

Интеллектуальные системы управления, анализ данных

© 2021 г. Ш.Х. ИШКИНА (shaura-ishkina@yandex.ru)
(Вычислительный центр им. А.А. Дородницына ФИЦ ИУ РАН, Москва),
К. В. ВОРОНЦОВ, д-р физ.-мат. наук (vokov@forecsys.ru)
(Московский физико-технический институт)

ИССЛЕДОВАНИЕ ЗАВЫШЕННОСТИ ОЦЕНОК ПЕРЕОБУЧЕНИЯ ПОРОГОВЫХ РЕШАЮЩИХ ПРАВИЛ¹

Данная статья посвящена проблеме вычисления точной верхней оценки функционалов обобщающей способности семейства одномерных пороговых решающих правил. Исследуется алгоритм, решающий поставленную задачу, полиномиальный по общему числу объектов выборки и по объему обучающей выборки. Доказывается теорема для вычисления оценки функционала ожидаемой переобученности семейства и оценки частоты ошибок метода минимизации эмпирического риска на контрольной выборке. Проводится сравнение точных оценок, вычисленных с помощью теоремы, с известными ранее быстро вычислимыми верхними оценками с целью оценить порядки их завышенности и выявить те оценки, которые можно было бы использовать в реальных задачах.

Ключевые слова: пороговый классификатор, обобщающая способность, комбинаторная теория, вероятность переобучения, полный скользящий контроль, Радемахеровская сложность.

DOI: 10.31857/S000523102105010X

1. Введение

При решении задачи обучения на основании обучающей выборки объектов, часто называемой обучением по прецедентам, строится алгоритм, восстанавливающий зависимость выходных переменных от входных на объектах из обучающей выборки. В задаче классификации выходная переменная одна и принимает бинарные значения, а алгоритмы называются классификаторами. Для успешного применения построенного классификатора он должен иметь высокую обобщающую способность, т.е. хорошо работать на произвольных объектах, не обязательно входящих в обучение. Если же качество классификатора на независимой выборке, называемой контрольной, оказывается значительно хуже, чем на обучающей выборке, то говорят, что произошло переобучение. В качестве характеристик обобщающей способности используются функционалы вероятности переобучения и полного скользящего контроля [1].

¹ Исследование выполнено при финансовой поддержке гранта Правительства РФ для государственной поддержки научных исследований, проводимых под руководством ведущих ученых (соглашение № 075-15-2019-1926).

Получение оценок обобщающей способности семейства классификаторов на основе информации об обучающей выборке и о структуре семейства с публикации [2] остается одной из основных задач теории статистического обучения. В конце 70-х годов XX века советские ученые В.Н. Вапник и А.Я. Червоненкис сформулировали основные статистические проблемы обучения в терминах проблемы минимизации среднего риска, т.е. вероятности ошибки классификатора на новом объекте, и предложили методы оценки среднего риска по эмпирическим данным [3, 4]. Вапник и Червоненкис получили равномерные по семействам классификаторов оценки, связывающие вероятность уклонения среднего риска от эмпирического с длиной обучающей выборки и сложностью семейства, над которыми минимизируется средний риск. Этот фундаментальный результат активно используется и сегодня.

Однако оценки Вапника–Червоненкиса являются завышенными. Экспериментальные исследования показывают, что оценки могут быть завышены на 6–12 порядков [5], что может приводить к неоптимальному выбору структурных параметров [6–8]. Кроме того, завышенные оценки не дают возможности исследовать явление переобучения, оценивать и контролировать его значения при решении реальных задач. В [5] исследуются причины завышенности оценок, из которых основной является независимость оценок от конкретной выборки. Оценка Вапника–Червоненкиса универсальна и, следовательно, является оценкой худшего случая.

Теория статистического обучения продолжает активно развиваться, последователи теории занимаются повышением точности равномерных оценок с учетом особенностей данных и конкретных алгоритмов классификации [9, 10]. Среди плодотворных подходов можно выделить оценки, адаптирующиеся к данным и использующие понятие Радемахеровской сложности, предложенной в 1999 г. В. Колчинским [11].

В комбинаторной теории [12], предложенной К.В. Воронцовым, для обобщающей способности была получена оценка расслоения–связности [13], учитывающая особенности способа построения классификатора по обучающей выборке, а также локальные свойства семейства классификаторов – эффекты расслоения и связности [14]. В [15] получены условия, при которых оценка расслоения–связности является точной. Имеются уточнения данной оценки за счет учета попарного взаимодействия классификаторов [16] и за счет учета сходства классификаторов [17], но оценки по-прежнему остаются завышенными.

В комбинаторной теории вероятностью переобучения называют долю разбиений конечного множества объектов на обучающую и контрольную выборки фиксированной длины, при которых произошло переобучение. Данное определение ранее появлялось в [18] для частного случая контрольной выборки, состоящей из одного объекта.

Точность эмпирических оценок функционалов обобщающей способности, полученных методом Монте-Карло, зависит от числа случайных разбиений. Вычисление точных оценок по определению требует экспоненциального по общему количеству объектов перебора всех возможных разбиений. Но для некоторых модельных семейств классификаторов удается аналитически вы-

числить точные оценки вероятности переобучения. К настоящему времени точные оценки получены для слоев и интервалов булева куба, монотонных и унимодальных цепей [19] и многомерных сетей [20], хэмминговых шаров и некоторых их разреженных подмножеств [21]. Разработан теоретико-групповой подход [22], который позволяет получать точные оценки для семейств с произвольными симметриями.

В [23] предложен способ аппроксимации вероятности переобучения стандартных методов классификации (нейронных сетей, решающих деревьев, ближайшего соседа) на реальных задачах с помощью монотонных сетей подходящей размерности. Оценки переобучения могут использоваться в качестве критерия отбора признаков при построении элементарных конъюнкций в логических алгоритмах классификации [13] или в качестве критерия ветвления в решающих деревьях [20].

