Министерство образования Российской Федерации
Московский государственный университет
им. М. В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра «Математические методы прогнозирования»

Дипломная работа

Методы предсказания информативности логических закономерностей

Работу выполнила
Студентка 517 группы
Дударенко Марина Алексеевна

Научный руководитель:
д.ф.-м.н.
Воронцов Константин Вячеславович

Москва
2012

Содержание

1 Введение **3**

2 Определения и обозначения **5**

3 Комбинаторные оценки обобщающей способности **6**
3.1 Принцип порождающих и запрещающих множеств 7
3.2 Оценки расслоения-связности 10

4 Логические алгоритмы классификации **13**
4.1 Понятие закономерности 14
4.2 Оценки расслоения-связности первого и второго рода 14
4.3 Наблюдаемый и ненаблюдаемый критерии предсказанной информативности 18
4.4 Эмпирическое обоснование перехода от ненаблюдаемых оценок к наблюдаемым 19
4.4.1 Результаты эксперимента 22

5 Заключение **26**

1 Введение

Комбинаторная теория надёжности обучения по прецедентам [3, 4] даёт общие оценки вероятности переобучения, учитывающие свойства расслоения и связности в параметрических семействах алгоритмов классификации. Эти свойства, как правило, наблюдаются при решении практических задач обучения по прецедентам, однако классические оценки [2] учитывают только простейшие размерные характеристики задачи — длину выборки и количество алгоритмов в семействе, различимых на заданной выборке.

Расслоение означает, что лишь малая часть алгоритмов из данного семейства допускают малое число ошибок, следовательно, имеют большую вероятность быть выбранными в результате обучения. Согласно комбинаторным оценкам, эти вероятности экспоненциально убывают с ростом номера слоя, то есть с ростом числа ошибок алгоритмов, составляющих слой. Таким образом, в каждой конкретной задаче эффективно используется лишь небольшое число нижних слоёв семейства алгоритмов, и комбинаторный подход позволяет учитывать это важное обстоятельство.

Связность означает, что для каждого алгоритма из данного семейства существует некоторое количество алгоритмов, отличающихся от него ровно на одном объекте. Такие алгоритмы называются *связанными*. Связность возникает в тех случаях, когда семейство алгоритмов классификации определяется разделяющей поверхностью, непрерывной по параметрам.

Для практического применения оценок расслоения–связности необходимо рассматривать конкретные семейства алгоритмов и решать комбинаторные задачи, связанные с перечислением всех алгоритмов, по-разному разделяющих заданную конечную выборку объектов.

В работах [3] была решена задача получения точных оценок вероятности переобучения для ряда достаточно простых модельных семейств. *Модельными семействами* называются семейства, задаваемые непосредственно своей матрицей ошибок «объекты–алгоритмы», при том, что множества объектов и алгоритмов в явном виде не задаются. *Реальными семействами* называются семейства, используемые на практике для решения прикладных задач классификации. В [1] были получены точные оценки вероятности переобучения для многомерных монотонных и унимодальных сетей алгоритмов — модельных семейств, обладающих основными свойствами реальных семейств — расслоением, связностью и размерностью. Показано, что точными оценками вероятности переобучения, полученными для этих модельных семейств, можно приближать вероятность переобучения реальных семейств. Этот подход позволил улучшить обобщающую способность решающих деревьев, что было подтверждено экспериментами на реальных данных.

В работах [5, 8] были получены оценки расслоения–связности для семейства логических закономерностей — конъюнкций пороговых правил над вещественными признаками. Эксперименты на 6 реальных задачах из репозитория UCI показали, что применение комбинаторных оценок позволяет добиться увеличения точности классификации контрольных данных на 1–2%. Для этого предлагалось взять за основу

любой стандартный метод поиска логических закономерностей и изменить в нём только критерий информативности правил, встроив в него процедуру оценивания обобщающей способности. При этом некоторые вопросы оставались открытыми.

Во-первых, большинство критериев информативности зависят от двух статистических характеристик правил: p — число выделяемых правилом объектов заданного класса y (число положительных примеров) и n — число выделяемых правилом объектов всех остальных классов (число отрицательных примеров). Известные оценки расслоения–связности непосредственно применимы только в случае, когда в качестве критерия информативности используется число ошибок $n + (P - p)$, где P — число объектов класса y. Однако на практике используются несколько десятков различных критериев [6], наиболее распространённые из них — энтропийный критерий, индекс Джини, критерий хи-квадрат и точный тест Фишера. В данной работе оценки расслоения–связности обобщаются на случай, когда требуется раздельно оценить переобучение для ошибок первого и второго рода, а оптимизируемым критерием в методе обучения является произвольный критерий информативности. Вводится понятие предсказанной информативности для оценки качества набора признаков (но не отдельной конъюнкции) с учётом переобучения, возникающего в результате оптимизации порогов. Оптимизация предсказанной информативности позволяет осуществлять отбор признаков. Заметим, что этот подход в точности соответствует классическому способу применения оценок обобщающей способности, широко используемому в машинном обучении для отбора моделей (model selection) и отбора признаков (features selection) [6].

Во-вторых оставался открытым вопрос, почему правомерна замена ненаблюдаемых оценок вероятности переобучения для неизвестной тестовой выборки длины K наблюдаемыми оценками, вычисленными путём разбиения наблюдаемой полной выборки длины L всеми C_L^ℓ способами на обучающую подвыборку длины ℓ и контрольную длины k. Интуитивно представляется, что характеристики расслоения и связности являются неотъемлемыми свойствами конкретной задачи, устойчивыми при изменении длины выборки. Поэтому переобученность, возникающая при разбиении полной выборки длины $L + K$ на наблюдаемую выборку длины L и скрытую длины K, оказывается с высокой точностью равной переобученности, возникающей при разбиении наблюдаемой выборки длины L на обучающую подвыборку длины ℓ и контрольную длины k. Теоретического обоснования этого факта пока не найдено. В данной работе получены его эмпирические обоснования.

Наконец, в-третьих, был реализован алгоритм эффективного вычисления оценок расслоения–связности для семейства пороговых конъюнкций, в том числе эффективного обхода классов эквивалентности конъюнкций и выбора числа нижних слоёв семейства.

Кроме того, вместо вероятности переобучения предлагается использовать функционалы полного скользящего контроля (CCV) или ожидаемой переобученности (EOF), которые гораздо эффективнее вычисляются. В данной работе получены новые оценки расслоения–связности для этих функционалов, также раздельно учитывающие ошибки первого и второго рода.

2 Определения и обозначения

Пусть имеется множество объектов $\mathbb{X}$ и множество ответов $\mathbb{Y}$. Объекты из $\mathbb{X}$ задаются набором из n признаков.

Для множеств $\mathbb{X}$ и $\mathbb{Y}$ существует *целевая функция* $y^*\colon \mathbb{X} \to \mathbb{Y}$, значения которой $y_i = y^*(x_i)$ известны на некотором подмножестве объектов множества $\mathbb{X}$: $\{x_1, \ldots, x_\ell\} \subset \mathbb{X}$.

Совокупность пар $X = (x_i, y_i)_{i=1}^{\ell}$ называется *обучающей выборкой*.

Требуется построить отображение $a\colon \mathbb{X} \to \mathbb{Y}$, совпадающее с целевой функцией y^* на обучающей выборке: $a(x_i) = y_i$, и приближающее её достаточно точно на всем множестве $\mathbb{X}$. Данное отображение называют *алгоритмом*.

Если $\mathbb{Y} = \{1, \ldots, M\}$, то восстановление целевой функции называется *задачей классификации на M непересекающихся классов*.

Для алгоритма классификации $a\colon \mathbb{X} \to \mathbb{Y}$ введем понятие *индикатора ошибки* — функции, возвращающей 1, если алгоритм a ошибается на объекте x, и 0 в противном случае:

$$I(a, x) = [a(x) \neq y^*(x)], \quad x \in \mathbb{X}.$$

Будем использовать ее в качестве функции потерь.

Введем понятие *числа ошибок* алгоритма a на множестве $X \subseteq \mathbb{X}$:

$$n(a, X) = \sum_{x \in X} I(a, x),$$

и *частоты ошибок* (или *эмпирического риска*) :

$$\nu(a, X) = \frac{1}{|X|} n(a, X).$$

Моделью алгоритмов называется параметрическое семейство отображений $A = \{\varphi(x, \gamma) | \gamma \in \Gamma\}$, где $\varphi\colon \mathbb{X} \times \Gamma \to \mathbb{Y}$ — некоторая фиксированная функция, Γ — множество допустимых значений параметра γ, называемое *пространством параметров*.

Процесс подбора параметров модели по обучающей выборке называют *обучением по прецедентам*.

Методом обучения называется отображение $\mu\colon 2^{\mathbb{X}} \to A$, которое произвольной конечной выборке $X \subseteq \mathbb{X}$ ставит в соответствие алгоритм $a\colon \mathbb{X} \to \mathbb{Y}$.

Метод минимизации эмпирического риска (МЭР) — метод обучения, при котором в заданной модели A ищется алгоритм a, доставляющий минимальное значение функционалу эмпирического риска ν на заданной обучающей выборке X:

$$\mu X \in A(X) = \operatorname{Arg}\min_{a \in A} \nu(a, X).$$

Метод минимизации эмпирического риска называется *пессимистичным*(ПМЭР), если среди всех алгоритмов, доставляющих минимум функционалу ν на обучающей

выборке X, выбирается алгоритм с наибольшим значением ν на контрольной выборке $\bar{X}$:

$$\mu X = \arg\max_{a \in A(X)} \nu(a, \bar{X}),$$

где *контрольная выборка* $\bar{X} = (x'_i, y'_i)_{i=1}^k$ — это совокупность новых объектов, не вошедших в состав обучающей выборки.

