Комбинаторные оценки качества обучения по прецедентам*

К. В. Воронцов
voron@ccas.ru

3 июля 2003 г.
Принято в печать в Докладах РАН

В сообщении рассматриваются функционалы скользящего контроля и их верхние оценки, характеризующие качество обучения алгоритмов по прецедентным эмпирическим данным. Случайность и независимость исходных данных не предполагаются. Описывается эффект локализации семейства алгоритмов и вводится понятие локальной функции роста. Приводятся оценки качества монотонных классификаторов, не вырожденные на малых выборках и не зависящие от сложности семейства.

Задача обучения по прецедентам состоит в следующем. Задано множество объектов X, множество ответов Y и множество $\mathfrak{A}$ отображений из X в Y, называемых алгоритмами. Существует отображение $y^*: X \to Y$, не обязательно принадлежащее $\mathfrak{A}$, значения которого $y_i = y^*(x_i)$ известны только на объектах конечной обучающей выборки $X^l = \{x_1, \ldots, x_l\}$. Требуется построить алгоритм $a^* \in \mathfrak{A}$, удовлетворяющий локальным ограничениям $a^*(x_i) = y_i$, $i = 1, \ldots, l$ и универсальным ограничениям $a^* \in \mathfrak{A}_u$, где множество алгоритмов $\mathfrak{A}_u \subseteq \mathfrak{A}$ определяется спецификой конкретной задачи [2]. Искомый алгоритм a^* должен приближать восстанавливаемую зависимость y^* не только на объектах обучающей выборки, но и на всём множестве X. Данное требование можно формализовать с помощью различных функционалов качества, некоторые из которых будут рассмотрены ниже.

Частота ошибок алгоритма $a \in \mathfrak{A}$ на выборке $X^p = \{x_1, \ldots, x_p\} \subset X$ есть

$$\nu(a, X^p) = \frac{1}{p} \sum_{i=1}^{p} I(x_i, a(x_i)),$$

где $I(x, y)$ — индикатор ошибки, принимающий значение 1, если ответ y является ошибочным для объекта x, и 0 в противном случае. Обычно индикатор ошиб-

*Работа поддержана Российским фондом фундаментальных исследований (№ 02-01-00325) и Фондом содействия отечественной науке.

ки определяют как функцию отклонения ответа y от правильного ответа $y^*(x)$, например $I(x,y) = \big[\,|y - y^*(x)| \geqslant \delta\,\big]$ при заданном $\delta > 0$. Здесь и далее квадратные скобки обозначают отображение логического результата в число: $[\,\text{Ложь}\,] = 0$, $[\,\text{Истина}\,] = 1$.

Определение 1. Методом обучения называется отображение μ, которое произвольной конечной обучающей выборке X^l с заданными на ней ответами $Y^l = \{y_1, \ldots, y_l\}$ ставит в соответствие определённый алгоритм $a = \mu(X^l, Y^l)$. Говорят также, что метод μ строит алгоритм a по обучающей выборке X^l.

Будем полагать, что метод μ строит алгоритмы, выбирая их из некоторого семейства алгоритмов $A \subseteq \mathfrak{A}_u$. Предполагая отображение y^* фиксированным, будем использовать сокращённое обозначение $\mu(X^l)$.

Алгоритм a называется корректным на выборке X^l, если $\nu(a, X^l) = 0$. Метод μ называется корректным на выборке X^l, если алгоритм $\mu(X^l)$ корректен на X^l. Корректность метода на обучающей выборке в общем случае не гарантирует, что построенный алгоритм будет столь же хорошо работать на остальных выборках.

Рассмотрим несколько функционалов, характеризующих качество алгоритмов вне обучающей выборки (говорят также о способности алгоритма к обобщению или экстраполяции).

1. Функционал частоты ошибок $\nu(\mu(X^l), X^k)$ на заданной контрольной выборке X^k. Недостаток этого функционала заключается в том, что фиксируется некоторое, вообще говоря, произвольное разбиение выборки $X^l \cup X^k$ на обучающую и контрольную части. Если значение $\nu(\mu(X^l), X^k)$ достаточно мало́, то нет никакой гарантии, что при другом разбиении той же выборки (X_1^l, X_1^k) значение $\nu(\mu(X_1^l), X_1^k)$ будет столь же мало́. Таким образом, при построении функционала качества целесообразно учитывать различные способы разбиения выборки. Далее будем считать, что l и k — произвольные фиксированные числа и $L = l + k$.

