Математическая биология и биоинформатика. 2011. Т. 6. № 2. С. 173–189.
URL: http:// www.matbio.org/2011/Guz2011(6_173).pdf

================ **ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ** ================

УДК: 519.7:004.852

Конструктивные оценки полного скользящего контроля для пороговой классификации

Гуз И.С.*

Факультет управления и прикладной математики, Московский физико-технический институт (государственный университет), 141700, Россия, Московская область, г. Долгопрудный, Институтский переулок, 9

Аннотация. Значительная часть задач классификации, в частности, задачи медицинской диагностики и биоинформатики, естественным образом сводятся к изучаемому в данной статье поиску оптимальных порогов для действительнозначных признаков. В ней приводятся комбинаторные верхняя и нижняя оценки функционала полного скользящего контроля (CCV) для одномерной задачи бинарной классификации. В качестве семейства алгоритмов рассматриваются монотонные пороговые классификаторы, учитывающие веса объектов при настройке. Обосновывается вычислительная процедура для расчета оценок CCV с полиномиальной от количества объектов сложностью. Эта процедура используется для выделения шумовых объектов, исключение которых из обучения приводит к уменьшению верхней оценки CCV.

Ключевые слова: *задачи бинарной классификации, точные оценки полного скользящего контроля, исключение шумовых объектов.*

ВВЕДЕНИЕ

Проблема синтеза алгоритмов на основе обучения по конечным выборкам прецедентов и изучения их качества на всем множестве прецедентов является одним из важнейших вопросов теории обучения систем. В качестве прецедентов рассматриваются пары: объект, описанный набором признаков, и класс, к которому принадлежит объект. Задача классификации состоит в том, чтобы на основании известного конечного множества прецедентов научиться определять априори неизвестную принадлежность объектов к классам. Обучение или настройка параметров алгоритма на обучающей выборке происходит путем решения задачи численной оптимизации. Практика показала, что качество настроенного алгоритма на контрольной выборке может существенно зависеть от разбиения всего множества прецедентов на обучающую и контрольную выборки. Таким образом, встает вопрос о качественной настройке алгоритма с точки зрения способности к обобщению.

На текущий момент существует несколько подходов к изучению данного вопроса. В основе первого подхода лежит статистическая теория Вапника–Червоненкиса [1]. В ней обобщающая способность определяется как вероятность ошибки алгоритма, полученного в результате обучения, либо как частота его ошибок на некоторой независимой и, вообще говоря, неизвестной контрольной выборке. К сожалению, оценки Вапника–Червоненкиса сильно завышены, что приводит к требованию слишком длинных обучающих выборок (10^5–10^6 объектов), а некоторые семейства алгоритмов находятся за пределами теории и для них данный подход не дает оценок. Второй

*ivan.guzz@gmail.com

подход использует комбинаторные методы, основанные на полном переборе всех возможных разбиений всего множества прецедентов на обучающую и контрольную выборки фиксированной длины, и анализе качества синтезированных алгоритмов при каждом таком разбиении. Оценки, полученные таким способом, справедливы для любого метода обучения и любой конечной выборки, не обязательно случайной, независимой, одинаково распределённой. Одной из таких оценок качества является функционал полного скользящего контроля (complete cross-validation, CCV). Этот функционал подсчитывает долю ошибок классификации синтезированного по обучающей выборке алгоритма на контрольной выборке, усредненную по всевозможным разбиениям всего множества прецедентов на выборки фиксированной длины. Функционал CCV позволяет судить о качестве конкретного семейства алгоритмов при решении задачи классификации. Оценки CCV для некоторых семейств алгоритмов были получены в [2].

Другой проблемой синтеза алгоритмов является обнаружение и исключение шумовых объектов из обучающей выборки для повышения качества классификации.

В статье исследуется часто используемое на практике семейство монотонных пороговых алгоритмов классификации. Примером этого семейства является монотонная корректирующая операция, применяемая к некоторому алгоритму бинарной классификации. Например, с помощью монотонного порогового алгоритма можно точнее выбирать порог в алгоритме логистической регрессии или бинарном решающем дереве.

Целью работы является построение оценок CCV для указанного выше алгоритмического семейства с учетом весов объектов и нахождение таких весов, при которых эта оценка минимальна. Чем больше вес объекта, тем сильнее он должен быть учтен методом обучения при синтезе алгоритма. Если вес некоторого объекта после процедуры минимизации верхней оценки CCV стал равен 0, то этот объект можно считать шумовым.

Работа состоит из четырех частей. В первой части работы вводятся определения, и формализуется постановка задачи. Вторая часть посвящена выводу верхней и нижней оценок CCV, меньших 1, для произвольного алгоритмического семейства. В основе этих оценок лежат определяемые в этой же части множества ошибочных и безошибочных выборок, определяющие, насколько часто на каждом объекте будут допускаться ошибки в случае его попадания в контроль. В третьей части обосновывается рекурсивная процедура расчета оценок CCV за полиномиальное время. В Заключении описывается алгоритм минимизации верхней оценки CCV для исключения шумовых объектов, а также результаты численных экспериментов его применения на модельной задаче. Полученные результаты позволяют судить о высокой точности полученных оценок CCV и о возможности прикладного применения доказанных методик исключения шумовых объектов.

1. ОПРЕДЕЛЕНИЯ

Пусть задано конечное множество $\mathbb{X} = \{x_0, x_1, \dots, x_{L-1}\}$, состоящее из L объектов, в котором каждый объект i описывается единственным признаком $x_i \in \mathbb{R}$, и соответствующее им множество классов $\mathbb{Y} = \{y_0, y_1, \dots, y_{L-1}\}$, в котором значения классов $y_i \in \{-1, +1\}$. Назовем множество $\mathbb{X}$ генеральной выборкой, и будем считать, что все его объекты $\mathbb{X}$ упорядочены по значениям их признаков, то есть $x_0 < x_2 < \dots < x_{L-1}$, и среди объектов нет совпадающих значений.

Пусть также задано множество A, элементы которого называются алгоритмами, где каждый алгоритм $a \in A: \mathbb{R} \to \{-1, +1\}$. Существует бинарная функция $I: A \times \mathbb{X} \to \{0,1\}$, называемая индикатором ошибки. Если $I(a,x)=1$, то будем считать, что алгоритм $a\in A$ допускает ошибку на объекте x. Вектором ошибок алгоритма $a\in A$ будем называть L-мерный бинарный вектор $\{I(a,x_i)\}_{i=0}^{L-1}$.

В качестве множества алгоритмов A в работе рассматривается семейство монотонных пороговых алгоритмов классификации. Каждый алгоритм a этого семейства описывается единственным вещественным параметром $c_a\in\mathbb{R}$, называемым порогом, и отображает произвольный объект $x\in\mathbb{R}$ в соответствующее значение класса $y=\{-1,+1\}$, по следующему правилу: $a(x)=sign(x-c_a)$.

Два алгоритма, у которых векторы ошибок на генеральной выборке $\mathbb{X}$ совпадают, являются неразличимыми на этой выборке. Будем полагать, что множество A состоит только из тех алгоритмов, у которых векторы ошибок на $\mathbb{X}$ попарно различны. Поскольку выборка одномерная, и рассматриваются монотонные пороговые классификаторы, то $|A|=L+1$. Пронумеруем алгоритмы множества $A=\{a_0,a_1,\dots,a_L\}$ в соответствии с возрастанием порога и для определенности выберем значение порога c_i для каждого алгоритма a_i по следующему правилу:

$$\begin{aligned}
&c_0=x_0-1;\\
&c_i=(x_{i-1}+x_i)/2,\quad 0<i<L-1;\\
&c_L=x_{L-1}+1.
\end{aligned}$$

Пороги можно выбирать произвольным образом, исходя из неравенств (рис. 1):

$$\begin{aligned}
&c_0<x_0;\\
&x_{i-1}<c_i<x_i,\quad 0<i<L;\\
&c_L>x_{L-1}.
\end{aligned}$$

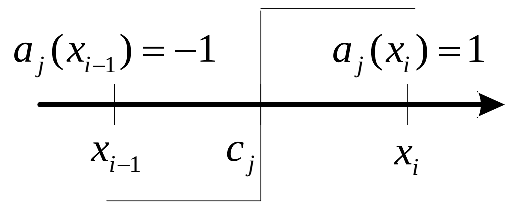

Рис. 1. Выбор порога монотонного порогового алгоритма классификации.

