Морфологический подход к синтезу метрических классификаторов и его реализация методом отыскания минимального разреза графа соседства для обучающей выборки*

Ю. В. Визильтер, В. С. Горбацевич
viz@gosniias.ru, gvs@gosniias.ru

Предложен и обоснован морфологический подход к синтезу классификаторов в задаче обучения с учителем, предполагающий решение двух последовательных подзадач — переразметки обучающей выборки и оптимальной корректной интерполяции (расширения) решающего правила. Доказана проективность морфологических операторов машинного обучения. Показано, что классы морфологических классификаторов и алгоритмов обучения образуют систему вложенных классов возрастающей сложности в смысле Пытьева и могут быть использованы для структурной минимизации риска переобучения. Показано, что для случая двухклассовой классификации метод минимального разреза графов позволяет отыскивать глобальный оптимум введенного критерия качества классификации.

Введение

Непосредственным толчком к разработке предлагаемого подхода послужило изучение работы [1], в которой (как и в [2]) распознающие алгоритмы (классификаторы) анализируются, будучи представлены лишь векторами решений на объектах обучающей выборки. Для перехода от анализа к синтезу необходимо дополнительно опереться на *принцип компактности* [3],[4], заключающийся в том, что соседние объекты выборки должны с большей вероятностью принадлежать к одному классу. Основанные на этом предположении алгоритмы будем называть *метрическими классификаторами.* Предлагается рассматривать задачу синтеза метрического классификатора как задачу *оптимальной сегментации* (labeling) точек обучающей выборки, а «форму» и «сложность» классификаторов трактовать в терминах «формы» и «сложности» изображений (образованных метками классов на точках выборки), то есть в терминах *математической морфологии.*

Основными источниками используемых далее морфологических конструкций и идей являются: *теория форм* М.Павель [5], *математическая морфология* Серра [6], *теория морфологического анализа* Пытьева [7] а также *критериальная проективная морфология* [8]. Для алгоритмической реализации процедур синтеза метрических классификаторов предлагается использовать технику построения минимальных разрезов графов [9]—[13], применяя ее к графам соседства элементов обучающей выборки.

Задача обучения с учителем

Пусть даны *пространство объектов* $\mathcal{A}$, конечное *множество классов* $C = \{c_1, \dots, c_l\}$, и известно разбиение объектов по классам $c_\mathcal{A}(a) : a \in \mathcal{A} \mapsto c \in C$. Обозначение $c_\mathcal{A}$ указывает на то, что функция определена на $\mathcal{A}$.

Производится описание объектов из $\mathcal{A}$ дескрипторами из *пространства описаний* (*признаков*) $\mathcal{X}$: $x_\mathcal{A}(a) : a \in \mathcal{A} \mapsto x \in \mathcal{X}$. Случайным образом формируется конечная *выборка объектов* $A \subseteq \mathcal{A}$, $\|A\| < +\infty$ и соответствующая *выборка описаний* $X \subseteq \mathcal{X}, \|X\| < +\infty$. Каждому значению x ставится в соответствие класс c породившего его объекта $c_X(x) : x_\mathcal{A}(a) \in X \mapsto c_\mathcal{A}(a) \in C$. По *обучающей выборке* c_X требуется построить такой *распознающий алгоритм* или *классификатор* $f_\mathcal{X}(x) : x \in \mathcal{X} \mapsto c \in C$, который обеспечивает *наилучшее разбиение* $\mathcal{X}$ на классы из C.

«Наилучшее разбиение» определим при помощи *тестовой выборки* $c'_Y(x) : x \in Y \mapsto c \in C$, $Y \subseteq \mathcal{X}$, $Y \cap X = \varnothing$, $\|X\| < +\infty$, и *критерия эмпирического риска* на выборке Y:

$J_Y(f_\mathcal{X}) = d_H(f_Y, c'_Y)/\|Y\|$,

$d_H(f_Y, c'_Y) = \sum_{x \in Y} 1(f(x) \neq c'(x))$,

где $1(true) = 1, 1(false) = 0$, $\|Y\| = \sum_{x \in Y} 1$.

