УДК 519.24

Н. К. Животовский

Московский физико-технический институт (государственный университет)
Институт проблем передачи информации им. А. А. Харкевича

Комбинаторные оценки переобучения с сублогарифмическим темпом роста

В рамках комбинаторной теории переобучения получены верхние оценки математического ожидания переобученности, имеющие в худшем случае порядок роста $O\big(\sqrt{\log|A|}\big)$, где $|A|$ — число алгоритмов в семействе. Также получены оценки, зависящие от характеристик расслоения и связности семейства алгоритмов, которые являются ещё более точными.

Ключевые слова: теория статистического обучения, комбинаторная теория переобучения, обобщающая способность, переобучение, расслоение, связность.

N. K. Zhivotovskiy

[1]Moscow Institute of Physics and Technology (State University)
[2]Institute for Information Transmission Problems (Kharkevich Institute)

Combinatorial bounds of overfitting with sublogarithmic order of growth

In terms of combinatorial theory of overfitting, new upper bounds of expected overfitting are obtained. They have the worst order of growth $O\big(\sqrt{\log|A|}\big)$, where $|A|$ is a number of algorithms. Also, some new bounds, depending on splitting and connectivity properties of algorithms, are obtained and are even tighter.

Key words: statistical learning theory, combinatorial theory of overfitting, generalization ability, overfitting, splitting, connectivity.

1. Введение

Получение верхних оценок вероятности ошибки на основе доступной информации о семействе алгоритмов, обучающей выборке и методе обучения — одна из центральных проблем теории статистического обучения [6, 9]. Минимизация таких оценок позволяет управлять обобщающей способностью алгоритма классификации на этапе его обучения по выборке. Проблема в том, что большинство оценок сильно завышены, что может приводить к неоптимальным решениям.

Комбинаторная теория переобучения [12–15] отличается от теории Вапника–Червоненкиса и других более современных подходов тем, что в ней оцениваются функционалы вероятности переобучения, полного скользящего контроля, ожидаемой переобученности, которые наиболее точно формализуют понятие переобучения и позволяют учитывать особенности конкретной выборки данных и метода обучения. Комбинаторные оценки показывают, что величина переобучения определяется не только сложностью семейства алгоритмов, но и более тонкими эффектами расслоения и связности, которые существенно снижают переобучение в реальных задачах классификации. Комбинаторный подход позволяет учитывать оба эти эффекта одновременно и получать оценки вероятности переобучения, которые в некоторых случаях оказываются точными равенствами [3]. Отметим, что частично учитывают структуру семейства алгоритмов и некоторые другие подходы, например, верхние оценки, основанные на локальных Радемахеровских сложностях в условиях малого шума [5].

Известные оценки вероятности переобучения, ожидаемой переобученности и полного скользящего контроля представляются в виде суммы по всем алгоритмам семейства [2, 14].

В данной работе развивается техника, основанная на построении верхних оценок производящих функций моментов оцениваемых величин. В результате получены существенно улучшенные оценки ожидаемой переобученности и полного скользящего контроля.

1.1. Основные определения

Введём основные понятия, придерживаясь обозначений, принятых в комбинаторной теории переобучения [12–15].

Задана конечная *генеральная совокупность объектов* $\mathbb{X} = \{x_1, \ldots, x_L\}$, конечное множество *алгоритмов* $A = \{a_1, \ldots, a_D\}$ и бинарная *функция потерь* $I\colon A \times \mathbb{X} \to \{0, 1\}$, где $I(a, x) = 1$ тогда и только тогда, когда алгоритм a ошибается на объекте x. *Вектором ошибок* алгоритма a называется бинарный вектор $\big(I(a, x_1), \ldots, I(a, x_L)\big)$ размерности L. Предполагается, что векторы ошибок всех алгоритмов из A попарно различны.

Определяются число ошибок и частота ошибок алгоритма $a \in A$ на выборке $X \subseteq \mathbb{X}$:

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

Методом обучения называется отображение $\mu\colon 2^{\mathbb{X}} \to A$, которое произвольной выборке $X \subset \mathbb{X}$ ставит в соответствие некоторый алгоритм $\mu X \in A$.

Метод обучения μ называется методом *минимизации эмпирического риска* (МЭР), если $\mu X \in A(X)$ для всех выборок $X \subset \mathbb{X}$, где

$$A(X) = \operatorname{Arg}\min_{a \in A} n(a, X), \quad X \subset \mathbb{X}.$$

Если множество $A(X)$ содержит более одного элемента, то выбор алгоритма методом МЭР не однозначен. Будем рассматривать худший случай. Метод МЭР называется *пессимистичным* (ПМЭР), если

$$\mu X \in \operatorname{Arg}\max_{a \in A(X)} n(a, \bar{X}), \quad X \subset \mathbb{X}.$$

Обычно в теории статистического обучения предполагается, что элементы выборки порождаются случайно и независимо из фиксированного неизвестного распределения. Комбинаторная теория переобучения основана на более слабой *гипотезе перестановочности*. Предполагается, что все $L!$ перестановок объектов конечной генеральной совокупности $\mathbb{X}$ равновероятны. Сами объекты предполагаются произвольными и неслучайными, никакой меры на множестве объектов не вводится, и даже не предполагается существование каких-то других объектов кроме $\mathbb{X}$. Случайным считается только порядок появления объектов. В момент, когда метод μ выбирает алгоритм $a = \mu X$, обучающая выборка X предполагается известной, выборка $\bar{X}$ из оставшихся $k = L - \ell$ объектов — скрытой. Нас интересует оценка частоты ошибок выбранного алгоритма $\nu(a, \bar{X})$ на будущих данных. Эта оценка характеризует обобщающую способность метода μ и на практике может использоваться в качестве критерия выбора модели алгоритмов A или метода обучения μ.

В комбинаторной теории понятие «вероятности ошибки» не определяется, оцениваются только частоты ошибок на конечных выборках. Все используемые функции выборок X и $\bar{X}$ инвариантны относительно перестановок объектов внутри этих выборок. Поэтому основное вероятностное предположение можно ещё немного ослабить, считая равновероятными все C_L^ℓ разбиений генеральной совокупности $\mathbb{X} = X \sqcup \bar{X}$ на две выборки — *наблюдаемую обучающую* X длины ℓ и *скрытую контрольную* $\bar{X}$ длины k.

