Локализация оценок избыточного риска в комбинаторной теории переобучения*

Толстихин И. О.
iliya.tolstikhin@gmail.com
Вычислительный центр им. А. А. Дородницына РАН

Предлагается использовать в рамках комбинаторной теории переобучения [7, 1] процедуру локализации оценок обобщающей способности, разработанную в теории статистического обучения за последнее десятилетие [5]. В отличие от классических подходов, основанных на равномерных по всему классу функций оценках, локализация последовательно сужает подмножество класса, по которому берется супремум.

Localized excess risk bounds in combinatorial theory of overfitting*

Tolstikhin I. O.
Dorodnicyn Computing Centre of RAS, Moscow, Russia

We study the localization of generalization bounds within the combinatorial theory of overfitting [7, 1]. The localization procedure was developed in statistical learning theory in last decade [5]. Unlike classical results, based on the uniform bounds over the whole function class, the localization sequentially confines the function subclasses the supremum is taken over.

За последнее десятилетие в теории статистического обучения (далее SLT) произошел качественный скачок, главным образом связанный с доказанным в середине 90-х концентрационным неравенством Талаграна для эмпирических процессов (подробный обзор неравенств концентрации меры приводится в [3]). Это привело к развитию нового подхода к получению оценок обобщающей способности, связанного с работами P. Bartlett, O. Bousquet, В. Колчинского, P. Massart и Д. Панченко. Классические подходы, начиная с теории Вапника–Червоненкиса, для оценки качества выбранного из семейства $\mathcal{F}$ на основе обучающей выборки $\{X_1, \ldots, X_n\}$ классификатора использовали равномерные по классу $\mathcal{F}$ оценки вида $\sup_{f\in\mathcal{F}} |Pf - P_n f|$, где Pf — математическое ожидание, а $P_n f$ — среднее значение функции f на обучающей выборке. В отличие от этого, новый подход последовательно уточняет подмножество класса $\mathcal{F}$, по которому берется супремум, за счет чего достигает существенных улучшений оценок. Эта процедура в литературе известна как *локализация* [5]. Ключевым ингредиентом локализации как раз и является концентрационное неравенство Талаграна, широко используемое последнее время в теории статистического обучения.

К сожалению, неравенство Талаграна рассматривает последовательности независимых и одинаково распределенных случайных величин, что не позволяет применять его в рамках комбинаторного подхода. В работе исследуется возможность использования процедуры локализации в рамках комбинаторного подхода к переобучению [7, 1]. В частности, вместо неравенств Талаграна предлагается использовать оценки концентрации [2], позволяющие повторить основные шаги локализации.

Работа поддержана РФФИ (проект № 11-07-00480), программой ОМН РАН «Алгебраические и комбинаторные методы математической кибернетики и информационные системы нового поколения».

Теория переобучения: SLT и комбинаторный подход

Классическая постановка основной задачи SLT заключается в следующем. Дано множество объектов $\mathbb{X}$ и множество ответов $\mathbb{Y}$. Предполагается, что на декартовом произведении $\mathbb{X} \times \mathbb{Y}$ задана неизвестная вероятностная мера P. Дана обучающая выборка $\{(X_1, Y_1), \ldots, (X_\ell, Y_\ell)\}$ — последовательность *независимых и одинаково распределенных* случайных величин из распределения P. Требуется решить задачу $\mathsf{E} r\bigl(g(X), Y\bigr) \to \min\limits_{g\in\mathcal{G}}$ для некоторой фиксированной функции потерь $r\colon \mathbb{Y}^2 \to \mathbb{R}^+$ и класса $\mathcal{G}$ отображений $g\colon \mathbb{X} \to \mathbb{Y}$, пользуясь обучающей выборкой.

В комбинаторном подходе, рассматриваемом в данной работе, вводятся следующие ограничения. Во-первых, предполагается, что множество объектов (генеральная выборка) конечно и состоит из L элементов: $\mathbb{X} = \{X_1, \ldots, X_L\}$. Во-вторых, рассматривается случай, когда классификация объектов генеральной выборки фиксирована: объект $X_i \in \mathbb{X}$ принадлежит классу $Y_i \in \mathbb{Y}$, $i = 1, \ldots, L$, где $\mathbb{Y}$ — множество меток классов. Это одно из возможных требований, к которым естественным образом ведет принятое в комбинаторном подходе отождествление классификаторов $a\colon \mathbb{X} \to \mathbb{Y}$ с *фиксированными* векторами их потерь на объектах обучающей выборки. Отметим, что это требование можно снять, например, разрешив генеральной выборке содержать повторяющиеся объекты с разными ответами на них.