3 Комбинаторные оценки обобщающей способности

Даже если для алгоритма a функционал эмпирического риска $\nu(a, X)$ достигает минимума на обучающей выборке X, его значение $\nu(a, \bar{X})$ на произвольной контрольной выборке $\bar{X}$ может быть очень велико — это значит, что алгоритм a плохо приближает целевую зависимость y^* на выборке $\bar{X}$. Когда качество работы алгоритма на контрольной выборке оказывается существенно хуже, чем на обучающей выборке, то говорят об эффекте *переобучения*.

Переобученностью алгоритма a относительно пары выборок $X, \bar{X}$ называется разность частот его ошибок на контроле и на обучении:

$$\delta(a, X) = \nu(a, \bar{X}) - \nu(a, X).$$

Переобученностью метода обучения μ называется переобученность алгоритма $a = \mu X$:

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

Определим *вероятность переобучения* метода μ для любого $\varepsilon \in (0, 1)$ как долю разбиений $\mathbb{X} = X \bigsqcup \bar{X}$, при которых переобученность метода превышает ε:

$$Q_\varepsilon(\mu, \mathbb{X}) = \mathbf{P}\left[\delta_\mu(X) \geqslant \varepsilon\right] = \mathbf{P}\left[\nu(\mu X, \bar{X}) - \nu(\mu X, X) \geqslant \varepsilon\right] = \eta(\varepsilon), \tag{3.1}$$

где знак вероятности $\mathbf{P}$ здесь и далее следует понимать как долю разбиений множества $\mathbb{X}$ на обучение и контроль:

$$\mathbf{P} \equiv \frac{1}{C_L^\ell} \sum_{X \sqcup \bar{X} = \mathbb{X}}, \quad |X| = \ell, \quad |\bar{X}| = k, \quad |\mathbb{X}| = L.$$

Вероятность переобучения является характеристикой *обобщающей способности* метода обучения μ на выборке $\mathbb{X}$.

Если получена верхняя оценка вероятности переобучения $Q_\varepsilon \leqslant \eta(\varepsilon)$, то с ее помощью может быть оценена частота ошибок на контрольной выборке: с вероятностью $1 - \eta$ выполняется неравенство

$$\nu(a, \bar{X}) \leqslant \nu(a, X) + \varepsilon(\eta), \tag{3.2}$$

где $\varepsilon(\eta)$ — функция, обратная к $\eta(\varepsilon) = Q_\varepsilon(\mu, \mathbb{X})$.

Параметр ε называют *точностью*, а η — *надежностью* оценки.

Также для оценки частоты ошибок на контрольной выборке могут использоваться следующие функционалы.

Полным скользящим контролем (complete cross-validation, CCV) называется средняя (по всем разбиениям) частота ошибок на контрольной выборке:

$$CCV(\mu, \mathbb{X}) = \mathbf{E}\nu(\mu X, \bar{X}). \tag{3.3}$$

Ожидаемой переобученностью (expected overfitting, EOF) называется среднее (по всем разбиениям) значение переобученности метода μ:

$$EOF(\mu, \mathbb{X}) = \mathbf{E}\delta_\mu(X) = \mathbf{E}(\nu(\mu X, \bar{X}) - \nu(\mu X, X)). \tag{3.4}$$

Заметим, что $C(\mu, \mathbb{X})$ можно выразить через $EOF(\mu, \mathbb{X})$:

$$CCV(\mu, \mathbb{X}) = \mathbf{E}\nu(\mu X, \bar{X}) = \mathbf{E}\nu(\mu X, X) + EOF(\mu, \mathbb{X}).$$

Запись функционала полного скользящего контроля в такой форме даёт связь с частотой ошибок на обучающей выборке.

3.1 Принцип порождающих и запрещающих множеств

Простая гипотеза ПЗМ. *Принцип порождающих и запрещающих множеств* (ПЗМ) позволяет получать точные (не завышенные, не асимптотические) оценки обобщающей способности. Он основан на предположении, что для каждого алгоритма можно выписать необходимые и достаточные условия того, что он является результатом обучения. Если же удаётся выписать лишь необходимые условия, то получаются верхние оценки.

Гипотеза 3.1. *Множество A, выборка $\mathbb{X}$ и метод μ таковы, что для каждого алгоритма $a \in A$ можно указать пару непересекающихся подмножеств $X_a \subset \mathbb{X}$ и $X'_a \subset \mathbb{X}$, удовлетворяющую условию*

$$\big[\mu X{=}a\big] = \big[X_a \subseteq X\big]\big[X'_a \subseteq \bar{X}\big], \quad \forall X \in [\mathbb{X}]^\ell. \tag{3.5}$$

Множество X_a будем называть *порождающим*, множество X'_a — *запрещающим* для алгоритма a. Гипотеза 3.1 означает, что метод μ выбирает алгоритм a тогда и только тогда, когда в обучающей выборке X находятся все порождающие объекты и ни одного запрещающего. Все остальные объекты $\mathbb{X}\backslash X_a \backslash X'_a$ будем называть *нейтральными* для алгоритма a. Наличие или отсутствие нейтральных объектов в обучающей выборке не влияет на результат обучения.

Для произвольного $a \in A$ обозначим через L_a число нейтральных объектов, ℓ_a — число нейтральных объектов, попадающих в обучающую выборку, m_a — число ошибок алгоритма a на нейтральных объектах; $s_a(\varepsilon)$ — наибольшее число ошибок

алгоритма a на нейтральных обучающих объектах $X \setminus X_a$, при котором имеет место большое отклонение частот ошибок, $\delta(a, X) \geqslant \varepsilon$:

$$\begin{aligned} L_a &= L - |X_a| - |X'_a|; \\ \ell_a &= \ell - |X_a|; \\ m_a &= n(a, \mathbb{X}\setminus X_a \setminus X'_a); \\ s_a(\varepsilon) &= \tfrac{\ell}{L}\big(n(a,\mathbb{X}) - \varepsilon k\big) - n(a, X_a). \end{aligned}$$

Введем также обозначение для *гипергеометрической функции распределения*, которая играет важную роль в вычислении последующих оценок:

$$\mathcal{H}_L^{\ell,\,m}(z) = \sum_{s=s_0}^{\lfloor z \rfloor} h_L^{\ell,\,m}(s) = \sum_{s=s_0}^{\lfloor z \rfloor} \frac{C_m^s C_{L-m}^{\ell-s}}{C_L^\ell}, \qquad s_0 = \max\{0, m-k\}. \tag{3.6}$$

Теорема 3.1. *Если гипотеза 3.1 справедлива, то вероятность переобучения вычисляется по формуле*

$$Q_\varepsilon(\mu, \mathbb{X}) = \sum_{a \in A} P_a \mathcal{H}_{L_a}^{\ell_a,\,m_a}\left(s_a(\varepsilon)\right), \tag{3.7}$$

где $P_a = \mathsf{P}\big[\mu X = a\big] = \mathsf{P}\big[X_a \subseteq X\big]\big[X'_a \subseteq \bar{X}\big] = \dfrac{C_{L_a}^{\ell_a}}{C_L^\ell}$.

Обобщенная гипотеза ПЗМ. Гипотеза 3.1 накладывает слишком сильные ограничения на выборку $\mathbb{X}$, семейство A и метод μ. Поэтому теорему 3.1 удаётся применять лишь в некоторых специальных случаях. Рассмотрим естественное обобщение гипотезы 3.1. Предположим, что для каждого алгоритма a существуют различные варианты выделения порождающих и запрещающих множеств.

Гипотеза 3.2. *Множество A, выборка $\mathbb{X}$ и метод μ таковы, что для каждого алгоритма $a \in A$ можно указать конечное множество индексов V_a, и для каждого индекса $v \in V_a$ можно указать* порождающее *множество $X_{av} \subset \mathbb{X}$,* запрещающее *множество $X'_{av} \subset \mathbb{X}$ и коэффициент $c_{av} \in \mathbb{R}$, удовлетворяющие условиям*

$$\big[\mu X{=}a\big] = \sum_{v \in V_a} c_{av}\big[X_{av} \subseteq X\big]\big[X'_{av} \subseteq \bar{X}\big], \quad \forall X \in [\mathbb{X}]^\ell. \tag{3.8}$$

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

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

Теорема 3.2. *Если гипотеза 3.2 справедлива, то вероятность переобучения вычисляется по формуле*

$$Q_\varepsilon(\mu, \mathbb{X}) = \sum_{a\in A} \sum_{v\in V_a} c_{av} P_{av} \mathcal{H}_{L_{av}}^{\ell_{av},\, m_{av}} \left(s_{av}(\varepsilon)\right), \tag{3.9}$$

где

$$P_{av} = \mathsf{P}\big[X_{av} \subseteq X\big]\big[X'_{av} \subseteq \bar{X}\big] = \frac{C_{L_{av}}^{\ell_{av}}}{C_L^\ell}. \tag{3.10}$$

Принцип ПЗМ для CCV и EOF. Для функционалов полного скользящего контроля (3.3) и ожидаемой переобученности (3.4) принцип порождающих и запрещающих множеств также даёт точную оценкупри условии гипотезы 3.2.