Вероятность переобучения метода μ на выборке $\mathbb{X}$ определяется как доля разбиений, при которых частота ошибок на контроле превосходит частоту ошибок на обучении на ε или более:

$$Q_\varepsilon(\mu, \mathbb{X}) = \mathsf{P}\big[\nu(\mu X, \bar{X}) - \nu(\mu X, X) \geq \varepsilon\big].$$

Введём для $\ell = k = \frac{L}{2}$ функционал равномерной *ожидаемой переобученности* (expectation of overfitting), равный средней по всем разбиениям разности между частотами ошибок на обучении и контроле (при этом в равномерном случае на каждом разбиении выбирается худший алгоритм):

$$\mathcal{EOF}_{\max}(\mathbb{X}) = \mathbf{E} \max_{a \in A} \left(\nu(a, \bar{X}) - \nu(a, X) \right),$$

и функционал ожидаемой переобученности, учитывающий метод обучения μ (которым в данной работе будет являться ПМЭР):

$$\mathcal{EOF}_{\mu}(\mathbb{X}) = \mathbf{E} \left(\nu(\mu X, \bar{X}) - \nu(\mu X, X) \right).$$

Введём также функционал полного скользящего контроля (complete cross-validation):

$$\mathcal{CCV}_{\mu}(\mathbb{X}) = \mathbf{E}\nu(\mu X, \bar{X}).$$

1.2. Оценки расслоения–связности

В отличие от статистической теории обучения, в комбинаторном подходе рассматриваются исключительно бинарные функции потерь. Это ограничение делает очень удобным рассмотрение метрических свойств множества A. Введём на множестве алгоритмов A, как на бинарных векторах ошибок, естественное отношение порядка и метрику Хэмминга: для любых $a, b \in A$

$$(a \leq b) \;\leftrightarrow\; (I(a,x) \leq I(b,x)\ \forall x \in \mathbb{X});$$
$$(a < b) \;\leftrightarrow\; (a \leq b \;\text{ и }\; a \neq b);$$
$$\rho(a,b) = \sum_{i=1}^{L} \big[I(a,x_i) \neq I(b,x_i) \big].$$

Если $a \leqslant b$ и $\rho(a,b) = 1$, то будем говорить, что a *предшествует* b и записывать $a \prec b$.

Графом расслоения–связности множества алгоритмов A будем называть направленный граф $\langle A, E \rangle$ с множеством рёбер $E = \big\{(a,b)\colon a \prec b\big\}$.

Граф расслоения–связности является многодольным, доли соответствуют *слоям алгоритмов* $A_m = \big\{a \in A\colon n(a,\mathbb{X}) = m\big\}$, рёбрами могут соединяться только алгоритмы соседних слоёв. Каждому ребру (a,b) соответствует единственный объект $x_{ab} \in \mathbb{X}$, такой, что $I(a,x_{ab}) = 0$ и $I(b,x_{ab}) = 1$.

Заметим, что если для любых $a, b \in A,\ a < b$ существует путь $a \prec a_1 \cdots \prec a_s = b$, то граф расслоения–связности совпадает с диаграммой Хассе [7] отношения порядка, введённого на множестве алгоритмов A. В общем случае он является лишь её подграфом. Заметим также, что граф расслоения–связности не обязательно является связным графом.

Порождающим множеством X_a алгоритма a называется множество объектов, соответствующих исходящим из вершины a рёбрам:

$$X_a = \big\{ x \in \mathbb{X} \;\big|\; \exists b \in A\colon a \prec b,\ I(a,x) < I(b,x) \big\}.$$

Запрещающим множеством X'_a алгоритма a называется множество объектов x, на которых алгоритм a ошибается, при том, что существует алгоритм $b \in A,\ b < a$, не ошибающийся на x:

$$X'_a = \big\{ x \in \mathbb{X} \;\big|\; \exists b \in A\colon b < a,\ I(b,x) < I(a,x) \big\}.$$

Верхней связностью алгоритма a называется число рёбер, исходящих из вершины a. Оно равно числу объектов x, на которых a не ошибается, при том, что существует алгоритм $b \in A,\ a \prec b$, ошибающийся на x:

$$q(a) = |X_a|.$$

Неполноценностью алгоритма a называется число объектов x, на которых a ошибается, при том, что существует алгоритм $b \in A,\ b < a$, не ошибающийся на x:

$$r(a) = |X'_a|.$$

Введем также удобное обозначение

$$m(a) = n(a, \mathbb{X}).$$

Определим функцию гипергеометрического распределения:

$$H_L^{\ell,\,m}(z) = \sum_{s=0}^{\lfloor z \rfloor} \frac{C_m^s C_{L-m}^{\ell-s}}{C_L^\ell}.$$

Следующая лемма, доказанная в [14], описывает важное свойство порождающего и запрещающего множеств.

Лемма 1.1. *Пусть μ — пессимистичная минимизация эмпирического риска, тогда $\forall a \in A$ выполнено следующее равенство:*

$$\big[\mu X = a\big] \leqslant \big[X_a \subset X\big]\big[X'_a \subset \bar{X}\big].$$

Таким образом, для того, чтобы пессимистичная минимизация эмпирического риска выбрала некоторый алгоритм, необходимо, чтобы его порождающее и запрещающее множества были соответственно подмножествами обучающей и контрольной подвыборок.

Эта оценка существенно зависит от характеристик $q(a)$, $r(a)$ каждого алгоритма $a \in A$. Она монотонно убывает по $q(a)$, $r(a)$. Если положить $q(a) = r(a) = 0$ и затем применить экспоненциальную верхнюю оценку функции гипергеометрического распределения, то получится классическая оценка Вапника–Червоненкиса.

Теорема 1.1 (оценка расслоения–связности [14])**.** *Для произвольной выборки $\mathbb{X}$, метода минимизации эмпирического риска μ и любого $\varepsilon \in (0, 1)$*

$$Q_\varepsilon(\mu, \mathbb{X}) \leq \sum_{a \in A} \frac{C_{L-q-r}^{\ell-q}}{C_L^\ell} H_{L-q-r}^{\ell-q,\,m-r}\left(\tfrac{\ell}{L}(m - \varepsilon k)\right),$$

где q — верхняя связность, r — неполноценность алгоритма a, $m = m(a)$.

