Вероятность переобучения плотных и разреженных семейств алгоритмов*

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

Показано, что для оценивания вероятности переобучения семейств алгоритмов, обладающих свойствами расслоения и связности, достаточно брать их подмножества, состоящие из небольшого числа существенно различных алгоритмов с малым числом ошибок.

The probability of overfitting for the compact and sparse sets of predictors*

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

We show that if a set of classifiers possesses the properties of splitting and similarity, then its probability of overfitting can be approximated by a small subset of essentially different classifiers with low error rate.

Оценивание вероятности переобучения — одна из основных задач теории статистического обучения [2]. Чем точнее её решение, тем больше возможностей открывается для создания алгоритмов машинного обучения с гарантированно высокой способностью к обобщению эмпирических данных.

Большинство известных верхних оценок вероятности переобучения представляют собой произведение двух сомножителей. Первый описывает сложность семейства алгоритмов. Например, в VC-оценках [1] это *коэффициент разнообразия* (shattering coefficient), равный числу алгоритмов семейства, различимых на заданной конечной выборке длины L. Второй сомножитель — это вероятность большого уклонения частот ошибок в двух непересекающихся подвыборках длины $L/2$ для одного отдельно взятого алгоритма. Обычно она оценивается сверху неравенствами Хёффдинга, Чернова, или другими оценками скорости сходимости в законе больших чисел или его обобщениях [2].

В комбинаторном подходе [5, 6] рассматривается *генеральная выборка* длины L и предполагается равновероятность всех её разбиений на две подвыборки — наблюдаемую *обучающую* длины ℓ и скрытую *контрольную* длины $k = L - \ell$. В этом случае второй сомножитель определяется точно, через гипергеометрическое распределение.

Первый сомножитель сильно завышен в большинстве известных оценок. Степень его завышенности измерялась в экспериментах на реальных задачах классификации [4]. Для этого вероятность переобучения оценивалась методом скользящего контроля. Поделив её на второй сомножитель, который известен точно, можно получить *эффективный локальный коэффициент разнообразия* (ЭЛКР). Он показывает, каким должен был бы быть первый сомножитель, чтобы оценка не была завышенной. Он также интерпретируется как число алгоритмов, «эффективно используемых» при решении данной конкретной задачи, то есть имеющих высокую вероятность быть выбранными в результате обучения. В экспериментах [4] ЭЛКР принимал значения порядка 10^0–10^1 при коэффициентах разнообразия порядка 10^6–10^9. Ни одна из известных теорий не могла объяснить столь низких значений ЭЛКР. В данной работе делается попытка дать такое объяснение.

В последующих работах [5, 6] было показано, что завышенность VC-оценок возможно устранить только путём одновременного учёта двух эффектов, наблюдаемых при решении реальных задач.

Эффект расслоения связан с тем, что при фиксации задачи семейство алгоритмов расслаивается по числу ошибок на генеральной выборке. Чем выше слой, тем меньше вероятность, что в результате обучения будет выбран алгоритм из этого слоя. Как правило, в нижних слоях находится ничтожно малая доля алгоритмов, однако именно они «эффективно используются» в данной задаче.

Эффект связности возникает в семействах алгоритмов, непрерывных по параметрам. В этом случае для каждого алгоритма в семействе найдётся некоторое число алгоритмов, различимых с ним только на одном каком-то объекте генеральной выборки. Будем называть такие алгоритмы связанными. Похожие (в частности, связанные) алгоритмы вносят меньший вклад в вероятность переобучения, чем непохожие. Поэтому увеличение связности способствует снижению переобучения.

Понимание этих эффектов позволяет выдвинуть следующую гипотезу: *вероятность переобучения расслоенного и связного множества алго-*

Работа поддержана РФФИ (проект №08-07-00422) и программой ОМН РАН «Алгебраические и комбинаторные методы математической кибернетики и информационные системы нового поколения».

ритмов может быть аппроксимирована вероятностью переобучения его подмножества, состоящего из существенно различных алгоритмов нижних слоёв. При этом малые значения ЭЛКР, наблюдавшиеся в экспериментах, позволяют надеяться, что для аппроксимации хватит нескольких десятков алгоритмов. В данной работе исследуется возможность аппроксимации «плотных» семейств алгоритмов их «разреженными» подсемействами.

Понятие вероятности переобучения

Задано множество объектов $\mathbb{X} = \{x_1, \ldots, x_L\}$, множество алгоритмов $A = \{a_1, \ldots a_D\}$ и бинарная функция $I\colon \mathbb{X} \times A \to \{0, 1\}$, называемая *индикатором ошибки*, такая, что $I(a, x) = 1$ тогда и только тогда, когда алгоритм a ошибается на объекте x.