В данной статье, являющейся продолжением публикации [24], рассматривается задача построения точных верхних оценок обобщающей способности одномерных пороговых решающих правил.

Практический интерес связан с тем, что данная задача возникает при построении алгоритмов классификации, в частности в решающих деревьях, логических закономерностях [25], алгоритмах вычисления оценок [26], а также при построении линейных классификаторов методом покоординатной оптимизации.

Теоретический интерес связан с тем, что в рамках комбинаторного подхода до сих пор не удавалось получать точные оценки обобщающей способности для данного семейства. Точные оценки были известны только для частных случаев задач классификации, где классы были линейно разделимы [19]. Также для частного случая, когда признак принимает попарно различные значения на объектах, был известен алгоритм вычисления верхней и нижней оценок обобщающей способности [27]. Однако завышенность верхней оценки остается неизученной.

В [24] был предложен алгоритм вычисления точных оценок вероятности переобучения и полного скользящего контроля одномерных пороговых решающих правил, полиномиальный по общему количеству объектов. Алгоритм допускает наличие связей во множестве значений признака и различные методы обучения.

В данной статье демонстрируется универсальность алгоритма по отношению к вычисляемым функционалам обобщающей способности. Доказывается теорема для вычисления оценки функционала ожидаемой переобученности порогового классификатора, т.е. разности доли ошибочно классифицированных объектов на обучающей и контрольной выборках.

Недостатком алгоритма является его большая вычислительная сложность. В данной статье проводится сравнение точных оценок, вычисленных с помощью алгоритма, с известными ранее быстро вычислимыми верхними оценками, с целью оценить порядки их завышенности и выявить те оценки, которые можно было бы использовать в реальных задачах.

2. Основные определения

Пусть задано конечное множество $\mathbb{X} = \{x_1, \dots, x_L\}$, элементы которого называются *объектами*, и конечное множество $\mathbb{A}$, элементы которого называются *классификаторами*. Множество $\mathbb{A}$ называется *семейством классификаторов*.

Пусть задана функция $I: \mathbb{A} \times \mathbb{X} \rightarrow \{0, 1\}$, называемая *индикатором ошибки*. Если $I(a, x) = 1$, то говорят, что на элементе x классификатор a *допускает ошибку*. Если $I(a, x) = 0$, то говорят, что классификатор *не ошибается* на данном элементе.

Предполагается, что каждому классификатору $a \in \mathbb{A}$ взаимно однозначно соответствует его вектор ошибок $(I(a, x_i))_{i=1}^L$, т.е. два классификатора с одинаковыми векторами ошибок отождествляются. Далее через a будет обозначаться как сам 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

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

где через $|X|$ обозначен объем выборки X .

Пусть множество $\mathbb{X}$ разбито на две выборки X и $\bar{X} = \mathbb{X} \setminus X$. Выборка X называется *обучающей*, выборка $\bar{X}$ – *контрольной*. *Переобученностью* классификатора a на двух выборках X и $\bar{X} = \mathbb{X} \setminus X$ называется величина

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

Методом обучения называется отображение $\mu: 2^{\mathbb{X}} \rightarrow \mathbb{A}$, которое обучающей выборке ставит в соответствие классификатор из заданного семейства.

Пусть $[\mathbb{X}]^\ell$ – множество всех выборок $X \subset \mathbb{X}$ без возвращения объема $\ell < L$. Введем на $[\mathbb{X}]^\ell$ равномерное распределение

$$P(X) = \frac{1}{C_L^\ell}.$$

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

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

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

Полным скользящим контролем называется функционал, равный математическому ожиданию числа ошибок метода обучения на контрольной выборке:

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

Ожидаемой переобученностью называется функционал, равный математическому ожиданию переобученности классификатора, выбранного методом обучения:

$$EOF(\mu, \mathbb{A}, \mathbb{X}, \ell) = \mathbb{E}\delta(\mu X, \bar{X}) = \frac{1}{C_L^\ell} \sum_{X \in [\mathbb{X}]^\ell} \nu(\mu X, \bar{X}) - \nu(\mu X, X).$$

Для краткости параметры, от которых зависят данные величины, опускаются.

В данной статье рассматривается метод обучения, *минимизирующий эмпирический риск* (МЭР)

$$\mu X \in M(X) = \operatorname{Arg} \min_{a \in \mathbb{A}} n(a, X),$$

и метод обучения, *максимизирующий переобученность* (МП)

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

Для получения верхних оценок Q_ε , CCV и EOF вводится понятие метода *пессимистичной минимизации эмпирического риска* (ПМЭР)

$$\mu X = \arg \max_{a \in M(X)} n(a, \mathbb{X}).$$

Это метод МЭР, который в случае неоднозначности среди $M(X)$ выбирает классификатор с наибольшим числом ошибок на множестве $\mathbb{X}$ [19].

3. Постановка задачи

Пусть дано семейство $\mathbb{A} = \{a_0, \dots, a_P\}$. Рассмотрим множество объектов, по которым различаются соседние классификаторы:

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

Определение 1. Прямой последовательностью называется семейство классификаторов, обладающее свойством

$$G_p \cap G_d = \emptyset \quad \forall p, d = 0, \dots, P-1.$$

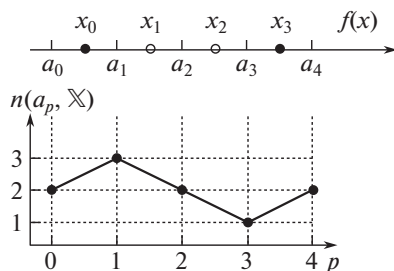


Рис. 1. Пример семейства одномерных пороговых решающих правил. По оси отложены объекты $x \in \mathbb{X}$, упорядоченные по значениям признака $f(x)$. Объекты из разных классов обозначены точками $\bullet$ и $\circ$. Пороги расположены между соседними объектами, каждому значению порога соответствует свой классификатор. Ниже оси изображен график числа ошибок классификаторов.

