Московский государственный университет им. М. В. Ломоносова

Дипломная работа

на тему:

Рекуррентное вычисление вероятности переобучения некоторых алгоритмов

Студент:
Ишкина Шаура Хабировна,
512 группа

Научные руководители:
д.ф.-м.н. Верещагин Николай Константинович,
д.ф.-м.н. Воронцов Константин Вячеславович

17 мая 2013г.

Содержание

1 Введение

Теория статистического обучения (SLT) занимается проблемами восстановления зависимостей по эмпирическим данным. Основная задача SLT заключается в том, чтобы количественно оценить способность алгоритмов классификации и прогнозирования к обобщению эмпирических фактов. Данные оценки нужны для повышения качества алгоритмов.

Единственным из направлений SLT, дающим, хотя и для узкого класса семейств алгоритмов, точные оценки обобщающей способности, является комбинаторный подход [3, 4].

В данной работе рассматривалось конкретное семейство – прямая цепь классификаторов.

Практический интерес связан с тем, что такие семейства возникают при использовании пороговых решающих правил в алгоритмах классификации, в частности, в решающих деревьях, логических закономерностях [6], алгоритмах вычисления оценок[5].

Теоретический интерес связан с тем, что в рамках комбинаторного подхода до сих пор не удавалось получать точные оценки вероятности переобучения для цепей произвольного вида. Были известны только частные оценки для монотонных и унимодальных цепей [9].

1.1 Задача обучения по прецедентам

Пусть задано конечное множество объектов $\mathbb{X} = \{x_1, \dots, x_L\}$, называемое *генеральной выборкой* и множество $\mathbb{A} = \{a_1, \dots, a_D\}$, элементы которого называются *классификаторами.*

Предполагается, что существует функция $I : \mathbb{A} \times \mathbb{X} \to \{0, 1\}$ – индикатор 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 a *не ошибается* на данном элементе. Предполагается, что каждому классификатору $a \in \mathbb{A}$ *однозначно* соответствует его вектор ошибок $\vec{a} = (a(x_i))_{i=1}^{L}$, то есть два классификатора с одинаковыми векторами ошибок отождествляются. Далее под a для упрощения записи будет пониматься его вектор ошибок.

Постановка задачи обучения по прецедентам. Пусть генеральная выборка разбита на две подвыборки $\mathbb{X} = X \bigsqcup \bar{X}$. Выборка X длины l называется *наблюдаемой* или *обучающей*, для объектов $x \in X$ известны значения индикатора ошибки $I(a, x)$. Выборка $\bar{X}$ длины $k = L - l$ называется *скрытой* или *контрольной*, и на ней значения индикатора ошибки неизвестны. Задача состоит в том, чтобы найти классификатор $a \in \mathbb{A}$ с минимальным числом ошибок на генеральной выборке, пользуясь только информацией о наблюдаемой выборке. Данная задача в общем случае не имеет точного решения, поскольку классификатор a выбирается по неполной информации. Поэтому ставится задача поиска приближенного решения и оценивания его точности.

1.2 Основные определения

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

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

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

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

Переобученностью классификатора a на двух выборках X и $\bar{X} = \mathbb{X} \backslash X$ называется величина

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

Методом обучения называется отображение, которое подмножеству генеральной выборки ставит в соответствие классификатор из заданного семейства $\mu\colon 2^{\mathbb{X}} \to \mathbb{A}$.

Пусть $[X]^l$ – множество всех подвыборок $X \subset \mathbb{X}$ без возвращения длины l. Введем на $[X]^l$ равномерное распределение

$$P_l(X) = \frac{1}{C_L^l}.$$

Для фиксированного метода обучения μ, семейства классификаторов $\mathbb{A}$, генеральной выборки $\mathbb{X}$ и длины обучающей выборки l целью является получение верхних оценок *вероятности переобучения метода* μ

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

где через $[B]$ обозначен индикатор события B, то есть предикат

$$[B] = \begin{cases} 1, & \text{имеет место событие B}, \\ 0, & \text{иначе}. \end{cases}$$

Далее для краткости будем опускать параметры, от которых зависит вероятность переобучения, и записывать Q_ε.

Получение оценок вероятности переобучения проводится в два этапа.

На первом этапе предполагается, что известны значения функции ошибок каждого классификатора на всей генеральной выборке, то есть задана *матрица ошибок* размера $L \times D$, строки которой соотвествуют объектам генеральной выборки, а столбцами являются векторы ошибок классификаторов семейства A на этих объектах. В результате получаются комбинаторные оценки, зависящие от некоторых статистических характеристик этой матрицы.

Следующим этапом является оценка этих статистических характеристик, зависящих от генеральной выборки $\mathbb{X}$, по случайной наблюдаемой выборке X. Эмпирический способ решения задачи второго этапа предложен в [7].

В статистической теории Вапника–Червоненкиса [1] верхние оценки Q_ε зависят только от размерных характеристик задачи L и D, поэтому второй этап не нужен. Однако это упрощение приводит к завышенности оценок Q_ε на 6–10 порядков, согласно экспериментам на реальных данных [3].

Комбинаторные оценки учитывают внутреннюю структуру матрицы ошибок, вследствие чего оказываются завышенными лишь на 1–2 порядка [7] и могут быть использованы непосредственно для решения реальных задач классификации [8].

1.3 Семейства классификаторов

Будем рассматривать семейства классификаторов, которые задаются непосредственно своими матрицами ошибок. В данной работе основное внимание уделяется конкретному семейству, а именно, *прямой цепи* классификаторов.

Расстоянием между классификаторами называется расстояние Хемминга между их векторами ошибок:

$$\rho(a, a') = \sum_{i=1}^{L} [I(a, x_i) \neq I(a', x_i)], \quad \forall a, a' \in \mathbb{A}.$$

Конечная последовательность классификаторов $\mathbb{A} = \{a_0, a_1, \ldots, a_D\}$ называется *цепью*, если $\rho(a_d, a_{d+1}) = 1,\ d = 0, \ldots, D-1$.

Цепь $\mathbb{A} = \{a_0, \ldots, a_D\}$ называется *прямой*, если $\rho(a_0, a_D) = D$.

Пример Семейство классификаторов, задаваемое следующей матрицей ошибок, является прямой цепью:

	a_0	a_1	a_2	a_3	a_4
x_1	1	0	0	0	0
x_2	1	1	1	1	0
x_3	0	0	1	1	1
x_4	1	1	1	0	0

На рис.1 изображен график, где по горизонтали отложены классификаторы цепи из последнего примера, по вертикали – значения числа ошибок соответствующих классификаторов на генеральной выборке.

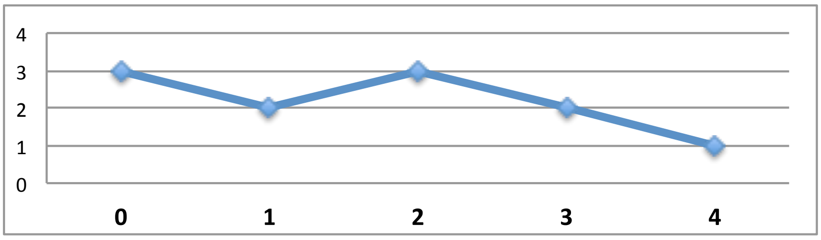

Рис. 1: Пример прямой цепи

Прямая цепь $\mathbb{A} = \{a_0, \ldots, a_D\}$ называется *возрастающей*, если каждый классификатор a_d допускает $m + d$ ошибок на множестве $\mathbb{X}$. Аналогично, прямая цепь $\mathbb{A} = \{a_0, \ldots, a_D\}$ называется *убывающей*, если каждый классификатор a_d допускает $m - d$ ошибок на генеральной выборке. Прямую цепь $\mathbb{A}$ будем называть *монотонной*, если она является убывающей или возрастающей.

Будем говорить, что прямая цепь $\mathbb{A}$ *составлена* из K монотонных, если ее можно представить в виде $\mathbb{A} = \bigcup\limits_{k=1}^{K} \mathbb{A}_k$, причем для каждого $k = 1, \ldots, K-1$ выполнены следующие требования:

1. A_k и A_{k+1} пересекаются ровно по одному классификатору
2. Среди A_k и A_{k+1} одна цепь является убывающей, другая – возрастающей

Пример Матрица ошибок

	a_0	a_1	a_2	a_3	a_4
x_1	1	0	0	0	0
x_2	1	1	0	0	0
x_3	0	0	0	1	1
x_4	0	0	0	0	1

задает прямую цепь, составленную из двух монотонных: $\{a_0, a_1, a_2\}$ – убывающей и $\{a_2, a_3, a_4\}$ – возрастающей. На рис.2 изображен график значений числа ошибок классификаторов на генеральной выборке.