Теорема 3.3. *Если гипотеза 3.2 справедлива, то значение функционала полного скользящего контроля вычисляется по формуле*

$$CCV(\mu, \mathbb{X}) = \frac{1}{k} \sum_{a\in A} \sum_{v\in V_a} c_{av} \frac{C_{L_{av}}^{\ell_{av}}}{C_L^\ell} \left(n(a, X'_{av}) + \frac{L_{av} - \ell_{av}}{L_{av}} m_{av} \right). \tag{3.11}$$

Доказательство. Запишем определения E и ν, затем подставим (3.8) согласно гипотезе 3.2 и переставим знаки суммирования:

$$\begin{aligned} CCV(\mu, \mathbb{X}) &= \frac{1}{C_L^\ell} \sum_{X\in[\mathbb{X}]^\ell} \sum_{a\in A} \big[\mu X{=}a\big] \frac{1}{k} \sum_{x_i\in\bar{X}} I(a, x_i) = \\ &= \frac{1}{k} \sum_{a\in A} \sum_{v\in V_a} c_{av} \sum_{i=1}^{L} I(a, x_i) \underbrace{\mathsf{P}\big[X_{av} \subseteq X\big]\big[X'_{av} \subseteq \bar{X}\big]\big[x_i \in \bar{X}\big]}_{p(a,v,x_i)}. \end{aligned}$$

Если $x_i \in X'_{av}$, то $\big[X'_{av} \subseteq \bar{X}\big]\big[x_i \in \bar{X}\big] = \big[X'_{av} \subseteq \bar{X}\big]$ и, согласно (3.10),

$$p(a, v, x_i) = \mathsf{P}\big[X_{av} \subseteq X\big]\big[X'_{av} \subseteq \bar{X}\big] = \frac{C_{L_{av}}^{\ell_{av}}}{C_L^\ell} = P_{av}.$$

Если $x_i \notin X'_{av}$ и $x_i \in X_{av}$, то $\big[X_{av} \subseteq X\big]\big[X'_{av} \subseteq \bar{X}\big]\big[x_i \in \bar{X}\big] = 0$.
Если же $x_i \notin X'_{av}$ и $x_i \notin X_{av}$ и, согласно (3.10),

$$p(a, v, x_i) = \mathsf{P}\big[X_{av} \subseteq X\big]\big[\{x_i\} \cup X'_{av} \subseteq \bar{X}\big] = \frac{C_{L_{av}-1}^{\ell_{av}}}{C_L^\ell} = P_{av} \frac{L_{av} - \ell_{av}}{L_{av}}.$$

Собирая вместе три случая, получим

$$p(a, v, x_i) = P_{av} \left(\big[x_i \in X'_{av}\big] + \big[x_i \notin X_{av}\big]\big[x_i \notin X'_{av}\big] \frac{L_{av} - \ell_{av}}{L_{av}} \right).$$

Подставляя это выражение в сумму по i, получаем

$$\sum_{i=1}^{L} I(a, x_i) p(a, v, x_i) = P_{av} \left(n(a, X'_{av}) + \frac{L_{av} - \ell_{av}}{L_{av}} n(a, \mathbb{X} \backslash X_{av} \backslash X'_{av}) \right),$$

откуда вытекает требуемое равенство (3.11). Теорема доказана. ∎

Следствие 3.3.1. *Если выполнена гипотеза 3.1, то значение функционала полного скользящего контроля равно*

$$CCV(\mu,\mathbb{X}) = \frac{1}{k}\sum_{a\in A}\frac{C_{L_a}^{\ell_a}}{C_L^{\ell}}\left(n(a,X'_a) + \frac{L_a-\ell_a}{L_a}m_a\right). \tag{3.12}$$

Теорема 3.4. *Если гипотеза 3.2 справедлива, то значение функционала ожидаемой переобученности вычисляется по формуле*

$$EOF(\mu,\mathbb{X}) = \sum_{a\in A}\sum_{v\in V_a} c_{av}\frac{C_{L_{av}}^{\ell_{av}}}{C_L^{\ell}}\left(\frac{1}{k}n(a,X'_{av}) - \frac{1}{\ell}n(a,X_{av}) + \left(\frac{1}{k}\frac{L_{av}-\ell_{av}}{L_{av}} - \frac{1}{\ell}\frac{\ell_{av}}{L_{av}}\right)m_{av}\right). \tag{3.13}$$

Доказательство. Используя определения $EOF(\mu,\mathbb{X})$ (3.4), E и ν, условие (3.8) гипотезы 3.2 и переставляя знаки суммирования, получим выражение:

$$\begin{aligned} EOF(\mu,\mathbb{X}) &= \frac{1}{C_L^{\ell}}\sum_{X\in[\mathbb{X}]^{\ell}}\sum_{a\in A}\big[\mu X{=}a\big]\left(\frac{1}{k}\sum_{x_i\in\bar{X}}I(a,x_i) - \frac{1}{\ell}\sum_{x_i\in X}I(a,x_i)\right) = \\ &= \sum_{a\in A}\sum_{v\in V_a} c_{av}\sum_{i=1}^{L} I(a,x_i)\underbrace{\mathsf{P}\big[X_{av}\subseteq X\big]\big[X'_{av}\subseteq\bar{X}\big]\left(\frac{1}{k}\big[x_i\in\bar{X}\big] - \frac{1}{\ell}\big[x_i\in X\big]\right)}_{p(a,v,x_i)}. \end{aligned}$$

Рассуждая аналогичным в доказательстве теоремы 3.11 образом, получим, что

$$p(a,v,x_i) = P_{av}\left(\frac{1}{k}\big[x_i\in X'_{av}\big] - \frac{1}{\ell}\big[x_i\in X_{av}\big] + \left(\frac{1}{k}\frac{L_{av}-\ell_{av}}{L_{av}} - \frac{1}{\ell}\frac{\ell_{av}}{L_{av}}\right)\big[x_i\notin X_{av}\big]\big[x_i\notin X'_{av}\big]\right).$$

Подставляя это выражение в сумму по i, получим требуемое равенство (3.13). Теорема доказана. ■

Следствие 3.4.1. *Если выполнена гипотеза 3.1, то значение функционала ожидаемой переобученности равно*

$$EOF(\mu,\mathbb{X}) = \sum_{a\in A}\frac{C_{L_a}^{\ell_a}}{C_L^{\ell}}\left(\frac{1}{k}n(a,X'_a) - \frac{1}{\ell}n(a,X_a) + \left(\frac{1}{k}\frac{L_a-\ell_a}{L_a} - \frac{1}{\ell}\frac{\ell_a}{L_a}\right)m_a\right). \tag{3.14}$$

3.2 Оценки расслоения-связности

Принцип *порождающих и запрещающих множеств* (ПЗМ) универсален, но не очень удобен в использовании. Если же наложить на множество алгоритмов дополнительные ограничения, то ПЗМ легко выписываются в терминах графа Хассе частично упорядоченного множества векторов ошибок.

Граф расслоения-связности. Определим расстояние между любыми двумя алгоритмами из семейства алгоритмов A как *расстояние Хэмминга* между их векторами ошибок: $\rho(a,b) = \sum_{i=1}^{L} |I(a,x_i) - I(b,x_i)|,\ \forall a,b \in A$.

Для множества алгоритмов A *m-ым слоем* назовем все алгоритмы, которые допускают ровно m ошибок на выборке $\mathbb{X}$: $A_m = \{a \in A \colon n(a,\mathbb{X}) = m\}$.

Будем полагать, что все векторы ошибок $\vec{a} = \big(I(a,x_i)\big)_{i=1}^{L}$, порождаемые алгоритмами a из A, попарно различны. Если это не так, то алгоритмы, соответствующие дублирующим векторам ошибок, исключим из множества A.

Введём на A естественное отношение порядка: $a \leqslant b$ тогда и только тогда, когда $I(a,x) \leqslant I(b,x)$ для всех $x \in \mathbb{X}$. Определим $a < b$ если $a \leqslant b$ и $a \neq b$.

Если $a < b$ и при этом $\rho(a,b) = 1$, то будем говорить, что a *предшествует* b и записывать $a \prec b$. Очевидно, что $n(a,\mathbb{X}) + 1 = n(b,\mathbb{X})$.

Графом расслоения–связности множества алгоритмов A называется направленный граф $\langle A, E\rangle$ с множеством ребер $E = \{(a,b) \colon a \prec b\}$.

Каждому ребру $a \prec b$ графа расслоения–связности соответствует один и только один объект $x_{ab} \in \mathbb{X}$, такой, что $I(a,x_{ab}) = 0$ и $I(b,x_{ab}) = 1$.

Граф расслоения–связности является подграфом графа транзитивной редукции отношения порядка $\leqslant$, называемого также *диаграммой Хáссе*. В графе $\langle A, E\rangle$ ребрами соединяются только алгоритмы, отличающиеся на одном объекте, тогда как в диаграмме Хассе ребрами соединяются также и алгоритмы a, b, отличающиеся более чем на одном объекте, если не существует такого $c \in A$, что $a < c < b$.

ПЗМ для монотонных методов обучения. Для получения точных оценок обобщающей способности было выписано необходимое и достаточное условие (3.8) того, что алгоритм $a \in A$ выдаётся методом μ в результате обучения по выборке $X \in [\mathbb{X}]^{\ell}$:

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

Чтобы упростить оценку можно ослабить условие до необходимого, заменив равенство неравенством, оставить по одной паре ПЗМ для каждого алгоритма и положить $c_{av} = 1$.

Гипотеза 3.3. *Множество A, выборка $\mathbb{X}$ и метод μ таковы, что для каждого алгоритма $a \in A$ можно указать* порождающее *множество $X_a \subset \mathbb{X}$ и* запрещающее *множество $X'_a \subset \mathbb{X}$, удовлетворяющие условиям*

$$\big[\mu X{=}a\big] \leqslant \big[X_a \subseteq X\big] \big[X'_a \subseteq \bar{X}\big], \quad \forall X \in [\mathbb{X}]^{\ell}. \tag{3.15}$$

Для каждого алгоритма $a \in A$ определим два множества объектов: X_a — множество объектов x_{ab}, соответствующих всевозможным рёбрам графа $(a,b) \in E$, исходящим из a:

$$X_a = \big\{x \in \mathbb{X} \;\big|\; \exists b \in A \colon a \prec b,\ I(a,x) < I(b,x)\big\}, \tag{3.16}$$

и X'_a — множество объектов $x \in \mathbb{X}$, таких, что a ошибается на x и существует лучший алгоритм $b \in A$, который не ошибается на x:

$$X'_a = \big\{x \in \mathbb{X} \;\big|\; \exists b \in A\colon b < a,\ I(b,x) < I(a,x)\big\}. \tag{3.17}$$

Лемма 3.1. *Если μ — монотонный пессимистичный метод обучения, то множество X_a (3.16) является порождающим, а множество X'_a (3.17) — запрещающим для алгоритма a в смысле гипотезы 3.3.*

Понятия связности и неполноценности алгоритма.

