* Statistical Methods for Detecting Latent Periodicity Patterns in Biological Sequences: The Case of Small-Size Samples // Pattern Recognition and Image Analysis. — 2009. — V. 19. — Pp. 358–367.

[19] *Venter J. C., et al.* The sequence of the human genome // Science. — 2001. — V. 291. — Pp. 1304–351.

Двоичный метод группового учета аргументов в задаче «структура–активность»*

Носеевич Ф. М., Девятьяров Д. А., Кумсков М. И., Апрышко Г. Н., Пермяков Е. А.
kumskov@mail.ru
Москва, МГУ им. М. В. Ломоносова, мехмат, кафедра вычислительной математики,
Москва, Российский онкологический научный центр им. Н. Н. Блохина,
Москва, Институт органической химии им. Н. Д. Зелинского РАН

В работе предложен новый эволюционный метод анализа матрицы «молекула–дескриптор» для решения задачи «структура–активность». Метод основан на эволюционном построении семейства ДНФ/КНФ и схож с методом группового учета аргументов. Метод анализирует бинарные данные с бинарным целевым вектором и позволяет обрабатывать матрицы с количеством дескрипторов, значительно превышающим число молекул. Вычислительные эксперименты показали, что при применении к выборке гликозидов предложенный метод строит прогнозирующие модели с меньшим числом выбросов и более высоким качеством прогноза, чем классический аналог.

Поиск взаимозависимостей между структурами химических соединений и их свойствами посредством построения математических моделей в задаче «структура–активность» разбивается на несколько этапов:

1) выбор описания пространственной структуры молекулы;
2) анализ числовой матрицы для выявления корреляции, зависимости между столбцами матрицы и столбцом свойства, результатом которого служит построение прогнозирующей функции;
3) верификация, проверка качества прогноза и выявление выбросов.

Для каждого из этих этапов существуют десятки вариантов реализации, в зависимости от постановки задачи и исходных данных. В данной работе идет речь только о втором этапе — анализе матрицы «молекула–дескриптор» (МД-матрицы). Детальная постановка задачи приведена в [5, в данном сборнике].

Целью работы является поиск зависимостей на двоичных векторах при двоичном целевом векторе с использованием эволюционного построения семейства ДНФ/КНФ на большом исходном пространстве признаков. Для решения данной задачи предлагается использовать метод по структуре похожий на метод группового учета аргументов (МГУА) [3]. По этой причине назовем метод *двоичный МГУА*.

Ниже приведено описание каждого этапа предложенного метода.

Иерархический кластерный анализ

Перед запуском классифицирующего метода необходимо содержательно разделить объекты-молекулы на кластеры, с тем чтобы запустить аналог МГУА отдельно на каждом из них с целью улучшить результат прогноза.

Структура данных такова, что при стандартных параметрах кластерного анализа почти все молекулы группируются в один кластер. Поэтому был проведен поиск наиболее оптимальных параметров, то есть параметров, позволяющих формировать несколько больших кластеров. Оптимальный тип метрики (1 − косинус угла между двумя векторами) и метод определения расстояния между кластерами (невзвешенное среднее расстояние) были найдены в результате ряда экспериментов и рекомендуются для использования. Процесс поиска оптимальных значений параметров для кластерного анализа может быть автоматизирован, так чтобы под каждую конкретную выборку использовать наилучшую кластеризацию.

Формирование двоичной матрицы

Двоичный МГУА предназначен для обработки бинарных матриц, поэтому МД-матрицу действительных чисел необходимо предварительно преобразовать к бинарному виду. Каждый столбец (соответствующий дескриптору) заменяется на несколько бинарных столбцов следующим образом: множество всех значений столбца разделяется на несколько крупных кластеров и формируется бинарный столбец для каждого кластера, отражающий, входит ли значение в соответствующий кластер (значение 1 в столбце) или нет (значение 0). Пример подобного преобразования приведен в таблице 1.

Автоматическое деление значений столбца на некоторое наперед заданное количество кластеров приводит к неоднородности разбиения. Предложены следующие модификации, которые позволяют сгладить эту неоднородность:

Модификация А. Каждый столбец-дескриптор преобразуется ровно в два бинарных столбца. Идет поиск такого разбиения на k кластеров, чтобы сумма числа элементов в двух самых больших кластерах была больше (а разность между ними — меньше) некоторого процента от общего количества элементов. Элементы остальных кластеров

*Работа выполнена при финансовой поддержке РФФИ, проект № 07-07-00282.

Таблица 1. Преобразование значений дескрипторов в бинарные векторы.

Значение дескриптора	Номер кластера	Столбец 1	Столбец 2
9	1	1	0
0	1	1	0
63	2	0	1
5	1	1	0

Таблица 2. Преобразование значений дескрипторов в бинарные векторы: модификация Б.

No.	1	2	3	4	5	6	7	8	9	10
1	*58*	*34*	*29*	2	1	1	1	1	1	1
2	*87*	16	9	6	5	3	1	1	1	1
3	*73*	*32*	6	5	4	4	2	1	1	1
4	*103*	11	6	2	2	1	1	1	1	1

No.	Число кластеров	Число элементов в кластерах	Кол-во выбросов
1	3	118	12 (9%)
2	1	88	42 (33%)
3	2	106	22 (17%)
4	1	104	26 (20%)

($< 20\%$) распределяются по наименьшему евклидову расстоянию до центров двух больших кластеров. Преимущества такого подхода состоят в простоте реализации и достаточной информативности, так как отражены особенности абсолютного большинства дескрипторов. Вместе с тем очевидны и его недостатки: для ряда случаев приходится жертвовать достаточно целостным кластером.

Модификация Б. Число бинарных столбцов, соответствующих столбцу-дескриптору, равно количеству кластеров в его структуре плюс отдельный столбец, соответствующий выбросам. При проведении кластерного анализа значимым считаем кластер, мощность которого превышает определенный процент от общего числа элементов. Формируем бинарный столбец для каждого значимого кластера, а также отдельно формируем группу выбросов (элементов, не вошедших в значимые кластеры) и строим отдельный бинарный столбец, отражающий принадлежность элемента к выбросам. С одной стороны, нет потери информации, с другой — наименее информативная часть элементов рассматривается отдельно.

В таблице 2 приведены типичные ситуации, анализ которых и приводит к необходимости использования этого подхода. Каждая строка соответствует одному из 4 дескрипторов, столбцы соответствуют кластерам. В верхней половине таблицы показаны мощности кластеров для разных дескрипторов. Курсивом выделены кластеры, которые считаются значимыми. Для каждого из них формируется отдельный бинарный дескриптор. Остальные кластеры объединены в группу выбросов, для которой также строится бинарный дескриптор. Таким образом, в таблице показано, как 4 дескриптора были преобразованы в 11 двоичных столбцов. При применении модификации А двоичных столбцов было бы ровно 8.

В наших экспериментах была реализована модификация А. Оптимальные параметры для метрики (евклидова) и метода определения расстояния между кластерами (метод кратчайшего расстояния) были найдены в результате серии вычислительных экспериментов. Следует отметить, что были выбраны разные параметры кластерного анализа при разбиении выборки на кластеры и при преобразовании действительнозначных дескрипторов в двоичные.

Двоичный МГУА на кластере

МГУА является эволюционным алгоритмом. Его принцип работы вместе с предшествующей ему процедурой иерархического кластерного анализа подробно изложен в [4].

Идея использования функций алгебры логики для поиска зависимостей не нова и была предложена, в частности, в работе [1]. Там в качестве одного из продолжений рассматривалось преобразование исходной МД-матрицы к матрице логического формата. Было сделано предположение, что при дискретизации интервала значений дескрипторов и использовании в качестве классификатора модификации МГУА, основанной на дизъюнктивной нормальной форме, результаты могут быть улучшены.

Сначала рассмотрим модели в виде конъюнкций и дизъюнкций набора дескрипторов. Итак, в нашем распоряжении имеется двоичная матрица. Производим полный перебор конъюнкций всевозможных пар дескрипторов (аналогичные алгоритмы будут использованы и для дизъюнкций, поэтому все рассуждения про конъюнкцию можно дословно перенести на дизъюнкцию). Для каждой конъюнкции вычисляем, насколько хорошо она предсказывает столбец активности. Ввиду удобного формата данных, это процедура проста: подсчитывается число совпадений нулей и единиц. При этом создаем структуру, где хранятся столбцы, дающие лучший прогноз. Также туда добавляется информация о том, каким образом они были получены (то есть номера столбцов матрицы, давшие такой результат), и с помощью какой операции был получен результат (конъюнкции или дизъюнкции).

Затем рассматриваем всевозможные конъюнкции моделей, полученных на предыдущем шаге (как конъюнкции пар бинарных столбцов), и одиночных бинарных столбцов. Таким образом, получаем конъюнкции троек бинарных столбцов, но не все возможные, а только построенные на основе лучших моделей на парах столбцов. Из конъ-

Рис. 1. Генетический алгоритм построения конъюнкций / дизъюнкций.

юнкций троек отбираем модели, дающие лучший прогноз для целевого вектора. Далее, аналогичным образом формируем четверки, пятерки и т. д., как показано на рис. 1.

В процессе качество прогноза улучшается до определенного числа задействованных дескрипторов, когда дальнейшее добавление дескрипторов не приводит к росту качества прогноза. Эксперименты показали, что довольно быстро качество выходит на постоянные значения, и дальнейшее добавление дескрипторов не имеет смысла.

Уже при использовании моделей вышеописанного вида, основанных только на конъюнкциях (или только на дизъюнкциях), достигается весьма высокое качество прогноза. Однако пока что речь не шла об аналоге ДНФ или КНФ: для этого нужно комбинировать логические операции. Соответствующие модификации алгоритма были проведены: на основе исходной матрицы и матрицы лучших конъюнкций были сформированы дизъюнкции, подобно алгоритму на рис. 1.

В результате можно получить целый набор структур, содержащих разной точности прогнозы целевого вектора активности. Мы выбираем модель с лучшим качеством прогноза. При этом на финальном этапе анализа устойчивости полученной модели необходимо рассматривать все многообразие моделей в совокупности. Вполне вероятно, что модель, дающая чуть худший прогноз, будет гораздо более устойчивой.

Вычисление параметров прогнозирующей модели и прогноз активности новых соединений

Отобрав модели с лучшим качеством прогноза, можно выделить дескрипторы, на основе которых были построены данные модели. Если ранее были сформированы МД-матрицы на различных параметрах описания, то качество прогноза позволяет выделить оптимальные параметры описания, при которых получены «удачные» матрицы. Тем самым, происходит выработка рекомендаций по параметрам детализации.

Получая на вход новое соединение, относим его тем или иным способом к одному из выделенных ранее кластеров. В противном случае оно попадает в выбросы и происходит отказ от прогноза. Прогноз строится с помощью конъюнкции и/или дизъюнкции по тем дескрипторам, что были выбраны.

Глобальная проблема подобного рода анализа — экстенсивный рост числа дескрипторов при усложнении описании молекул. То есть, если мы хотим учитывать дополнительные факторы для соединения (например, электростатический заряд на молекулярной поверхности), то необходимо иметь в виду, что это приведет к значительному увеличению пространству дескрипторов — МД-матрица станет еще более широкой. По этой причине большую значимость приобретает предварительный анализ матрицы.

Двоичный МГУА может быть использован не только для прогноза, но и для подобного предварительного анализа. Например, предлагается рассмотреть объединение дескрипторов по каждому столбцу. Полученное множество будет в некотором роде информационной выжимкой из матрицы. Число дескрипторов для анализа будет снижено в 10–20 раз, и к ним уже, в свою очередь, может быть применен отдельный анализ при использовании другого прогнозатора. Объединив дескрипторы, мы потеряли связь со столбцами-дескрипторами двоичного МГУА. Однако, теперь алгоритм двоичного МГУА может быть запущен по этому небольшому набору дескрипторов. Это рациональный путь для выявления наиболее информативных дескрипторов для дальнейшей работы.

Результаты расчетов

В таблице 3 приведены результаты вычислительных экспериментов для 3 матриц, построенных для выборки гликозидов при разных параметрах описания. В работах [2] и [6] можно найти описание построения молекулярной поверхности и подробное описание цепочки вычислений, результатом которых становится МД-матрица. В частности, были применены структурные трехмерные дескрипторы — пары и тройки особых точек, определенных на триангулированной молекулярной поверхности химического соединения [7, 8]

В первой части таблицы показаны результаты для обычного МГУА, во второй — двоичного МГУА, описанного выше. При запуске обычного МГУА были использованы неоптимальные параметры кластерного анализа, в результате чего число получаемых выбросов было, во-первых, вели-

Таблица 3. Сравнение результатов двоичного МГУА на динамически формируемых кластерах молекулярных графов с результатами классического МГУА: N — количество элементов в кластере или группе выбросов, R^2 — показатель качества прогноза на скользящем контроле.

Классический МГУА					
	Кластер 1		Кластер 2		Выбросы
Матрица	N	R^2	N	R^2	N
1	102	72%	9	67%	18
2	70	76%	12	92%	47
3	69	78%	21	60%	39
Двоичный МГУА					
	Кластер 1		Кластер 2		Выбросы
Матрица	N	R^2	N	R^2	N
1	67	80%	57	81%	5
2	67	87%	57	84%	5
3	78	90%	46	85%	5

ко, во-вторых, крайне непостоянно. В случае, если выбросы составляют около 50% выборки (как в матрице 2), сложно говорить о том, что построенной модели стоит доверять. Это означает, что число данных, которые использовались для прогноза, становится очень незначительным. Новая молекула в такой ситуации, скорее всего, получит отказ от прогноза. При использовании двоичного МГУА, были получены более однородные кластеры и меньшее количество выбросов.

В целом, из таблицы видно, что двоичный МГУА выдает более высокое качество прогноза.

Выводы

В работе предложен новый эволюционный метод поиска логических зависимостей для обработки МД-матриц с числом дескрипторов, сильно превышающим число молекул, для чего реализован метод формирования логических переменных на основе дескриптора-столбца. Это привело к тому, что меньшее число дескрипторов относится к числу неинформативных и исключается из прогноза, либо же просто неверно интерпретируется. Результаты вычислительных экспериментов позволили получить модели, сопоставимые с классическими, которые отличаются меньшим количеством выбросов и более высоким качеством прогноза.

Могут быть выделены следующие дальнейшие направления работы.

1. Накопление статистики по большему числу различных выборок, которое позволит предоставить более полные рекомендации математикам, участвующим в поиске зависимости между структурой и свойством.
2. Поиск химической интерпретации полученных результатов, то есть перехода от формального математического описания, используемого для построения модели, к физико-химической составляющей задачи.

Литература

[1] Алгоритмы и программы восстановления зависимостей / Под редакцией В. Н. Вапника. — Москва: Наука, Главная редакция физико-математической литературы, 1984.— 816 с.

[2] *Григорьева С. С., Чичуа В. Т., Деветьяров Д. А., Кумсков М. И.* Выбор оптимального описания структуры молекулы в задаче структура-свойство для заданной биологической активности // Вестник Московского университета. Серия 2. Химия. — 2007. — Т. 48, No. 5. — С. 305–307.

[3] *Ивахненко А. Г., Зайченко Ю. П., Дмитриев В. Д.* Принятие решений на основе самоорганизации — Москва: Сов. Радио, 1976. — 220 с.

[4] *Кумсков М. И., Митюшев Д. Ф.* Применение метода группового учета аргументов для построения коллективных оценок свойств органических соединений на основе индуктивного перебора их «структурных спектров» // Проблемы управления и информатики. — 1996. — No. 4. — С. 127–149.

[5] *Прохоров Е. И., Перевозников А. В., Воропаев И. Д., Кумсков М. И., Пономарёва Л. А.* Поиск представления молекул и методы прогнозирования активности в задаче «структура–свойство» // Всероссийская конференция ММРО-14. — М.: МАКС Пресс, 2009. — С. 589–591 (в настоящем сборнике).

[6] *Kumskov M. I., Mityushev D. F.* Group Method of Data Handling (GMDH) as Applied to Collective Property Estimation of Organic Compounds by an Inductive Search of Their Structural Spectra // Pattern Recognition and Image Analysis. — 1996. — Vol. 6, No. 3. — Pp. 497–509.

[7] *Svitanko I. V., Devetyarov D. A., Tcheboukov D. E., Dolmat M. S., Zakharov A. M., Grigoryeva S. S., Chichua V. T., Ponomareva L. A., Kumskov M. I.* QSAR Modeling on the Basis of 3D Descriptors Representing the Electrostatic Molecular Surface (Ambergris Fragrances) // Mendeleev Communications. — 2007. — Vol. 17, No 2. — Pp. 90–91.

[8] *Svitanko I. V., Kumskov M. I., Tcheboukov D. E., Dolmat M. S., Zakharov A. M., Ponomareva L. A., Grigoryeva S. S., Chichua V. T.* QSAR Modeling on the Base of Electrostatic Molecular Surface (Amber Fragrances) // 16th European Symposium on Quantum Structure-Activity Relationships and Molecular Modelling, Italy: EuroQSAR. — 2007.

Топология ДНК вблизи бактериальных промоторов*

Панюков В. В., Озолинь О. Н.
panjukov@impb.psn.ru
Пущино, Институт математических проблем биологии РАН, Институт биофизики клетки РАН

В ранних работах в качестве трёхмерного образа протяжённого фрагмента двойной спирали ДНК использовали наименьший по объёму прямой цилиндр, содержащий все атомы биополимера. Однако в дальнейшем было обнаружено, что эта модель не учитывает многих функционально значимых особенностей пространственного строения природных ДНК. В данной работе в качестве структурной модели фрагмента ДНК предлагается изоцилиндр (геометрическое тело, полученное посредством изгибной деформации прямого цилиндра) и метод, позволяющий оценивать степень непрямолинейности модельных изоцилиндров. Обнаруженные особенности трехмерной структуры ДНК могут служить дискриминирующим критерием в алгоритмах поиска промоторов. Использование этого метода для промоторов Escherichia coli позволило обнаружить участок с предпочтительной конфигурацией промоторной ДНК.

Метаболизм любого организма определяется набором экспрессируемых в его клетках генов, зависящей от внешних условий частотой транскрипции каждого из них и эффективностью синтеза белка с матричных РНК. Принято считать, что основная регуляция генной экспрессии реализуется во время синтеза РНК, который осуществляют эволюционно консервативные белки — ДНК-зависимые РНК-полимеразы. Эти ферменты способны мигрировать вдоль молекулы ДНК, обнаруживать в ней сигнальные последовательности оснований (промоторы) и взаимодействовать с ними с образованием транскрипционного комплекса. Несмотря на то, что механизм полимераза-промоторного взаимодействия изучается уже много лет, до сих пор нет удовлетворительной модели, позволяющей предсказать его эффективность на конкретном участке ДНК.

Не вызывает ни малейшего сомнения, что основными сигнальными элементами промоторов являются консервативные гексануклеотиды, образующие специфические контакты с определенными структурными модулями в молекуле РНК-полимеразы. Но из-за высокой степени допустимых вариаций в последовательностях этих элементов, модели промоторов, построенные на их основе, обладают очень низкой специфичностью, то есть обнаруживают огромное число фальшивых сигналов транскрипции в непромоторных участках ДНК.

Начиная с 1982 года [1], когда было обнаружено, что многие природные молекулы ДНК имеют анизотропные изгибы, появляются работы, которые свидетельствуют в пользу того, в механизме полимераза-промоторного взаимодействия не последнюю роль играют топологические особенности ДНК [2], а также ее способность подвергаться адаптивным конформационным изменениям. Это дало толчок к разработке алгоритмов поиска промоторов ДНК, которые принимают во внимание пространственную структуру ДНК вблизи мест инициации транскрипции [3, 4, 5].

Предсказательная сила таких алгоритмов оказалась выше, чем у компьютерных программ, базирующихся на особенностях первичной структуры ДНК. Их дополнительным преимуществом является возможность предсказания регуляторных участков в геномах с ограниченной информацией о консервативных элементах промоторов.

Поэтому дальнейшее совершенствование методов пространственной топологической характеристики промоторов представляется исключительно целесообразным и перспективным.

Постановка задачи

Информация о функциональных и структурных свойствах промоторов необходима для полноценной реконструкции регуляторных сетей, определяющих метаболизм биологических объектов разного уровня организации. Самой используемой моделью в настоящее время является Escherichia coli (E.coli). Это связано с наличием богатого экспериментального материала, накопленного по этой бактерии. В частности, известны координаты более чем 1000 мест инициации транскрипции в её геноме. Последнее обстоятельство обусловило наш выбор объекта для моделирования особенностей пространственной структуры промоторов.

В отличие от ранее предложенных моделей, учитывающих анизотропные изгибы промоторной области, основной задачей данной работы было разработать подход, позволяющий оценить степень отклонения изоцилиндра двойной спирали промоторной ДНК от прямого модельного цилиндра. Для этого были использованы трехмерные модели 171 фрагмента геномной ДНК E.coli, содержащих экспериментально картированные промоторы.

Пространственные координаты всех атомов фрагмента вычисляли с помощью Интернет ресурса DNA Tools [7]. Размер моделируемого фрагмента для каждого промотора составляет 200 пар оснований и охватывает область ДНК от -150 до $+50$ относительно стартовой точки транскрипции.

*Работа выполнена при финансовой поддержке РФФИ, проекты №09-07-00455 и №07-04-01066.

Для каждого фрагмента, используя координаты его атомов, строили модельный изоцилиндр. Полученные изоцилиндры фрагментов сравнивали для выявления участков предпочтительной деформации. Предпочтительная деформации может отражать биологические функциональные особенности промоторов. Неформально, предпочтительная деформация обнаруживается «наложением» изоцилиндров с последующим их перемещением относительно друг друга для совпадения «похожей» локальной деформации, присущей всем изучаемым фрагментам. Так обнаруженная деформация в множестве фрагментов называется согласованной деформацией.

Дадим формальное определение согласованности деформации.

Пусть I — некоторое конечное множество изоцилиндров в трёхмерном пространстве и $d(v)$ — некоторая функция координат, $v = (x, y, z)$, которую мы будем называть деформацией. Ограничение d на изоцилиндре $g \in I$ даёт деформацию изоцилиндра. Если $v \in g$, тогда $d(v)$ есть деформация изоцилиндра g в точке v. Обозначим $G = g_1 \times \ldots \times g_n$, где n — число промоторов. Для каждого $V = (v_1, \ldots, v_n) \in G$ имеем размах $R(V)$ множества $\{d(v_1), \ldots, d(v_n)\}$. Согласованность деформации есть число $\min\{R(V) \colon V \in G\}$.

Требуется оценить согласованность деформаций заданного множества изоцилиндров.

Метод

Напомним, что фрагмент — это часть двойной спирали ДНК, представленная совокупностью атомов, заданных своими координатами. Ему соответствует некоторый изоцилиндр. Поэтому далее термин «фрагмент» будет означать модельный изоцилиндр.

Поставленную выше трёхмерную задачу оценки согласовнной деформации фрагментов будем решать методом сведения задачи к одномерной форме. Метод основан на простом наблюдении, которое заключается в следующем.

Пусть имеется фрагмент молекулы ДНК, координаты атомов которого заданы в трёхмерной системе координат. Можно считать, что в изогнутое состояние фрагмент перешёл из прямого состояния в результате непрерывной деформации фрагмента. Выберем какую-либо образующую боковой поверхности прямого фрагмента, которая является отрезком прямой. Будем следить за изгибом образующей, чтобы оценить деформацию фрагмента. Тем самым исходная задача сводится к одномерной.

Итак, фрагменту молекулы мы сопоставляем линейный объект и оцениваем деформацию этого линейного объекта.

Рассмотрим недеформированный фрагмент. Элемент фрагмента — это пара атомов фосфора, которые лежат на поверхности цилиндра. Соединим фосфоры отрезком прямой, получим «гантель». Центр тяжести гантели находится на оси цилиндра. Ось следующей «гантели» фрагмента повернута относительно предыдущей приблизительно на 36 градусов. Центры тяжести соседних «гантелей» лежат приблизительно на одинаковом расстоянии (34 ангстрема) друг от друга, при этом все центры лежат на оси цилиндра. Учитывая этот формализм, будем говорить о совокупности центров тяжести «гантелей» как о цепочке фосфоров.

Будем оценивать деформацию фрагмента ДНК, анализируя расположение цепочки фосфоров в пространстве. Вопрос заключался в следующем: имеют ли промотор-содержащие фрагменты ДНК длиной 200 нуклеотидных пар какую-то предпочтительную деформацию?

Исследование проводилось в среде программного продукта, специально разработанного для этой цели. Топологические деформации оценивались для сегментов разной длины, находящихся в разных участках промотор-содержащего фрагмента.

Топология сегментов характеризовалась следующими параметрами.

- *Реальная длина* — длина цепочки фосфоров.
- *Спрямлённая длина* — длина отрезка прямой, соединяющей концы цепочки.
- *Максимальное отклонение* — максимальное расстояние фосфора от спрямляющего отрезка прямой.
- *Ведущий треугольник* — треугольник, образованный концами цепочки и фосфором, который даёт максимальное отклонение от спрямляющего отрезка прямой.
- *Максимальный выступ* — максимальное расстояние фосфора от плоскости ведущего треугольника.
- *Центральный угол* — угол при вершине ведущего треугольника.

Распознавание согласованной деформации. Для выбранного топологического параметра t и длины фрагмента L мы имеем n функций значения параметра t, зависящего от позиции сегмента внутри промотор-содержащих фрагментов ДНК, где n — число промоторов. Каждая функция определена в области своего промотора. Совместив позиции промоторов, мы приведём функции к единой системе координат. Теперь в каждой целочисленной позиции интервала $[-150, 50 - L]$ мы имеем n значений параметра t.

Зафиксируем позицию P. Позволим графикам рассматриваемых функций смещаться в пределах интервала $[P, P + \Delta]$, где Δ — параметр задачи.