Каждому алгоритму соответствует бинарный *вектор ошибок* $\big(I(a, x_i)\big)_{i=1}^{L}$ длины L. *Матрицей ошибок* называется $D \times L$-матрица, строками которой являются векторы ошибок алгоритмов из A. В дальнейшем будем считать, что строки матрицы ошибок попарно различны, и отождествлять алгоритмы с их векторами ошибок.

Числом ошибок алгоритма a на выборке $X \subseteq \mathbb{X}$ называется величина $n(a, X) = \sum\limits_{x \in X} I(a, x)$.

Частотой ошибок (или *эмпирическим риском*) алгоритма a на выборке $X \subseteq \mathbb{X}$ называется величина $\nu(a, X) = \frac{1}{|X|} n(a, X)$.

Обозначим через $[\mathbb{X}]^\ell$ множество всех C_L^ℓ подмножеств $X \subseteq \mathbb{X}$ мощности ℓ.

Методом обучения называется отображение $\mu\colon [\mathbb{X}]^\ell \to A$, которое произвольной выборке X длины ℓ ставит в соответствие некоторый алгоритм.

Метод обучения μ называется методом *минимизации эмпирического риска* (МЭР), если

$$\mu X \in A(X) \equiv \operatorname{Arg}\min_{a \in A} n(a, X).$$

В случае $|A(X)| > 1$ выбор алгоритма из $A(X)$ неоднозначен. Метод МЭР называется *пессимистическим* (ПМЭР), если

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

Метод МЭР называется *рандомизированным* (РМЭР), если он выбирает произвольный алгоритм из $A(X)$ случайно и равновероятно [3].

Величина $\delta_\mu(X) = \nu\left(\mu X, \mathbb{X}\backslash X\right) - \nu\left(\mu X, X\right)$ называется *переобученностью* метода μ на выборке X. Если $\delta_\mu(X) \geqslant \varepsilon$, где ε — фиксированный *порог переобучения*, то говорят, что *метод μ переобучен на выборке X*.

Следуя комбинаторному подходу [4, 5, 6], будем полагать, что при фиксированном ℓ все C_L^ℓ разбиений *генеральной выборки* $\mathbb{X}$ на наблюдаемую обучающую выборку X длины ℓ и скрытую контрольную $\bar{X} = \mathbb{X}\backslash X$ длины $k = L - \ell$ равновероятны.

Нашей основной задачей будет вычисление вероятности переобучения для ПМЭР:

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

или для РМЭР:

$$Q_\varepsilon(A) = \sum_{a \in A} Q_{\varepsilon,a}(A),$$

где величина $Q_{\varepsilon,a}(A)$ называется *вкладом алгоритма a* в вероятность переобучения:

$$Q_{\varepsilon,a}(A) = \frac{1}{C_L^\ell} \sum_{X \in [\mathbb{X}]^\ell} \frac{[a \in A(X)]}{|A(X)|} \big[\delta(a, X)) \geqslant \varepsilon\big].$$

Заметим, что ПМЭР не реализуем на практике, т. к. он «подглядывает» в скрытую выборку на этапе обучения. Теоретически он интересен тем, что даёт верхние оценки вероятности переобучения. Введение РМЭР упрощает вывод точных оценок вероятности переобучения для семейств алгоритмов, обладающих определённой симметрией [3].

Шар алгоритмов и его подмножества

Пусть $d(a, a')$ — расстояние Хэмминга между алгоритмами a и a' как L-мерными бинарными векторами. Введём множества алгоритмов:

$B_r(a_0) = \big\{a \in \{0,1\}^L\colon d(a_0, a) \leqslant r\big\}$ — *шар алгоритмов* с центром в $a_0 \in \{0,1\}^L$ и радиусом r;

$S_r(a_0) = \big\{a \in B_r(a_0)\colon d(a_0, a) = r\big\}$ — *сфера алгоритмов* с центром a_0 и радиусом r;

$A_m = \big\{a \in \{0,1\}^L\colon n(a, \mathbb{X}) = m\big\}$ — *m-й слой*;

$B_r^m(a_0) = B_r(a_0) \cap A_m$ — *m-й слой шара* $B_r(a_0)$;

$S_r^m(a_0) = S_r(a_0) \cap A_m$ — *m-й слой сферы* $S_r(a_0)$.

