ISSN 2074-1871 Уфимский математический журнал. Том 10. № 1 (2018). С. 50-65.

УДК 519.25

КОМБИНАТОРНЫЕ ОЦЕНКИ ПЕРЕОБУЧЕНИЯ ПОРОГОВЫХ РЕШАЮЩИХ ПРАВИЛ

Ш.Х. ИШКИНА

Аннотация. Оценивание обобщающей способности является фундаментальной задачей теории статистического обучения. Тем не менее, точные и вычислительно эффективные оценки до сих пор не известны даже для многих простых частных случаев. В данной работе исследуется семейство одномерных пороговых решающих правил. Применяется комбинаторная теория переобучения, основанная на единственном вероятностном допущении, что все разбиения множества объектов на обучающую и тестовую выборки равновероятны. Предлагается полиномиальный алгоритм для вычисления функционалов вероятности переобучения и полного скользящего контроля. Алгоритм основан на рекуррентном подсчете числа допустимых траекторий при блуждании по трехмерной сетке между двумя заданными точками с ограничениями специального вида. Проведенное сравнение полученных точных оценок обобщающей способности демонстрирует завышенность существующих верхних оценок и их неприменимость для реальных задач.

Ключевые слова: статистическое обучение, минимизации эмпирического риска, комбинаторная теория переобучения, вероятность переобучения, полный скользящий контроль, обобщающая способность, пороговое правило, вычислительная сложность.

Mathematics Subject Classification: 68Q32, 60C05

1. Введение

Рассмотрим следующую математическую модель принятия решений в условиях неполноты информации. Задана бинарная матрица, строки которой соответствуют объектам, столбцы — правилам принятия решений, называемым также классификаторами или гипотезами. В ячейке матрицы находится единица тогда и только тогда, когда данный классификатор ошибается на данном объекте. Из множества $\mathbb{X}$ всех строк матрицы случайно и равновероятно выбирается наблюдаемая обучающая выборка — подмножество $X \subset \mathbb{X}$ фиксированной мощности. Затем из множества $\mathbb{A}$ всех столбцов матрицы выбирается классификатор с минимальной частотой ошибок на X. Требуется оценить частоту ошибок этого классификатора на скрытой контрольной выборке $\bar{X} = \mathbb{X} \setminus X$. Если разность частот ошибок на контрольной и обучающей выборках превышает ε, то говорят, что произошло переобучение. Получение верхних оценок вероятности переобучения является одной из основных задач теории статистического обучения [1]–[3].

Классические оценки Вапника–Червоненкиса [1] зависят только от размера матрицы ошибок. Будучи оценками «худшего случая», они завышены на порядки и плохо согласуются с результатами экспериментов [4]. Более тонкие оценки зависят от свойств отношения частичного порядка на множестве вектор-столбцов матрицы ошибок [5]. В комбинаторной

Sh.Kh. Ishkina, Combinatorial bounds of overfitting for threshold classifiers.

Работа выполнена при финансовой поддержке РФФИ, проекты № 15-37-50350 мол_нр и № 14-07-00847.

Поступила 21 декабря 2016 г.

теории переобучения [6]–[8] обосновывается необходимость сочетания двух свойств, расслоения и связности [9, 12]. Благодаря расслоению, классификаторы с высокой вероятностью ошибки вносят пренебрежимо малый вклад в переобучение. Благодаря связности, у классификаторов с близкими векторами ошибок резко снижается вклад в переобучение.

В [13] получены условия, при которых оценка расслоения–связности является точной. Им удовлетворяют, в частности, монотонные и унимодальные цепи классификаторов [9]. В практических задачах статистического обучения такие цепи могут порождаться элементарными пороговыми правилами, используемых в таких алгоритмах классификации, как решающие деревья, логические закономерности [14], алгоритмы вычисления оценок [15], а также при построении линейных классификаторов методом покоординатной оптимизации. Но при этом делается предположение о существовании безошибочного правила, практически не выполнимое в реальных задачах. В общем случае пороговые правила порождают последовательности классификаторов, называемые прямыми цепями.

Ранее для них были известны лишь верхние оценки ожидаемой частоты ошибок на контрольной выборке [16]. Различные уточнения оценок расслоения–связности, например, учитывающие попарную конкуренцию между классификаторами [17] или послойную кластеризацию множества классификаторов [18, 19], также остаются завышенными для прямых цепей.

В данной работе предлагается алгоритм полиномиальной сложности для вычисления вероятности переобучения произвольной прямой цепи. Алгоритм основан на рекуррентном подсчете числа допустимых траекторий при блуждании по трехмерной сетке между двумя заданными точками с ограничениями специального вида.

1.1. Основные определения. Задано конечное множество $\mathbb{X} = \{x_1, \ldots, x_L\}$, элементы которого называются *объектами*, и конечное множество $\mathbb{A}$, элементы которого называются *классификаторами*. Множество $\mathbb{A}$ называется *семейством классификаторов*.

Задана функция $I\colon \mathbb{A} \times \mathbb{X} \to \{0, 1\}$, называемая *индикатором ошибки*. Если $I(a, x) = 1$, то говорят, что классификатор a *допускает ошибку* на объекте x. Бинарная матрица $\big(I(a, x)\colon x \in \mathbb{X}, a \in \mathbb{A}\big)$ размера $|\mathbb{X}|\times|\mathbb{A}|$ называется *матрицей ошибок*.

Предполагается, что каждому классификатору $a \in \mathbb{A}$ взаимно однозначно соответствует его вектор ошибок $(I(a, x_i))_{i=1}^{L}$, т.е. в матрице ошибок не может быть двух равных столбцов. Будем считать, что порядок строк в матрице ошибок не важен. Договоримся обозначать через a как классификатор, так и его вектор ошибок.

Числом ошибок классификатора a на выборке $X \subset \mathbb{X}$ называется величина

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

Частотой ошибок классификатора a на выборке $X \subset \mathbb{X}$ называется величина

$$\nu(a, X) = n(a, X)/|X|.$$

Обозначим через $[X]^l$ множество всех подмножеств $\mathbb{X}$ мощности $l < L$. Подмножества $X \in [X]^l$ будем называть *обучающими выборками*, а их дополнения $\bar{X} = \mathbb{X}\backslash X$ — *контрольными выборками*. Введем на множестве $[X]^l$ равномерное распределение вероятностей:

$$\mathsf{P}(X) = 1/C_L^l, \qquad X \in [X]^l.$$

Переобученностью классификатора a на разбиении $(X, \bar{X})$ называется величина

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

Если $\delta(a, X) > \varepsilon$, то будем говорить, что классификатор a переобучен на X.

Методом обучения называется отображение $\mu\colon [X]^l \to \mathbb{A}$, которое каждой обучающей выборке X ставит в соответствие классификатор $a = \mu X$ из семейства $\mathbb{A}$.

Пессимистичной минимизацией эмпирического риска (ПМЭР) называется метод обучения, который выбирает классификатор, допускающий наименьшее число ошибок на обучающей выборке X, а если таких классификаторов в семействе несколько, то выбирает из них классификатор с наибольшим числом ошибок на контрольной выборке $\bar{X}$ [9].

Для фиксированного метода обучения μ, семейства классификаторов $\mathbb{A}$, множества $\mathbb{X}$ и объема обучающей выборки l *вероятностью переобучения* называется функционал

$$Q_\varepsilon(\mu, \mathbb{A}, \mathbb{X}, l) = \mathsf{P}[\delta(\mu X, X) \geqslant \varepsilon] = \frac{1}{C_L^l} \sum_{X \in [X]^l} [\delta(\mu X, X) \geqslant \varepsilon].$$

Здесь и далее квадратные скобки будут использоваться для преобразования логического условия в числовое значение по правилу [истина] $= 1$, [ложь] $= 0$.

Полным скользящим контролем (complete cross-validation, CCV) называется функционал, равный математическому ожиданию числа ошибок на контрольной выборке:

$$CCV(\mu, \mathbb{A}, \mathbb{X}, l) = \mathsf{E}\nu(\mu X, \bar{X}) = \frac{1}{C_L^l} \sum_{X \in [X]^l} \nu(\mu X, \bar{X}).$$

Эффективное вычисление Q_ε и CCV непосредственно по определению возможно только при малых $|\bar{X}| = L - l$. Если l близко к $L/2$, то число слагаемых экспоненциально по L.