Путем перебора вариантов целочисленных смещений мы добьёмся максимального выравнивания значений функций в позиции P. Трудоёмкость ис-

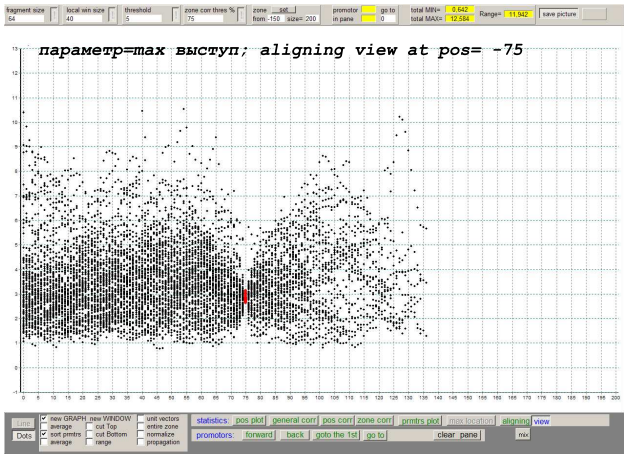


Рис. 1. Результат выравнивания максимальных выступов для фрагментов длиной 64 основания.

пользуемого нами алгоритма выравнивания оценивается как $nL\ln(nL)$.

На рис. 1 показан результат выравнивания максимального выступа для фрагментов ДНК длиной 64 пар оснований при $\Delta = 40$. Близкие значения данного топологического параметра, сосредоточенные около -75-й позиции промоторов, говорят о предпочтительной конфигурации промоторной ДНК в этом участке.

Выводы

Степень непрямолинейности промоторосодержащих фрагментов ДНК, определяемая максимальным выступом, варьировала в диапазоне 10 ангстрем и проявляла зависимость от расположения анализируемого сегмента относительно стартовой точки транскрипции. В позициях -74 оказалось возможным выравнивание указанного топологического параметра с минимальной вариацией 0,5 ангстрем, рис. 1.

Малая вариация значений в полученном выравнивании свидетельствует о наличие в промоторах участка с предпочтительной конфигурацией двойной спирали и допускает возможность использования степени непрямолинейности фрагментов ДНК в качестве критерия, позволяющего выбрать способные к инициации синтеза РНК промоторы среди непродуктивных мест связывания РНК-полимеразы, которые могут выполнять в геноме функции, не связанные с синтезом РНК.

Литература

[1] *Marini J. C., Levene S. D., Crothers D. M., Englund P. T.* Bent helical structure in kinetoplast DNA // Proc. Natl. Acad. Sci. USA. — 1982. — Vol. 79. — Pp. 7664–7667.

[2] *Ohyama T.* Intrinsic DNA bends: an organizer of local chromatin structure for transcription. // Bioessays. — 2001. — Vol. 23. — Pp. 708–715.

[3] *Kanhere A., Bansal M.* A novel method for prokaryotic promoter prediction based on DNA stability. // BMC Bioinformatics. — 2005. — Vol. 6. — Pp. 1471–2105.

[4] *Wang H., Benham C. J.* Promoter prediction and annotation of microbial genomes based on DNA sequence and structural responses to superhelical stress. // BMC Bioinformatics. — 2006. — Vol. 7. — Pp. 248–263.

[5] *Ozoline O. N., Deev A. A.* Predicting antisense RNAs in the genomes of Escherichia coli and Salmonella typhimurium using promoter-search algorithm PlatProm. // J. Bioinf. Comput. Biol.. — 2006. — Vol. 4. — Pp. 443–454.

[6] *Vlahovicek K., Kajan L. O., Pongor S. O.* DNA analysis servers: plot.it, bend.it, model.it and IS. // Nucleic Acids Res. — 2003. — Vol. 31, No. 13. — Pp. 3686–3687.

[7] `http://hydra.icgeb.trieste.it/dna/bend_it.html`

Построение трехмерной модели мозга мыши по набору двумерных изображений из Алленовского Атласа*

Осокин А. А., Ветров Д. П., Кропотов Д. А.
osokin.anton@gmail.com, vetrovd@yandex.ru
Москва, МГУ ВМК

В рамках данной работы был разработан полностью автоматический метод трехмерной реконструкции мозга мыши по набору двумерных гистологических коронарных срезов. Для построения трехмерной модели используются нелинейные деформации соседних срезов друг в друга и дальнейший морфинг. Построенная трехмерная модель позволяет получать виртуальное сечение произвольной плоскостью. В качестве исходных данных для построения трехмерной модели был взят Алленовский коронарный атлас мышиного мозга.

В настоящее время исследования мозга занимают важное место в медицине и биологии [1, 2, 3]. Исследуется как анатомическая структура мозга, так и физиологические процессы, происходящие внутри мозга. Одним из методов исследования является замораживание мозга в жидком азоте и последующая нарезка. Такой метод предоставляет в распоряжение исследователя набор фотографий двумерных срезов.

Поскольку данная процедура очень трудоемка, невозможно сделать «полный» комплект срезов. В процессе экспериментов срезы получают с достаточно большим интервалом. Такого набора двумерных срезов недостаточно для полноценного отображения структуры головного мозга. Требуется построение трехмерной модели, которая бы отражала внутреннюю структуру мозга. В рамках данного исследования и был разработан метод построения трехмерной модели мозга.

Исходные данные

В качестве исходных данных для трехмерной модели был взят коронарный Алленовский Атлас мозга [4], представляющий из себя 132 среза мозга. Пример среза мозга изображен на рис. 1.

Эти изображения содержат информацию как о гистологической структуре среза (левая часть изображения), так и об анатомических структурах, выделенных на срезах экспертами (правая часть изображения).

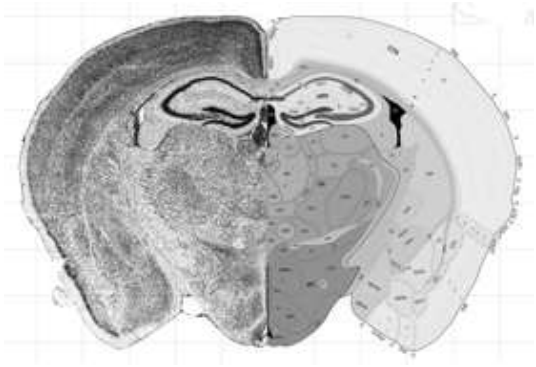


Рис. 1. Изображение из Алленовского атласа мозга.

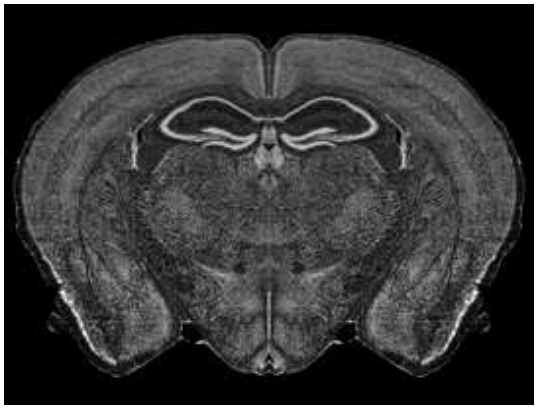


Рис. 2. Предобработанное изображение Алленовского атласа.

Предобработка набора срезов

Для построения трехмерной модели мышиного мозга требуется провести предобработку изображений Алленовского атласа.

Первым этапом предобработки является предобработка каждого изображения по отдельности:

1. Получение полного гистологического изображения путем отражения левой половины атласных изображений (рис. 1) относительно центральной оси. После этого для удобства реализации проводится переход от позитивов к негативам.
2. Производится существенное уменьшение разрешения изображений (разрешение оригинальных атласных изображений — 5690×4418; новое разрешение — 270×204). Во-первых, это позволяет подавить шум и слишком мелкие элементы мозга (на оригинальных изображениях видна клеточная структура мозга). Во-вторых, уменьшение разрешения существенно уменьшает время работы алгоритмов дальнейшей предобработки и построения трехмерной модели.
3. Выделение мозга на изображении и шумоподавление (под шумом понимаются пятна на изобра-

*Работа выполнена при финансовой поддержке РФФИ, проекты №08-01-00405, №08-01-90016, №08-01-90427, №09-04-12215-офи-м.

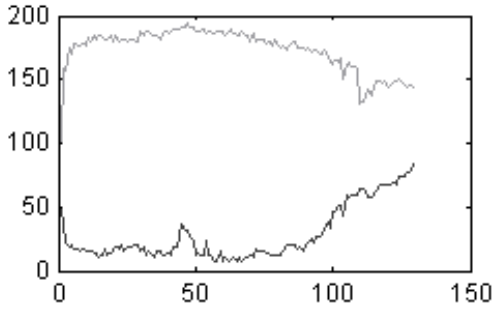


Рис. 3. Верхняя и нижняя границы трехмерной модели мозга без выравнивания.

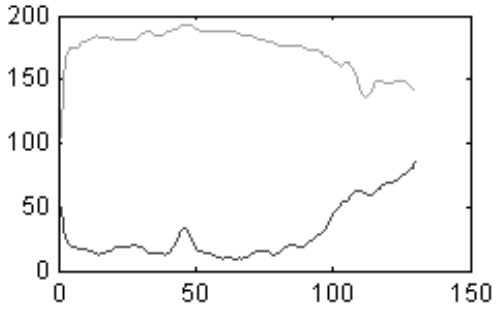


Рис. 4. Верхняя и нижняя границы трехмерной модели мозга с выравниванием.

жении вне мозга) проводится при помощи алгоритма, основанного на использовании разрезов графов [5, 6, 7].

4. Выравнивание освещенности на изображении проводится путем деления атласного изображения на карту освещенности, полученную применением Гауссовского фильтра большого радиуса. Такое преобразование позволяет как выровнять освещенность в пределах одного изображения, так и сделать одинаковой среднюю освещенность изображений по всему набору срезов. Пример предобработанного изображения приведен на рис. 2.

Из-за особенностей технологии получения атласных срезов, а именно симметризации, срезы плохо выровнены между собой. Поэтому вторым этапом предобработки является выравнивание изображений друг относительно друга. Выравнивание осуществляется при помощи полиномиального сглаживающего фильтра Савицкого–Голея, примененного к границам прямоугольников, обрамляющих срезы мозга.

На рис. 3 показаны верхняя и нижняя границы обрамляющих прямоугольников до выравнивания.

На рис. 4 показаны соответствующие границы после выравнивания.

Трехмерная модель мозга

Под трехмерной моделью мозга подразумевается функция $F\colon \mathbb{R}^3 \to [0,1]$. Атлас мозга позволяет восстановить значения F только на некотором дискретном множестве точек. В плоскостях срезов интерполяция непрерывной функции по дискретной проводится при помощи взвешенной суммы интенсивностей в соседних точках. Интерполяция в других плоскостях может быть проведена похожим образом (взвешенная сумма соседних точек). Но трехмерная модель, полученная таким образом, недостаточно гладкая. Еще одним существенным ее недостатком являются размытые границы внутренних структур. Это делает такую трехмерную модель непригодной для использования в биологических исследованиях.

Предлагаемый ниже метод «заполнения пустот», основанный на нелинейных деформациях соседних атласных срезов друг в друга, практически устраняет размытие границ, поскольку при деформациях структуры хорошо соотносятся. Гладкость модели также существенно повышается.

После нахождения всех нелинейных деформаций соседних срезов друг в друга, пространство между атласными срезами заполняется следующим образом:

$$F(x,y,z) = \alpha f^{\alpha}_{1,k-1}(x,y) + (1-\alpha) f^{1-\alpha}_{2,k}(x,y).$$

Здесь $\alpha = \frac{z-z_{k-1}}{z_k - z_{k-1}}$, $z_{k-1} \leqslant z < z_k$, z_k — z-координата среза номер k;

$$f^{\alpha}_{1,k-1}(x,y) = f_{k-1}((x,y) + \alpha(g^{k}_{k-1}(x,y) - (x,y)));$$
$$f^{\alpha}_{2,k}(x,y) = f_{k}((x,y) + \alpha(g^{k-1}_{k}(x,y) - (x,y)));$$

$g^j_i(x,y)$ — деформационная функция i-го среза в j-й.

Преобразования срезов

Методы построения нелинейных деформаций можно разделить на две группы. К первой группе относятся непараметрические локальные методы. Деформационная функция принадлежит функциональному пространству с очень мягкими ограничениями. Эти методы могут быть сформулированы при помощи скалярного критерия, который полностью определяет итоговое решение [8].

Вторая группа методов представляет собой параметрические модели, представляющие деформации умеренным числом параметров. Например это иерархические модели [9, 10], вейвлеты [11], базис тригонометрических функций [12], метод основанный на принципе динамического программирования [13].

В данной работе был использован подход, основанный на использовании B-сплайнов в качестве базиса, описанный в [14, 15].

Построение нелинейных деформаций

Исходные изображения представлены в виде пары двумерных дискретных функций

$$f_1, f_2\colon I \to [0,1],$$

где $I \subset \mathbb{Z}^2$ — двумерный дискретный интервал, покрывающий все пиксели двух изображений.

Целью является построение деформации изображения f_1 в изображение f_2:

$$f_1^c(g(x,y)) \approx f_2(x,y),$$

где $g(x,y)\colon \mathbb{R}^2 \to \mathbb{R}^2$ — деформационная функция, f_1^c — непрерывное продолжение f_1.

В качестве меры различия двух изображений выбрана сумма квадратов разностей интенсивностей:

$$E = \sum_{(i,j)\in I} \big(f_1^c(g(i,j)) - f_2(i,j)\big)^2. \qquad (1)$$

Представим деформационную функцию в виде линейной комбинации базисных функций.

$$g(x,y) = \sum_{k\in K} c_k b_k(x,y),$$

где K — множество индексов базисных функций, $c_k \in \mathbb{R}^2$. В качестве базисных функций b_k выбраны однородные кубические B-сплайны [1].

Использование разложения по базисным функциям позволяет свести задачу оптимизации в функциональном пространстве к задаче оптимизации сравнительно небольшого количества параметров.

Таким образом, деформационные функции ищутся в виде:

$$g(x,y) = \sum_{(k_x,k_y)\in K} \beta_3(x/h_x - k_x)\beta_3(y/h_y - k_y).$$

Центры B-сплайнов расположены на равномерной решетке $(k_x h_x, k_y h_y)$. Обработка однородных сплайнов производится значительно быстрее, чем обработка неоднородных. Чтобы получить полный контроль над деформационной функцией g, часть узлов решетки расположены за пределами изображений.

Итак, требуется решить оптимизационную задачу с функционалом (1) по параметрам c. Для оптимизации функционала (1) использовался метод градиентного спуска. Правило изменения параметров на каждом шаге: $\Delta c = -\mu \nabla_c E(c)$. Если такой шаг не приводит к уменьшению значения функционала (1), то μ делится на $\mu_f > 1$, иначе шаг выполняется и μ умножается на $\mu_f^* > 1$.

Пример деформационного поля, полученного при помощи описанного алгоритма, приведен на рис. 5.

[1] Кубический B-сплайн представляет собой функцию:

$$\beta_3(x) = \begin{cases} 2/3 - (1-|x|/2)x^2, & 0 < |x| \leqslant 1, \\ (2-|x|)^3/6, & 1 < |x| < 2, \\ 0, & |x| \geqslant 2. \end{cases}$$

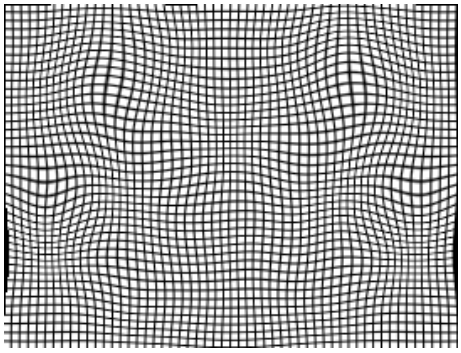


Рис. 5. Пример деформационного поля.

Результаты экспериментов

На рис. 6 изображен синтетический срез трехмерной модели, построенной тривиальным методом — при помощи интерполяции по взвешенной сумме соседних точек. Синтетический срез модели, построенной описанным выше методом приведен на рис. 7. Плоскость сечения этих срезов перпендикулярна плоскости атласных срезов.

Иллюстрации показывают недостаточную гладкость модели и размытые границы внутренних структур на рис. 6. На рис. 7 гладкость модели существенно выше. Границы внутренних контуров также отражены более четко.

Присутствует некоторая «ребристость» модели. Это вызвано деформациями срезов на этапе их получения. Устранение «ребристости» — одно из направлений дальнейших исследований.

Поскольку в Алленовском атласе мозга [4] присутствует разметка внутренних структур мозга, то по синтетическому гистологическому срезу можно построить синтетический срез с разметкой внутренних структур. Пример такого среза, изображен на рис. 8. Такие срезы могут приносить отдельную пользу в исследованиях мозга.

Выводы

В рамках данного исследования был разработан метод построения трехмерной модели мышиного мозга. Была построена трехмерная модель по данным Алленовского атласа мозга. Метод построения нелинейных деформаций достаточно стабилен и совмещает структуры мозга с требуемой точностью.

Недостатком трехмерной модели является недостаточное выравнивание срезов друг относительно друга. Устранение этого недостатка является дальнейшей целью.

Литература

[1] *Ng L., et al.* Neuroinformatics for Genome-Wide 3D Gene Expression Mapping in the Mouse Brain // IEEE Transactions on Computational Biology and Bioinfomatics, vol. 4, no. 3, 2007 — Pp. 382–393.

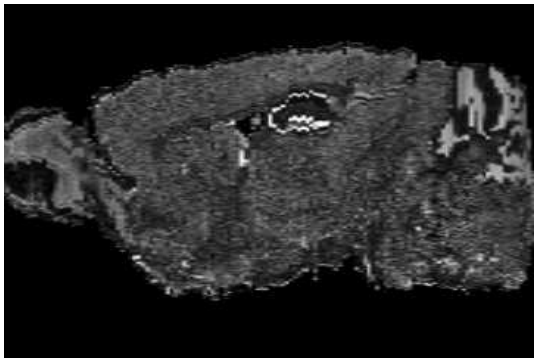


Рис. 6. Синтетический гистологический сагиттальный срез, построенный при помощи интерполяции по взвешенной сумме соседних точек.

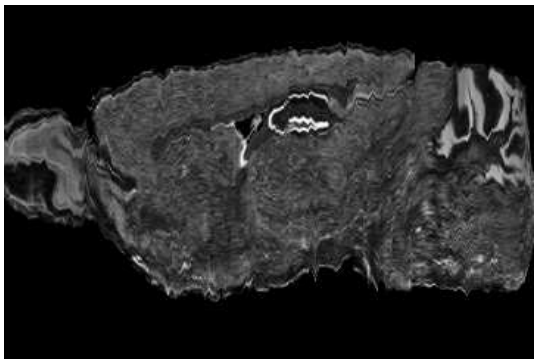


Рис. 7. Синтетический гистологический сагиттальный срез построенной трехмерной модели.

Рис. 8. Синтетический структурный сагиттальный срез построенной трехмерной модели.

[2] *Bolyne J., Lee E. F., Toga A. W* Digital atlases as a framework for data sharing // Frontiers in neuroscience, VOL. 2, 2008 — Pp. 100–106.

[3] *Lein E. S., et al.* Genome-wide atlas of gene expression in the adult mouse brain // Nature 445, 2007 — Pp. 168–176.

[4] Allen Brain Atlas [Internet]. Seattle (WA): Allen Institute for Brain Science. 2008. (`http://www.brain-map.org`).

[5] *Boykov Y., Veksler O., Zabih R.* Efficient Approximate Energy Minimization via Graph Cuts // In IEEE transactions on Pattern Analysis and Machine Intelligence (PAMI), 20(12), November 2001 — Pp. 1222–1239.

[6] *Kolmogorov V., Zabih R.* Energy Functions can be Minimized via Graph Cuts? // In IEEE transactions on Pattern Analysis and Machine Intelligence (PAMI), 26(2), February 2004 — Pp. 147–159.

[7] *Boykov Y., Kolmogorov V.* An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision // In IEEE transactions on Pattern Analysis and Machine Intelligence (PAMI), 26(9), September 2004 — Pp. 1124–1137.

[8] *Bajcsy R., Kovacic S.* Multiresolution elastic matching // Comput., Vis., Graph., Image Process., vol. 46, 1989 — Pp. 1-–21.

[9] *Moulin P., Krishnamurthy R., Woods J.* Multiscale modeling and estimation of motion fields for video coding // IEEE Trans. Image Processing, vol. 6, Dec. 1997 — Pp. 1606—1620.

[10] *Musse O., Heitz F., Armspach J.-P.* Topology preserving deformable image matching using constrained hierarchical parametric models // IEEE Trans. Med. Imag., vol. 10, July 2001 — Pp. 1081-–1093.

[11] *Wu Y., Kanade T., Li C., Cohn J.* Image registration using waveletbased motion model
Int. J. Comput. Vis., J. Le Moigne, Ed., vol. 38, 2000 — Pp. 129-–152.

[12] *Ashburner J., Friston K., Penny W.* Human brain function.

[13] *Ju T., et al.* 3D volume reconstruction of a mouse brain from histological sections using warp filtering // Journal of Neuroscience Methods, vol. 156, 2006 — Pp. 84—100.

[14] *Kybic J., Thevenaz P., Nirkko A., Unser M.* Unwarping of Unidirectionally Distorted EPI Images // IEEE Transactions on Medical Imaging, vol. 19, no. 2, 2000 — Pp. 80–93.

[15] *Kybic J., Unser M.* Fast Parametric Elastic Image Registration // IEEE Transactions on Image Processing, vol. 12, no. 11, 2003 — Pp. 1427–1442.

Спектральный подход в задаче распознавания и визуализации нечётких повторов в генетических последовательностях*

Панкратов А. Н., Горчаков М. А., Дедус Ф. Ф., Долотова Н. С., Куликова Л. И., Махортых С. А., Назипова Н. Н., Новикова Д. А., Ольшевец М. М., Пятков М. И., Руднев В. Р., Тетуев Р. К., Филиппов В. В.

pan@impb.ru

Москва, Факультет ВМиК МГУ им. М. В. Ломоносова;
Пущино, Институт математических проблем биологии РАН

На примере генетических последовательностей рассматривается задача поиска повторяющихся фрагментов в сигналах. Предложен принцип обнаружения повторов, основанный на сравнении спектров разложения сигнала по классическим ортогональным полиномам. Разрабатывается программное обеспечение и база данных для распознавания повторов в геномах.

Поиск повторов (гомологий) в нуклеотидных последовательностях — одна из основных вычислительных задач биоинформатики. Схожесть генетических текстов позволяет выдвигать гипотезы об их эволюционной и функциональной близости. Изучение повторов может внести существенный вклад в понимание структурно-функциональной организации геномов, наиболее интригующими вопросами которой являются проблема информационной избыточности геномов и проблема фрагментированности информации, кодирующей белки.

Исторически самым первым методом нахождения повторов в двух последовательностях является метод построения точечной матрицы сходства $M = (m_{ij})$ двух последовательностей ДНК (дот-матрицы), где $m_{ij} = 0$, если i-ый элемент первой последовательности не равен j-му элементу второй последовательности и $m_{ij} = 1$, если наоборот [1]. С помощью этого простого метода исследователь мог на графике идентифицировать участки сходства двух последовательностей. Непрерывные участки, состоящие из единиц и параллельные основной диагонали, соответствуют участкам сходства последовательностей. Позже было предложено множество разных фильтров для получения значимых результатов и исключения одиночных точек, зашумляющих рисунок. Самый простой — это сканирование диагоналей окном заранее определенной длины W и нанесение на плоскость рисунка отрезка диагонали длиной W только тогда, когда на протяжении окна встретилось заданное число B совпадений. Тогда речь идет о визуализации участков с B/W-процентным уровнем сходства [2]. Несколько другая схема визуализации используется для последовательностей, заданных на более длинных алфавитах, чем четырехбуквенный, т. к. точное совпадение символов становится более редким событием; тогда применяют матрицы весов замен символов для взвешивания каждой возможной пары символов, в долях единицы, а не просто в нулях и единицах. Если нормировать в каждом окне веса замен таким образом, чтобы в сумме они давали W единиц, то можно использовать тот же самый фильтр — окно считается подходящим, если сумма нормированных весов пар символов в этом окне превышает B единиц. Таким образом, с введением весовых матриц замен символов не надо смотреть, одинаковые ли символы стоят на пересечении соответствующих вертикали и горизонтали. Нужно просто подставить соответствующее значение из весовой матрицы.

В то время метод стал настолько популярным, что большинство универсальных пакетов программ для обработки генетических последовательностей обязательно включали в себя построение точечной матрицы гомологии. Например, пакет программ SAMSON [3, 4] обеспечивал дополнительный сервис — на одной и той же диаграмме выводились разными цветами прямые, инвертированные, комплементарно-инвертированные, комплементарные повторы. Участки сходства выдавались в отдельный файл в виде попарно выровненных последовательностей.

Решением проблемы зашумления рисунка статистически малозначимыми участками занимался ряд исследователей [8, 5, 6, 7, 9]. Проблеме статистической значимости найденных хитов (hit — участок локального сходства) на дот-матрицах посвящена работа Рейха и Мейске [10]. В ней выводится функция распределения случайной величины появления хита в окне.

К настоящему времени выросли объемы последовательностей, подлежащих сравнению. Теперь используются усовершенствованные методы построения дот-матриц для сравнения длинных (свыше 100 кБ) участков геномов и геномов целиком [11]. В основном усовершенствования заключаются в предварительной обработке сравниваемых последовательностей с тем, чтобы выявить протяженные блоки сходства и использовать их в построении матрицы. Известно современное программное обеспечение, предназначенное для построения матриц сходства, такое как OWEN [12] и DPView.

*Работа выполнена при финансовой поддержке РФФИ, проекты №08-01-12030, 08-07-00353, 07-01-00564.

Предлагаемый в данной работе алгоритм поиска повторов также основан на построении матрицы сходства, однако, сравнение производится целыми фрагментами генетических последовательностей, а само сравнение основано на применении обобщенного спектрально-аналитического метода [13]. Получаемые матрицы отличаются от полученных классическими методами, поэтому предложенные матрицы можно назвать матрицами спектральной гомологии по аналогии с классическими матрицами точечной гомологии.