Определение 2. Прямая последовательность называется прямой цепью, если каждая пара соседних классификаторов различается по одному объекту

$$|G_p| = 1, \quad p = 0, \dots, P - 1.$$

В [24] доказано, что прямые последовательности порождаются семейством одномерных пороговых решающих правил $a_\theta(x) = [f(x) \geq \theta]$ над признаком f , полученных варьированием порога θ вдоль значений признака. В случае когда числовой признак принимает попарно различные значения на объектах множества $\mathbb{X}$, семейство является прямой цепью.

На рис. 1 изображен пример данного семейства. По горизонтальной оси отложены объекты множества $\mathbb{X}$, упорядоченные по значениям признака f . Объекты класса 0 обозначены точками $\circ$, объекты класса 1 — точками $\bullet$. Пороги θ расположены между соседними объектами, каждому порогу соответствует свой классификатор. Объекты, расположенные правее порога, классификатор относит к классу 1. Ниже оси изображен график числа ошибок классификаторов. Например, число ошибок классификатора a_2 равно 2: классификатор ошибочно относит объект x_0 к классу 0, а объект x_2 — к классу 1.

Покажем, что переобучение цепи зависит от геометрической структуры классов.

Эмпирической оценкой функции $\phi(X, \bar{X})$, не зависящей от порядка элементов в выборках X и $\bar{X}$, называется величина $\hat{E}\phi(X, \bar{X})$, полученная методом Монте-Карло путем усреднения по некоторому случайному подмножеству выборок $N \subset [\mathbb{X}]^l$

$$\hat{E}\phi(X, \bar{X}) = \frac{1}{|N|} \sum_{X \in N} \phi(X, \bar{X}).$$

На рис. 2 изображены прямые цепи различной формы. График частоты ошибок классификаторов на полном множестве изображен линией с точкой. Цепи упорядочены по возрастанию количества классификаторов в нижних

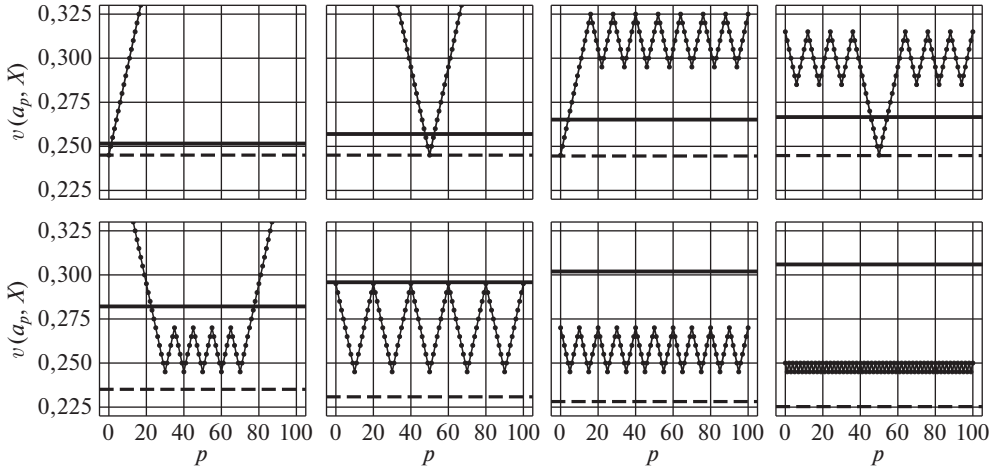


Рис. 2. Сравнение переобучения прямых цепей различной формы. По горизонтали отложены номера p классификаторов цепи. Условия эксперимента: $L = 200$, $\ell = 150$, $\varepsilon = 0,05$. Минимальная частота ошибок равна 0,245.

слоях, где слои определяются по числу ошибок. Пилообразные участки соответствуют шуму в данных, где объекты из разных классов чередуются друг с другом. Чем чаще пила, тем выше уровень шума. Рассматриваются случаи, когда пилообразные участки расположены вдали от границы классов (верхний ряд) и вблизи от границы (нижний ряд).

Горизонтальными линиями показаны эмпирические оценки частоты ошибок метода ПМЭР, вычисленные методом Монте-Карло по 10^5 разбиениям. Пунктирной линией изображена оценка на обучающей выборке $\hat{E}\nu(\mu X, X)$ и сплошной линией – оценка на контрольной выборке $\hat{E}\nu(\mu X, \bar{X})$. Чем больше расстояние между ними, равное $\hat{E}\delta(\mu X, X)$, тем сильнее переобучается семейство.

Данный эксперимент показывает, что одни семейства переобучаются значительно сильнее, чем другие: переобучение тем выше, чем больше классификаторов находится в нижних слоях семейства и чем более они различны.

Вследствие этого ставится следующая задача: для прямой последовательности $\mathbb{A}$ общего вида, методов обучения ПМЭР и МП вычислить точные значения вероятности переобучения, полного скользящего контроля и ожидаемой переобученности.

4. Обзор известных оценок вероятности переобучения

При построении оценок используется понятие *гипергеометрической функции распределения*

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

где через $\lfloor x \rfloor$ обозначена целая часть x , т.е. наибольшее целое число, не превосходящее x . Гипергеометрическая функция распределения $H_L^{\ell, m}(s)$ 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 . Предполагается $C_n^i = 0$ при невыполнении условия $0 \leq i \leq n$.

4.1. Оценка Вапника–Червоненкиса

Верхняя оценка вероятности переобучения, полученная Вапником и Червоненкисом, является функцией от мощности множества объектов и сложности семейства алгоритмов. Мерой сложности на заданном множестве объектов является коэффициент разнообразия, определяемый как число попарно различных бинарных векторов ошибок, индуцируемых классификаторами семейства. В комбинаторной теории коэффициент разнообразия равен мощности семейства.