2. Функционал скользящего контроля:

$$Q_c^{l,k}(\mu, X^L) = \frac{1}{N} \sum_{n=1}^{N} \nu(\mu(X_n^l), X_n^k),$$

где (X_n^l, X_n^k), $n = 1, \ldots, N$ — всевозможные разбиения выборки X^L на обучающую и контрольную подвыборки длиной l и k соответственно, $N = C_L^l$.

3. Функционал скользящего контроля, терпимый к незначительной доле ошибок ε на контрольной подвыборке, $0 \leqslant \varepsilon < 1$:

$$Q_\varepsilon^{l,k}(\mu, X^L) = \frac{1}{N} \sum_{n=1}^{N} \big[\,\nu(\mu(X_n^l), X_n^k) > \varepsilon\,\big]\,.$$

Теорема 1. Функционалы $Q_c^{l,k}$ и $Q_\varepsilon^{l,k}$ связаны двусторонними оценками

$$\varepsilon Q_\varepsilon^{l,k} \leqslant Q_c^{l,k} \leqslant \varepsilon + (1-\varepsilon) Q_\varepsilon^{l,k}.$$

4. Функционал скользящего контроля, терпимый к незначительным отклонениям частоты ошибок на контрольной выборке от частоты ошибок на обучении:

$$Q_{\nu,\varepsilon}^{l,k}(\mu, X^L) = \frac{1}{N} \sum_{n=1}^{N} \left[\nu(\mu(X_n^l), X_n^k) - \nu(\mu(X_n^l), X_n^l) > \varepsilon \right].$$

Если метод μ корректен на всевозможных подвыборках длины l, то функционалы $Q_\varepsilon^{l,k}$ и $Q_{\nu,\varepsilon}^{l,k}$ совпадают. В общем случае они связаны неравенством $Q_{\nu,\varepsilon}^{l,k} \leqslant Q_\varepsilon^{l,k}$.

5. Пусть множество объектов X является вероятностным пространством, выборки X^l и X^k случайные, независимые и одинаково распределенные, A — заданное семейство алгоритмов. Вапником и Червоненкисом предложен вероятностный функционал равномерного отклонения частоты ошибок в двух выборках, для которого в случае $l = k$ ими получена верхняя оценка [1]:

$$P_{\nu,\varepsilon}^{l,k}(A) = \mathsf{P}\Big\{ \sup_{a \in A} \big(\nu(a, X^k) - \nu(a, X^l) \big) > \varepsilon \Big\} \leqslant 1.5\, \Delta^A(L)\, e^{-\varepsilon^2 l},$$

где $\Delta^A(L)$ — функция роста семейства алгоритмов A, определяемая как число различных бинарных векторов $(\beta_1, \ldots, \beta_L)$, $\beta_i = I(x_i, a(x_i))$, порождаемых всевозможными алгоритмами $a \in A$ на всевозможных выборках X^L. Если семейство A имеет конечную ёмкость h, то $\Delta^A(L) \leqslant 1.5\, \frac{L^h}{h!}$.

Функционалы скользящего контроля $Q_c^{l,k}$, $Q_\varepsilon^{l,k}$, $Q_{\nu,\varepsilon}^{l,k}$ будем называть комбинаторными. В отличие от вероятностного функционала $P_{\nu,\varepsilon}^{l,k}$ они зависят от метода обучения и конкретной выборки, которая не обязана быть случайной. При соответствующих теоретико-вероятностных предположениях возможен переход от комбинаторных функционалов к вероятностным с помощью операции математического ожидания:

$$\begin{aligned}
&\mathsf{E} Q_c^{l,k}(\mu, X^L) = \mathsf{P}\{ I(\mu(X^l), x) = 1 \}; \\
&\mathsf{E} Q_\varepsilon^{l,k}(\mu, X^L) = \mathsf{P}\big\{ \nu(\mu(X^l), X^k) > \varepsilon \big\}; \\
&\mathsf{E} Q_{\nu,\varepsilon}^{l,k}(\mu, X^L) = \mathsf{P}\big\{ \nu(\mu(X^l), X^k) - \nu(\mu(X^l), X^l) > \varepsilon \big\} \leqslant P_{\nu,\varepsilon}^{l,k}(A). \qquad (1)
\end{aligned}$$

Отсюда следует, что любые верхние оценки комбинаторных функционалов легко переносятся на соответствующие вероятностные функционалы. Кроме того, из неравенства (1) следует, что оценка Вапника-Червоненкиса верна и для $\mathsf{E} Q_{\nu,\varepsilon}^{l,k}$.