Для функционала полного скользящего контроля аналогичная оценка получена в [2].

Теорема 1.2. *Для произвольной выборки $\mathbb{X}$ и метода минимизации эмпирического риска μ*

$$\mathcal{CCV}_\mu(\mathbb{X}) \leq \sum_{a \in A} \frac{C_{L-q-r}^{\ell-q}}{C_L^\ell} \left(\frac{m}{k} - \frac{m-r}{k}\frac{\ell-q}{L-q-r}\right).$$

Там же для функционала ожидаемой переобученности получена аналогичная оценка.

Теорема 1.3. *Для произвольной выборки $\mathbb{X}$ и метода минимизации эмпирического риска μ*

$$\mathcal{EOF}_\mu(\mathbb{X}) \leq \sum_{a \in A} \frac{C_{L-q-r}^{\ell-q}}{C_L^\ell} \left(\frac{m}{k} - \frac{m-r}{k}\frac{\ell-q}{L-q-r}\frac{L}{\ell}\right).$$

Все три оценки расслоения–связности имеют схожую структуру. Каждая из них представляет собой сумму по алгоритмам семейства, и в худшем случае ($q = r = 0$) все три оценки имеют порядок $O(|A|)$. Целью данной работы является получение оценок $\mathcal{EOF}_\mu$, имеющих меньший порядок роста и также учитывающих расслоение и связность.

2. Оценки ожидаемой переобученности

2.1. Равномерная оценка ожидаемой переобученности

В теории статистического обучения используются различные неравенства концентрации меры, основанные на гипотезе независимости элементов выборки [9]. В комбинаторной теории переобучения необходимо использовать аналогичные неравенства, справедливые при предположении о перестановочности.

Функция ${}_2F_1$, определённая в круге $|z| < 1$ и заданная выражением

$$ {}_2F_1\big(a,b,c,z\big) = 1 + \sum_{k=1}^{\infty}\left[\prod_{l=0}^{k-1}\frac{(a+l)(b+l)}{(1+l)(c+l)}\right]z^k, $$

где a, b, c — действительные параметры, называется *гипергеометрической*.

Лемма 2.1. *Для целых чисел m, ℓ, таких что $0 \leqslant m \leqslant \ell$, и действительного $z \in [0,1]$*

$$ {}_2F_1\left(\frac{1-m}{2}, -\frac{m}{2}, \frac{1}{2} - \ell, z\right) \leqslant 1. $$

Доказательство данной леммы чисто техническое и вынесено в конец статьи.

Следующая лемма является одной из основных в используемом подходе. Она для фиксированного алгоритма даёт верхнюю оценку на производящую функцию моментов разности числа его ошибок на обучении и контроле. В несколько более общей постановке производящая функция моментов оценивалась для выборок без возвращений в работе [10].

Лемма 2.2. *Пусть $a \in A$, $\ell = k = \frac{L}{2}$ и $m(a) = m \leqslant \ell$, тогда для всех $\lambda > 0$*

$$ \mathbf{E}\exp\big(\lambda(n(a,\bar{X}) - n(a,X))\big) \leqslant (\cosh(\lambda))^{\ell}\,{}_2F_1\left(\frac{1-m}{2}, -\frac{m}{2}, \frac{1}{2} - \ell, (\tanh(\lambda))^2\right). $$

Доказательство. Произвольному разбиению генеральной выборки сопоставим вектор $\sigma = (\sigma_1, \ldots, \sigma_L)$, где на половине позиций стоят -1, которые соответствуют элементам обучающей выборки в $\mathbb{X}$, а на остальных позициях – 1. Без ограничения общности перенумеруем генеральную выборку так, чтобы алгоритм ошибался на первых m объектах, а на оставшихся не допускал ошибок. Запишем производящую функцию моментов в виде

$$ \mathbf{E}_\sigma \exp\left(\lambda\sum_{i=1}^{L}\sigma_i I(a,x_i)\right). $$

Обозначим $\hat{x}_i = \sigma_i I(a,x_i)$, тогда с учётом $m(a) \leqslant \ell$

$$ \mathbf{E}_\sigma \exp\big(\lambda\sum_{i=1}^{L}\sigma_i I(a,x_i)\big) = \mathbf{E}_\sigma \prod_{i=1}^{L}\exp(\lambda\hat{x}_i) = \mathbf{E}_\sigma \prod_{i=1}^{\ell}\exp(\lambda\hat{x}_i). $$

Очевидно, что для всех i: $\lambda\hat{x}_i \in [-\lambda, \lambda]$. Используя выпуклость экспоненты, получаем:

$$ \exp(\lambda\hat{x}_i) \leqslant \frac{\lambda\hat{x}_i + \lambda}{\lambda+\lambda}\exp(\lambda) + \frac{-\lambda\hat{x}_i + \lambda}{\lambda+\lambda}\exp(-\lambda) = \hat{x}_i\sinh(\lambda) + \cosh(\lambda). $$

Подставляем полученное неравенство в предыдущее выражение:

$$ \mathbf{E}_\sigma\left(\prod_{i=1}^{\ell}\exp(\lambda\hat{x}_i)\right) \leqslant \mathbf{E}_\sigma \prod_{i=1}^{\ell}\left(\hat{x}_i\sinh(\lambda) + \cosh(\lambda)\right). $$

Раскроем скобки в полученном выражении и учтём, что последние $\ell - m$ значений $\hat{x}_i$ тождественно равны нулю, так как алгоритм не ошибается на этих объектах. Учтём, что

$$\mathbf{E}_\sigma \prod_{i=1}^{\ell}(\hat{x}_i \sinh(\lambda) + \cosh(\lambda)) = (\cosh(\lambda))^{\ell-m}\mathbf{E}_\sigma \prod_{i=1}^{m}(\hat{x}_i \sinh(\lambda) + \cosh(\lambda)).$$

Очевидно, что для всех $i \leqslant m$ математическое ожидание произведения одинакового числа объектов $\hat{x}_i$, соответствующих различным индексам, одинаково. Таким образом,

$$\begin{aligned}
&(\cosh(\lambda))^{\ell-m}\mathbf{E}_\sigma \prod_{i=1}^{m}(\hat{x}_i \sinh(\lambda) + \cosh(\lambda)) = \\
&= (\cosh(\lambda))^{\ell-m}\mathbf{E}_\sigma(\cosh^m(\lambda) + C_m^1\hat{x}_1 \sinh^1(\lambda)\cosh^{m-1}(\lambda) + \\
&+ C_m^2\hat{x}_1\hat{x}_2 \sinh^2(\lambda)\cosh^{m-2}(\lambda) + \ldots + \hat{x}_1 \ldots \hat{x}_m \sinh^m(\lambda)) = \\
&= (\cosh(\lambda))^{\ell}\mathbf{E}_\sigma(1 + C_m^1\hat{x}_1 \tanh^1(\lambda) + \ldots + \hat{x}_1 \ldots \hat{x}_m \tanh^m(\lambda)).
\end{aligned}$$

Рассмотрим для неотрицательного целого r выражение $\mathbf{E}(\hat{x}_1 \ldots \hat{x}_{2r+1})$.

