О ВЛИЯНИИ РАЗЛИЧНОСТИ АЛГОРИТМОВ НА ВЕРОЯТНОСТЬ ПЕРЕОБУЧЕНИЯ[1]

К. В. Воронцов

ВЦ РАН, Москва, voron@ccas.ru

Получение точных верхних оценок вероятности переобучения остаётся открытой проблемой в теории статистического обучения уже более 40 лет, начиная с работ В. Н. Вапника и А. Я. Червоненкиса. Наиболее точные из известных оценок всё ещё сильно завышены. Данное экспериментальное исследование нацелено на поиск новых путей повышения точности теоретических оценок. Показано, что на вероятность переобучения решающим образом влияет различность алгоритмов. Получение адекватных оценок возможно только на пути совместного учёта двух факторов: локализации (расслоения) семейства и различности алгоритмов.

Пусть задано множество объектов $\mathbb{X}$, множество алгоритмов A, и существует функция $I: A\times\mathbb{X}\to\{0,1\}$, называемая *индикатором ошибки*; $I(a,x)=1$ означает, что алгоритм a допускает ошибку на объекте x. *Вектором ошибок* алгоритма a на выборке объектов $X^\ell=\{x_i\}_{i=1}^\ell$ из $\mathbb{X}^\ell$ называется ℓ-мерный бинарный вектор $\vec{a}=\big(I(a,x_i)\big)_{i=1}^\ell$. *Частота ошибок* $\nu(a,X^\ell)$ алгоритма a на выборке X^ℓ определяется как доля единиц в этом векторе.

Задан *метод обучения* $\mu:\mathbb{X}^\ell\to A$, который произвольной *обучающей* выборке X^ℓ из $\mathbb{X}^\ell$ ставит в соответствие некоторый алгоритм a из A. Пусть X^k — произвольная выборка из $\mathbb{X}^k$, называемая *контрольной*. *Переобученностью* метода μ на паре выборок (X^ℓ,X^k) назовём отклонение частот ошибок алгоритма $a=\mu(X^\ell)$:

$$\delta(\mu,X^\ell,X^k)=\nu(a,X^k)-\nu(a,X^\ell).$$

В слабой вероятностной аксиоматике [1] рассматривается конечная выборка X^L и предполагается, что все её разбиения на наблюдаемую обучающую выборку X_n^ℓ длины ℓ и скрытую контрольную выборку X_n^k длины k равновероятны, где $n=1,...,N$ — номер разбиения, $N=C_L^\ell$.

Основная задача — получить не сильно завышенные верхние оценки *вероятности переобучения* $Q_\varepsilon=\mathsf{P}_n\big\{\delta(\mu,X_n^\ell,X_n^k)\ge\varepsilon\big\}$.

Коэффициентом разнообразия (shatter coefficient) множества алгоритмов $A'\subseteq A$ на выборке X^L называется число попарно различных векторов $\vec{a}$, порождаемых всеми алгоритмами a из A' на выборке X^L. В слабой аксиоматике доказана следующая теорема [1], аналогичная основной теореме Вапника-Червоненкиса из [2].

Теорема 1. Для любых X^L, μ и $\varepsilon\in[0,1)$

$$Q_\varepsilon\le\sum_{m=1}^L D_m H_L^{\ell,m}(\varepsilon)\le\Delta_L^\ell\max_{m=1,..,L}H_L^{\ell,m}(\varepsilon), \quad (1)$$

где $H_L^{\ell,m}(\varepsilon)$ — функция гипер–геометрического распределения, D_m — коэффициент разнообразия множества $A_m=\Big\{a_n=\mu(X_n^\ell)\Big|\nu(a_n,X^L)=\frac{m}{L},\ n=1,...,N\Big\}$, Δ_L^ℓ — коэффициент разнообразия множества $A_L^\ell=\Big\{a_n=\mu(X_n^\ell)\Big|n=1,...,N\Big\}$.

Если оценивается вероятность большого уклонения частот для фиксированного алгоритма $a=\mu(X_n^\ell)=\mathrm{const}(n)$, то оценка (1) является точной. В общем случае она сильно завышена. В [1] экспериментально показано, что наиболее существенны два фактора завышенности.

[1] Работа выполнена при поддержке РФФИ, проект 08-07-00422.

Во-первых, не вполне учитывается эффект *локализации* или *расслоения* множества A_L^ℓ на подмножества A_m. Коэффициенты D_m с ростом m, как правило, возрастают, однако вероятность получить алгоритм a_n из A_m с достаточно малым $\nu(a_n, X_n^\ell)$ падает экспоненциально по m. Таким образом, «эффективно используемое» подмножество A_L^ℓ в каждой задаче сильно локализовано. *Оценки расслоения* (shell bounds) исследовались Дж. Лэнгфордом в [3].