1.2. Прямые последовательности классификаторов. Рассмотрим множества объектов, по которым различаются соседние классификаторы семейства $\mathbb{A} = \{a_0, \ldots, a_P\}$:

$$G_p = \big\{x \in \mathbb{X} \,|\, I(a_p, x) \neq I(a_{p+1}, x)\big\}, \quad p = 0, \ldots, P-1. \tag{1}$$

Определение 1. *Семейство классификаторов называется прямой последовательностью, если множества G_p попарно не пересекаются.*

Заметим, что из определения следует, что порядок классификаторов важен. Действительно, рассмотрим два семейства классификаторов, первое из которых является прямой последовательностью $\mathbb{A} = \{a_0, \ldots, a_P\}$, а второе получается из первого перестановкой классификаторов a_p и a_{p+1} для некоторого p: $\mathbb{A}' = \{a_0, \ldots, a_{p-1}, a_{p+1}, a_p, a_{p+2}, \ldots, a_P\}$. Определим множества G_p по (1). Тогда семейство $\mathbb{A}'$ не является прямой последовательностью, поскольку соседние классификаторы a_{p-1} и a_{p+1} различаются по множеству объектов $G_{p-1} \sqcup G_p$, а классификаторы a_{p+1} и a_p – по множеству объектов G_p, т.е. эти множества пересекаются.

Определение 2. *Прямая последовательность $\mathbb{A} = \{a_0, \ldots, a_P\}$ называется прямой цепью, если каждая пара соседних классификаторов различается по одному объекту: $|G_p| = 1$, $p = 0, \ldots, P-1$. Число P называется длиной прямой цепи $\mathbb{A}$.*

Определение 3. *Одномерным пороговым классификатором над множеством $\mathbb{X} \subset \mathbb{R}$ называется семейство пороговых правил $a(x, \theta) = [x \geqslant \theta]$, где $\theta \in \mathbb{R}$ – параметр, называемый порогом.*

Согласно следующей теореме, понятия прямой последовательности и одномерного порогового классификатора являются синонимами.

Теорема 1. *Определим множество V прямых последовательностей $\mathbb{A} = \{a_0, \ldots, a_P\}$, таких, что $\sum_{p=0}^{P-1} |G_p| = L$, где G_p определены по (1), и множество U одномерных пороговых классификаторов над множеством $\mathbb{X} = \{x_1, \ldots, x_L\}$ точек числовой оси, таким, что каждому x_i соответствует истинная метка класса $y_i \in \{0, 1\}$. Тогда между этими множествами имеется биекция.*

Доказательство. Во множествах V и U объекты определены с точностью до переименования объектов множества $\mathbb{X}$.

Каждый объект $u \in U$ однозначно определяется распределением объектов двух классов $\{0,1\}$ на числовой оси, т.е. расположением точек множества $\mathbb{X}$ на оси $\mathbb{R}$ и набором правильных ответов $\{y_1,\ldots,y_L\}$. Значения порогов выбираются так, чтобы они всеми возможными различными способами разбивали множество $\mathbb{X}$ на два класса.

Каждый объект множества V однозначно определяется количеством единиц в векторе a_0, т.е. $n(a_0,\mathbb{X})$, и последовательностью пар $(n_0^p, n_1^p)_{p=0}^{P-1}$, где n_0^p – количество нулей в векторе a_p, являющихся единицами в a_{p+1}, и n_1^p – количество единиц в векторе a_p, являющихся нулями в a_{p+1}. При наличии данной информации матрица ошибок $\{a_0,\ldots,a_P\}$ строится следующим образом. Вектор a_0 задается так, что на первых $n(a_0,\mathbb{X})$ позициях стоят единицы, затем нули. Для каждого p последовательно, начиная с $p=0$, вектор a_{p+1} получается из вектора a_p путем инвертирования n_0^p нулей и n_1^p единиц.

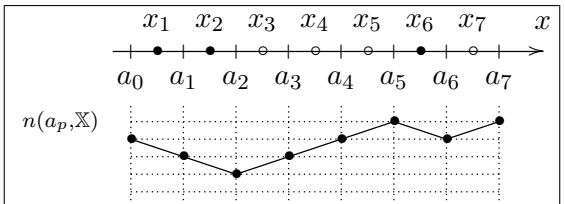
Построим отображение $f: U \to V$ следующим образом. Пусть дан объект $u \in U$, т.е. набор точек $x_1 \leqslant \cdots \leqslant x_L$ и правильных ответов $y_1,\ldots,y_L$. Поставим ему в соответствие прямую последовательность $v = f(u) \in V$.

Для этого введём индикатор ошибки $I(a,x_i) = \big[a(x_i,\theta) \neq y_i\big]$. Варьирование θ порождает не более $L+1$ классификаторов с попарно различными векторами ошибок. Они образуют прямую последовательность. Если все объекты x_i попарно различны, $x_1 < x_2 < \cdots < x_L$, то прямая последовательность является прямой цепью.

Отображение f однозначно определяет прямую последовательность по семейству пороговых правил, т.е. оно является инъекцией. Докажем, что оно является сюръекцией.

Пусть дана прямая последовательность $v \in V$, т.е. величина $n(a_0,\mathbb{X})$ и набор пар $(n_0^p, n_1^p)_{p=0}^{P-1}$. Построим матрицу ошибок $\{a_0,\ldots,a_P\}$. Определим семейство пороговых правил $u \in U$ следующим образом. Поставим в соответствие каждому множеству G_p точки $x_p^1 = \cdots = x_p^{|G_p|}$ и положим, что $x_0^1 < x_1^1 < \cdots < x_{P-1}^1$. Положим $y_p^i = 1$, если $I(a_p, x_p^i) = 0$, и $y_p^i = 0$ в противном случае. Легко проверить, что построенное семейство u является прообразом v при отображении f, т.е. $v = f(u)$. Таким образом, отображение f является биекцией. □

Пример 1. На рис. 1 показан пример прямой цепи. По оси x отложены объекты x_i. Правильные решения y_i показаны точками ∘ и •. Пороги θ выбраны посередине между соседними объектами. Ниже показан график числа ошибок классификаторов и матрица ошибок.



x_1	1	0	0	0	0	0	0	0
x_2	1	1	0	0	0	0	0	0
x_3	0	0	0	1	1	1	1	1
x_4	0	0	0	0	1	1	1	1
x_5	0	0	0	0	0	1	1	1
x_6	1	1	1	1	1	1	0	0
x_7	0	0	0	0	0	0	0	1

Рис. 1: Пример прямой цепи

Определение 4. *Прямая цепь* $\mathbb{A} = \{a_0, \ldots, a_P\}$ *называется возрастающей (убывающей), если каждый классификатор* a_p *допускает* $m+p$ *(соответственно,* $m-p$*) ошибок на множестве* $\mathbb{X}$ *при некотором значении* m*. Прямую цепь* $\mathbb{A}$ *будем называть монотонной, если она является убывающей или возрастающей.*

Прямая цепь $\mathbb{A}$ может состоять из нескольких участков монотонности. Например, в цепи, показанной на рис. 1, имеется четыре участка монотонности: $\{a_0, a_1, a_2\}$ и $\{a_5, a_6\}$ — убывающие, $\{a_2, a_3, a_4, a_5\}$ и $\{a_6, a_7\}$ — возрастающие.

1.3. Постановка задачи. Найти способ вычисления функционалов вероятности переобучения Q_ε и полного скользящего контроля CCV за полиномиальное по L время для ПМЭР μ и произвольной прямой последовательности $\mathbb{A}$.

2. Переобучение произвольного семейства

Пусть дано произвольное подмножество $\mathbb{D} \subseteq \mathbb{X}$ множества $\mathbb{X}$. Каждое разбиение $(X, \bar{X})$ множества $\mathbb{X} = X \sqcup \bar{X}$ индуцирует разбиение $(X \cap \mathbb{D}, \bar{X} \cap \mathbb{D})$ подмножества $\mathbb{D}$. Также любая пара разбиений $(D', \bar{D}')$ и $(D'', \bar{D}'')$ подмножеств $\mathbb{D}' \subseteq \mathbb{X}$ и $\mathbb{D}'' = \mathbb{X} \backslash \mathbb{D}'$ соответственно определяет разбиение $(X, \bar{X})$ множества $\mathbb{X}$ по правилу $X = D' \cup D''$ и $\bar{X} = \bar{D}' \cup \bar{D}''$.

Назовем пару классификаторов a и a' *неразличимыми на множестве* $\mathbb{X}' \subset \mathbb{X}$, если $I(a, x) = I(a', x)$ для всех $x \in \mathbb{X}'$.