Теорема 1 (см. [2]). Для любого метода обучения, множества $\mathbb{X}$, семейства классификаторов $\mathbb{A}$, объема ℓ обучающей выборки и порога $\varepsilon \in (0, 1)$

$$Q_\varepsilon \leq |\mathbb{A}| \max_{n=1, \dots, L} H_L^{\ell, n} \left(\frac{\ell}{L} (n - \varepsilon(L - \ell)) \right).$$

4.2. Оценка расслоения–связности

В комбинаторном подходе учитываются геометрические свойства булевых векторов ошибок классификаторов – расслоение и связность.

Под *расслоением* семейства понимается распределение классификаторов по слоям ошибок. *Слоем* называется множество классификаторов, допускающих на множестве $\mathbb{X}$ равное число ошибок. Чем меньше ошибок допускает классификатор, тем ниже его слой.

Связность предполагает, что для каждого классификатора в семействе найдется множество похожих классификаторов, отличающихся от него только на одном объекте выборки.

Пусть дано семейство классификаторов $\mathbb{A} = \{a_0, \dots, a_P\}$ с известными векторами ошибок на множестве $\mathbb{X}$. На множестве классификаторов, как векторов ошибок, существует отношение лексикографического порядка $\leq$. Говорят, что классификатор a *предшествует* b и записывают $a \prec b$, если $a \leq b$ и расстояние Хемминга между ними равно 1.

Для каждого $a \in \mathbb{A}$ через u , q и n будут обозначаться:

$$\begin{aligned} u &= |\{b \in \mathbb{A} \mid a \prec b\}|, \\ q &= |\{x \in \mathbb{X} \mid I(a, x) = 1, \exists b \prec a : I(b, x) = 0\}|, \\ n &= n(a, \mathbb{X}). \end{aligned}$$

Теорема 2 (см. [13]). Для произвольного множества $\mathbb{X}$, семейства классификаторов $\mathbb{A}$, метода обучения ПМЭР, объема ℓ обучающей выборки и по-

рога $\varepsilon \in (0, 1)$ справедливы оценки

$$(2) \quad \begin{aligned} Q_\varepsilon &\leq \sum_{a \in \mathbb{A}} \frac{C_{L-u-q}^{l-u}}{C_L^\ell} H_{L-u-q}^{\ell-u, n-q} \left(\frac{\ell}{L} (n - \varepsilon(L - \ell)) \right), \\ CCV &\leq \sum_{a \in \mathbb{A}} \frac{C_{L-u-q}^{l-u}}{C_L^\ell} \left(\frac{n}{L - \ell} - \frac{(n - q)(\ell - u)}{(L - u - q)(L - \ell)} \right). \end{aligned}$$

В [16] оценка была улучшена за счет более тонкого анализа эффекта связности.

Теорема 3 (см. [16]). Пусть S – множество истоков семейства $\mathbb{A}$, т.е. множество классификаторов s таких, что нет классификаторов $a \prec s$. Тогда верна оценка

$$(3) \quad Q_\varepsilon \leq \sum_{p=0}^P \min_{s \in S} \left\{ \sum_{i=0}^{\min\{|A_{ps}|, |B_{ps}|\}} \frac{C_{|B_{ps}|}^i C_{L-u-|B_{ps}|}^{\ell-u-i}}{C_L^\ell} \times \right. \\ \left. \times H_{L-u-|B_{ps}|}^{\ell-u-i, n-|B_{ps}|} \left(\frac{\ell}{L} (n - \varepsilon(L - \ell)) - i \right) \right\},$$

где

$$\begin{aligned} A_{ij} &= \{x \in \mathbb{X} \mid I(a_i, x) = 0, I(a_j, x) = 1\}, \\ B_{ij} &= \{x \in \mathbb{X} \mid I(a_i, x) = 1, I(a_j, x) = 0\}. \end{aligned}$$

5. Обзор известных оценок полного скользящего контроля

В комбинаторной теории наряду с оценкой расслоения–связности (2) для частного случая семейства имеются оценки Гуза.

5.1. Оценки Гуза

Рассмотрим семейство одномерных пороговых решающих правил над числовым признаком, принимающим попарно различные значения на объектах множества $\mathbb{X}$. Пусть порог пробегает все возможные значения. Для данного семейства в [27] был предложен полиномиальный алгоритм вычисления верхней и нижней оценок полного скользящего контроля.

Теорема 4 (см. [27]). Для произвольного множества $\mathbb{X}$, семейства классификаторов $\mathbb{A}$, метода обучения ПМЭР, объема ℓ обучающей выборки справедливы верхняя и нижняя оценки полного скользящего контроля:

$$\frac{1}{L - \ell} \frac{1}{C_L^\ell} \sum_{i=1}^L |E^1(i)| \leq CCV \leq 1 - \frac{1}{L - \ell} \frac{1}{C_L^\ell} \sum_{i=1}^L |E^0(i)|,$$

где для каждого i множества $E^0(i)$ безошибочных выборок и $E^1(i)$ ошибочных выборок определены как

$$\begin{aligned} E^0(i) &= \{X \mid x_i \in \bar{X}, \forall \mu X \in M(X) I(\mu X, x_i) = 0\} \subseteq [\mathbb{X}]^\ell, \\ E^1(i) &= \{X \mid x_i \in \bar{X}, \forall \mu X \in M(X) I(\mu X, x_i) = 1\} \subseteq [\mathbb{X}]^\ell. \end{aligned}$$

6. Переобучение прямой последовательности классификаторов

Пусть дана прямая последовательность $\mathbb{A} = \{a_0, \dots, a_P\}$. Определим множество

$$(4) \quad \mathbb{D} = \{x \in \mathbb{X} \mid \exists a, a' \in \mathbb{A}: I(a, x) \neq I(a', x)\},$$

объекты которого будем называть *ребрами* последовательности $\mathbb{A}$.

Объекты множества $\mathbb{N} = \mathbb{X} \setminus \mathbb{D}$ назовем *нейтральными*. На множестве $\mathbb{N}$ классификаторы семейства допускают одинаковое число ошибок

$$(5) \quad m = n(a, \mathbb{N}) \quad \forall a \in \mathbb{A}.$$

Рассмотрим классификатор a_p . Определим величину $D_p(t, e)$, равную числу разбиений множества $\mathbb{D}$, таких что a_p является результатом обучения:

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

Такие разбиения назовем *допустимыми*.