Здесь *расстояние Хэмминга* d_H имеет смысл *числа ошибок классификации* на тестовой выборке Y. Отсюда *критерий среднего ожидаемого эмпирического риска* имеет вид

$$J_\mathcal{X}(f_\mathcal{X}) = E_{Y \subseteq \mathcal{X}}\{J_Y(f_\mathcal{X})\},$$

где $E_{Y \subseteq \mathcal{X}}\{.\}$ — математическое ожидание по всем возможным выборкам $Y \subseteq \mathcal{X}$.

Таким образом, задача построения *оператора оптимального синтеза* θ, заключается в минимизации критерия $J_\mathcal{X}(f_\mathcal{X}) = J_\mathcal{X}(\theta_{C_X})$:

$$\theta : c_X \in \Omega_X \mapsto f_\mathcal{X} \in \Omega_\mathcal{X},$$

$$\theta : \arg\min_{\theta'}\{J_\mathcal{X}(\theta'_{C_X})\}. \qquad (1)$$

Здесь Ω_X и $\Omega_\mathcal{X}$ — *множества всех возможных разбиений* X и $\mathcal{X}$ по классам из C.

Как правило, от *задачи синтеза* (1) сразу переходят к *задаче обучения* классификаторов заданного класса при помощи *обучающего правила* известного типа:

$$\theta \in \Theta : c_X \in \Omega_X \mapsto f_\mathcal{X} \in F_\mathcal{X} \subseteq \Omega_\mathcal{X},$$

Работа выполнена при финансовой поддержке РФФИ, проекты № 11-08-01114-а, 11-08-01039-а.

$$\theta = \arg\min_{\theta' \in \Theta} \{J_Y(\theta'{}_{C_X})\}, \qquad (2)$$

где $F_{\mathcal{X}}$ — класс классификаторов, Θ — класс алгоритмов обучения для $F_{\mathcal{X}}$ на выборках $X \subseteq \mathcal{X}$.

Кроме того, вместо недоступного критерия $J_{\mathcal{X}}(f_{\mathcal{X}})$, на практике используется критерий *наблюдаемого эмпирического риска* $J_X(\theta_{c_X})$, который имеет глобальный минимум при $f_X \equiv c_X$, непригодный для неизвестной тестовой выборки Y. Этой проблеме посвящена *теория оценки и контроля переобучения* [2]. Здесь риск оценивается по обучающей выборке, но сложность решающего правила искусственно ограничивается. Для этого вводится понятие *сложности классификатора* $Q(f_{\mathcal{X}})$, а точнее *сложности класса классификаторов* $Q(F_{\mathcal{X}})$. Соответственно вместо (2) решается задача *задача минимизации наблюдаемого риска с регуляризацией по сложности класса обучаемого классификатора*:

$$\theta \in \Theta : c_X \in \Omega_X \mapsto f_{\mathcal{X}} \in F_{\mathcal{X}} \subseteq \Omega_{\mathcal{X}},$$

$$\theta = \arg\min_{\theta' \in \Theta} \{J_X(\theta'{}_{C_X}) + \alpha Q(F_{\mathcal{X}})\}, \qquad (3)$$

где $\alpha \geqslant 0$ — *параметр регуляризации.*

Решение задачи (3) сводится к следующему (*«метод структурной минимизации риска»*). Пусть в рамках некоторого суперкласса $\mathbf{F}_{\mathcal{X}}$ определена последовательность *вложенных классов классификаторов нарастающей сложности*:

$$F^0_{\mathcal{X}} \subseteq F^1_{\mathcal{X}} \subseteq ... \subseteq F^j_{\mathcal{X}} \subseteq \mathbf{F}_{\mathcal{X}} \subseteq \Omega_{\mathcal{X}} :$$

$$Q(F^0_{\mathcal{X}}) \leqslant Q(F^1_{\mathcal{X}}) \leqslant ... \leqslant Q(F^j_{\mathcal{X}}) \leqslant ... \qquad (4)$$

Тогда задача (3) последовательно решается для $F^j_{\mathcal{X}}$, $j = 0, 1, 2, \dots$, пока значения критерия не перестанут улучшаться. Значение α подбирается *методом кросс-валидации с валидационной выборкой* $Z \subseteq \mathcal{X}$, $Y \cap X = \varnothing$, $\|Z\| < +\infty$.