Для каждого алгоритма $a\in A$ определим частоту его ошибок v на произвольной выборке $X\subseteq\mathbb{X}$:

$$v(a,X)=\frac{1}{|X|}\sum_{x\in X}I(a,x).$$

Пусть ℓ – фиксированное натуральное число, $\ell<L$. Обозначим через $[\mathbb{X}]^\ell$ множество всех ℓ-элементных подмножеств генеральной выборки $\mathbb{X}$. Очевидно, что $\left|[\mathbb{X}]^\ell\right|=C_L^\ell$.

Методом обучения называется отображение $\mu:[\mathbb{X}]^\ell\to A$, которое произвольной обучающей выборке $X\in[\mathbb{X}]^\ell$ ставит в соответствие некоторый алгоритм $a\in A$. Выборка $\bar{X}=\mathbb{X}\backslash X$ называется контрольной.

Во введенных обозначениях функционал полного скользящего контроля CCV запишется следующим образом:

$$Q_k(\mu,\mathbb{X}) = \frac{1}{C_L^\ell}\sum_{X\cup\bar{X}=\mathbb{X}} \nu(\mu(X),\bar{X}) = E_{\mathbb{X}}\nu(\mu(X),\bar{X}).$$

В этой формуле символом $E_{\mathbb{X}}$ обозначена операция усреднения по всевозможным разбиениям генеральной выборки $\mathbb{X}$ на обучающую и контрольную выборки.

При решении задачи классификации объекты могут иметь разную ценность. Шумовые объекты мешают решать задачу с точки зрения обобщающей способности, тогда как неинформативные периферийные объекты не помогают. Параметризуем метод обучения μ с помощью введения множества весов объектов $\mathbb{W} = \{w_0, w_1, \ldots, w_{L-1}\}$, где вес каждого объекта $w_i \in \mathbb{Z}$ фиксирован.

Для упрощения записи будем идентифицировать объекты, входящие в обучающую выборку X, по их номеру в обучающей выборке в соответствии с их порядком с помощью верхнего индекса. Например, x^0 означает первый объект, входящий в рассматриваемую обучающую выборку X. Будем рассматривать только методы обучения из семейства $\mathrm{M} = \{\mu(X)\}$, минимизирующие взвешенный эмпирический риск:

$$\mu(X) = \arg\min_{a\in A}\left(\sum_{i=0}^{\ell-1} w^i[a(x^i)y^i < 0]\right). \qquad (0.1)$$

Докажем лемму 1, упрощающую нахождение порога оптимального алгоритма в методе обучения (0.1).

Лемма 1. Пусть метод обучения μ выбирает алгоритм из семейства монотонных пороговых классификаторов, минимизируя взвешенный эмпирический риск для выборки $X \in [\mathbb{X}]^\ell$, состоящей из ℓ упорядоченных объектов с индексами $0,1,\ldots,\ell-1$. Тогда индекс j оптимального алгоритма $a_j = \mu(X)$ равен: $j = \arg\min_{k=0,\ldots,\ell} \sum_{i=0}^{k-1} w^i y^i$.

◦ Из свойств монотонных пороговых классификаторов следует цепочка равенств:

$$[a^j(x^i)y^i < 0] = [y^i < 0][i \ge j] + [y^i > 0][i < j] = [y^i < 0](1 - [i < j]) + [y^i > 0][i < j] =$$

$$= [y^i < 0] + ([y^i > 0] - [y^i < 0])[i < j] = [y^i < 0] + y^i[i < j].$$

Преобразуем формулу минимизации взвешенного эмпирического риска (0.1):

$$\mu(X) = a_j : j = \arg\min_{k=0,\ldots,\ell}\left(\sum_{i=0}^{\ell-1} w^i[a^k(x^i)y^i < 0]\right) = \arg\min_{k=0,\ldots,\ell}\left(\sum_{i=0}^{\ell-1} w^i[y^i < 0] + w^i y^i[i < k]\right) =$$

$$= \arg\min_{k=0,\ldots,\ell}\sum_{i=0}^{k-1} w^i y^i . \qquad \bullet$$

Доказанная лемма позволяет наглядно с помощью графика определять оптимальные алгоритмы, обученные с помощью метода минимизации взвешенного эмпирического риска. Действительно, для каждого объекта обучающей выборки X отложим на графике по горизонтальной оси его индекс k, а по вертикальной оси взвешенную сумму классов объектов с индексами от 0 до $k-1$. В соответствии с порядком объектов, последовательно соединим полученные на графике точки. Согласно лемме 1 порог оптимального алгоритма следует расположить сразу после значения абсциссы, при котором значение ординаты минимально (рис. 2).

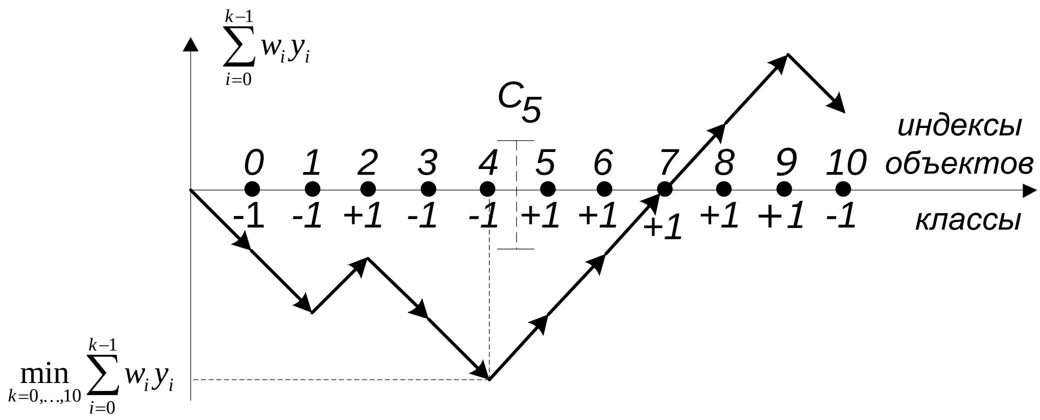

Рис 2. Поиск порога оптимального алгоритма (веса всех объектов равны 1). Алгоритм с порогом c_5 минимизирует взвешенный эмпирический риск.

В зависимости от выбора конкретного метода обучения $\mu \in \mathrm{M}$, CCV будет принимать различные значения. Максимальное значение CCV будет принимать при использовании пессимистичного метода минимизации взвешенного эмпирического риска μ^{pes}. Этот метод в случае неоднозначности выбора в (0.1) выбирает алгоритм, максимизирующий частоту ошибок на контрольной выборке. Минимальное же значение CCV будет достигаться при использовании оптимистичного метода минимизации взвешенного эмпирического риска μ^{opt}. Этот метод в случае неоднозначности выбора в (0.1) выбирает алгоритм, минимизирующий частоту ошибок на контрольной выборке. Запишем, используя введенные обозначения, достижимые оценки CCV:

$$Q_k(\mu^{opt}, \mathbb{X}) \le Q_k(\mu, \mathbb{X}) \le Q_k(\mu^{pes}, \mathbb{X}). \tag{0.2}$$

Цель работы состоит в определении такого набора весов объектов $\mathbb{W}$, при котором значение CCV было бы минимальным. Однако значение CCV зависит от выбора конкретного метода обучения из семейства M, поэтому будем минимизировать его верхнюю оценку. Нижняя оценка CCV позволит судить о наилучшем качестве, достижимом с помощью построенного алгоритма.