Верхней связностью $u(a)$(up-connectivity) алгоритма $a \in A$ будем называть число ребер графа, исходящих из вершины a:

$$u(a) = \#\big\{x_{ab} \in \mathbb{X}\colon a \prec b\big\} = |X_a|.$$

Нижней связностью $d(a)$(down-connectivity) алгоритма $a \in A$ будем называть число ребер графа, входящих в вершину a:

$$d(a) = \#\big\{x_{ba} \in \mathbb{X}\colon b \prec a\big\}.$$

Связность $u(a)$ (или $d(a)$) есть реализуемое семейством A число способов изменить алгоритм a так, чтобы он стал делать на одну ошибку больше (или меньше). Связность можно интерпретировать как число степеней свободы семейства A в локальной окрестности алгоритма $a \in A$.

Неполноценностью $q(a)$(inferiority) алгоритма $a \in A$ будем называть число объектов $x \in \mathbb{X}$, на которых алгоритм a ошибается, при том, что существует алгоритм $b \in A$, лучший, чем a (то есть $b < a$), не ошибающийся на x:

$$q(a) = |X'_a|.$$

В терминах графа расслоения–связности $q(a)$ равно числу различных объектов x_{bc}, соответствующих всевозможным рёбрам (b,c) на путях, ведущих к вершине a.

Верхние оценки расслоения–связности.

Теорема 3.5 (оценка расслоения–связности для вероятности переобучения). *Для произвольной выборки $\mathbb{X}$, произвольного монотонного метода обучения μ и произвольного $\varepsilon \in (0,1)$*

$$Q_\varepsilon(\mu,\mathbb{X}) \;\leqslant\; \sum_{a\in A} \frac{C^{\ell-u}_{L-u-q}}{C^{\ell}_{L}} \mathcal{H}^{\ell-u,\,m-q}_{L-u-q}\left(\tfrac{\ell}{L}(m-\varepsilon k)\right), \tag{3.18}$$

где $u = u(a)$ — верхняя связность, $q = q(a)$ — неполноценность, $m = n(a,\mathbb{X})$ — число ошибок алгоритма a на генеральном множестве объектов.

Гипотеза 3.4. *Пусть при справедливости гипотезы 3.3, выполняется следующее соотношение:*

$$EOF(\mu,\mathbb{X}) \;\leqslant\; \sum_{a\in A}\frac{C_{L_a}^{\ell_a}}{C_L^{\ell}}\left(\frac{1}{k}n(a,X'_a)-\frac{1}{\ell}n(a,X_a)+\left(\frac{1}{k}\frac{L_a-\ell_a}{L_a}-\frac{1}{\ell}\frac{\ell_a}{L_a}\right)m_a\right). \tag{3.19}$$

Замечание 3.1. Данное утверждение является гипотезой, так как для его строгого доказательства необходимо иметь оценки снизу для $\mathbf{E}\nu(\mu X, X)$ в определении функционала $EOF = \mathbf{E}\nu(\mu X,\bar{X}) - \mathbf{E}\nu(\mu X, X)$. Однако существует предположение, что частота ошибок на контроле в большинстве случаев превышает частоту ошибок на обучении (данное утверждение будет показано экспериментально), поэтому данная оценка справедлива почти всегда, так как оцениваемая разница частот положительна.

Теорема 3.6 (оценка расслоения-связности для EOF). *Для выборки $\mathbb{X}$ и монотонного метода обучения μ таких, что выполнена гипотеза 3.4, справедлива оценка:*

$$EOF(\mu,\mathbb{X}) \;\leqslant\; \sum_{a\in A}\frac{C_{L-u-q}^{\ell-u}}{C_L^{\ell}}\left(\frac{m}{k}-\left(\frac{1}{k}+\frac{1}{\ell}\right)\frac{(m-q)(\ell-u)}{L-u-q}\right), \tag{3.20}$$

где $u=u(a)$ — верхняя связность, $q=q(a)$ — неполноценность, $m=n(a,\mathbb{X})$ — число ошибок алгоритма a на генеральном множестве объектов.

Доказательство. Согласно лемме 3.1, для каждого алгоритма $a\in A$ множество X_a (3.16) является порождающим, а множество X'_a (3.17) — запрещающим. При этом a ошибается на всех объектах из X'_a и не ошибается на всех объектах из X_a: $|X_a|=u(a),\ n(a,X_a)=0,\ |X'_a|=n(a,X'_a)=q(a)$. Тогда общий вид оценки в гипотезе 3.4 примет вид:

$$\begin{aligned}
EOF(\mu,\mathbb{X}) \;&\leqslant\; \sum_{a\in A}\frac{C_{L_a}^{\ell_a}}{C_L^{\ell}}\left(\frac{1}{k}n(a,X'_a)-\frac{1}{\ell}n(a,X_a)+\left(\frac{1}{k}\frac{L_a-\ell_a}{L_a}-\frac{1}{\ell}\frac{\ell_a}{L_a}\right)n(a,\mathbb{X}\backslash X_a\backslash X'_a)\right) \;= \\
&=\; \sum_{a\in A}\frac{C_{L_a}^{\ell_a}}{C_L^{\ell}}\left(\frac{1}{k}n(a,X'_a)+\left(\frac{1}{k}\frac{L_a-\ell_a}{L_a}-\frac{1}{\ell}\frac{\ell_a}{L_a}\right)n(a,\mathbb{X}\backslash X'_a)\right) \;= \\
&=\; \sum_{a\in A}\frac{C_{L_a}^{\ell_a}}{C_L^{\ell}}\left(\frac{1}{k}n(a,\mathbb{X})-\left(\frac{1}{k}+\frac{1}{\ell}\right)\frac{\ell_a}{L_a}n(a,\mathbb{X}\backslash X'_a)\right)
\end{aligned}$$

Подставляя в полученную формулу выражения для $\ell_a,\ L_a,\ n(a,\mathbb{X}),\ n(a,\mathbb{X}\backslash X'_a)$, получим требуемую оценку. Теорема доказана. ■

4 Логические алгоритмы классификации

Рассмотрим задачу классификации. Допустим, что каждому объекту $x_i\in\mathbb{X}$ соответствует *правильный ответ* $y_i\in\mathbb{Y}$, где $\mathbb{Y}$ — конечное множество имён классов. Объекты описываются набором n числовых *признаков* $f_j\colon\mathbb{X}\to\mathbb{R},\ j=1,\ldots,n$.

Логические методы классификации основаны на построении композиций информативных, хорошо интерпретируемых логических закономерностей.

4.1 Понятие закономерности

Отображение вида $r\colon \mathbb{X} \to \{0,1\}$ назовем *предикатом*, определённым на множестве объектов $\mathbb{X}$. Предикат r *выделяет* объект x, если $r(x) = 1$.

Правилом будем называть предикат из некоторого фиксированного семейства предикатов R, обладающих свойством *интерпретируемости* — они достаточно просты и допускают запись на естественном языке в терминах предметной области. Эти требования формализуются в самой конструкции семейства R.

В данной работе будет рассматриваться наиболее распространённый вид правил — *пороговые конъюнкции*:

$$r(x;\theta) = \prod_{j\in J} \big[f_j(x) \lesseqgtr_j \theta^j\big], \tag{4.1}$$

где $J \subseteq \{1,\dots,n\}$ — подмножество признаков, $\theta^j \in \mathbb{R}$ — *порог* по j-му признаку, $\theta = (\theta^j)_{j\in J}$ — *вектор порогов*, $\lesseqgtr_j$ — один из знаков сравнения $\{\leqslant, \geqslant\}$.

Логической закономерностью класса $y \in \mathbb{Y}$ будем называть правило $r \in R$, выделяющее в заданной выборке $X \subseteq \mathbb{X}$ достаточно много объектов класса y и мало объектов всех остальных классов. Для формализации этого требования вводят два критерия: $p(r,X)$ — число *положительных примеров* — объектов класса y, выделяемых правилом r, и $n(r,X)$ — число *отрицательных примеров* — объектов всех остальных классов, выделяемых правилом r. Для поиска закономерностей в семействе правил R по обучающей выборке X естественно ставить задачу двухкритериальной оптимизации:

$$p(r,X) = \#\big\{x_i \in X \;\big|\; r(x_i) = 1,\ y_i = y\big\} \to \max;$$
$$n(r,X) = \#\big\{x_i \in X \;\big|\; r(x_i) = 1,\ y_i \neq y\big\} \to \min.$$

Введём также число положительных и отрицательных примеров в выборке X:

$$P(X) = \#\{x_i \in X\colon y_i = y\};$$
$$N(X) = \#\{x_i \in X\colon y_i \neq y\}.$$

На практике два критерия p, n сворачивают в один *эвристический критерий информативности* $\mathscr{H}(p,n) \to \max$, то есть лучшей считается та закономерность, у которой значение $\mathscr{H}(p,n)$ выше.

4.2 Оценки расслоения-связности первого и второго рода

Хороший алгоритм $a(x)$ можно построить только из непереобученных закономерностей, имеющих высокую информативность как на обучающей, так и на контрольной выборке.

Нашей целью будет построение такого критерия $\mathscr{H}'$, который предсказывал бы значение заданного стандартного критерия $\mathscr{H}$ на скрытой выборке после оптимизации порогов θ^j. *Предсказанную информативность* $\mathscr{H}'$ предполагается затем использовать в качестве критерия выбора подмножеств признаков J.

Введём для фиксированного класса $y \in \mathbb{Y}$ на множестве правил R индикатор ошибки

$$I(r, x_i) = \big[r(x_i) \neq [y_i{=}y]\big], \quad r \in R, \quad x_i \in \mathbb{X}. \tag{4.2}$$

Теорема 4.1. *Если функция $\mathscr{H}(p, n)$ строго монотонно возрастает по p и строго монотонно убывает по n, то критерий $M(r, X) = -\mathscr{H}\big(p(r, X), n(r, X)\big)$ является монотонным относительно индикатора ошибки* (4.2)*, и, соответственно, метод максимизации информативности является монотонным методом обучения правил.*

Число ошибок правила r относительно индикатора ошибки (4.2) договоримся обозначать буквой m вместо обычного n, чтобы не путать с числом отрицательных примеров:

$$m(r, X) = \sum_{x \in X} I(r, x) \;=\; n(r, X) + P(X) - p(r, X).$$

В теореме 3.5была выписана оценка расслоения-связности, обращая которую можно получить оценку частоты ошибок на контрольной выборке. Однако для логических алгоритмов классификации большее значение имеет не количество ошибок, а количество положительных ($p(r, \bar{X})$) и отрицательных ($n(r, \bar{X})$) примеров. В общем случае простой взаимосвязи между критерием $\mathscr{H}$ и частотой ошибок ν нет, и нужно раздельно оценивать $p(r, \bar{X})$ снизу и $n(r, \bar{X})$ сверху.