Оказывается, эта же оценка, и даже более сильная, верна непосредственно для функционала $Q_{\nu,\varepsilon}^{l,k}(\mu, X^L)$ при произвольных μ и X^L. Усиление оценки связано с эффектом «локализации» функции роста, который состоит в том, что при фиксированной выборке лишь конечная часть семейства A может быть получена в результате обучения, а остальные алгоритмы остаются незадействованными.

Определение 2. Локальным семейством алгоритмов, порождённым методом μ на выборке X^L, называется множество алгоритмов

$$A_L^l(\mu, X^L) = \left\{\mu(X_n^l) \;\middle|\; n = 1, \ldots, N\right\}, \quad N = C_L^l.$$

Определение 3. Локальной функцией роста $\Delta_L^l(\mu, X^L)$ метода μ на выборке X^L называется число различных бинарных векторов $(\beta_1, \ldots, \beta_L)$, $\beta_i = I(x_i, a(x_i))$, порождаемых всевозможными алгоритмами $a \in A_L^l(\mu, X^L)$.

Локальная функция роста не превосходит $\Delta^A(L)$ и ограничена сверху числом C_L^l, в то время как $\Delta^A(L) \leqslant 2^L$.

Определение 4. Степенью некорректности метода μ на выборке X^L называется максимальная частота ошибок на всевозможных обучающих подвыборках длины l:

$$\sigma_L^l(\mu, X^L) = \max_{n=1,\ldots,N} \nu(\mu(X_n^l), X_n^l).$$

Если метод μ является корректным на всех подвыборках длины l, то $\sigma_L^l = 0$. В дальнейшем будут использоваться сокращённые обозначения Δ_L^l, A_L^l и σ_L^l с опущенными аргументами (μ, X^L).

Теорема 2. При любых μ и X^L справедлива оценка

$$Q_{\nu,\varepsilon}^{l,k}(\mu, X^L) < \Delta_L^l(\mu, X^L) \cdot \Gamma_L^l(\varepsilon, \sigma_L^l), \tag{2}$$

где $\Gamma_L^l(\varepsilon, \sigma) = \max\limits_m \sum\limits_s \dfrac{C_m^s C_{L-m}^{l-s}}{C_L^l}$, индекс m пробегает значения от $\lceil \varepsilon k \rceil$ до $k + \sigma l$, индекс s пробегает значения от $\max\{0, m-k\}$ до $\min\left\{\left\lfloor \frac{l}{L}(m - \varepsilon k)\right\rfloor, \sigma l\right\}$.

Величину $\Gamma_L^l(\varepsilon, \sigma)$ будем называть комбинаторным множителем.

Следствие 1. Оценка (2) не убывает по σ, поскольку $\Gamma_L^l(\varepsilon, \sigma)$ не убывает по σ. Наименьшее значение достигается при $\sigma = 0$, когда метод является корректным:

$$\Gamma_L^l(\varepsilon, 0) = \frac{C_{L-\lceil \varepsilon k \rceil}^l}{C_L^l} \leqslant \left(\frac{k}{L}\right)^{\varepsilon k}.$$

Следствие 2. При $l = k$ и любых (μ, X^L) для функционала качества $Q_{\nu,\varepsilon}^{l,k}$ справедлива оценка Вапника-Червоненкиса с точностью до замены функции роста всего семейства на локальную функцию роста:

$$Q_{\nu,\varepsilon}^{l,k}(\mu, X^L) < 1.5\, \Delta_L^l(\mu, X^L)\, e^{-\varepsilon^2 l} \leqslant 1.5\, \Delta^A(L)\, e^{-\varepsilon^2 l}. \tag{3}$$

Отметим, что нет особых оснований приравнивать l и k, за исключением удобства оценивания Γ_L^l, поэтому мы рассматриваем общий случай произвольных l и k.