Расчет значений точной верхней и нижней оценки CCV в (0.2) является сложной вычислительной задачей. В следующем разделе доказываются верхние и нижние оценки CCV, которые в общем случае могут не являться достижимыми. Для этих оценок удается построить вычислительно эффективную процедуру их расчета, с полиномиальной сложностью вычисления.

2. ПОСТРОЕНИЕ ВЕРХНЕЙ И НИЖНЕЙ ОЦЕНКИ CCV

Для нахождения оценок CCV свяжем с каждым объектом генеральной выборки x_i два множества:

1. Множество безошибочных выборок $E^0(i)$:

$$E^0(i) = \{X : x_i \in \bar{X},\ \forall \mu \in \mathrm{M}\ \ I(\mu(X), x_i) = 0\} \subseteq [\mathbb{X}]^\ell.$$

2. Множество ошибочных выборок $E^1(i)$:

$$E^1(i) = \{X : x_i \in \bar{X},\ \forall \mu \in \mathrm{M}\ \ I(\mu(X), x_i) = 1\} \subseteq [\mathbb{X}]^\ell.$$

Теорема 1. Справедливы следующие верхняя и нижняя оценки CCV:

$$\frac{1}{k}\frac{1}{C_L^\ell}\sum_{i=0}^{L-1}|E^1(i)| \le Q_k(\mu^{opt},\mathbb{X}) \le Q_k(\mu,\mathbb{X}) \le Q_k(\mu^{pes},\mathbb{X}) \le 1-\frac{1}{k}\frac{1}{C_L^\ell}\sum_{i=0}^{L-1}|E^0(i)|,$$

где μ^{opt} и μ^{pes} обозначают методы оптимистичной и пессимистичной минимизации взвешенного эмпирического риска соответственно.

$\circ$ С каждым объектом генеральной выборки x_i свяжем множество нейтральных выборок $E^2(i)$:

$$E^2(i)=\{X : x_i\in\bar{X},\ \exists\mu_1\in\mathrm{M}\ \exists\mu_2\in\mathrm{M}\ \ I(\mu_1(X),x_i)=1\, I(\mu_2(X),x_i)=0\}\subseteq[\mathbb{X}]^\ell.$$

Множества $E^0(i)$, $E^1(i)$ и $E^2(i)$ не пересекаются по построению, а их объединение равно множеству всевозможных разбиений генеральной выборки, при которых объект i попадает в контрольную выборку. Количество таких разбиений равно C_{L-1}^ℓ. Таким образом, получаем следующее тождество:

$$|E^0(i)|+|E^1(i)|+|E^2(i)|=C_{L-1}^\ell. \tag{1.1}$$

Для методов оптимистичной и пессимистичной минимизации взвешенного эмпирического риска свяжем с каждым объектом аналогичные множества ошибочных выборок:

$$E^{pes}(i)=\{X : x_i\in\bar{X},\ I(\mu^{pes}(X),x_i)=1\}\subseteq[\mathbb{X}]^\ell;$$
$$E^{opt}(i)=\{X : x_i\in\bar{X},\ I(\mu^{opt}(X),x_i)=1\}\subseteq[\mathbb{X}]^\ell.$$

Из определений этих множеств вытекают включения $E^1(i)\subseteq E^{opt}(i),\ E^{pes}(i)\subseteq E^1(i)\bigcup E^2(i)$. Следовательно, справедливы неравенства:

$$|E^1(i)|\le|E^{opt}(i)|,\qquad |E^{pes}(i)|\le|E^1(i)|+|E^2(i)|. \tag{1.2}$$

Преобразуем теперь значение функционала CCV:

$$Q_k(\mu,\mathbb{X})=E_{\mathbb{X}}\nu(\mu(X),\bar{X})=E_{\mathbb{X}}\frac{1}{k}\sum_{x\in\bar{X}}I(\mu(X),x)=E_{\mathbb{X}}\frac{1}{k}\sum_{i=0}^{L-1}[x_i\in\bar{X}]I(\mu(X),x_i)=$$
$$=\frac{1}{k}\frac{1}{C_L^\ell}\sum_{i=0}^{L-1}\underbrace{\sum_{X\bigcup\bar{X}=\mathbb{X}}[x_i\in\bar{X}]I(\mu(X),x_i))}_{N_i}=\frac{1}{k}\frac{1}{C_L^\ell}\sum_{i=0}^{L-1}N_i. \tag{1.3}$$

Здесь N_i равно числу разбиений генеральной выборки $\mathbb{X}$ на обучающую выборку X и контрольную выборку $\bar{X}$, при которых объект i попадает в контрольную выборку и выбранный по обучающей выборке алгоритм с помощью метода μ на нем ошибается. Поскольку при вычислении N_i используется некий произвольный метод обучения $\mu\in\mathrm{M}$, то справедливы следующие неравенства:

$$|E^{opt}(i)|\le N_i\le|E^{pes}(i)|.$$

Воспользовавшись доказанным тождеством (1.1) и неравенством (1.2) получаем:

$$|E^1(i)|\le|E^{opt}(i)|\le N_i\le|E^{pes}(i)|\le|E^1(i)|+|E^2(i)|=C_{L-1}^\ell-|E^0(i)|.$$

Подставим эти оценки N_i в (1.3), получим

$$\frac{1}{k}\frac{1}{C_L^\ell}\sum_{i=0}^{L-1}|E^1(i)| \le \frac{1}{k}\frac{1}{C_L^\ell}\sum_{i=0}^{L-1}|E^{opt}(i)| \le \frac{1}{k}\frac{1}{C_L^\ell}\sum_{i=0}^{L-1}N_i \le \frac{1}{k}\frac{1}{C_L^\ell}\sum_{i=0}^{L-1}|E^{pes}(i)| \le \frac{1}{k}\frac{1}{C_L^\ell}\sum_{i=0}^{L-1}\left(C_{L-1}^\ell - |E^0(i)|\right).$$

Упростив верхнюю оценку:

$$\frac{1}{k}\frac{1}{C_L^\ell}\sum_{i=0}^{L-1}\left(C_{L-1}^\ell - |E^0(i)|\right) = \frac{L}{(L-\ell)}\frac{C_{L-1}^\ell}{C_L^\ell} - \frac{1}{k}\frac{1}{C_L^\ell}\sum_{i=0}^{L-1}|E^0(i)| = 1 - \frac{1}{k}\frac{1}{C_L^\ell}\sum_{i=0}^{L-1}|E^0(i)|,$$

и воспользовавшись тождеством

$$\frac{1}{k}\frac{1}{C_L^\ell}\sum_{i=0}^{L-1}|E^{opt}(i)| = \frac{1}{k}\frac{1}{C_L^\ell}\sum_{i=0}^{L-1}\sum_{X\cup\bar{X}=\mathbb{X}}[x_i \in \bar{X}]I(\mu^{opt}(X), x_i)) = Q_k(\mu^{opt}, \mathbb{X})$$

получаем утверждение теоремы. •

Теорема 1 формулирует важный принцип расчета значения CCV – от полного перебора по всем разбиениям на обучающую и контрольную выборки можно перейти к расчету вклада в значение оценки CCV каждого объекта, с помощью расчета мощности множества безошибочных и ошибочных выборок. Полученные оценки CCV могут не быть достижимыми, однако они всегда меньше 1. Возможность расчета как нижней, так и верхней оценки CCV позволяет судить о среднем значении CCV.

Рассмотрим задачу минимизации верхней оценки CCV для нахождения оптимального набора весов объектов. Для этого необходимо рассчитать мощность множества безошибочных выборок для каждого объекта генеральной выборки. В следующем разделе доказываются необходимые и достаточные условия, которыми должны удовлетворять обучающие выборки, чтобы являться безошибочными для конкретного объекта. На основе доказанных условий будет построена процедура, непосредственно рассчитывающая их мощность, с полиномиальной от количества объектов в генеральной выборке сложностью.