Ранее в [24] были доказаны теоремы о выражении функционалов вероятности переобучения и скользящего контроля через мощность множества допустимых разбиений. В данной статье доказывается аналогичная теорема для функционала ожидаемой переобученности.

Теорема 5. Для методов обучения μ МП и ПМЭР, произвольной прямой последовательности классификаторов $\mathbb{A} = \{a_0, \dots, a_P\}$, множества $\mathbb{X}$ мощности L , объема обучающей выборки ℓ выражение для ожидаемой переобученности имеет вид

$$(6) \quad EOF = \frac{1}{C_L^\ell} \sum_{p=0}^P \sum_{(t,e) \in \Psi_p} D_p(t, e) F_p(t, e),$$

где множество $\mathbb{D}$ и параметр m определяются по (4) и (5) и

$$(7) \quad F_p(t, e) = \sum_{s=0}^{\min\{\ell-t, m\}} C_m^s C_{L-P-m}^{\ell-t-s} \left(\frac{1}{L-\ell} (n(a_p, \mathbb{X}) - (s+e)) - \frac{1}{\ell} (s+e) \right);$$

$$(8) \quad \Psi_p = \{(t, e) \mid 0 \leq t \leq \min\{l, |\mathbb{D}|\}, 0 \leq e \leq \min\{t, n(a_p, \mathbb{D})\}\}.$$

Доказательство данной теоремы приведено в Приложении.

Алгоритм из [24] легко дополняется формулами (6)–(8), добавляя возможность вычисления функционала ожидаемой переобученности. Во избежание дублирования текста публикации [24] полное описание алгоритма не приводится.

Алгоритм вычисления всех трех функционалов обобщающей способности для методов ПМЭР и МП реализован на языке C++ и доступен в репозитории GitHub [28].

7. Оценка частоты ошибок метода ПМЭР на контрольной выборке

В данном разделе доказывается теорема для вычисления оценок частоты ошибок классификатора, выбранного методом ПМЭР, на контрольной выборке.

Пусть задана вероятностная мера P_σ . Для семейства $\mathbb{A}$ и множества $\mathbb{X}$ определим *Радемахеровскую сложность*

$$\mathcal{R}_L(\mathbb{A}, \mathbb{X}) = 2E_\sigma \sup_{a \in \mathbb{A}} \frac{1}{L} \sum_{i=1}^L \sigma_i I(a, x_i),$$

где $I(a, x_i)$ обозначает ошибку классификатора, E_σ – математическое ожидание по мере P_σ , а радемахеровские случайные величины $\sigma_1, \dots, \sigma_L$, независимые в совокупности, определяются как

$$\sigma_i = \begin{cases} +1, P_\sigma = \frac{1}{2}, \\ -1, P_\sigma = \frac{1}{2}. \end{cases}$$

Радемахеровская сложность описывает сложность семейства классификаторов. Чем больше Радемахеровская сложность, тем лучше ошибки классификаторов семейства могут коррелировать со случайным шумом σ_i .

В [29] было продемонстрировано, что функционал ожидаемой переобученности возникает в выражении Радемахеровской сложности, таким образом связывая комбинаторную теорию с теорией эмпирических процессов и неравенств концентрации вероятностной меры:

Лемма 1 (см. [29]). *Для метода обучения μ_δ МП, конечного семейства классификаторов $\mathbb{A}$, множества $\mathbb{X}$ мощности L , объема обучающей выборки $\ell = \frac{L}{2}$ верно*

$$EOF(\mu_\delta, \mathbb{A}, \mathbb{X}, \ell) = \mathcal{R}_L(\mathbb{A}, \mathbb{X}).$$

Из определения ожидаемой переобученности и метода МП следует лемма 2.

Лемма 2. *Для методов обучения μ_δ МП и μ ПМЭР, конечного семейства классификаторов $\mathbb{A}$, множества $\mathbb{X}$ мощности L , объема обучающей выборки ℓ верно*

$$EOF(\mu, \mathbb{A}, \mathbb{X}, \ell) \leq EOF(\mu_\delta, \mathbb{A}, \mathbb{X}, \ell).$$

Из лемм 1 и 2 следует теорема 6.

Теорема 6. *Для метода обучения μ ПМЭР, конечного семейства классификаторов $\mathbb{A}$, множества $\mathbb{X}$ мощности L с вероятностью ε верно*

$$(9) \quad \nu(\mu X, \bar{X}) \leq \nu(\mu X, X) + EOF(\mu, \mathbb{A}, \mathbb{X}, \ell) + \eta(\varepsilon),$$

$$(10) \quad \nu(\mu X, \bar{X}) \leq \nu(\mu X, X) + \mathcal{R}_L(\mathbb{A}, \mathbb{X}) + \eta(\varepsilon),$$

где поправка $\eta(\varepsilon) = \sqrt{-2 \ln \frac{\varepsilon}{2}}$ и объем обучающей выборки $\ell = \frac{L}{2}$.

Доказательство данной теоремы приведено в Приложении.

Оценки, предложенные в теореме 6, различаются следующим. Из леммы 2 следует, что оценка (9) является более точной, чем оценка (10). Однако недостатком оценки (9), вычисляемой по теореме 5, является большая вычислительная сложность $O(L^5)$, тогда как для Радемахеровской сложности возможно построить быстро вычисляемую верхнюю оценку, основанную на лемме Массара [10].

В данной статье мы сравним точные значения ожидаемой переобученности методов обучения ПМЭР μ и МП μ_δ на примере семейства прямых последовательностей. Если зазор между значениями $EOF(\mu_\delta)$ и $EOF(\mu)$ окажется достаточно малым, то в правой части неравенства (10) можно заменить величину $\mathcal{R}_L(\mathbb{A}, \mathbb{X})$ на ее быстро вычисляемую оценку и использовать полученную оценку в практических задачах.