Морфологический подход к синтезу классификаторов

Рассмотрим поход к машинному обучению, направленный непосредственно на решение задачи (1). Решение (1) предлагается отыскивать в виде композиции решений подзадач:

$$\theta_\alpha = \delta_\alpha \psi_\alpha, \qquad (5)$$

где ψ_α — оператор (процедура) *синтеза оптимального отклика классификатора на обучающей выборке X с учетом его сложности (локальной некомпактности)*

$$\psi_\alpha : c_X \in \Omega_X \mapsto f_X \in \Omega_X,$$

$$\psi_\alpha = \arg\min_{\psi'} \{J_X(\psi' c_X) + \alpha Q_X(\psi' c_X)\}; \qquad (6)$$

δ_α — оператор (процедура) *оптимальной корректной интерполяции (расширения)* классификатора f_X на $\mathcal{X}$ *с учетом сложности получаемого классификатора* $f_{\mathcal{X}}$:

$$\delta_\alpha : f_X \in \Omega_X \mapsto f_{\mathcal{X}} \in \Omega_{\mathcal{X}},$$

$$\delta_\alpha = \arg\min_{\delta'} \{J_{NN}(\delta' f_X) + \beta Q(\delta' f_X)\}; \qquad (7)$$

Здесь

$$J_{NN}(\delta f_X) = \begin{cases} +\infty, \text{если } \exists x \in X : \delta f_X(X) \neq f_X(X); \\ d_H(\delta f_X(X)), \delta^{NN} f_X(X)) \text{– иначе,} \end{cases}$$

d_H — расстояние Хэмминга; δ^{NN} — оператор интерполяции, соответствующий *правилу ближайшего соседа* (Nearest Neighbor).

Как решается в рамках такого подхода проблема переобучения? Ответ связан с областью *критериальной морфологии* [8], где доказано следующее

Утверждение 1. ***Если*** *ψ_α — морфологический фильтр*

$$\psi_\alpha A = \arg\min_B \{J(A, B) + \alpha Q(B)\};$$

где A — исходный образ, B — выходной образ, $J(A, B)$ — критерий ошибки аппроксимации, $Q(B)$ — критерий сложности образа B, $\alpha \geqslant 0$ — параметр регуляризации, и критерий $J(A, B)$ обладает свойствами метрики, ***То***

1) *оператор ψ_α является проектором: $\psi^2_\alpha = \psi_\alpha$;*
2) *проектор ψ_α определяет систему вложенных классов, монотонную по параметру α: $\alpha \geqslant \beta \Rightarrow \psi_\beta \psi_\alpha = \psi_\alpha$.*

Поскольку в задаче (1) функционал J имеет вид расстояния Хэмминга, из Утверждения 1 следует, что θ_α (5) является *алгебраическим проектором*:

$$\theta^2_\alpha = \theta_\alpha \Rightarrow \forall x \in X : \theta_\alpha f_X(x) = f_X(x) \qquad (8)$$

Кроме того, из Утверждения 1 следует, что на основе θ_α образуется система вложенных классов решающих правил, монотонная относительно α:

$$\forall \alpha \geqslant \beta \Rightarrow F^\alpha_{\mathcal{X}} \subseteq F^\beta_{\mathcal{X}} : Q(F^\alpha_{\mathcal{X}}) \leqslant Q(F^\beta_{\mathcal{X}}), \qquad (9)$$

где $F^\alpha_{\mathcal{X}} = \{f_X(x) : \theta_\alpha f_X(x) = f_X(x)\}$ — множество классификаторов (разбиений), стабильное относительно проектора θ_α. В морфологиях изображений такая система вложенных проективных классов рассматривается как множество Пытьевских «форм» нарастающей сложности. В задаче синтеза классификаторов последовательность «форм» может быть использована для решения проблемы переобучения методом минимизации структурного риска. Необходимо лишь определить $Q_X(f_X)$ как критерий компактности.