Ошибки I и II рода. Если r — закономерность класса y, то невыделение объекта класса y называют «пропуском цели» или *ошибкой первого рода*, а выделение объекта чужого класса — «ложной тревогой» или *ошибкой второго рода*. Генеральная выборка разбивается на два подмножества $\mathbb{X}' = \{x_i \in \mathbb{X} \colon y_i = y\}$ и $\mathbb{X}'' = \{x_i \in \mathbb{X} \colon y_i \neq y\}$. Таким образом, возникают ещё два определения индикатора ошибки:

$$\begin{aligned} I'(r, x_i) \;&=\; \big[r(x_i) = 0\big]\big[y_i = y\big] \;=\; I(r, x_i)\big[x_i \in \mathbb{X}'\big]; \\ I''(r, x_i) \;&=\; \big[r(x_i) = 1\big]\big[y_i \neq y\big] \;=\; I(r, x_i)\big[x_i \in \mathbb{X}''\big]; \end{aligned}$$

и, соответственно, два определения числа и частоты ошибок:

$$\begin{aligned} m'(r, X) &= P(X) - p(r, X); & \nu'(r, X) &= m'(r, X)/|X|; \\ m''(r, X) &= n(r, X); & \nu''(r, X) &= m''(r, X)/|X|. \end{aligned}$$

В силу тождества $I'(r, x_i) + I''(r, x_i) = I(r, x_i)$ имеет место разложение общего числа ошибок и общей частоты ошибок на ошибки I и II рода:

$$m'(r, X) + m''(r, X) = m(r, X); \qquad \nu'(r, X) + \nu''(r, X) = \nu(r, X).$$

Оценка расслоения–связности для ошибок I и II рода функционала вероятности переобучения.

Теорема 4.2 (оценка расслоения–связности для ошибок I и II рода). *Для произвольной выборки $\mathbb{X}$, произвольного монотонного метода обучения μ и произвольного $\varepsilon \in (0,1)$*

$$Q'_\varepsilon = \mathsf{P}\big[\nu'(r,\bar{X}) - \nu'(r,X) \geqslant \varepsilon\big] \;\leqslant\; \sum_{r\in R} \frac{C^{\ell-u}_{L-u-q}}{C^{\ell}_{L}} \mathcal{H}^{\ell-u,\,m'-q'}_{L-u-q}\big(\tfrac{\ell}{L}(m'-\varepsilon k)\big)\,; \qquad (4.3)$$

$$Q''_\varepsilon = \mathsf{P}\big[\nu''(r,\bar{X}) - \nu''(r,X) \geqslant \varepsilon\big] \;\leqslant\; \sum_{r\in R} \frac{C^{\ell-u}_{L-u-q}}{C^{\ell}_{L}} \mathcal{H}^{\ell-u,\,m''-q''}_{L-u-q}\big(\tfrac{\ell}{L}(m''-\varepsilon k)\big)\,; \qquad (4.4)$$

где $u = |X_r|$ — верхняя связность правила r, $q = |X'_r|$, $q' = |X'_r \cap \mathbb{X}'|$, $q'' = |X'_r \cap \mathbb{X}''|$ — неполноценность правила r относительно индикаторов ошибки I, I', I'' соответственно, $m' = m'(r,\mathbb{X})$, $m'' = m''(r,\mathbb{X})$ — число ошибок правила r на генеральной выборке относительно индикаторов ошибки I', I'' соответственно.

Доказательство. Рассмотрим функционал Q'_ε. Введём в выражении $\mathsf{P}\big[\delta'_\mu(x) \geqslant \varepsilon\big]$ под знак суммирования по X ещё два вспомогательных суммирования: первый — по всем правилам r из R при условии $\mu X = r$, второй — по всем значениям s числа ошибок I рода правила r на подвыборке $X\backslash X_r$. Значение Q'_ε от этого не изменится:

$$Q'_\varepsilon = \mathsf{P}\big[\delta'_\mu(X) \geqslant \varepsilon\big] = \mathsf{P}\sum_{r\in R}\big[\mu X = r\big]\sum_{s=0}^{\ell_r}\big[m'(r,X\backslash X_r) = s\big]\big[\delta'(r,X) \geqslant \varepsilon\big]. \qquad (4.5)$$

Для каждого правила $r \in R$ множества X_r и X'_r задаются леммой 3.1, при этом $m(r,X_r) = m'(r,X_r) = 0$, $|X_r| = u, |X\backslash X_r| = \ell - u = \ell_r, |\mathbb{X}\backslash X_r \backslash X'_r| = L - u - q = L_r$. Тогда отклонение частот ошибок I рода выражается в виде

$$\delta'(r,X) = \frac{m'(r,\mathbb{X}) - s - m'(r,X_r)}{k} - \frac{s + m'(r,X_r)}{\ell} = \frac{m'(r,\mathbb{X}) - s}{k} - \frac{s}{\ell}$$

следовательно,

$$\big[\delta'(r,X) \geqslant \varepsilon\big] = \big[s \leqslant \tfrac{\ell}{L}\big(m'(r,\mathbb{X}) - \varepsilon k\big) - m'(r,X_r)\big] = \big[s \leqslant \tfrac{\ell}{L}\big(m'(r,\mathbb{X}) - \varepsilon k\big)\big] = \big[s \leqslant s'_r(\varepsilon)\big].$$

Подставим полученное выражение в (4.5), затем заменим $\big[\mu X = r\big]$ правой частью неравенства (3.15) и переставим знаки суммирования (очевидно, P также можно рассматривать как суммирование):

$$Q'_\varepsilon \;\leqslant\; \sum_{r\in R}\sum_{s=0}^{\ell_r} \underbrace{\mathsf{P}\big[X_r \subseteq X\big]\big[X'_r \subseteq \bar{X}\big]\big[m'(r,X\backslash X_r) = s\big]}_{M'(r)}\big[s \leqslant s'_r(\varepsilon)\big]. \qquad (4.6)$$

Выделенное в данной формуле выражение $M'(r)$ есть доля разбиений генеральной выборки $\mathbb{X} = X \sqcup \bar{X}$ таких, что множество объектов X_r целиком лежит в X,

множество объектов X'_r целиком лежит в $\bar{X}$, и в подвыборку $X\backslash X_r$ длины ℓ_r попадает ровно s объектов, на которых правило r допускает ошибку I рода.

Число ошибок I рода правила r на объектах, не попадающих ни в X_r, ни в X'_r, равно $m'_r = m'(r, \mathbb{X}\backslash X_r\backslash X'_r) = m'(r, \mathbb{X}) - m'(r, X'_r \cap \mathbb{X}')$. Существует $C^s_{m'_r}$ способов выбрать из них s объектов, которые попадут в $X\backslash X_r$. Для каждого из этих способов имеется ровно $C^{\ell_r - s}_{L_r - m'_r}$ способов выбрать $\ell_r - s$ объектов, на которых правило r не допускает ошибку, и которые также попадут в $X\backslash X_r$. Тем самым однозначно определяется состав выборки $X\backslash X_r$, а, значит, и состав выборки $\bar{X}\backslash X'_r$. Таким образом, $M'(r) = C^s_{m'_r} C^{\ell_r - s}_{L_r - m'_r} / C^\ell_L$. Подставим это выражение в (4.6) и выделим в нём формулу гипергеометрической функции вероятности:

$$Q'_\varepsilon \leqslant \sum_{r\in R} \frac{C^{\ell_r}_{L_r}}{C^\ell_L} \sum_{s=s_0}^{\ell_r} \big[s \leqslant s'_r(\varepsilon)\big] \frac{C^s_{m'_r} C^{\ell_r - s}_{L_r - m'_r}}{C^{\ell_r}_{L_r}} = \sum_{r\in R} \frac{C^{\ell_r}_{L_r}}{C^\ell_L} \mathcal{H}^{\ell_r,\, m'_r}_{L_r}\left(s'_r(\varepsilon)\right).$$

С учётом того, что $L_r = L - u - q,\ \ell_r = \ell - u,\ s'_r(\varepsilon) = \frac{\ell}{L}\big(m' - \varepsilon k\big),\ m'_r = m' - q'$, где $m' = m'(r, \mathbb{X}),\ q' = |X'_r \cap \mathbb{X}'|$, получаем оценку из условия теоремы:

$$Q'_\varepsilon \leqslant \sum_{r\in R} \frac{C^{\ell-u}_{L-u-q}}{C^\ell_L} \mathcal{H}^{\ell-u,\, m'-q'}_{L-u-q}\left(\tfrac{\ell}{L}\big(m' - \varepsilon k\big)\right).$$

Аналогично выводится оценка для ошибки II рода с учётом соотношений $s''_r(\varepsilon) = \frac{\ell}{L}\big(m'' - \varepsilon k\big),\ m''_r = m'' - q''$, где $m'' = m''(r, \mathbb{X}),\ q'' = |X'_r \cap \mathbb{X}''|$.

Теорема доказана. ∎

Замечание 4.1. Отметим, что вероятности получения правила $r \in R$ в функционалах $Q_\varepsilon, Q'_\varepsilon, Q''_\varepsilon$ совпадают. Это произошло потому, что использовался один и тот же метод обучения, минимизирующий общую ошибку, порождающие и запрещающие множества одинаковые для всех функционалов. Различия появляются из-за использования разных индикаторов ошибки, следовательно, различных вероятностей переобучения правила r относительно заданного типа ошибки.