8. Вычислительные эксперименты

Напомним обозначения. Дана прямая последовательность $\mathbb{A} = \{a_0, \dots, a_P\}$, множество $\mathbb{X}$ мощности L , множество $\mathbb{D}$ ребер последовательности. Объем обучающей выборки ℓ . Точность ε . Параметр t равен числу ошибок на множестве $\mathbb{X} \setminus \mathbb{D}$.

8.1. Модельные данные

В экспериментах используются случайные прямые последовательности. Для порождения таких семейств генерируются класс нулей $X_0 \sim \mathcal{N}(0, 1)$ и класс единиц $X_1 \sim \mathcal{N}(\Delta, 1)$ как выборки равной мощности $\frac{L}{2}$ из нормальных распределений. Параметр Δ влияет на минимальное количество ошибок классификаторов семейства. Объединение выборок является множеством $\mathbb{X}$. Множество $\mathbb{D}$ соответствует значениям, которые пробегает порог.

В экспериментах по сравнению оценок полного скользящего контроля в силу ограниченности оценок Гуза рассматриваются прямые цепи. Для порождения таких цепей из выборок X_0 и X_1 удаляются совпадающие элементы. Также оценки Гуза справедливы только для случая, когда порог пробегает все возможные значения, т.е. множество $\mathbb{D}$ совпадает с $\mathbb{X}$.

8.2. Сравнение с существующими оценками вероятности переобучения

На рис. 3 в логарифмической шкале отложены значения оценки Вапника–Червоненкиса (точки ■), оценки расслоения–связности (точки ♦) и оценки

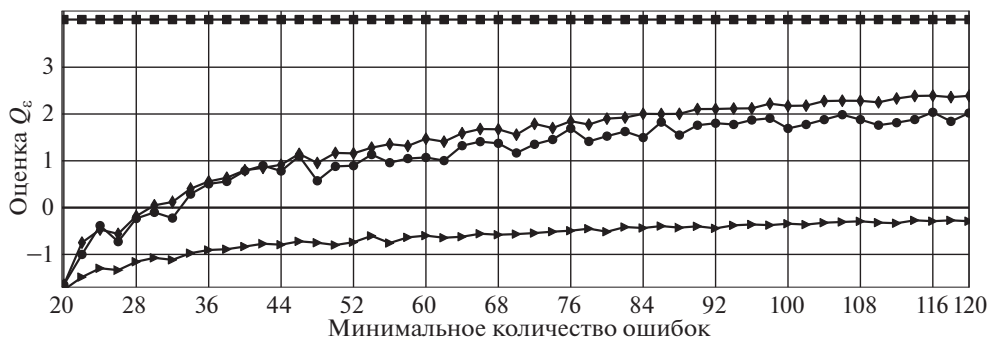


Рис. 3. Сравнение верхних оценок вероятности переобучения в логарифмической шкале. Условия эксперимента: $L = 240$, $\ell = 160$, $m = 20$, $\varepsilon = 0,05$. По горизонтали отложено минимальное количество ошибок классификаторов.

Соколова (точки $\bullet$) в сравнении с точной верхней оценкой вероятности переобучения прямой последовательности (точки $\blacktriangleright$). Горизонтальной линией указано значение $Q_\varepsilon = 1$.

Оценки расслоения–связности и Соколова являются точными только в одном случае, когда минимальное количество ошибок совпадает с параметром m . В этом случае семейство является унимодальной цепью (второй график в верхнем ряду на рис. 2), т.е. на границе классов отсутствует шум и граница определяется четко. С увеличением минимального количества ошибок оценка (3) начинает превосходить реальное значение вероятности переобучения. Оценка Валника–Червоненкиса для рассматриваемой последовательности оказывается завышенной при любом значении минимального количества ошибок.

8.3. Сравнение с существующими оценками полного скользящего контроля

На рис. 4 по вертикали в логарифмической шкале отложены значения нижней (точки $\bullet$) и верхней оценок Гуза (точки $\blacksquare$) и оценки расслоения–связности (точки $\blacklozenge$) в сравнении с точной верхней оценкой (точки $\blacktriangleright$).

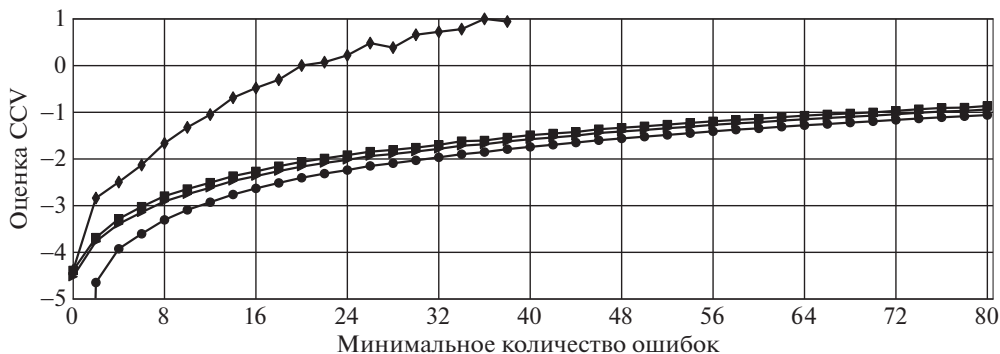


Рис. 4. Сравнение верхних оценок полного скользящего контроля в логарифмической шкале. Условия эксперимента: $L = 240$, $\ell = 160$, $m = 0$. По горизонтали отложено минимальное количество ошибок классификаторов.