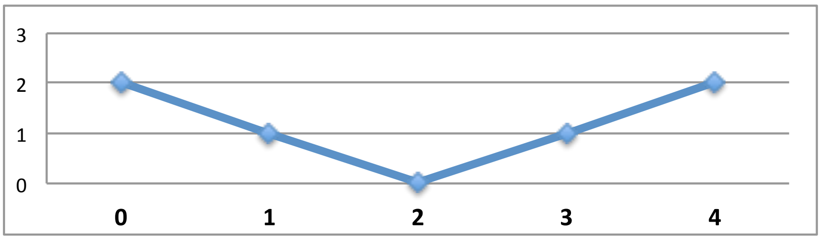

Рис. 2: Пример прямой цепи, составленной из двух монотонных

1.4 Постановка задачи

Рассмотрим метод обучения $\mu : [X]^l \to \mathbb{A}$, который выбирает классификатор, допускающий наименьшее число ошибок на обучающей выборке $X \in [X]^l$. Предположим, что в случае неоднозначности среди таких классификаторов μ выбирает классификатор с наибольшим числом ошибок на генеральной выборке. Данный метод назовем *пессимистичной минимизацией эмпирического риска (ПМЭР).*

Для генеральной выборки $\mathbb{X}$ длины L, длины обучающей выборки l, метода обучения μ – ПМЭР, произвольной прямой цепи $\mathbb{A}$ длины D ставится задача точного вычисления вероятности переобучения за полиномиальное по D, l и L время.

1.5 Основные результаты

Для двух частных случаев прямых цепей явно выписаны выражения для вероятности переобучения. Для произвольной прямой цепи $\mathbb{A}$ длины D предложен алгоритм вычисления вероятности переобучения с временем работы

$$O\left(Dl^4L\prod_{i=1}^{K}(1+h_i)\right),$$

где $h_1, \ldots, h_K$ – длины монотонных цепей, из которых составлена $\mathbb{A}$.

При условии

$$K = O(1)$$

данный алгоритм полиномиален по D.

2 Частные случаи прямых цепей

Будем обозначать через k $k = L - l$ – длину контрольной выборки, через $\lfloor x \rfloor$ обозначать целую часть x, то есть наибольшее целое число, не превосходящее x.

Гипергеометрической функцией распределения будем называть величину

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

В терминах множеств гипергеометрическая функция распределения $H_L^{l,m}(s)$ для данной генеральной выборки $\mathbb{X}$ длины L и подвыборки $X_0 \subset \mathbb{X}$ длины m равна доле подвыборок длины l, содержащих не более s элементов из X_0.

Для сокращения записей положим биномиальные коэффициенты C_n^k равными нулю при невыполнении условия $0 \leqslant k \leqslant n$. Гипергеометрическую функцию распределения $H_L^{l,m}(s)$ положим равной нулю при отрицательных s, а также при ее вычислении будем придерживаться описанных правил для биномиальных коэффициентов.

Семейству классификаторов $\mathbb{A}$ можно поставить в соответствие граф $G_{\mathbb{A}} = \langle V, E \rangle$, множество вершин V которого совпадает с $\mathbb{A}$, а множество ребер есть

$$E = \{e = (a, a') \mid \rho(a, a') = 1\}.$$

Заметим, что каждому ребру графа $G_{\mathbb{A}}$ можно поставить в соответствие объект генеральной выборки $\mathbb{X}$:

$$e = (a, a') \to x \in \mathbb{X}\colon I(a, x) \neq I(a', x).$$

В случае прямой цепи данное отображение инъективно, поэтому далее ребра в графе $G_{\mathbb{A}}$ будем отождествлять с соответствующими им элементами $\mathbb{X}$.

2.1 Симметричная цепь с максимумом

Определение *Симметричной цепью с максимумом* будем называть прямую цепь

$$\mathbb{A} = \{a_0, \ldots, a_{D-1}, a_D, a'_{D-1}, \ldots, a'_0\},$$

такую, что $\{a_0, \ldots, a_{D-1}, a_D\}$ – возрастающая цепь, а $\{a'_0, \ldots, a'_{D-1}, a'_D\}$, где $a'_D = a_D$, – убывающая.

Подцепь $\{a_0, \ldots, a_D\}$ будем называть *левой ветвью*, а $\{a'_0, \ldots, a'_D\}$ – *правой*.

Определим две выборки $Z = \{x_1, \ldots, x_D\} \subset \mathbb{X}$ и $Z' = \{x'_1, \ldots, x'_D\} \subset \mathbb{X}$, такие, что для каждого d a_{d-1} и a_d соединены ребром x_d в $G_{\mathbb{A}}$, a'_{d-1} и a'_d соединены ребром x'_d.

Обозначим через $Z_0 = Z \cup Z'$. Выборку $Z_1 = \mathbb{X} \backslash Z_0$ назовем *нейтральной*, на ней классификаторы цепи $\mathbb{A}$ неразличимы.

Обозначим через m число ошибок на нейтральной выборке. Таким образом, классификаторы a_d и a'_d допускают $m + d$ ошибок на $\mathbb{X}$.

Пример График числа ошибок на генеральной выборке симметричной цепи с максимумом при $D = 2$ и $m = 0$ изображен на рис.3.

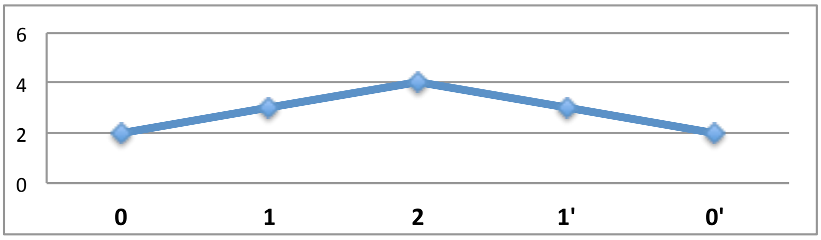

Рис. 3: Пример симметричной цепи с максимумом

Будем считать, что в случае неоднозначности, при которой a_d и a'_d одновременно могут быть выбраны методом обучения, выбирается классификатор из правой ветви, то есть a'_d.

Теорема 2.1 *Пусть* $\mathbb{A} = \{a_0, \dots, a_{D-1}, a_D, a'_{D-1}, \dots, a'_0\}$ *– симметричная цепь с максимумом,* $m = n(a_0, \mathbb{X}), D \geqslant 1, L \geqslant 2D + m$*. Тогда вероятность переобучения равна*

$$Q_\varepsilon = \sum_{z_0=1}^{\min\{l,2D\}} \sum_{d=0}^{D-1} \Bigl(2 \sum_{z=0}^{\hat{z}(z_0)} C_D^z C_{D-d-1}^{z_0-z-1} \widetilde{H}_{z_0}(s_{d,z}(\varepsilon)) + \\ + \delta(z_0) C_{D-d-1}^{z_0/2-1} (2C_D^{z_0/2} - C_{D-d}^{z_0/2} - C_{D-d-1}^{z_0/2}) \widetilde{H}_{z_0}(s_{d,z_0/2}(\varepsilon)) \Bigr) + \\ + \widetilde{H}_0(s_{D,0}(\varepsilon)), \qquad (1)$$

где

$$\delta(z_0) = [z_0 \equiv 0 \bmod 2],$$
$$\hat{z}(z_0) = \lfloor z_0/2 \rfloor - \delta(z_0),$$
$$\widetilde{H}_{z_0}(s) = \frac{C_{L-2D}^{l-z_0}}{C_L^l} H_{L-2D}^{l-z_0,m}(s),\ \forall s \leqslant m,$$
$$s_{d,z}(\varepsilon) = \frac{l}{L}(m + D + d - \varepsilon k) - z.$$

2.2 Несимметричная цепь с максимумом

Пусть дан неотрицательный параметр $0 \leqslant h \leqslant D$.

Определение *Несимметричной цепью с максимумом* с параметром h будем называть прямую цепь

$$\mathbb{A} = \{a_h, a_{h+1}, \dots, a_D, a'_{D-1}, \dots, a'_0\},$$

такую, что $\{a_h, a_{h+1}, \dots, a_D\}$ – возрастающая цепь, а $\{a'_0, \dots, a'_{D-1}, a'_D\}$, где $a'_D = a_D$, – убывающая.

Так же, как и в случае симметричной цепи с максимумом, обозначим через

$$\begin{aligned}
Z &= \{x_{h+1}, \ldots, x_D\},\\
Z' &= \{x'_0, \ldots, x'_D\},\\
Z_0 &= Z \cup Z',
\end{aligned}$$

где x_i – объект на котором a_{i-1} лучше, чем a_i, x'_i - объект на котором a'_{i-1} лучше, чем a'_i. Согласно определению прямой цепи, все элементы в Z_0 различны.

Через $Z_1 = \mathbb{X} \backslash Z_0$ обозначим нейтральную выборку, на которой классификаторы неразличимы.

Будем считать, что в случае неоднозначности, при которой a_d и a'_d одновременно могут быть выбраны методом обучения, выбирается классификатор из правой ветви, то есть a'_d.