3. СОСТАВ МНОЖЕСТВА БЕЗОШИБОЧНЫХ ВЫБОРОК

Учтем специфику семейства монотонных пороговых классификаторов для определения состава множества безошибочных выборок. Так как в работе рассматриваются только задачи классификации на 2 класса из множества $\{-1,+1\}$, то условие отсутствия ошибок классификации на объекте x_i алгоритмом a_j эквивалентно условию $a_j(x_i)y_i > 0$. Для семейства монотонных пороговых классификаторов это условие можно представить в следующем виде:

$$a_j(x_i)y_i > 0 \Leftrightarrow \begin{cases} j > i,\ y_i < 0, \\ j \le i,\ y_i > 0. \end{cases}$$

Используя это выражение, перепишем определение множества безошибочных выборок:

$$E^0(i) = \left\{ X : x_i \in \bar{X},\ \forall \mu \in \mathrm{M}\ \ \mu(X) = a_j : \begin{Bmatrix} j > i,\ y_i < 0 \\ j \le i,\ y_i > 0 \end{Bmatrix} \right\}. \qquad (1.4)$$

Задачей этого раздела является нахождение необходимых и достаточных условий, которым должны удовлетворять безошибочные обучающие выборки X для выполнения определения (1.4).

Без ограничения общности будем считать, что $t \in \{0,1,\dots,\ell\}$ объектов обучающей выборки X с индексами $j_1, j_2, \dots j_t$ расположены правее объекта x_i и $\ell - t$ объектов с индексами $n_{\ell-t}, n_{\ell-t-1}, \dots n_1$ расположены левее объекта x_i (рис. 3).

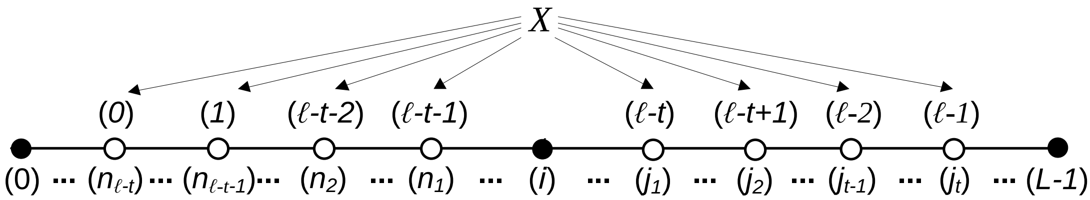

Рис. 3. Расположение объектов обучающей выборки X относительно объекта x_i. Под объектами подписаны их индексы в генеральной выборке. Над объектами – номера в обучающей выборке.

На основе введенных таким образом индексных множеств определим функции, которые будут часто использоваться в дальнейших расчетах:

$$f_+(X) = \min_{k=\ell-t+1,\dots,\ell} \left(\sum_{p=\ell-t}^{k-1} y^p w^p \right); \quad f_+^0(X) = \min_{k=\ell-t,\dots,\ell} \left(\sum_{p=\ell-t}^{k-1} y^p w^p \right) = \min(0, f_+(X));$$

$$f_-(X) = \max_{k=0,\dots,\ell-t-1} \left(\sum_{p=k}^{\ell-t-1} w^p y^p \right); \quad f_-^0(X) = \max_{k=0,\dots,\ell-t} \left(\sum_{p=k}^{\ell-t-1} w^p y^p \right) = \max(0, f_-(X)).$$

Используя введенные обозначения, докажем теорему, определяющую необходимые и достаточные условия, которым должна удовлетворять обучающая выборка для того, чтобы быть безошибочной для объекта x_i.

Теорема 2. Обучающая выборка X, в которой t объектов лежат правее объекта x_i, будет являться безошибочной для этого объекта, тогда и только тогда, когда одновременно выполняются следующие условия:

$$E^0(i) = \left\{ X : \begin{array}{l} x_i \in \bar{X} \\ t \in [\min(\ell, L-i-1), \max(\ell-i, 0)] \\ f_-(X) + f_+^0(X) > 0, \text{ если } y_i = +1 \\ f_-^0(X) + f_+(X) < 0, \text{ если } y_i = -1 \end{array} \right\}.$$

◦ Предположим, что $y_i = +1$. Согласно определению, обучающая выборка X будет безошибочной тогда и только тогда, когда выполнено условие $\forall \mu \in \mathrm{M} \colon \mu(X) = a_j : j \le i$ (см. рисунок 2). Возможны два случая. В первом случае для обучающей выборки X выполняется условие $\forall \mu \in \mathrm{M} \colon \mu(X) = a_j : n_1 < j \le i$. Если в этом случае взять метод обучения, выбирающий среди оптимальных алгоритмов алгоритм с самым правым положением порога, то получим условие $j > i$, что приводит к противоречию. Следовательно, возможен только второй случай, когда для обучающей выборки X должно выполняться условие $\forall \mu \in \mathrm{M} \colon \mu(X) = a_j : j \le n_1 < i$. Запишем формально это условие, воспользовавшись леммой 1 и используя номера объектов в обучающей выборке:

$$\arg\min_{k=0,\dots,\ell} \sum_{i=0}^{k-1} w^i y^i \le \ell - t - 1. \tag{1.5}$$

Это условие будет выполнено тогда и только тогда, когда минимум любой суммы в (1.5) при $k<\ell-t$ будет меньше минимума любой суммы при $k\ge\ell-t$. Получаем эквивалентное условие:

$$\min_{k=0,\dots,\ell-t-1}\left(\sum_{i=0}^{k-1}w^i y^i\right)<\min_{k=\ell-t,\dots,\ell}\left(\sum_{i=0}^{k-1}w^i y^i\right)=\sum_{i=0}^{\ell-t-1}w^i y^i+\min_{k=\ell-t,\dots,\ell}\left(\sum_{i=\ell-t}^{k-1}w^i y^i\right).$$

Перенося все члены в правую сторону неравенства и воспользовавшись тождеством $\forall a\in\mathbb{R},b\in\mathbb{R}:-\min(a,b)=\max(-a,-b)$, получаем:

$$\min_{k=\ell-t,\dots,\ell}\left(\sum_{i=\ell-t}^{k-1}w^i y^i\right)+\max_{k=0,\dots,\ell-t-1}\left(\sum_{i=0}^{\ell-t-1}w^i y^i-\sum_{i=0}^{k-1}w^i y^i\right)=\min_{k=\ell-t,\dots,\ell}\left(\sum_{i=\ell-t}^{k-1}w^i y^i\right)+\max_{k=0,\dots,\ell-t-1}\left(\sum_{i=k}^{\ell-t-1}w^i y^i\right)>0.$$

Во введенных обозначениях, это условие принимает вид $f_-(X)+f_+^0(X)>0$.

Проводя аналогичную цепочку доказательств, можно показать, что при $y_i=-1$ для обучающей выборки должно выполняться условие $f_-^0(X)+f_+(X)<0$.

Теперь учтем тот факт, что параметр t не может принимать произвольные значения. С одной стороны, значение параметра t должно быть меньше числа объектов, расположенных правее объекта i, то есть меньше $L-i-1$, а также меньше длины обучающей выборки ℓ. С другой стороны, значение параметра $\ell-t$ должно быть меньше числа объектов, расположенных левее объекта i, то есть меньше i. Из этих рассуждений получаем ограничения на параметр t. •

Теорема 2 позволяет эффективно проверять принадлежность любой обучающей выборки X к множеству безошибочных выборок для объекта x_i. Докажем аналогичную теорему для множества ошибочных выборок для объекта x_i, необходимую для для расчета нижней оценки CCV.