Пусть дано произвольное семейство классификаторов $\mathbb{A}$. Пусть на множестве $\mathbb{A} \times \mathbb{A} \times [X]^\ell$ имеется отношение строгого порядка $a \succ_X a'$. Назовем его *финитным*, если для любых классификаторов $a, a' \in \mathbb{A}$, неразличимых на множестве $\mathbb{X}' \subset \mathbb{X}$, отношение $a \succ_X a'$ не зависит от выбора разбиения множества $\mathbb{X}'$.

Пример 2. Определенные по следующим правилам отношения порядка являются финитными:

1. $a \succ_X a' \iff n(a, X) < n(a', X)$;
2. $a \succ_X a' \iff \delta(a, X) > \delta(a', X)$.

Действительно, для любого $X \in [X]^\ell$ и для любого $\mathbb{X}'$ справедливо равенство $n(a, X) = n(a, X \cap \mathbb{X}') + n(a, X \setminus \mathbb{X}')$. Если классификаторы a и a' неразличимы на множестве $\mathbb{X}'$, то $n(a, \mathbb{X}' \cap X) = n(a', \mathbb{X}' \cap X)$, откуда следует финитность отношения 1.

Для доказательства финитности отношения 2 перепишем переобученность как $\delta(a, X) = \frac{1}{L-\ell} n(a, \mathbb{X}) - \frac{L}{(L-\ell)\ell} n(a, X)$. Тогда утверждение следует из первого пункта.

Из определения вытекает следующее свойство:

Лемма 1. *Пусть классификаторы семейства* $\mathbb{A}' \subseteq \mathbb{A}$ *неразличимы на множестве* $\mathbb{N}'$*. Тогда для любого* $a \in \mathbb{A}'$ *выполнение финитного отношения* $a \succ_X a'$ *одновременно для всех* $a' \in \mathbb{A}' \setminus \{a\}$ *не зависит от выбора разбиения множества* $\mathbb{N}'$*.*

Будем говорить, что на выборке X классификатор a *лучше, чем* a', если $a \succ_X a'$. Назовем метод обучения $\mu\colon [X]^\ell \to \mathbb{A}$ *финитным*, если результатом обучения является лучший с точки зрения финитного отношения $\succ_X$ классификатор:

$$a = \mu X \quad \Leftrightarrow \quad a \succ_X a', \quad \forall a' \neq a. \tag{2}$$

Пример 3. Метод минимизации эмпирического риска (МЭР), выбирающий классификатор с минимальным числом ошибок на обучающей выборке, и метод максимизации переобученности (МП), выбирающий классификатор с максимальной переобученностью, являются финитными.

Метод МП возникает в задаче комбинаторного вычисления рademахеровской сложности класса решающих правил [10]. Действительно, при $\ell = \frac{L}{2}$ случайные величины

$$\sigma_i = \begin{cases} +1, \text{если } x_i \in \bar{X}, \\ -1, \text{если } x_i \notin \bar{X}, \end{cases}$$

подчиняются рademахеровскому распределению $P(\sigma_i = 1) = P(\sigma_i = -1) = \frac{1}{2}$. Тогда рademахеровская сложность семейства равна ожидаемой переобученности метода МП μ [11]:

$$\mathcal{R}_L(\mathbb{A}, \mathbb{X}) = \mathsf{E} \sup_{a \in \mathbb{A}} \frac{2}{L} \sum_{i=1}^{L} \sigma_i a_i = \mathsf{E} \sup_{a \in \mathbb{A}} \nu(a, \bar{X}) - \nu(a, X) = \mathsf{E}\delta(\mu, \bar{X}).$$

Рademахеровскую сложность можно рассматривать как величину, описывающую сложность класса решающих правил. Чем больше Рademахеровская сложность, тем лучше ошибки классификаторов семейства могут коррелировать со случайным шумом σ_i.

Обозначим через $\mathbb{D}$ подмножество объектов, по которым классификаторы семейства $\mathbb{A} = \{a_0, \ldots, a_P\}$ различимы:

$$\mathbb{D} = G_0 \cup \cdots \cup G_{P-1} = \big\{x \in \mathbb{X} \,|\, \exists a, a' \in \mathbb{A} \colon I(a, x) \neq I(a', x)\big\}, \tag{3}$$

где множества G_p определяются согласно (1).

Объекты множества $\mathbb{N} = \mathbb{X} \setminus \mathbb{D}$ назовем *нейтральными*. На множестве $\mathbb{N}$ классификаторы семейства неразличимы и допускают одинаковое число ошибок m. Через m_p обозначим число ошибок классификатора a_p на множестве $\mathbb{D}$:

$$\begin{aligned} m &= n(a, \mathbb{N}), \quad \forall a \in \mathbb{A}; \\ m_p &= n(a_p, \mathbb{D}). \end{aligned} \tag{4}$$

Сведем задачу вычисления вероятности переобучения Q_ε и полного скользящего контроля CCV к нахождению числа разбиений множества $\mathbb{D}$ с некоторыми ограничениями.

Будем обозначать через t число объектов из $\mathbb{D}$, попавших в обучающую выборку X, а через e — число ошибок классификатора a_p на этих объектах. Введём две функции от t и e: число разбиений множества $\mathbb{N}$, таких, что классификатор a_p переобучен на X

$$N_p(t, e) = \#\big\{(X \cap \mathbb{N}, \bar{X} \cap \mathbb{N}) \;\big|\; \delta(a_p, X) \geqslant \varepsilon,\ t = |X \cap \mathbb{D}|,\ e = n(a_p, X \cap \mathbb{D})\big\},$$

и число разбиений множества $\mathbb{D}$, таких, что a_p является результатом обучения:

$$D_p(t, e) = \#\big\{(X \cap \mathbb{D}, \bar{X} \cap \mathbb{D}) \;\big|\; \mu X = a_p,\ t = |X \cap \mathbb{D}|,\ e = n(a_p, X \cap \mathbb{D})\big\}.$$

Введём *гипергеометрическую функцию распределения*

$$H_L^{l,m}(s) = \frac{1}{C_L^l} \sum_{i=0}^{\min\{\lfloor s \rfloor, l, m\}} C_m^i C_{L-m}^{l-i},$$

где $\lfloor x \rfloor$ — целая часть x, т.е. наибольшее целое число, не превосходящее x. Гипергеометрическая функция распределения $H_L^{l,m}(s)$ для данного множества $\mathbb{X}$ мощности L и выборки $X_0 \subset \mathbb{X}$ объема m равна доле выборок множества $\mathbb{X}$ объема l, содержащих не более s элементов из X_0. Будем полагать $C_n^i = 0$ при невыполнении условия $0 \leqslant i \leqslant n$.

Теорема 2. *Для произвольного семейства классификаторов $\mathbb{A} = \{a_0, \ldots, a_P\}$, финитного метода обучения μ, множества $\mathbb{X}$ мощности L, объема обучающей выборки l, точности $\varepsilon \in (0, 1)$ вероятность переобучения имеет вид*

$$Q_\varepsilon = \frac{1}{C_L^l} \sum_{p=0}^{P} \sum_{(t,e) \in \Psi_p} D_p(t, e) N_p(t, e), \tag{5}$$

где множество $\mathbb{D}$, *параметры* m_p *и* m *определяются по* (3) *и* (4) *и*

$$\Psi_p = \big\{(t,e)\,|\,0 \leqslant t \leqslant \min\{l, |\mathbb{D}|\},\ 0 \leqslant e \leqslant \min\{t, m_p\}\big\}; \tag{6}$$

$$N_p(t,e) = C^{l-t}_{L-|\mathbb{D}|}\, H^{l-t,m}_{L-|\mathbb{D}|}(s_p(e)); \tag{7}$$

$$s_p(e) = \frac{l}{L}(n(a_p,\mathbb{X}) - \varepsilon(L-l)) - e.$$

Доказательство. Представим вероятность переобучения в виде

$$Q_\varepsilon = \sum_{p=0}^{P} \mathsf{P}\big[\mu X = a_p \text{ и } \delta(a_p, X) \geqslant \varepsilon\big].$$

Рассмотрим множество разбиений $(X, \bar{X})$ с фиксированными значениями t и e:

$$t = |X \cap \mathbb{D}|, \quad e = n(a_p, X \cap \mathbb{D}). \tag{8}$$

Множество допустимых значений (t,e) есть Ψ_p, согласно (6).