Теорема 2.2 *Пусть* $\mathbb{A} = \{a_h, a_{h+1}, \ldots, a_D, a'_{D-1}, \ldots, a'_0\}$ *– несимметричная цепь с максимумом с параметром* h, $m = n(a_0, \mathbb{X}), D \geqslant h, L \geqslant 2D - h$. *Тогда вероятность переобучения равна*

$$\begin{aligned}
Q_\varepsilon = \sum_{z_0=1}^{2(D-h)} \sum_{d=h}^{D-1} \Bigg(\sum_{z=0}^{\hat{z}(z_0)} (C_{D-h}^{z} + C_{D}^{z}) C_{D-d-1}^{z_0-z-1} \widetilde{H}_{z_0}(s_{d,z}(\varepsilon)) + \\
+ \delta(z_0) C_{D-d-1}^{z_0/2-1} (C_{D}^{z_0/2} + C_{D-h}^{z_0/2} - C_{D-d}^{z_0/2} - C_{D-d-1}^{z_0/2}) \widetilde{H}_{z_0}(s_{d,z_0/2}(\varepsilon)) \Bigg) + \\
+ \sum_{d=0}^{h-1} \sum_{z=0}^{\hat{z}(z_0)} C_{D-h}^{z} C_{D-d-1}^{z_0-z-1} \widetilde{H}_{z_0}(s_{d,z}(\varepsilon)) + \\
+ \sum_{z_0=2(D-h)+1}^{\min\{2D-h, l\}} \sum_{d=0}^{D-1} \sum_{z=0}^{D-h} C_{D-h}^{z} C_{D-d-1}^{z_0-z-1} \widetilde{H}_{z_0}(s_{d,z}(\varepsilon)) + \widetilde{H}_0(s_{D,0}(\varepsilon)),
\end{aligned} \tag{2}$$

где

$$\begin{aligned}
\delta(z_0) &= [z_0 \equiv 0 \bmod 2],\\
\hat{z}(z_0) &= \lfloor z_0/2 \rfloor - \delta(z_0),\\
\widetilde{H}_{z_0}(s) &= \frac{C_{L-2D+h}^{l-z_0}}{C_L^l} H_{L-2D+h}^{l-z_0, m}(s),\ \forall s \leqslant m,\\
s_{d,z}(\varepsilon) &= \frac{l}{L}(m + D + d - h - \varepsilon k) - z.
\end{aligned}$$

Заметим, что оценка (1) является частным случаем оценки (2) при $h = 0$.

3 Произвольная прямая цепь

Пусть дана прямая цепь $\mathbb{A} = \{a_0, \ldots, a_D\}$. Через Z_0 обозначим множество объектов, по которым различаются классификаторы семейства, причем пронумеруем эти объекты так, чтобы x_i обозначал элемент, на котором различаются a_{i-1} и a_i. В силу того, что цепь прямая, в Z_0 ровно D элементов. Через Z_1 будем обозначать нейтральную выборку, то есть $\mathbb{X}\backslash Z_0$. Через m будем обозначать число ошибок классификаторов на Z_1 (классификаторы неразличимы на Z_1, поэтому данное число определено однозначно). Будем считать, что верно

$$L \geqslant m + D.$$

Через m_d будем обозначать число ошибок классификатора a_d на Z_0. Таким образом, классификатор a_d допускает на $\mathbb{X}$ $n(a_d, \mathbb{X}) = m + m_d$ ошибок.

По определению условной вероятности, для всех d верно равенство:

$$P\big[\mu X = a_d, \delta(\mu X, X) \geqslant \varepsilon\big] = P\big[\mu X = a_d\big] P\big[\delta(\mu X, X) \geqslant \varepsilon \big| \mu X = a_d\big].$$

Заметим, что в силу того, что классификаторы неразличимы на Z_1, от Z_1 зависит только событие $\big[\delta(\mu X, X) \geqslant \varepsilon \big| \mu X = a_d\big]$.

Обозначим через $M_d(z_0, z)$ число разбиений Z_1, таких, что выполнено $[\delta(\mu X, X) \geqslant \varepsilon \big| \mu X = a_d]$ при условии $|Z_0 \cap X| = z_0$, $n(a_d, Z_0 \cap X) = z$.

Обозначим через $N_d(z_0, z)$ число разбиений Z_0, таких, что выполнено $[\mu X = a_d]$ при условии $z_0 = |Z_0 \cap X|$ и $z = n(a_d, Z_0 \cap X)$.

Очевидно, что границы изменения параметров z_0 и z следующие:

$$z_0 \leqslant \min\{l, D\},\ z \leqslant \min\{l, m_d, z_0\}.$$

Обозначим через $\hat{z}_0 = min\{l, D\}$ и $\hat{z}(z_0, d) = min\{l, m_d, z_0\}$.

Легко проверить, что выражение вероятности переобучения имеет вид:

$$Q_\varepsilon = \sum_{d=0}^{D} \sum_{z_0=0}^{\hat{z}_0} \sum_{z=0}^{\hat{z}(z_0,d)} \frac{1}{C_L^l} N_d(z_0, z) M_d(z_0, z).$$

Лемма 3.1 *Пусть* $s_{d,z}(\varepsilon)) = \frac{l}{L}(m + m_d - \varepsilon k) - z$. *Тогда*

$$M_d(z_0, z) = C_{L-D}^{l-z_0} H_{L-D}^{l-z_0, m}(s_{d,z}(\varepsilon)),$$

Обозначим через

$$\widetilde{H}_{z_0}(s) = \frac{1}{C_L^l} M_d(z_0, z) = \frac{C_{L-D}^{l-z_0}}{C_L^l} H_{L-D}^{l-z_0, m}(s).$$

Таким образом, мы доказали следующую теорему:

Теорема 3.2 *Формула для вычисления вероятности переобучения принимает вид*

$$Q_\varepsilon = \sum_{d=0}^{D} \sum_{z_0=0}^{\hat{z}_0} \sum_{z=0}^{\hat{z}(z_0,d)} N_d(z_0, z) \widetilde{H}_{z_0}(s_{d,z}(\varepsilon)). \tag{3}$$

Задача свелась к нахождению $N_d(z_0, z)$ для всех d, z_0, z.

4 Алгоритм вычисления вероятности переобучения произвольной прямой цепи

Опишем, как в общем случае найти величины $N_d(z_0, z)$ для каждой тройки d, z_0, z.

Будем считать, что из классификаторов, минимизирующих число ошибок на обучающей выборке и допускающих равное число ошибок на генеральной выборке, выбирается классификатор с наибольшим номером.

Зафиксируем d, z_0, z. Рассмотрим классификатор a_d. Относительно него цепь разбивается на две: правую и левую. В правой подцепи лежат классификаторы с номерами $d, d+1, \dots, D$, в левой – $0, 1, \dots, d$.

Будет называть классификатор называется *лучшим*, если он выбирается методом обучения. Классификатор a_d – лучший в $\mathbb{A}$, если он является лучшим в левой и правой относительно него цепях.

Обозначим через Z_{right} объекты $x_{d+1}, \dots, x_D$, $Z_{left} = Z_0 \backslash Z_{right} = x_1, \dots, x_d$. В силу того, что цепь прямая, множества Z_{left} и Z_{right} не пересекаются, классификаторы левой цепи неразличимы на Z_{right}, классификаторы правой цепи неразличимы на Z_{left}. Из этого следует, что общее число разбиений Z_0, в которых метод обучения выбирает a_d, является произведением числа разбиений Z_{left} и Z_{right}, при которых классификатор a_d является лучшим в соответствующих цепях.

Так мы разбили задачу на две подзадачи:

1. найти число разбиений Z_{left}, таких, что выполнено $[\mu X = a_d]$ в левой цепи
2. найти число разбиений Z_{right}, таких, что выполнено $[\mu X = a_d]$ в правой цепи

Напомним, что у нас имеется дополнительное ограничение на число элементов из X в Z_0 ($|Z_0 \cap X| = z_0$), а также на число ошибок $n(a_d, Z_0 \cap X) = z$.

Для решения подзадач добавим требования

$$|Z_{left} \cap X| = z_{0,left},$$
$$n(a_d, Z_{left} \cap X) = z_{left}.$$

Обозначим через $z_{0,right} = z_0 - z_{0,left}$, $z_{right} = z - z_{left}$. Сразу определим границы изменения параметров:

$$z_{0,left} \leqslant \hat{z}_0 = \min(z_0, d),$$
$$z_{left} \leqslant \hat{z} = \min(z_0, n(a_d, Z_{left})).$$

Пусть $L(z_{0,left}, z_{left})$ – решение подзадачи на левой цепи с параметрами $z_{0,left}, z_{left}$, $R(z_{0,right}, z_{right})$ – решение подзадачи на правой подцепи с параметрами $z_{0,right}, z_{right}$.

Тогда искомое решение задачи поиска числа разбиений с параметрами z_0, z находится по формуле

$$N_d(z_0, z) = \sum_{z_{0,left}=0}^{\hat{z}_0} \sum_{z_{left}=0}^{\hat{z}} L(z_{0,left}, z_{left}) R(z_0 - z_{0,left}, z - z_{left}). \qquad (4)$$

Есть два крайних случая: $d = 0$ и $d = D$. При $d = 0$ решение подзадачи на левой цепи тривиально: параметры $z_{0,left} = z = 0$ и $L(0, 0) = 1$. Аналогично, при $d = D$ $R(0, 0) = 1$.