Эти ограничения соответствуют частному случаю в SLT, когда условное распределение $\mathsf{P}(Y \mid X_j)$ для фиксированного элемента X_j из конечного множества объектов целиком концентрируется на значении Y_j:

$$\mathsf{P}(Y \mid X_j) = \begin{cases} 1, & Y = Y_j; \\ 0, & Y \neq Y_j. \end{cases}$$

Наиболее существенное отличие комбинаторного подхода заключается в следующем третьем предположении: всевозможные разбиения генеральной выборки $\mathbb{X} = X^\ell \cup X^k$ на не пересекающиеся между собой обучающую X^ℓ длиной ℓ и контрольную X^k длиной k равновероятны.

Теперь фиксируем в общем случае бесконечное множество A классификаторов $a\colon \mathbb{X} \to \mathbb{Y}$, которые мы будем отождествлять с векторами их *потерь* на объектах генеральной выборки $\mathbb{X}$:

$$a \equiv (a_i)_{i=1}^L = \bigl(r\bigl(a(X_i), Y_i\bigr)\bigr)_{i=1}^L, \quad i = 1, \dots, L,$$

где $r\colon \mathbb{Y}^2 \to \mathbb{R}^+$ — неотрицательная функция потерь. Будем рассматривать только *ограниченные* функции потерь, что позволяет без ограничения общности полагать $r\colon \mathbb{Y}^2 \to [0, 1]$. В SLT класс

$$\mathcal{F} = \bigl\{f(X, Y) = r\bigl(g(X), Y\bigr), g \in \mathcal{G}\bigr\},$$

соответствующий множеству классификаторов $\mathcal{G}$, и фиксированной функции потерь r принято называть *классом потерь*.

До сих пор в рамках комбинаторного подхода по ряду технических причин рассматривалась только *бинарная* функция потерь

$$r(y, y^*) = [y \neq y^*],$$

штрафующая единицей отнесение объекта класса y^* к другому классу y. Результаты, изложенные дальше, не требуют этого ограничения. Наша цель — пользуясь обучающей выборкой, выбрать алгоритм a из класса A, который бы допускал как можно меньше ошибок на генеральной выборке $\mathbb{X}$. Таким образом, нас интересует задача минимизации *риска*

$$P_L a \equiv \frac{1}{L} \sum_{i=1}^{L} r\bigl(a(X_i), Y_i\bigr) = \frac{1}{L} \sum_{i=1}^{L} a_i \to \min_{a \in A}. \qquad (1)$$

Сформулированная постановка задачи чрезвычайно похожа на классическую постановку SLT. Единственное но важное отличие — в способе получения обучающей выборки. В SLT, в случае конечного множества объектов $\mathbb{X}$, полагается, что обучающая выборка $\{X_1, \dots, X_\ell\}$ — *простая* из распределения P, то есть объекты вытягиваются независимо и случайно *с возвращениями* из множества объектов $\mathbb{X}$. Слабая вероятностная аксиоматика, принятая в комбинаторном подходе, утверждает, что объекты обучающей выборки вытягиваются случайно и равновероятно *без возвращений* из конечного множества объектов. Очевидно, что если в постановке SLT вместо выборки с возвращениями использовать выборку без возвращений, а также положить а) множество объектов конечным; б) маргинальное распределение $P(X)$ равномерным и в) условное распределение $P(Y \mid X_j)$ сконцентрировать на правильном ответе Y_j, то мы придем к постановке комбинаторного подхода к переобучению. Этот взгляд на отличие двух подходов к оцениванию обобщающей способности играет решающую роль в изложенных дальше результатах.