Для таких разбиений выполнение условия $\delta(a_p, X) \geqslant \varepsilon$ не зависит от выбора разбиения множества $\mathbb{D}$, а выполнение условия $\mu X = a_p$ по лемме 1 не зависит от выбора разбиения множества $\mathbb{N}$, поскольку классификаторы неразличимы на множестве $\mathbb{N}$. Поэтому для каждой тройки параметров p, t, e число разбиений множества $\mathbb{X}$, таких, что одновременно выполнены условия $\mu X = a_p$ и $\delta(\mu X, X) \geqslant \varepsilon$, равно произведению $N_p(t,e)D_p(t,e)$.

Докажем (7). Пусть $n(a_p, X \cap \mathbb{N}) = s$, тогда $n(a_p, X) = e + s$. Условие $\delta(a_p, X) \geqslant \varepsilon$ эквивалентно условию $n(a_p, X) \leqslant \frac{l}{L}(n(a_p,\mathbb{X}) - \varepsilon(L-l))$, значит, $s \leqslant s_p(e)$. Число разбиений множества $\mathbb{N}$ при данных t и s равно $C^s_m C^{l-t-s}_{L-|\mathbb{D}|-m}$, откуда следует

$$N_p(t,e) = \sum_{s=0}^{s_p(e)} C^s_m C^{l-t-s}_{L-|\mathbb{D}|-m} = C^{l-t}_{L-|\mathbb{D}|} \frac{1}{C^{l-t}_{L-|\mathbb{D}|}} \sum_{s=0}^{s_p(e)} C^s_m C^{l-t-s}_{L-|\mathbb{D}|-m} = C^{l-t}_{L-|\mathbb{D}|}\, H^{l-t,m}_{L-|\mathbb{D}|}(s_p(e)).$$

□

Для функционала полного скользящего контроля имеет место аналогичная теорема.

Теорема 3. *Для произвольного семейства классификаторов* $\mathbb{A} = \{a_0, \ldots, a_P\}$, *финитного метода обучения* μ, *множества* $\mathbb{X}$ *мощности* L, *объема обучающей выборки* l, *функционал полного скользящего контроля имеет вид*

$$CCV = \frac{1}{(L-l)C^l_L} \sum_{p=0}^{P} \sum_{(t,e)\in\Psi_p} D_p(t,e)F_p(t,e), \tag{9}$$

где

$$F_p(t,e) = \sum_{s=0}^{\min\{l-t,m\}} C^s_m C^{l-t-s}_{L-|\mathbb{D}|-m}\big(n(a_p,\mathbb{X}) - s - e\big), \tag{10}$$

множества $\mathbb{D}$ *и* Ψ_p *определяются по* (3) *и* (6), *параметры* m_p *и* m *определяются по* (4).

Доказательство. Запишем формулу полного скользящего контроля и переставим в ней знаки суммирования:

$$CCV = \frac{1}{C^l_L} \sum_{X\in[X]^l} \sum_{p=0}^{P} [\mu X = a_p]\, \nu(a_p, \bar{X}) = \frac{1}{C^l_L} \sum_{p=0}^{P} \sum_{X\in[X]^l} [\mu X = a_p]\, \nu(a_p, \bar{X}).$$

Выполнение условия $\mu X = a_p$ по лемме 1 не зависит от выбора разбиения множества $\mathbb{N}$.

Представим число ошибок классификатора a_p на контрольной выборке в виде

$$n(a_p, \bar{X}) = n(a_p,\mathbb{X}) - n(a_p, X) = n(a_p,\mathbb{X}) - n(a_p, X\cap\mathbb{D}) - n(a_p, X\cap\mathbb{N}).$$

Определим параметры t и e по формулам (8). Обозначим $s = n(a_p, X \cap \mathbb{N})$. Из ограничений $s + t \leqslant l$ и $s \leqslant m$ следует верхняя оценка параметра s в (10).

Легко проверить, что число разбиений множества $\mathbb{N}$ при данных t и s равно $C_m^s C_{L-|\mathbb{D}|-m}^{l-t-s}$, откуда следует утверждение теоремы. □

Таким образом, задача сводится к вычислению для каждого p значений $D_p(t, e)$ на всем множестве Ψ_p. Для случая прямой последовательности далее будет описан рекуррентный алгоритм вычисления $D_p(t, e)$.

3. Вычисление количества разбиений множества ребер прямой последовательности

Пусть теперь семейство $\mathbb{A} = \{a_0, \ldots, a_P\}$ является прямой последовательностью. Объекты множества $\mathbb{D}$ будем называть *ребрами прямой последовательности* $\mathbb{A}$.

3.1. Сведение к задачам на левой и правой последовательностях. Рассмотрим классификатор a_p и зафиксируем точку $(t, e) \in \Psi_p$. Относительно a_p прямая последовательность $\mathbb{A}$ разбивается на две: левую $a_0, a_1, \ldots, a_p$ и правую $a_p, a_{p+1}, \ldots, a_P$.

Сведем задачу вычисления $D_p(t, e)$ к нахождению числа разбиений множества ребер левой и правой последовательностей с некоторыми ограничениями.

Теорема 4. *Пусть μ – финитный метод обучения. Для каждого p для всех $(t, e) \in \Psi_p$ число разбиений множества $\mathbb{D}$, таких, что $t = |X \cap \mathbb{D}|$, $e = n(a_p, X \cap \mathbb{D})$ и $\mu X = a_p$, равно*

$$D_p(t, e) = \sum_{t'+t''=t} \sum_{e'+e''=e} L_p(t', e') R_p(t'', e''), \tag{11}$$

где

$$L_p(t', e') = \#\left\{ (X \cap \mathbb{L}_p, \bar{X} \cap \mathbb{L}_p) \;\middle|\; \begin{array}{l} \forall d = 0, \ldots, p \;\; a_p \succ_X a_d, \\ t' = |X \cap \mathbb{L}_p|, \;\; e' = n(a_p, X \cap \mathbb{L}_p) \end{array} \right\}, \tag{12}$$

$$R_p(t'', e'') = \#\left\{ (X \cap \mathbb{R}_p, \bar{X} \cap \mathbb{R}_p) \;\middle|\; \begin{array}{l} \forall d = p+1, \ldots, P \;\; a_p \succ_X a_d, \\ t'' = |X \cap \mathbb{R}_p|, \;\; e'' = n(a_p, X \cap \mathbb{R}_p) \end{array} \right\}, \tag{13}$$

множества $\mathbb{L}_p$ и $\mathbb{R}_p$ – множества ребер левой и правой последовательностей соответственно, точки (t', e') и (t'', e'') являются элементами множеств Ψ'_p и Ψ''_p соответственно, где

$$\Psi'_p = \big\{ (t', e') \;\big|\; 0 \leqslant t' \leqslant \min\{l, |\mathbb{L}_p|\}, 0 \leqslant e' \leqslant \min\{t', n(a_p, \mathbb{L}_p)\} \big\}, \tag{14}$$

$$\Psi''_p = \big\{ (t'', e'') \;\big|\; 0 \leqslant t'' \leqslant \min\{l, |\mathbb{R}_p|\}, 0 \leqslant e'' \leqslant \min\{t'', n(a_p, \mathbb{R}_p)\} \big\}. \tag{15}$$

Доказательство. Множества $\mathbb{L}_p$ и $\mathbb{R}_p$ не пересекаются, значит, классификаторы левой последовательности неразличимы на $\mathbb{R}_p$, классификаторы правой последовательности неразличимы на $\mathbb{L}_p$. Тогда выполнение условия (2) для всех классификаторов левой последовательности по лемме 1 не зависит от выбора разбиения множества $\mathbb{R}_p$. Аналогично, выполнение условия (2) для всех классификаторов правой последовательности не зависит от выбора разбиения множества $\mathbb{L}_p$. Значит, общее число разбиений множества $\mathbb{D}$, в которых метод обучения выбирает a_p, является произведением числа разбиений множеств $\mathbb{L}_p$ и $\mathbb{R}_p$, в которых a_p лучше всех классификаторов левой и правой последовательностей соответственно. Параметры t', t'', e', e'' необходимы для выполнения условий, задаваемых параметрами t и e. □

Назовем разбиения множеств $\mathbb{L}_p$ и $\mathbb{R}_p$, удовлетворяющие условиям (12) и (13) соответственно, *допустимыми*.