Так как по условию $m \leqslant \ell$, то $2r + 1 \leqslant \ell$. Тогда для каждого разбиения генеральной выборки на обучение и контроль знак $\hat{x}_1 \ldots \hat{x}_{2r+1}$ зависит лишь от четности числа объектов $x_1, \ldots, x_{2r+1}$, попавших в контроль. Число разбиений, на которых всё нечётное число этих объектов попадает в обучение, равно числу разбиений, когда все эти объекты попадают в контроль. Вклад таких разбиений в математическое ожидание просто противоположен по знаку. Также одинаков по модулю и противоположен по знаку вклад разбиений, где лишь один из перечисленных объектов в обучении и где все, кроме одного, в обучении. И так далее компенсируем вклады всех 2^{2r+1} вариантов помещения части объектов $\hat{x}_1 \ldots \hat{x}_{2r+1}$ в обучение.

Пусть $j = 2r$. Выведем точную формулу для $\mathbf{E}(\hat{x}_1 \ldots \hat{x}_j)$. Как и ранее, сосчитаем вклады разбиений в математическое ожидание. Введём суммирование по i — числу объектов $x_1, \ldots, x_j$, попавших в обучение. Очевидно, что вклад разбиения равен $(-1)^i$. Число разбиений сосчитать не сложно: сначала выберем i позиций среди j для помещения в обучение, оставшиеся объекты помещаем в контроль. При этом мы еще не учли все возможные разбиения, а именно: нужно среди оставшихся $2l - j$ объектов выбрать $\ell - j + i$ и поместить их в контроль. Объединяя результат, получаем формулу

$$\mathbf{E}\ \hat{x}_1 \ldots \hat{x}_j = \frac{1}{C_{2\ell}^{\ell}} \sum_{i=0}^{j} (-1)^i C_{2\ell-j}^{\ell-j+i} C_j^i.$$

Это же выражение верно и для нечётных j, при этом оно тождественно равно нулю. В итоге получаем

$$\mathbf{E}(1 + C_m^1\hat{x}_1 \tanh^1(\lambda) + \ldots + \hat{x}_1 \ldots \hat{x}_m \tanh^m(\lambda)) = \frac{1}{C_{2\ell}^{\ell}} \sum_{j=0}^{m} C_m^j (\tanh(\lambda))^j \sum_{i=0}^{j} (-1)^i C_{2\ell-j}^{\ell-j+i} C_j^i.$$

Далее остается доказать, что выполнено равенство

$$\frac{1}{C_{2\ell}^{\ell}} \sum_{j=0}^{m} C_m^j (\tanh(\lambda))^j \sum_{i=0}^{j} (-1)^i C_{2\ell-j}^{\ell-j+i} C_j^i = {}_2F_1\left(\frac{1-m}{2}, -\frac{m}{2}, \frac{1}{2} - \ell, (\tanh(\lambda))^2\right).$$

Это утверждение чисто техническое и доказывается индукцией по параметрам.

В условиях предыдущей леммы можно получить более точную верхнюю оценку:

$$\mathbf{E} \exp\left(\lambda(n(a, \bar{X}) - n(a, X))\right) \leqslant (\cosh(\lambda))^{m(a)} {}_2F_1\left(\frac{1-m}{2}, -\frac{m}{2}, \frac{1}{2} - \ell, (\tanh(\lambda))^2\right).$$

Тем не менее во избежание чрезмерной громоздкости мы будем использовать более грубый результат леммы 2.2.

После этого можно сформулировать основную теорему данного раздела. Это некоторое обобщение леммы об ожидаемом максимуме субгауссовских случайных величин [8, 9].

Теорема 2.1. *Пусть* $\ell = k = \frac{L}{2}$ *и* $\max\limits_{a\in A} m(a) \leqslant \frac{L}{2}$, *тогда*

$$\begin{aligned}
&\mathcal{EOF}_{\max} \leqslant \\
&\leqslant \inf_{\lambda>0} \frac{1}{\lambda}\left(\ln(\cosh(\lambda)) + \frac{1}{\ell}\ln\left(\sum_{s=0}^{l}\Delta_s\ {}_2F_1\left(\frac{1-s}{2}, -\frac{s}{2}, \frac{1}{2}-\ell, (\tanh(\lambda))^2\right)\right)\right) \leqslant \\
&\leqslant \sqrt{\frac{2\ln|A|}{\ell}},
\end{aligned}$$

где Δ_s *— число алгоритмов в* s*-м слое семейства алгоритмов.*

Доказательство. Математическое ожидание по совокупности σ_i будем обозначать символом $\mathbf{E}$. В обозначениях предыдущей леммы оценивается величина

$$\lambda\mathbf{E}\max_{a\in A}\left(\sum_{i=1}^{L}\sigma_i I(a,x_i)\right) = \ln\left(\exp\left(\lambda\mathbf{E}\max_{a\in A}\left(\sum_{i=1}^{L}\sigma_i I(a,x_i)\right)\right)\right).$$

По неравенству Йенсена:

$$\ln\left(\exp\left(\lambda\mathbf{E}\max_{a\in A}\left(\sum_{i=1}^{L}\sigma_i I(a,x_i)\right)\right)\right) \leqslant \ln\left(\mathbf{E}\exp\left(\lambda\max_{a\in A}\left(\sum_{i=1}^{L}\sigma_i I(a,x_i)\right)\right)\right).$$