Обозначим правые части неравенств (4.3) и (4.4) через $\eta'_{\ell,k}(\varepsilon)$ и $\eta''_{\ell,k}(\varepsilon)$ соответственно, а обратные к ним функции — через $\varepsilon'_{\ell,k}(\eta)$ и $\varepsilon''_{\ell,k}(\eta)$. Тогда с вероятностью не менее $(1-\eta)$ справедливы оценки

$$\nu'(r, \bar{X}) \leqslant \nu'(r, X) + \varepsilon'_{\ell,k}(\eta),$$
$$\nu''(r, \bar{X}) \leqslant \nu''(r, X) + \varepsilon''_{\ell,k}(\eta).$$

Возвращаясь от частот ошибок ν', ν'' к обозначениям p, n, получим, с вероятностью не менее $1 - \eta$,

$$p(r, \bar{X}) \geqslant k\big(p(r, X)/\ell - \varepsilon'_{\ell,k}(\eta) - \delta\big),$$
$$n(r, \bar{X}) \leqslant k\big(n(r, X)/\ell + \varepsilon''_{\ell,k}(\eta)\big),$$

где $\delta = \frac{1}{\ell}P(X) - \frac{1}{k}P(\bar{X})$ — поправка на нестратифицированность классов, которая равна нулю, если доли положительных примеров в обучении и контроле одинаковы.

Подставляя правые части этих оценок в заданный стандартный критерий информативности $\mathscr{H}$, получим *критерий предсказанной информативности*:

$$\mathscr{H}'(p,n) = \mathscr{H}\Big(k\big(p/\ell - \varepsilon'_{\ell,k}(\eta) - \delta\big),\ k\big(n/\ell + \varepsilon''_{\ell,k}(\eta)\big)\Big). \tag{4.7}$$

Оценка расслоения–связности для ошибок I и II рода функционала ожидаемой переобученности. Аналогичные оценки можно получить для функционала ожидаемой переобученности EOF.

Теорема 4.3 (оценка расслоения–связности для ошибок I и II рода для EOF). *При условии выполнения гипотезы 3.4*

$$EOF'(\mu,\mathbb{X}) \ \leqslant\ \sum_{a\in A} \frac{C^{\ell-u}_{L-u-q}}{C^{\ell}_{L}} \left(\frac{m'}{k} - \left(\frac{1}{k}+\frac{1}{\ell}\right) \frac{(m'-q')(\ell-u)}{L-u-q} \right), \tag{4.8}$$

$$EOF''(\mu,\mathbb{X}) \ \leqslant\ \sum_{a\in A} \frac{C^{\ell-u}_{L-u-q}}{C^{\ell}_{L}} \left(\frac{m''}{k} - \left(\frac{1}{k}+\frac{1}{\ell}\right) \frac{(m''-q'')(\ell-u)}{L-u-q} \right), \tag{4.9}$$

где $u = |X_r|$ — верхняя связность правила r, $q = |X'_r|$, $q' = |X'_r \cap \mathbb{X}'|$, $q'' = |X'_r \cap \mathbb{X}''|$ — неполноценность правила r относительно индикаторов ошибки I, I', I'' соответственно, $m' = m'(r,\mathbb{X})$, $m'' = m''(r,\mathbb{X})$ — число ошибок правила r на генеральной выборке относительно индикаторов ошибки I', I'' соответственно.

Доказательство. Данные оценки следуют непосредственно из общей оценки 3.20 при замене индикатора ошибки и связанных с ним понятий ($m \to m'(m'')$, $q \to q'(q'')$), а так же из замечания 4.1 к теореме 4.2. ■

Повторив рассуждения, аналогичные приведённым выше для вероятности переобучения и величины $\varepsilon_{\ell k}(\eta)$, в среднем получим:

$$p(r,\bar{X}) \geqslant k\big(p(r,X)/\ell - EOF' - \delta\big),$$
$$n(r,\bar{X}) \leqslant k\big(n(r,X)/\ell + EOF''\big),$$

откуда следует аналогичный (4.7) *критерий предсказанной по EOF информативности*:

$$\mathscr{H}'(p,n) = \mathscr{H}\Big(k\big(p/\ell - EOF'_{\ell k} - \delta\big),\ k\big(n/\ell + EOF''_{\ell k}\big)\Big). \tag{4.10}$$

4.3 Наблюдаемый и ненаблюдаемый критерии предсказанной информативности

Переход от ненаблюдаемой оценки к наблюдаемой. Функции $\varepsilon'_{\ell,k}$, $\varepsilon''_{\ell,k}$ предсказывают число ошибок I и II рода на контрольной выборке длины k по числу ошибок I и II рода на обучающей выборке длины ℓ и графу расслоения–связности. При этом для построения графа используется полная выборка $\mathbb{X}$ с известными классификациями y_i всех объектов $x_i \in \mathbb{X}$. Значит, выборка $\mathbb{X}$ всё же предполагается

наблюдаемой. Хотелось бы использовать все имеющиеся данные $\mathbb{X}$ для обучения закономерностей, а информативность предсказывать для некоторой неизвестной выборки $\bar{\mathbb{X}}$.

Допустим, что реализовалось одно из равновероятных разбиений *супервыборки* $\mathscr{X} = \mathbb{X} \sqcup \bar{\mathbb{X}}$ длины $L + K$ на наблюдаемую $\mathbb{X}$ длины L и скрытую $\bar{\mathbb{X}}$ длины K. Если бы мы знали скрытую выборку, то могли бы, повторив в точности все выкладки, построить функции $\varepsilon'_{L,K}$, $\varepsilon''_{L,K}$ для предсказания информативности на скрытой выборке $\bar{\mathbb{X}}$ по наблюдаемым значениям $p = p(r, \mathbb{X})$ и $n = n(r, \mathbb{X})$. По аналогии с (4.7) имеем *ненаблюдаемый (unobservable) критерий предсказанной информативности*:

$$\mathscr{H}'_{\text{un}}(p, n) = \mathscr{H}\Big(K\big(p/L - \varepsilon'_{L,K}(\eta) - \delta\big),\ K\big(n/L + \varepsilon''_{L,K}(\eta)\big)\Big). \tag{4.11}$$

Однако мы не можем вычислить величины $\varepsilon'_{L,K}$, $\varepsilon''_{L,K}$, поскольку выборка $\bar{\mathbb{X}}$ неизвестна. Заменим их наблюдаемыми $\varepsilon'_{\ell,k}$, $\varepsilon''_{\ell,k}$ и пренебрежём поправкой δ. Получим *наблюдаемый (observable) критерий предсказанной информативности*:

$$\mathscr{H}'_{\text{ob}}(p, n) = \mathscr{H}\Big(K\big(p/L - \varepsilon'_{\ell,k}(\eta)\big),\ K\big(n/L + \varepsilon''_{\ell,k}(\eta)\big)\Big). \tag{4.12}$$

4.4 Эмпирическое обоснование перехода от ненаблюдаемых оценок к наблюдаемым

Для проверки гипотезы о возможности замены ненаблюдаемых оценок наблюдаемыми, описанной в параграфе 4.3, был проведен следующий эксперимент(Алгоритм 4.1). Он основывается на вычислении коэффициента линейной корреляции Пирсона для наблюдаемого и ненаблюдаемого критериев предсказанной информативности.

Для каждого S в пределах рассматриваемого диапазона от S_{min} до S_{max} генерировались супервыборки $\mathscr{X}$ длины S (шаги 7, 8), на которые накладывался шум, то есть классификация некоторых объектов выборки инвертировалась относительно чистой закономерности. Уровень шума варьировался от 0% до 50% объектов. Менялось также распределение шума — граничный (вблизи границы классов), равномерный (по всей области), периферийный (вдали от границы классов). Для генерации зашумленной выборки использовался алгоритм 4.2. При этом на шагах 4-8 и 12-15 объекты создавались и выбирались так, чтобы количество пложительных и отрицательных примеров в выборке было одинаково. Параметр λ, отвечающий за позицию шума, может быть любым положительным вещественным числом. В данной работе он полагался равным 0.95 для генерации граничного шума, 1 — равномерного и 1.05 — периферийного.

Из каждой супервыборки многократно выбиралась обучающая подвыборка $\mathbb{X}$ длины L, $L = K$ (шаги 11, 12). (Значения были подобраны так, чтобы всего в результате получилось 40 разбиений, так как в этом случае удобно вычислять 95%-ный доверительный интервал для оцениваемой величины, в нашем случае — корреляции). Для метода обучения μ и множеств $\mathscr{X}$ и $\mathbb{X}$ вычислялись оценки расслоения-связности

Алгоритм 4.1. Эксперимент для проверки близости наблюдаемого и ненаблюдаемого критериев.

Вход:

n — размерность задачи, $noisePct$ — процент зашумления выборки, $noisePos$ — позиция шума, μ — метод обучения, $\mathscr{H}$ — критерий информативности.

Выход:

график корреляции.