Полученный результат означает, что качество обучения можно описывать не только на языке теории вероятностей, но и с помощью комбинаторных функционалов, зависящих от выборки, и основанных на идее скользящего контроля. Оценка (3) справедлива для любой выборки, не обязательно случайной и независимой.

В теории вероятностей независимость означает инвариантность вероятностной меры относительно всевозможных перестановок элементов выборки. В комбинаторной постановке вместо независимости выборки достаточно предположить инвариантность функционала качества относительно всевозможных перестановок выборки (симметричность функционала). Заметим, что все введенные выше комбинаторные функционалы симметричны. Данное ограничение является существенно более слабым, поскольку оно накладывается не на исходные данные, а на используемый функционал качества. Таким образом, природа оценки (3) оказывается исключительно комбинаторной и вытекает из дискретности индикатора ошибки $I(x, y)$ и симметричности функционала качества.

Для комбинаторных функционалов возможно получение даже более точных оценок, зависящих от свойств конкретной выборки. В частности, такие оценки позволяют учесть эффект локализации функции роста. В силу неравенства (1) вероятностный функционал равномерного отклонения частот может рассматриваться как верхняя оценка функционала скользящего контроля. Потеря точности в этой оценке связана с избыточностью требования равномерной сходимости.

Представим отношение правой и левой частей неравенства (3) в следующем виде:

$$\frac{\Delta(A)\,1.5\,e^{-\varepsilon^2 l}}{Q^{l,k}_{\nu,\varepsilon}} = \frac{\Delta(A)}{\Delta^l_L} \cdot \frac{1.5\,e^{-\varepsilon^2 l}}{\Gamma^l_L} \cdot \frac{\Delta^l_L \Gamma^l_L}{Q^{l,k}_{\nu,\varepsilon}}.$$

В каждой из дробей числитель является верхней оценкой знаменателя. Три сомножителя в правой части равенства описывают соответственно три основные причины завышенности вероятностных оценок качества. Первая причина — пренебрежение эффектом локализации. Сложность конечного подсемейства алгоритмов A^l_L, реально получаемых в результате обучения, может оказаться существенно меньше сложности всего семейства A. Вторая причина — относительная погрешность экспоненциальной оценки комбинаторного множителя, которая, в отличие от абсолютной погрешности, заметно увеличивается с ростом l. Третья причина — погрешность разложения функционала скользящего контроля в произведение локальной функции роста Δ^l_L и комбинаторного множителя Γ^l_L.

Перспективным подходом к повышению точности оценок представляется полный отказ от использования сложностных характеристик семейства алгоритмов. Оценки такого вида известны для стабильных алгоритмов [5] и выпуклых комбинаций классификаторов [4]. Мы рассматриваем ещё один случай — когда имеется априорная информация о монотонности или почти монотонности восстанавливаемой зависимости. Практическая значимость монотонных классификаторов обсуждается в [6]. Методы построения монотонных алгоритмов по конечным выборкам рассматриваются в [3] для задач классификации и восстановления регрессии.

Рассмотрим задачу классификации, в которой множество X частично упо-

рядочено, $Y = \{0, 1\}$, индикатор ошибки имеет вид $I(x, y) = |y^*(x) - y|$, метод обучения μ выбирает алгоритмы из множества A всех монотонных отображений из X в Y.

Определение 5. Степенью немонотонности выборки X^L называется наименьшая частота ошибок, допускаемых на ней монотонными алгоритмами:

$$\delta(X^L) = \min_{a \in A} \nu(a, X^L).$$

Выборка X^L называется монотонной, если из $x_i \leqslant x_j$ следует $y_i \leqslant y_j$ для всех $i, j = 1, \ldots, L$. Выборка монотонна тогда и только тогда, когда $\delta(X^L) = 0$. Если метод μ строит алгоритмы с минимальной частотой ошибок на обучающей выборке в классе всех монотонных функций A, то этот метод будет корректным на любой монотонной выборке [3].