Максимум может быть вынесен

$$\ln\left(\mathbf{E}\exp\left(\lambda\max_{a\in A}\left(\sum_{i=1}^{L}\sigma_i I(a,x_i)\right)\right)\right) = \ln\left(\mathbf{E}\max_{a\in A}\exp\left(\lambda\left(\sum_{i=1}^{L}\sigma_i I(a,x_i)\right)\right)\right).$$

Заменяем максимум неотрицательных величин на их сумму:

$$\ln\left(\mathbf{E}\max_{a\in A}\exp\left(\lambda\left(\sum_{i=1}^{L}\sigma_i I(a,x_i)\right)\right)\right) \leqslant \ln\left(\mathbf{E}\sum_{a\in A}\exp\left(\lambda\sum_{i=1}^{L}\sigma_i I(a,x_i)\right)\right).$$

По лемме 2.2

$$\begin{aligned}
&\ln\left(\mathbf{E}\sum_{a\in A}\exp\left(\lambda\sum_{i=1}^{L}\sigma_i I(a,x_i)\right)\right) \leqslant \\
&\leqslant \ln\left(\sum_{a\in A}(\cosh(\lambda))^{\ell}\,{}_2F_1\left(\frac{1-m(a)}{2}, -\frac{m(a)}{2}, \frac{1}{2}-\ell, (\tanh(\lambda))^2\right)\right).
\end{aligned}$$

Из цепочки неравенств на предыдущем шаге имеем

$$\mathbf{E}\max_{a\in A}\left(\sum_{i=1}^{L}\sigma_i I(a,x_i)\right) \leqslant \frac{1}{\lambda}\ln\left((\cosh(\lambda))^{\ell}\sum_{a\in A}{}_2F_1\left(\frac{1-m}{2}, -\frac{m}{2}, \frac{1}{2}-\ell, (\tanh(\lambda))^2\right)\right).$$

Для доказательства первого из неравенств теперь достаточно прологарифмировать данное выражение и минимизировать его по λ. Затем необходимо ввести суммирование по слоям, так как в каждом слагаемом суммы алгоритм характеризуется лишь числом ошибок.

Чтобы доказать верхнее неравенство, воспользуемся элементарными неравенствами

$$\cosh(\lambda) \leqslant \exp\left(\frac{\lambda^2}{2}\right) \quad \text{и} \quad (\cosh(\lambda))^\ell \leqslant \exp\left(\frac{\ell\lambda^2}{2}\right).$$

С учётом этих неравенств и леммы 2.1

$$\mathbf{E}\max_{a\in A}\left(\sum_{i=1}^{L}\sigma_i I(a,x_i)\right) \leqslant \inf_{\lambda>0}\frac{1}{\lambda}\left(\ln|A| + \frac{\ell\lambda^2}{2}\right) = \sqrt{2\ell\ln|A|}.$$

Поделив обе части неравенства на ℓ, получаем правое неравенство из утверждения теоремы.

2.2. Учёт структуры семейства алгоритмов

Оценки, полученные в предыдущем разделе, мало использовали структуру семейства алгоритмов. В комбинаторном подходе лучшие результаты получались благодаря явному учёту метрической структуры семейства алгоритмов, выраженной свойствами расслоения и связности. В данном разделе для предложенных оценок с помощью техник комбинаторного подхода удастся учесть метод обучения.

Следующая теорема даёт оценку ожидаемой переобученности, учитывающую структуру графа расслоения–связности.

Теорема 2.2. *Пусть метод обучения μ – ПМЭР, $\ell = k = \frac{L}{2}$ и $\max\limits_{a\in A} m(a) \leqslant \frac{L}{2}$, тогда*

$$\begin{aligned}
&\mathcal{EOF}_\mu \leqslant \\
&\leqslant \inf_{\lambda>0}\frac{1}{\lambda}\left(\ln(\cosh(\lambda)) - \frac{\ln(C_{2\ell}^{\ell})}{\ell} + \frac{1}{\ell}\ln\left(\sum_{a\in A}\varphi(\ell, m(a), q(a), u(a), \lambda)\right)\right) \leqslant \\
&\leqslant \sqrt{\frac{2\ln|A|}{\ell}},
\end{aligned}$$

где $\varphi(\ell,m,q,u,\lambda) = \sum\limits_{j=0}^{m-q}\sum\limits_{i=0}^{j}(-1)^i C_j^i C_{2\ell-j-u-q}^{\ell-q+i-j} C_{m-q}^{j}(1+\tanh(\lambda))^q(\tanh(\lambda))^j$.

Доказательство. Доказательство до определённого шага полностью повторяет шаги теоремы 2.1. Отличие возникает на шаге, где максимум заменяется на сумму. При учёте метода обучения этот шаг несколько уточняется. Для вектора σ подвыборка X получается выбором из $\mathbb{X}$ всех объектов, позициям которых в σ соответствуют -1.

$$\begin{aligned}
&\lambda\mathbf{E}\left(\sum_{i=1}^{L}\sigma_i I(\mu X, x_i)\right) \leqslant \\
&\leqslant \ln\left(\mathbf{E}\left(\sum_{a\in A}[\mu X = a](\cosh(\lambda))^\ell \prod_{i=1}^{m(a)}(\hat{x}_i\tanh(\lambda)+1)\right)\right).
\end{aligned}$$

Рассмотрим отдельно выражение

$$\mathbf{E}\left(\sum_{a\in A}[\mu X = a]\prod_{i=1}^{m(a)}(\hat{x}_i\tanh(\lambda)+1)\right).$$

Благодаря лемме 1.1 оно мажорируется выражением

$$\mathbf{E}\left(\sum_{a\in A}[X_a \subset X][X'_a \subset \bar{X}]\prod_{i=1}^{m(a)}(\hat{x}_i\tanh(\lambda)+1)\right).$$

Последний переход очень важен. Как и в случае равномерных оценок, на этом шаге может накапливаться большая завышенность.