В случае классической постановки SLT риск классификатора a определяется стандартным образом — как математическое ожидание потерь, связанных с использованием этого классификатора:

$$P\,a \equiv \mathsf{E}\, r\bigl(a(X), Y\bigr) = \int_{\mathbb{X}\times\mathbb{Y}} r\bigl(a(X), Y\bigr) d\,\mathsf{P}\,.$$

Замечание 1. Заметим, что если принять ограничения а)–в) из предыдущего абзаца, то в постановке SLT для любого классификатора $Pa = P_L a$.

Поскольку нам доступна лишь обучающая выборка, заменим задачу (1) следующей задачей *минимизации эмпирического риска*:

$$P_\ell\, a \equiv \frac{1}{\ell} \sum_{i \in I_\ell} a_i \to \min_{a \in A}, \qquad (2)$$

при этом здесь и далее

$$I_\ell = \bigl\{i \in \{1, \dots, L\}\colon X_i \in X^\ell\bigr\}, \;\; I_k = \{1, \dots, L\} \setminus I_\ell.$$

Обозначим решение задачи минимизации эмпирического риска $a^\ell = \arg\min_{a \in A} P_\ell\, a$. Закон больших чисел дает основания рассматривать минимизатор эмпирического риска a^ℓ в качестве хорошего кандидата для решения задачи (1). Следующую величину, отражающую, насколько точно функция a решает задачу (1), будем называть *избыточным риском* функции a:

$$\mathcal{E}_L(a) = P_L a - \min_{a \in A} P_L a.$$

В классической постановке SLT

$$\mathcal{E}(a) = Pa - \min_{a \in A} Pa.$$

Всюду далее мы будем рассматривать величину $\mathcal{E}_L(a^\ell)$ (или $\mathcal{E}(a^\ell)$ в случае SLT), характеризующую качество приближения решения задачи (2) к решению задачи (1). Поскольку $\mathcal{E}_L(a^\ell)$ — случайная величина (зависит от обучающей выборки), нас будут интересовать вероятностные неравенства вида

$$\mathbb{P}\{\mathcal{E}_L(a^\ell) \geqslant t\} \leqslant \eta(A, \ell, t),$$

где $\mathbb{P}$ — равномерное вероятностное распределение на множестве всевозможных разбиений генеральной выборки на обучающую и контрольную: $\mathbb{X} = X^\ell \cup X^k$, а η — некая неотрицательная убывающая функция. В частности нас будут интересовать случаи, когда η экспоненциально убывает с ростом t.

Локализация оценок

В этом разделе мы будем существенно опираться на лекциии В. И. Колчинского [5].

Назовем δ-*минимальными множествами* следующие подмножества класса A:

$$A_L(\delta) = \{a \in A \colon \mathcal{E}_L(a) \leqslant \delta\};$$
$$A(\delta) = \{a \in A \colon \mathcal{E}(a) \leqslant \delta\}.$$

В случае бинарной функции потерь r, $A_L(\delta)$ — просто фиксированное число нижних слоев семейства алгоритмов A [7]. Выпишем элементарную цепочку неравенств, предположив, что минимум риска на классе A достигается в $\bar{a}$:

$$\begin{aligned}\delta^\ell = \mathcal{E}_L(a^\ell) = P_L a^\ell - P_L \bar{a} = \\ = P_\ell\, a^\ell - P_\ell\, \bar{a} + (P_L - P_\ell)(a^\ell - \bar{a}) \leqslant \\ \leqslant (P_L - P_\ell)(a^\ell - \bar{a}).\end{aligned}$$

Последнее неравенство — следствие определения функции a^ℓ. Отсюда получаем:

$$\delta^\ell \leqslant \sup_{a,b \in A_L(\delta^\ell)} |(P_L - P_\ell)(a - b)|, \qquad (3)$$

поскольку $a^\ell, \bar{a} \in A_L(\delta^\ell)$. Все те же шаги можно проделать и в рамках SLT для $\mathcal{E}(a^\ell)$ и $A(\delta^\ell)$.

Основная идея локализации в рамках SLT заключается в построении неслучайной оценки $U_\ell(\delta)$, для которой с большой вероятностью относительно случайных реализаций обучающей выборки равномерно по параметру δ справедливо неравенство

$$U_\ell(\delta) \geqslant \sup_{a,b \in A(\delta)} |(P_L - P_\ell)(a - b)|. \qquad (4)$$

Тогда с той же вероятностью избыточный риск $\mathcal{E}(a^\ell)$ оценивается сверху максимальным решением неравенства $\delta \leqslant U_\ell(\delta)$.