Теорема 3. Обучающая выборка X, в которой t объектов лежат правее объекта x_i, будет являться ошибочной для этого объекта, тогда и только тогда, когда одновременно выполняются следующие условия:

$$E^1(i)=\left\{X:\ \begin{array}{c} x_i\in\bar{X}\\ t\in[\min(\ell,L-i-1),\max(\ell-i,0)]\\ f_-^0(X)+f_+(X)<0,\text{ если } y_i=+1\\ f_-(X)+f_+^0(X)>0,\text{ если } y_i=-1\end{array}\right\}.$$

◦ Доказательство полностью аналогично доказательству теореме 2. •

Не смотря на то, что теоремы 2 и 3 позволяют проверять обучающие выборки на принадлежность к соответствующим множествам, вычисление мощности множеств ошибочных и безошибочных выборок по-прежнему остается вычислительно трудной задачей. Действительно, для решения этой задачи с помощью полного перебора пришлось бы применить условия теорем 2 и 3 последовательно ко всем обучающим выборкам длины ℓ, число которых равно C_L^ℓ. Однако, оказывается, существует рекурсивная процедура расчета мощностей множеств ошибочных и безошибочных выборок для объекта x_i, имеющая полиномиальную по количеству объектов генеральной выборки сложность. В следующем разделе доказывается алгоритм расчета оценок CCV, который является основной целью работы.

4. МОЩНОСТЬ МНОЖЕСТВА БЕЗОШИБОЧНЫХ ВЫБОРОК

Обозначим операцию суммирования по различным выборкам $X \subset [\mathbb{X}]^\ell$, у которых t объектов лежит правее объекта x_i, символом $\sum_{i,t}^{X}$ и воспользуемся теоремой 2 для расчета мощности множества безошибочных выборок:

$$|E^0(i)| = \sum_{t=\min(\ell,L-i-1)}^{\max(\ell-i,0)} \sum_{i,t}^{X} \Big([f_-(X) + f_+^0(X) > 0][y_i = +1] + [f_-^0(X) + f_+(X) < 0][y_i = -1] \Big). \quad (2.1)$$

Для упрощения этой формулы заметим, что объект $x_i \in \mathbb{X}$ разделяет любую обучающую выборку X на две выборки, одна из которых расположена левее объекта x_i, а другая правее:

$$X = X^- \cup X^+, \quad \text{где} \begin{cases} X^- = \{x \in X : x < x_i\}, \\ X^+ = \{x \in X : x > x_i\}. \end{cases}$$

Учитывая, что функции f_+, f_+^0, f_-, f_-^0, введенные в предыдущем разделе, определены на множествах X^- и X^+, справедливы тождества:

$$\begin{aligned} f_+(X) = f_+(X^+), \quad & f_+^0(X) = f_+^0(X^+), \\ f_-(X) = f_-(X^-), \quad & f_-^0(X) = f_-^0(X^+). \end{aligned} \quad (2.2)$$

Для упрощения формулы расчета мощности множества безошибочных выборок, введем дополнительные обозначения:

$$\mathrm{f}_+(i,t,s) = \sum_i^+ [f_+(X^+) = s], \quad \mathrm{f}_+^0(i,t,s) = \sum_i^+ [f_+^0(X^+) = s],$$
$$\mathrm{f}_-(i,\ell-t,s) = \sum_i^- [f_-(X^-) = s], \quad \mathrm{f}_-^0(i,\ell-t,s) = \sum_i^- [f_-^0(X^-) = s].$$

Здесь символом $\sum_i^+$ обозначена операция суммирования по всевозможным выборкам $X^+ \subset [\mathbb{X}]^t$, лежащим правее объекта x_i, а символом $\sum_i^-$ обозначена операция суммирования по всевозможным выборкам $X^- \subset [\mathbb{X}]^{\ell-t}$, лежащим левее объекта x_i. Из определений функций f_+, f_+^0, f_-, f_-^0 следуют ограничения на возможные значения s, которые могут принимать эти функции:

$$s \in S = \left\{ -\sum_{i=0}^{L-1} w_i [y_i = -1], \dots, \sum_{i=0}^{L-1} w_i [y_i = +1] \right\}.$$

Из тождеств (2.2) с использованием введенных обозначений следует цепочка равенств для одного слагаемого, входящего в (2.1):

$$\sum_{i,t}^{X} [f_-(X) + f_+^0(X) > 0] =$$
$$= [t=0] \sum_{i,t}^{X} [f_-(X) > 0] + [0 < t < \ell] \sum_{i,t}^{X} \sum_{s \in S} [f_+^0(X) = s][f_-(X) > -s] + [t = \ell] \sum_{i,t}^{X} [f_+^0(X) > 0].$$

Выражение при множителе $[t = \ell]$ равно 0, так как согласно определению $f_+^0(X) < 0 \ \forall X \in [\mathbb{X}]^\ell$. Упростим выражение при множителе $[0 < t < \ell]$:

$$\sum_{s \in S} \Big(\sum_{i,t}^{X} [f_+^0(X) = s][f_-(X) > -s] \Big) = \sum_{s \in S} \Big(\sum_{i,t}^{X} [f_+^0(X^+) = s][f_-(X^-) > -s] \Big) =$$
$$= \sum_{s \in S} \Big(\sum_i^+ [f_+^0(X^+) = s] \sum_i^- [f_-(X^-) > -s] \Big) = \sum_{s \in S} \left(\mathrm{f}_+^0(i,t,s) \sum_{s' > -s} \mathrm{f}_-(i,\ell-t,s) \right).$$

Используя доказанные тождества, получаем:

$$\sum\nolimits_{i,t}^{X}[f_-(X)+f_+^0(X)>0]=$$
$$=[t=0]\sum_{s>0}\mathrm{f}_-(i,\ell-t,s)+[0<t<\ell]\sum_{s\in S}\left(\mathrm{f}_+^0(i,t,s)\sum_{s'>-s}\mathrm{f}_-(i,\ell-t,s)\right). \quad (2.3)$$

Повторяя аналогичный вывод для другого слагаемого, входящего в (2.1), получаем тождество:

$$\sum\nolimits_{i,t}^{X}[f_-^0(X)+f_+(X)<0]=[0<t<\ell]\sum_{s\in S}\left(\mathrm{f}_-^0(i,\ell-t,s)\sum_{s'<-s}\mathrm{f}_+(i,t,s)\right)+[t=\ell]\sum_{s<0}\mathrm{f}_+(i,t,s). \quad (2.4)$$

На основе доказанных тождеств (2.3), (2.4) и формулы (2.1) рассчитывается мощность множества безошибочных выборок.

Аналогично, с использованием теоремы 3 и тождеств (2.2), может быть доказана формула вычисления мощности множества ошибочных выборок, необходимая для расчета нижней оценки CCV.

Таким образом, на основе значений функций $\mathrm{f}_+,\mathrm{f}_+^0,\mathrm{f}_-,\mathrm{f}_-^0$, мощность множеств ошибочных и безошибочных выборок может быть рассчитана по доказанным выше формулам. Поэтому основная задача состоит в расчете значений этих функций на сетке их аргументов. Функции f_+^0 и f_-^0 вычисляются на основе их определений по рассчитанным значениям функций f_+ и f_-:

$$\mathrm{f}_+^0(i,t,s)=\sum\nolimits_i^+[f_+^0(X^+)=s]=\sum\nolimits_i^+[\min(0,f_+(X^+))=s]=\begin{cases}\mathrm{f}_+(i,t,s), & s<0,\\ \sum_{s'\geq 0}\mathrm{f}_+(i,t,s'), & s=0,\\ 0, & s>0.\end{cases} \quad (2.5)$$

Расчет функции f_-^0 по известным значениям функции f_- аналогичен:

$$\mathrm{f}_-^0(i,t,s)=\sum\nolimits_i^-[f_-^0(X^-)=s]=\sum\nolimits_i^-[\max(0,f_-(X^-))=s]=\begin{cases}\mathrm{f}_-(i,t,s), & s>0,\\ \sum_{s'\leq 0}\mathrm{f}_-(i,t,s'), & s=0,.\\ 0, & s<0.\end{cases} \quad (2.6)$$

Теоремы 4 и 5, доказываемые ниже, определяют рекурсивный способ расчета функций f_+ и f_-.