4.1 Решение подзадач

Будем называть классификатор $a_j, 1 \leqslant j \leqslant D-1$ *локальным максимумом*, если он является локальным максимумом функции $n(a_i, \mathbb{X})$, то есть имеет место неравенство

$$n(a_{j-1}, \mathbb{X}) = n(a_{j+1}, \mathbb{X}) < n(a_j, \mathbb{X}).$$

a_0 – локальный максимум, если $n(a_1, \mathbb{X}) < n(a_0, \mathbb{X})$, a_D – если $n(a_{D-1}, \mathbb{X}) < n(a_{\hat{D}}, \mathbb{X})$.

Аналогично, $a_j, 1 \leqslant j \leqslant D-1$ – *локальный минимум*, если точка $(j, n(a_j, \mathbb{X}))$ является локальным минимумом функции $n(a_i, \mathbb{X})$. a_0 – локальный минимум, если $n(a_0, \mathbb{X}) < n(a_1, \mathbb{X})$, a_D – если $n(a_D, \mathbb{X}) < n(a_{D-1}, \mathbb{X})$.

Выпишем в список A_{MaxMin} порядковые номера всех локальных максимумов и минимумов по возрастанию, начиная с 0, заканчивая D (a_0 и a_D всегда являются либо локальными максимумами, либо минимумами). Заметим, что в данном списке максимумы с минимумами чередуются.

В решении подзадач мы будем использовать понятие *запаса ошибок* данного классификатора a:

$$\Delta(a) = n(a, X) - n(a_d, X).$$

Для классификаторов левой цепи, в силу того, что они неразличимы на $\mathbb{X} \backslash Z_{left}$, данную величину можно вычислить следующим образом:

$$\Delta(a) = n(a, Z_{left} \cap X) - n(a_d, Z_{left} \cap X),$$

для правой:

$$\Delta(a) = n(a, Z_{right} \cap X) - n(a_d, Z_{right} \cap X).$$

Подчеркнем, в чем отличие подзадач для левой и правой цепей. Мы договорились считать, что из классификаторов, допускающих наименьшее число ошибок на обучающей выборке и равное число ошибок на генеральной выборке, выбирается классификатор с наибольшим номером. Значит, в левой цепи разрешено, чтобы классификатор, допускающий столько же ошибок, что и a_d, на обучающей выборке, допускал столько же ошибок и на генеральной выборке (в этом случае будет выбран a_d), тогда как в правой цепи это запрещено.

Таким образом, получаем

Теорема 4.1 *Необходимыми и достаточными условиями выбора классификатора a_d методом обучения, то есть события $[\mu X = a_d]$, являются:*

1. *Либо* $\Delta(a) > 0$;
2. *Либо* $\Delta(a) = 0$, *а из левой цепи и* $n(a, Z_0) \leqslant n(a_d, Z_0)$;
3. *Либо* $\Delta(a) = 0$, *а из правой цепи и* $n(a, Z_0) < n(a_d, Z_0)$.

4.2 Постановка и решение подзадачи для левой цепи

Обозначим через $\mathbb{B}$ список классификаторов из A_{MaxMin}, которые находятся в левой цепи, кроме a_d, причем записанных в порядке убывания номеров. Если в $\mathbb{B}$ нет классификатора a_{d-1}, добавим его *в начало.*

Введем обозначения: $\mathbb{B}$ состоит из $\{b_0, b_1, \ldots, b_T\}$, где $b_0 = a_{d-1}$, $b_T = a_0$; $a = a_d$, $\mathbb{Y} = \{x_{d-1}, \ldots, x_1\}$, $Y_0 = \{x_{d-2}, \ldots, x_1\}$, $y_0 = |Y_0 \cap X|$, $y = n(a_d, Y_0 \cap X)$.

Для каждого i пара (b_i, b_{i+1}) задает монотонную цепь. Поэтому $\mathbb{B}$ – способ записи левой цепи в виде последовательности монотонных. Далее, когда будем говорить *цепь* $\mathbb{B}$, мы будем иметь в виду именно некоторую подцепь левой цепи.

Заметим, что однозначно определяются параметры $\Delta(b_0)$, y_0, y:

1. Если $n(b_0, \mathbb{Y}) < n(a, \mathbb{Y})$, то $x_d \in \bar{X}$. В противном случае классификатор b_0 допускает на X меньше ошибок, чем a_d. Тогда $\Delta(b_0) = 0$, $y_0 = z_{0,left}$, $y = z_{left}$.

2. Если $n(b_0, \mathbb{Y}) > n(a_d, \mathbb{Y})$, то $x_d \in X$. Иначе классификатор b_0 допускает столько же ошибок на обучении, что a_d, но на больше контроле. Тогда $\Delta(b_0) = 1$, $y_0 = z_{0,left} - 1$, $y = z_{left}$.

Заметим, что вектора ошибок классификаторов $a = a_d$ и $b_0 = a_{d-1}$, ограниченные на Y_0, совпадают. Действительно, по построению a_d и a_{d-1} различаются только на объекте x_d. Будем поддерживать это свойство в качестве инварианта. Значит, условие $y = n(a, Y_0 \cap X)$ равносильно $y = n(b_0, Y_0 \cap X)$.

Будем называть в рамках решения подзадачи классификатор c *лучшим*, чем c', если c либо допускает меньшее число ошибок на обучающей выборке, либо столько же ошибок на обучении, но больше на контроле.

Постановка задачи: для данных

$$\begin{aligned}
&Y_0 = \{x_{d-1}, \ldots, x_1\} \subset \mathbb{Y},\\
&y_0 = |Y_0 \cap X|,\\
&y = n(b_0, Y_0 \cap X),\\
&\Delta_0 = \Delta(b_0),
\end{aligned}$$

найти число разбиений ребер цепи $\mathbb{B} = \{b_0, b_1, \ldots, b_T\}$, таких что классификатор a лучше любого классификатора из $\mathbb{B}$. Обозначим число искомых разбиений через $M(\mathbb{B}, \Delta_0, Y_0, y_0, y)$.

4.2.1 Описание алгоритма решения задачи для левой цепи

Даны $\mathbb{B}, Y_0, y_0 = |Y_0 \cap X|$, $y = n(b_0, Y_0 \cap X)$, $\Delta_0 = \Delta(b_0) \geqslant 0$.

Рассмотрим крайние случаи.

Лемма 4.2 *При данных $\Delta(b_0), y_0, y$ запас ошибок классификатора b_T равен:*

$$\Delta(b_T) = \Delta(b_0) + y_0 - 2y.$$

Из леммы 4.2 следует:

1. Если $\Delta_0 < 0$ или $\Delta(b_T) = \Delta_0 + y_0 - 2y < 0$, то $M(\mathbb{B}, \Delta_0, y_0, y) = 0$.

2. Если $\mathbb{B}$ состоит из одного классификатора, то есть число ребер равно нулю, то

$$M(\mathbb{B}, \Delta_0, Y_0, y_0, y) = \begin{cases} 0, & y_0 \neq 0, \\ 1, & y_0 = y = 0. \end{cases}$$

Будем говорить, что параметры $\Delta(b_0), y_0, y$ *согласованы*, если $\Delta(b_0) \geqslant 0$, $\Delta(b_T) = \Delta(b_0) + y_0 - 2y \geqslant 0$.

Пусть $\mathbb{B}$ состоит хотя бы из двух классификаторов и параметры задачи согласованы.

Рассмотрим пару b_0, b_1. Два классификатора задают монотонную цепь. Будем записывать ее как (b_0, b_1). Обозначим $m_0 = n(b_0, \mathbb{Y})$, $m_1 = n(b_1, \mathbb{Y})$, $m_a = n(a, \mathbb{Y})$.

Обозначим через $h = |m_0 - m_1|$ длину цепи (b_0, b_1). Обозначим классификаторы цепи через $c_0, \dots, c_h$, ребра через $z_1, \dots, z_h$, нумерация в порядке от b_0 к b_1, то есть $c_0 = b_0, c_h = b_1$.

Введем параметр s – число ребер цепи (b_0, b_1), лежащих в X. Обозначим через $\check{s}$ нижнюю границу значений параметра, через $\hat{s}$ – верхнюю.

При данных m_0, m_1, y_0, y, $\Delta(b_0)$ и s хотим найти число разбиений, таких что никакой c_i не лучше, чем a. В силу того, что исходная цепь прямая, это условие зависит только от способа разбиения $z_1, \dots, z_h$.

Лемма 4.3 *В зависимости от вида цепи (b_0, b_1), m_0, m_1, m_a и $\Delta(b_0)$ однозначно определяются $\Delta(b_1)$ и границы значений $\check{s}$, $\hat{s}$:*

1. *Если цепь возрастающая, то $\Delta(b_1) = \Delta(b_0) + s, \hat{s} = \min\{m_1 - m_0, y_0\}$.*
   (a) *Если $m_0 \leqslant m_a < m_1$:*
      i. *Если $\Delta(b_0) > 0$, то $\check{s} = 0$.*
      ii. *Если $\Delta(b_0) = 0$, то $\check{s} = 1$.*
   (b) *Иначе нижняя граница $\check{s} = 0$.*
2. *Если цепь убывающая, то $\Delta(b_1) = \Delta(b_0) - s, \check{s} = 0$.*
   (a) *Если $m_a < m_1$, то $\hat{s} = \min\{\Delta(b_0) - 1, m_0 - m_1, y\}$.*
   (b) *Иначе $\hat{s} = \min\{\Delta(b_0), m_0 - m_1, y\}$.*

Обозначим через $Y_0' = Y_0 \backslash \{z_1, \dots, z_h\}$, $y_0' = |Y_0' \cap X|$, $y' = n(a, Y_0' \cap X)$.