Во-вторых, для выделения сомножителей D_m вероятность объединения событий $\left\{\delta(\mu, X_n^\ell, X_n^k) \geq \varepsilon\right\}$ оценивается сверху суммой их вероятностей (union bound). Эта оценка тем сильнее завышена, чем более схожи векторы ошибок алгоритмов. Влияние различности алгоритмов на вероятность переобучения изучалось в [4,5], однако радикального улучшения оценок добиться так и не удалось.

Цель данной работы — показать, что различность алгоритмов решающим образом влияет на переобучение, и что адекватные верхние оценки вероятности переобучения возможно получить только путём одновременного учёта и локализации (расслоения) семейства, и различности.

Эффективный локальный коэффициент разнообразия (ЭЛКР) — это такое значение коэффициента Δ_L^ℓ, при котором оценка (1) не является завышенной [1]. Наряду с вероятностью переобучения Q_ε удобно рассматривать верхнюю оценку ЭЛКР,

$$\overline{\Delta}_L^\ell = \frac{\mathsf{P}_n\left\{\delta(\mu, X_n^\ell, X_n^k) \geq \varepsilon\right\}}{\min\limits_{a \in A_L^\ell} \mathsf{P}_n\left\{\delta(a, X_n^\ell, X_n^k) \geq \varepsilon\right\}}.$$

показывающую, во сколько раз вероятность переобучения метода μ превышает вероятность большого уклонения частот у наилучшего алгоритма a.

В экспериментах вероятности $\mathsf{P}_n\{\bullet\}$ оцениваются методом Монте-Карло по случайному подмножеству разбиений.

Далее предполагается, что μ — это метод *минимизации эмпирического риска*:

$$\mu(X^\ell) = \arg\min_{a \in A} \nu(a, X^\ell).$$

Частный случай: два алгоритма

Рассмотрим семейство из двух алгоритмов $A = \{a_1, a_2\}$. В случаях неоднозначности $\nu(a_1, X_n^\ell) = \nu(a_2, X_n^\ell)$, будем полагать, что выбирается алгоритм с бóльшим числом ошибок на полной выборке.

Теорема 2. Пусть в выборке X^L имеется m_0 объектов, на которых оба алгоритма допускают ошибку; m_1 объектов, на которых только a_1 допускает ошибку; m_2 объектов, на которых только a_2 допускает ошибку; для определённости $m_1 \leq m_2$. Тогда для любого $\varepsilon \in [0,1)$ справедлива точная оценка:

$$Q_\varepsilon = \sum_{s_0, s_1, s_2} \frac{C_{m_0}^{s_0} C_{m_1}^{s_1} C_{m_2}^{s_2} C_{L-m_0-m_1-m_2}^{\ell-s_0-s_1-s_2}}{C_L^\ell} \times$$
$$\times\Big(\big[s_1 < s_2\big]\big[s_0 + s_1 \leq \tfrac{\ell}{L}(m_0 + m_1 - \varepsilon k)\big] +$$
$$+\big[s_2 \leq s_1\big]\big[s_0 + s_2 \leq \tfrac{\ell}{L}(m_0 + m_2 - \varepsilon k)\big]\Big),$$

где суммирование производится по всем $s_0 = 0,\ldots,m_0$, $s_1 = 0,\ldots,m_1$, $s_2 = 0,\ldots,m_2$ таким, что $m_0 + m_1 + m_2 \leq s_0 + s_1 + s_2 + k \leq L$.

На рис. 1 показаны зависимости ЭЛКР от хэммингова расстояния между векторами алгоритмов $\rho(\vec{a}_1, \vec{a}_2) = m_1 + m_2$. Графики позволяют сделать следующие выводы.

1. Переобучение неизбежно даже когда выбор делается только из двух алгоритмов.
2. Если алгоритмы допускают на X^L одинаковое число ошибок $m_1 = m_2$, и при этом максимально различны ($m_0 = 0$), то вапниковская оценка $\overline{\Delta}_L^\ell = 2$ достигается или почти достигается (верхний график).
3. Если алгоритмы схожи, то значение $\overline{\Delta}_L^\ell$ близко к 1. Иными словами, два схожих алгоритма ведут себя практически так же, как один алгоритм.
4. Если алгоритмы различны по числу ошибок, $\delta = m_2 - m_1 > 0$, то вапниковская оценка не достигается. Чем больше δ, тем меньше вероятность переобучения (нижний график). Значит, эффект расслоения проявляется уже при двух алгоритмах.

[1] Работа выполнена при поддержке РФФИ, проект 08-07-00422.