Рассмотрим метод ПМЭР. Докажем, что он является финитным, значит, для него справедливы теоремы 2 — 4 и для каждого p задача сводится к вычислению числа допустимых разбиений $L_p(t', e')$ и $R_p(t'', e'')$ для всех точек множеств Ψ'_p и Ψ''_p.

Будем считать, что из классификаторов, минимизирующих число ошибок на обучающей выборке X и допускающих равное число ошибок на контрольной выборке $\bar{X}$, выбирается классификатор с наибольшим номером. Данное ограничение не влияет на оценку вероятности переобучения и полного скользящего контроля, но позволяет точно вычислить искомое количество разбиений.

Определение 5. *Запасом ошибок классификатора a относительно a_p на выборке X назовем величину $\Delta_p(a, X) = n(a, X) - n(a_p, X)$.*

Лемма 2. *Метод ПМЭР является финитным с отношением порядка $\succ_X$, определенным следующим образом: классификатор a_p лучше, чем классификатор a, на выборке X тогда и только тогда, когда выполнено одно из следующих условий:*

1) $\Delta_p(a, X) > 0$;
2) $\Delta_p(a, X) = 0$ и a находится в левой последовательности и $n(a, \mathbb{X}) \leqslant n(a_p, \mathbb{X})$;
3) $\Delta_p(a, X) = 0$ и a находится в правой последовательности и $n(a, \mathbb{X}) < n(a_p, \mathbb{X})$.

Лемма следует из определения ПМЭР.

Далее рассматривается случай, когда прямая последовательность $\mathbb{A}$ является прямой цепью. Тогда левая и правая последовательности $\mathbb{L}_p$ и $\mathbb{R}_p$ также являются цепями. Рассматривается метод ПМЭР μ с определенным по лемме 2 отношением порядка $\succ_X$.

3.2. Нахождение числа допустимых разбиений множества ребер левой цепи. Найдём $L_p(t', e')$ для каждого p в каждой точке $(t', e') \in \Psi'_p$. Заметим, что при $p = 0$ решение задачи тривиально: множество Ψ'_0 состоит из одной точки $(0, 0)$ и $L_0(0, 0) = 1$. Всюду далее считаем $1 \leqslant p \leqslant P$.

Перенумеруем классификаторы так, чтобы последовательность начиналась в a_p и заканчивалась в a_0. Обозначим $\{b_0, \ldots, b_p\}$, где $b_d = a_{p-d}$ для каждого $d = 0, \ldots, p$. Запас ошибок относительно a_p запишем как $\Delta_0(b_d, X) = \Delta_p(a_{p-d}, X)$ для каждого d.

Левая цепь $\mathbb{L}_p$ составлена из возрастающих и убывающих монотонных участков. Обозначим множество всех ребер возрастающих монотонных участков цепи через $\mathbb{C}_p$, убывающих монотонных участков цепи — через $\mathbb{I}_p$. Верно, что $\mathbb{C}_p \sqcup \mathbb{I}_p = \mathbb{L}_p$.

Цепь прямая, следовательно, b_0 не ошибается на всех объектах $\mathbb{C}_p$, т.е.

$$\begin{aligned} \mathbb{C}_p &= \{x \in \mathbb{L}_p \colon I(b_0, x) = 0\}, \\ \mathbb{I}_p &= \{x \in \mathbb{L}_p \colon I(b_0, x) = 1\}. \end{aligned} \tag{16}$$

Тогда верно, что $e' = |X \cap \mathbb{I}_p|$, а $|X \cap \mathbb{C}_p| = t' - e'$.

Заметим, что, поскольку классификаторы левой цепи различимы только на объектах множества $\mathbb{L}_p$, то для любого классификатора b из левой цепи верно

$$\Delta_0(b, X) = \Delta_0(b, X \cap \mathbb{L}_p), \quad \forall X \subseteq \mathbb{X}.$$

Отсюда следует, что, зафиксировав разбиение множества $\mathbb{L}_p$, мы определим запас ошибок на всех соответствующих обучающих выборках X.

Введём трехмерную сетку $\Omega_p = \{0, \ldots, |\mathbb{L}_p|\} \times \{-|\mathbb{L}_p|, \ldots, |\mathbb{L}_p|\} \times \{0, \ldots, |\mathbb{L}_p|\}$.

Определение 6. *Определим на Ω_p множество $\mathbb{T}_p$ траекторий, выходящих из точки $(0, 0, 0)$ и образованных переходами трех видов:*

1) из точки (d, Δ, i) в точку $(d + 1, \Delta, i)$ – «вправо»;
2) из точки (d, Δ, i) в точку $(d + 1, \Delta + 1, i)$ – «вправо-вверх»;
3) из точки (d, Δ, i) в точку $(d + 1, \Delta - 1, i + 1)$ – «вправо-вниз»;

причем для каждого d переход из точки (d, Δ, i) удовлетворяет условию: пусть классификаторы b_d и b_{d+1} соединены ребром x, тогда

1) если $x \in \mathbb{C}_p$, то это переход вида «вправо» или «вправо-вверх»;
2) если $x \in \mathbb{I}_p$, то это переход вида «вправо» или «вправо-вниз».

Теорема 5. *Между разбиениями множества $\mathbb{L}_p$ и траекториями из множества $\mathbb{T}_p$ имеется взаимно однозначное соответствие. Траектория, соответствующая разбиению $(X \cap \mathbb{L}_p, \bar{X} \cap \mathbb{L}_p)$, проходит через точки (d, Δ, i), где для каждого $d = 0, \ldots, p$ координата $\Delta = \Delta_0(b_d, X)$, а координата i равна числу ребер из $X \cap \mathbb{I}_p$ между b_0 и b_d.*

Доказательство. Пусть классификаторы b_{d-1} и b_d соединены ребром x.

Если $x \in \bar{X}$, то $\Delta_0(b_d, X) = \Delta_0(b_{d-1}, X)$, так как запас ошибок зависит только от X.

Пусть x лежит в X. Если x лежит в возрастающей цепи, то b_{d-1} не ошибается на этом ребре, тогда как b_d ошибается. Тогда $\Delta_0(b_d, X) = \Delta_0(b_{d-1}, X) + 1$. Если же x лежит в $\mathbb{I}_p$, то b_{d-1} ошибается на этом объекте, а b_d — нет. Значит, $\Delta_0(b_d, X) = \Delta_0(b_{d-1}, X) - 1$.

Поставим в соответствие разбиению множества $\mathbb{L}_p$ траекторию по следующему правилу. Пусть траектория проходит через точку (d, Δ, i). При $d = 0$ полагаем, что это точка $(0, 0, 0)$. Из этой точки вдоль траектории выполняется переход вида «вправо», если $x \in \bar{X}$; «вправо-вверх», если $x \in X \cap \mathbb{C}_p$; «вправо-вниз», если $x \in X \cap \mathbb{I}_p$.

Тогда для каждого d координаты Δ и i имеют смысл, указанный в условии теоремы, и при описанных переходах изменяются не более, чем на 1. Значит, траектория действительно целиком лежит на сетке Ω_p и, следовательно, во множестве $\mathbb{T}_p$ и однозначно определена.

По тем же правилам каждой траектории из $\mathbb{T}_p$ можно однозначно поставить в соответствие разбиение множества $\mathbb{L}_p$. Значит, отображение из множества разбиений во множество траекторий $\mathbb{T}_p$ сюръективно и инъективно, т.е. оно биективно. □

Пример 4. На рис. 2 на нижнем графике изображена цепь, где выделены ребра, попавшие в обучающую выборку. Такому разбиению ребер цепи соответствует траектория, проекция которой на плоскость (d, Δ) изображена на верхнем графике. В данном примере траектория проходит через точки, у которых координата Δ отрицательна. Значит, в цепи имеются классификаторы с отрицательным запасом ошибок. Следовательно, по лемме 2 и условиям (2), при таком разбиении классификатор b_0 не будет выбран методом обучения. Исключив из рассмотрения траектории, не удовлетворяющие лемме 2, мы отбросим и разбиения, не являющиеся допустимыми.