Лемма 4.4 *В зависимости от вида цепи (b_0, b_1) и значений параметров s, y_0 и y однозначно определяются y_0' и y':*

1. *Если цепь возрастающая, то $y_0' = y_0 - s$, $y' = y$.*
2. *Если цепь убывающая, то $y_0' = y_0 - s$, $y' = y - s$.*

Итак, по данному параметру $s : \check{s} \leqslant s \leqslant \hat{s}$ и $\Delta(b_0)$ однозначно вычисляется $\Delta(b_1)$. Определим

$$\Delta = \begin{cases} \Delta(b_0), & m_0 < m_1, \\ \Delta(b_1), & m_1 < m_0. \end{cases}$$

Δ зависит от $m_0, m_1, s, \Delta(b_0)$.

Лемма 4.5 *Рассмотрим цепь (b_0, b_1). Пусть s – число ребер, попавших в обучающую выборку, $\check{s} \leqslant s \leqslant \hat{s}$. Тогда число разбиений ребер цепи (b_0, b_1), в которых никакой классификатор цепи не лучше, чем a, зависит только от s, m_a, Δ, $M = \max\{m_0, m_1\}, m = \min\{m_0, m_1\}$ (обозначим через $N(M, m, \Delta, m_a, s)$) и вычисляется по следующему правилу:*

1. *Если $m_a < m$ или $M \leqslant m_a$, то $N(M, m, \Delta, m_a, s) = C^s_{M-m}$.*
2. *Если $m \leqslant m_a < M$, то*
   (a) *Если $\Delta > 0$, то $N(M, m, \Delta, m_a, s) = C^s_{M-m}$.*
   (b) *Если $\Delta = 0$, то $N(M, m, \Delta, m_a, s) = C^s_{M-m} - C^s_{M-m_a-1}$.*

Обозначим через $\mathbb{B}' = \mathbb{B}\backslash\{b_0\} = \{b_1, \ldots, b_T\}$, через $\Delta'_0 = \Delta(b_1)$.

Мы научились искать число способов распределить ребра цепи (b_0, b_1) так, что никакой классификатор этой цепи не был лучше, чем a. Если теперь мы решим задачу на $\mathbb{B}'$ с параметрами y'_0, y', зависящими от s, то, в силу того, что $\mathbb{A}$ – прямая цепь, то искомое число разбиений при данном s является произведением ответов. Таким образом, мы доказали теорему:

Теорема 4.6 *Решение $M(\mathbb{B}, \Delta_0, Y_0, y_0, y)$ исходной задачи следующее:*

1. *Если $\Delta_0 < 0$ или $\Delta_0 + y_0 - 2y < 0$, то $M(\mathbb{B}, \Delta_0, Y_0, y_0, y) = 0$.*
2. *Если $\mathbb{B}$ состоит из одного классификатора, то*
$$M(\mathbb{B}, \Delta_0, Y_0, y_0, y) = \begin{cases} 0, & y_0 \neq 0, \\ 1, & y_0 = y = 0. \end{cases}$$
3. *Иначе*
$$M(\mathbb{B}, \Delta_0, Y_0, y_0, y) = \sum_{s=\check{s}}^{\hat{s}} N(M, m, \Delta, m_a, s) M(\mathbb{B}', \Delta'_0, Y'_0, y'_0, y').$$

Перейдем к доказательству корректности.

4.2.2 Доказательство корректности

Напомним, что в качестве инварианта мы выбрали такое свойство: вектор ошибок b_0 и $a = a_d$, ограниченные на Y_0, совпадают.

Лемма 4.7 *Инвариант алгоритма сохраняется при переходе от $\mathbb{B}$ к $\mathbb{B}'$.*

Доказательство Докажем по индукции по порядковому номеру классификатора b_0 в исходной цепи $\mathbb{A}$:

1. Пусть $b_0 = a_{d-1}$. Очевидно, база индукции выполнена, поскольку a_d и a_{d-1} различаются только на x_d, который не лежит в Y_0.

2. Пусть для текущего b_0 это верно. Докажем, что если в $\mathbb{B}$ хотя бы два классификатора, то при переходе к $\mathbb{B}'$ и Y_0' первый классификатор цепи $\mathbb{B}'$ также обладает данным свойством.

   Первым классификатором $\mathbb{B}'$ является b_1. b_0 и b_1 различаются только на ребрах цепи (b_0, b_1). Поэтому, если эти ребра удалить из Y_0, что и проходит при переходе к Y_0', то b_0 и b_1 будут совпадать на Y_0', а значит, и a с b_1 тоже. Переход и лемма доказаны. ■

Переход от $\mathbb{B}$ к $\mathbb{B}'$ сохраняет инвариант, на основе которого строятся рассуждения, приводящие к решению задачи для цепи (b_0, b_1). Алгоритм останавливается, поскольку длина списка $\mathbb{B}$ каждый раз уменьшается на 1. Следовательно, при условии истинности утверждений лемм, построенный алгоритм действительно решает поставленную задачу для левой цепи.

4.2.3 Сложность алгоритма

Будем считать, что арифметические операции и сравнение чисел осуществляются за $O(1)$.

Имеем $\mathbb{B} = \{b_0, \dots, b_T\}$. Обозначим время работы процедуры $M(\mathbb{B}, \Delta_0, Y_0, y_0, y)$ через $Time(\mathbb{B}, \Delta_0, Y_0, y_0, y)$.

Если $\mathbb{B}$ состоит из одного классификатора или параметры несогласованы, то

$$Time(\mathbb{B}, \Delta_0, Y_0, y_0, y) = O(1).$$

Если $\mathbb{B}$ состоит не меньше чем из двух классификаторов, то для каждого s : $\check{s} \leqslant s \leqslant \hat{s}$ мы запускаем процедуру $M(\mathbb{B}', \Delta_0'(s), Y_0', y_0'(s), y'(s))$. Можем считать, что $N(M, m, \Delta, m_a, s)$ работает за $O(1)$ – биномиальные коэффициенты посчитаем перед началом работы алгоритма.

Тогда

$$\begin{aligned} Time(\mathbb{B}, \Delta_0, Y_0, y_0, y) &= \sum_{s=\check{s}}^{\hat{s}} Time(\mathbb{B}', \Delta_0'(s), Y_0', y_0'(s), y'(s)) \leqslant \\ &\leqslant (|m_0 - m_1| + 1) \max_s Time(\mathbb{B}', \Delta_0'(s), Y_0', y_0'(s), y'(s)). \end{aligned}$$

Перейдя по индукции к оценке $Time(\mathbb{B}', \Delta_0'(s), Y_0', y_0'(s), y'(s))$ получаем оценку

$$Time(\mathbb{B}, \Delta_0, Y_0, y_0, y) \leqslant \prod_{i=0}^{i=T-1} (1 + h_i),$$

где h_i - длина цепи (b_i, b_{i+1}).

4.3 Постановка и решение подзадачи для правой цепи

Записываем в $\mathbb{B}$ все классификаторы из $\mathbb{A}_{MaxMin}$ с номерами, большими d, и a_{d+1}, в порядке возрастания номеров.

Вводим обозначения $\mathbb{B} = \{b_0, \dots, b_T\}$, где $b_0 = a_{d+1}$, $b_T = a_D$; $a = a_d$, $\mathbb{Y} = \{x_{d+1}, \dots, x_D\}$, $Y_0 = \{x_{d+2}, \dots, x_D\}$.

Однозначно определяются $\Delta_0 = \Delta(b_0)$, $y_0 = |Y_0 \ \cap \ X|$ и $y = n(b_0, Y_0 \cap X)$. Если $n(b_0, \mathbb{Y}) < n(a, \mathbb{Y})$, тогда $x_{d+1} \in \bar{X}$, $\Delta_0 = 0$, $y_0 = z_{0,right}$, $y = z_{right}$. Если $n(b_0, \mathbb{Y}) > n(a, \mathbb{Y})$, то $x_{d+1} \in X$, $\Delta_0 = 1$, $y_0 = z_{0,right} - 1$, $y = z_{right}$.

Для данных a, Y_0, y_0, y, Δ_0 требуется найти число разбиений $M(\mathbb{B}, \Delta_0, Y_0, y_0, y)$ цепи $\mathbb{B}$ так, чтобы никакой классификатор цепи не был лучше a.

4.3.1 Описание алгоритма

Крайние случаи те же.

Рассмотрим пару b_0, b_1. Напомним обозначения: $m_0 = n(b_0, \mathbb{Y})$, $m_1 = n(b_1, \mathbb{Y})$, $m_a = n(a, \mathbb{Y})$. Длина цепи (b_0, b_1) есть $h = |m_0 - m_1|$, классификаторы $c_0, \dots, c_h$ $(c_0 = b_0,\ c_h = b_1)$, ребра $z_1, \dots, z_h$,

Вводим параметр s – число ребер цепи (b_0, b_1), лежащих в X. Обозначаем через $\check{s}$ нижнюю границу значений параметра, через $\hat{s}$ – верхнюю.