Теорема 4. Функция $\mathrm{f}_+(i,t,s)$ вычисляется по рекурсивной формуле:

$$\mathrm{f}_+(i,t,s)=\mathrm{f}_+^0(i+1,t-1,s-w_{i+1}y_{i+1})+\mathrm{f}_+(i+1,t,s)$$

◦ Введем дополнительную функцию $\overline{f_+}(X^+)=\min\left(w_iy_i,w_iy_i+f_+(X^+)\right)$.

От функции $f_+(X^+)$ она отличается тем, что суммирование в ней начинается с объекта x_i. Аналогично определим функцию $\overline{\mathrm{f}_+}$:

$$\overline{\mathrm{f}_+}(i,t,s)=\sum\nolimits_i^+[\overline{f_+}(X^+)=s].$$

Используя определения функций f_+ и $\overline{f_+}$, перепишем определение функции $\overline{\mathrm{f}_+}$:

$$\overline{\mathrm{f}_+}(i,t,s) = \sum\nolimits_i^+ \left[\min\left(w_i y_i, w_i y_i + f_+(X^+) \right) = s \right].$$

Далее воспользуемся очевидным тождеством, которое справедливо для любых вещественных a и b: $\min(a, a+b) = a + \min(0,b)$:

$$\overline{\mathrm{f}_+}(i,t,s) = \sum\nolimits_i^+ \left[\min\left(0, f_+(X^+)\right) = s - w_i y_i \right] = \sum\nolimits_i^+ \left[f_+^0(X^+) = s - w_i y_i \right] = \mathrm{f}_+^0(i,t,s - w_i y_i). \quad (2.7)$$

Для вывода следующего тождества введем явно длину t обучающей выборки X^+, на которой она определена, в список аргументов функции f_+: $f_+(X^+) = f_+(X^+, t)$. Аналогично (2.7) запишем формально определение для функции $\mathrm{f}_+(i,t,s)$ и преобразуем его:

$$\begin{aligned} \mathrm{f}_+(i,t,s) &= \sum\nolimits_i^+ [\, f_+(X^+,t) = s] = \\ &= \sum\nolimits_{i+1}^+ \left[\min\left(w_{i+1} y_{i+1}, w_{i+1} y_{i+1} + f_+(X^+, t-1) \right) = s \right] + \sum\nolimits_{i+1}^+ [\, f_+(X^+,t) = s] = \\ &= \overline{\mathrm{f}_+}(i+1, t-1, s) + \mathrm{f}_+(i+1,t,s) \end{aligned}$$

Подставляя (2.7) в эту формулу, получаем утверждение теоремы. •

Совершенно аналогично доказывается теорема 5, описывающая вычислительно эффективный способ расчета значений функции f_-.

Теорема 5. Вектор $F_-(i,t)$ вычисляется по рекурсивной формуле:

$$\mathrm{f}_-(i,t,s) = \mathrm{f}_-^0(i-1, t-1, s - w_{i-1} y_{i-1}) + \mathrm{f}_-(i-1,t,s).$$

На основе Теорем 4 и 5, а также полученных формул расчета мощности множеств безошибочных выборок, выпишем в явном виде алгоритм расчета оценок CCV.

5. АЛГОРИТМ РАСЧЕТА ОЦЕНОК CCV

Для упрощения записи алгоритма расчета оценок CCV обозначим операции расчета значений функций f_+^0 и f_-^0 в (2.5) и (2.6) с помощью символов $[\square]^-$ и $[\square]^+$:

$$\mathrm{f}_-^0(i,t,s) = [\mathrm{f}_-(i,t,s)]^-;$$
$$\mathrm{f}_+^0(i,t,s) = [\mathrm{f}_+(i,t,s)]^+.$$

Ниже приведен алгоритм, вычисляющий верхнюю и нижнюю оценки CCV с полиномиальной по количеству объектов генеральной выборки сложностью.

Алгоритм 1.
Вход:

- Упорядоченная генеральная выборка объектов $\mathbb{X} = \{x_0, x_1, \ldots, x_{L-1}\},\ x \in \mathbb{R}$.
- Классы объектов $\mathbb{Y} = \{y_0, y_1, \ldots, y_{L-1}\},\ y \in \{-1,+1\}$.
- Веса объектов $\mathbb{W} = \{w_0, w_1, \ldots, w_{L-1}\},\ w \in \mathbb{Z}$.
- Длина обучающей выборки равна ℓ.

Шаг 1. Инициализация значений для длины обучающей выборки длиной $t = 1$:

$$s_{\max} = \sum_{i=0,\ldots,L-1} w_i [y_i = +1], \quad s_{\min} = -\sum_{i=0,\ldots,L-1} w_i [y_i = -1]$$

$\forall i = L-1,\dots,0$

$\forall s = s_{\min},\dots,s_{\max}$

$$\mathrm{f}_+(i,1,s) = \sum_{i<j}[w_j y_j = s]$$

$$\mathrm{f}_+^0(i,1,s) = [\mathrm{f}_+(i,1,s)]^+$$

$\forall i = 0,\dots,L-1$

$\forall s = s_{\min},\dots,s_{\max}$

$$\mathrm{f}_-(i,1,s) = \sum_{j<i}[w_j y_j = s]$$

$$\mathrm{f}_-^0(i,1,s) = [\mathrm{f}_-(i,1,s)]^-$$

Шаг 2. Расчет значений функций $\mathrm{f}_+, \mathrm{f}_+^0, \mathrm{f}_-, \mathrm{f}_-^0$ для длин обучающей выборки $t>1$:

$\forall t = 2,\dots,\ell$

$\forall i = L-t-1, L-t-2,\dots,1,0$

$\forall s = s_{\min},\dots,s_{\max}$

$$\mathrm{f}_+(i,t,s) = \mathrm{f}_+^0(i+1,t-1,s-w_{i+1}y_{i+1}) + \mathrm{f}_+(i+1,t,s)$$

$$\mathrm{f}_+^0(i,t,s) = [\mathrm{f}_+(i,t,s)]^+$$

$\forall i = t, t+1,\dots,L-1$

$$\mathrm{f}_-(i,t,s) = \mathrm{f}_-^0(i-1,t-1,s-w_{i-1}y_{i-1}) + \mathrm{f}_-(i-1,t,s)$$

$$\mathrm{f}_-^0(i,t,s) = [\mathrm{f}_-(i,t,s)]^-$$

Шаг 3. $\forall i = 0,\dots,L-1$:

$$P = \sum_{t=\min(\ell,L-i-1)}^{\max(\ell-i,0)} \left([t=0]\sum_{s>0}\mathrm{f}_-(i,\ell,s) + [0<t<\ell]\sum_{s\in S}\left(\mathrm{f}_+^0(i,t,s)\sum_{s'>-s}\mathrm{f}_-(i,\ell-t,s') \right) \right)$$

$$N = \sum_{t=\min(\ell,L-i-1)}^{\max(\ell-i,0)} \left([0<t<\ell]\sum_{s\in S}\left(\mathrm{f}_-^0(i,\ell-t,s)\sum_{s'<-s}\mathrm{f}_+(i,t,s') \right) + [t=\ell]\sum_{s<0}\mathrm{f}_+(i,\ell,s) \right)$$

$$|E^0(i)| = \begin{cases} P, & y_i = +1, \\ N, & y_i = -1, \end{cases}$$

$$|E^1(i)| = \begin{cases} N, & y_i = +1, \\ P, & y_i = -1. \end{cases}$$

Шаг 4. Расчет верхней и нижней оценок CCV:

$$\frac{1}{k}\frac{1}{C_L^\ell}\sum_{i=0}^{L-1}|E^1(i)| \le Q_k(\mu,\mathbb{X}) \le 1 - \frac{1}{k}\frac{1}{C_L^\ell}\sum_{i=0}^{L-1}|E^0(i)|$$

Рассчитаем сложность предложенного алгоритма.