Алгоритм удобно разделить на несколько относительно самостоятельных этапов. На первом этапе алгоритм создает функцию-профиль из ДНК-последовательности методом скользящего окна, в котором ведется расчет доли гуанина G и цитозина C. Этот способ построения профиля имеет физический смысл силы связи двойной спирали ДНК, поскольку комплементарная связь G и C, образованная тремя парами водородных связей, сильнее связи A и T, образованной двумя парами водородных связей. Кроме того, построенный профиль инвариантен по отношению к комплементарным заменам в повторах. Этим решается проблема поиска комплементарных повторов. Параметром этого этапа является ширина окна статистического усреднения последовательности.

На втором этапе функция-профиль переводится в спектральное представление с использованием классических полиномиальных базисов семейства Якоби. Это представление на следующей стадии используется для сравнения на основе некоторого специально разработанного критерия. Заметим, что в данной задаче обоснованно использование аппроксимативных свойств такого базиса, весовая функция которого стремится плавно к нулю на концах интервала аппроксимации, что может компенсировать размывание границ повтора, которое происходит на этапе построения статистического профиля. Параметрами этого этапа становятся ширина окна и глубина спектрального оценивания. Вычисления показали, что метод устойчиво работает, если ширина окна аппроксимации более чем в два превышает ширину окна статистического оценивания.

На третьем этапе критерий оценивает отличия векторов коэффициентов разложения. Построение критерия состоит в выборе метрики в функциональном пространстве и ее нормировании. Выбор метрики обусловлен в основном выбором базиса, его параметров и глубины разложения. Нормирование имеет целью получить характеристику, инвариантную по отношению к величине сигнала. Рассматривались два варианта нормирования метрики:

$$\theta_1(x(t), y(t)) = \frac{\|x(t) - y(t)\|}{\|x(t)\| + \|y(t)\|};$$

$$\theta_2(x(t), y(t)) = \frac{\|x(t) - y(t)\|}{1 + \|x(t) - y(t)\|};$$

где $x(t)$ и $y(t)$ — две сравниваемых функции, соответствующие различным участкам функционального профиля, а $\|\cdot\|$ — евклидова норма в N-мерном пространстве коэффициентов разложения сигнала. Обе нормировки пригодны для идентификации повторов при сравнении их с некоторыми пороговыми значениями, которые становятся параметрами метода.

Перечислим основные свойства предложенного алгоритма, которые характеризуют его эффективность. Использование индексирования последовательности позволяет существенно ускорить попарное сравнение всех фрагментов покрытия последовательности. Спектральное представление позволяет получать спектры инвертированного образца непосредственно из спектра прямого шаблона, если используемый базис состоит из четных и нечетных базисных функций. Таким образом, поиск инвертированных повторов практически не добавляет вычислительной сложности алгоритму. Метрика монотонна по числу коэффициентов разложения, что придает дискриминантный характер процессу сравнения. Например, если сравнение по первым коэффициентам превышает пороговое значение, то дальнейшее вычисление метрики необязательно. Алгоритм полностью построен на вычислениях с плавающей точкой и хорошо векторизуется и распараллеливается.

Предложена и опробована векторно-параллельная реализация алгоритма. Методология этой реализации алгоритма основана на экономном использовании оперативной памяти для достижения линейной масштабируемости по числу ядер вычислительной системы с общей памятью. Это достигается отказом от хранения матрицы преобразования Фурье взамен вычисления ее при каждом вычислении коэффициентов разложения. Эффективное вычисление этой матрицы возможно благодаря рекуррентным соотношениям, соединяющим строки матрицы. В то же время, возможно и более полное использование векторных операций современных процессоров. Запись алгоритма в векторном виде позволяет практически полностью избавиться от циклов. Как показали пробные расчеты, использование библиотеки векторных операций Intel IPP позволяет повысить быстродействие на порядок. В сочетании с использованием эффективного распараллеливания задачи это является залогом высокой эффективности алгоритма.

Разработана модель представления данных для базы данных структурно-функциональных элемен-

тов геномов, спроектирована и построена специализированная база данных GENome REVisor. База данных реализована с достаточным уровнем общности, что позволяет хранить генетические последовательности и структурно-функциональные элементы геномов различных организмов, как эукариот, так и прокариот. Модульная, расширяемая структура базы данных позволяет хранить разные типы структурных элементов геномов и в перспективе даст возможность построить базу знаний. Разработан формат входных/выходных файлов для базы данных на основе стандарта XML, что даёт возможность пополнения базы данных сразу большим количеством разнородных структурно-функциональных элементов. Прототип базы данных доступен в сети Интернет, `http://www.jcbi.ru/`.

В настоящее время ведется работа по дальнейшему совершенствованию этого подхода, в частности,

1) оптимизация выбора базиса из семейства Якоби;
2) редукции числа параметров задачи;
3) разработка метода «обучения» алгоритма поиска повторов;
4) составления «азбуки» повторов;
5) определения координат повторов;
6) определения статистической значимости повторов;
7) разработке алгоритма для массивно параллельных вычислительных систем;
8) разработке базы знаний, построенной на основе найденных повторов различных типов.

Литература

[1] *Gibbs A. J., McIntyre G. A.* The diagram, a method for comparing sequences. Its use with amino acid and nucleotide sequences // Euro. J. Biochem., 1970, V. 16, Pp. 1–11.

[2] *Staden R.* An interactive graphics program for comparing and aligning nucleic acid and amino acid sequences // Nucl. Acids Res., 1982, V. 10, Pp. 2951–2961.

[3] *Вернослов С. Е., Кондрашов А. С., Ройтберг М. А., Шабалина С. А., Юрьева О. В., Назипова Н. Н.* Пакет программ для анализа первичных структур биополимеров САМСОН // Мол. биология, 1990, Т. 24, С. 524–529.

[4] *Nazipova N. N., Shabalina S. A., Ogurtsov A. Yu., Kondrashov A. S., Roytberg M. A., Buryakov G. V., Vernoslov S. E.* SAMSON: a software package for the biopolymer primary structure analysis // CABIOS, 1995, V. 11, No. 4, Pp.423–426.

[5] *McLachlan A. D., Bosswell D. R.* Confidence Limits for Homology in Protein or Gene Sequences. The c-myc Oncogene and Adenovirus E1a Protein // J. Mol. Biol., 1985, V. 185, Pp. 39–49.

[6] *Brooks L. D., Weir B. S., Schaffer H. E.* The Probabilities of Similarities in DNA Sequence Comparisons // Genomics, 1988, V. 3, Pp. 207–216.

[7] *Queen C. L., Korn L. J.* Computer analysis of nucleic acids and proteins // In «Methods in Enzymology» Eds. L. Grossman and K. Moldave, Academic Press, New York, 1980, V. 65, Pp. 595–609.

[8] *Arratia R., Gordon L., Waterman M. S.* An extreme value theory for sequence matching // Ann. Stat., 1985, V. 14, Pp. 971–993.

[9] *Smith T. F., Waterman M. S., Burks C.* The statistical distribution of nucleic acid similarities // Nucleic Acids Res., 1985, V. 13, Pp. 645–656.

[10] *Reich J. G., Meiske W.* A simple statistical significance test of window scores in large dot matrices obtained from protein or nucleic acid sequences // Comput. Appl. Biosci., 1987, V. 3, No. 1, Pp.25–30.

[11] *Szafranski K., Jahn N., Platzer M.* Fast pairwise nucleotide sequence comparison with noise suppression // Bioinformatics, 2006, V. 22, No. 15, Pp. 1917–1918.

[12] *Ogurtsov A. Y., Roytberg M. A., Shabalina S. A., Kondrashov A. S.* OWEN: aligning long collinear regions of genomes // Bioinformatics, V. 18, No. 12, 2002, Pp. 1703–1704.

[13] *Дедус Ф. Ф., Куликова Л. И., Махортых С. А., Назипова Н. Н., Панкратов А. Н., Тетуев Р. К.* Аналитические методы распознавания повторяющихся структур в геномах // Доклады Академии Наук, 2006, Т. 411, № 5, С. 599–602.

Поиск представления молекул и методы прогнозирования активности в задаче «структура–свойство»*

Прохоров Е. И., Перевозников А. В., Воропаев И. Д., Кумсков М. И., Пономарёва Л. А.
qsar_msu@mail.ru, eugeny.prokhorov@gmail.com
Москва, Механико-математический факультет МГУ им. М. В. Ломоносова

В работе рассматривается проблема поиска представления молекул и методы прогнозирования активности в задаче «структура–свойство». Приводятся постановки основных задач, решаемых QSAR-анализом. Предлагаются методы построения прогностических моделей, основанные на кластерной структуре обучающей выборки. Экспериментальные результаты получены для выборки гликозидов и используются для дальнейшей работы по моделированию химических веществ.

Задача «структура–свойство» (QSAR — Quantitative Structure Activity Relationship) — актуальная задача распознавания образов — состоит в том, чтобы по структуре химического соединения предсказать его активность (химическую или биологическую) [9, 7, 8].

Объектами распознавания в данном случае являются молекулы, а классами — классы активности химических веществ.

Особенностью QSAR-задачи является необходимость описать структуру химического соединения в виде дескрипторов — любых свойств молекулы, выраженных численно. Дескрипторы выступают в роли признаков объекта распознавания.

Поэтому решение задачи разбивается на два основных этапа: этап построения описания обучающей выборки, на котором формируется матрица «молекула–дескриптор» и этап поиска функциональной зависимости.

Для анализа была предложена выборка гликозидов, по которой построены модели с высокой прогностической способностью. В ходе работы авторы использовали новый подход к построению дескрипторов молекул, использующий нечёткие функции принадлежности [3]. Для этапа поиска функциональной зависимости предложены новые эволюционные методы распознавания.

Определения и постановка задачи

M-граф (меченый молекулярный граф $G = \{E, V\}$) — это помеченный граф, вершины которого интерпретируются как атомы молекулы, а ребра — как валентные связи между парами атомов. Метки вершин и ребер (числа или символы) отражают локальные свойства атомов и химических связей. В качестве меток вершин могут быть использованы любые характеристики соответствующих атомов (например, трехмерные координаты, символ химического элемента, заряд ядра, поляризуемость, атомный вес, атомный радиус и др.), а в качестве меток ребер — любые характеристики соответствующих связей (кратность, длины, порядки связей, полученные из квантово-химических расчетов, и т. д.)

Задача «структура–свойство» заключается в следующем.

Задана обучающая (эталонная) выборка — база данных из N химических соединений, в которой:

— каждое i-ое соединение представлено меченым молекулярным графом G_i, имеющим укладку в трехмерном пространстве (т. е. для каждой вершины в качестве меток заданы её трехмерные координаты);
— соединение G_i либо отнесено к C_i — одному из K классов активности (например, «активных», «слабоактивных», «неактивных» веществ) согласно исследуемому свойству A_i, либо для G_i задано численное значение свойства A_i.

Требуется построить классифицирующую функцию F, получающую в качестве аргумента произвольный молекулярный граф с метками того же типа, и «наилучшим образом» относящую это соединение к одному из классов активности, либо «наилучшим образом» предсказывающую численное значение исследуемого свойства.

Функционал качества $\varphi(F)$ задан как доля молекул из обучающей выборки, верно классифицированных функцией F:

$$\varphi(F) = 1 - \frac{1}{N}\sum_{i=1}^{N}\varepsilon_i, \qquad (1)$$

где $\varepsilon_i = 0$ если $F(G_i) = C_i$, и $\varepsilon_i = 1$ в противном случае. В случае, когда функция должна предсказывать численное значение свойства,

$$\varphi(F) = 1 - \sum_{i=1}^{N}\big(F(G_i) - A_i\big)^2 \Big/ \sum_{i=1}^{N} A_i^2. \qquad (2)$$

Поставленную таким образом задачу поиска классифицирующей функции будем называть задачей «структура–свойство» или QSAR-задачей.

Дескриптором будем называть какое-либо свойство, численное значение которого может быть вычислено для произвольного молекулярного графа G (в распознавании это принято называть признаком).

*Работа выполнена при финансовой поддержке РФФИ, проект № 07-07-00282.

Алфавитом дескрипторов будем называть множество всех дескрипторов, используемых для анализа обучающей выборки, обозначенных различными символьными метками.

Пусть алфавит дескрипторов состоит из M элементов. *Вектором признаков* молекулярного графа G будем называть вектор $\boldsymbol{x} = (x_1, \ldots, x_M) \in \in \mathbb{R}^M$, где x_j — значение j-ого дескриптора, вычисленное для G.

МД-матрицей или матрицей «молекула–дескриптор» (матрицей признаков) для рассматриваемой обучающей выборки будем называть матрицу размера $N \times M$, в i-ой строке которой стоит вектор признаков i-го соединения $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iM})$.

Задача построения описания обучающей выборки включает в себя выбор алфавита дескрипторов, построение отображения из множества молекулярных графов в признаковое пространство $\mathbb{R}^M$ и формирование матрицы «молекула–дескриптор» для обучающей выборки. Подробнее о поиске представления молекул в QSAR-задаче см. в [2, 3].

Для каждой молекулы из обучающей выборки известно также значение её активности (целевое свойство) y_i, $i = 1, \ldots, N$. Задача распознавания состоит в определении активности нового соединения молекулы x по её описанию $\boldsymbol{x} = = (x_1, \ldots, x_M)$ и информации об обучающей выборке $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)$.

Используя обучающую выборку, будем строить модели, предсказывающие активность молекул. Для оценки прогностической способности моделей будем использовать коэффициент скользящего контроля R^2_{cv} (cross validation) [3, 5], вычисляемый в ходе следующей процедуры:

— в цикле по числу молекул из выборки удаляется текущая молекула;
— по оставшимся в обучающей выборке молекулам строится модель;
— с помощью модели предсказывается значение активности удалённой молекулы.

Процент успешных прогнозов в этом случае и есть коэффициент R^2_{cv}.

Методы решения

Задача построения описания обучающей выборки была подробно описана в [1, 2]. Отдельно стоит отметить проведённый анализ зависимости качества прогноза от типа функции принадлежности, на основе которой формируются значения дескрипторов [3].

Решение задачи поиска функциональной зависимости разбивается на несколько этапов: кластеризация и обработка выбросов; отбор значимых дескрипторов; построение модели; прогноз. Опишем подробно каждый этап.

Кластеризация позволяет строить модель локально, внутри небольшой группы сходных соединений, что часто оказывается очень полезным. Задача кластеризации чрезвычайно важна также для ускорения вычислений, ведь построение моделей внутри каждого кластера может идти параллельно, сам алгоритм кластеризации также может быть распараллелен.

Эффективность проводимых вычислений важна, так как необходимо обрабатывать большие массивы молекулярных структур, собранных во многих организациях, например, таких, как National Cancer Institute [www.cancer.gov], где выборки содержат сотни молекул, для каждой из которых обрабатываются тысячи дескрипторов.

На этом этапе идёт также обработка выбросов.

Выброс — химическое соединение в выборке, признаки которого существенно отличаются от признаков остальных соединений. Такие молекулы могут попасть в выборки из-за ошибки составителей, но могут сами по себе являться содержательными с точки зрения химии. Присутствие выбросов существенно ухудшает качество прогноза. Поэтому при построении модели мы их не учитываем.

Из-за огромного числа дескрипторов многие алгоритмы распознавания становятся неприменимыми для нашей задачи, поэтому их число необходимо сократить, отобрав при этом наиболее существенные для прогноза. В рассматриваемых реализациях были выбраны разные подходы к этой проблеме. Например, результирующую модель на основе системы нечёткого логического вывода ANFIS [5] предлагалось строить на главных компонентах матрицы «молекула–дескриптор». В другом случае была применена модификация МГУА (метода группового учёта аргументов) для алгоритма kNN (k ближайших соседей) с ограничением по радиусу [3]. Особое внимание стоит обратить на двоичный МГУА [4], реализующий поиск зависимостей на двоичных векторах при двоичном целевом векторе с использованием эволюционного построения семейства ДНФ/КНФ на большом исходном пространстве признаков. Вычислительные эксперименты показали, что при применении к выборке гликозидов предложенный метод строит прогнозирующие модели с меньшим числом выбросов и более высоким качеством прогноза, чем классический аналог.

На рис. 1 представлена общая схема схема решения задачи.

Полученные результаты

Для анализа была представлена выборка гликозидов. Гликозид — органическое вещество, молекулы которого состоят из углевода и неуглеводного компонента (агликона), соединенных гликозидной связью. Гликозиды служат формой переноса и хранения различных веществ растений. Сердечные гликозиды наперстянки применяют в медицине.

Рис. 1. Общая схема схема решения задачи (k — число кластеров).

Гликозиды представляют собой обширную группу органических веществ, встречающихся в растительном (реже в животном) мире и/или получаемых синтетическим путём.

По выборке были сформированы 24 матрицы с различными параметрами, описывающие 76 молекул, в зависимости от способа разбиения интервала электростатического заряда (2 варианта), типа функции принадлежности — четкие, нечеткие треугольные, нечеткие трапециевидные (3 варианта) и количества разбиений интервала расстояний между особыми точками (ОТ) и между ОТ парой ОТ (еще 4 варианта).

Число дескрипторов в построенных матрицах (порядка 2000) варьировалось в зависимости от того, с какими настройками была построена матрица. К указанным матрицам были применены описанные подходы. На каждом этапе вычисления проводились с различными параметрами. Исходя из качества полученного прогноза, формировались рекомендации по изменению параметров, использованных на этапе описания молекул и формирования дескрипторов. Далее весь алгоритм запускается заново уже с новыми параметрами.

Подробные цифры и сравнение результатов для различных подходов приведены в [3, 4]. Здесь же отметим, что коэффициенты R^2_{cv} для построенных моделей высоки для задач типа «структура–свойство», и позволяют рассчитывать на дальнейший успех в работе с этим химическим свойством.

Выводы

Для выборки гликозидов построены модели с высокой прогностической способностью. Проведено сравнение моделей. Нашли применение результаты, касающиеся построения нечетких дескрипторов, предложенных в ранних работах авторов. Предложен новый эволюционный метод поиска логических зависимостей для обработки МД-матриц. Накоплен опыт, позволяющий быстро строить качественные (в смысле коэффициента R^2_{cv}) модели, прогнозирующие активность химических веществ.

Литература

[1] *Кохов В. А.* Метод количественного определения сходства графов на основе структурных спектров // Известия РАН, Техническая кибернетика. — 1994. — № 5. — С.143–159.

[2] *Devetyarov D. A., Zaharov A. M., Kumskov M. I., Ponomareva L. A.* Fuzzy logic application for construction of 3D descriptors of molecules in QSAR problem // 8th Intern. Conf. «Pattern Recognition and Image Analysis: New Information Technologies». — 2007. — Vol. 2. — Pp. 249–252.

[3] *Деветьяров Д. А., Кумсков М. И., Апрышко Г. Н., Носеевич Ф. М. и др.* Сравнительный анализ применения нечетких дескрипторов при решении задачи «структура–активность» для выборки гликозидов // Всеросс. конф. ММРО-14. — М.: МАКС Пресс, 2009. — С. 511–514.

[4] *Носеевич Ф. М., Деветьяров Д. А., Кумсков М. И., Апрышко Г. Н., Пермяков Е. А.* Двоичный метод группового учета аргументов в задаче «структура–активность» // Всеросс. конф. ММРО-14. — М.: МАКС Пресс, 2009. — С. 575–578.

[5] *Штовба С. Д.* Введение в теорию нечетких множеств и нечеткую логику. — Винница: Изд-во винницкого гос. техн. университета, 2001. — 198 с.

[6] *Журавлев Ю. И, Рязанов В.В., Сенько О.В.* «Распознавание». Математические методы. Программная система. Практические применения — Москва: ФАЗИС, 2006.

[7] *Кумсков М. И., Смоленский Е. А., Пономарева Л. А., Митюшев Д. Ф., Зефиров Н. С.* Системы структурных дескрипторов для решения задач «структура-свойство» // М.: Наука Доклады Академии Наук, 1994. — С. 336.

[8] *Деветьяров Д. А., Григорьева С. С., Пермяков Е. А., Кумсков М. И., Понаморёва Л. А., Свитанко И. В.* Решение задачи «структура–свойство» для молекул с множеством пространственных конформаций // Система прогнозирования свойств химических соединений: Алгоритмы и модели: Сборник научных работ / Под ред. М. И. Кумскова. Москва: МАКС Пресс, 2008.

[9] *Григорьева С. С., Кумсков М. И., Захаров А. М.* Применение метода главных компонент при построении кластерной структуры обучающей выборки молекул // 13-я всеросс. конф. ММРО-13: Сборник докладов. — Москва: МАКС Пресс, 2007.

Локальная модель случайных эволюционных преобразований белков и вероятностное обобщение задачи множественного выравнивания аминокислотных последовательностей

Разин Н. А., Сулимова В. В., Моттль В. В., Мучник И. Б.
nrmanutd@gmail.com, vsulimova@yandex.ru, vmottl@yandex.ru, muchnikilya@yahoo.com
Москва, ВЦ РАН, МФТИ,
Тула, Тульский Государственный Университет,
США, Нью Брансвик, Университет Ратгерс

В данной работе предлагается вероятностная модель эволюционных преобразований аминокислотных последовательностей, являющаяся прямым обобщением общепринятой в биоинформатике модели эволюции аминокислот, предложенной Маргарет Дэйхофф. На основе данной модели сформулирована принципиально новая вероятностная постановка задачи множественного выравнивания как задача поиска наиболее правдоподобной общей последовательности-прародителя для группы аминокислотных последовательностей. Численные эксперименты с модельными данными подтвердили эффективность предложенного решения сформулированной задачи.

Молекулярная биология является массовым источником данных в виде массивов последовательностей разной длины, в частности, последовательностей аминокислот, образующих полимерные молекулы белков. Аминокислотная последовательность (так называемая первичная структура белка), представляет собой символьную последовательность индивидуальной длины над 20-буквенным алфавитом аминокислот и содержит, как правило, несколько сотен элементов, а иногда больше тысячи.

Одной из фундаментальных проблем современной биоинформатики, возникающей при решении целого ряда важнейших задач анализа белковых данных, является проблема измерения группового сходства аминокислотных последовательностей. В качестве инструментов для ее решения в настоящее время используются алгоритмы множественного выравнивания [9, 4, 6, 7]. Многие из них сопровождают результат так называемым профилем анализируемой совокупности последовательностей, под которым понимается некоторый самостоятельный «обобщенный» белок в виде последовательности дискретных распределений вероятностей над множеством всех аминокислот. Однако проблема заключается в том, что существующие алгоритмы, во-первых, не основаны на какой-либо формальной постановке задачи и, во-вторых, не базируются на какой-либо единой модели эволюционной модификации белков.

В данной работе предложена простейшая модель эволюционных преобразований аминокислотных последовательностей. На ее основе сформулирована принципиально новая вероятностная постановка задачи множественного выравнивания группы последовательностей как задача поиска для них наиболее правдоподобной общей последовательности-прародителя известной длины. При этом прародитель ищется в виде последовательности независимых дискретных распределений вероятностей на множестве аминокислот, что соответствует общепринятому понятию профиля группы последовательностей.

Модель эволюции аминокислот PAM

В качестве теоретического прототипа модели эволюционных преобразований аминокислотных последовательностей в данном исследовании используется вероятностная теория эволюционных чередований в алфавите аминокислот, предложенная Маргарет Дэйхофф [3] и широко известная в современной биоинформатике под названием PAM (Point Accepted Mutation).

Пусть $A = \{\alpha^{(1)}, \ldots, \alpha^{(20)}\}$ — конечное множество аминокислот. Эволюционная модель PAM предполагает, что склонность аминокислот к взаимному мутационному превращению количественно выражается квадратной матрицей условных вероятностей

$$\mathbf{\Psi} = \left[\psi_{ij}\right]_{i=1}^{20}{}_{j=1}^{20},\ \psi_{ij} = \psi(\alpha^j \,|\, \alpha^i),\ \alpha^i, \alpha^j \in A,$$

интерпретируемых как вероятность того, что аминокислота α^i на очередном шаге эволюции превратится в аминокислоту α^j.

Основным математическим понятием модели PAM является понятие марковской цепи истории эволюционного изменения аминокислоты в отдельно взятой позиции h_t, $t = 1, 2, 3, \ldots$ при некоторой достаточно сложной интерпретации понятия «величины шага» $(\ldots, t, t{+}1, \ldots)$, которую мы здесь не рассматриваем. Предполагается, что данный марковский процесс обладает двумя свойствами.

1. Эргодичность — существование финального распределения вероятностей:

$$\xi(\alpha^j) = \sum_{\alpha^i \in A} \xi(\alpha^i)\psi(\alpha^j \,|\, \alpha^i).$$

2. Обратимость: $\xi(\alpha^i)\psi(\alpha^j \,|\, \alpha^i) = \xi(\alpha^j)\psi(\alpha^i \,|\, \alpha^j)$.

Локальная вероятностная модель происхождения совокупности белков

Пусть $A = \{\alpha^{(1)}, \ldots, \alpha^{(20)}\}$ — множество аминокислот, $\bar{\boldsymbol{\Omega}} = \{\boldsymbol{\omega}_j\}_{j=1}^{M}$ — анализируемая совокупность аминокислотных последовательностей $\boldsymbol{\omega}_j = (\omega_{jt})_{t=1}^{N_j}$ индивидуальной длины N_j.

Предлагаемая в данной работе вероятностная модель происхождения группы последовательностей основана на следующих гипотезах.

Гипотеза 1. Все белки получены независимо друг от друга из некоторой общей аминокислотной последовательности-прародителя $\boldsymbol{\vartheta} = (\vartheta_i)_{i=1}^{n}$ известной длины n.

Гипотеза 2. Элементы ϑ_i последовательности-прародителя $\boldsymbol{\vartheta}$ выбраны из алфавита аминокислот независимо в соответствии с неизвестными наблюдателю распределениями вероятностей $\boldsymbol{\beta}_i = (\beta_{ik})_{k=1}^{20}$, где β_{ik} удовлетворяют условиям $0 \leqslant \beta_{ik} \leqslant 1$ для всех i, k и $\sum\limits_{k=1}^{m} \beta_{ik} = 1$ для всех i.