Определение 6. Верхним и нижним клином объекта $x_i \in X^L$ называются соответственно множества

$$W_0(x_i) = \left\{ x_k \in X^L \;\middle|\; x_i < x_k \text{ и } y_k = 0 \right\};$$
$$W_1(x_i) = \left\{ x_k \in X^L \;\middle|\; x_k < x_i \text{ и } y_k = 1 \right\}.$$

Мощность клина $w_i = |W_{y_i}(x_i)|$ характеризует глубину погружения объекта x_i в тот класс, которому он принадлежит. Чем меньше w_i, тем ближе объект к границе класса. Для граничных объектов $w_i = 0$. Если монотонный алгоритм допускает ошибку на объекте x_i, то он допускает ошибку и на всех объектах из клина $W_{y_i}(x_i)$. Данный факт существенно используется при доказательстве следующей теоремы.

Теорема 3. Если метод μ строит алгоритм с минимальной частотой ошибок на обучающей выборке в классе всех монотонных функций, и если степень немонотонности выборки X^L равна δ, то

$$Q_c^{l,k}(\mu, X^L) \leqslant \frac{1}{L} \sum_{\substack{i=1 \\ w_i < \delta L + k}}^{L} \sum_{s=0}^{\min\{\delta L, l, w_i\}} \frac{C_{w_i}^s C_{L-1-w_i}^{l-s}}{C_{L-1}^l}. \tag{4}$$

Следствие. Оценка монотонно не убывает по δ, достигая наименьшего значения при $\delta = 0$, когда выборка монотонна и метод μ является корректным:

$$Q_c^{l,k}(\mu, X^L) \leqslant \frac{1}{L} \sum_{\substack{i=1 \\ w_i < k}}^{L} \frac{C_{L-1-w_i}^l}{C_{L-1}^l} \leqslant \frac{1}{L} \sum_{\substack{i=1 \\ w_i < k}}^{L} \left(\frac{k}{L}\right)^{w_i}.$$

Полученная оценка, в отличие от сложностных, всегда не превышает 1. Наибольшее значение 1 достигается, если $w_i = 0$ для всех $i = 1, \ldots, L$. Это тот случай, когда оба класса состоят из попарно несравнимых объектов, и вся выборка распадается на две антицепи. Наименьшее значение достигается, когда выборка монотонна и линейно упорядочена. В этом случае число клиньев мощности w не превышает 2 для всех $w = 1, \ldots, k$, откуда вытекает $Q_c^{l,k} \leqslant 2/l$.

Ёмкость класса монотонных классификаторов бесконечна, поскольку на выборке длины L, состоящей из попарно несравнимых элементов, реализуется ровно 2^L дихотомий. Таким образом, классическая теория Вапника-Червоненкиса вообще не даёт оценок качества для данного случая. Известно [6], что эффективная ёмкость класса монотонных функций не превосходит длины максимальной антицепи в выборке X^L. Оценка (4) существенно более точная, особенно при малых выборках.

Автор выражает глубокую признательность академику РАН Ю. И. Журавлёву за оказанную поддержку и своему Учителю чл.-корр. РАН К. В. Рудакову за постоянное внимание к работе и ценные замечания.

Список литературы

[1] *Вапник В. Н.* Восстановление зависимостей по эмпирическим данным.— М. Наука, 1979.

[2] *Журавлёв Ю. И., Рудаков К. В.* Об алгебраической коррекции процедур обработки (преобразования) информации // *Проблемы прикладной математики и информатики.*— 1987.— С. 187–198.
http://www.ccas.ru/frc/papers/zhurrud87correct.pdf.

[3] *Рудаков К. В., Воронцов К. В.* О методах оптимизации и монотонной коррекции в алгебраическом подходе к проблеме распознавания // *Доклады РАН.*— 1999.— Т. 367, № 3.— С. 314–317.
http://www.ccas.ru/frc/papers/rudvoron99dan.pdf.

[4] *Bartlett P.* The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network // *IEEE Transactions on Information Theory.*— 1998.— Vol. 44, no. 2.— Pp. 525–536.
http://discus.anu.edu.au/~bartlett.

[5] *Bousquet O., Elisseeff A.* Stability and generalization // *Journal of Machine Learning Research.*— 2002.— no. 2.— Pp. 499–526.
http://citeseer.nj.nec.com/article/bousquet00stability.html.