1: $numberSuper = 10$;
2: $numberTrain = 4$;
3: **для всех** S от S_{min} до S_{max} с шагом S_{step}
4: $L = 0.5 * S$;
5: $\ell = 0.5 * L$;
6: $Corr_S = \emptyset$;
*п*олные выборки
7: **для всех** $k = 1..numberSuper$
8: $\mathscr{X} =$ СгенерироватьСупервыборку$(n, S, noisePct, noisePos)$;
9: $Q_S =$ ВычислитьОценкиРасслоенияСвязности-I-II-рода$(\mu, \mathscr{X})$;
10: $\{\varepsilon'_{LK}, \varepsilon''_L K\} = Q_S(\eta = 0.5)$;
*в*ыделение наблюдаемой выборки
11: **для всех** $j = 1..numberTrain$
12: $\mathbb{X} =$ СгенерироватьПодвыборку$(\mathscr{X}, L)$;
13: $Q_L =$ ВычислитьОценкиРасслоенияСвязности-I-II-рода$(\mu, \mathbb{X})$;
14: $\{\varepsilon'_{\ell k}, \varepsilon''_{\ell k}\} = Q_L(\eta = 0.5)$;
*в*ычисление корреляции
15: $H_{un} = \emptyset,\ H_{ob} = \emptyset$;
16: **для всех** $d = 1..100$
*в*ыделение ненаблюдаемой выборки
17: $\bar{\mathbb{X}} =$ СгенерироватьПодвыборку$(\mathscr{X}, L)$;
18: $R =$ НайтиЛучшие Правила$(\bar{\mathbb{X}})$;
19: **для всех** правил r из множества лучших правил R
20: $h_{un} =$ ВычислитьЗначениеКритерия$(r, \mathscr{H}, \varepsilon'_{LK}, \varepsilon''_{LK})$;
21: $h_{ob} =$ ВычислитьЗначениеКритерия$(r, \mathscr{H}, \varepsilon'_{\ell k}, \varepsilon''_{\ell k})$;
22: $H_{un} =$ Добавить(h_{un});
23: $H_{ob} =$ Добавить(h_{ob});
24: $corr =$ ВычислитьКорреляцию(H_{un}, H_{ob});
25: $Corr_S =$ Добавить$(corr)$;
26: ОтобразитьГрафикКорреляции$(Corr_S)$;

I и II рода, которые затем обращались при $\eta = 0.5$ (9, 10 и 13, 14). То есть использовались *медианные оценки*. Таким образом, получили две пары величин: $\{\varepsilon'_{LK}, \varepsilon''_{LK}\}$ и $\{\varepsilon'_{\ell k}, \varepsilon''_{\ell k}\}$.

Алгоритм 4.2. Генерация зашумленных данных.

Вход:

n — размерность задачи, L — мощность выборки, $pct \in [0,1]$ — процент зашумления выборки, $\lambda \in \mathbb{R}^+$ — позиция шума.

Выход:

$\mathbb{X}$ — зашумленная выборка.

---

*г*енерируем точку, задающую исходное правило

1: **для всех** $j = 1..n$

2: $\quad a^j = rand([0,1])$;

*г*енерируем незашумленную закономерность с четкой границей

3: $\mathbb{X} = \varnothing$;

4: **для всех** $i = 1..L$

5: $\quad$ **для всех** $j = 1..n$

6: $\quad\quad x_i^j = rand([0,1])$;

7: $\quad class(x_i) = \bigwedge_{j=1}^{n}[x_i^j < a^j]$;

8: $\quad$ Добавить$(\mathbb{X}, x_i)$;

*у*порядочиваем объекты по близости к границе классов a

9: $XSort = sort(\mathbb{X}, a)$

*г*енерируем распределение, из которого будут выбираться объекты для шума

*р*азбиваем отрезок $[0,1]$ на L непересекающихся сегментов δ_i

10: **для всех** $i = 1..L$

11: $\quad$ длина $(\delta_i) = \frac{\lambda^i}{\sum_{k=1}^{L}\lambda^k}$;

*в*ыбираем объекты, метка класса которых изменится, согласно распределению и близости к границе

12: **для всех** $m = 1..L * pct$

13: $\quad p = rand([0,1])$;

14: $\quad k = \sum_{i=1}^{L} i * \mathbb{I}(p \in \delta_i)$;

15: $\quad$ Инвертировать метку класса k-ого объекта в выборке $XSort$;

16: $\mathbb{X} = XSort$;

Теперь нужно проверить согласованность наблюдаемого и ненаблюдаемого критериев. Для этого был предложен следующий способ(шаги алгоритма с 16 по 24). Из супервыборки генерируется реализация «ненаблюдаемой» подвыборки (шаг 17). Она не является ненаблюдаемой в полной мере, так как хотя величины $\varepsilon'_{\ell k}, \varepsilon''_{\ell k}$ были вычислены по другой, наблюдаемой, подвыборке, две этих выборки могут иметь общие объекты. Для каждой «ненаблюдаемой» подвыборки ищутся лучшие правила с помощью метода обучения μ (шаг 18). Для этих правил вычисляется два критерия предсказанной информативности с поправками, полученными по супервыборке и по наблюдаемой подвыборке (шаги 19-21). После получения 100 различных случайных реализаций «ненаблюдаемой» подвыборки, находим корреляцию значений предсказанных критериев информативности (шаг 24).

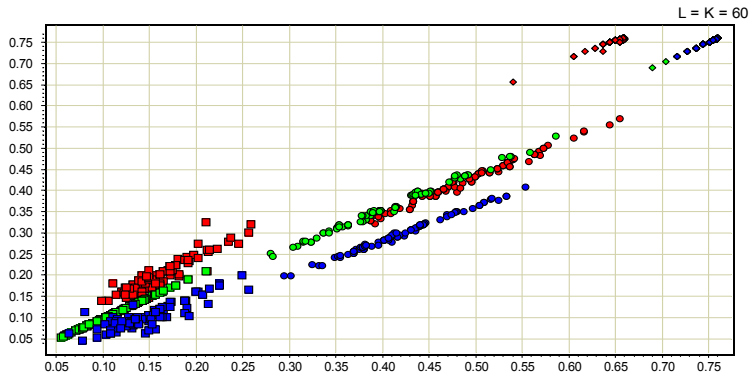

Рис. 1: Зависимость $\mathscr{H}'_{\text{un}}$ от $\mathscr{H}'_{\text{ob}}$ при $L = K = 60$ для двумерных модельных выборок при различном уровне шума ($\diamond$ — 0%, $\circ$ — 10%, $\square$ — 50%) и различном распределении шумовых объектов (синий — на границе классов, зелёный — равномерно, красный — вдали от границы классов). Точки соответствуют разбиениям супер-выборки на наблюдаемую и скрытую.

Если корреляция стремится к 1, значит значения критериев согласованы, они принимают большее значение на более хороших правилах и меньшие значения на плохих. В этом случае переход от ненаблюдаемой оценки к наблюдаемой будет обоснованным. Если же корреляция будет ближе к 0, значит данный переход делать нельзя.

В представленном эксперименте есть несколько входных параметров:
n — размерность задачи,
$noisePct$ — процент зашумления выборки,
$noisePos$ — позиция шума.
Крайне интересным представляется исследование зависимости корреляции критериев при варьировании значений данных параметров.

Кроме того, аналогичный эксперимент был проведен с поправками, которые вычислялись с помощью функционала ожидаемой переобученности EOF I и II рода, то есть вместо шагов 9, 10 будет шаг
9-10: $\{EOF'_{LK}, EOF''_{LK}\} = \text{ВычислитьОценкиРасслоенияСвязностиEOF-I-II-рода}(\mu, \mathscr{X})$;
а вместо 13, 14 —
13-14: $\{EOF'_{\ell k}, EOF''_{\ell k}\} = \text{ВычислитьОценкиРасслоенияСвязностиEOF-I-II-рода}(\mu, \mathbb{X})$.

4.4.1 Результаты эксперимента

Рис. 1 показывает, что зависимость ненаблюдаемой информативности от наблюдаемой близка к линейной, даже несмотря на малый объём данных $L = K = 60$.

В ходе проведения экспериментов было выяснено, что сходимость корреляции к 1 незначительно зависит от позиции шума. Однако все же при его локации на границе сходимость самая плохая. Поэтому все графики в данной работе приведены для такого типа шума, как иллюстрация самого худшего из возможных случаев. На рисунках 2, 3 показаны графики зависимости корреляции наблюдаемого и ненаблюдаемого критериев информативности с поправками, посчитанными по функционалам Q_{eps} и EOF соответственно. Вычислениия проводились на двумерной выборке

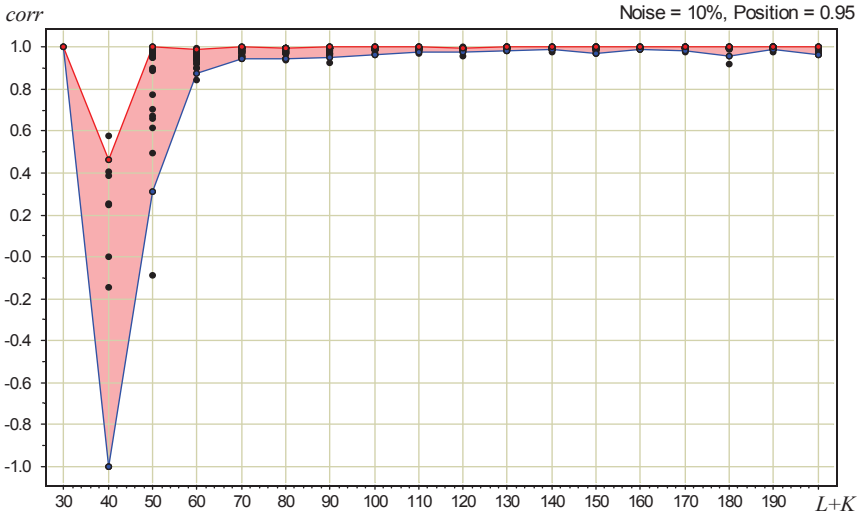

Рис. 2: 95%-ный доверительный интервал зависимости корреляции наблюдаемого и ненаблюдаемого критериев информативности от длины супервыборки, скорректированных по функционалу вероятности переобучения Q_{eps}. Вычисления проводились на двумерных выборках с 10%-ным граничным шумом.

с 10%-ным граничным шумом. Из этих графиков видно, что корреляция достаточно быстро сходится к 1 в обоих случаях, быть может, несколько быстрее для EOF.

На рисунке 4 показана связь поправок, вычисленных при обращении оценок Q_{eps} и посчитанных по EOF для двумерной выборки из 100 объектов с 10%-ным граничным шумом. Видно, что в среднем их значения примерно равны. При увеличении длины выборки, уменьшении процента шума либо изменении его позиции на более благоприятную (периферийный либо равномерный) поправки по EOF становятся меньше поправок по Q_{eps} (опять же в среднем). С учетом того, что вычислять EOF быстрее, предполагается, что использование поправок, посчитанных по EOF, будет давать более хороший результат при отборе закономерностей.