Совокупность распределений $\bar{\boldsymbol{\beta}} = (\boldsymbol{\beta}_i)_{i=1}^{n}$, определяющая последовательность-прародитель $\boldsymbol{\vartheta} = (\vartheta_i)_{i=1}^{n}$, соответствует общепринятому понятию профиля группы последовательностей.

Гипотеза 3. Каждый белок $\boldsymbol{\omega}_j$, $j = 1, \ldots, M$ получен из последовательности-прародителя $\boldsymbol{\vartheta}$ в два этапа в соответствии со следующей моделью преобразования последовательностей.

На первом этапе генерируется структура $\boldsymbol{v}$ преобразования последовательности-прародителя $\boldsymbol{\vartheta}$ фиксированной длины n в формируемую последовательность $\boldsymbol{\omega}_j$ длины N_j. Данная структура представляет собой монотонно возрастающую последовательность длины n из целых чисел $\boldsymbol{v} = (v_i)_{i=1}^{n}$, $v_1 \geqslant 1$, однозначно определяющую позиции элементов формируемой последовательности, в которые будут преобразованы соответствующие элементы исходной последовательности-прародителя ϑ_i. То есть структура преобразования определяет выравнивание элементов исходной и формируемой последовательностей. В связи с этим структуру преобразования $\boldsymbol{v}$ будем также называть выравниванием. При этом следует обратить внимание, что не любая структура преобразования позволяет получить конкретную последовательность $\boldsymbol{\omega}_j$ длины N_j из последовательности $\boldsymbol{\vartheta}$ длины n, в частности, не допустимыми являются преобразования, для которых $v_n > N_j$.

Структура преобразования генерируется в соответствии с некоторым распределением вероятностей $q_{N_j}(\boldsymbol{v})$ на множестве всех допустимых вариантов преобразования V_{N_j}, удовлетворяющим обычным условиям $0 \leqslant q_{N_j}(\boldsymbol{v}) \leqslant 1$ и $\sum\limits_{\boldsymbol{v} \in V_{N_j}} q_{N_j}(\boldsymbol{v}) = 1$.

Распределение $q_{N_j}(\boldsymbol{v})$ может выбираться различными способами. В данной работе оно выбирается так, что вероятность конкретного выравнивания $\boldsymbol{v}$ зависит только от разностей $(v_i - v_{i-1})$, $i = 2, \ldots, n$, и не зависит от их абсолютных значений, т. е. она инвариантна к одновременному сдвигу позиций v_i вдоль формируемой последовательности. В связи с этим предложенная модель случайного преобразования названа локальной, по аналогии с общепринятым термином локального выравнивая последовательностей.

На втором этапе генерируются аминокислоты, из которых будет состоять формируемая последовательность. При этом каждый элемент исходной последовательности ϑ_i преобразуется, согласно сгенерированной на первом этапе структуре $\boldsymbol{v}$, в элемент, находящийся в позиции v_i формируемой последовательности, в соответствии с некоторым распределением вероятностей $\eta(\omega_{v_i} \,|\, \vartheta_i)$. Данное условное распределение, естественно строить на основе случайного преобразования аминокислот, принятого в модели PAM, учитывая при этом случайность выбора аминокислот последовательности-прародителя:

$$\eta(\omega_{v_i} \,|\, \vartheta_i) = \sum_{k=1}^{20} \beta_{ik} \psi\big(\omega_{v_i} \,|\, \alpha^{(k)}\big).$$

Остальные элементы формируемой последовательности ω_t, $t \neq v_i$, соответствующие вставкам новых элементов в исходную последовательность, данная модель не учитывает.

В соответствии с предложенной моделью, функция правдоподобия происхождения одного белка $\boldsymbol{\omega}_j$ из последовательности-прародителя $\boldsymbol{\vartheta}$, задаваемой последовательностью распределений $\bar{\boldsymbol{\beta}}$, будет иметь вид:

$$f(\boldsymbol{\omega}_j \,|\, \bar{\boldsymbol{\beta}}) \propto \sum_{\boldsymbol{v} \in V_{N_j}} q_{N_j}(\boldsymbol{v})\, \eta_n(\boldsymbol{\omega}_j \,|\, \bar{\boldsymbol{\beta}}, \boldsymbol{v}),$$

где $\eta_n(\boldsymbol{\omega}_j \,|\, \bar{\boldsymbol{\beta}}, \boldsymbol{v}) = \prod\limits_{i=1}^{n} \sum\limits_{k=1}^{20} \beta_{ik} \psi\big(\omega_{v_i} \,|\, \alpha^{(k)}\big)$.

Оценивание вероятностной модели прародителя

Цель анализа: оценивание последовательности распределений вероятностей $\bar{\boldsymbol{\beta}} = (\boldsymbol{\beta}_1 \ldots \boldsymbol{\beta}_n)$ по всей совокупности анализируемых белков $\bar{\boldsymbol{\Omega}} = \{\boldsymbol{\omega}_j\}_{j=1}^{M}$.

Оценивать вероятностный профиль общей последовательности-прародителя будем в соответствии с принципом максимального правдоподобия, максимизируя логарифмированную функцию правдоподобия совместного порождения всех ана-

лизируемых белков $F(\bar{\boldsymbol{\omega}} \mid \bar{\boldsymbol{\beta}})$:

$$\hat{\bar{\boldsymbol{\beta}}} = \arg\max \ln F(\bar{\boldsymbol{\omega}} \mid \bar{\boldsymbol{\beta}}) = \arg\max \sum_{j=1}^{M} \ln f(\boldsymbol{\omega}_j \mid \bar{\boldsymbol{\beta}}) =$$

$$= \arg\max \sum_{j=1}^{M} \ln \sum_{\boldsymbol{v} \in V_{N_j}} q_{N_j}(\boldsymbol{v})\, \eta_n(\boldsymbol{\omega}_j \mid \bar{\boldsymbol{\beta}}, \boldsymbol{v}). \quad (1)$$

В основу решения данной задачи положена итерационная EM-процедура, впервые предложенная М. И. Шлезингером в 1965 году [2], которая применима к широкому классу функций правдоподобия для вероятностных моделей со скрытыми параметрами. В данном случае в качестве таких скрытых параметров выступают выравнивания $\boldsymbol{v} \in V_{N_j}$ скрытой последовательности-прародителя с анализируемыми белками.

Пусть на s-м шаге получено приближение к искомому профилю последовательности-прародителя $\bar{\boldsymbol{\beta}}^s = (\boldsymbol{\beta}_1^s, ..., \boldsymbol{\beta}_n^s)$ и пусть найдено апостериорное распределение на множестве выравниваний j-го белка $p(\boldsymbol{v} \mid \boldsymbol{\omega}_j, \bar{\boldsymbol{\beta}}^s)$, $\boldsymbol{v} \in V_{nN_j}$, в предположении, что $\bar{\boldsymbol{\beta}}^s$ — истинный профиль исходной последовательности. Выберем $\bar{\boldsymbol{\beta}}^{s+1}$ по правилу:

$$\bar{\boldsymbol{\beta}}^{s+1} =$$

$$= \arg\max \sum_{j=1}^{M} \sum_{\boldsymbol{v} \in V_{N_j}} p(\boldsymbol{v} \mid \boldsymbol{\omega}_j, \bar{\boldsymbol{\beta}}^s) \ln \eta_n(\boldsymbol{\omega}_j \mid \bar{\boldsymbol{\beta}}, \boldsymbol{v}). \quad (2)$$

Теорема 1. *При определении $\bar{\boldsymbol{\beta}}^{s+1}$ в соответствии с (2) справедливо неравенство $F(\bar{\boldsymbol{\omega}} \mid \bar{\boldsymbol{\beta}}^{s+1}) \geqslant F(\bar{\boldsymbol{\omega}} \mid \bar{\boldsymbol{\beta}}^s)$, причем $F(\bar{\boldsymbol{\omega}} \mid \bar{\boldsymbol{\beta}}^{s+1}) = F(\bar{\boldsymbol{\omega}} \mid \bar{\boldsymbol{\beta}}^s)$, тогда и только тогда, когда $\nabla_{\boldsymbol{\beta}_i} F(\bar{\boldsymbol{\omega}} \mid \bar{\boldsymbol{\beta}}^s) = 0$ для всех элементов прародителя $i = 1, \ldots, n$.*

Теорема 2. *Задача (2) эквивалентна совокупности независимых задач для отдельных элементов профиля:*

$$\boldsymbol{\beta}_i^{s+1} = \underset{\beta_{i1}, \ldots, \beta_{i20}}{\arg\max} \sum_{l=1}^{20} h_i(\bar{\boldsymbol{\beta}}^s, \bar{\boldsymbol{\omega}}) \ln \sum_{k=1}^{20} \beta_{ik} \psi\big(\alpha^{(l)} \mid \alpha^{(k)}\big),$$

где

$$h_i(\bar{\boldsymbol{\beta}}^s, \bar{\boldsymbol{\omega}}) = \sum_{j=1}^{M} \sum_{\substack{t=1 \\ \omega_{jt} = \alpha^{(l)}}}^{N_j} p^{it}\big(\bar{\boldsymbol{\beta}}^s, \boldsymbol{\omega}_j\big)$$

и $p^{it}(\bar{\boldsymbol{\beta}}^s, \boldsymbol{\omega}_j)$ — апостериорная относительно $\bar{\boldsymbol{\beta}}^s$ вероятность того, что в случайном выравнивании j-го белка i-й элемент последовательности-прародителя будет связан с t-й позицией белка.

Решение задачи (2) очевидно. Компоненты элемента профиля $\bar{\boldsymbol{\beta}}_i^{s+1} = (\beta_{i1}^{s+1}, \ldots, \beta_{i,20}^{s+1})$ являются решением системы линейных алгебраических уравнений с матрицей условных вероятностей эволюционного чередований аминокислот:

$$\sum_{k=1}^{20} \psi\big(\alpha^{(l)} \mid \alpha^{(k)}\big)\, \beta_{ik}^{s+1} = u_{il}, \quad l = 1, \ldots, 20,$$

$$u_{il} = \frac{h_i(\bar{\boldsymbol{\beta}}^s, \boldsymbol{\omega}_j)}{\sum\limits_{r=1}^{20} h_r\big(\bar{\boldsymbol{\beta}}^s, \boldsymbol{\omega}_j\big)}.$$

Вероятностная модель данных

Экспериментальное исследование предложенной схемы было проведено на модельных данных, сгенерированных следующим образом. В каждом эксперименте изначально задавалась длина последовательности-прародителя. Затем в соответствии с равномерным распределением генерировалась сама последовательность. Анализируемые последовательности генерировались на основе исходной по следующему правилу. Последовательность-прародитель случайным образом разбивалась на две подпоследовательности. Между ними и по краям добавлялись фрагменты последовательностей случайной длины, элементы которых также генерировались в соответствии с равномерным распределением.

Результаты экспериментов

Параметры генерации модельных данных: количество анализируемых последовательностей — 20, длина последовательностипрародителя — 20, количество частей, на которые делился домен — 2.

Для такой модели данных точность восстановления последовательности-прародителя составила в среднем 90%. Однако следует отметить, что эта точность зависит от числа анализируемых последовательностей, по которым ищется общий прародитель, от длины этих последовательностей и от длины «сорных» фрагментов. Исследования показали, что алгоритм более стабильно выделяет общего прародителя, если анализируемых последовательностей достаточно много, либо длина последовательности-прародителя не слишком мала по отношению к длинам анализируемых последовательностей.

Выводы

В данной работе поставлена и решена задача множественного выравнивания группы последовательностей как задача поиска их общего прародителя фиксированной длины в виде вероятностного профиля (упорядоченной совокупности независимых дискретных распределений на множестве аминокислот). Предложенный подход выгодно отличается от существующих алгоритмов множественного

выравнивания, в силу того, что он основан на эволюционной модели преобразования последовательностей, а также тем, что он принимает во внимание все возможные пути эволюции, а не один, что представляется более естественным. Однако в связи с этим решение поставленной задачи не дает само по себе множественного выравнивания в привычном смысле в виде совокупности записанных друг под другом выровненных последовательностей. Фактически, предлагаемый подход позволяет выделить консервативные фрагменты анализируемых последовательностей в виде упорядоченной совокупности из n элементов в каждой из них, никак не определяя при этом выравнивание остальных элементов анализируемых последовательностей. Тем не менее, традиционное визуальное представление множественного выравнивания можно получить, например, путем обычного парного выравнивания найденного вероятностного профиля с каждой из анализируемых последовательностей.

В рамках данной работы был проведен ряд экспериментов на модельных данных, результаты которых показали достаточно высокую эффективность работы предложенной процедуры. В дальнейшем планируется провести серию экспериментов на реальных данных, а также усовершенствовать данный алгоритм, добавив процедуру автоматического выбора наиболее адекватной длины общего прародителя. Также на основании данной работы планируется создать алгоритм автоматической классификации белков.

Литература

[1] *Гельфанд М. С.* Апология биоинформатики // Биофизика, 2005. — Т. 50, № 4. — С. 752–766.

[2] *Шлезингер М. И.* О самопроизвольном различении образов // Сб. Читающие автоматы. — Киев: Наукова думка, 1965.

[3] *Dayhoff M. O., Schwartz R. M., Orcutt B. C.* A model for evolutionary change in proteins // Atlas for Protein Sequence and Structure (M. O. Dayhoff, ed.). — 1978, Vol. 5, Pp. 345–352.

[4] *Edgar R. C., Batzoglou S.* Multiple sequence alignment // Current Opinion in Structural Biology 2006. — No. 16. — Pp. 368–373.

[5] *Mottl V. V., Dvoenko S. D., Seredin O. S., Kulikowski C. A., Muchnik I. B* Alignment scores in a regularized support vector classification method for fold recognition of remote protein families // DIMACS Technical Report 2001-01, Center for Discrete Mathematics and Theoretical Computer Science, Rutgers University, the State University of New Jersey, USA, January 2001, 33 p.

[6] *Notredame C.* Recent evolutions of multiple sequence alignment algorithms PLoS Comput Biol 3(8): e123. doi:10.1371/journal.pcbi.0030123.

[7] *Pei J. Kim B. H., Grishin N. V.* PROMALS3D: a tool for multiple protein sequence and structure alignments Nucleic Acids Research, 2008. — Vol. 36, No. 7 — Pp. 2295–2300.

[8] *Smith T. F., Waterman M. S.* Identification of common molecular subsequences. // Journal of Molecular Biology, 1981. — Vol. 147, No. 1. — Pp. 195–197

[9] *Wallance I. M., Blackshields G., Higgins D. G.* Multiple sequence alignments // Current opinion in structural biology, 15, 2005, Pp. 261–266.

О разрешимости формальной задачи распознавания вторичной структуры белка*

Рудаков К. В., Торшин И. Ю.
Москва, ВЦ РАН им. А. А. Дородницына
Москва, РСЦ Института микроэлементов ЮНЕСКО

Настоящий доклад посвящен рассмотрению вопроса о возможности применения конструкций алгебраического подхода к проблеме распознавания вторичной структуры белка. Анализ экспериментальных данных выявил необходимость разработки исходного формализма для корректной постановки изучаемой проблемы. Разработанный формализм позволил (1) сформулировать корректное описание принятой у биологов гипотезы о локальном характере зависимости вторичной структуры от первичной и (2) получить конструктивный критерий разрешимости задачи распознавания вторичной структуры, названный нами «условием $(L(M), R(M))$-корректности».

Введение

Распознавание вторичной структуры белка на основе его первичной структуры (аминокислотной последовательности) — одна из важнейших задач вычислительной биологии [1]. Вкратце, задача может быть описана как перевод последовательности символов из одного алфавита в другой:

Последовательность символов в алфавите 1
(«первичная структура»)

```
··· V L S P A D K T N V ···
    ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓
··· L L L H H H H H H H ···
```

Последовательность символов в алфавите 2
(«вторичная структура»)

Изучение общедоступных экспериментальных данных по первичной и вторичной структуре (основанных на результатах рентгеноструктурного анализа более 50 000 белков [2]) и существующих методов предсказания вторичной структуры [3] выявило необходимость разработки специализированного формализма для корректной постановки изучаемой проблемы. Предлагаемый формализм позволил получить конструктивный критерий разрешимости исследуемой задачи.

Определения исходных множеств

Пусть заданы два алфавита, A и B. Алфавит A соответствует множеству 20 типов аминокислот, образующих белки; алфавит B — множеству типов вторичной структуры. В общем случае $A = \{a_1, \ldots, a_n\},\ n > 0,\ B = \{b_1, \ldots, b_m\},\ m > 0$. В случае задачи предсказания вторичной структуры, $A = \{A, C, D, E, F, G, H, I, K, L, M, N, P, R, S, T, V, W, Y\}$ и $B = \{H, S, L\}$. Обозначим множества слов длины k в каждом из алфавитов A^k и B^k. Тогда множество всех исходных слов $A^* = \bigcup\limits_{l=1}^{\infty} A^l$, а множество всех слов в алфавите B есть $B^* = \bigcup\limits_{l=1}^{\infty} B^l$.

Решение задачи распознавания вторичной структуры белка сводится к поиску некоторой функции $F\colon A^* \to B^*$, причем $|F(\vec{a})| = |\vec{a}|$ ($|\vec{a}|$ — длина слова $\vec{a}$). Пусть $\mathfrak{F} = \{F\colon A^* \to B^*;\ |F(\vec{a})| = |\vec{a}|\}$ — множество функций этого типа.

Пусть Δ — неопределенность. Введем расширенные алфавиты $\widetilde{A} = A \cup \{\Delta\}$ и $\widetilde{B} = B \cup \{\Delta\}$ и расширенные множества слов $\widetilde{A}^* = \bigcup\limits_{l=1}^{\infty} \widetilde{A}^l$ и $\widetilde{B}^* = \bigcup\limits_{l=1}^{\infty} \widetilde{B}^l$ соответственно.

Пусть задано множество прецедентов $\mathrm{Pr} \subseteq \widetilde{A}^* \times \widetilde{B}^*,\ \mathrm{Pr} \neq \varnothing$, где «$\times$» обозначает декартово произведение. Прецедент, таким образом, представляет собой пару слов $(\vec{a}, \vec{b}) \in \mathrm{Pr},\ |\vec{a}| = |\vec{b}|$. Назовем $V = \vec{a}$ «верхним словом», а $W = \vec{b}$ — «нижним словом» прецедента.

Функция F корректна, если $\underset{\mathrm{Pr}}{\forall}(\vec{a}, \vec{b})\colon F(\vec{a}) = \vec{b}$. Очевидно, что F существует тогда и только тогда, когда $\underset{\mathrm{Pr}}{\forall}(\vec{a}_1, \vec{b}_1)(\vec{a}_2, \vec{b}_2)\colon (\vec{a}_1 = \vec{a}_2) \Rightarrow (\vec{b}_1 = \vec{b}_2)$.

Маски и оператор выбора подслова (окрестности)

Пусть дано слово $\vec{U} = \{u_1, \ldots, u_n\}$ длины n. Это может быть верхнее слово (V) или нижнее слово (W). Определим некую *ведущую позицию* i, $1 \leqslant i \leqslant n$. Дана также «маска» $\hat{m} = \{\mu_1, \ldots, \mu_m\}$, где $\mu_i \in \mathbb{Z},\ \mu_1 < \ldots < \mu_m$. Параметр m назовем *размерностью* маски $\hat{m}$, а параметр $(\mu_m - \mu_1 + 1)$ — *протяженностью маски*. Определим *оператор выбора окрестности* или *оператор выбора подслова* как $\eta(i, \hat{m}, \vec{U}) = u_{i+\mu_1} \cdots u_{i+\mu_m}$, если $1 \leqslant i + \mu_1$ и $i + \mu_m \leqslant n$; в противном случае $\eta(i, \hat{m}, \vec{U}) = \varnothing$. Иначе говоря, оператор η выбирает определенную подпоследовательность слова $\vec{U}$ по маске $\hat{m}$, помещенной в позицию i.

(L,R)-корректность F

Пусть $L, R \in \mathbb{N} \cup \{0\}$.

Определение 1. *Функцию (алгоритм) F назовем (L, R)-корректной, если для всех* $(\vec{a}, \vec{b}) \in \mathrm{Pr}$

*Работа выполнена при финансовой поддержке гранта РФФИ, проект № 09-07-12098 офи_м.

выполнено $F(\vec{a}) = \vec{b}'$, *где* $b'_1 = \ldots = b'_L = b'_{|\vec{a}|-R+1} = \ldots = b'_{|\vec{a}|} = \Delta$ *и* $b'_i = b_i$ *при* $L < i \leqslant |\vec{a}| - R$.

Определение L и R на наборе масок

Ниже считаем, что дан набор масок $M = \{\hat{m}_1, \ldots, \hat{m}_N\}$. Определим числовые параметры $L(M)$ и $R(M)$. Будем считать, что $\hat{m}_1 = (\mu^1_1, \ldots, \mu^1_{|m_1|})$, $\hat{m}_N = (\mu^N_1, \ldots, \mu^N_{|m_N|})$. Задача заключается в нахождении такого минимального i (значения ведущей позиции), при котором применима по крайней мере одна маска из набора M. Иначе говоря, $L(M) + 1 = \min(i)\colon \overset{N}{\underset{k=1}{\exists}} (i + \mu^k_1 = 1)$. Тогда

$$L(M) = \max\Big(-\max_{k=1,N} \mu^k_1, 0\Big).$$

Аналогично, $R(M)$ определяется как

$$R(M) = \max\Big(\min_{k=1,N} \mu^k_{|m_k|}, 0\Big).$$

Условие локальности

Решение задачи о распознавании вторичной структуры будет искаться в форме локальной функции. Определим класс локальных функций $\mathfrak{f}(M) \subseteq \mathfrak{F}$. Некая функция F принадлежит $\mathfrak{f}(M)$ тогда и только тогда, когда существует функция $f\colon (\widetilde{A}^*)^P \to \widetilde{B}^*$ при $P = \sum\limits_{k=1}^{N} m_k$ такая, что для всякого $\vec{a} = (a_1, \ldots, a_n)$ имеем $F(\vec{a}) = \vec{b} = (b_1, \ldots, b_n)$, где $b_1, \ldots, b_{L(M)} = b_{n-R(M)+1}, \ldots, b_n = \Delta$, а при $i = L(M)+1, \ldots, n-R(M)$

$$b_i = f\big(\eta(i, \hat{m}_1, \vec{a}), \ldots, \eta(i, \hat{m}_N, \vec{a})\big).$$

Условие существования локальных функций

При поиске решения $F \in \mathfrak{f}(M)$ естественно требовать $(L(M), R(M))$-корректности. Функция F существует (т. е. алгоритм $(L(M), R(M))$-корректен), если выполняется условие:

$$\underset{\Pr}{\forall}(\vec{V}_1, \vec{W}_1),\ (\vec{V}_2, \vec{W}_2),\ \underset{i,j \in N}{\forall}(i,j)\ \overset{|M|}{\underset{k=1}{\forall}}\colon$$
$$\eta(i, \hat{m}_k, \vec{V}_1) = \eta(j, \hat{m}_k, \vec{V}_2) \ \Rightarrow\ W^i_1 = W^j_2, \quad (1)$$

где $L(M) < i \leqslant |\vec{V}_1| - R(M),\ L(M) < j \leqslant |\vec{V}_2| - R(M)$.

Разрешимость задачи и монотонность условия разрешимости

При выполнении условия существования локальных функций (1), задача прогнозирования вторичной структуры *разрешима*, в противном случае — *неразрешима*. Очевидно, что разрешимость задачи зависит от набора прецедентов Pr и набора масок M как параметров разрешимости. Таким образом, если имеются Pr и M, можно определить наличие разрешимости.

Примем Pr = const и рассмотрим возможности варьирования M. Варьирование M заключается в добавлении или удалении отдельных масок. В общем случае, условие разрешимости (1) не монотонно по M, т. е. из существования разрешимости на M не следует разрешимость на произвольном M' таком, что $M \subseteq M'$. Иначе говоря, может быть возможным нахождение маски $\hat{m} \notin M$ такой, что при включении этой маски в M изменятся значения $L(M)$ и $R(M)$, так что условие разрешимости (1) также нарушится вследствие нарушения требования $(L(M), R(M))$-корректности. Иначе говоря, в общем случае утверждение о монотонности условия разрешимости неверно.

Поэтому, целесообразно скорректировать утверждение о монотонности условия разрешимости (1) с условием неизменности значений $L(M)$ и $R(M)$:

Утверждение 1. *Если для* $\langle \Pr, M \rangle$ *есть разрешимость,* $M \subseteq M'$, $L(M) = L(M')$ *и* $R(M) = R(M')$, *то разрешимость есть и для* $\langle \Pr, M' \rangle$.

Замечание 1. Если $L(M') < L(M)$ или $R(M') > R(M)$, то задача может быть неразрешима для $\langle \Pr, M' \rangle$.

Выводы

Существующие методы распознавания вторичной структуры белка имеют в своей основе ряд неверифицируемых предположений касательно «подобия последовательностей», «эволюции белков», «роли контекста» последовательности и т. д. В настоящей работе был сформулирован строгий математический формализм, основанный на предварительно проведенном нами анализе экспериментальных данных и свободный от произвольных предположений псевдо-биологического характера. Предложенная нами система терминов позволяет осмыслить задачу о распознавании вторичной структуры белка с точки зрения алгебраического подхода к проблемам распознавания.

Литература

[1] *Torshin I. Y.* Bioinformatics in the Post-Genomic Era: The Role of Biophysics // 2006 Nova Biomedical Books, NY, ISBN: 1-60021-048

[2] *Berman H. M., Henrick K.* Announcing the worldwide Protein Data Bank // Nature Structural Biology — 2003. — Vol. 10, № 12. — 980 p.

[3] *Simossis V. F., Herringa J.* Integrating protein secondary structure prediction and multiple sequence alignment // Curr Protein Pept Sci. — 2004. Vol. 5. — Pp. 249–266.