При данных m_0, m_1, y_0, y, $\Delta(b_0)$ и s хотим найти число разбиений, таких что никакой c_i не лучше, чем a.

Лемма 4.8 *В зависимости от вида цепи (b_0, b_1), m_0, m_1, m_a и $\Delta(b_0)$ однозначно определяются $\Delta(b_1)$ и границы значений $\check{s}$, $\hat{s}$:*

1. *Если цепь возрастающая, то $\Delta(b_1) = \Delta(b_0) + s$, $\hat{s} = \min\{m_1 - m_0, y_0\}$.*
   (a) *Если $m_0 < m_a \leqslant m_1$*
      i. *Если $\Delta(b_0) > 0$, то $\check{s} = 0$.*
      ii. *Если $\Delta(b_0) = 0$, то $\check{s} = 1$.*
   (b) *Иначе нижняя граница $\check{s} = 0$.*
2. *Если цепь убывающая, то $\Delta(b_1) = \Delta(b_0) - s$, $\check{s} = 0$.*
   (a) *Если $m_a \leqslant m_1$, то $\hat{s} = \min\{\Delta(b_0) - 1, m_0 - m_1, y\}$.*
   (b) *Иначе $\hat{s} = \min\{\Delta(b_0), m_0 - m_1, y\}$.*

Обозначим через $Y_0' = Y_0 \backslash \{z_1, \dots, z_h\}$, $y_0' = |Y_0' \cap X|$, $y' = n(a, Y_0' \cap X)$. Правила получения y' и y_0' те же, что и в лемме 4.4.

Определим

$$\Delta = \begin{cases} \Delta(b_0), & m_0 < m_1, \\ \Delta(b_1), & m_1 < m_0. \end{cases}$$

Лемма 4.9 *Рассмотрим цепь (b_0, b_1). Пусть s – число ребер, попавших в обучающую выборку, $\check{s} \leqslant s \leqslant \hat{s}$. Тогда число разбиений ребер цепи (b_0, b_1), в которых никакой классификатор цепи не лучше, чем a, зависит только от s, m_a, Δ, $M = \max\{m_0, m_1\}, m = \min\{m_0, m_1\}$ (обозначим через $N(M, m, \Delta, m_a, s)$) и вычисляется по следующему правилу:*

1. *Если $m_a \leqslant m$ или $M < m_a$, то $N(M, m, \Delta, m_a, s) = C^s_{M-m}$.*
2. *Если $m < m_a \leqslant M$, то*
   (a) *Если $\Delta > 0$, то $N(M, m, \Delta, m_a, s) = C^s_{M-m}$,*

(b) *Если* $\Delta = 0$, *то* $N(M, m, \Delta, m_a, s) = C^s_{M-m} - C^s_{M-m_a}$.

Обозначим через $\mathbb{B}' = \mathbb{B} \backslash \{b_0\} = \{b_1, \ldots, b_T\}$, через $\Delta'_0 = \Delta(b_1)$.

Теорема 4.10 *В указанных обозначениях решение* $M(\mathbb{B}, \Delta_0, Y_0, y_0, y)$ *исходной задачи следующее:*

1. *Если* $\Delta_0 < 0$ *или* $\Delta_0 + y_0 - 2y < 0$, *то* $M(\mathbb{B}, \Delta_0, Y_0, y_0, y) = 0$.

2. *Если* $\mathbb{B}$ *состоит из одного классификатора, то*

$$M(\mathbb{B}, \Delta_0, Y_0, y_0, y) = \begin{cases} 0, & y_0 \neq 0, \\ 1, & y_0 = y = 0. \end{cases}$$

3. *Иначе*

$$M(\mathbb{B}, \Delta_0, Y_0, y_0, y) = \sum_{s=\check{s}}^{\hat{s}} N(M, m, \Delta, m_a, s) M(\mathbb{B}', \Delta'_0, Y'_0, y'_0, y').$$

Доказательство корректности повторяет доказательство для левой цепи.

Сложность такая же:

$$Time(\mathbb{B}, \Delta_0, Y_0, y_0, y) \leqslant \prod_{i=0}^{i=T-1} (1 + h_i),$$

где h_i – длина цепи (b_i, b_{i+1}).

4.4 Сложность алгоритма решения подзадач

Вернемся к цепи $\mathbb{A} = \{a_0, \ldots, a_D\}$ и формуле (4).

Обозначим через $h_1, \ldots h_K$ длины монотонных цепей, из которых составлена $\mathbb{A}$.

Сложность поиска $L(z_{0,left}, z_{left})$ органичена сверху произведением длин монотонных участков, из которых составлена левая цепь, увеличенных на 1. Данную величину можно оценить сверху как:

$$Time(L) = O(\prod_{i=1}^{K}(1 + h_i)).$$

Величина, стоящая в правой части, не зависит от $z_{0,left}, z_{left}$.

Аналогично, сложность вычисления $R(z_{0,right}, z_{right})$ есть

$$Time(R) = O(\prod_{i=1}^{K}(1 + h_i)).$$

Для данных $z_{0,left}, z_{left}$ произведение вычисляется за

$$Time(L) + Time(R) = O(\prod_{i=1}^{K}(1 + h_i)).$$

Оценка не зависит от значения параметров, перебор по всем парам $z_{0,left}$, z_{left} есть $O(l^2)$. Тогда $N_d(z_0, z)$ для данных z_0, z вычисляется за

$$O\left(2l^2\left(\prod_{i=1}^{K}(1 + h_i)\right)\right).$$

4.5 Сложность алгоритма решения исходной задачи

Итак, мы установили оценку сложности вычисления $N_d(z_0, z)$.

Вспомним формулу (3). Множители $\hat{H}_{z_0}()$ вычисляются за $O(l+m) = O(L)$. Перебор по всем парам z_0, z есть $O(l^2)$. Перебор по всем классификаторам цепи есть $O(D)$. Тогда сложность вычисления формулы (3) есть

$$O\left(Dl^4L\prod_{i=1}^{K}(1+h_i)\right)$$

4.6 Доказательство лемм

Доказательство леммы 4.2

Заметим, что Y_0 состоит из тех и только тех объектов генеральной выборки, на которых различимы классификаторы *цепи* $\mathbb{B}$, которая является прямой. Из того, что она прямая, следует, что мощность Y_0 *равна длине цепи.* Хеммингово расстояние между классификаторами b_0 и b_T также равно длине цепи $\mathbb{B}$. Значит, вектора ошибок данных классификаторов, *ограниченные на* Y_0 (на остальной части выборки классификаторы цепи $\mathbb{B}$ неразличимы), получаются друг из друга инвертированием, то есть заменой нуля на единицу, а единицы на нуль. По условию, из тех объектов, на которых b_0 ошибается, в X попадают ровно y, а из тех, на которых он не ошибается, ровно $y_0 - y$. Для b_T это значит, что в X попадает ровно $y_0 - y$ объектов, на которых он ошибается, и y, на которых нет. Остальные объекты попадают в контрольную выборку. Зная, что запас ошибок b_0 равен $\Delta(b_0)$, получаем тогда, что запас ошибок b_T равен

$$\Delta(b_T) = (\Delta(b_0) - y) + (y_0 - y) = \Delta(b_0) + y_0 - 2y.$$

Лемма доказана. ■

Леммы 4.4, 4.3, 4.5 докажем для случая, когда цепь (b_0, b_1) является возрастающей. Рассуждения для убывающей цепи (b_0, b_1) и для правой цепи $\{a_{d+1}, \dots, a_D\}$ аналогичны.

Имеем возрастающую цепь $(b_0, b_1) = \{c_0, \dots, c_h\}$, где $c_0 = b_0, c_h = b_1$, $z_1, \dots, z_h$ – соответствующие ребра цепи, $Y_0' = Y_0 \backslash \{z_1, \dots, z_h\}$, $m_0 = n(b_0, \mathbb{Y})$, $m_1 = n(b_0, \mathbb{Y})$, $m_a = n(a, \mathbb{Y})$.

Сначала найдем y_0' и y'. Очевидно, $y_0' = y_0 - s$: s ребер из (b_0, b_1) попали в обучение, поэтому осталось распределить $y_0 - s$ на Y_0'.

По условию, $y = n(b_0, Y_0 \cap X)$. Классификатор b_0 не ошибается на объектах $z_1, \dots, z_h$. Значит, он ошибается только на Y_0', то есть $y = n(b_0, Y_0' \cap X)$. Так как b_1 является первым классификатором цепи $\mathbb{B}'$, то $y' = n(b_1, Y_0' \cap X)$. Классификаторы b_0 и b_1 неразличимы на Y_0', следовательно, $y' = y$.