Сложность шага 1 равна $O(L(s_{\max}-s_{\min}))$, так как необходимо в общем случае для L объектов рассчитать все $s_{\max}-s_{\min}$ значений компонент соответствующих векторов. Сложность шага 2 равна $O(L\ell(s_{\max}-s_{\min}))$, так как необходимо повторить процедуру описанную на шаге 1 ℓ раз. Сложность шага 3 равна $O(L\ell\max(s_{\min},s_{\max}))$, так как для каждого из L объектов необходимо рассчитать соответствующие мощности при всех значениях t, которых не более ℓ и каждый расчет имеет сложность $O(\max(s_{\min}, s_{\max}))$. Сложность шага 4 равна $O(L)$.

Поскольку все шаги выполняются последовательно, то сложность алгоритма равна максимальной сложности шага и равна:

$$O(L\ell(s_{\max} - s_{\min})) = O(L\ell(|w_0| + |w_1| + \ldots + |w_{L-1}|)).$$

Для вычисления мощностей безошибочных множеств для конкретного объекта x_i на шаге 3 требуется знание значений функций f_- и f_+ для всех значений $t = 1, \ldots, \ell$. Заметим, что рекурсивные процедуры расчета этих векторов, описанные в теоремах 1 и 2, начинают свой расчет с разных противоположных концов генеральной выборки. Это приводит к тому, что не удается организовать последовательного вычисления $|E^0(i)|$, не сохраняя все значения функций f_- и f_+ в памяти для всех значений $i = 0, \ldots, L-1$ и $t = 1, \ldots, \ell$. Это приводит к затратам памяти:

$$O(L\ell(s_{\max} - s_{\min})) = O(L\ell(|w_0| + |w_1| + \ldots + |w_{L-1}|)).$$

Если брать веса объектов из множества $w \in \{0,1\}$, то есть принимать решение об использовании того или иного объекта для обучения, то вычислительная сложность алгоритма 1 будет равна $O(\ell L^2)$, а затраты памяти $O(\ell L^2)$.

Таким образом, алгоритм 1 позволяет рассчитывать оценки CCV за приемлемое для небольших выборок (до 1000 объектов) время. Основываясь на этом расчете, построим алгоритм 2, реализующий метод жадного спуска для подбора оптимального набора весов объектов, минимизирующего значение верхней оценки CCV.

Алгоритм 2.

Дано:

- Упорядоченная генеральная выборка объектов $\mathbb{X} = \{x_0, x_1, \ldots, x_{L-1}\},\ x \in \mathbb{R}$.
- Классы объектов $\mathbb{Y} = \{y_0, y_1, \ldots, y_{L-1}\},\ y \in \{-1, +1\}$.
- Длина обучающей выборки равна ℓ.
- Максимальное значение $W_{\max}$, которое могут принимать веса объектов.
- Минимальное значение $W_{\min}$, которое могут принимать веса объектов.

Шаг 1. Инициализация весов:

$$\forall i = 0, \ldots, L-1$$
$$w_i = 1$$

Шаг 5. Повторять, пока $\Delta Q_c^{\ell k} > 0$, одна итерация обязательна.

$$\forall i = 0, \ldots, L-1$$

$$w_i^+ = \begin{cases} w_i + 1, & w_i + 1 \le W_{\max}, \\ 100, & w_i > W_{\max}, \end{cases}$$

$$w_i^- = \begin{cases} w_i - 1, & w_i - 1 \ge W_{\min}, \\ -100, & w_i < W_{\min}, \end{cases}$$

$$w_i' = \operatorname*{arg\,min}_{\hat{w}_i = \{w_i, w_i^+, w_i^-\}} \left(Q_c^{\ell k}(\mu(w_0, \ldots, \hat{w}_i, \ldots, w_{L-1}), \mathbb{X}) \right)$$

$$i^* = \arg\min_{i=0,\dots,L-1}\left(Q_c^{\ell k}(\mu(w_0,\dots,w_i',\dots,w_{L-1}),\mathbb{X})\right)$$

$$\Delta Q_c^{\ell k} = Q_c^{\ell k}(\mu(w_0,\dots,w_{i^*}',\dots,w_{L-1}),\mathbb{X}) - Q_c^{\ell k}(\mu(w_0,\dots,w_{i^*},\dots,w_{L-1}),\mathbb{X})$$

$$w_{i^*} = w_{i^*}'$$

Перейдем теперь к рассмотрению результатов численных экспериментов, целью которых является вычисление оптимального набора весов для модельной задачи, минимизирующего оценку CCV с помощью алгоритма 2.

РЕЗУЛЬТАТЫ ЧИСЛЕННЫХ ЭКСПЕРИМЕНТОВ

Рассмотрим модельную задачу, состоящую из 100 объектов класса −1 и 100 объектов класса +1, в которой признаки объектов созданы с помощью двух нормальных распределений:

$$x_i \in N(25,25), \quad y_i = -1;$$
$$x_i \in N(75,25), \quad y_i = +1.$$

На рисунке 4 показана сходимость верхней оценки CCV в зависимости от номера итерации по выбору весов объектов с помощью алгоритма 2. Длина обучающей выборки считалась равной 140. Из рисунка видно, что алгоритм сходится достаточно быстро, требуется не больше 30 итераций.

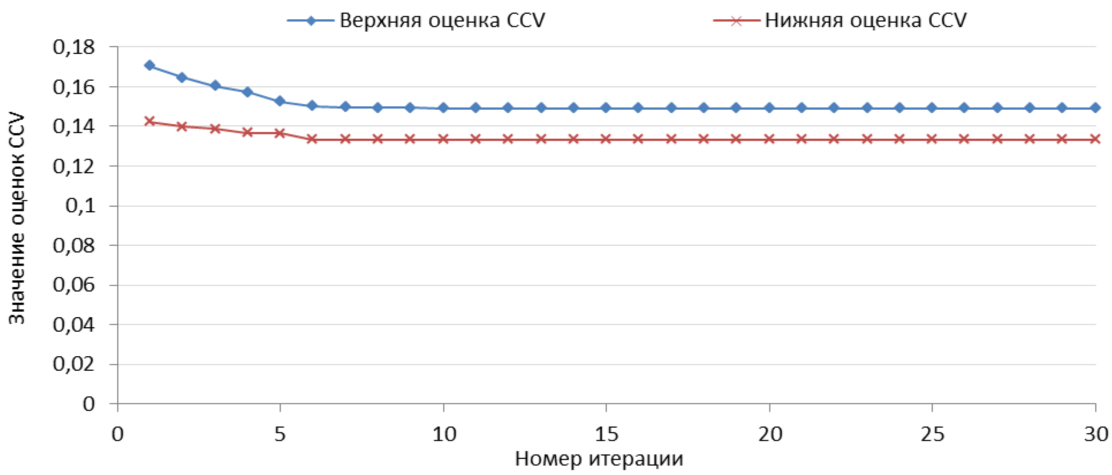

Рис. 4. Сходимость верхней оценки CCV при выборе весов объектов с помощью алгоритма 2.

Заметим, что при оптимальном значении весов (итерация 30), интервал возможных значений CCV становится достаточно узким. Относительное отклонение верхней оценки CCV от нижней составляет 12%. Таким образом, при оптимальном значении весов на данной модельной задаче предложенный метод позволяет рассчитывать CCV с максимальной относительной ошибкой равной 12%, что говорит о высокой точности полученных оценок.