Потенциальные функции на множестве аминокислот на основе модели эволюции М. Дэйхофф

Сулимова В. В., Моттль В. В., Куликовский К. А., Мучник И. Б.
vsulimova@yandex.ru, vmottl@yandex.ru, kulikows@cs.rutgers.edu, muchnikilya@yahoo.com
Тула, Тульский Государственный Университет;
Москва, ВЦ РАН, МФТИ;
USA, New Jersey, New Brunswick, Rutgers University

Решение задач анализа белков, представленных аминокислотными последовательностями, неизбежно должно основываться на сравнении аминокислот. При этом важно, чтобы используемая мера сходства, выражаемая в виде подстановочной матрицы, имела эволюционный смысл и обладала свойствами потенциальной функции. В данной работе доказывается, что: 1) модель эволюции PAM (Point Accepted Mutation), разработанная М. Дэйхофф, содержит в себе всё необходимое для формирования потенциальных функций на множестве аминокислот, и построенные на ее основе подстановочные матрицы теряют свойства потенциальных функций только вследствие неподходящего окончательного представления; 2) другой статистический подход к измерению сходства аминокислот BLOSUM (Block Summation Matrices), предложенный С. и Дж. Хениковыми, может быть выражен в терминах модели эволюции М. Дэйхофф, и соответствующие подстановочные матрицы изначально также являются потенциальными функциями.

Введение

Одной из фундаментальных проблем современной биоинформатики является проблема измерения сходства аминокислотных последовательностей, составляющих полимерные молекулы белков. Ее решение неизбежно должно основываться на сравнении аминокислот.

Парное сходство 20-ти существующих аминокислот принято выражать в виде квадратной матрицы 20×20, называемой подстановочной матрицей. В данной работе рассматриваются две общепринятые меры сходства. Первая из них представлена семейством подстановочных матриц PAM [1], предложенных Маргарет Дэйхофф и основанных на модели эволюции аминокислот. Параметры этой модели оцениваются по эмпирическим данным, полученным из филогенетических деревьев над семействами близких белков. Вторая из рассматриваемых мер сходства, получившая распространение под названием подстановочных матриц BLOSUM [2], согласно ее авторам Стивену и Джорджии Хениковым, принципиально отличается от матриц PAM и основана на чисто статистическом подходе, исходными данными для которого являются блоки консервативных регионов аминокислотных последовательностей, полученные в результате множественного выравнивания семейств существенно более далеких белков.

Согласно многочисленным публикациям, последняя мера сходства является более адекватной для сравнения далеких белков, однако обладает по сравнению с PAM существенным недостатком — она не имеет эволюционного обоснования [3]. В связи с этим, важным результатом данной работы является доказательство того, что мера сходства BLOSUM может быть выражена в терминах модели эволюции аминокислот PAM и отличие между соответствующими подстановочными матрицами заключается только в разных исходных данных, по которым оцениваются параметры одной и той же модели.

Вторым аспектом, исследуемым в данной работе являются условия, при которых постановочные матрицы, построенные на основе модели эволюции PAM, являются положительно определенными. В этом случае соответствующая мера сходства называется потенциальной функцией [4]. Она погружает множество аминокислот в некоторое гипотетическое линейное пространство, в котором играет роль скалярного произведения [5]. Важность построения потенциальных функций на множестве аминокислот обусловлена тем, что они являются естественной основой для построения потенциальных функций на множестве аминокислотных последовательностей [6, 7, 8, 9], которые, в свою очередь, позволяют применять для решения задач классификации белков классические методы анализа данных, разработанные для линейных пространств, в частности, такой удобный и эффективный инструмент, как метод опорных векторов [10].

Меры сходства семейств PAM и BLOSUM в их традиционном форме представления не являются потенциальными функциями. В литературе известны многочисленные попытки скорректировать эти подстановочные матрицы с целью сделать их неотрицательно определенными. Однако при этом либо теряется биологический смысл меры сходства [6, 7], либо нет гарантии отсутствия отрицательных собственных чисел [8].

В данной работе доказывается, что модель эволюции М. Дэйхофф содержит в себе всё необходимое для построения потенциальных функций. Более того, подстановочные матрицы семейств PAM и BLOSUM являются потенциальными функциями по своей природе и теряют соответствующие

свойства исключительно вследствие неподходящего окончательного представления.

Также следует отметить, что любая новая подстановочная матрица, построенная на основе данной модели эволюции по технике Дэйхофф или Хениковых, также будет обладать свойствами потенциальной функции. Этот факт имеет особенно большое значение в связи с ростом интереса к построению специфических подстановочных матриц, предназначенных для решения конкретных задач [11]. Такие матрицы, в отличие от классических, строятся только по белкам определенной группы организмов или функций, и, как правило, позволяют получить лучшие результаты при их дальнейшем анализе.

Модель эволюции аминокислот PAM

Вероятностная модель эволюции Маргарет Дэйхофф, получившая название PAM (Point Accepted Mutation), играет роль основной теоретической концепции сравнения аминокислот, а затем и белков, по их сходству. Центральная гипотеза этой модели заключается в том, что эволюционные изменения в аминокислотной последовательности складываются из случайных независимых изменений (mutation) отдельных аминокислот цепи (point), причем таких изменений, которые закрепились в ходе дальнейшего естественного отбора (accepted).

Основным математическим понятием модели Дэйхофф является понятие марковской цепи истории эволюции аминокислоты в отдельно взятой позиции последовательности h_t, $t = 1, 2, \ldots$, при некоторой достаточно сложной интерпретации понятия «величины шага» $(\ldots, t-1, t, t+1, \ldots)$, которое здесь не рассматривается.

Пусть $A = \{\alpha^i\}_{i=1}^{20}$ — множество аминокислот. Эволюционная модель PAM предполагает, что склонность аминокислот к взаимному мутационному превращению количественно выражается квадратной матрицей условных вероятностей

$$\Psi = \big(\psi(\alpha^j \,|\, \alpha^i)\big)_{i=1\,j=1}^{20\quad 20},$$
$$\psi(\alpha^j \,|\, \alpha^i) = P\big(h_t{=}\alpha^j \,|\, h_{t-1}{=}\alpha^i\big),$$

интерпретируемых как вероятность того, что аминокислота α^i на очередном шаге эволюции превратится в аминокислоту α^j.

При этом предполагается, что данная марковская цепь является:

1) эргодической, т. е. существует финальное распределение вероятностей:

$$\xi(\alpha^j) = \sum_{\alpha^i \in A} \xi(\alpha^i)\psi(\alpha^j \,|\, \alpha^i), \quad j = 1, \ldots, 20.$$

Это означает, что эволюция началась уже очень давно, случайный процесс мутаций успел установиться и не зависит от неизвестных начальных вероятностей состояний;

2) обратимой, т. е. выполняется условие

$$\xi(\alpha^i)\psi(\alpha^j \,|\, \alpha^i) = \xi(\alpha^j)\psi(\alpha^i \,|\, \alpha^j).$$

Это означает, что эволюция не имеет направления и принципиально невозможно определить, какая из двух аминокислот является потомком, а какая прародителем.

Потенциальные функции на основе модели эволюции М. Дэйхофф

В основе построения семейства потенциальных функций на множестве аминокислот лежит факт, что эргодический и обратимый марковский процесс эволюции аминокислот, наблюдаемый с любым шагом s и определяемый матрицей переходных вероятностей $\Psi_{[s]} = \underbrace{\Psi \times \ldots \times \Psi}_{s}$ остается эргодическим и обратимым марковским процессом с тем же финальным распределением вероятностей $\xi(\alpha^i)$, $i = 1, \ldots, 20$.

Сходство двух аминокислот $\alpha^i, \alpha^j \in A$ естественно оценивать значением правдоподобия гипотезы об их происхождении в результате двух независимых преобразований одного и того же неизвестного случайно выбранного объекта $\alpha^k \in A$, играющего роль общего прототипа:

$$\mu_{[2s]}(\alpha^i, \alpha^j) = \sum_{\alpha^k \in A} \xi(\alpha^k)\, \psi_{[s]}(\alpha^i \,|\, \alpha^k)\, \psi_{[s]}(\alpha^j \,|\, \alpha^k). \quad (1)$$

Двухместная функция правдоподобия (1) по своей структуре есть скалярное произведение двух векторов и, следовательно, является потенциальной функцией на множестве аминокислот для любого шага s.

Однако существенно более удобным является ее эквивалентное представление, возможное благодаря свойствам эргодичности и обратимости марковского процесса:

$$\mu_{[2s]}(\alpha^i, \alpha^j) = \sum_{\alpha^k \in A} \underbrace{\xi(\alpha^k)\, \psi_{[s]}(\alpha^i \,|\, \alpha^k)}_{\xi^i \psi_{[s]}(\alpha^k \,|\, \alpha^i)} \psi_{[s]}(\alpha^j \,|\, \alpha^k) =$$
$$= \xi(\alpha^i) \underbrace{\sum_{\alpha^k \in A} \psi_{[s]}(\alpha^j \,|\, \alpha^k)\, \psi_{[s]}(\alpha^k \,|\, \alpha^i)}_{\psi_{[2s]}(\alpha^j \,|\, \alpha^i)} =$$
$$= \xi(\alpha^i)\, \psi_{[2s]}(\alpha^j \,|\, \alpha^i). \quad (2)$$

Кроме того, часто полезным оказывается применение потенциальной функции, нормированной

на финальные вероятности:

$$\bar{\mu}_{[2s]}(\alpha^i,\alpha^j)=\frac{\mu_{[2s]}(\alpha^i,\alpha^j)}{\xi(\alpha^j)\xi(\alpha^i)}= \\ =\frac{\psi_{[s]}(\alpha^j\,|\,\alpha^i)}{\xi(\alpha^j)}=\frac{\psi_{[s]}(\alpha^i,\alpha^j)}{\xi(\alpha^i)}. \quad (3)$$

С учетом того, что функция $\mu_{[2s]}(\alpha^i,\alpha^j)$ изначально имеет вид (1), очевидно, что при нормировании на финальные вероятности она остается потенциальной функцией.

Положительно определенные матрицы PAM

Исходными данными для оценивания параметров модели PAM являются фрагменты реализации марковской цепи эволюции аминокислот, полученные из филогенетических деревьев над близкими аминокислотными последовательностями (не менее 95% похожести). По этим данным в классе обратимых марковских цепей находятся оценки вектора финальных вероятностей $\hat{\boldsymbol{\xi}}=\big(\hat{\xi}(\alpha^i)\big)_{i=1}^{20}$ и одношаговой матрицы переходных вероятностей $\hat{\Psi}_{[1]}=\big(\hat{\psi}_{[1]}(\alpha^j|\alpha^i)\big)_{i=1}^{20}{}_{j=1}^{20}$, соответствующей эволюционному шагу 1 PAM, для которого из ста случайно выбранных аминокислот изменится только одна: $\sum\limits_{i=1}^{20}\hat{\xi}(\alpha^i)\big(1-\hat{\psi}_{[1]}(\alpha^i,\alpha^i)\big)=0{,}01$.

По этим оценкам для любой s-шаговой матрицы переходных вероятностей $\hat{\Psi}_{[s]}$ вычисляется мера сходства $\pi_{[s]}(\alpha^j,\alpha^i)=\hat{\psi}_{[s]}(\alpha^j|\alpha^i)/\hat{\xi}(\alpha^j)$, которая является потенциальной функцией, поскольку по своей структуре совпадает с (3). Однако традиционно семейство подстановочных матриц представляется в логарифмической форме с последующим округлением до целых $d_{[s]}(\alpha^i,\alpha^j)=\big\lfloor 10\,\log_{10}\pi_{[s]}(\alpha^i,\alpha^j)\big\rfloor$. Свойства потенциальной функции при этом, естественно, теряются.

Семейство подстановочных матриц BLOSUM

Исходными данными для построения подстановочных матриц BLOSUM являются совокупность консервативных столбцов множественных выравниваний последовательностей существенно более далеких, по сравнению с PAM, белков. На основе этих данных Хениковы вычисляют следующие статистики:

1) наблюдаемая вероятность совместного появления аминокислот α^i и α^j в одном столбце множественного выравнивания:

$$p^{ij}=M^{ij}/M, \quad (4)$$

где M^{ij} — частоты встречаемости неупорядоченных пар (α^i,α^j) в сумме по всем столбцам и M — общее число неупорядоченных пар;

2) ожидаемая вероятность появления аминокислоты α^i в неупорядоченной паре (α^i,α^j)

$$q^i=p^{ii}+\frac{1}{2}\sum_{i\neq j}p^{ij}. \quad (5)$$

На основе данных статистик подстановочные матрицы BLOSUM $\boldsymbol{B}=\big(b^{ij},\ i,j=1,\ldots,20\big)$ вычисляются следующим образом:

$$b^{ij}=\begin{cases}\big\lfloor 2\,\log_2\big(p^{ij}/q^iq^j\big)\big\rfloor, & \text{для } i=j;\\ \big\lfloor 2\,\log_2\big(p^{ij}/2q^iq^j\big)\big\rfloor, & \text{для } i\neq j.\end{cases} \quad (6)$$

Различные матрицы семейства BLOSUM отличаются друг от друга степенью сходства аминокислотных последовательностей, по множественным выравниваниям которых вычисляются статистики.

Полученные таким образом подстановочные матрицы имеют отрицательные собственные числа и, согласно их авторам, а также другим публикациям, не основаны на какой-либо модели эволюции. Однако в данной работе показано, что при определенной интерпретации происхождения исходных данных, по которым вычисляются статистики (4) и (5), мера сходства BLOSUM имеет тот же вероятностный смысл, что и мера сходства PAM.

Модель эволюции М.Дэйхофф для BLOSUM

Предположим, что процесс эволюции аминокислот $A=\{\alpha^i\}_{i=1}^{20}$ может быть описан эргодическим и обратимым марковским процессом с матрицей переходных вероятностей $\boldsymbol{\Psi}_{[1]}=\big(\psi_{[1]}(\alpha^j\,|\,\alpha^i)\big)_{i=1}^{20}{}_{j=1}^{20}$ и вектором финальных вероятностей $\boldsymbol{\xi}=(\xi^i)_{i=1}^{20}$.

Будем также полагать, что все столбцы множественного выравнивания, являющиеся исходными данными для построения подстановочных матриц BLOSUM, были порождены независимо друг от друга по следующей схеме. Сначала для каждого столбца, в соответствии с финальным распределением на множестве аминокислот $\xi(\vartheta)$, была выбрана некоторая неизвестная аминокислота-прародитель $\vartheta\in A$. Затем из этой аминокислоты в результате независимых опытов были получены все аминокислоты данного столбца множественного выравнивания в соответствии со случайным преобразованием на множестве аминокислот $\psi_{[1]}(\alpha|\vartheta)$, $\alpha,\vartheta\in A$.

При такой модели порождения данных, статистики, вычисляемые Хениковыми, являются несмещенными оценками параметров модели Дэйхофф, о чем говорят следующие две теоремы, доказательства которых приведены в работе [12].

Теорема 1. *Статистика* (5) *является несмещенной оценкой* $\xi(\alpha^i)$.

В соответствии с (2), вероятность совместного появления пары аминокислот $\alpha^i, \alpha^j \in A$ из одного общего неизвестного прародителя $\vartheta \in A$ в результате двух независимых реализаций случайного преобразования $\psi_{[1]}(\alpha|\vartheta)$ будет иметь вид

$$\mu_{[2]}(\alpha^i, \alpha^j) = \xi(\alpha^i)\, \psi_{[2]}(\alpha^j \,|\, \alpha^i).$$

Теорема 2. *Статистика p^{ij} (4) для $i = j$ совместно с $\frac{1}{2}p^{ij}$ для $i \neq j$ является несмещенной оценкой вероятности $\mu_{[2]}(\alpha^i, \alpha^j)$.*

Положительно определенные матрицы BLOSUM

В соответствии с теоремами 1 и 2 каждый элемента матрицы BLOSUM (6) может быть выражен в терминах модели эволюции аминокислот PAM, в основе которой лежит понятие эргодического и обратимого марковского процесса с переходной матрицей $\mathbf{\Psi}_{[2]}$ и финальным распределением $\xi(\alpha^i)$, $i = 1, \ldots, 20$, следующим образом:

$$b^{ij} = \left\lfloor 2\log_2 \left(\frac{\mu_{[2]}(\alpha^i, \alpha^j)}{\xi(\alpha^i)\xi(\alpha^j)} \right) \right\rfloor = \left\lfloor 2\log_2 \left(\frac{\psi_{[2]}(\alpha^i|\alpha^j)}{\xi(\alpha^i)} \right) \right\rfloor = \left\lfloor 2\log_2 \left(\frac{\psi_{[2]}(\alpha^j|\alpha^i)}{\xi(\alpha^j)} \right) \right\rfloor. \quad (7)$$

Заметим, что выражение под знаком логарифма полностью совпадает по своей структуре с (3) и, следовательно, является потенциальной функцией. Таким образом, мера сходства BLOSUM, так же как и PAM, изначально является положительно определенной и теряет свойства потенциальной функции только из-за логарифмирования и последующего округления до целых.

Погружение множества аминокислот в линейное пространство

Обладая свойствами потенциальной функции на алфавите аминокислот $A = \{\alpha^1, \ldots, \alpha^{20}\}$, каждая из мер сходства $\mu_{[s]}(\alpha^i, \alpha^j)$ погружает его в гипотетическое двадцатимерное линейное пространство $\tilde{A}_s \supset A$. В качестве полного линейно независимого базиса удобно принять сам исходный алфавит, интерпретируемый как конечное подмножество точек в этом пространстве. Совокупность, вообще говоря, воображаемых элементов линейного пространства $\tilde{\alpha} = \sum\limits_{i=1}^{20} c^i \alpha^i \in \tilde{A}_s$ естественно интерпретировать как некоторые «обобщенные аминокислоты». В частности, если $c^i \geqslant 0$ для всех i и $\sum\limits_{i=1}^{20} c^i = 1$, то обобщенная аминокислота $\tilde{\alpha}$ имеет смысл вероятностной смеси реальных аминокислот с вектором вероятностей $(c^1, \ldots, c^m)$.

Выводы

В данной работе доказано, что две общепринятые меры сходства аминокислот, как PAM, так и BLOSUM, численно выражают правдоподобие гипотезы об общем происхождении двух указанных аминокислот от одной неизвестной аминокислоты в результате двух независимых шагов эволюции. Такая двухместная функция на алфавите аминокислот всегда является потенциальной функцией по своей структуре. Практически используемые подстановочные матрицы PAM и BLOSUM не являются положительно определенными только в силу логарифмического представления результата, к тому же округленного до целого значения.

Литература

[1] *Dayhoff M. O., Schwartz R. M., Orcutt B. C.* A model for evolutionary change in proteins // Atlas for Protein Sequence and Structure (M. O. Dayhoff, ed.). — 1978, Vol. 5, Pp. 345–352.

[2] *Henikoff S., Henikoff J.* Amino acid substitution matrices from protein blocks // Proc. Natl. Acad. Sci., 1992, 10915-10919.

[3] *Altschul S. F.* The Statistics of Sequence Similarity Scores // `http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html`.

[4] *Айзерман М. А., Браверман Э. М., Розоноэр Л. И.* Метод потенциальных функций в теории обучения машин // М.: Наука, 1970, 384 с.

[5] *Моттль В. В.* Метрические пространства, допускающие введение линейных операций и скалярного произведения // ДАН, 2003. — Т. 67, № 1.

[6] *Vanschoenwinkel B., Manderic B.* Substitution matrix based kernel functions for protein secondary structure prediction // Int. conf. on Machine Learning and Applications, 2004. — Pp. 388–396.

[7] *Wu F., Oslon B., Dobbs D., Honavar V.* Comparing kernels for predicting protein binding sites from amino acid sequence // Newral Netwarks, 2006, IJCNN'OC, pp. 1612-1616.

[8] *Vert J. P., Saigo H., Akutsu T.* Local alignment kernels for biological sequences // Kernel methods in computational biology, MIT Press, 2004, Pp. 131–154.

[9] *Vincent M., Passerini A., Labbe M., Frasconi P.* A simplified approach to disulfide connectivity prediction from protein sequences // BMC Bioinformatics, January 2008, 9:20 doi:10.1186/1471-2105-9-20.

[10] *Vapnik V.* Statistical Learning Theory // New York: John-Wiley and Sons, Inc.,1998, 732 p.

[11] *Ng P. C., Henikoff J. G., Henikoff S.* PHAT: a transmembrane-specific substitution matrix // Bioinformatics, 2000, 16, 760-766.

[12] *Sulimova V. V., Mottl V. V., Kulikowski C., Muchnik I.* Probabilistic evolutionary model for substitution matrices of PAM and BLOSUM families // DIMACS Tech. Rep. 2008-16, Rutgers University, 17 p. — 2008 — `ftp://dimacs.rutgers.edu/pub/dimacs/TechicalReports/TechReports/2008/2008-16.pdf`

Структурный анализ поведенческой динамики*

Темлянцев А. В., Ветров Д. П., Кропотов Д. А.
alexander.temlyantsev@gmail.com
Москва, ВМиК МГУ, Вычислительный Центр РАН

В данной работе построено решение задачи анализа поведения в рамках конкретного биологического эксперимента методами структурного распознавания. Предлагаемый подход формализует взаимосвязи между поведением и эмоциональной динамикой живой системы в виде *графической модели*, в рамках которой удается построить эффективные алгоритмы обучения и распознавания.

Развитие биологических представлений об организме как о живой системе явилось стимулом для возникновения целого ряда аналогий и применения методов биологической науки в смежных дисциплинах. Преимущество изучения целостного организма с особой силой выражено уже в очень ранних работах И. П. Павлова. Так, например, еще в конце прошлого столетия он выдвинул идею, что «наиболее нормальные функции организма можно изучать не у ограниченного в подвижности животного, т. е. в условиях вивисекции, а у целостного, ненаркотизированного животного» [1]. Одна лишь видеозапись активности лабораторного животного содержит в себе существенную информацию о динамике обуславливающих ее скрытых функциональных систем. Масштаб и относительная простота экспериментов, вкупе с их научной значимостью обуславливают возрастающий интерес к построению систем автоматического анализа результирующих данных.

В данной работе предложены алгоритмы распознавания эмоционального состояния животного по видеозаписи на базе композитно-иерархических скрытых марковских моделей, сочетающих в себе ряд фундаментальных биологических представлений:

— Динамика поведения животного может быть представлена последовательностью поведенческих актов. Идея *поведенческого акта* достаточно точно передается соответствующим ему глаголом: «идти», «бежать», «красться», выражающим действие вкупе с целью, и по достижении которой оно будет прекращено [1].

— Эмоциональное состояние есть набор независимых элементарных эмоций, влияющих на поведение особи композитно. Особи, находящейся в сложном эмоциональном состоянии, одновременно присущи свойства, характерные для каждой активной в данный момент элементарной эмоции.

Байесовские сети

Определение 1. *Пусть $G = \langle V, \Gamma\rangle$ — ориентированный граф с множеством вершин V и множеством ребер Γ. Будем говорить, что случайный вектор $(x_1 \dots x_n)$, $n = |V|$ удовлетворяет графической структуре G, если существует такое взаимнооднозначное соответствие $\{x_1, \dots, x_n\} \overset{f}{\leftrightarrow} V$, что*

$$p(x_1 \dots x_n) = \prod_{i=1}^{|V|} p\big(x_i \,|\, f^{-1}(\pi(f(x_i)))\big), \qquad (1)$$

где $\pi(v) = \big\{v_{prev} \colon (v_{prev}, v) \in \Gamma\big\}$ — множество родителей вершины $v \in V$ в графе G.

Говорят, что семейство распределений удовлетворяет G, если каждое распределение этого семейства удовлетворяет G

Язык графов оказывается удобным и гибким инструментом для отражения в вероятностной модели причинно-следственных связей, присущих описываемому объекту. Зафиксировав графическую структуру вероятностной модели, в достаточно общем случае удается предложить эффективные методы решения основных задач статистического анализа сразу для всех удовлетворяющих этой структуре семейств распределений. Однако наиболее завершенных прикладных результатов удается достичь, если ввести дополнительные аналитические ограничения, например, зафиксировав явно параметрическое семейство распределений для каждой вершины графа (например, для применения ряда критериев классической статистики необходимо потребовать нормальности элементов выборки):

Определение 2. *Байесовской сетью называется система $\mathfrak{A} = \langle G{=}\langle V, \Gamma\rangle, \mathcal{P}\rangle$ где G есть ориентированный граф,*

$$\mathcal{P} = \big\{(v, p) \colon v \in V,\ p = p(v \,|\, \pi(v), \theta_v)\big\}, \quad \theta_v \in \Theta_v,$$

причём каждой вершине $v \in V$ соответствует единственная пара $(v, p) \in \mathcal{P}$. Говорят, что случайный вектор $(v_1, \dots, v_n)$, $n = |V|$ удовлетворяет байесовской сети, если при некоторых $\theta_1, \dots, \theta_n$

$$p(v_1 \dots v_n) = \prod_{i=1}^{|V|} p\big(v_i \,|\, \pi(v_i), \theta_i\big). \qquad (2)$$

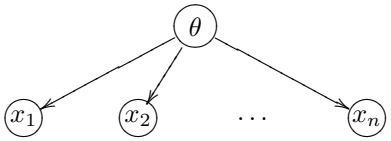
Представление (2) называется *факторизацией совместного распределения*. Отметим, что каждая

*Работа выполнена при финансовой поддержке РФФИ, проект №08-01-00405

вершина v_i входит в факторизацию, как минимум, один раз в множителе $p\big(v_i \,|\, \pi(v_i), \theta_i\big)$ При этом она может входить и в другие сомножители, являясь предком какой-либо другой вершины графа G. Будем далее говорить, что этот множитель соответствует вершине v_i.

Часто удобно заранее разбить множество вершин на наблюдаемые (observable) и скрытые (hidden), обозначая $V = \{v_1^{obs}, \ldots, v_r^{obs}, v_1^{hid}, \ldots, v_t^{hid}\}$.

Пример 1. В классической статистике под выборкой понимают совокупность независимых одинаково распределенных случайных величин. Такая ситуация описывается графической структурой



Здесь наблюдаемые вершины $x_1, \ldots, x_n$ соответствуют объектам выборки, а скрытая θ — неизвестному параметру распределения.

В статистическом анализе данных особенную значимость имеют следующие задачи.