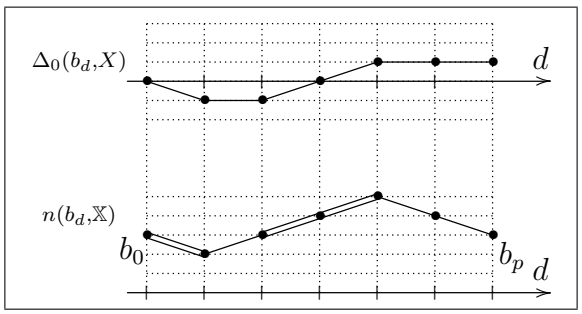


Рис. 2: Соответствие разбиения цепи (нижний график) проекции траектории (верхний график). Двойными линиями выделены ребра цепи, попавшие в обучающую выборку

Определим множество

$$\Omega'_p = \left\{ (d, \Delta, i) \in \Omega_p \;\middle|\; \begin{array}{l} 0 \leqslant i \leqslant d \text{ и } |\Delta| \leqslant d \text{ и} \\ \big(\text{либо } \Delta > 0, \text{ либо } (\Delta = 0 \text{ и } n(b_d, \mathbb{X}) \leqslant n(b_0, \mathbb{X}))\big) \end{array} \right\}. \tag{17}$$

Лемма 3. *Всякая точка (d, Δ, i) траектории из $\mathbb{T}_p$, соответствующей допустимому разбиению множества $\mathbb{L}_p$, принадлежит множеству $\Omega'_p \subseteq \Omega_p$.*

Доказательство. Выполнение первых двух условий из определения (17) является следствием теоремы 5. Третье условие есть повторение условий леммы 2. □

Пусть $T_p(d, \Delta, i)$ есть число траекторий из $\mathbb{T}_p$, соединяющих точку $(0, 0, 0)$ с (d, Δ, i) и проходящих только через точки множества Ω'_p. Из правил построения траектории по разбиению множества $\mathbb{L}_p$ следует

Лемма 4. *В каждой точке (d, Δ, i) на трехмерной сетке Ω_p величина $T_p(d, \Delta, i)$ вычисляется рекуррентно.*

1) Начальное условие $T_p(0, 0, 0) = 1$.
2) Если $(d, \Delta, i) \notin \Omega'_p$, то $T_p(d, \Delta, i) = 0$.
3) Пусть b_{d-1} и b_d соединены ребром x. Тогда

$$T_p(d, \Delta, i) = \begin{cases} T_p(d-1, \Delta, i) + T_p(d-1, \Delta-1, i), & \text{если } x \in \mathbb{C}_p, \\ T_p(d-1, \Delta, i) + T_p(d-1, \Delta+1, i-1), & \text{если } x \in \mathbb{I}_p, \end{cases} \tag{18}$$

где множества $\mathbb{C}_p$ и $\mathbb{I}_p$ определяются по (16).

Теорема 6. *Пусть даны метод ПМЭР μ, множество $\mathbb{X}$ мощности L, объем обучающей выборки l и прямая цепь $\mathbb{A} = \{a_0, \ldots, a_P\}$. Тогда для каждого $p = 1, \ldots, P$ в каждой точке (t', e') множества Ψ'_p, определенного в* (14), *число $L_p(t', e')$ допустимых разбиений множества $\mathbb{L}_p$, определяемое по* (12), *равно*

$$L_p(t', e') = T_p(|\mathbb{L}_p|, t' - 2e', e')$$

и вычисляется рекуррентно по правилам, описанным в лемме 4, где $b_d = a_{p-d}$ для каждого d, при краевых условиях $L_0(0, 0) = 1$.

Доказательство. Из теоремы 5 следует, что

$$\Delta_p(a_0, X) = |X \cap \mathbb{C}_p| - |X \cap \mathbb{I}_p| = t' - 2e'.$$

Между разбиениями множества ребер левой цепи и траекториями из $\mathbb{T}_p$ имеется биекция. Таким образом, число траекторий, проходящих через точку $(p, t' - 2e', e')$, равно числу разбиений, удовлетворяющих условиям $t' = |X \cap \mathbb{L}_p|$ и $e' = n(a_p, X \cap \mathbb{L}_p)$. Оставив среди них те, которые проходят только через точки множества $\Omega'_p(t', e')$, мы оставим траектории, соответствующие допустимым разбиениям. Их число равно $T_p(|\mathbb{L}_p|, t' - 2e', e')$. □

Замечание 1. Ограничения $i \leqslant e'$ и $\Delta \leqslant t' - e'$, являющиеся следствием теоремы 5, выполняются автоматически для тех траекторий, которые соединяют точки $(0, 0, 0)$ и $(p, t' - 2e', e')$. Действительно, поскольку величины i и $\Delta + i$ не возрастают, значит, не превосходят значений в конечной точке, т.е. $i \leqslant e'$ и

$$\Delta + i \leqslant t' - 2e' + e' = t' - e'.$$

Координата $i \geqslant 0$, значит, $\Delta \leqslant \Delta + i \leqslant t' - e'$. В силу этого замечания, в определение множества Ω'_p данные ограничения не входят.

Таким образом, мы научились решать задачу для левой цепи.

3.3. Нахождение допустимых разбиений множества ребер правой цепи. Решаем задачу вычисления $R_p(t'', e'')$ для каждого p в каждой точке $(t'', e'') \in \Psi''_p$. Решение практически повторяет решение задачи для левой цепи после замены $\mathbb{L}_p$ на $\mathbb{R}_p$ и точки (t', e') на (t'', e''). Также имеются краевые условия: при $p = P$ множество $\Psi''_P = \{(0,0)\}$ и $R_P(0,0) = 1$. Далее полагаем, что $0 \leqslant p \leqslant P-1$.

Обозначим классификаторы цепи через $b_d = a_{p+d}$ для каждого $d = 0, \ldots, P-p$. Из леммы 2 следует, что для справедливости леммы 4 для правой цепи множество Ω'_p необходимо заменить на множество Ω''_p, определяемое следующим образом:

$$\Omega''_p = \left\{ (d, \Delta, i) \in \Omega_p \;\middle|\; \begin{array}{l} 0 \leqslant i \leqslant d \text{ и } |\Delta| \leqslant d \text{ и} \\ \big(\text{либо } \Delta > 0, \text{ либо } (\Delta = 0 \text{ и } n(b_d, \mathbb{X}) < n(b_0, \mathbb{X}))\big) \end{array} \right\}. \tag{19}$$

По аналогии с теоремой 6, для правой цепи верна следующая теорема.

Теорема 7. *Пусть даны метод ПМЭР μ, множество $\mathbb{X}$ мощности L, объем обучающей выборки l и произвольная прямая цепь $\mathbb{A} = \{a_0, \ldots, a_P\}$. Тогда для каждого $p = 0, \ldots, P-1$ в каждой точке (t'', e'') множества Ψ''_p, определенного в (15), число $R_p(t'', e'')$ допустимых разбиений множества $\mathbb{R}_p$, определяемое по (13), равно*

$$R_p(t'', e'') = T_p(|\mathbb{R}_p|, t'' - 2e'', e'')$$

и вычисляется рекуррентно по правилам, описанным в лемме 4, с заменой множества Ω'_p на Ω''_p и b_d на a_{p+d} для каждого d. Краевые условия $R_P(0,0) = 1$.

Замечание 2. По лемме 2, для всех $d = 0, \ldots, P$ запас ошибок классификатора a_d цепи должен быть неотрицателен для допустимых разбиений множества ребер левой и правой цепи. В частности, $\Delta_p(a_0, X) = t' - 2e' \geqslant 0$ и $\Delta_p(a_P, X) = t'' - 2e'' \geqslant 0$. Значит, границы изменения вторых координат точек множеств Ψ_p, Ψ'_p, Ψ''_p имеют вид

$$0 \leqslant e \leqslant \min\{\tfrac{1}{2}t,\, m_p\}, \qquad 0 \leqslant e' \leqslant \min\{\tfrac{1}{2}t', n(a_p, \mathbb{L}_p)\}, \qquad 0 \leqslant e'' \leqslant \min\{\tfrac{1}{2}t'', n(a_p, \mathbb{R}_p)\}.$$

3.4. Нахождение числа допустимых разбиений множества ребер прямой последовательности. Рассмотрим общий случай прямой последовательности $\mathbb{A} = \{a_0, \ldots, a_P\}$. Сведем задачу вычисления количества допустимых разбиений левой и правой последовательностей к аналогичным задачам для прямых цепей.