Также исследовалась сходимость корреляции при увеличении процента шума и размерности правил. На рисунке 5 показан 95%-ный доверительный интервал корреляции для двумерных выборок с 20% шума (поправки посчитаны по EOF). По сравнению с 10% (рис. 3) требуется большая длина выборки для выхода на постоянный уровень. При увеличении размерности и том же уровне шума (рис. 6, трехмерные выборки с 10%) увеличивается доверительный интервал для корреляции.

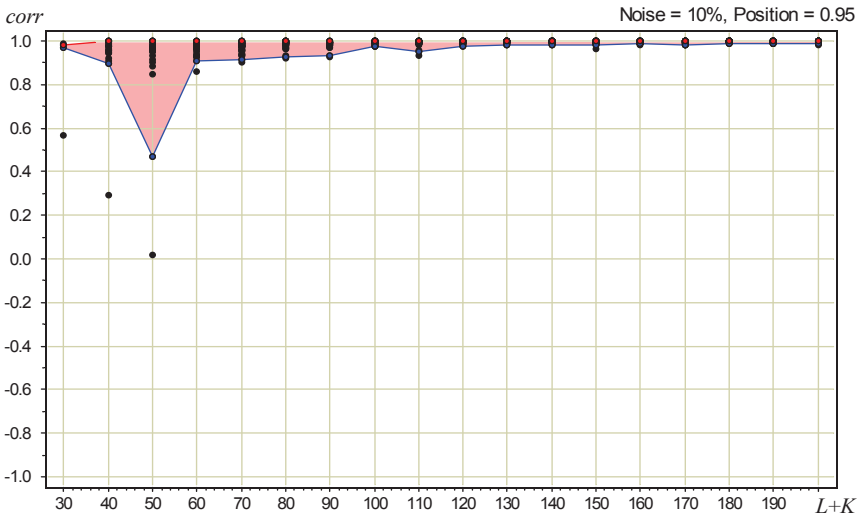

Рис. 3: 95%-ный доверительный интервал зависимости корреляции наблюдаемого и ненаблюдаемого критериев информативности от длины супервыборки, скорректированных по функционалу ожидаемой переобученности EOF. Вычисления проводились на двумерных выборках с 10%-ным граничным шумом.

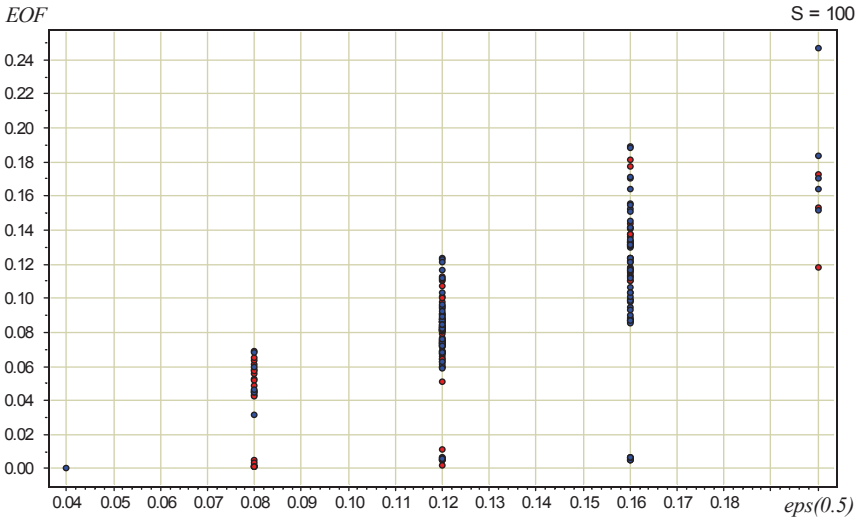

Рис. 4: Связь поправок, вычисленных при обращении оценок Q_{eps} и посчитанных по EOF. Вычисления проводились на двумерной выборке с 10%-ным граничным шумом. Длина супервыборки равна 100 объектам.

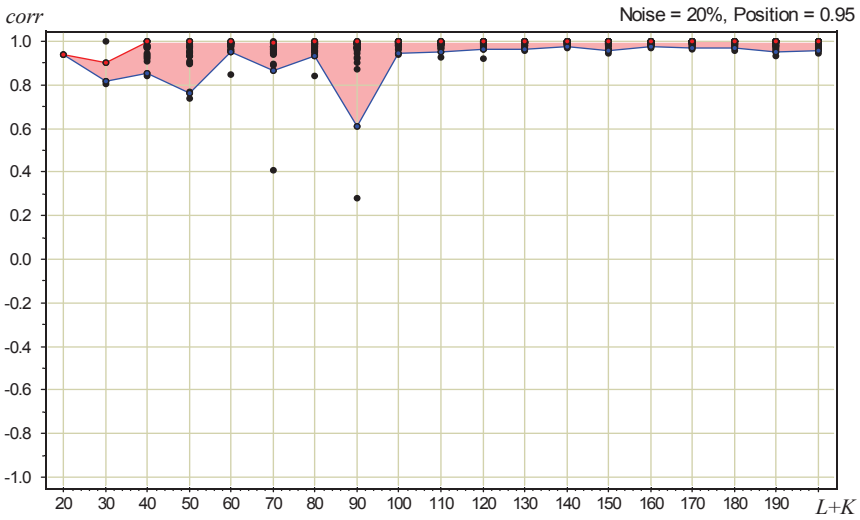

Рис. 5: 95%-ный доверительный интервал зависимости корреляции наблюдаемого и ненаблюдаемого критериев информативности от длины супервыборки, скорректированных по функционалу ожидаемой переобученности EOF. Вычисления проводились на на двумерных выборках с 20%-ным граничным шумом.

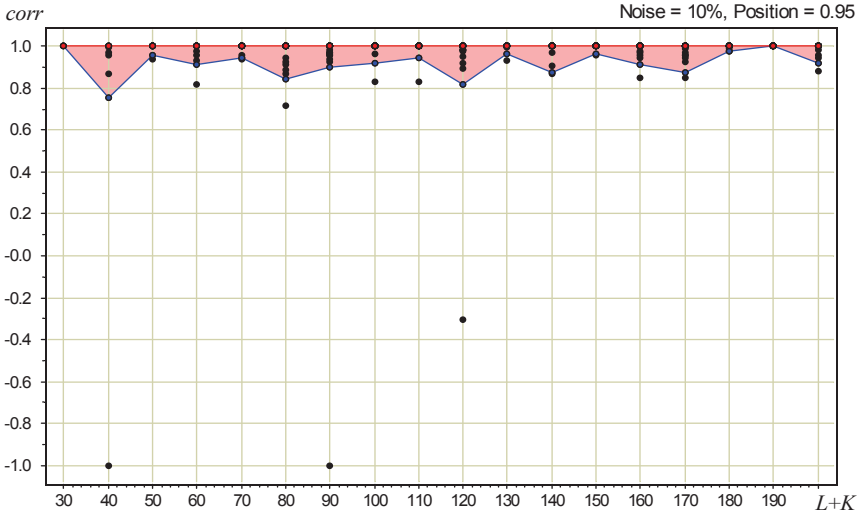

Рис. 6: 95%-ный доверительный интервал зависимости корреляции наблюдаемого и ненаблюдаемого критериев информативности от длины супервыборки, скорректированных по функционалу ожидаемой переобученности EOF. Вычисления проводились на трехмерных выборках с 10%-ным граничным шумом.

5 Заключение

Расслоение и связность — очень важные характеристики семейства алгоритмов, определяющие их обобщающую способность. С помощью оценок расслоения связности возможно делать более аккуратный отбор признаков в семействе пороговых конъюнкций, таким образом понижая переобученность закономерностей. На практике это реализуется с помощью модификации критерия информативности.

Основные результаты работы. Для семейства конъюнкций пороговых условий исследуется число положительных и отрицательных объектов закономерности, выписываются их оценки на неизвестной контрольной выборке с помощью функционалов вероятности переобучения и ожидаемой переобученности. Вводятся понятия модифицированного наблюдаемого и ненаблюдаемого критериев информативности, исследуются их свойства и возможность использования при оценке набора признаков. Экспериментально обосновывается переход от ненаблюдаемого критерия к наблюдаемому. Один из важных фактов, подтвержденных в данной работе — связность и расслоение являются универсальным характеристиками выборки, зависящими от структуры самой выборки и практически не зависящими от ее размера. То есть эти два параметра сохраняют информацию о структурных особенностях данных.

Направления дальнейших исследований. Цель дальнейших исследований — выяснение условий и ограничений, при которых модифицированный критерий информативности будет давать улучшение при отборе закономерностей и построении алгоритмов классификации.

Список литературы

[1] П. Ботов. Оценки вероятности переобучения многомерных семейств алгоритмов классификации // Диссертация на соискание ученой степени кандидата физико-математических наук. ВЦ РАН, 2011.

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

[3] К. В. Воронцов. Комбинаторная теория надежности обучения по прецедентам // Диссертация на соискание ученой степени доктора физико-математических наук. ВЦ РАН, 2010.

[4] К. В. Воронцов. Комбинаторная теория переобучения: результаты, приложения и открытые проблемы // 15-я всероссийская конференция «Математические методы распознавания образов», Петрозаводск, 2011. — С. 40–43.

[5] А. А. Ивахненко. Комбинаторные оценки вероятности переобучения и их применение в логических алгоритмах классификации
МФТИ, 2010.

[6] J. Fürnkranz and P. A. Flach. Roc 'n' rule learning-towards a better understanding of covering algorithms // *Machine Learning*, 58(1):39–77, 2005.

[7] T. Hastie, R. Tibshirani, J. Friedman. The Elements of Statistical Learning, 2nd ed. Springer. 2009.

[8] K. V. Vorontsov and A. A. Ivahnenko. Tight combinatorial generalization bounds for threshold conjunction rules // *4th International Conference on Pattern Recognition and Machine Intelligence (PReMI'11). June 27 – July 1, 2011*, 2011.