Оценка компактности и минимизация сложности классификаторов

Для каждого $x \in X \subseteq \mathcal{X}$ определим систему вложенных окрестностей $O_k(x) \subseteq X$,

$k = 1, \ldots, ||X|| - 1$, состоящих из k ближайших соседей. Принцип компактности предполагает, что близкие соседи должны с большей вероятностью принадлежать к одному классу. Введем локальную меру k-некомпактности f_X:

$$Q_k(x, f_X) = q_H(O_k(x))/||O_k(x)||;$$

$$q_H(O_k(x)) = \Sigma_{y \in O_k(x))} 1(f_X(x) \neq f_X(y)), \quad (10)$$

и глобальную меру k-некомпактности f_X:

$$Q_X^k(f_X) = Q_H(X, f_X)/||X||,$$

$$Q_H(X, f_X) = \Sigma_{x \in X} Q_k(x, f_X). \quad (11)$$

Значение $Q_X^k(f_X)$ (11) характеризует эмпирическую оценку вероятности того, что один из k ближайших соседей в разбиении $f_X(x)$ будет отнесен к другому классу. При любых фиксированных k и X усложнению классификатора f_X соответствует нарастание меры k -некомпактности $Q_X^k(f_X)$. Кроме того, как показано в [11], при увеличении параметра k в (11) преимущество получают более простые и «гладкие» разделяющие поверхности.

С учетом (11) задача (6) сводится к известной задаче оценивания параметров скрытой Марковской модели [14], для которой существует эффективное приближенное решение методом нахождения максимального потока / минимального разреза графа. Для случая двух классов метод разрезания графа может давать точное оптимальное решение. Алгоритм нахождения минимального разреза на графе с двумя терминальными вершинами позволяет минимизировать функционал энергии вида:

$$E(T) = E_0 + \sum_{i=1..N} E_i(t_i) + \sum_{(i,j) \in V} E_{ij}(t_i, t_j), \quad (12)$$

где N — число нетерминальных вершин графа; $T = \langle t_1, \ldots, t_N \rangle, t_1, \ldots, t_N \in \{0, 1\}$ — метки ассоциирования нетерминальных вершин с терминальными; $E_i(0), E_i(1) \in \{0, 1\}$ — унарные потенциалы; $E_{ij}(t_i, t_j)$ — парные потенциалы $E_{ij}(0,0), E_{ij}(0,1), E_{ij}(1,0), E_{ij}(1,1)$; V — подмножество пар индексов, задающее соседство на T.

Энергия (12) *субмодулярна* [13], если $\forall (i,j) \in V$:

$$E_{ij}(0,0) + E_{ij}(1,1) \leqslant E_{ij}(0,1) + E_{ij}(1,0). \quad (13)$$

Для субмодулярной энергии (12)—(13) метод минимального разрезания графа [9], [12] гарантирует нахождение точного минимума [10], [13].

Для задачи синтеза двухклассового классификатора (6), (11) примем: $C = \{0,1\}, N = ||X||, T = \langle f_X(x_1), \ldots, f_X(x_N) \rangle$, $E_i(x) = 1(f_X(x) \neq c_X(x)), E_{ij}(t_i, t_j) = 1(f_X(x_i) \neq f_X(x_j))$, $V = (i,j) : j \in O_k(x_i)$. Легко убедиться, что соответствующая энергия (12) будет субмодулярной, а значит, метод действительно порождает α-семейства проекторов из Утверждения 1.