Для этого построим прямую цепь $\mathbb{A}_c$, такую, что $\mathbb{A} \subseteq \mathbb{A}_c$ и первый и последний классификаторы семейств совпадают, следующим образом: для каждого i, такого, что $|G_i| > 1$, добавим в последовательность $\mathbb{A}$ прямую цепь $\mathbb{G}_i$

$$\{a_0, \ldots, a_{i-1}\} \cup \mathbb{G}_i \cup \{a_{i+2}, \ldots, a_P\},$$

где прямая цепь $\mathbb{G}_i$ такова, что первым классификатором цепи является a_i, последним — a_{i+1}. Для определенности будем считать, что $\mathbb{G}_i$ строится как прямая цепь, составленная из двух монотонных: убывающей цепи длины n_1 и возрастающей длины n_0, где

$$n_1 = \#\{x \in G_i \,|\, I(a_i, x) = 1\},$$
$$n_0 = \#\{x \in G_i \,|\, I(a_i, x) = 0\}.$$

Назовем построенную цепь $\mathbb{A}_c$ *интерполяцией* последовательности $\mathbb{A}$. Ее длина равна $|\mathbb{D}|$.

Для каждого $a_p \in \mathbb{A}$ рассмотрим левую последовательность $\{a_p, \ldots, a_0\} \subseteq \mathbb{A}$ и левую цепь $\{a_p, \ldots, a_0\} \subseteq \mathbb{A}_c$. По построению множества ребер данных семейств совпадают, вследствие чего множества допустимых разбиений левой цепи и левой последовательности, определенные по (12), также совпадают. Вычислим их количество по теоремам 6 и 7 с единственным отличием.

Согласно (2), условие $a_p \succ_X a$ должно быть выполнено только для $a \in \mathbb{A}$. Данное ограничение определяет строение множеств Ω'_p и Ω''_p, задаваемых в (17) и (19). Переопределим

их для случая интерполяции последовательности $\mathbb{A}$:

$$\Omega'_p = \left\{ (d,\Delta,i) \in \Omega_p \;\middle|\; \begin{array}{c} b_d \in \mathbb{A}_c\backslash\mathbb{A} \text{ или } \big(b_d \in \mathbb{A} \text{ и } 0 \leqslant i \leqslant d \text{ и } |\Delta| \leqslant d \\ \text{и } \big(\Delta > 0 \text{ или } (\Delta = 0 \text{ и } n(b_d,\mathbb{X}) \leqslant n(b_0,\mathbb{X}))\big)\big) \end{array} \right\}; \tag{20}$$

$$\Omega''_p = \left\{ (d,\Delta,i) \in \Omega_p \;\middle|\; \begin{array}{c} b_d \in \mathbb{A}_c\backslash\mathbb{A} \text{ или } \big(b_d \in \mathbb{A} \text{ и } 0 \leqslant i \leqslant d \text{ и } |\Delta| \leqslant d \\ \text{и } \big(\Delta > 0 \text{ или } (\Delta = 0 \text{ и } n(b_d,\mathbb{X}) < n(b_0,\mathbb{X}))\big)\big) \end{array} \right\}. \tag{21}$$

Теорема 8. *Пусть даны метод ПМЭР μ, множество $\mathbb{X}$ мощности L, объем обучающей выборки l и прямая последовательность $\mathbb{A} = \{a_0,\ldots,a_P\}$. Пусть прямая цепь $\mathbb{A}_c = \{c_0,\ldots,c_{|\mathbb{D}|}\}$ является интерполяцией последовательности $\mathbb{A}$. Каждому классификатору $a_p \in \mathbb{A}$ соответствует $c_{i_p} \in \mathbb{A}_c$.*

Тогда для каждого $p = 1,\ldots,P$ в каждой точке (t',e') множества Ψ'_p, определенного в (14), число $L_p(t',e')$ допустимых разбиений множества $\mathbb{L}_p$, определяемое по (12), равно

$$L_p(t',\,e') = T_p(|\mathbb{L}_p|,\,t'-2e',\,e') \tag{22}$$

и вычисляется рекуррентно по правилам, описанным в лемме 4, где $b_d = c_{i_p-d}$ для каждого d и множество Ω'_p определено по (20). Краевые условия $L_0(0,0) = 1$.

Для каждого $p = 0,\ldots,P-1$ в каждой точке (t'',e'') множества Ψ''_p, определенного в (15), число $R_p(t'',e'')$ допустимых разбиений множества $\mathbb{R}_p$, определяемое по (13), равно

$$R_p(t'',e'') = T_p(|\mathbb{R}_p|,t''-2e'',e'') \tag{23}$$

и вычисляется рекуррентно по правилам, описанным в лемме 4, с заменой множества Ω'_p на Ω''_p, определенного по (21), и b_d на c_{i_p+d} для каждого d. Краевые условия $R_P(0,0) = 1$.

4. Алгоритм вычисления вероятности переобучения и полного скользящего контроля

Итак, в теореме 8 описан алгоритм нахождения количества допустимых разбиений множеств ребер правой и левой последовательностей для каждого p. Остается подставить найденные значения в формулы (11), (5) и (9). Для сокращения вычислений по теоремам 2 и 3 для каждого p предлагается заранее вычислить $L_p(t',e')$, $R_p(t'',e'')$, $N_p(t,e)$ и $F_p(t,e)$, после чего сложить полученные значения. Схема вычислений показана в алгоритме 1.

4.1. Сложность алгоритма. Оценим сложность выполнения шагов 5–11 алгоритма 1.

При вычислении $L_p(t',e')$ по теореме 6 на шагах 5–6 один раз для всех $(d,\Delta,i) \in \Omega'_p$ вычисляются $T_p(d,\Delta,i)$, затем для каждого $(t',e') \in \Psi'_p$ величина $L_p(t',e')$ полагается равной $T_p(d,t'-2e',e')$. Множество Ω'_p вложено в куб со стороной $O(|\mathbb{L}_p|)$, поскольку каждая координата ограничена по модулю количеством ребер в левой последовательности. Следовательно, сложность выполнения шагов 5–6 составляет $O(|\mathbb{L}_p|^3)$. Аналогично, сложность выполнения шагов 7–8 составляет $O(|\mathbb{R}_p|)^3)$.

Для нахождения $N_p(t,e)$ и $F_p(t,e)$ необходимо вычислить биномиальные коэффициенты C_m^i и C_{L-P-m}^i при всех возможных i за $O(L)$. Биномиальные коэффициенты для каждого p не пересчитываются. При известных значениях биномиальных коэффициентов искомые $N_p(t,e)$ и $F_p(t,e)$ вычисляются за $O(L)$. Множество Ψ_p вложено в квадрат со стороной L, значит, выполнение шагов 9–11 выполняется за $O(L^3)$. Следовательно, сложность выполнения шагов 5–11 составляет $O(|\mathbb{D}|^3 + L^3) = O(L^3)$ для каждого p.

Множества Ψ'_p и Ψ''_p вложены в квадрат со стороной P, значит, шаги 12–13 выполняются за $O(L^5)$, и сложность алгоритма 1 также составляет $O(L^5)$.

Алгоритм 1: Вычисление вероятности переобучения и полного скользящего контроля

Вход: матрица ошибок прямой последовательности $\mathbb{A} = \{a_0, \ldots, a_P\}$, параметры l, ε.

Выход: вероятность переобучения Q_ε и полный скользящий контроль CCV.

построить прямую цепь $\mathbb{A}_c$ — интерполяцию последовательности $\mathbb{A}$;
определить m по (4);
для всех $p = 0, \ldots, P$
  разделить цепь $\mathbb{A}_c$ на две — левую $\{a_p, \ldots, a_0\}$ и правую $\{a_p, \ldots, a_P\}$;
  для всех *точек* (t', e') *множества* Ψ'_p, *определенного по* (14)
    найти $L_p(t', e')$ по формулам (22), (18) и (20);
  для всех *точек* (t'', e'') *множества* Ψ''_p, *определенного по* (15)
    найти $R_p(t'', e'')$ по формулам (23), (18) и (21);
  для всех *точек* (t, e) *множества* Ψ_p, *определенного по* (6)
    вычислить $N_p(t, e)$ по формуле (7);
    вычислить $F_p(t, e)$ по формуле (10);

$$Q_\varepsilon := \frac{1}{C_L^l} \sum_{p=0}^{P} \sum_{(t',e')\in\Psi'_p} \sum_{(t'',e'')\in\Psi''_p} L_p(t', e') R_p(t'', e'') N_p(t' + t'', e' + e'');$$

$$CCV := \frac{1}{(L-l)C_L^l} \sum_{p=0}^{P} \sum_{(t',e')\in\Psi'_p} \sum_{(t'',e'')\in\Psi''_p} L_p(t', e') R_p(t'', e'') F_p(t' + t'', e' + e'');$$

5. Сравнение с существующими оценками вероятности переобучения

Рассмотрим семейство одномерных пороговых решающих правил в задаче классификации с классами равной мощности. Покажем, что для данной задачи существующие верхние оценки вероятности переобучения являются завышенными.