Задача 1. Обучение: оценить значения параметров по известным значениям наблюдаемых.

Задача 2. Восстановление значений скрытых переменных:

$$p(v_1^{obs}, \ldots, v_r^{obs}, v_1^{hid}, \ldots, v_t^{hid} \,|\, \theta) \to \max_{v_1^{hid} \ldots v_t^{hid}}.$$

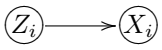
Следует упомянуть и задачу выбора наиболее правдоподобной структуры графа (задача выбора графической модели), однако в существующих сегодня постановках в этом направлении еще не получено значительных результатов; предложенные методы ее решения носят, как правило, характер перебора.

Построение графической модели поведенческой динамики

Коротко проследим ряд основных этапов на пути к цели исследования — построению эмоциональной динамики животного по видеозаписи его активности. Каждая вводимая здесь случайная переменная войдет как вершина в итоговую графическую модель. Таким образом, можно говорить об итоговом совместном распределении.

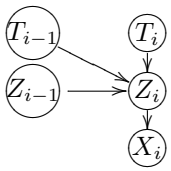
1. Устройство видеозаписи в клетке выдает последовательность кадров-изображений $I_1, \ldots, I_N$, на основании которой строится признаковое описание активности животного $X_1, \ldots, X_N$, представляющее собой набор механических характеристик его движения (производные координат центра масс и других характерных точек — носа, хвоста и т. п.).

2. Предполагается, что $X_1, \ldots, X_N$ представляют собой реализации случайных величин. Каждая случайная величина X_i зависит от активного в i-й момент времени поведенческого акта, номер которого в свою очередь полагается реализацией случайной величины Z_i:



Соответствующий X_i множитель итогового совместного распределения имеет вид $p(x_i|z_i) = \Phi\big(x_i, \mu(z_i), \Sigma(z_i)\big)$, где $\Phi(x, \mu, \Sigma)$ есть функционал плотности нормального закона с математическим ожиданием μ и матрицей ковариации Σ.

3. Z_i зависит от активного поведенческого акта в предыдущий момент, а также от локальной эмоциональной динамики (см. следующий пункт):



Ввиду того, что множества возможных поведенческих актов и эмоциональных состояний конечны (см. следующий пункт), соответствующий z_i сомножитель итогового распределения представляет собой функцию дискретных переменных

$$a^i_{\zeta|\eta,\sigma,\tau} = p\big(z_i = \zeta \,|\, z_{i-1} = \eta,\ T_i = \sigma,\ T_{i-1} = \tau\big).$$

Здесь и далее $\zeta, \eta \in \{1, \ldots, k\}$, σ, τ — булевы векторы размерности l, σ_i — булев вектор размерности l с единицей в i-й позиции и нулями в остальных.

4. Эмоциональное состояние в каждый момент времени моделируется бинарным вектором $T_i = (t_i^1, \ldots, t_i^l)$, где $t_i^k = 1$ если и только если k-я элементарная эмоция активна в i-й момент времени. Эти переменные в совокупности и определяют эмоциональную динамику — восстановление их значений по заданным значениям наблюдаемых решит исходную задачу. Предполагается, что t_i^k зависит от t_{i-1}^k, то есть соответствующий каждому t_i^k сомножитель итогового распределения есть

$$b^{i,k}_{p,q} = p\big(t_i^k = p \,|\, t_{i-1}^k = q\big), \quad p, q \in \{0, 1\}.$$

5. Предполагается, что продолжительность видеосъемки достаточно мала для того, чтобы считать условные распределения введенных величин не зависящими от времени, что мотивирует *ограничения гомогенности*: для любых $i, j \in \{2, \ldots, N\}$, $k \in \{1, \ldots, l\}$, $p, q \in \{0, 1\}$, $\zeta, \eta, \sigma, \tau$ имеет место

$$\begin{cases} a^i_{\zeta|\eta,\sigma,\tau} = a^j_{\zeta|\eta,\sigma,\tau} = a_{\zeta|\eta,\sigma,\tau}; \\ b^{i,k}_{p,q} = b^{j,k}_{p,q} = b^k_{p,q}; \\ \mu(z_i{=}\zeta) = \mu(z_j{=}\zeta); \\ \Sigma(z_i{=}\zeta) = \Sigma(z_j{=}\zeta). \end{cases} \tag{3}$$

6. Предполагается, что выполнена *гипотеза композитности*: если в данный момент времени, не являющийся моментом смены эмоционального состояния, несколько элементарных эмоций активны одновременно, то поведению животного в одинаковой степени присущи свойства каждой из них по отдельности. Это соображение можно моделировать следующим образом: если в данный момент времени активны k элементарных эмоций, то для каждого $i = 1, \ldots, k$ особь с вероятностью $1/k$ ведет себя как если бы активна была только одна i-я:

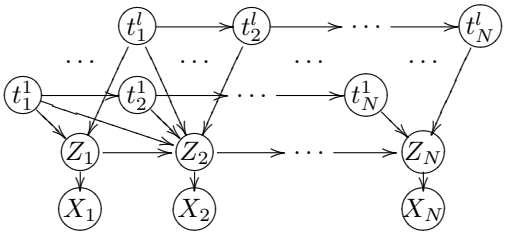
$$a_{\zeta\,|\,\eta,\sigma,\sigma} = \frac{1}{|N(\sigma)|} \sum_{i \in N(\sigma)} a_{\zeta|\eta,\sigma_i,\sigma_i}. \tag{4}$$

В момент смены эмоционального состояния поведение особи не зависит от предыстории:

$$a_{\zeta\,|\,\eta,\sigma,\tau} = a^0_{\zeta\,|\,\sigma}. \tag{5}$$

Подытожим сказанное в определении.

Определение 3. *Композитно-иерархической скрытой марковской моделью (КИСММ) эмоциональной динамики будем называть байесовскую сеть с графом*



дополненную ограничениями гомогенности (3) и композитности (4,5)

Отметим, что задача восстановления эмоциональной динамики по наблюдаемым автоматически решается с решением задачи 2 для КИСММ как байесовской сети, что на практике требует известных значений параметров модели a, b, μ, σ. Параметры, в свою очередь, могут быть восстановлены как решения для КИСММ задачи 1.

Задача определения значений скрытых переменных

Приведем формальную постановку задачи 2 для КИСММ:

Задача 3. $p(Z, T, X \,|\, \theta) \to \max\limits_{Z,T}$.

Учет графической структуры модели позволяет существенно упростить процедуру максимизации функционала плотности.

Задача 3 эквивалентна следующей

$$\log p(Z, T, X) \to \max_{Z,T}. \tag{6}$$

Согласно определению модели справедливо представление:

$$\begin{aligned}\log p(Z,T,X) &= \\ &= \log p(z_1 \,|\, T_1) + \log p(T_1) + \log p(x_1 \,|\, z_1) + \\ &\quad + \sum_{i=2}^{N} \Big(\log p(z_i \,|\, z_{i-1}, T_i, T_{i-1}) + \\ &\qquad + \log p(T_i \,|\, T_{i-1}) + \log p(x_i \,|\, z_i) \Big).\end{aligned}$$

Введем следующие обозначения:

$$\begin{aligned}&\mathfrak{C}(z_1, T_1) = \log p(T_1) + \log p(z_1|T_1) + \log p(x_1|z_1); \\ &\mathfrak{C}(z_n, T_n) = \max_{\substack{z_1,\ldots,z_{n-1} \\ T_1,\ldots,T_{n-1}}} \Big(\mathfrak{C}(z_1, T_1) + \\ &\qquad + \sum_{i=2}^{n} \log p(z_i, T_i, x_i \,|\, z_{i-1}, T_{i-1}) \Big).\end{aligned} \tag{7}$$

Опираясь только на дистрибутивность операции max относительно сложения и умножения, можно обосновать справедливость следующего утверждения:

Утверждение 1. *Для всех* $n = 1, \ldots, N$

$$\begin{aligned}\mathfrak{C}(z_n, T_n) = \max_{z_{n-1}, T_{n-1}} \Big(&\mathfrak{C}\big(z_{n-1}, T_{n-1}\big) + \\ &+ p\big(z_n, T_n, x_n \,|\, z_{n-1}, T_{n-1}\big) \Big).\end{aligned} \tag{8}$$

Утверждение 1 позволяет вычислять $\mathfrak{C}(z_n, T_n)$ итерационно, существенно упрощая процедуру максимизации.

Обозначим

$$\begin{aligned}\mathfrak{B}\big(z^* k, T^*_k\big) = \operatorname*{arg\,max}_{z_{n-1}, T_{n-1}} \Big(&\mathfrak{C}(z_{n-1}, T_{n-1}) + \\ &+ p\big(z_n, T_n, x_n \,|\, z_{n-1}, T_{n-1}\big) \Big).\end{aligned} \tag{9}$$

Теорема 2.

$$\max_{Z,T} p(Z, T, X) = \max_{z_N, T_N} \mathfrak{C}(z_N, T_N).$$

Пусть z^*, T^* *таковы, что*

$$(z^*_N, T^*_N) = \operatorname*{arg\,max}_{z_N, T_N} \mathfrak{C}(z_N, T_N); \tag{10}$$

$$(z^*_k, T^*_k) = \mathfrak{B}(z^* k + 1, T^*_{k+1}); \tag{11}$$

тогда

$$\max_{Z,T} \log p(Z, T, X) = p(z^*, T^*, X).$$

Последние два утверждения мотивируют эффективный алгоритм нахождения максимума в задаче 4. Алгоритм 1 представляет собой обобщение

Алгоритм 1. Восстановление значений скрытых переменных.

1: вычислить $\mathfrak{C}(z_1, T_1)$ по формуле (7);
2: **для** $i := 2, \ldots, N$
3: вычислить $\mathfrak{C}(z_k, T_k)$, $\mathfrak{B}(z_k, T_k)$ по формулам (8), (9); все требуемые для вычисления величины уже определены на предыдущих шагах;
4: $(z_N^*, T_N^*) := \arg\max_{z_N, T_N} \mathfrak{C}(z_N, T_N)$;
5: **для** $i := 1, \ldots, N-1$
6: $(z_{N-i}^*, T_{N-i}^*) := \mathfrak{B}(z_{N-i+1}^*, T_{N-i+1}^*)$;

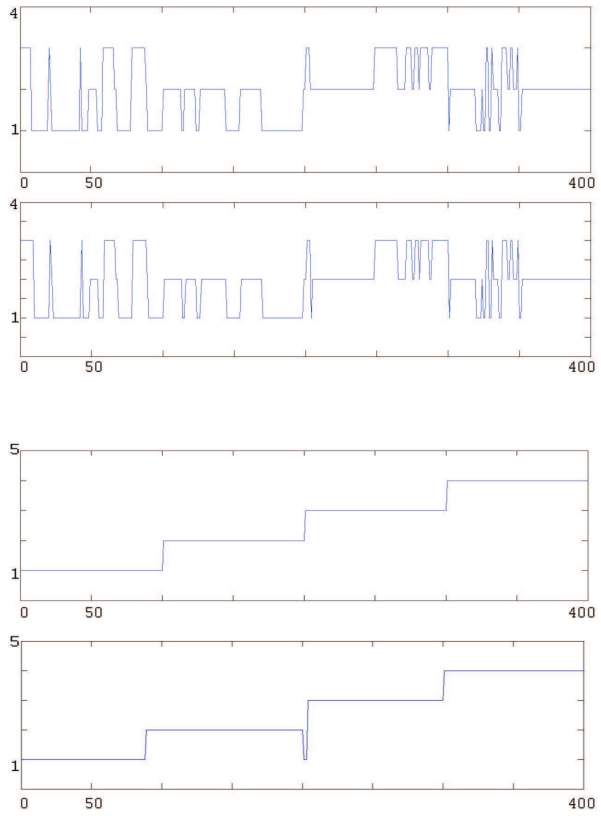


Рис. 1. Результат модельного тестирования Алгоритма 1. Сверху вниз: эмоциональная динамика, сгенерированные значения; эмоциональная динамика, ответ алгоритма; динамика поведенческого акта, сгенерированные значения; динамика поведенческого акта, ответ алгоритма.

известного алгоритма Витерби, решающего ту же самую задачу для скрытых марковских моделей.

Данный алгоритм был протестирован на модельных данных: по фиксированному θ генерировались значения скрытых и наблюдаемых переменных T_0, Z_0, X_0. Значения θ и наблюдаемых подавались на вход алгоритма. Результат работы последнего сравнивался с Z_0, T_0 (см. рис. 1.)

Задача обучения без учителя

Построение эффективного метода настройки параметров КИСММ, подаваемых далее вместе со значениями наблюдаемых на вход алгоритма 1, позволяет создать автономную программную систему поддержки биологического эксперимента. Естественно возникают два подхода к решению этой проблемы:

1. Получить соответствующие некоторому вектору наблюдаемых значения скрытых переменных из экспертной разметки и решить задачу $p(X, Z, T|\theta) \to \max_\theta$. Зафиксировав решение θ^*, применять Алгоритм 1 для произвольного вектора наблюдаемых. Теоретически, эксперту требуется один раз разметить видеозапись, чтобы далее Алгоритм 1 осуществлял разметку автоматически. Однако эксперт далеко не всегда способен провести такую разметку с достаточной для корректной работы алгоритма точностью.

2. Решать напрямую задачу обучения без учителя $p(X|\theta) \to \max_\theta$, что потребует явного вычисления $p(X \,|\, \theta) = \sum_{Z,T} p(X, Z, T \,|\, \theta)$ при неизвестном значении параметра θ, не допускающего использования преимуществ структурного метода.

Компромиссный подход состоит в построении итерационного процесса, способного по заданному θ_i построить новое приближение θ_{i+1} так, чтобы $p(X|\theta_{i+1}) > p(X|\theta_i)$. Выбирая в качестве начального значения θ^0, построенное по экспертной разметке или из иных эвристических соображений, можно далее автоматически найти более точное в смысле правдоподобия значение параметра.

Означенным свойством обладает широко используемый в приложениях EM-алгоритм [2]. Соответствующая итерационная последовательность $\{\theta_i\}_{i \in \mathbb{N} \cup \{0\}}$ строится по правилу

$$Q(\theta_i, \theta_{i-1}) \to \max_{\theta_i}, \; i \in \mathbb{N}$$

где $Q(\theta_i, \theta_{i-1}) = \sum_Z \sum_T \ln p(X, Z, T|\theta_i) p(Z, T|\theta_{i-1})$.

Учет структуры КИСММ позволяет существенно упростить вид функционала Q и, как следствие, реализацию полученного итерационного процесса.

Литература

[1] *Анохин П. К.* Очерки по физиологии функциональных систем // Москва, «Медицина», 1975. — Стр. 17-62

[2] *Bishop C. M.* Pattern Recognition and Machine Learning // Springer. 2006.

[3] *Elliott R. J., Aggoun L., Moore J. B.* Hidden Markov Models: Estimation and Control // Berlin: Springer-Verlag, 1995.

[4] *Ghahramani Z., Jordan M. I.* Factorial hidden Markov models // Machine Learning, 1997. — Vol. 29 — Pp. 245–275.

[5] *Fine S., Singer Y., Tishby N.* The Hierarchical Hidden Markov Model: Analysis and Applications // Machine Learning, 1998. — Vol. 32 — Pp. 41–62.

Пространственно-временная фильтрация данных магнитной энцефалографии*

Устинин М. Н., Панкратова Н. М., Ольшевец М. М.
ustinin@impb.psn.ru
Пущино, ИМПБ РАН

Предложен метод обнаружения полезного сигнала, без использования внешней информации о моментах его возникновения, на фоне спонтанной активности головного мозга.

Задача исследования

Современные научные исследования проводятся с помощью сложного оборудования и порождают большие объемы экспериментальных данных. В медицине также наблюдается тенденция к усложнению экспериментального оборудования. Задача неинвазивной энцефалографии состоит в том, чтобы узнать, как работает мозг по магнитному или электрическому полю, регистрируемому на поверхности головы. Это означает, что необходимо восстановить распределение в пространстве и во времени элементарных токовых диполей, порождающих внешние поля. В решении этой задачи большие надежды возлагаются на магнитную энцефалографию (МЭГ). Магнитные энцефалографы располагаются в магнитоизолируемых помещениях и строятся с использованием высокочувствительных физических приборов — СКВИДов, которые позволяют измерять слабое магнитное поле на поверхности головы с высокой точностью. Большое количество каналов (сотни) и высокая частота регистрации (сотни герц) дают возможность получать весьма подробную пространственно-временную картину поля, отражающую электрическую активность мозга с максимальной полнотой. Однако при анализе этой картины следует учитывать, что в каждый момент времени головной мозг решает множество задач одновременно. Это проявляется в сложной структуре суммарной электрической активности и приводит к необходимости выделения ее компонент, отвечающих разным задачам.

Таким образом, для получения новой информации о работе мозга необходимо применение эффективных методов обработки данных и постановка целенаправленных экспериментов. При этом предполагается, что избавление от внешних, по отношению к мозгу, магнитных полей обеспечивается, с одной стороны, условиями регистрации МЭГ (экранированием или конфигурацией датчиков), а с другой стороны, предобработкой МЭГ, устраняющей шумы от работы сердца, дыхания и т. п. При анализе МЭГ производится выделение полезного сигнала на фоне общей спонтанной активности мозга, а затем решается обратная задача: по магнитному полю определяется расположение электрических источников на магниторезонансной томограмме головного мозга испытуемого и делаются научные или диагностические выводы.

Под полезным сигналом понимается пространственно-временная последовательность магнитных полей $m(k,t)$, регистрируемых в результате изучаемой активности головного мозга, например, отклика на какой-либо внешний стимул. Здесь $m(k,t)$ — магнитное поле, измеренное k-тым датчиком, $k = 1,\ldots,K$, в момент времени $t = t_1,\ldots,t_n$, K — общее число датчиков (каналов регистрации), $(t_n - t_1)$ — длительность полезного сигнала, например, интервал между моментами подачи одинаковых стимулов.

Шумом (в целях данной работы) $\chi(k,t)$ считается вся активность мозга, не связанная с проявлением изучаемой нами в данном эксперименте. Считается, что сигнал и шум аддитивны, так что МЭГ за время длительности отклика записывается как $m(t,k) + \chi(k,t)$, при этом $m(k,t)$ повторяется без изменений в ответ на каждый отклик, а $\chi(k,t)$ — случайна. Многократно повторяя стимул и регистрируя отклик на него, можно «очистить» сигнал усреднением:

$$m(t_i,k) + \bar{\chi}(t_i,k) = \frac{1}{L}\sum_{l=1}^{L} m(t_{li},k) + \frac{1}{L}\sum_{l=1}^{L}\chi(t_{li},k),$$

где t_{li} — i-я точка по времени в отклике на l-ый стимул, t_{l1} — l-я опорная точка для усреднения, k — номер канала, $i = 1,\ldots,n$ — номер отсчета по t в полезном сигнале. Такая процедура позволяет улучшить отношение сигнал/шум в $\sqrt{L}$ раз, хотя и требует многократного повторения стимула ($L \sim 10^3 - 10^4$).

Методика

Типичной для магнитной энцефалографии является ситуация, когда искомый сигнал на порядок слабее спонтанной активности и лежит в той же полосе частот. Как правило, для выделения сигналов малой амплитуды используются либо внешние проявления патологической активности, например, запись миограммы при паркинсоническом треморе, либо запись стимула при экспериментах с вызванной активностью. По этим данным определяются

*Работа выполнена при финансовой поддержке РФФИ, проект № 07-07-00280, 07-01-00490, 09-07-12108.

опорные точки, по которым полезный сигнал выделяется усреднением. Этот подход доказал свою эффективность во многих экспериментах и широко используется в энцефалографии. Однако, остается актуальной задача выделения полезного сигнала при отсутствии дополнительной информации. Особенно остро эта проблема возникает при обработке данных, снятых у пациентов с какой-либо патологией. Например, при патологии tinnitus (слуховые галлюцинации) мы не располагаем внешними проявлениями болезни.

В настоящей работе предлагается взять в качестве опорных точек моменты обнаружения сигнала по его признакам на фоне спонтанной активности, что позволит затем очистить сигнал усреднением.

В качестве тестовой задачи использовались данные контрольного аудиторного эксперимента, полученные на 148-канальном измерительном стенде Magnes 2500 WH в больнице Бельвью, в Центре нейромагнетизма Медицинского факультета Нью-Йоркского университета [1]. Здоровому добровольцу подавался акустический стимул 7 раз в секунду. По моментам подачи стимула был выделен аудиторный отклик, а решение обратной задачи дало локализацию источников вызванной активности в слуховой зоне коры головного мозга. Таким образом была получена подробная информация о структуре магнитного поля, возникающего в ответ на аудиторный стимул. В качестве признака для обнаружения сигнала было выбрано пространственное распределение магнитного поля, соответствующее максимуму аудиторного отклика. Было выполнено разложение МЭГ по собственным функциям Карунена-Лоэва:

$$M(k,t) = \sum_{p=1}^{K} C_p(t) f_p(k),$$

где K — число каналов, $t = 1, \ldots, T$ — все время эксперимента.

Было найдено, что $f_3(k)$ соответствует по пространственной структуре аудиторному отклику, и в качестве признака был выбран $C_3(t)$. Затем была выполнена фильтрация $C_3(t)$ с помощью специально построенных линейных цифровых фильтров [4, 5].

Результаты

Были рассчитаны корреляционные функции между отфильтрованными коэффициентами разложения Карунена-Лоэва магнитной энцефалограммы и соответственными коэффициентами аудиторного отклика. Эти корреляционные функции были введены в среду обработки магнитных энцефалограмм MRIAN [6]. По максимумам корреляционных функций были найдены опорные точки и проведено усреднение по моментам распознавания. Решение обратной задачи показало совпадение координат источника с найденными по внешнему стимулу [7]. При этом очистка сигнала не является идеальной. Вместе с полезным сигналом были выделены шумовые компоненты, впрочем, разделенные во времени с хорошо локализуемой активностью.

По результатам работы можно сделать следующие выводы:

— предложенный подход перспективен для выделения полезных сигналов с известными свойствами из общей спонтанной активности;
— метод нуждается в развитии, прежде всего, в использовании дополнительных признаков для обнаружения сигнала.

Литература

[1] *Llinas R., Ribary U., Jeanmonod D., Kronberg E., Mitra P.* Thalamocortical dysrhythmia: A neurological and neuropsychiatric syndrome characterized by magnetoencephalography // Proc. of the National Academy of Sciences of the USA. 1999; 96: 15222.

[2] *Karhunen K.* Uber lineare Methoden in der Wahrscheinlichkeitsrechnung // Ann. Acad. Sci. Fennicae. Ser. A. I. Math.-Phys., — 1947. — № 37.— P. 1–79.

[3] *Loeve M.* Probability theory // Vol. II, 4th ed., Graduate Texts in Mathematics, Vol. 46, Springer-Verlag, 1978.

[4] *Беликова Т. П.* Моделирование линейных фильтров для обработки рентгеновских изображений в задачах медицинской диагностики // Цифровая оптика. Обработка изображений и полей в экспериментальных исследованиях. Сборник научных трудов. М.: Наука, 1990.

[5] *Беликова Т. П.* Синтез линейных фильтров для выделения диагностически важных объектов в задачах медицинской интроскопии // Цифровая оптика в медицинской интроскопии. М.: Наука, 1992.

[6] *Устинин М. Н., Махортых С. А., Молчанов А. М., Ольшевец М. М., Панкратов А. Н., Панкратова Н. М., Сухарев В. И., Сычев В. В.* Задачи анализа данных магнитной энцефалографии // Компьютеры и суперкомпьютеры в биологии. Под ред. В. Д. Лахно и М. Н. Устинина. Москва–Ижевск: Институт компьютерных исследований, 2002 — С. 327–348.

[7] *Устинин М. Н.* Спектрально-аналитические методы обработки данных вычислительного и натурного эксперимента // Дисс. д. ф.-м. н., Пущино, 2004.

Анализ текстур гистологических изображений*

Федотов Н. Г., Мокшанина Д. А., Романов С. В.
fedotov@pnzgu.ru
Пенза, Пензенский государственный университет

Предложен новый подход к анализу текстур гистологических изображений, основанный на аппарате стохастической геометрии и функционального анализа. Приведены результаты апробации данного подхода. Выделена наиболее эффективная применительно к рассматриваемой проблеме группа признаков.

Рост числа онкологических заболеваний обуславливает необходимость развития методов их ранней диагностики. Высокую степень достоверности имеет гистологический анализ. Квалифицированный врач на основе своего многолетнего опыта может по изображению гистологического препарата сделать экспертное заключение о наличии и типе заболевания. Однако, в силу огромного разнообразия форм заболеваний, многочисленности имеющихся эталонных образцов изображений, возможны экспертные ошибки. Кроме того, число специалистов-экспертов столь высокого класса невелико. В связи с этим для распознавания гистологических изображений целесообразно построить автоматизированную систему с высокой надежностью принятия решения.

Введение

На первом этапе работы такой системы проводится предобработка изображения, одной из задач которой является выделение фиброзной ткани. Для отсечения областей фиброзной ткани целесообразно применять методы анализа текстур изображения, так как они позволяют различать объекты одинакового цвета и формы.

Важной задачей анализа текстур является выделение признаков. Можно отметить три основных подхода к описанию текстур, на основании которых могут быть сформированы их признаки.

1. Статистический подход, при котором наличие или отсутствие пространственного взаимодействия между непроизводными элементами оценивается вероятностным образом. Наиболее распространенными методами, относящимся к этому подходу, являются:
   (а) метод, основанный на матрицах смежности, характеризующих статистики второго порядка и описывающих пространственные связи пар яркостей элементов в цифровом изображении текстуры [6];
   (б) метод, основанный на использовании длин серий, где под серией понимается непроизводный элемент, состоящий из максимальной связанной совокупности вытянутых в прямую линию пикселей одинаковой яркости [6].
2. Структурный подход, при котором непроизводные элементы явно определены. В терминах данного подхода текстура составлена из регулярно или почти регулярно распределенных по пространству непроизводных элементов. Поэтому анализ текстуры, с точки зрения такого подхода, должен состоять из описания непроизводных элементов и правил их размещения [6].
3. Фрактальный подход, при котором распознаваемый объект (текстуру) называют фракталом. Под фракталом понимают структуру, состоящую из частей, которые в каком-то смысле подобны целому. В основу этого метода положено выведенное Мандельбротом соотношение между периметром и площадью объекта, которое заключается в следующем. Если линия, ограничивающая объект, является фрактальной, то отношение ее длины к квадратному корню из площади ограниченной ею расходится. В противном случае это отношение есть константа. Анализ фракталов, по существу, дает характеристику текстуры [5].