Шар алгоритмов интересен тем, что это «максимально плотное» множество алгоритмов, обладающее наибольшей связностью среди всех множеств такой же мощности. Следовательно, оценки вероятности переобучения, получаемые для шара или его слоёв, являются оценками «лучшего случая» и могут служить ориентировочными нижними оценками и для других множеств алгоритмов.

Далее приводятся точные оценки вероятности переобучения для шара, его слоёв и некоторых их разреженных подмножеств. Будет показано, что в m-м слое сферы можно взять совсем небольшое подмножество алгоритмов, для которого вероятность переобучения лишь немного отличается от вероятности переобучения соответствующего слоя шара, содержащего на несколько порядков большее число алгоритмов.

Обозначим через $H_L^{l,m}(z)$ гипергеометрическую функцию распределения:

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

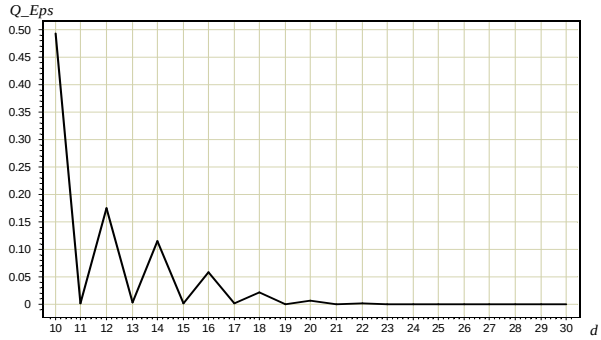

Рис. 1. Вклады слоёв шара в вероятность переобучения, при $l = k = 100,\ m = 20,\ r_0 = 10,\ \varepsilon = 0{,}05$.

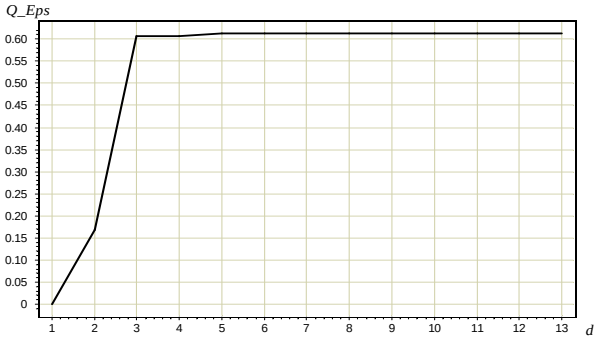

Рис. 2. Зависимость Q_ε от числа d нижних слоев шара, при $l = k = 100,\ m = 10,\ r_0 = 6,\ \varepsilon = 0,05$.

Теорема 1. *Пусть μ — РМЭР, $A = B_{r_0}(a_0)$ — шар алгоритмов, $n(a_0, \mathbb{X}) = m$ и $r_0 \leqslant \min(m, L-m)$. Тогда для любого $\varepsilon \in [0,1]$:*

$$Q_\varepsilon = \sum_{i=0}^{r_0} h_L^{\ell,m}(i)\, \frac{\sum\limits_{r=0}^{r_0}\sum\limits_{n=0}^{r} S(n,r,i)\big[m+r-2n \geqslant \varepsilon k\big]}{\sum\limits_{r=0}^{r_0}\sum\limits_{n=0}^{r} S(n,r,i)} + \sum_{i=r_0+1}^{\lfloor s_d(\varepsilon)\rfloor} h_L^{\ell,m}(i),$$

$$S(n,r,i) = C_{m-i}^{n-i} C_{k-m+i}^{r-n},\quad s_d(\varepsilon) = \tfrac{\ell}{L}(m-\varepsilon k) + \tfrac{r_0 k}{L}.$$

На рис. 1 показаны значения вкладов слоев шара в вероятность его переобучения. Видно, что наибольший вклад приходится на первые несколько слоев. Поэтому отдельный интерес представляет оценка вероятности переобучения для нескольких нижних слоев шара алгоритмов.

Теорема 2. *Пусть μ — РМЭР, $A = \big\{a \in B_{r_0}(a_0) : m - r_0 \leqslant n(a,\mathbb{X}) \leqslant m - r_0 + d - 1\big\}$ — d нижних слоёв шара, $n(a_0,\mathbb{X}) = m$ и $r_0 \leqslant \min(m, L-m)$. Тогда для любого $\varepsilon \in [0,1]$:*

$$Q_\varepsilon = \sum_{i=0}^{r_0} h_L^{\ell,m}(i)\, \frac{\sum\limits_{r=0}^{r_0}\sum\limits_{n=0}^{r} S'(n,r,i)\,\big[m+r-2n \geqslant \varepsilon k\big]}{\sum\limits_{r=0}^{r_0}\sum\limits_{n=0}^{r} S'(n,r,i)} + \sum_{i=r_0+1}^{\lfloor s_d(\varepsilon)\rfloor} h_L^{\ell,m}(i),$$

где $S'(n,r,i) = C_{m-i}^{n-i} C_{k-m+i}^{r-n}\,[r + r_0 + 1 \leqslant 2n + d]$.