На рис. 3 в логарифмической шкале отложены значения оценки Вапника–Червоненкиса [1], оценки расслоения-связности [12] и оценки Соколова [17] в сравнении с точной верхней оценкой вероятности переобучения прямой последовательности. Оценка расслоения-связности и Соколова является точной только в одном случае, когда минимальное количество ошибок совпадает с параметром m. В этом случае граница между классами определяется четко, и семейство является унимодальной цепью [9]. С увеличением минимального количества ошибок оценка Соколова начинает превосходить точную верхнюю оценку. Оценка Вапника–Червоненкиса для рассматриваемой последовательности оказывается завышенной при любом значении минимального количества ошибок.

6. Заключение

Введено понятие финитного метода обучения, для которого разработан алгоритм вычисления вероятности переобучения и полного скользящего контроля прямых последовательностей классификаторов, порождаемых элементарными пороговыми правилами при варьировании параметра порога. Показано, что финитными являются метод минимизации эмпирического риска (МЭР) и метод максимизации переобученности (МП). Для МЭР показано, что существующие верхние оценки вероятности переобучения прямых последовательностей являются завышенными и неприменимыми для реальных задач.

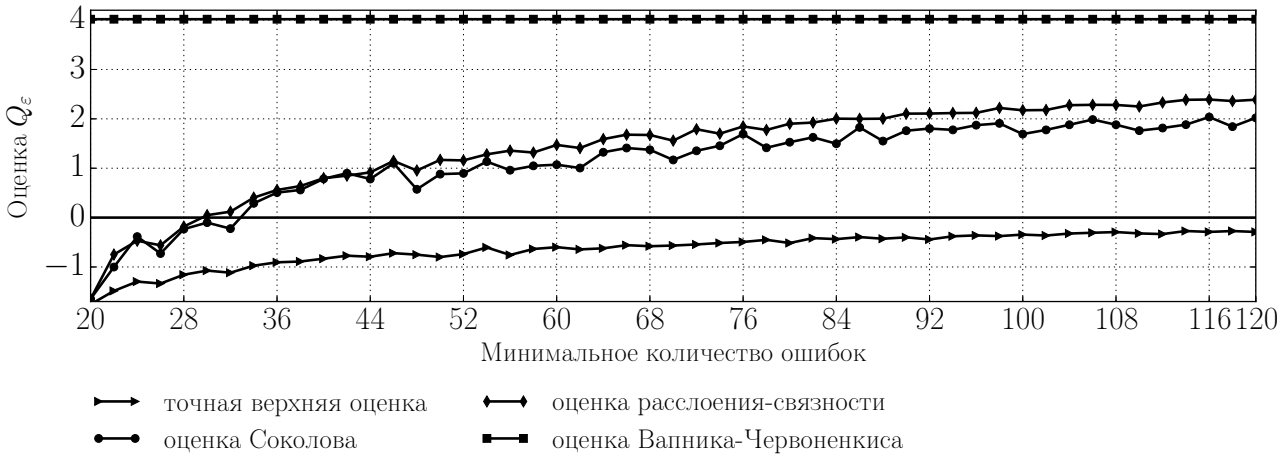


Рис. 3: Сравнение верхних оценок вероятности переобучения в логарифмической шкале. Горизонтальной линией указано значение $Q_\varepsilon = 1$. Условия эксперимента: $L = 240$, $\ell = 160$, $m = 20$, $\varepsilon = 0.05$. По горизонтали отложено минимальное количество ошибок классификаторов

Задачей будущего исследования является применение данного алгоритма для повышения обобщающей способности методов статистического обучения, в частности, для совершенствования критериев отбора признаков, методов поиска логических закономерностей в данных, линейных и логических алгоритмов классификации. Другим направлением работы является обобщение данного алгоритма на другие функционалы обобщающей способности, в частности, на функционал ожидаемой переобученности метода МП, равный радемахеровской сложности семейства и связывающий комбинаторную теорию переобучения с теорией эмпирических процессов и с теорией неравенств концентрации вероятностной меры.

Автор выражает глубокую признательность научному руководителю К. В. Воронцову за постоянное внимание к работе и ценным замечаниям.

СПИСОК ЛИТЕРАТУРЫ

1. Вапник В.Н., Червоненкис А.Я. *О равномерной сходимости частот появления событий к их вероятностям* // Теория вероятностей и ее применения. 1971. Т. 16, № 2. С. 264–280.
2. S. Boucheron, O. Bousquet, G. Lugosi *Theory of classification: A survey of some recent advances* // ESAIM: Probability and Statistics. 2005. Vol. 9. P. 323–375.
3. V. Koltchinskii *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: École d'Été de Probabilités de Saint-Flour XXXVIII-2008.* Lecture Notes in Mathematics. Springer, 2011.
4. K. V. Vorontsov *Combinatorial probability and the tightness of generalization bounds* // Pattern Recognition and Image Analysis. 2008. Vol. 18, no. 2. P. 243–259.
5. D. Haussler, N. Littlestone, M.K. Warmuth *Predicting* $\{0,1\}$*-functions on randomly drawn points* // Inf. Comput. December 1994. Vol. 115. P. 248–292.
6. Воронцов К.В. *Комбинаторные оценки качества обучения по прецедентам* // Доклады РАН. 2004. Т. 394, № 2. С. 175–178.
7. Воронцов К.В. *Точные оценки вероятности переобучения* // Доклады РАН. 2009. Т. 429, № 1. С. 15–18.
8. K.V. Vorontsov *Splitting and similarity phenomena in the sets of classifiers and their effect on the probability of overfitting* // Pattern Recognition and Image Analysis. 2009. Vol. 19, no. 3. P. 412–420.

9. K.V. Vorontsov *Exact combinatorial bounds on the probability of overfitting for empirical risk minimization* // Pattern Recognition and Image Analysis. 2010. Vol. 20, no. 3. P. 269–285.
10. V. Koltchinskii *Rademacher Penalties and Structural Risk Minimization* // IEEE Trans. Inf. Theory. 2001. Vol. 47, no. 5. P. 1902–1914.
11. K.V. Vorontsov *Combinatorial Theory of Overfitting: How Connectivity and Splitting Reduces the Local Complexity* // 9th IFIP WG 12.5 International Conference, AIAI 2013, Paphos, Cyprus, September 30 – October 2, 2013, Proceedings. Springer-Verlag Berlin Heidelberg, 2013.
12. K.V. Vorontsov, 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. Lecture Notes in Computer Science. Springer-Verlag, 2011. P. 66–73.
13. Животовский Н.К., Воронцов К.В. *Критерии точности комбинаторных оценок обобщающей способности* // Интеллектуализация обработки информации (ИОИ-2012): Докл. Москва: Торус Пресс, 2012. С. 25–28.
14. Журавлёв Ю.И., Рязанов В.В., Сенько О.В. *«Распознавание». Математические методы. Программная система. Практические применения.* М.: Фазис, 2006. 176 с.
15. Журавлёв Ю.И. *Об алгебраическом подходе к решению задач распознавания или классификации* // Проблемы кибернетики: Вып. 33. 1978. С. 5–68.
16. Гуз И.С. *Конструктивные оценки полного скользящего контроля для пороговой классификации* // Математическая биология и биоинформатика. 2011. Т. 6, № 2. С. 173–189.
17. Воронцов К.В., Фрей А.И., Соколов Е.А. *Вычислимые комбинаторные оценки вероятности переобучения* // Машинное обучение и анализ данных. 2013. Т. 1, № 6. С. 734–743.
18. Фрей А.И., Толстихин И.О. *Комбинаторные оценки вероятности переобучения на основе кластеризации и покрытий множества алгоритмов* // Машинное обучение и анализ данных. 2013. Т. 1, № 6. С. 761–778.
19. Фрей А.И., Толстихин И.О. *Комбинаторные оценки вероятности переобучения на основе покрытий множества алгоритмов* // Доклады РАН. 2014. Т. 455, № 3. С. 265–268.

Шаура Хабировна Ишкина,
ФИЦ «Информатика и управление» РАН,
ул. Вавилова, д. 44/2
119333, г. Москва, Россия
E-mail: shaura-ishkina@yandex.ru