Перечисленные выше подходы предполагают использование небольшого числа признаков, сознательно выделенных экспертом-аналитиком в качестве характеристик. Мы предлагаем новый подход к данной проблеме, основанный на аппарате стохастической геометрии и функционального анализа.

Новый подход к анализу текстур гистологических изображений

Подход с позиции стохастической геометрии позволяет автоматически, без непосредственного участия эксперта генерировать большое число признаков, имеющих не только медицинскую интерпретацию, но и являющихся математической абстрактной характеристикой изображения [2, в настоящем сборнике]. Опора на большое количество признаков повышает надежность распознавания. Эффективность аппарата стохастической геометрии была подтверждена в [1] и [3].

Признаки изображения в рассматриваемом подходе имеют структуру в виде композиции трех функционалов:

$$\Pi(F) = \Theta \circ P \circ T(F \cap l(\rho,\ \theta)) \qquad (1)$$

*Работа выполнена при финансовой поддержке РФФИ, проект № 09-07-00089.

где ρ, θ — нормальные координаты сканирующей прямой $l(\rho,\theta)$, с которыми связаны функционалы P и Θ соответственно; функционал T связан с параметром t, задающем точку на сканирующей прямой $l(\rho,\theta)$; $F(x,\ y)$ — функция изображения на плоскости (x,y). В связи с характерной структурой такие признаки были названы триплетными, их подробное описание приведено в [3].

Рассмотрим данный подход применительно к задаче анализа текстур гистологических изображений ткани щитовидной железы.

Гистологические изображения получают под микроскопом при увеличении в диапазоне от 50- до 1000-кратного, при этом каждый шаг увеличения дает свою долю диагностической информации. При 50-кратном увеличении основным выделяемым объектом является фиброзная ткань, отсечение которой необходимо для дальнейшей обработки фолликул.

Для формирования признаков фиброзной ткани нами использовалось её полутоновое изображение. Причем однородные сегменты сканирующей прямой выделялись следующим образом:

— определялась яркость в каждой точке сканирующей прямой $l(\rho,\theta)$;
— полученные значения представляют функцию яркости $I(x)$ для данной прямой;
— вычислялось значение производной функции яркости $\frac{dI(x)}{dx}$. По её экстремумам определялись резкие перепады яркости, то есть граничные точки однородных по яркости отрезков сканирующей прямой $l(\rho,\theta)$.

Принцип выделения однородных сегментов секущей прямой демонстрирует рис. 1.

Для решения поставленной задачи нами были выделены три группы признаков:

— признаки, характеризующие геометрические особенности изображения;
— признаки, характеризующие его яркость;
— признаки, характеризующие как геометрические, так и яркостные особенности изображения. Признаки данной группы не являются триплетными. Они определяются как евклидово расстояние между триплетными признаками первых двух групп.

Наиболее эффективная применительно к рассматриваемой проблеме группа признаков была определена нами в ходе эксперимента.

Признаки первых двух групп имеют одинаковую структуру вида (1). Отличие между ними заключается лишь в подходе к заданию параметра t, от которого зависит функционал T. Для построения признаков, характеризующих геометрические особенности изображения, параметр t задавал точку на сканирующей прямой $l(\rho,\ \theta)$ её декартовыми координатами. Для построения признаков, характеризующих яркости однородных сегментов изображения, параметр t задавал точку на сканирующей прямой $l(\rho,\ \theta)$ значением её яркости.

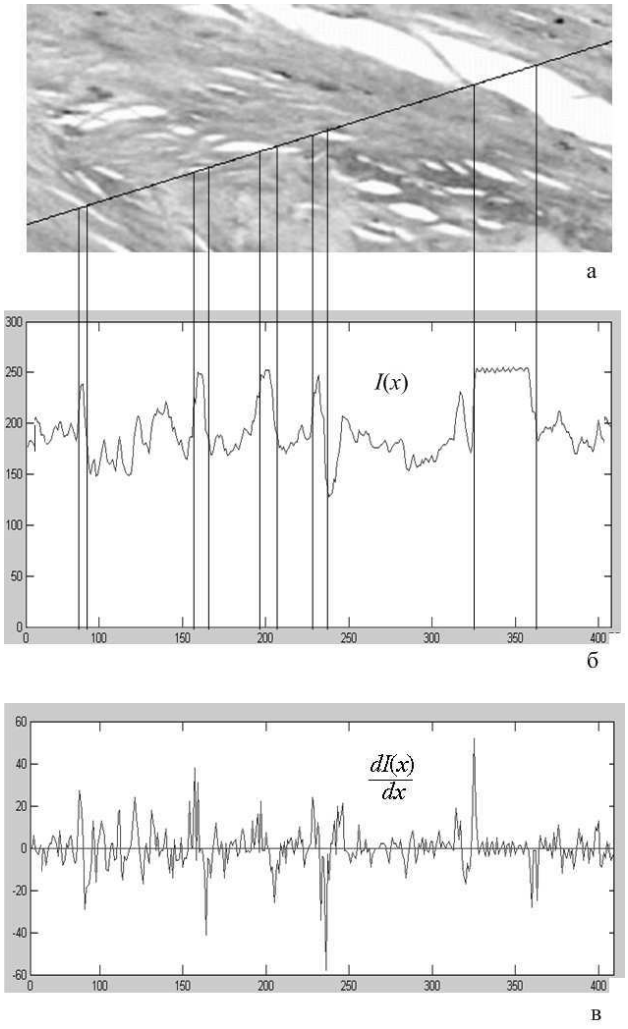


Рис. 1. Фрагмент фиброзной ткани со сканирующей прямой $l(\rho,\ \theta)$ (а), функция яркости $I(x)$ (б), производная $\frac{dI(x)}{dx}$ вдоль сканирующей прямой (в).

После генерации признаков была проведена процедура минимизации признакового пространства, обеспечивающая выделение минимального набора эффективных поисковых признаков, основанная на методах кластеризации. Данная процедура подробно рассмотрена в [4].

Проведенный эксперимент показал, что средняя ошибка классификации для группы информативных признаков первого типа составляет 4,8%, для группы признаков второго типа — 10%, для группы признаков третьего типа — 6.5%. Полученный результат говорит о том, что для выделения фиброзной ткани гистологического изображения следует применять признаки, характеризующие геометрические особенности изображения.

Выводы

Таким образом, методы стохастической геометрии позволяют автоматически, без непосредственного участия эксперта, генерировать большое число признаков, что повышает надежность распознавания. Причем хорошо различают текстуры гистологических изображений признаки, характеризующие геометрические особенности изображения.

Описанный подход обнаруживает высокую эффективность применительно к анализу текстур гистологических изображений.

Литература

[1] *Федотов Н. Г.* Теория признаков распознавания образов на основе стохастической геометрии и функционального анализа. М.: Физматлит, 2009. — 304 с.

[2] *Федотов Н. Г.* Трейс-преобразование как источник признаков распознавания // Всероссийская конференция ММРО-14. — С. 457–460.

[3] *Федотов Н. Г., Шульга Л. А.* Теория распознавания и понимания образов на основе стохастической геометрии // Искусственный интеллект.— 2002. № 2. — С. 282–289.

[4] *Федотов Н. Г., Курынов Д. А., Петренко А. Г., Кольчугин А. С., Смолькин О. А.* Интеллектуальная система поиска биометрических изображений в базе данных на основе стохастической геометрии // Надежность и качество.— 2006. — Т. 2. — С. 245–247.

[5] *Пьетронеро Л., Тозитти Э.* Фракталы в физике. Москва: Мир, 1988. — 644 с.

[6] *Харалик Р. М.* Статистический и структурный подходы к описанию текстур // ТИИЭР.— 1979. — Т. 67, № 5 — С. 98–118.

Сегментация гистологических изображений. Выделение фолликулов и ядер*

Федотов Н. Г., Романов С. В., Мокшанина Д. А.
fedotov@pnzgu.ru
Пенза, Пензенский государственный университет

В данном докладе рассматривается методика сегментации гистологических изображений. Для выделения информативных объектов предлагается оригинальный подход, основанный на анализе динамики изменения цвета пикселей. Поэтапно описан процесс выделения фолликулов и ядер на гистологическом изображении. Предложенный метод отличается низкой ресурсоемкостью и высокой точностью, что позволяет эффективно решить задачу сегментации гистологических изображений.

Компьютерный анализ гистологических изображений является важным этапом в развитии методов диагностики онкологических заболеваний.

Гистологический препарат представляет собой тонкий срез ткани. Для изучения препарат окрашивается гематоксилином и эозином. В зависимости от толщины препарата, качества и концентрации красителя, цитоплазма получает розоватую или красноватую окраску, а ядра — синюю или черную. После окраски препарат помещают под микроскоп, где осуществляется его фотосъёмка. Пример гистологического изображения представлен на рис. 1.

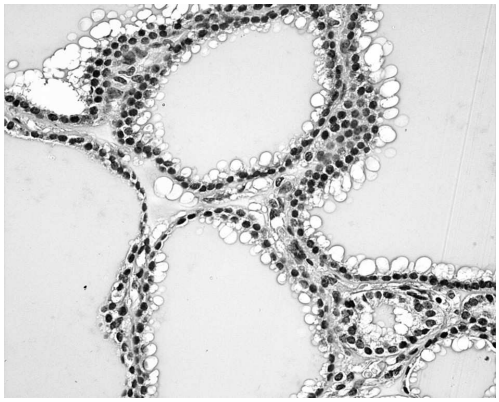


Рис. 1. Пример гистологического изображения щитовидной железы.

Для анализа гистологических изображений использовались методы стохастической геометрии, основанные на применении трейс-преобразования [4, в настоящем сборнике]. Для эффективного применения данного аппарата, необходимо выделить на исходном изображении наиболее информативные элементы.

Наиболее информативными элементами гистологического изображения являются ядра и фолликулы. Ядра представляют собой овальные или круглые элементы темного цвета. Фолликулы — овальные области, ограниченные ядрами. При анализе гистологических изображений рассматриваются размер, форма, и оптические свойства ядер. Аналогичные показатели применяются к фолликулам. Очевидно, что качество анализа напрямую связано с точностью сегментации.

Сложность сегментации гистологических изображений обусловлена рядом их особенностей. Как и большинство изображений биологических объектов, они слабо структурированы и отличаются значительной вариабельностью элементов по форме и размеру. Особенности подготовки гистологического препарата приводят к заметным вариациям цвета и яркости.

Выделение ядер

Для анализа гистологического изображения необходимо выделить ядра и представить их в виде отдельных изображений. Ключевым вопросом является обнаружение ядер. Попытки применить хорошо известные методы поиска объекта не дали хороших результатов. Сегментация по цветовым или яркостным характеристикам неэффективна ввиду большой вариабельности цвета и яркости. Поиск объекта по описанию или шаблону позволяет выделить лишь небольшое количество ядер ввиду значительной вариабельности форм, существования «незаполненных» ядер, а также явления наложения одного ядра на другое, возникающего в результате слишком большой толщины среза гистологического препарата.

При визуальном изучении гистологического изображения не составляет сложности выделить на нем ядра и фолликулы, поэтому рассмотрим более детально цветовые характеристики ядер. На рис. 2 представлен график изменения интенсивностей каждого компонента цвета при прохождении прямой через центр ядра.

На графике хорошо видно, что при прохождении через область ядра происходит значительное снижение интенсивности всех компонентов цвета, кроме того, при пересечении границы ядра наблюдаются характерные колебания.

Для оценки динамики изменений преобразуем исходный вектор интенсивности цвета по формуле $X_i = M_i - M_{i-1},\ i = 1, \ldots, n$, где X_i — вектор

*Работа выполнена при финансовой поддержке РФФИ, проект № 09-07-00089.

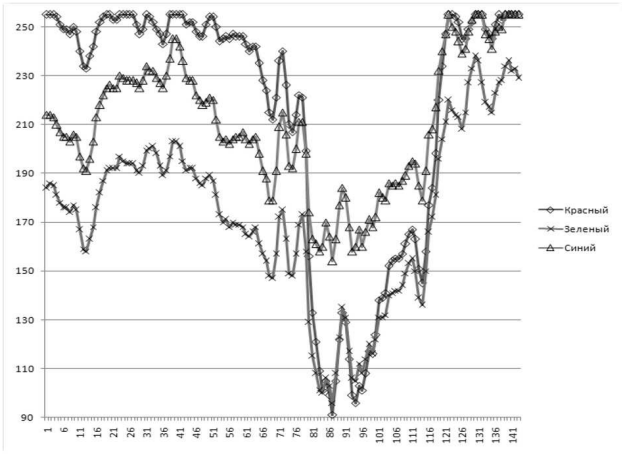


Рис. 2. Изменение компонентов цвета при прохождении прямой через ядро.

динамики изменения, а M_i — вектор значений интенсивности цвета. Затем происходит формирование вектора D путем суммирования соседних элементов X, имеющих одинаковый знак. В результате при рассмотрении вектора D для любой точки можно определить тенденцию изменения интенсивности цвета и амплитуду данного изменения на всем отрезке.

Для оценки частоты колебания введем вектор F, $F_i = 100\frac{1}{T}{}_i$, где T_i — вектор периода, вычисляемый как количество точек изображения между точками перегиба вектора D.

Данные преобразования осуществляются независимо для всех компонентов цвета (красный, зеленый, синий). В результате экспериментов и изучения гистологических изображений были выявлены следующие признаки ядер:

— частота изменения интенсивности синего цвета $FB_i > 30$;
— частота изменения интенсивности красного цвета $FR_i > 50$;
— амплитуда изменения интенсивности синего цвета $|DB_i| > 100$;
— амплитуда изменения интенсивности зеленого цвета $|DG_i| > 50$.

Точка, соответствующая любому из перечисленных выше признаков, считается принадлежащей ядру. Для уменьшения количества ошибок признаки вычисляются только для областей изображения с яркостью ниже среднего значения. На рис. 3 представлена маска изображения после применения указанных выше признаков.

Как видно из рис. 3, кроме точек, принадлежащих ядрам, на маске сформировался незначительный шум. Для его удаления исключим из маски связанные области с малой площадью (менее 200 точек).

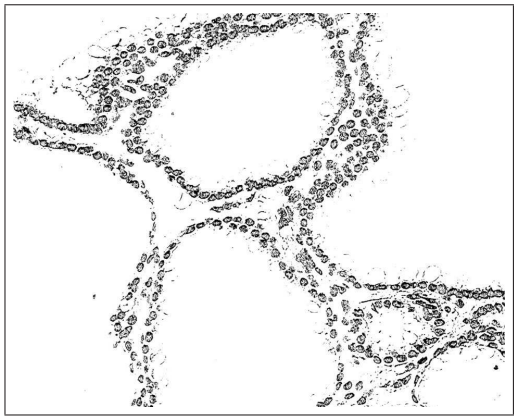


Рис. 3. Результат выделения точек, принадлежащих ядрам.

На рис. 4 представлен увеличенный фрагмент изображения после удаления шума.

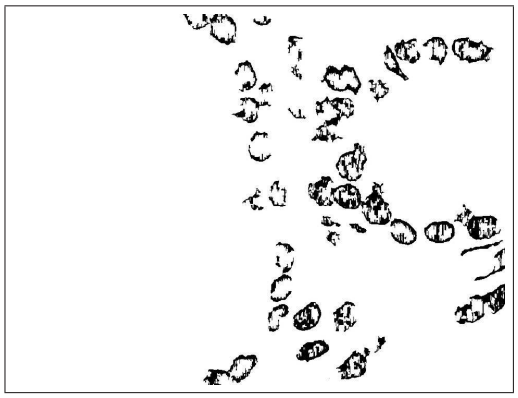


Рис. 4. Результат удаления шума с маски.

Кроме удаления шума по минимальной площади связанной области, на данном этапе обработки можно удалить и области со слишком большой площадью. Фильтрация элементов со слишком большой площадью ускорит дальнейшую обработку и исключит из маски области с наложением ядер друг на друга.

В данный момент выделена только часть точек, принадлежащих ядрам. Далее необходимо уточнить контур ядра. Решить данную задачу можно большим количеством методов. Очень хороший результат дает использование метода активных контуров. Но данный метод отличается очень низкой скоростью работы, поэтому уточнение контура осуществлялось по принципу цветовой близости. На рис. 2 хорошо видно, что при прохождении прямой через ядро колебания цвета внутри него значительно ниже колебаний цвета на границе. Основываясь на этом, для каждой точки маски проверяем соседние на предмет близости по цвету и уточняем контур каждого ядра. Результат уточнения контуров ядер представлен на рис. 5.

Рис. 5. Маска ядер после уточнения контуров.

После уточнения контуров практически вся площадь ядер включена в маску. В маску не вошли только часть центральных точек ядер. Это связано с особенностью строения ядра. Внутри ядра располагаются его элементы, например, ядрышки, которые значительно отличаются по цвету. При обработке некоторых изображений после данного этапа формируется только замкнутый контур ядра с абсолютно незаполненным центром. Для решения этой задачи применим заливку к замкнутой области ядра. Подойдет любой алгоритм, например, построчное заполнение. Результат заливки представлен на рис. 6.

Рис. 6. Маска ядер после заливки замкнутых областей.

На заключительном этапе для каждой замкнутой области маски осуществляется определение описывающей её прямоугольной области и создание отдельного изображения ядра путем копирования исходного изображения по сформированной маске. В результате формируется множество изображений, представляющих ядра. Для анализа взаимного расположения ядер на гистологическом препарате удобно использовать сформированную маску. Для уменьшения эффекта наложения ядер к ней применяется морфологическая операция эрозии.

Выделение фолликулов

Для выделения фолликулов можно использовать различные методы. По сравнению с ядрами, они обладают достаточно однородным цветом и являются наиболее светлыми областями изображения. Площадь фолликулов достаточно велика, что позволяет легко убрать шум, возникающей при сегментации по признакам яркости или цвета объекта.

Для ускорения и уточнения процесса выделения фолликулов можно использовать имеющуюся маску ядер. Если посмотреть на рис.1, то можно заметить, что гистологическое изображение состоит из фолликулов, ограниченных цепочкой ядер. В некоторых случаях на изображении могут находится области фиброзной ткани. Фиброзная ткань очень близка к фолликулам по цветовым и яркостным характеристикам, и для её удаления необходимо использовать анализ текстуры.

На первом этапе выделения фолликулов маска выделения ядер инвертируется и наращивается. Для увеличения точности анализа гистологических изображений из маски необходимо удалить элементы, соприкасающиеся с краем изображения. Далее, с использованием данной маски происходит уточнение контуров по цвету, заливка и представление фолликулов в виде отдельных изображений.

Выводы

Выполнена сегментация гистологического изображения с целью выделения наиболее информативных объектов — фолликулов и ядер. Предложенный метод отличается высоким быстродействием и простотой реализации. Параметры метода могут быть легко адаптированы для решения задач сегментации в других областях.

Направлением дальнейших исследований является улучшения качества сегментации гистологических изображений. Улучшить результат планируется путем введения адаптивного изменения параметров выделения ядер, основанного на анализе размеров замкнутых областей, полученных в результате выделения. Вторым направлением развития является разработка алгоритма разделения ядер, перекрывающих друг друга.

Литература

[1] *Sonka M., Hlavac V., Boyle R.* Image Processing, Analysis, and Mashine Vision. Brooks and Cole Publishing, 1998.

[2] *Serra J.* Image Analysis and Mathematical Morphology. Vol. 2, Academic Press, 1988.

[3] *Прэтт У. К.* Цифровая обработка изображений: В 2 т. — Москва: Мир, 1982.

[4] *Федотов Н. Г.* Трейс-преобразование как источник признаков распознавания // Всероссийская конференция ММРО-14. — С. 457–460.

Скрытая профильная периодичность как новый тип периодичности генома

Чалей М. Б., Кутыркин В. А.
maramaria@yandex.ru, vladkutyr@yandex.ru
Пущино, Институт математических проблем биологии РАН
Москва, Московский Государственный Технический Университет им. Н. Э. Баумана

В работе рассматривается новое понятие скрытой периодичности в текстовых строках, названное профильной периодичностью или профильностью, расширяющее понятие размытого тандемного повтора. На основе оригинального подхода предлагаются методы распознавания скрытой профильности в геноме.

В каждой клетке организма, не связанной с репродуктивной функцией, содержится характерный для данного организма набор пар гомологичных хромосом, определяющий его наследственную информацию. Половина хромосом из этого набора, полученная от одного из родителей и дополненная до комплекта хромосом, связанных с полом, образует характерную совокупность хромосом, представляющих геном организма. Носителем информации в каждой хромосоме является биополимерная молекула дезоксирибонуклеиновой кислоты (ДНК), образованная мономерными звеньями — нуклеотидами четырёх типов, называемых аденин, гуанин, цитозин и тимин (часто обозначаемых буквами a, g, c, t). Таким образом, молекула ДНК каждой хромосомы может быть представима в виде уникальной последовательности букв (нуклеотидов) исходного четырёхбуквенного алфавита. Расшифровка генома (или, используя специальный термин, — секвенирование генома) организма означает определение уникальных нуклеотидных последовательностей всех хромосом из характерной совокупности. Многочисленные проекты секвенирования геномов человека, мыши, крысы, дрожжей, растения арабидопсис, плодовой мушки, и др. за последнее десятилетие предоставили возможность анализа различных уровней организации и состава нуклеотидных последовательностей хромосом. Такие последовательности также называют генетическими текстами. Длина генетических текстов полных геномов и отдельных хромосом достигает иногда нескольких миллионов и даже миллиардов нуклеотидов. Необходимость анализа таких громадных объемов информации способствует развитию математических методов как для поиска уже известных структурных особенностей генома, так и для выявления его новых структурно-функциональных свойств.

Одними из важных структурных объектов в геноме являются тандемные повторы (последовательно повторяющиеся копии некоторого фрагмента нуклеотидной последовательности — паттерна периодичности) и, в том числе, размытые тандемные повторы, копии паттерна периодичности в которых повреждены заменами, вставками и делециями букв.

С тандемными повторами могут быть связаны конкретные функции (такие, как формирование участков взаимодействия ДНК с белками) и особые свойства генома, например, гибкость локальных районов ДНК. Тандемные повторы с паттерном от 2 до 6 букв (микро-сателлиты) влияют на регуляцию генов (участков, кодирующих, в основном, белки). С помощью варьирующих по длине районов микро- и мини-сателлитов (длина паттерна периодичности не превышает 30 букв) идентифицируют микроорганизмы и определяют родственные связи между отдельными личностями.

Районы тандемных повторов являются участками генетической нестабильности (они способны как удлиняться так и сокращаться в длине) и, как следствие, — источниками риска проявления наследственных заболеваний. Хорошо известен феномен экспансии триплетов, т.е. неконтролируемого удлинения районов триплетных тандемных повторов, ведущего к ряду генетических неврологических заболеваний.

Поскольку размытые тандемные повторы бывает весьма трудно, порой невозможно, идентифицировать глазом, они получили название скрытой периодичности. Существование в нуклеотидных последовательностях (текстовых строках в алфавите из четырех букв) скрытой периодичности рассматривается как базовое свойство геномной ДНК [1].

Ранее распознавание скрытой периодичности было основано только на понятии размытого тандемного повтора, периодическая структура которого определяется текстовым консенсус-паттерном. В настоящее время для геномных последовательностей введено новое понятие скрытой периодичности, названное профильной периодичностью [2] или профильностью. Было показано, что это понятие расширяет понятие размытого тандемного повтора. В настоящей работе предложены оригинальные методы распознавания скрытой профильной периодичности. Предлагаемые методы также хорошо распознают и размытые тандемные повторы. Методы распознавания скрытой профильности основаны на спектрально-статистическом подходе (2С подходе), с помощью которого в последовательностях генома выделяются спектрально-статистические характеристики, чувствительные к иско-

мому типу периодичности. Введение таких характеристик стало возможным благодаря предложенной ранее модели проявления скрытой профильной периодичности [2]. Согласно этой модели последовательности генома (текстовые строки) рассматриваются как реализации специальных случайных строк, названных профильными строками. По определению профильная строка является совершенным тандемным повтором со случайным паттерном периодичности, составленным из независимых случайных букв.

Структура случайных и профильных строк

Далее $A = \langle a_1, \ldots, a_K \rangle$ — упорядоченный алфавит анализируемой текстовой строки, $p = (p^1, \ldots, p^K)$ — столбец частот букв алфавита A, $\sum\limits_{i=1}^{K} p^i = 1$, и $\mathrm{Chr}(p, A)$ — случайная буква, принимающая с вероятностью (частотой) p^i значение буквы a_i алфавита A. В частности, для последовательностей ДНК генома $K = 4$, $a_1 = a$ (аденин), $a_2 = g$ (гуанин), $a_3 = c$ (цитозин), $a_4 = t$ (тимин). Строка

$$\mathrm{Str}_L(\pi, A) = \mathrm{Chr}(\pi_1, A)\,\mathrm{Chr}(\pi_2, A)\ldots\mathrm{Chr}(\pi_L, A)$$

из перечисленных независимых случайных букв далее называется случайной строкой длины L в алфавите A. Эта случайная строка индуцируется L-профильной матрицей $\pi = (\pi_1, \ldots, \pi_L) = (\pi_j^i)_L^K$. Если b_j — независимая реализация случайной буквы $\mathrm{Chr}(\pi_j, A)$, $j = 1, \ldots, L$, строки $\mathrm{Str}_L(\pi, A)$, то текстовая строка $b_1 b_2 \ldots b_L$ является реализацией случайной строки $\mathrm{Str}_L(\pi, A)$.