Очевидно, что случай $d = 2r_0 + 1$ соответствует всему шару алгоритмов $B_{r_0}(a_0)$. На рис. 2 представлена зависимость вероятности переобучения нижних слоев шара от их числа. Видно, что вероятность переобучения шара достигается уже на 3–4 его нижних слоях.

Далее мы покажем возможность аппроксимации вероятности переобучения «плотного» семейства $B_{r_0}^m(a_0)$ его разреженным подмножеством.

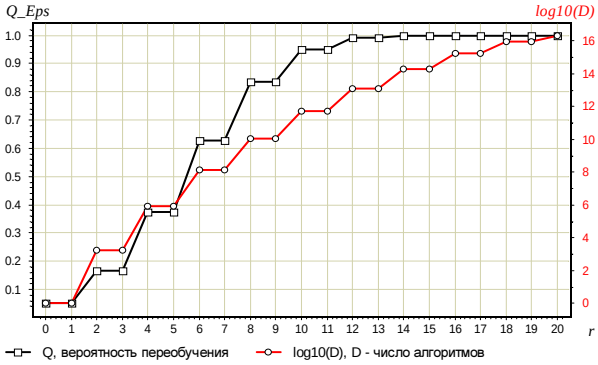

Рис. 3. Зависимость вероятности переобучения Q_ε и $\log_{10}|A|$ от радиуса шара r для m-го слоя шара, при $l = k = 100,\ m = 10,\ \varepsilon = 0{,}05$.

Теорема 3. *Пусть μ — МЭР, $A = B_{r_0}^m(a_0)$ — m-й слой шара, $m \leqslant k$ и $n(a_0,\mathbb{X}) = m$. Тогда для любого $\varepsilon \in [0,1]$:*

$$Q_\varepsilon = \begin{cases} H_L^{\ell,m}\big(\frac{\ell}{L}(m-\varepsilon k) + \lfloor r_0/2\rfloor\big), & m \geqslant \varepsilon k, \\ 0, & m < \varepsilon k. \end{cases}$$

На рис. 3 представлена зависимость вероятности переобучения и логарифма числа алгоритмов семейства $B_r^m(a_0)$ от радиуса r. Благодаря значительной «плотности» данного семейства вероятность переобучения остаётся на приемлемо низком уровне при мощности семейства порядка тысяч.

Следующая лемма показывает, что вероятность переобучения множества алгоритмов, лежащего в слое, может только увеличиться при добавлении к нему новых алгоритмов из того же слоя.

Лемма 4. *Пусть μ — МЭР, $A \subset A_m$ и $a \in A_m \setminus A$. Тогда $Q_\varepsilon(A) \leqslant Q_\varepsilon(A \cup \{a\})$.*

Зададимся целью найти в m-м слое шара подмножество B' малой мощности, чтобы их вероятности переобучения совпадали или хотя бы не сильно отличались: $Q_\varepsilon(B') \lessapprox Q_\varepsilon\big(B_{r_0}^m(a_0)\big)$. Следующая теорема ограничивает область поиска такого подмножества m-м слоем сферы $S_{2\lfloor r_0/2\rfloor}^m(a_0)$.

Введем обозначение $\delta = \lfloor r_0/2\rfloor$.

Теорема 5. *Пусть μ — МЭР, $m = n(a_0, \mathbb{X})$ и $k \geqslant m + \delta$. Тогда $Q_\varepsilon\big(B^m_{r_0}(a_0)\big) = Q_\varepsilon\big(S^m_{2\delta}(a_0)\big)$.*

Итак, вероятность переобучения слоя шара «сосредоточена на его внешней окружности». Построим два подмножества $S^m_{2\delta}(a_0)$, приближающих вероятность переобучения слоя шара. Обозначим через X^m множество объектов, на которых алгоритм a_0 допускает ошибку, $\bar{X}^m = \mathbb{X} \setminus X^m$.

Подмножество алгоритмов $B'(m, r_0)$ образуется всеми C^δ_m различными алгоритмами, которые допускают $m-\delta$ ошибок на X^m и δ ошибок на фиксированных объектах подвыборки $\bar{X}^m$. Очевидно, что оно целиком лежит в $S^m_{2\delta}(a_0)$.