Перейдем теперь к анализу оптимального значения весов объектов. На рисунке 5 для рассматриваемой задачи построены аналоги графиков из рисунка 1 для двух итераций – первой, когда веса всех объектов равны 1 и оптимальной, когда верхняя оценка CCV минимальна.

Из рисунка 5 и леммы 1 следует, что при оптимальном значении весов, порог оптимального алгоритма следует провести после объекта с индексом 100, то есть посередине между центрами двух нормальных распределений. Этот результат согласуется со здравым смыслом. При этом из рисунка 4 также видно, что веса всех объектов класса +1, находящихся левее порога оптимального алгоритма равны 0. Аналогично, веса всех объектов класса −1, лежащих правее порога оптимального алгоритма, равны 0. Таким образом, алгоритмы 1 и 2 позволяют не только оценивать CCV, но и выделять шумовые объекты, мешающие классификации, обнуляя их веса.

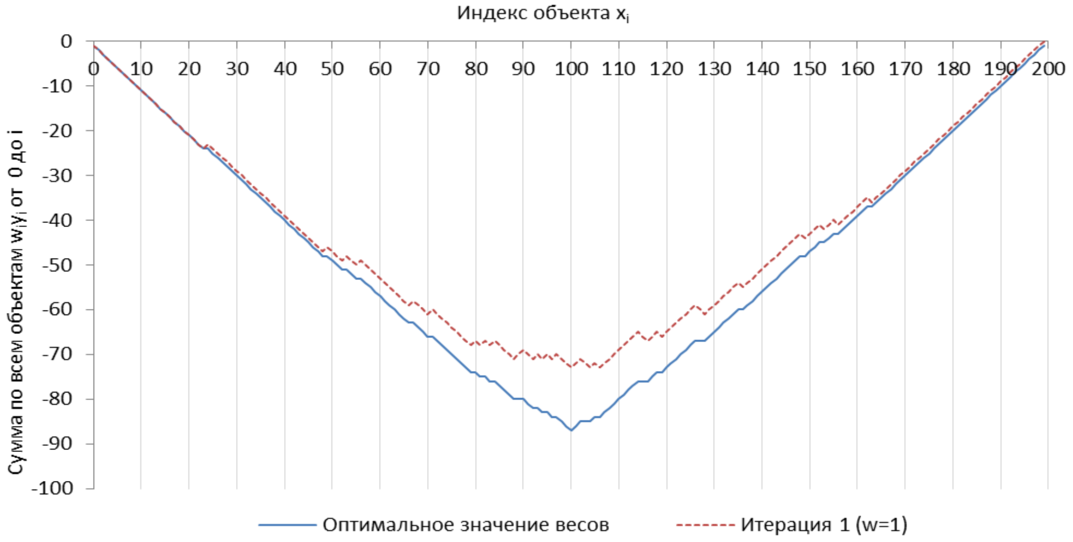

Рис. 5. Взвешенная сумма классов в зависимости от индекса объекта, до которого выполняется суммирование.

ЗАКЛЮЧЕНИЕ

В работе предложен комбинаторный метод оценки обобщающей способности алгоритмических семейств, основанный на расчете верхней и нижней оценок CCV, меньших 1. В основе метода лежит анализ состава и определение мощности множества ошибочных и безошибочных выборок. Множество ошибочных выборок для некоторого объекта генеральной выборки состоит из таких обучающих выборок, при которых данный объект попадает в контрольную выборку и синтезированный по обучающей выборке алгоритм всегда на нем ошибается. Аналогично определяется множество безошибочных выборок. Предложенный метод позволяет перейти от перебора по экспоненциальному числу слагаемых для расчета CCV к расчету мощности ошибочных и безошибочных выборок для каждого из L объектов генеральной выборки.

Для рассматриваемого в работе семейства монотонных пороговых классификаторов сложность предложенного метода расчета оценок CCV равна $O(L^3)$, что делает его пригодным для применения на выборках небольшой длины. Результаты численного эксперимента показали высокую точность полученных оценок CCV – относительное отклонение верхней оценки от нижней на рассматриваемой задаче составило не более 12%.

Предложенный метод расчета оценок CCV также позволяет учитывать веса объектов при синтезе алгоритмов по обучающей выборке. Минимизация верхней оценки CCV по множеству весов объектов позволила выделить и исключить шумовые объекты для рассмотренной в работе модельной задачи.

Открытым остается вопрос: возможно ли построить аналогичную процедуру расчета оценок CCV с полиномиальной по числу объектов сложностью для случая $x_i \in \mathbb{R}^n$ и семейства монотонных бинарных классификаторов?

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

СПИСОК ЛИТЕРАТУРЫ

1. Вапник В.Н., Червоненкис А.Я. *Теория распознавания образов*. М.: Наука, 1974.
2. Воронцов К.В. Оптимизационные методы линейной и монотонной коррекции в алгебраическом подходе к проблеме распознавания. *ЖВМ и МФ*. 2000. Т. 40. № 1. С. 166–176.
3. Гуз И.С. Нелинейные монотонные композиции классификаторов. В: *Математические методы распознавания образов-13*. М.: МАКС Пресс. 2007. С. 111–114.
4. Рудаков К.В., Воронцов К.В. О методах оптимизации и монотонной коррекции в алгебраическом подходе к проблеме распознавания. *Доклады РАН*. 1999. Т. 367. № 3. С. 314–317.
5. Воронцов К.В. Комбинаторный подход к оценке качества обучаемых алгоритмов. *Математические вопросы кибернетики*. М.: Физматлит, 2004. Т. 13. С. 5–36.
6. Воронцов К. В. Комбинаторные обоснования обучаемых алгоритмов. *ЖВМ и МФ*. 2004. Т. 44. № 11. С. 2099–2112.
7. Воронцов К.В. Комбинаторные оценки качества обучения по прецедентам. *Доклады РАН*. 2004. Т. 394. № 2. С. 175-178.
8. Vorontsov K.V. Combinatorial probability and the tightness of generalization bounds. *Patters Recognition and Image Analysis*. 2008. V. 18. № 2. P. 243-259.
9. Vorontsov K.V. Splitting and similarity phenomena in the sets of classifiers and their effect on the probability of overfitting. *Pattern Recognotion and Image Ananlysis*. 2009. V. 19. № 3. P. 412-420.
10. Воронцов К.В. Точные оценки вероятности переобучения. *Доклады РАН*. 2009. Т. 429. № 1. С. 15-18.
11. Вапник В.Н. *Восстановление зависимостей по эмпирическим данным*. М.: Наука, 1979.
12. Журавлёв Ю.И. Об алгебраическом подходе к решению задач распознавания или классификации. *Проблемы кибернетики*. 1978. Т. 33. С. 5–68.
13. Рудаков К.В. Алгебраическая теория универсальных и локальных ограничений для алгоритмов распознавания. *Диссертация на соискание учёной степени д.ф.-м.н.*, М.: ВЦ РАН, 1992.
14. Hastie T., Tibshirani R., Friedman J. *The elements of statistical learning*. Springer-Verlag, 2001.
15. Vapnik V. *The nature of statistical learning theory*. 2 edition. New York: Springer-Verlag, 2000.
16. Рудаков К.В. Монотонные и унимодальные корректирующие операции для алгоритмов распознавания. В: *Математические методы распознавания образов–VII*: Тез. докл. М., 1995.
17. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection.14th International Joint Conference on Artificial Intelligence, Palais de Congres Montreal, Quebec, Canada. 1995. P. 1137–1145. URL: http://citeseer.ist.psu.edu/kohavi95study.html.
18. Mullin M., Sukthankar R. Complete cross-validation for nearest neighbor classifiers. In: *Proceedings of International Conference on Machine Learning*. 2000. URL: http://citeseer.ist.psu.edu/309025.html

Материал поступил в редакцию 19.05.2011, опубликован 25.07.2011.