Каждый алгоритм a ошибается на всех объектах X'_a. Без ограничения общности можно считать, что X'_a соответствуют последние $q(a)$ объектов. Для них в условиях $X'_a \subset \bar{X}$ соответствующие $\hat{x}_i$ тождественно равны единице. Учитывая это и раскрывая скобки, имеем

$$\mathbf{E}\left(\sum_{a\in A}[X_a \subset X][X'_a \subset \bar{X}](\tanh(\lambda)+1)^{q(a)}\prod_{i=1}^{m(a)-q(a)}(\hat{x}_i\tanh(\lambda)+1)\right) =$$

$$= \mathbf{E}\Bigg(\sum_{a\in A}[X_a \subset X][X'_a \subset \bar{X}](\tanh(\lambda)+1)^{q(a)}\Big(1 + C^1_{m(a)-q(a)}\hat{x}_1\tanh(\lambda) + \ldots +$$

$$+ \hat{x}_1\ldots\hat{x}_{m(a)-q(a)}(\tanh(\lambda))^{m(a)-q(a)}\Big)\Bigg).$$

Теперь нужно проанализировать для каждого алгоритма a и j, такого что $j \leqslant m(a) - q(a)$ выражение

$$\mathbf{E}\left([X_a \subset X][X'_a \subset \bar{X}]\cdot\hat{x}_1\ldots\hat{x}_j\right).$$

Простые комбинаторные рассуждения, аналогичные приводимым ранее, приводят к выражению

$$\mathbf{E}\left([X_a \subset X][X'_a \subset \bar{X}]\cdot\hat{x}_1\ldots\hat{x}_j\right) = \frac{1}{C^\ell_{2\ell}}\sum_{i=0}^{j}(-1)^i C^i_j C^{\ell-q+i-j}_{2\ell-j-u-q}.$$

Таким образом,

$$\mathbf{E}\left(\sum_{a\in A}[X_a \subset X][X'_a \subset \bar{X}]\prod_{i=1}^{m(a)}(\hat{x}_i\tanh(\lambda)+1)\right) =$$

$$= \frac{1}{C^\ell_{2\ell}}\sum_{a\in A}(1+\tanh(\lambda))^q\sum_{j=0}^{m-q}\sum_{i=0}^{j}(-1)^i C^i_j C^{\ell-q+i-j}_{2\ell-j-u-q}C^j_{m-q}\tanh^j(\lambda).$$

Подставляя в ранее выписанные выражения полученную формулу, получаем первое неравенство теоремы. Теперь на основании того, что

$$\mathbf{E}\left(\sum_{a\in A}[X_a \subset X][X'_a \subset \bar{X}]\prod_{i=1}^{m(a)}(\hat{x}_i\tanh(\lambda)+1)\right) \leqslant \mathbf{E}\left(\sum_{a\in A}\prod_{i=1}^{m(a)}(\hat{x}_i\tanh(\lambda)+1)\right)$$

получаем и второе неравенство.

Нетрудно показать, что если обнулить в оценке все u и q, то получится та же оценка, что и в теореме 2.1.

3. Приложение

3.1. Оценка complete cross-validation

Примененная техника очень общая. Продемонстрируем её на примере функционала полного скользящего контроля:

$$\mathcal{CCV}_{\max} = \mathbf{E}\max_{a\in A}\nu(a,\bar{X}).$$

Теорема 3.1. *Пусть* $\ell = k = \frac{L}{2}$ *и* $\max\limits_{a\in A} m(a) \leqslant \frac{L}{2}$, *тогда*

$$\mathcal{CCV}_{\max} \leqslant$$

$$\leqslant \inf_{\lambda>0}\frac{1}{\lambda}\left(\ln(\cosh(\lambda)) + \frac{1}{\ell}\ln\left(\sum_{s=0}^{l}\Delta_s\ {}_2F_1\left(-\ell,-s,-2\ell,-\tanh(\lambda)\right)\right)\right),$$

где Δ_s *— число алгоритмов в* s*-м слое семейства алгоритмов.*

Доказательство. Все шаги доказательства полностью аналогичны теореме 2.1. Легко показать, что в данной теореме в выражении

$$\mathbf{E}(\hat{x}_1 \dots \hat{x}_j) = \frac{1}{C_{2\ell}^{\ell}} \sum_{i=0}^{j} (-1)^i C_{2\ell-j}^{\ell-j+i} C_j^i$$

нужно учитывать лишь слагаемое, соответствующее $i = 0$, что и заменит разность частот на одну частоту. После этого вместо ограниченной функции

$${}_2F_1\left(\frac{1-m}{2}, -\frac{m}{2}, \frac{1}{2} - \ell, (\tanh(\lambda))^2\right)$$

получится уже неограниченная:

$${}_2F_1\left(-\ell, -m, -2\ell, -\tanh(\lambda)\right).$$

Далее шаги доказательства опять повторяются.

3.2. Получение оценок вероятности переобучения

В теореме 2.1 доказано даже большее, а именно: оценена производящая функция моментов для $\max\limits_{a\in A}(\nu(a, \bar{X}) - \nu(a, X))$. Грубая из оценок этой теоремы даёт для $\lambda > 0$

$$\mathbf{E} \exp\left(\lambda \max_{a\in A}(\nu(a, \bar{X}) - \nu(a, X))\right) \leqslant |A| \exp\left(\frac{\lambda^2 \ell}{2}\right).$$

Отсюда с помощью неравенства Маркова можно получить хорошо известное неравенство. Для $t > 0$

$$\begin{aligned}
&\mathsf{P}\left(\max_{a\in A}(\nu(a, \bar{X}) - \nu(a, X)) > t\right) = \\
&= \mathsf{P}\left(\exp\left(\lambda \max_{a\in A}(\nu(a, \bar{X}) - \nu(a, X))\right) > \exp(\lambda t)\right) \leqslant |A| \exp\left(\frac{\lambda^2 \ell}{2} - \lambda t\right).
\end{aligned}$$

Оптимизируя по λ, получаем

$$\mathsf{P}\left(\max_{a\in A}(\nu(a, \bar{X}) - \nu(a, X)) > t\right) \leqslant |A| \exp\left(\frac{-t^2 \ell}{2}\right).$$

Аналогичные оценки получались напрямую и в [11], и в комбинаторном подходе. Нетрудно показать, что аналогичное неравенство верно и для величины $\max\limits_{a\in A}(\nu(a, X) - \nu(a, \bar{X}))$.

Можно обобщать и результат теоремы 2.2. Как и в случае теоремы 2.1, здесь мы оцениваем производящую функцию моментов величины $\nu(\mu X, \bar{X}) - \nu(\mu X, X)$. Поэтому, используя неравенство Маркова и оптимизируя по параметру λ, можно получить практически точный аналог оценки расслоения–связности 1.1. Получаемая оценка будет также линейна по вкладам алгоритмов.