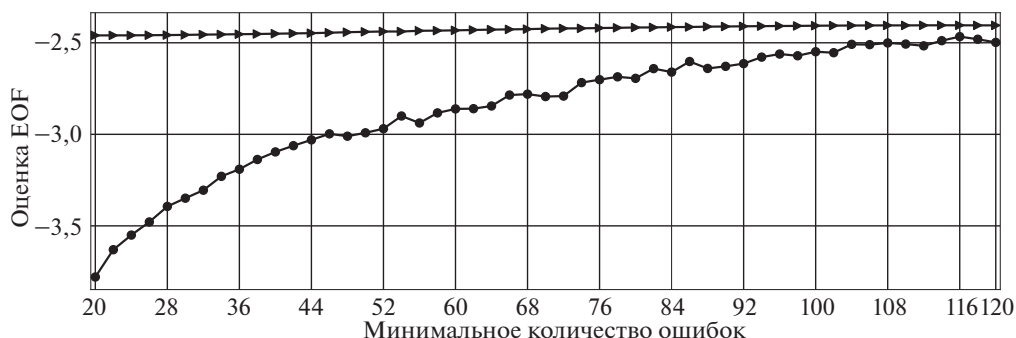


Рис. 5. Сравнение оценок ожидаемой переобученности для методов ПМЭР и МП в логарифмической шкале. Условия эксперимента: $L = 240$, $\ell = 120$, $m = 0$. По горизонтали отложено минимальное количество ошибок классификаторов.

Оценка расслоения–связности оказывается точной только в случае, когда минимальное количество ошибок равно нулю, т.е. когда классы линейно разделимы, для остальных значений параметра она является завышенной. Верхняя оценка Гуза практически совпадает с точной, из чего можно сделать вывод о высокой точности оценок.

8.4. Сравнение с Радемахеровской сложностью

На рис. 5 по вертикали в логарифмической шкале отложены значения ожидаемой переобученности классификатора, выбираемого методами обучения ПМЭР (точки $\bullet$) и МП (точки $\blacktriangleright$). Можно отметить, что с увеличением минимального числа ошибок классификаторов в семействе, т.е. с увеличением шума, два метода обучения начинают давать близкие значения ожидаемой переобученности. В этом случае для получения верхней оценки частоты ошибок классификатора, выбранного методом ПМЭР, на контрольной выборке, можно использовать оценку (10). Но при малом уровне шума ожидаемая переобученность метода ПМЭР оказывается ниже на два порядка. В этом случае для повышения точности оценок необходимо пользоваться оценкой (9).

9. Заключение

Проведено исследование обобщающей способности семейства одномерных пороговых решающих правил, определяемой с помощью функционалов вероятности переобучения, полного скользящего контроля и ожидаемой переобученности.

Показано, что имеющиеся верхние оценки вероятности переобучения являются завышенными на 1–2 порядка. Оценки Вапника–Червоненкиса не учитывают геометрическую структуру классов, от которой, как показано в экспериментах, зависит обобщающая способность семейства. Комбинаторные оценки, несмотря на то что принимают во внимание такие геометрические свойства, как расслоение и связность, все равно являются завышенными для данного семейства.

Оценки Гуза для полного скользящего контроля, полученные в рамках комбинаторной теории переобучения и вычислимые за полиномиальное от общего количества объектов время, демонстрируют высокую точность, что обосновывает возможность применения данных оценок в реальных задачах.

Оценки Радемахеровской сложности оказываются достаточно точными только для задач с высоким уровнем шума на границе классов. В противном случае, когда граница между классами определяется четко, оценки Радемахеровского типа являются завышенными на несколько порядков и неприменимыми на практике.

Поскольку имеющиеся верхние оценки обобщающей способности, за исключением оценок Гуза, не продемонстрировали высокую точность, возникает задача улучшения алгоритма вычисления точных верхних оценок и уменьшения его вычислительной сложности. Также следующей задачей является применение точных оценок для повышения качества алгоритмов классификации, в частности для модификации критериев отбора признаков в жадных алгоритмах индукции конъюнктивных логических закономерностей и других логических алгоритмах классификации.

Авторы выражают глубокую признательность рецензентам за тщательное рассмотрение и ценные замечания, которые были учтены при редактировании и способствовали улучшению изложения.

ПРИЛОЖЕНИЕ

Доказательство теоремы 5. Запишем формулу переобученности и переставим в ней знаки суммирования:

$$\begin{aligned} EOF &= \frac{1}{C_L^\ell} \sum_{X \in [\mathbb{X}]^l} \sum_{p=0}^P [\mu X = a_p] \delta(a_p, X) = \\ &= \frac{1}{C_L^\ell} \sum_{p=0}^P \sum_{X \in [\mathbb{X}]^l} [\mu X = a_p] \delta(a_p, X). \end{aligned}$$

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

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

Множество допустимых значений (t, e) есть Ψ_p согласно (8).

Обозначим $s = n(a_p, X \cap \mathbb{N})$. Из ограничений $s + t \leq l$ и $s \leq m$ следует верхняя оценка параметра s в (7).

Поскольку число ошибок классификатора a_p на контрольной выборке равно

$$\begin{aligned} n(a_p, \bar{X}) &= n(a_p, \mathbb{X}) - n(a_p, X) = \\ &= n(a_p, \mathbb{X}) - n(a_p, X \cap \mathbb{D}) - n(a_p, X \cap \mathbb{N}), \end{aligned}$$

переобученность классификатора a_p для данных s и e представляется в виде

$$\begin{aligned}\delta(a_p, X) &= \frac{1}{L-\ell} n(a_p, \bar{X}) - \frac{1}{\ell} n(a_p, X) = \\ &= \frac{1}{L-\ell} (n(a_p, \mathbb{X}) - (s+e)) - \frac{1}{\ell} (s+e).\end{aligned}$$

В [24] доказано, что выполнение условия $\mu X = a_p$ не зависит от выбора разбиения множества $\mathbb{N}$, и показано, что количество разбиений множества $\mathbb{N}$ для данных t и s равно $C_m^s C_{L-P-m}^{\ell-t-s}$, откуда следует утверждение теоремы 5.