Если $\mathrm{Str}_L(\pi, A) = \mathrm{Str}\,\mathrm{Str}\ldots\mathrm{Str}$, где Str — некоторая случайная строка, то случайная строка $\mathrm{Str}_L(\pi, A)$ называется периодической. Непериодическая случайная строка $\mathrm{Str}_L(\pi, A)$ называется случайным паттерном периодичности и для неё используется обозначение $\mathrm{Ptn}_L(\pi, A) = \mathrm{Str}_L(\pi, A)$.

Случайная строка

$$\mathrm{Tdm}_L(\pi, A, n) = \mathrm{Ptn}_L(\pi, A)\ldots\mathrm{Ptn}_L(\pi, A)\,\mathrm{Str}_m(\pi')$$

называется совершенным тандемным повтором длины n (в алфавите A) со случайным паттерном периодичности $\mathrm{Ptn}_L(\pi, A)$, где $\pi' = (\pi_1, \ldots, \pi_m) = (\pi_j^i)_m^K$ — профильная матрица, $0 \leqslant m < L$. Такой тандемный повтор называется L-профильной строкой (в алфавите A) длины n с главной профильной матрицей π. В этом случае число L называется основным периодом профильной строки $\mathrm{Tdm}_L(\pi, A, n)$.

Профильная строка $\mathrm{Tdm}_L(\pi, A, n)$, где $L = 1$, называется однородной строкой с заданным столбцом π частот встречаемости букв алфавита A.

Базовые спектрально-статистические характеристики

А) Рассмотрим последовательное разбиение строки $\mathrm{Str} = \mathrm{Str}_n(\pi, A) = \mathrm{Chr}(\pi_1, A)\,\mathrm{Chr}(\pi_2, A)\ldots\mathrm{Chr}(\pi_n, A)$ на подстроки длины L. Такое разбиение называется горизонтальным L-профилем строки Str с тест-периодом $L \leqslant n$ и тест-экспонентом $R_L = n/L$. Для текстовой строки горизонтальный L-профиль строится аналогичным образом.

Б) Каждый горизонтальный L-профиль строки можно представить в виде вертикального L-профиля (и наоборот). Для этого подстроки горизонтального L-профиля последовательно располагаются друг под другом, образуя строки вертикального L-профиля.

В) Вертикальный L-профиль случайной строки $\mathrm{Str} = \mathrm{Str}_n(\pi, A)$ позволяет вычислить её L-профильную матрицу $\Pi_{\mathrm{Str}}(L) = (\theta_j^i)_L^K$, в которой каждый элемент θ_j^i равен вероятности встречаемости i-той буквы алфавита A в j-том столбце вертикальных L-профилей реализаций случайной строки $\mathrm{Str} = \mathrm{Str}_n(\pi, A)$. Таким образом, для строки $\mathrm{Str} = \mathrm{Str}_n(\pi, A)$ введена функция Π_{Str}, которая каждому тест-периоду L ставит в соответствие L-профильную матрицу Π_{Str}. Эта функция называется профильно-матричным спектром строки $\mathrm{Str} = \mathrm{Str}_n(\pi, A)$. Аналогично, для текстовой строки str строится (выборочный) профильно-матричный спектр Π_{str} с диапазоном тест-периодов $1, \ldots, L_{\max} \backsim \frac{n}{5K}$.

Критерий проверки статистической неотличимости строк

Пусть str и Str — две строки длины n в алфавите A, где str — текстовая и Str — случайная строки. Эти строки определяют профильно-матричные спектры Π_{str} и Π_{Str} соответственно, с единым диапазоном тест-периодов от 1 до $L_{\max} \backsim \frac{n}{5K}$. Для наглядности изложения предполагается, что $L_{\max}$ не более 100. Согласно спектрам Π_{str} и Π_{Str} создается статистический спектр $H_{(\mathrm{str},\,\mathrm{Str})}$ сравнения строк str и Str. Процедура его построения следующая.

Для λ-профильных матриц $\Pi_{\mathrm{str}}(\lambda) = (\pi_j^{*i})_\lambda^K$ и $\Pi_{\mathrm{Str}}(\lambda) = (\pi_j^i)_\lambda^K$, $1 \leqslant \lambda \leqslant L_{\max}$, вычисляется статистика Пирсона:

$$\psi\big(\Pi_{\mathrm{str}}(\lambda), \Pi_{\mathrm{Str}}(\lambda)\big) = R_\lambda \sum_{j=1}^{\lambda}\sum_{i=1}^{K}(\pi_j^{*i} - \pi_j^i)^2/\pi_j^i, \quad (1)$$

где $R_\lambda = n/\lambda$. При достаточно большом R_λ статистика (1) имеет стандартное χ^2-распределение с $N = (K-1)\lambda$ степенями свободы, то есть

$$\psi\big(\Pi_{\mathrm{str}}(\lambda), \Pi_{\mathrm{Str}}(\lambda)\big) \sim \chi^2. \quad (2)$$

Статистика (1) измеряет различие профильных матриц $\Pi_{\mathrm{str}}(\lambda)$ и $\Pi_{\mathrm{Str}}(\lambda)$. С помощью формул (1)

и (2) определим статистику

$$H_{(\mathrm{str},\,\mathrm{Str})}(\lambda) = \frac{\psi\big(\Pi_{\mathrm{str}}(\lambda), \Pi_{\mathrm{Str}}(\lambda)\big)}{\chi^2_{\mathrm{crit}}((K-1)\lambda, \alpha)}, \qquad (3)$$

где $\chi^2_{\mathrm{crit}}((K-1)\lambda, \alpha)$ — критическое значение для стандартной статистики χ^2 с $N = (K-1)\lambda$ степенями свободы на уровне значимости $\alpha = 0{,}05$.

Спектр $H_{(\mathrm{str},\,\mathrm{Str})}$ позволяет проверить гипотезу о статистической неотличимости строк str и Str. В частности, если строка str является реализацией случайной строки Str, то, как правило, для любого тест-периода $\lambda = 1, \ldots, L_{\max}$ выполняется условие: $H_{(\mathrm{str},\,\mathrm{Str})}(\lambda) \leqslant 1$. Следовательно, если выполняется условие: $H_{(\mathrm{str},\,\mathrm{Str})} \leqslant 1$, то (на выбранном уровне значимости) указанные строки можно признавать статистически неотличимыми. В противном случае, следует принять гипотезу о различии указанных строк. Спектр $H_{(\mathrm{str},\,\mathrm{Str})}$ индуцирует (выборочное) множество $\mathrm{Sp}_{H}(\mathrm{str},\,\mathrm{Str})$ значимых отличий указанных строк. По определению:

$$\mathrm{Sp}_{H}(\mathrm{str},\,\mathrm{Str}) = \\ = \big\{\lambda\colon 1 \leqslant \lambda \leqslant L_{\max},\ H_{(\mathrm{str},\,\mathrm{Str})}(\lambda) > 1\big\}. \qquad (4)$$

Выборочное множество $\mathrm{Sp}_{H}(\mathrm{str},\,\mathrm{Str})$ позволяет сформулировать следующий критерий проверки статистической неотличимости строк str и Str. Если $\mathrm{Sp}_{H}(\mathrm{str},\,\mathrm{Str}) = \varnothing$, то для указанных строк принимается гипотеза статистической неотличимости. В противном случае эта гипотеза отвергается.

Методы распознавания скрытой профильной периодичности

Распознать профильную периодичность в анализируемой текстовой строке str — значит найти такую профильную строку, реализацией которой, согласно критерию неотличимости строк, является эта текстовая строка. Для этого строка str сравнивается с профильными строками вида $\mathrm{Tdm}_{\Lambda}(\Pi_{\mathrm{str}}(\Lambda), A, n)$, где тест-период Λ выбирается из диапазона $\Lambda = 1, \ldots, L_{\max}$. Статистическим показателем неотличимости сравниваемых строк является спектр $D_{(\Lambda,\,\mathrm{str})} = H_{(\mathrm{str},\,\mathrm{Tdm}_{\Lambda})} = \mathtt{D}\Lambda$ отклонения строки от Λ-профильности. Первый тест-период L из диапазона $1, \ldots, L_{\max}$, для которого выполняется критерий неотличимости сравниваемых строк: $\mathrm{Sp}_{H}(\mathrm{str},\,\mathrm{Str}) = \varnothing$ (см. (4)) — позволяет предположить существование скрытой L-профильности в анализируемой строке str.

При выявленной L-профильности и наличии достаточного статистического материала для анализируемой текстовой строки возможна статистическая реконструкция спектра $D_{(\Lambda,\,\mathrm{str})} = \mathtt{D}\Lambda$ отклонения от Λ-профильности, например, $D_{(1,\,\mathrm{str})} = \mathtt{D1}$ — от однородности.

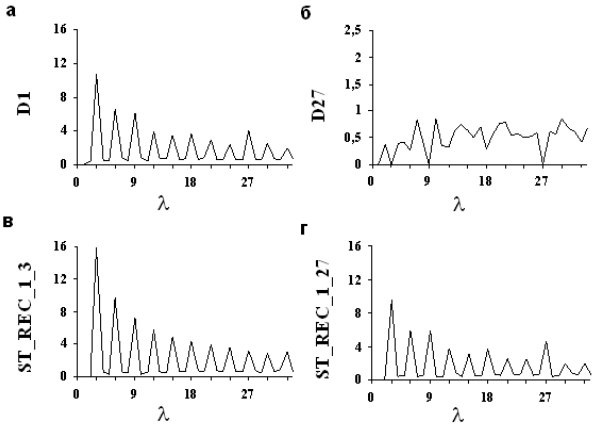


Рис. 1. Спектрально-статистические характеристики фрагмента последовательности (5046-5742 нукл.) гена «cya»: кальмодулин-чувствительной аденилатциклазы *Bordetella pertussis* (GenBank, locus BPCYA). В последовательности проявляется 27-профильность.
(а) Спектр D1 отклонения от 1-профильности.
(б) Спектр отклонения от 27-профильности.
(в) Статистическая реконструкция спектра D1 фрагмента в предположении его скрытой 3-профильности.
(г) Статистическая реконструкция спектра D1 фрагмента в предположении его скрытой 27-профильности.

Сходство исходного и реконструированного спектров отклонения будет служить дополнительным подтверждением найденной оценки случайного паттерна профильной периодичности. Для спектра статистической реконструкции вводится обозначение $\mathtt{ST\_REC\_}\Lambda\mathtt{\_}L$.

Также для дополнительного подтверждения найденной оценки случайного паттерна профильной периодичности можно использовать спектр $H_{(\mathrm{Tdm}_{L},\,\mathrm{Tdm}_{\Lambda})}$, который будет называться теоретической реконструкцией спектра $D_{(\Lambda,\,\mathrm{str})}$ для тест-периода L. Для этого спектра справедливы те же замечания, что и для спектра статистической реконструкции. Теоретическая реконструкция полезна в условиях ограниченного статистического материала, когда статистическая реконструкция может приводить к значительному искажению реконструированного спектра.

В качестве примера распознавания скрытой профильности рассмотрим фрагмент последовательности бактериального генома, спектрально-статистические характеристики которого приведены на рис. 1. Этот фрагмент не является размытым тандемным повтором. Ранее высказывались предположения о наличии в этом фрагменте скрытой периодичности неизвестного типа в 3 нуклеотида (Фурье-анализ) и 27 нуклеотидов [3].

Согласно рис. 1(а) в рассматриваемом фрагменте наблюдается значимая неоднородность на тест-периодах в $3M$ нуклеотидов, где $M = 1, \ldots, 11$. Проведенный анализ показал, что скрытая про-

фильность проявляется только на тест-периоде в 27 нуклеотидов (см. рис. 1(б)), так как значения спектра D27 отклонения от 27-профильности нигде не превышают 1. Существование в последовательности фрагмента скрытой 27-профильности подтверждается сходством статистической реконструкции ST_REC_1_27 на тест-периоде в 27 нукл. со спектром D1 анализируемой последовательности (см. рис. 1(г)). Для сравнения на рис. 1(в) показана статистическая реконструкция ST_REC_1_3 спектра D1 на тест-периоде в 3 нукл.

Выводы

Применение предлагаемого спектрального статистического подхода (2С подхода) к распознаванию нового типа скрытой периодичности в текстовых строках — профильной периодичности было показано в настоящей работе на примере последовательностей геномов организмов. С помощью 2С подхода можно получить достоверное подтверждение выдвигаемых ранее гипотез о наличии в геноме скрытой периодичности неизвестного типа. При этом для анализируемой последовательности распознается стохастическая структура профильной периодичности. В случае достаточного статистического материала, параметры этой структуры получают достоверное подтверждение на основе статистической реконструкции выявленных в работе спектрально-статистических характеристик анализируемой последовательности, чувствительных к скрытой профильной периодичности. В условиях ограниченного статистического материала такая реконструкция становится менее наглядной и носит качественный характер, поэтому проблема улучшения количественных показателей реконструкции требует дальнейшей разработки.

Литература

[1] *Bolshoy A.* Revisiting the relationship between compositional sequence complexity and periodicity // Comput. Biol. Chem. — 2008. — V. 32. — P. 17–28.

[2] *Chaley M., Kutyrkin V.* Model of perfect tandem repeat with random pattern and empirical homogeneity testing poly-criteria for latent periodicity revelation in biological sequences // Math. Biosci. — 2008. — V. 211. — P. 186–204.

[3] *Korotkov E. V., Korotkova M. A., Kudryashov N. A.* Information decomposition method to analyze symbolical sequences // Phys. Lett. A — 2003. — V. 312. — P. 198–210.

О классификационном подходе к имитационному моделированию транспортных потоков*

Чехович Ю. В.
d_yura@ccas.ru
Москва, Вычислительный центр РАН

В работе предлагается так называемый классификационный подход к моделированию автомобильных транспортных потоков. Модели, построенные на основе предлагаемого подхода, позволят изучить зависимость пропускной способности транспортной сети от таких параметров как плотность потока, манеры вождения водителей, способов организации дорожного движения, случайных факторов, влияющих на пропускную способность. Обсуждаются вопросы идентификации моделей на реальных транспортных системах и ситуациях, возникающих при движении транспортных средств.

Введение

Рассматривается задача синтеза и идентификации имитационной модели транспортного потока на фиксированной транспортной сети. Моделирование транспортных потоков — задача чрезвычайно актуальная и сложная. Актуальность обусловлена непрекращающимся ростом автопарка во всем мире и, в особенности, крупных городах. В настоящее время практически каждый мегаполис испытывает проблемы, связанные с пропускной способностью транспортных систем [5]. При этом значительная часть проблем обусловлена либо ошибками проектирования транспортной сети, либо неоптимальным управлением движением, либо, чаще всего, сочетанием того и другого. Таким образом, создание моделей, способных помочь при проектировании сетей и при управлении движением, является чрезвычайно актуальной задачей.

В то же время, следует отметить, что исследователями в этой области разработано большое количество разнообразных математических моделей, которые в соответствии с [1] можно разбить на три класса: модели-аналоги, модели следования за лидером и вероятностные модели. Модели-аналоги, которые также называются «макроскопическими моделями», уподобляют транспортный поток какому-либо физическому аналогу (газу, сжимаемой или несжимаемой жидкости, «замерзающей жидкости» и т. п.). Модели следования за лидером или «микроскопические модели» рассматривают каждое транспортное средство обособленно и изучают взаимосвязи между движением каждого транспортного средства и движением его соседей по потоку. Вероятностные модели подходят к задаче с позиций теории массового обслуживания и рассматривают взаимодействия элементов транспортной сети между собой, с характерными для каждого элемента ограничениями. В то же время, сейчас по-видимому не существует модели, которая позволяла бы решать поставленные задачи с качеством, удовлетворяющим потребителя.

В настоящей работе предлагается класс моделей транспортных потоков, основанный на гипотезе существования ярко выраженных и устойчивых типов поведения у водителей транспортных средств, которые проявляются в виде реакций водителя на складывающиеся вокруг него в процессе движения локальные ситуации [6]. Предлагаемый класс моделей сочетает в себе особенности «микроскопического моделирования» с вероятностным подходом.

Описание объекта моделирования

В соответствии с [3], далее будем предполагать, что задана некоторая транспортная система, а для каждого водителя транспортного средства известен полный маршрут движения по транспортной системе. Необходимо моделировать оперативные действия каждого водителя, направленные на реализацию указанного маршрута движения.

Водитель в процессе движения получает информацию о локальной ситуации в окрестности своего транспортного средства. Под описанием ситуации здесь понимаются данные о положениях, скоростях и ускорениях тех участников дорожного движения, которые видны водителю. Кроме того, в описание локальной ситуации включается описание участка транспортной системы (количество и ширина полос, качество покрытия, угол кривизны по горизонтали и вертикали, наличие препятствий и т. п.) на котором находится водитель в данный момент.

В ходе движения водитель управляет автомобилем, выбирая типы действия (принимая решения) из некоторого множества. Множество принимаемых решений может быть, например, таким: увеличить скорость, затормозить, перестроится в другой ряд, совершить обгон, повернуть, ничего не менять. Предполагается, что каждый водитель выбирает тип действия на основе анализа локальной ситуации вокруг транспортного средства и в одинаковых или сходных ситуациях он выбирает один и тот же тип действия. Следует отметить, что в рамках реальной транспортной системы присутствуют водители, которые в одинаковых или сходных локальных ситуациях действуют по-разному (выбирают различные типы действий). Это позволяет утвер-

*Работа выполнена при финансовой поддержке РФФИ, проект № 08-07-00304.

ждать, что существует несколько типов водителей (много меньше общего числа водителей), при этом водители одного типа в сходных ситуациях действуют одинаково, а реакции водителей разных типов в сходных ситуации могут различаться.

Способы описания локальных ситуаций

Очевидно, что количество возможных различных описаний локальных ситуаций чрезвычайно велико. В то же время, разумным выглядит предположение о существовании типов (классов) ситуаций таких, что с точки зрения водителей различные ситуации одного типа неразличимы. Такое предположение позволяет провести факторизацию всего множества описаний ситуаций и перейти к весьма ограниченному множеству типов ситуаций.

Для факторизации множества принимаемых ситуаций можно использовать несколько подходов. Одним из таких подходов является введение на пространстве описания ситуаций метрики, с последующей кластеризацией ситуаций или экспертным выделением опорных ситуаций для каждого типа ситуаций.

Можно также использовать описания ситуаций, сформулированные в виде набора логических признаков. Для введения такого описания пространство вокруг моделируемого автомобиля в системе координат, привязанной к этому автомобилю, разбивается определенным образом на зоны. Затем формируется набор элементарных высказываний про каждую из зон. Например, «зона содержит автомобиль», «автомобиль в зоне едет быстрее (медленнее) моделируемого», «автомобиль в зоне сигнализирует о повороте направо», «автомобиль в зоне сигнализирует о торможении» и т. п. Такой способ описания ситуаций, с одной стороны, позволяет получить содержательную интерпретацию каждого типа локальной ситуации. С другой стороны, в ряде случаев использование логического способа описания локальных ситуаций потребует существенно бóльших вычислительных ресурсов при решении задачи классификации ситуаций. Данный способ описания локальных ситуаций был предложен в работе И. М. Селезнёва «Имитационное моделирование транспортных потоков на основе классификации локальных ситуаций».

Таким образом, проведя разбиение множества локальных ситуаций на классы, зафиксировав множество типов действий водителя, информацию, необходимую для идентификации модели можно представить в виде матрицы, столбцам которой соответствуют типы локальных ситуаций, а строкам — типы водителей. Каждой паре ⟨тип ситуации, тип водителя⟩ должен быть поставлен в соответствие тип действия, которое предпримет водитель данного типа в ситуации данного типа. Идентификация имитационной модели сводится к полному или частичному заполнению данной матрицы.

Процесс моделирования

Построенная на описанном подходе модель, может структурно быть устроена следующим образом. Для каждого моделируемого транспортного средства зафиксирован тип водителя и характеристики, не зависящие от типа водителя: габаритные размеры, предельные ускорения и т. п. В модель вводится некоторое количество транспортных средств с заданным распределением характеристик и типов водителей. Далее с некоторым шагом по времени для каждого моделируемого транспортного средства производится анализ локальной ситуации, в результате которого определяется тип локальной ситуации. Для полученного типа локальной ситуации однозначным образом определяется тип действия водителя, выполняется данной действие, после чего производится переход на следующий шаг.

Способы идентификации модели

Следует отметить, что, как правило, идентификация имитационных моделей является самым дорогим и трудоемким этапом создания модели [4]. Рассматриваемый класс моделей, симулирующих движение транспортных средств, не является исключением из этого правила.

Основой предлагаемого подхода к моделированию является согласованное решение двух задач кластеризации: выделение типов ситуаций и выделение типов водителей. Можно утверждать, что качество решения этих задач определяет возможность последующей адекватной идентификации модели и качество построенной модели в целом.

Идентифицированная модель характеризуется способом классификации локальных ситуаций, зафиксированным набором типов водителей и типов действий, соответствующих паре ⟨тип ситуации, тип водителя⟩. Также необходимо зафиксировать долю водителей каждого типа относительно общего числа водителей.

Чтобы получить описанные данные необходимо, но, вообще говоря, не достаточно, располагать последовательными описаниями локальных ситуаций, возникающих в окрестности моделируемого транспортного средства, а также протоколом решений, реализуемых водителем в качестве реакций на изменение локальных ситуаций.

Для получения данных о решениях, которые принимаются водителем в реальных ситуациях можно использовать специализированные мобильные стенды. Такой стенд представляет собой автомобиль, оборудованный для видеозаписи всего

происходящего вокруг автомобиля, таким образом, как это видит водитель, записи действий водителя путем фиксации управляющих воздействий водителя на автомобиль, а также записи изменения положения автомобиля в пространстве (например, на основе координат, получаемых от системы GPS) [2]. Обработка данных, собираемых таким стендом, позволяет получить описание локальных ситуаций и синхронизировать их с протоколом действий водителя и положений автомобиля в рамках транспортной сети.

Недостатком такого способа сбора данных является невозможность оборудования таким измерительным комплексом всех или значительного количества участников движения, а также осведомленность водителя о сборе данных, что, безусловно, влияет на стиль управления транспортным средством.

Для идентификации модели можно также использовать данные по реальным транспортным потокам, записанные в так называемом «трековом виде». Такая запись содержит координаты всех транспортных средств на участке транспортной сети в каждый момент времени и позволяет с некоторой разумной погрешностью восстанавливать скорости и ускорения всех транспортных средств. Треки транспортных средств можно зафиксировать, обработав специальными методами данные видеосъемок реальных транспортных потоков. При этом видеосъемка может производиться несколькими синхронизированными видеокамерами с перекрывающимися или не перекрывающимися полями зрения, что позволяет одновременно охватывать достаточно большие по протяженности участки транспортной сети.

Полученные трековые данные можно использовать как для выявления типов решений, принимаемых водителями, и моментов времени, в которые эти решения принимаются, так и для восстановления описаний локальных ситуаций, которые наблюдает каждый водитель в каждый момент времени. Таким образом, обработка трековых данных позволяет провести полную идентификацию модели. К недостаткам данного подхода к идентификации следует отнести объективные сложности в получении видеозаписей участков транспортной сети значительной протяженности. Например, получение такого рода данных по транспортной сети большого города или, хотя бы, района города представляется нереалистичным.

Верификация модели

Верификация модели производится путем сопоставления определенного набора рассчитываемых характеристик реальных транспортных потоков, на которых производилась идентификация модели, с характеристиками потоков, которые возникают в результате моделирования. Также используется экспертная оценка адекватности моделирования, во-первых, для качественного анализа результатов моделирования, во-вторых, для исследования поведения модели в критических ситуациях: аварии, возникновение заторов, приближение к предельной плотности потока, возникновение непредвиденных препятствий и т. п.

Выводы

В работе предложен подход к синтезу и идентификации моделей транспортных потоков, который позволит изучать зависимость пропускной способности транспортной сети от плотности потока, манеры вождения водителей, способов организации дорожного движения, различного рода случайных факторов (погодные условия, припаркованные автомобили, аварии и т. п.), влияющих на пропускную способность. Также предложены способы идентификации (настройки) моделей на реальных транспортных системах и ситуациях, возникающих при движении реальных транспортных средств.

Литература

[1] *Брайловский Н. О., Грановский Б. И.* Моделирование транспортных систем // М.: Транспорт, 1978 — 125 с.

[2] *Буслаев А. П., Кузьмин Д. М., Яшина М. В.* Компьютерные методы обработки информации и распознавания образов в задачах транспорта и связи. Часть 5: Мобильный Улично-Дорожный РЕЦептор «МУДРец» // МТУСИ. — М., 2008. — 101 с.

[3] *Иванов Г. Е., Рудаков К. В., Чехович Ю. В.* Алгебраический подход к имитационному моделированию транспортных потоков // Некоторые проблемы фундаментальной и прикладной математики. М., 2007 С. 96–102.

[4] *Павловский Ю. Н.* Имитационные модели и системы (Математическое моделирование. Вып. 2) — М.: ФАЗИС: ВЦ РАН, 2000. — 134 с.

[5] *Семенов В. В.* Математическое моделирование динамики транспортных потоков мегаполиса // Препринт Института прикладной математики им. Келдыша, 2004.
`http://spkurdyumov.narod.ru/Semenov.pdf`

[6] *Чехович Ю. В.* Классификационный подход к имитационному моделированию транспортных потоков // Интеллектуализация обработки информации (ИОИ-2008): Тезисы докл. — Симферополь: КНЦ НАН Украины, 2008. — С. 239–240.

Содержание

Алфавитный указатель

Л

М

Н

О

П

Р

Научное издание

Математические методы
распознавания образов

Сборник докладов
14-й Всероссийской конференции

Напечатано с готового оригинал-макета

Издательство ООО «МАКС Пресс»
Лицензия ИД №00510 от 01.12.1999
Подписано к печати 21.08.2009
Печать офсетная. Бумага офсетная.
Формат 60×88 1/8. Усл. печ. л. 79,0. Тираж 300 экз. Изд. № 426 Заказ .

119992, ГСП-2, Москва, Ленинские горы, МГУ им. М. В. Ломоносова,
2-й учебный корпус, 627 к.
Тел. 939-3890, 393-3891, Тел./Факс. 939-3891.