[6] *Sill J.* The capacity of monotonic functions // *Discrete Applied Mathematics (special issue on VC dimension).*— 1998.— Vol. 86.— Pp. 95–107.
http://citeseer.nj.nec.com/49191.html.

ПРИЛОЖЕНИЕ

1 Доказательство теоремы 1

Докажем справедливость двусторонних оценок $\varepsilon Q_\varepsilon^{l,k} \leqslant Q_c^{l,k} \leqslant \varepsilon + (1-\varepsilon)Q_\varepsilon^{l,k}$.

Введём сокращённое обозначение $\nu_n^{lk} = \nu(\mu(X_n^l), X_n^k), \quad n = 1, \ldots, N$. Соотношение $\varepsilon Q_\varepsilon^{l,k} \leqslant Q_c^{l,k}$ вытекает из очевидного неравенства $\varepsilon \left[\nu_n^{lk} > \varepsilon \right] \leqslant \nu_n^{lk}$, справедливого для любого ε из отрезка $[0, 1]$.

Для получения второго соотношения разобьём множество индексов $\{1, \ldots, N\}$ на два подмножества $N_1 = \{n \in N \mid \nu_n^{lk} > \varepsilon\}$ и $N_2 = \{n \in N \mid \nu_n^{lk} \leqslant \varepsilon\}$. Тогда

$$
\begin{aligned}
Q_\varepsilon^{l,k} &= \frac{1}{N} \sum_{n=1}^{N} \left[\nu_n^{lk} > \varepsilon \right] = \frac{|N_1|}{N}; \\
Q_c^{l,k} &= \frac{1}{N} \left(\sum_{n \in N_1} \nu_n^{lk} + \sum_{n \in N_2} \nu_n^{lk} \right) \leqslant \\
&\leqslant \frac{|N_1|}{N} + \frac{\varepsilon |N_2|}{N} = \frac{|N_1|}{N} + \varepsilon \frac{N - |N_1|}{N} = \varepsilon + (1-\varepsilon) Q_\varepsilon^{l,k}.
\end{aligned}
$$

Теорема доказана.

В силу доказанной теоремы функционалы $Q_c^{l,k}$ и $Q_\varepsilon^{l,k}$ могут рассматриваться как взаимозаменяемые. Для оценивания качества можно выбирать тот функционал, с которым удобнее работать в каждом конкретном случае.

2 Доказательство теоремы 2

Идея доказательства в целом следует Вапнику [1], за исключением того, что рассматривается комбинаторный функционал скользящего контроля вместо вероятностного функционала равномерной сходимости частот в двух подвыборках.

Введём на локальном семействе алгоритмов A_L^l отношение эквивалентности, положив, что произвольные алгоритмы a и a' из A_L^l эквивалентны тогда и только тогда, когда они допускают ошибки на одних и тех же объектах выборки X^L:

$$
a \sim a' \Leftrightarrow (\forall x \in X^L)\ I(x, a) = I(x, a').
$$

Обозначим через Δ_m число способов получить m ошибок на данной выборке всевозможными алгоритмами из A_L^l.

Отношение эквивалентности разбивает множество A_L^l на классы эквивалентности, обозначаемые далее A_{mi}, где m — число ошибок, допускаемых на выборке X^L алгоритмами данного класса, i — порядковый номер класса среди всех классов, алгоритмы которых допускают m ошибок, $i = 1, \ldots, \Delta_m$.

Таким образом, множество A_L^l представляется в виде объединения непересекающихся классов эквивалентности:

$$A_L^l = \bigcup_{m=0}^{L} \bigcup_{i=1}^{\Delta_m} A_{mi}.$$

Очевидно, количество всех классов эквивалентности равно локальной функции роста метода μ на выборке X^L:

$$\Delta_L^l = \sum_{m=0}^{L} \Delta_m. \tag{5}$$

Эквивалентность на алгоритмах порождает эквивалентность на разбиениях, если для произвольных n и n' из $\{1,\ldots,N\}$ положить $n \sim n' \Leftrightarrow \mu(X_n^l) \sim \mu(X_{n'}^l)$. При этом на множестве разбиений образуются классы эквивалентности, взаимно однозначно соответствующие классам A_{mi}:

$$N_{mi} = \left\{ n \;\middle|\; \mu(X_n^l) \in A_{mi} \right\}.$$

Множество всех разбиений также представляется в виде объединения непересекающихся классов эквивалентности:

$$\{1,\ldots,N\} = \bigcup_{m=0}^{L} \bigcup_{i=1}^{\Delta_m} N_{mi}.$$

Запишем функционал качества $Q_{\nu+\varepsilon}(\mu, X^L)$, суммируя разбиения отдельно по каждому классу эквивалентности:

$$Q_{\nu,\varepsilon}^{l,k}(\mu, X^L) = \frac{1}{N} \sum_{m=0}^{L} \sum_{i=1}^{\Delta_m} \sum_{n \in N_{mi}} [\nu(\mu(X_n^l), X_n^k) > \nu(\mu(X_n^l), X_n^l) + \varepsilon].$$

Согласно определению эквивалентностей на алгоритмах и разбиениях, значение функционала не изменится, если алгоритм $\mu(X_n^l)$, $n \in N_{mi}$, заменить на произвольный элемент a_{mi} из класса A_{mi}. Учтём также, что при $m \leqslant \varepsilon k$ и при $m > k + \sigma_L^l l$ под знаком суммы оказывается нуль. Поэтому суммирование можно производить не по всем m, а только от $m_0 = \lceil \varepsilon k \rceil$ до $m_1 = k + \sigma_L^l l$.

Итак,

$$Q_{\nu,\varepsilon}^{l,k}(\mu, X^L) = \sum_{m=m_0}^{m_1} \sum_{i=1}^{\Delta_m} \underbrace{\frac{1}{N} \sum_{n \in N_{mi}} [\nu(a_{mi}, X_n^k) > \nu(a_{mi}, X_n^l) + \varepsilon]}_{\gamma_{mi}}. \tag{6}$$

Обозначим внутреннюю сумму через γ_{mi} и оценим её сверху, заменив суммирование по классу эквивалентности N_{mi} суммированием по всем разбиениям. Обозначим полученную оценку через $\tilde{\gamma}_m$:

$$\gamma_{mi} \leqslant \tilde{\gamma}_m \equiv \frac{1}{N} \sum_{n=1}^{N} [\nu(a_{mi}, X_n^k) > \nu(a_{mi}, X_n^l) + \varepsilon].$$

Обозначим через s число ошибок, допускаемых алгоритмом a_{mi} на обучающей подвыборке X_n^l. Тогда $\tilde{\gamma}_m$ есть доля разбиений выборки X^L, при которых $\frac{m-s}{k} > \frac{s}{l} + \varepsilon$. Выражая отсюда s, получаем $s < (m - \varepsilon k)l/L$. Имеется также второе ограничение $s < \sigma_L^l l$, вытекающее из определения показателя некорректности. Таким образом, имеем:

$$s \leqslant s_1 \equiv \min\left(\lfloor (m - \varepsilon k)l/L \rfloor, \sigma_L^l l\right). \tag{7}$$

Кроме того, на s накладываются два ограничения снизу: $s \geqslant 0$ и $s \geqslant m - k$, поскольку общее число ошибок m не может превышать суммы $s + k$. Таким образом,

$$s \geqslant s_0 \equiv \max(0, m - k). \tag{8}$$

Обозначим через $\Gamma_{L,l}^{m,s}$ долю разбиений выборки длины L на две подвыборки, при которых из m ошибок в первую подвыборку длины l попадают s ошибок. Из комбинаторики известно, что

$$\Gamma_{L,l}^{m,s} = \frac{C_m^s C_{L-m}^{l-s}}{C_L^l}. \tag{9}$$

Следовательно, величина $\tilde{\gamma}_m$ представляется в виде суммы $\tilde{\gamma}_m = \sum_s \Gamma_{L,l}^{m,s}$ по всем s, удовлетворяющим (7) и (8). Заметим, что она не зависит от i, поэтому в (6) её можно вынести за знак суммирования по i. Оценивая, согласно (5), сумму величин Δ_m локальной функцией роста, приходим к неравенству:

$$Q_{\nu+\varepsilon}(\mu, X^L) \leqslant \sum_{m=m_0}^{m_1} \sum_{i=1}^{\Delta_m} \tilde{\gamma}_m = \sum_{m=m_0}^{m_1} \Delta_m \tilde{\gamma}_m < \Delta_L^l \underbrace{\max_{m_0 \leqslant m \leqslant m_1} \sum_{s=s_0}^{s_1} \Gamma_{L,l}^{m,s}}_{\Gamma_L^l(\varepsilon, \sigma_L^l)}.$$

Подставляя сюда $\Gamma_{L,l}^{m,s}$ из (9), получаем требуемую оценку.