Доказательство теоремы 6. С помощью неравенства Хёфдинга [30] с вероятностью ε можно оценить отклонение $\eta = \eta(\varepsilon)$ переобученности от ее математического ожидания:

$$\delta(\mu X, \bar{X}) \leq EOF(\mu) + \eta(\varepsilon),$$

где отклонение $\eta = \sqrt{-2 \ln \frac{\varepsilon}{2}}$.

Тогда частоту ошибок классификатора, выбранного методом ПМЭР, на контрольной выборке, можно оценить непосредственно через частоту ошибок на обучающей выборке и математическое ожидание переобученности:

$$\nu(\mu X, \bar{X}) = \nu(\mu X, X) + \delta(\mu X, \bar{X}) \leq \nu(\mu X, X) + EOF(\mu) + \eta(\varepsilon),$$

что обосновывает неравенство (9). Отсюда и из лемм 1 и 2 следует неравенство (10). Теорема 6 доказана.

СПИСОК ЛИТЕРАТУРЫ

1. *Kohavi R.* A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection // Proc. Int. Joint Conf. on Artificial Intelligence. 1995. P. 1137–1143.
2. *Вапник В.Н., Червоненкис А.Я.* О равномерной сходимости частот появления событий к их вероятностям // Теория вероятности и ее применения. 1971. Т. 16. № 2. С. 264–280.
3. *Вапник В.Н., Червоненкис А.Я.* Теория распознавания образов. М.: Наука, 1974.
4. *Вапник В.Н.* Восстановление зависимостей по эмпирическим данным. М.: Наука, 1979.
5. *Vorontsov K.V.* Combinatorial probability and the tightness of generalization bounds // Patt. Rec. and Image An. 2008. V. 18. No. 2. P. 243–259.
6. *Langford J.* Quantitatively Tight Sample Complexity Bounds: Ph.D. thesis // Carnegie Mellon Thesis, 2002.
7. *Freund Y., Schapire R.E.* A decision-theoretic generalization of on-line learning and an application to boosting // Journal of Computer and System Sciences. 1997. V. 55. No. 1. P. 119–139.
8. *Kearns M.J., et.al.* An experimental and theoretical comparison of model selection methods // Computational Learning Theory. 1995. P. 21–30.
9. *Boucheron S., Bousquet O., Lugosi G.* Theory of classification: A survey of some recent advances // ESAIM: Probability and Statistics. 2005. V. 9. P. 323–375.

10. *Shalev-Shwartz S., Ben-David S.* Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, 2014.
11. *Koltchinskii V.* 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.
12. *Воронцов К.В.* Точные оценки вероятности переобучения // Докл. РАН. 2009. Т. 429. № 1. С. 15–18.
13. *Vorontsov K.V., Ivahnenko A.A.* Tight combinatorial generalization bounds for threshold conjunction rules // 4th Int. Conf. on Pattern Recognition and Machine Intelligence, 2011. Lecture Notes in Computer Science. Springer-Verlag, 2011. P. 66–73.
14. *Vorontsov K.V.* Splitting and similarity phenomena in the sets of classifiers and their effect on the probability of overfitting // Pattern Recogn. and Image Anal. 2009. V. 19. No. 3. P. 412–420.
15. *Животовский Н.К., Воронцов К.В.* Критерии точности комбинаторных оценок обобщающей способности // Интеллектуализация обработки информации (ИОИ-2012). М.: Торус Пресс, 2012. С. 25–28.
16. *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.
17. *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.
18. *Haussler D., Littlestone N., Warmuth M.K.* Predicting 0, 1-functions on randomly drawn points // Information and Computation. 1994. V. 115. No. 2. P. 248–292.
19. *Vorontsov K.V.* Exact combinatorial bounds on the probability of overfitting for empirical risk minimization // Pattern Recogn. and Image Anal. 2010. V. 20. No. 3. P. 269–285.
20. *Botov P.V.* Exact estimates of the probability of overfitting for multidimensional modeling families of algorithms // Pattern Recogn. and Image Anal. 2010. V. 20. No. 4. P. 52–65.
21. *Толстихин И.О.* Вероятность переобучения некоторых разреженных семейств алгоритмов // Междунар. конф. ИОИ-8. М.: МАКС Пресс, 2010. С. 83–86.
22. *Frei A.I.* Accurate estimates of the generalization ability for symmetric set of predictors and randomized learning algorithms // Pattern Recogn. and Image Anal. 2010. V. 20. No. 3. P. 241–250.
23. *Ботов П.В.* Точные оценки вероятности переобучения для монотонных и уни-модальных семейств алгоритмов // Математические методы распознавания образов-14. М.: МАКС Пресс, 2009. С. 7–10.
24. *Ишкина Ш.Х.* Комбинаторные оценки переобучения пороговых решающих правил // Уфимск. матем. журн. 2018. Т. 10. № 1. С. 50–65.
25. *Журавлёв Ю.И., Рязанов В.В., Сенько О.В.* Распознавание. Математические методы. Программная система. Практические применения. М.: ФАЗИС, 2005.
26. *Журавлёв Ю.И.* Об алгебраическом подходе к решению задач распознавания или классификации // Проблемы кибернетики. 1978. Т. 33. С. 5–68.
27. *Гуз И.С.* Конструктивные оценки полного скользящего контроля для пороговой классификации // Математическая биология и биоинформатика. 2011. Т. 6. № 2. С. 173–189.

28. GitHub Project <https://github.com/shaurushka/threshold-clfs-gen-bound>
29. Vorontsov K. V. Combinatorial Theory of Overfitting: How Connectivity and Splitting Reduces the Local Complexity // 9th IFIP WG 12.5 Int. Conf., AIAI, 2013. Springer Berlin Heidelberg, 2013.
30. Hoeffding W. Probability inequalities for sums of bounded random variables // Journal of the American Statistical Association. 1963. No. 58. P. 13–30.

Статья представлена к публикации членом редколлегии А.И. Михальским.

Поступила в редакцию 29.06.2020

После доработки 09.12.2020

Принята к публикации 15.01.2021