Далее, (b_0, b_1) – возрастающая цепь, поэтому каждое ребро, попавшее в обучающую выборку, увеличивает запас ошибок очередного классификатора. Следовательно, для каждого i запас ошибок c_i и число ошибок c_i на $\mathbb{Y}$ выражаются через $\Delta(b_0)$ и m_0 по следующим формулам:

$$\begin{aligned}&\Delta(c_i) = \Delta(b_0) + |z_1, \dots, z_i \cap X|,\\ &n(c_i, \mathbb{Y}) = m_0 + i.\end{aligned}$$

Следовательно, запас ошибок $\Delta(b_1) = \Delta(b_0) + s$.

Из формулы вычисления y_0' следует ограничение на s: $0 \leqslant s \leqslant \min\{m_1 - m_0, y_0\}$. Поэтому положим $\check{s} = 0$, $\hat{s} = \min\{m_1 - m_0, y_0\}$.

Максимальным из m_0 и m_1 является m_1, минимальным – m_0. $\Delta = \Delta(b_0)$. Найдем $N(m_1, m_0, \Delta(b_0), m_a, s)$ и, по мере необходимости, скорректируем $\hat{s}$ и $\check{s}$. Ответ зависит от соотношения между m_0, m_a, m_1:

1. $m_a < m_0$. В этом случае $\Delta(b_0)$ обязательно положительно (иначе b_0 лучше, чем a), на s дополнительных ограничений нет.
$$N(m_1, m_0, \Delta(b_0), m_a, s) = C^s_{m_1 - m_0}.$$
2. $m_0 \leqslant m_a < m_1$. Требуем, чтобы $\Delta(b_0) + s > 0$. Иначе классификатор b_1 лучше, чем a.
   (a) Если $\Delta(b_0) > 0$, то на s дополнительных ограничений нет, число разбиений равно
   $$N(m_1, m_0, \Delta(b_0), m_a, s) = C^s_{m_1 - m_0}.$$
   (b) Если $\Delta(b_0) = 0$, то на s накладывается дополнительное требование $s > 0$, то есть $\check{s} = 1$. Также требуется, чтобы у тех классификаторов c_i, которые допускают на $\mathbb{Y}$ больше ошибок, чем a, запас ошибок был положителен. Это значит выполнение двух условий
   $$\begin{aligned} &m_0 + i > m_a, \\ &|z_1, \dots, z_i \cap X| > 0. \end{aligned}$$
   Отсюда следует, что минимальное i, при котором допустимо равенство $|z_1, \dots, z_i \cap X| = 0$, есть $m_a - m_0$. А для $i = m_a - m_0 + 1$ среди $y_1, \dots, y_i$ должен быть хотя бы один элемент из X. Таким образом, получаем условие
   $$\Big[|z_1, \dots, z_{m_0 - m_a + 1} \cap X| > 0\Big] \Leftrightarrow 1 - [z_1, \dots, z_{m_a - m_0 + 1} \in \bar{X}].$$
   В итоге число всевозможных разбиений ребер цепи равно
   $$N(m_1, m_0, \Delta(b_0), m_a, s) = C^s_{m_1 - m_0} - C^s_{m_1 - m_a - 1}.$$
3. $m_1 \leqslant m_a$. В этом случае при любых $\Delta(b_0)$ и s нет классификатора, лучшего, чем a. Поэтому число всевозможных разбиений равно
$$N(m_1, m_0, \Delta(b_0), m_a, s) = C^s_{m_1 - m_0}.$$
На s дополнительных ограничений нет.

5 Доказательство теорем

5.1 Доказательство теоремы 3.2

Лемма 5.1 *Событие $[\delta(\mu X, X) \geqslant \varepsilon]$ эквивалентно следующему условию:*
$$n(\mu X, X) \leqslant \frac{l}{L}(n(\mu X, \mathbb{X}) - \varepsilon k).$$

Доказательство Пусть s - число ошибок классификатора μX на X, тогда число ошибок на контрольной выборке равно $n(\mu X, \bar{X}) = n(\mu X, \mathbb{X}) - s$. Несложными преобразованиями переобученность алгоритма, выбранного методом обучения, выражается как

$$\delta(\mu X, X) = \frac{1}{k}(n(\mu X, \mathbb{X}) - s\frac{L}{l}).$$

Тогда

$$\delta(\mu X, X) \geqslant \varepsilon \Leftrightarrow s \leqslant \frac{l}{L}(n(\mu X, \mathbb{X}) - \varepsilon k).$$

Лемма доказана ■

Доказательство леммы 3.1

Пусть классификатор a_d был выбран методом обучения. Тогда, согласно обозначениям, μX допускает на $\mathbb{X}$

$$n(\mu X, \mathbb{X}) = m_d + m$$

ошибок.

Пусть $n(a_d, Z_1 \cap X) = s$, то есть классификатор, выбранный методом обучения, допускает на обучающей выборке $z + s$ ошибок. Тогда, по предыдущей лемме, значение s ограничено сверху величиной, которая, по условию, есть $s_{d,z}(\varepsilon)$.

В X уже имеется z_0 элементов из Z_0. Тогда число разбиений Z_1 на $Z_1 \cap X$ и $Z_1 \cap \bar{X}$ равно

$$M_d(z_0, z) = \sum_{s=0}^{s_{d,z}(\varepsilon)} C_m^s C_{L-D-m}^{l-z_0-s}.$$

Вспомним формулу гипергеометрической функции распределения и заметим, что, поделив и домножив каждое слагаемое на $C_{L-D}^{l-z_0}$, получим в точности

$$M_d(z_0, z) = C_{L-D}^{l-z_0} H_{L-D}^{l-z_0, m}(s_{d,\ z}(\varepsilon)).$$

Лемма доказана. ■

5.2 Доказательство теоремы 2.2

Дана несимметричная цепь с максимумом $\mathbb{A} = \{a_h, \dots, a_D, a'_{D-1}, \dots, a'_0\}$.

Через Z_0 обозначаем множество объектов, на которых классификаторы цепи различимы. Для данного d $m_d = n(a_d, Z_0) = n(a'_d, Z_0)$.

При данных $z_0 = |Z_0 \cap X|$ и $z = n(a_d, Z_0 \cap X)$ обозначим через $N_d(z_0, z)$ число разбиений Z_0 так, что выполнено $[\mu X = a_d]$, через $N'_d(z_0, z)$ – выполнено $[\mu X = a'_d]$.

Тогда по (3) вероятность переобучения несимметричной цепи $\mathbb{A}$ равна

$$Q_\varepsilon = \sum_{d=h}^{D-1} \sum_{z_0=0}^{\hat{z}_0} \sum_{z=0}^{\hat{z}} N_d(z_0, z) \widetilde{H}_{z_0}(s_{d,z}(\varepsilon)) + \sum_{d=0}^{D-1} \sum_{z_0=0}^{\hat{z}_0} \sum_{z=0}^{\hat{z}} N'_d(z_0, z) \widetilde{H}_{z_0}(s_{d,z}(\varepsilon)) + \quad (5)$$

$$+ \sum_{z_0=0}^{\hat{z}_0} \sum_{z=0}^{\hat{z}} N_D(z_0, z) \widetilde{H}_{z_0}(s_{D,z}(\varepsilon)),$$

где $s_{d,z}(\varepsilon) = \frac{l}{L}(m + D + d - \varepsilon k)$.

Рассмотрим левую цепь относительно классификатора a_d: $\{a_d, \dots, a_h\}$. Она является убывающей. Запас ошибок $\Delta(a_{d-1}) = 0$. Тогда по лемме 4.3 в левой цепи нет ребер из обучающей выборки. Число разбиений левой цепи равно 1.

Рассмотрим правую цепь относительно a_D. Она также является убывающей, поэтому среди ребер цепи ни одно не попадает в обучающую выборку, значит, условие на параметры $z_0 = z = 0$, число разбиений ребер Z_0 равно

$$N_D(z_0, z) = [z_0 = 0, z = 0].$$

Рассмотрим правую цепь относительно классификатора a_d при $d < D$. Она составлена из двух монотонных цепей: возрастающей $\{a_{d+1}, \dots, a_D\}$ и убывающей $\{a'_D, \dots, a_0\}$.

Дано, что число ошибок a_d на $Z_0 \cap X$ равно z. По определению несимметричной цепи с максимумом, классификатор a_d ошибается на ребрах левой относительно него цепи и на ребрах цепи $\{a'_D, \dots, a'_0\}$, то есть на $\{x'_D, \dots, x'_1\}$. Поскольку ни одно из ребер левой цепи не попадает в X, то $z = |x'_D, \dots, x'_1 \cap X|$.

Тогда среди ребер возрастающей цепи $\{a_{d+1}, \dots, a_D\}$ в X попадают $z_0 - z$.

Запас ошибок $\Delta(a_{d+1}) = 1$, так как $x_{d+1} \in X$, иначе классификатор a_{d+1} оказывается лучше, чем a_d. Запас ошибок $\Delta(a_D) = z_0 - z$. Имеем ограничение на z вида $\Delta(a_D) \geqslant 1$, поскольку запас ошибок на возрастающей цепи не убывает. Число разбиений ребер возрастающей цепи равно $C_{D-d-1}^{z_0-z-1}$, так как $x_d \in X$.