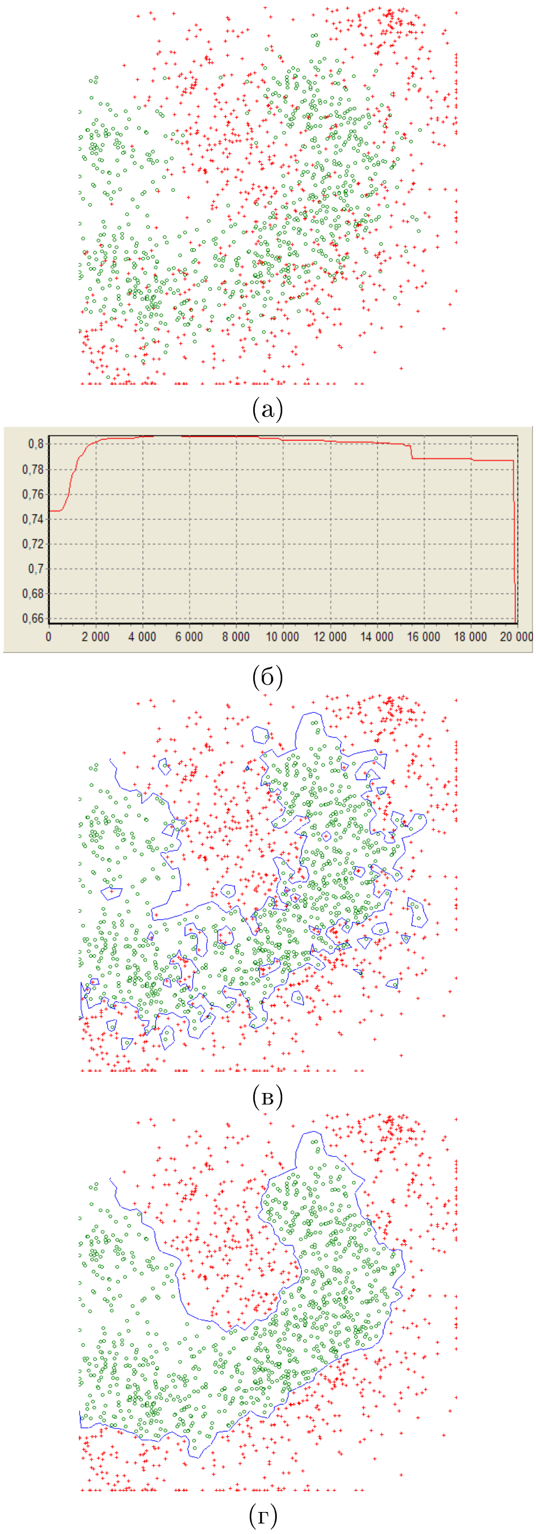
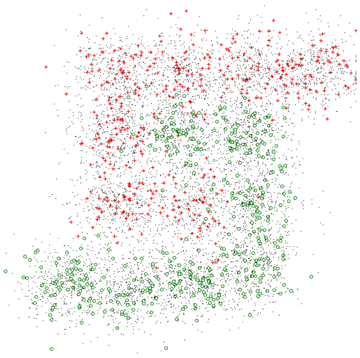


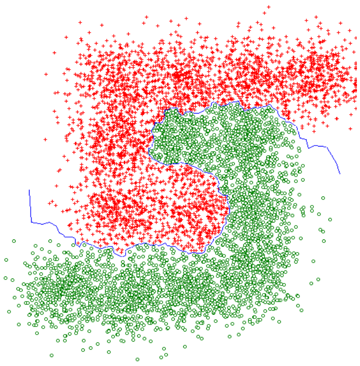
Рис. 1. Пример морфологического обучения: а) обучающая выборка; б) зависимость вероятности распознавания от параметра α; в) переразметка выборки при $\alpha = 1000$ (переобучение); г) $\alpha = 4500$ (оптимум).

Моделирование процедур морфологического синтеза классификаторов

Для наглядности моделировались двумерные данные. Выборка формировалась на основе сме-



(а)



(б)

Рис. 2. Пример обучения на частично (на 10%) размеченной обучающей выборке. Показаны: а) обучающая выбрка; б) результат переразметки выборки.