Один из общепринятых способов построения подобных оценок $U_\ell(\delta)$ в классической постановке SLT основан на использовании неравенства Талаграна. В частности, в [5] используется следующая его версия, предложенная в [4]. Далее с помощью P^ℓ и E^ℓ будем обозначать продакт-меру на обучающих выборках (ℓ-ю декартову степень распределения P) и математическое ожидание относительно нее.

В рамках SLT введем следующие обозначения:

$$\varphi_\ell(\delta) = \mathsf{E}^\ell \sup_{a,b \in A(\delta)} |(P - P_\ell)(a - b)|;$$
$$D^2(\delta) = \sup_{a,b \in A(\delta)} P(a - b)^2;$$
$$U_\ell(\delta, t) = K\left(\varphi_\ell(\delta) + D(\delta)\sqrt{\frac{t}{\ell}} + \frac{t}{\ell}\right), \qquad (5)$$

где K — некоторая константа. Фиксируем произвольную $\delta > 0$. Тогда, согласно неравенству Талаграна в версии [4], для некоторой универсальной константы $K > 0$ и для любого $t > 0$ выполнено:

$$\mathsf{P}^\ell\left\{\sup_{a,b \in A(\delta)} |(P - P_\ell)(a - b)| \geqslant U_\ell(\delta, t)\right\} \leqslant 1 - e^{-t}.$$

С помощью функции $U_\ell(\delta, t)$ можно относительно легко конструировать $U_\ell(\delta)$, с большой вероятностью удовлетворяющую неравенству (4). Решая затем неравенство $\delta \leqslant U_\ell(\delta)$, мы получим оценку $\bar{\delta}_n(A)$, с большой вероятностью ограничивающую $\mathcal{E}(a^\ell)$ сверху.

Описанный подход позволяет получать оценки асимптотически точного порядка для широкого класса задач машинного обучения (например, [5, глава 5]). Это становится возможным благодаря учету диаметра $D(\delta)$ δ-минимального множества класса функций: если выполнено условие $D(\delta) \to 0$ при $\delta \to 0$, то в ряде случаев функция $U_\ell(\delta, t)$ убывает как $o(1/\sqrt{\ell})$ при одновременном стремлении $\ell \to \infty$ и $\delta \to 0$. Например, это справедливо для так называемых P-Донскеровских классов функций ([5, Теорема 4.5]).

В комбинаторном подходе условие стремления $D(\delta)$ к нулю при $\delta \to 0$ означает, что в нижних слоях рассматриваемого семейства содержатся существенно похожие алгоритмы. Таким образом мы снова наблюдаем необходимость одновременного учета эффектов *связности* и *расслоения* семейства алгоритмов, которая обоснована экспериментами [6] и комбинаторными оценками [7, 1]. Сужая дельта-минимальное подмножество, по которому берется супремум, мы учитываем расслоение. При этом, если мы не будем учитывать диаметр этого множества, то мы не получим оценки лучшие, чем порядка $O(1/\sqrt{\ell})$.

Замечание 2. Обобщения неравенств типа Хевдинга для суммы независимых центрированных случайных величин, использовавшихся в ранних подходах (например, неравенство МакДиармида), учитывают лишь ограниченность случайных величин и никак не учитывают их *дисперсий*. В то же время, чтобы учитывать диаметр множества функций, как это было сделано выше ($D(\delta)$ — это $L_2(\mathsf{P})$-диаметр δ-минимального множества), необходимо использовать оценки типа Бернштейна.

Неравенство Талаграна как раз и является обобщением неравенства Бернштейна на равномерный по классу функций случай.

Замечание 3. Оценки, получаемые описанным способом, *не вычислимы* — они зависят от параметров неизвестного распределения (например, в определении $\varphi_\ell(\delta)$). Оказывается, существует метод, позволяющий в большинстве случаев получить *вычислимые по данным* варианты этих оценок. Этот метод главным образом основан на использовании *радемахеровского процесса* и одной из версий *неравенства симметризации*, связывающей математическое ожидание его супремума с математическим ожиданием супремума эмпирического процесса [5, раздел 4.2].