По лемме 4.8 параметр z ограничен сверху: $z \leqslant \Delta(a_D)$. В случае строгого неравенства число разбиений ребер убывающей цепи по лемме 4.9 равно C_D^z, иначе $(C_D^z - C_{D-h}^z)$. Равенство возможно только при четных z_0.

Таким образом, для a_d получаем ответ

$$\begin{aligned}
d = D \Rightarrow & N_D(z_0, z) = [z_0 = z = 0]; \\
h \leqslant d < D \Rightarrow & N_d(z_0, z) = \begin{cases} C_{D-d-1}^{z_0-z-1} C_D^z, & z \leqslant \hat{z}(z_0), \\ \delta(z_0) C_{D-d-1}^{z_0/2-1} (C_D^{z_0/2} - C_{D-d}^{z_0/2}), & z = z_0/2; \end{cases}
\end{aligned}$$

Аналогично, рассмотрев правую и левую цепи относительно $a'_d, d < D$, получаем:

$$\begin{aligned}
h \leqslant d < D \Rightarrow & N_d(z_0, z) = \begin{cases} C_{D-d-1}^{z_0-z-1} C_{D-h}^z, & z \leqslant \hat{z}(z_0), \\ \delta(z_0) C_{D-d-1}^{z_0/2-1} (C_{D-h}^{z_0/2} - C_{D-d-1}^{z_0/2}), & z = z_0/2, \end{cases} \\
0 \leqslant d < h \Rightarrow & N_d(z_0, z) = \begin{cases} C_{D-d-1}^{z_0-z-1} C_{D-h}^z, & z \leqslant \hat{z}(z_0), \\ 0, & z = z_0/2. \end{cases}
\end{aligned}$$

Подставляем данные выражения в (5), группируем и получаем искомое равенство.

6 Вычислительные эксперименты

На примере симметричной цепи с максимумом покажем, что существующие оценки являются завышенными для данного семейства.

Напомним обозначения. Дана цепь $\mathbb{A} = \{a_0, \dots, a_D, \dots, a'_0\}$ на генеральной выборке $\mathbb{X}$ длины L с параметром $m = n(a_0, \mathbb{X})$. Длина обучающей выборки полагается равной l. Точность ε.

Для начала продемонстрируем правильность полученной формулы с помощью метода Монте-Карло. На рис. 4 показаны результаты вычисления вероятности переобучения методом Монте-Карло на 100000 случайных разбиений и по формуле (1). Близость точек позволяет утверждать, что формула верна.

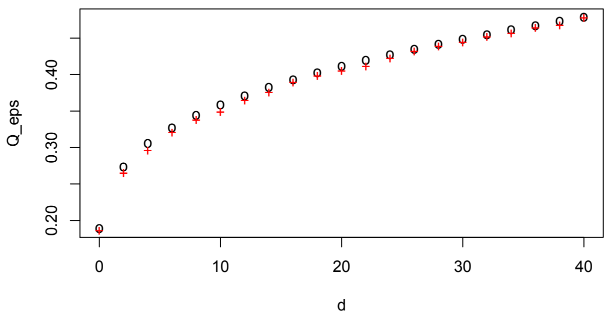

Рис. 4: Красным отмечены значения вероятности переобучения, полученные методом Монте-Карло, черным – полученные по формуле (1). Условия эксперимента: $L = 200$, $l = 100$, $m = 40$, $\varepsilon = 0.05$. Параметр D принимает последовательные значения от 0 до 40 с шагом 2.

6.1 Оценка Вапника–Червоненкиса

Пусть дано семейство классификаторов $\mathbb{B} = \{b_1, \ldots, b_T\}$ с известной матрицей ошибок на генеральной выборке $\mathbb{X}$ длины L с попарно различными столбцами – векторами ошибок классификаторов. Длина обучающей выборки равна l.

Тогда верна оценка Вапника–Червоненкиса [2]:

$$Q_\varepsilon \leqslant T \max_{m=1,\ldots,L} H_L^{l,m}\Big(\frac{l}{L}(m - \varepsilon k)\Big). \tag{6}$$

На рис.5 показано, как сильно завышена оценка Вапника–Червоненкиса для унимодальной цепи с максимумом. Как мы видим, оценка быстро приходит к насыщению – превышению значения 1.

6.2 Улучшенная оценка расслоения – связности

Пусть дано семейство классификаторов $\mathbb{B} = \{b_1, \ldots, b_T\}$ с известной матрицей ошибок на генеральной выборке $\mathbb{X}$ длины L. Длина обучающей выборки равна l.

На множестве классификаторов, как векторов ошибок, существует отношение лексикографического порядка $\leqslant$. Будем говорить, что классификатор a *предшествует* b и записывать $a \prec b$, если $a \leqslant b$ и расстояние Хемминга между ними равно 1.

Будем называть классификатор s *истоком*, если нет классификаторов b, таких что $b \prec s$.

Через $u(a)$ будем обозначать

$$u(a) = |\{b \in \mathbb{B} \mid a \prec b\}|.$$

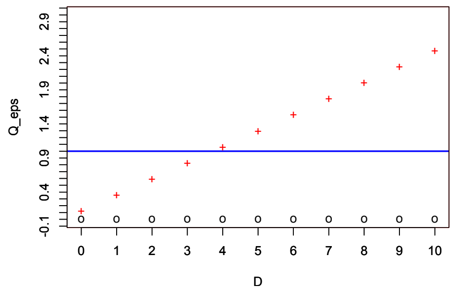

Рис. 5: Красным отмечены значения оценки Вапника–Червоненкиса, черным – точные значения вероятности переобучения Условия эксперимента: $L = 1000$, $l = 800$, $m = 20$, $\varepsilon = 0.05$. Параметр D принимает значения от 0 до 10..

Через $m(a)$ будем обозначать число ошибок a на генеральной выборке.

Пусть даны два классификатора a_i и a_j. Тогда через A_{ij} будем обозначать множество

$$A_{ij} = \{x \in \mathbb{X} |\ I(a_i, x) = 0,\ I(a_j, x) = 1\},$$

а через

$$B_{ij} = \{x \in X |\ I(a_i, x) = 1,\ I(a_j, x) = 0\}.$$

Следующую теорему приведем без доказательства.

Теорема 6.1 *Пусть S – множество истоков семейства $\mathbb{B}$. Тогда верна следующая оценка*

$$Q_\varepsilon \leqslant \sum_{i=1}^{T} \min_{s \in S} \Bigg\{ \sum_{t=0}^{\min\{|A_{is}|,\ |B_{is}|\}} \frac{C^t_{|B_{is}|} C^{l-u-t}_{L-u-|B_{is}|}}{C^l_L} H^{l-u-t,\ m-|B_{is}|}_{L-u-|B_{is}|} \left(\frac{l}{L}(m - \varepsilon k) - t \right) \Bigg\} \tag{7}$$

Как мы видим на рис.6, данная оценка сильно завышена для унимодальной цепи с максимумом.

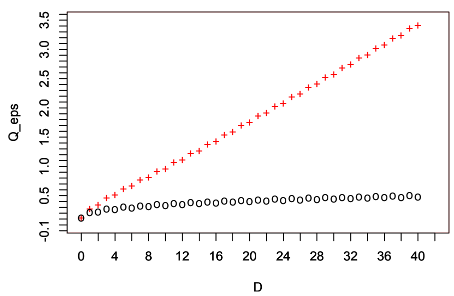

Рис. 6: Красным отмечены значения оценки (7), черным – точные значения вероятности переобучения. Условия эксперимента: $L = 200$, $l = 100$, $m = 20$, $\varepsilon = 0.05$. Параметр D принимает значения от 0 до 40.

Список литературы

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

[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] *Воронцов К.В.* Комбинаторная теория надежности переобучения по прецедентам: Дис. док. физ.-мат. наук: 05-13-17: Ph.D. thesis / Вычислительный центр РАН. - 2010. - С.271

[4] *Воронцов К.В.* Теория надежности обучения по прецедентам (комбинаторная теория переобучения). - Курс лекций. - 2012. - С.177

[5] *Журавлёв, Ю. И.* Об алгебраическом подходе к решению задач распознавания или классификации // Проблемы кибернетики: Вып.33. — 1978. — С. 5–68.

[6] *Журавлёв, Ю. И., Рязанов, В. В., Сенько, О. В.* «Распознавание». Математические методы. Программная система. Практические применения. — М.: ФАЗИС, 2006. — 176 с.

[7] *Ивахненко А. А., Воронцов К. В.* Критерии информативности пороговых логических правил с поправкой на переобучение порогов // 15-я Всеросс. конф. Математические методы распознавания образов. — М.: МАКС Пресс, 2011. —С. 48–51.

[8] *Vorontsov K. V., Ivahnenko A. A.* Tight combinatorial generalization bounds for threshold conjunction rules // 4-th Int'l Conf. on Pattern Recognition and Machine Intelligence (PReMI'11), June 27 – July 1, 2011. Lecture Notes in Computer Science. Springer-Verlag, 2011. — Pp. 66–73.

[9] *Vorontsov K. V.* Exact combinatorial bounds on the probability of overfitting for empirical risk minimization // *Pattern Recognition and Image Analysis.* — 2010. — Vol. 20, no. 3. — Pp. 269–285.