Теорема 6. *Пусть μ — МЭР, $A = B'(m, r_0)$. Тогда для любого $\varepsilon \in [0, 1]$:*

$$Q_\varepsilon = \sum_{\substack{i=0\\ i+j+p=\ell}}^{m} \sum_{j=0}^{h} \sum_{p=0}^{\delta} \frac{C^i_m C^j_h C^p_\delta}{C^\ell_L} \times$$

$$\times \Big(\big[i < \delta\big]\big[p \leqslant s_d(\varepsilon)\big] + \big[i \geqslant \delta\big]\big[i + p \leqslant \delta + s_d(\varepsilon)\big] \Big),$$

где $h = L - m - \delta$, $s_d(\varepsilon) = \frac{\ell}{L}(m - \varepsilon k)$.

Подмножество алгоритмов $B''(m, r_0)$**.** Пусть $L - m$ кратно δ. Семейство B'' — это все $C^\delta_m \frac{L-m}{\delta}$ алгоритмов, допускающих $m - \delta$ ошибок на X^m и δ ошибок на $\bar{X}^m$ при том, что для любого $x \in \bar{X}^m$ существует единственный $a \in B''$: $I(a, x) = 1$. Семейство B'' также является подмножеством $S^m_{2\delta}(a_0)$.

Теорема 7. *Пусть μ — МЭР и $\ell < \frac{L-m}{\delta}$. Тогда $Q_\varepsilon\big(B''(m, r_0)\big) = Q_\varepsilon\big(B^m_{r_0}(a_0)\big)$.*

Численный эксперимент

Вероятность переобучения Q_ε нетрудно оценить эмпирически методом Монте-Карло, если вместо доли всех разбиений выборки взять долю разбиений из заданного случайного подмножества разбиений. В данном эксперименте бралась тысяча случайных разбиений. Сравнивались точные оценки вероятности переобучения множеств $B^m_{r_0}(a_0)$ мощности 809 876 и $B'(m, r_0)$ мощности 45 с эмпирическими оценками $B''(m, r_0)$ мощности 4 275 и множества, состоящего из D случайных представителей $S^m_{2\delta}(a_0)$. Результаты представлены на рисунке 4.

В этом примере условие теоремы 7 не выполняется. Тем не менее семейство B'' достаточно хорошо приближает вероятность переобучения слоя шара. Семейство B' показало худший результат. Это объясняется тем, что алгоритмы из B' не различаются на подвыборке $\bar{X}^m$ и, в отличие от B'', заполняют множество $S^m_{2\delta}(a_0)$ неравномерно. Стоит также заметить, что Q_ε случайного подмножества $S^m_{2\delta}(a_0)$ достигает значения слоя шара при $D \approx 70$, что указывает на избыточность семейства B''.

Рис. 4. Оценки Q_ε для множеств $B^m_{r_0}(a_0)$, B', B'' и случайного подмножества D алгоритмов из $S^m_{2\delta}(a_0)$, при $l = k = 100$, $m = 10$, $r_0 = 4$, $\varepsilon = 0{,}05$.

Выводы

Получены точные формулы для вероятности переобучения модельных семейств алгоритмов — хэммингова шара радиуса r и некоторых его подмножеств. Показано, что на внешней окружности слоя хэммингова шара можно выбрать «разреженное» подсемейство из небольшого числа (порядка десятков) алгоритмов, вероятность переобучения которого будет очень близка к вероятности переобучения всего слоя, состоящего из огромного числа алгоритмов. Таким образом, дано косвенное объяснение экспериментов [4], в которых измеренные значения эффективного локального коэффициента разнообразия также не превосходили нескольких десятков.

Есть основания полагать, что аппроксимация «плотных» семейств их «разреженными» подсемействами является перспективным подходом, который позволит оценивать вероятность переобучения в практических ситуациях. Данная работа является первым шагом в этом направлении.

Литература

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

[2] *Boucheron S., Bousquet O., Lugosi G.* Theory of classification: A survey of some recent advances // ESAIM: Prob. and Stat. — 2005. — No. 9. — Pp. 323–375.

[3] *Frey A. I.* Accurate estimates of the generalization ability for symmetric sets of predictors and randomized learning algorithms // Pattern Recognition and Image Analysis. — 2010. — Vol. 20, No. 3. — Pp. 241–250.

[4] *Vorontsov K. V.* Combinatorial probability and the tightness of generalization bounds // Patt. Rec. and Image An. — 2008. — Vol. 18, No. 2. — Pp. 243–259.

[5] *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.

[6] *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.