3.3. Доказательство технической леммы

Докажем лемму 2.1.

Для действительного параметра α многочлены $\{C_n^{(\alpha)}(x)\}_{n=0}^{\infty}$, определенные на отрезке $[-1, 1]$, производящая функция которых равна

$$\frac{1}{(1 - 2xt + t^2)^{\alpha}} = \sum_{n=0}^{\infty} C_n^{(\alpha)}(x) t^n,$$

называются *ультрасферическими*. Нам понадобится следующая рекуррентная формула [4]:

$$C_0^{(\alpha)}(x) = 1,$$
$$C_1^{(\alpha)}(x) = 2\alpha x,$$
$$C_m^{(\alpha)}(x) = \frac{1}{m}(2x(m+\alpha-1)C_{m-1}^{(\alpha)}(x) - (m+2\alpha-2)C_{m-2}^{(\alpha)}(x)).$$

Доказательство. Легко видеть, что в данном случае гипергеометрическая функция является многочленом от z, причем для $z \in [0,1]$ согласно [4] имеет место равенство

$${}_2F_1\left(\frac{1-m}{2}, -\frac{m}{2}, \frac{1}{2}-\ell, z\right) = (-1)^m \frac{m!}{(\frac{1}{2}-\ell)_m}\left(\frac{z}{4}\right)^{\frac{m}{2}} C_m^{(\frac{1}{2}+\ell-m)}\left(\frac{1}{\sqrt{z}}\right),$$

где в правой части значение в нуле определено по непрерывности, а $(x)_m$ – нижний факториал числа x. Далее нас будут интересовать лишь точки экстремумов данной функции, поэтому мы будем работать только с $(z)^{\frac{m}{2}} C_m^{(\frac{1}{2}+\ell-m)}\left(\frac{1}{\sqrt{z}}\right)$, так как остальная часть выражения при данных соотношениях на параметры неотрицательна и не зависит от z. Обозначим $x^2 = z,\ \ x \in [-1,1]$.

Обозначим также $\hat{C}_m^{\alpha}(x) = x^m C_m^{(\alpha)}(\frac{1}{x})$. Тогда для данного многочлена легко получить рекуррентные соотношения, аналогичные тем, что имеются для ультрасферического:

$$\hat{C}_0^{(\alpha)}(x) = 1,$$
$$\hat{C}_1^{(\alpha)}(x) = 2\alpha,$$
$$\hat{C}_m^{(\alpha)}(x) = \frac{1}{m}(2(m+\alpha-1)\hat{C}_{m-1}^{(\alpha)}(x) - x^2(m+2\alpha-2)\hat{C}_{m-2}^{(\alpha)}(x)).$$

Докажем по индукции, что если $\alpha \geqslant \frac{1}{2}$, то на всем отрезке $[-1,1]$ имеет место неравенство $\hat{C}_m^{(\alpha)}(x) \geqslant \hat{C}_{m-1}^{(\alpha)}(x)$. База индукции очевидна. Для $m \geqslant 3$ рассмотрим разность

$$\hat{C}_m^{(\alpha)}(x) - \hat{C}_{m-1}^{(\alpha)}(x) =$$
$$= \frac{1}{m}(2(m+\alpha-1)\hat{C}_{m-1}^{(\alpha)}(x) - x^2(m+2\alpha-2)\hat{C}_{m-2}^{(\alpha)}(x) - m\hat{C}_{m-1}^{(\alpha)}(x)) =$$
$$= \frac{1}{m}((m+2\alpha-2)(\hat{C}_{m-1}^{(\alpha)} - x^2\hat{C}_{m-2}^{(\alpha)})).$$

Но $(m+2\alpha-2) \geqslant 0$, а по предположению индукции $\hat{C}_{m-1}^{(\alpha)} - \hat{C}_{m-2}^{(\alpha)} \geqslant 0$, но так как по индукции $\hat{C}_{m-2}^{(\alpha)} \geqslant 0$ и $x^2 \leqslant 1$, то $\hat{C}_{m-1}^{(\alpha)} - x^2\hat{C}_{m-2}^{(\alpha)} \geqslant 0$.

Рассмотрим теперь производные $\hat{C}_m^{\alpha}(x)$. Имеет место рекуррентное соотношение

$$\hat{C'}_0^{(\alpha)}(x) = 0,$$
$$\hat{C'}_1^{(\alpha)}(x) = 0,$$
$$\hat{C'}_m^{(\alpha)}(x) = \frac{1}{m}\Big(2(m+\alpha-1)\hat{C'}_{m-1}^{(\alpha)}(x) - x^2(m+2\alpha-2)\hat{C'}_{m-2}^{(\alpha)}(x) -$$
$$- 2x(m+2\alpha-2)\hat{C}_{m-2}^{(\alpha)}(x)\Big).$$

Из формы рекуррентных соотношений легко видеть, что $\hat{C}_m^{(\alpha)}(x)$ – чётная функция. Поэтому удобно производить анализ производных только на $[0,1]$.

Докажем по индукции, что если $\alpha \geqslant \frac{1}{2}$, то на всем отрезке $[0,1]$ имеет место неравенство $\hat{C'}_m^{(\alpha)}(x) \leqslant \hat{C'}_{m-1}^{(\alpha)}(x)$.

База индукции опять же очевидна. Аналогично предыдущему случаю

$$\hat{C'}_m^{(\alpha)}(x) - \hat{C'}_{m-1}^{(\alpha)}(x) = \\ = \frac{1}{m}((m+2\alpha-2)(\hat{C'}_{m-1}^{(\alpha)} - x^2\hat{C'}_{m-2}^{(\alpha)} - 2x\hat{C}_{m-2}^{(\alpha)}(x))).$$

Действительно, $(m+2\alpha-2) \geqslant 0$, $\hat{C'}_{m-1}^{(\alpha)} - x^2\hat{C'}_{m-2}^{(\alpha)} \leqslant 0$, так как $\hat{C'}_{m-1}^{(\alpha)} - \hat{C'}_{m-2}^{(\alpha)} \leqslant 0$, $\hat{C'}_{m-2}^{(\alpha)} \leqslant 0$ и $x^2 \leqslant 1$. Также по ранее доказанному $2x\hat{C}_{m-2}^{(\alpha)}(x) \geqslant 0$ на $[0,1]$.