3 Доказательство теоремы 3

Запишем в функционале скользящего контроля частоту ошибок через сумму индикаторов ошибки и поменяем местами знаки суммирования:

$$\begin{aligned} Q_c^{l,k}(\mu, X^L) &= \frac{1}{N} \sum_{n=1}^{N} \frac{1}{k} \sum_{x \in X_n^k} I(x, \mu(X_n^l)) = \\ &= \frac{1}{k} \sum_{i=1}^{L} \frac{1}{N} \underbrace{\sum_{n=1}^{N} [x_i \in X_n^k] I(x_i, \mu(X_n^l))}_{N_i}. \end{aligned} \tag{10}$$

Внутренняя сумма, обозначенная через N_i, выражает число разбиений выборки X^L, при которых объект x_i оказывается в контрольной части выборки, и построенный для данного разбиения алгоритм допускает на нём ошибку.

Оценим N_i, воспользовавшись следующим свойством клиньев, вытекающим непосредственно из определения. Если алгоритм a монотонный и допускает ошибку на объекте x_i, то он допускает ошибку и на всех объектах из клина $W_i = W_{y_i}(x_i)$. В зависимости от соотношения мощности клина $w_i = |W_i|$ и степени немонотонности выборки возможны три случая.

Если $w_i \leqslant \delta L$, то N_i оценим сверху тривиальным образом — числом разбиений, при которых x_i попадает в контрольную выборку:

$$N_i \leqslant C_{L-1}^{l}. \tag{11}$$

Если $w_i \geqslant \delta L + k$, то ни при каком разбиении монотонный алгоритм не будет ошибаться на x_i, поскольку $\delta L + k$ есть максимальное число ошибок, которое может допустить монотонная функция на всей выборке X^L. Это вытекает из допущения, что метод μ строит алгоритм с минимимальным числом ошибок на обучающей выборке в классе всех монотонных функций. Минимальное число ошибок на любой подвыборке X_n^l не превосходит минимального числа ошибок на всей выборке X^L. Следовательно число ошибок на обучении не превышает δL. Таким образом, в этом случае $N_i = 0$.

Рассмотрим третий случай, когда $\delta L < w_i < \delta L + k$. Пусть s — число объектов из W_i, находящихся в обучающей выборке, $0 \leqslant s \leqslant \min\{\delta L, l\}$. Имеется $C_{w_i}^{s}$ способов поделить клин W_i на s обучающих объектов и $s - w_i$ контрольных. Для каждого из этих способов имеется $C_{L-1-w_i}^{l-s}$ вариантов выбрать $l - s$ обучающих объектов из оставшихся $L - 1 - w_i$. В итоге получаем оценку числа разбиений:

$$N_i \leqslant \sum_{s=0}^{\min\{\delta L, l, w_i\}} C_{w_i}^{s} C_{L-1-w_i}^{l-s}. \tag{12}$$

Это неравенство справедливо также и в первом случае, поскольку (12) обращается в (11), когда верхний предел суммирования равен w_i (в соответствии с известным комбинаторным тождеством — свёрткой Вандермонда).

Представим N в виде $N = C_L^l = \frac{L}{k} C_{L-1}^l$ и подставим найденную оценку для N_i в (10), учитывая, что $N_i = 0$ при $w_i \geqslant \delta L + k$:

$$Q_c^{l,k}(\mu, X^L) \leqslant \frac{1}{k} \sum_{\substack{i=1 \\ w_i < \delta L + k}}^{L} \frac{k}{L} \sum_{s=0}^{\min\{\delta L, l, w_i\}} \frac{C_{w_i}^{s} C_{L-1-w_i}^{l-s}}{C_{L-1}^{l}},$$

откуда непосредственно вытекает требуемое (4).