К сожалению, неравенство Талаграна применимо только к простым выборкам с возвращением, и не может быть непосредственно использовано в постановке комбинаторного подхода. В следующем разделе показано, что в рамках слабой вероятностной аксиоматики существует адекватная замена неравенству Талаграна.

Неравенства концентрации меры в слабой вероятностной аксиоматике

Приведем без доказательства основной результат данной работы — экспоненциальное концентрационное неравенство, рассматривающее выборки без возвращений, и учитывающее одну из характеристик *разброса* случайных величин. Неравенство получено с помощью более общих неравенств, приведенных в работе [2].

Рассмотрим в рамках комбинаторного подхода *счетное* множество классификаторов A и следующую случайную величину:

$$Z = \sup_{a\in A} |(P_L - P_\ell)\, a|.$$

Введем обозначение

$$\sigma_A^2 = \sup_{\mathbb{X}=X^\ell\cup X^k} \Bigl(\frac{k}{L\ell}\sup_{a\in A}\sum_{i\in I_\ell}(a_i)^2 + \frac{1}{L}\sup_{a\in A}\sum_{j\in I_k}(a_j)^2\Bigr).$$

Обозначим через $\mathbb{P}$ равномерное вероятностное распределение на множестве всех разбиений генеральной выборки $\mathbb{X} = X^\ell \cup X^k$, а $\mathbb{E}$ — соответствующее ему математическое ожидание.

Теорема 1. *Для любых $h > 0$ справедливо:*

$$\mathbb{P}\{Z - \mathbb{E}Z \geqslant h\} \leqslant \exp\Bigl(-\frac{\ell h^2}{16\sigma_A^2}\Bigr); \qquad (6)$$

$$\mathbb{P}\{|Z - \mathbb{E}Z| \geqslant h\} \leqslant 2\exp\Bigl(-\frac{\ell h^2}{16\sigma_A^2}\Bigr).$$

В частности, после обращения вероятности, из (6) следует, что для любого $t > 0$ с вероятностью не меньше $1 - e^{-t}$ справедливо

$$Z \leqslant \mathbb{E}Z + 4\sigma_A\sqrt{\frac{t}{\ell}}\,.$$

Замечание 4. Очевидно, что $\sigma_A^2 < 2$, поскольку мы рассматриваем ограниченные единицей функции потерь. Эта тривиальная оценка вместе с теоремой 1 дает неравенство типа МакДиармида для случайной величины Z.

Заключение

В данной работе была преодолена первая преграда на пути применения процедуры локализации в рамках комбинаторного подхода — невозможность использования неравенства Талаграна из-за принципиальных различий в постановках двух подходов к теории переобучения.

В дальнейшем предполагается в рамках комбинаторной теории последовательно совершить оставшиеся шаги (вкратце описанные в данной работе и подробно изложенные в [5]), ведущие к получению зависящих от распределения оценок избыточного риска порядка $o(1/\sqrt{\ell})$, а также их вычислимых по данным аналогам. Помимо улучшения нынешних оценок, это позволит сравнить комбинаторные оценки вероятности переобучения [1] с последними результатами, полученными в рамках классической постановки SLT.

Литература

[1] *Воронцов К. В.* Комбинаторная теория переобучения: результаты, приложения и открытые проблемы. // Математические методы распознавания образов: 15-ая Всеросс. конф.: Докл. — М.: МАКС Пресс, 2011. — С. 40–43.

[2] *Bobkov S. G.* Concentration of normalized sums and a central limit theorem for noncorrelated random variables // The Annals of Probability. — 2004. — V. 32, N. 4. — Pp. 2884–2907.

[3] *Boucheron S., Lugosi G., and Bousquet O.* Concentration inequalities // Lecture Notes in Computer Science. — 2004. — V. 3176. — Pp. 208–240.

[4] *Bousquet O.* A Bennett Concentration Inequality and Its Application to Suprema of Empirical Processes // CR. Acad. Sci. Paris, Ser. I. — 2002. — V. 334. — Pp. 495–500.

[5] *Koltchinskii V.* Oracle inequalities in empirical risk minimization and sparse recovery problems. — École d'Été de Probabilités de Saint-Flour XXXVIII-2008, 2011.

[6] *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. — Vol. 19, No. 3. — Pp. 412–420.

[7] *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.