Таким образом, с учётом чётности, $\hat{C}_m^{(\alpha)}(x)$ на $[-1,1]$ не превосходит своего значения в нуле. Это соответствует тому, что для $z \in [0,1]$

$${}_2F_1\left(\frac{1-m}{2}, -\frac{m}{2}, \frac{1}{2} - \ell, z\right)$$

ограничена своим значением в $z = 0$, то есть единицей.

Работа выполнена при поддержке РФФИ (проект 11-07-00480) и программы ОМН РАН «Алгебраические и комбинаторные методы математической кибернетики и информационные системы нового поколения».

Литература

1. *Вапник В.Н., Червоненкис А.Я.* О равномерной сходимости частот появления событий к их вероятностям. *ДАН СССР*. 1968. Т. 181, 4. С. 781–784.
2. *Воронцов К.В., Решетняк И.М.* Точные комбинаторные оценки обобщающей способности онлайнового обучения. Интеллектуализация обработки информации (ИОИ-2010): Докл. М.: МАКС Пресс, 2010. С. 24–27.
3. *Животовский Н.К., Воронцов К.В.* Критерий точности комбинаторных оценок вероятности переобучения // Сборник докладов 9-й международной конференции «Интеллектуализация обработки информации». М.: Торус Пресс, 2012. С. 25–28.
4. *Прудников А.П., Брычков Ю.А., Маричев О.И* Интегралы и ряды. Том 3. Специальные функции. Дополнительные главы. М.: Физматлит, 2003.
5. *Bartlett P.L., Bousquet O., Mendelson S.* Local Rademacher complexities. Annals of Statistics // 33(4):1497–1537, 2005.
6. *Boucheron S., Bousquet O., Lugosi G.* Theory of classification: A survey of some recent advances // ESAIM: Probability and Statistics. 2005. N 9. P. 323–375.
7. *Grätzer G.* General Lattice Theory. Basel, Switzerland: Birkhäuser, 1978. ISBN 978-0-12-295750-5.
8. *Devroye L., Lugosi G.* Combinatorial Methods in Density Estimation. Springer Series in Statistics. Springer-Verlag, 2001.
9. *Koltchinskii V.* Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems. École d'Été de Probabilités de Saint-Flour XXXVIII-2008. Lecture Notes in Mathematics. Springer-Verlag, 2011.
10. *Serfling R.J.* Probability inequalities for the sum in sampling without replacement // Ann. Statist. V. 2, N 1 (1974), 39-48.
11. *Vapnik V.* Statistical Learning Theory. New York: John Wiley and Sons, 1998.
12. *Vorontsov K.V.* Combinatorial probability and the tightness of generalization bounds // Pattern Recognition and Image Analysis. 2008. V. 18, N 2. P. 243–259.
13. *Vorontsov K.V.* Splitting and similarity phenomena in the sets of classifiers and their effect on the probability of overfitting // Pattern Recognition and Image Analysis. 2009. V. 19, N 3. P. 412–420.

14. *Vorontsov K. V.* Exact combinatorial bounds on the probability of overfitting for empirical risk minimization // *Pattern Recognition and Image Analysis.* 2010. V. 20, N 3. P. 269–285.

15. *Vorontsov K.V., Ivahnenko A.A.* Tight combinatorial generalization bounds for threshold conjunction rules // 4th International Conference on Pattern Recognition and Machine Intelligence (PReMI'11). June 27 – July 1, 2011. Lecture Notes in Computer Science. Springer-Verlag. 2011. P. 66–73.

References

1. *Vapnik V.N., Chervonenkis A.Y.* On the uniform convergence of relative frequencies of events to their probabilities. *Proceedings of the USSR Academy of Sciences.* 1968. T. 181, 4. P. 781–783.

2. *Vorontsov K., Reshetnyak I.* Exact combinatorial bounds of generalization ability of online learning. Intellectualization of information processing (IIP-2010). M.: MAKS Press, 2010. P. 24–27.

3. *Zhivotovskiy N., Vorontsov K.* The criterion of the exactness of combinatorial bounds of overfitting. Intellectualization of information processing (IIP-2012). M.: Torus Press, 2012. P. 25–28.

4. *Prydnikov A, Brychkov Y., Marichev O.* Integrals and Series: Special functions. Additional chapters. M.: Fizmatlit, 2003.

5. *Bartlett P.L., Bousquet O., Mendelson S.* Local Rademacher complexities. Annals of Statistics, 33(4):1497–1537, 2005.

6. *Boucheron S., Bousquet O., Lugosi G.* Theory of classification: A survey of some recent advances. ESAIM: Probability and Statistics. 2005. N 9. P. 323–375.

7. *Grätzer G.* General Lattice Theory. Basel, Switzerland: Birkhäuser, 1978. ISBN 978-0-12-295750-5.

8. *Devroye L., Lugosi G.* Combinatorial Methods in Density Estimation. Springer Series in Statistics. Springer-Verlag, 2001.

9. *Koltchinskii V.* Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems. École d'Été de Probabilités de Saint-Flour XXXVIII-2008. Lecture Notes in Mathematics. Springer-Verlag, 2011.

10. *Serfling R.J.* Probability inequalities for the sum in sampling without replacement. Ann. Statist. 1974. V. 2, N 1. 39–48.

11. *Vapnik V.* Statistical Learning Theory. New York: John Wiley and Sons, 1998.

12. *Vorontsov K.V.* Combinatorial probability and the tightness of generalization bounds. Pattern Recognition and Image Analysis. 2008. V. 18, N 2. P. 243–259.

13. *Vorontsov K.V.* Splitting and similarity phenomena in the sets of classifiers and their effect on the probability of overfitting. Pattern Recognition and Image Analysis. 2009. V. 19, N 3. P. 412–420.

14. *Vorontsov K.V.* Exact combinatorial bounds on the probability of overfitting for empirical risk minimization. *Pattern Recognition and Image Analysis.* 2010. V. 20, N 3. P. 269–285.

15. *Vorontsov K.V., Ivahnenko A.A.* Tight combinatorial generalization bounds for threshold conjunction rules. 4th International Conference on Pattern Recognition and Machine Intelligence (PReMI'11). June 27 – July 1, 2011. Lecture Notes in Computer Science. Springer-Verlag. 2011. P. 66–73.

Поступила в редакцию 24.02.2014.