си гауссовых распределений. При построении графа соседства использовался алгоритм триангуляции Делоне с динамическим кэшированием [15]. Разрезы графов вычислялись при помощи библиотеки [16], что обеспечило время обучения не более 1 секунды. На Рис.1. показан пример морфологического обучения с учителем. Обучающая выборка содержала 1000 объектов двух классов. Приведена полученная методом Монте-Карло зависимость вероятности правильного распознавания от параметра α. На Рис.2 показан пример обучения по частично размеченной выборке – т.н. semi-supervised learning (размечено 10% из 10000 объектов).

Заключение

Предложенный подход к машинному обучению назван «*морфологическим*» в силу его алгоритмического и методического соответствия известному морфологическому подходу к анализу изображений. Дальнейшие работы в рамках предложенного подхода должны быть связаны с исследованием его поведения на многомерных модельных и реальных данных, связанных с задачами распознавания сложных образов.

Литература

[1] *Воронцов К. В.* Комбинаторная теория надёжности обучения по прецедентам. Диссертация на соискание ученой степени доктора физико-математических наук. ВЦ им. А. А. Дородницына РАН, Москва, 2010.

[2] *Вапник В. Н.* Восстановление зависимостей по эмпирическим данным. — Москва: Наука, 1979.

[3] *Айзерман М. А., Браверман Э. М., Розоноэр Л. И.* Метод потенциальных функций в теории обучения машин — Москва: Наука, 1970 — 314 с.

[4] *Хачай М. Ю.* Топологический подход к выводу условий равномерной по классу событий сходимости частот к вероятностям // Интеллектуализация обработки информации: 8-я международная конференция (ИОИ-8), Кипр, г. Пафос, 2010 г.: Сборник докладов, Москва: МАКС Пресс, , 2010, С. 91–94.

[5] *Pavel M.* Fundamentals of Pattern Recognition // Marcel Dekker. Inc., New York, 1989.

[6] *Serra J.* Image Analysis and Mathematical Morphology // Academic Press, London, 1982.

[7] *Пытьев Ю. П., Чуличков А. И.* Методы морфологического анализа изображений // Москва: Физматлит, 2010 — 336 с.

[8] *Визильтер Ю. В.* Обобщенная проективная морфология. // Компьютерная оптика — 2008. — Т. 32, № 4. — С. 384–399

[9] *L. Ford, D. Fulkerson.* Flows in Networks // Princeton University Press, 1962.

[10] *D. Greig, B. Porteous,A. Seheult.* Exact maximum a posteriori estimation for binary images // Journal of the Royal Statistical Society 1989. — Vol. 51, No. 2. — Pp. 271–279.

[11] *Y. Boykov, V. Kolmogorov.* Computing geodesics and minimal surfaces via graph cuts // In Proc. IEEE International Conf. Computer Vision (ICCV) 2003. — Pp. 26–33.

[12] *Y. Boykov, V. Kolmogorov.* An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision // IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI) 2004. — Vol. 26, No. 9. — Pp. 1124–1137.

[13] *V. Kolmogorov, R. Zabih* What energy functions can be minimized via graph cuts? // IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI) 2004. — Vol. 26, No. 2. — Pp. 147–159.

[14] *Geman S., Geman D.* Stochastic relaxation, Gibbs distributions, the Bayesian restoration of images // IEEE Trans. Pattern Analysis and Machine Intelligence 1984. — No. 6. — Pp. 721–741.

[15] *Скворцов А. В.* Обзор алгоритмов построения триангуляции Делоне // Вычислительные методы и программирование 2002. — Т. 3. — С. 14–39.

[16] *Y. Boykov, V. Kolmogorov.* MAXFLOW - software for computing mincut/maxflow in a graph. V. 3.01. — `http://www.cs.ucl.ac.uk/staff/V.Kolmogorov/software.html`