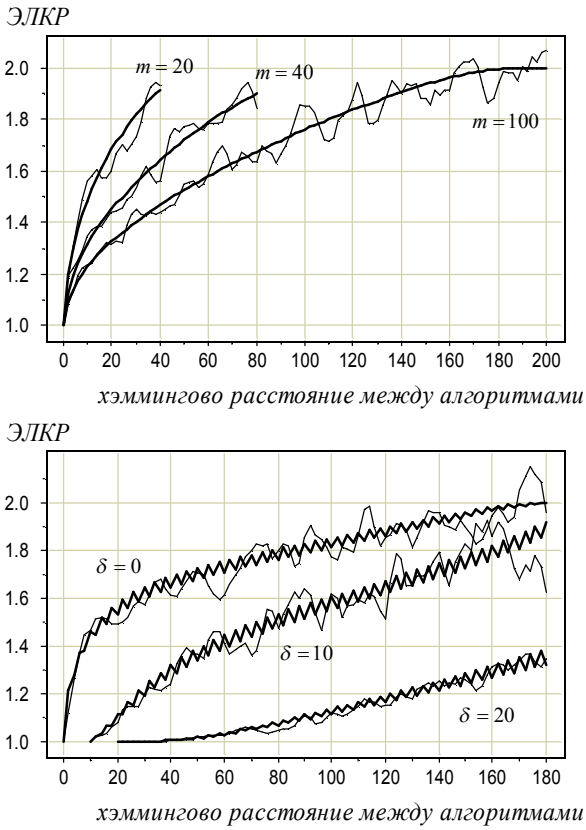

Рис. 1. Жирные линии — зависимость ЭЛКР от различности алгоритмов, тонкие линии — оценки ЭЛКР методом Монте-Карло по 1000 случайных разбиений; $\ell = k = 100$, $\varepsilon = 0.05$. Верхний график: $m_1 = m_2$, три кривые соответствуют трём значениям $m = Lv(a_i, X^L) = m_0 + m_1 = 20, 40, 100$. Нижний график: $m_2 = m_1 + \delta$, $m_0 = 20$; три кривые соответствуют трём значениям $\delta = 0, 10, 20$.

Эксперимент с цепочками алгоритмов

Многие широко используемые на практике параметрические семейства алгоритмов $A = \left\{ a(x,\gamma) \middle| \gamma \in \mathbb{R}^d \right\}$ обладают следующим свойством: при изменении вектора параметров γ по некоторой непрерывной траектории $\gamma(t)$ каждое изменение вектора ошибок $\vec{a}(x,\gamma)$ происходит только на одном объекте. Одновременное изменение нескольких координат возможно, но только для «редких» траекторий $\gamma(t)$, образующих в пространстве траекторий множество меры нуль. В частности, этим свойством обладают классификаторы с непрерывной по параметрам разделяющей поверхностью: линейные классификаторы, SVM с непрерывными ядрами, нейронные сети с непрерывными функциями активации, решающие деревья с пороговыми условиями ветвления, и многие другие. В [5] такие семейства называются связными, и множество векторов ошибок всех алгоритмов семейства представляется в виде связного графа. В [4] семейство кластеризуется на группы схожих классификаторов. В данной работе исследуются цепочки алгоритмов.

Цепочкой алгоритмов будем называть такую последовательность $A = \{a_1, ..., a_D\}$, в которой хэммингово расстояние между векторами ошибок алгоритмов $\vec{a}_{t-1}$ и $\vec{a}_t$ равно 1 для всех $t = 2, ..., D$.

В данной работе экспериментально исследовалась зависимость вероятности переобучения от длины цепочки D. Модельные цепочки задавались непосредственно набором векторов ошибок $\vec{a}_1, ..., \vec{a}_D$. Каждый следующий вектор $\vec{a}_t$ получался из $\vec{a}_{t-1}$ путём инверсии одной случайно выбранной координаты.

Строились цепочки двух типов:

1) *цепочки без расслоения*: число ошибок алгоритмов на полной выборке, чередуясь, принимало значения то m, то $m+1$;

2) *цепочки с расслоением*: число ошибок равнялось m для алгоритма a_1 и постепенно приближалось к $L/2$ для последующих алгоритмов.

Число m в обоих случаях являлось параметром эксперимента.

Кроме того, для каждой цепочки строилась соответствующая ей *не-цепочка*, состоящая из существенно различных алгоритмов $A' = \{a'_1, ..., a'_D\}$. Векторы $\vec{a}'_t$ генерировались случайным образом, но так, чтобы $v(a'_t, X^L) = v(a_t, X^L)$ для всех $t = 1, ..., D$.

Таким образом, в эксперименте строилось четыре конечных семейства алгоритмов. Их сопоставление позволяет разделить влияние *различности* (цепочки или не-цепочки) и *расслоения* (m ошибок у всех алгоритмов или только у лучшего) на вероятность переобучения.

На рис. 2 показаны графики зависимости вероятности переобучения и эффективного локального коэффициента разнообразия $\overline{\Delta}_L^\ell$ от числа алгоритмов в семействе D. На рис. 3 более детально показан начальный участок тех же графиков.

[1] Работа выполнена при поддержке РФФИ, проект 08-07-00422.

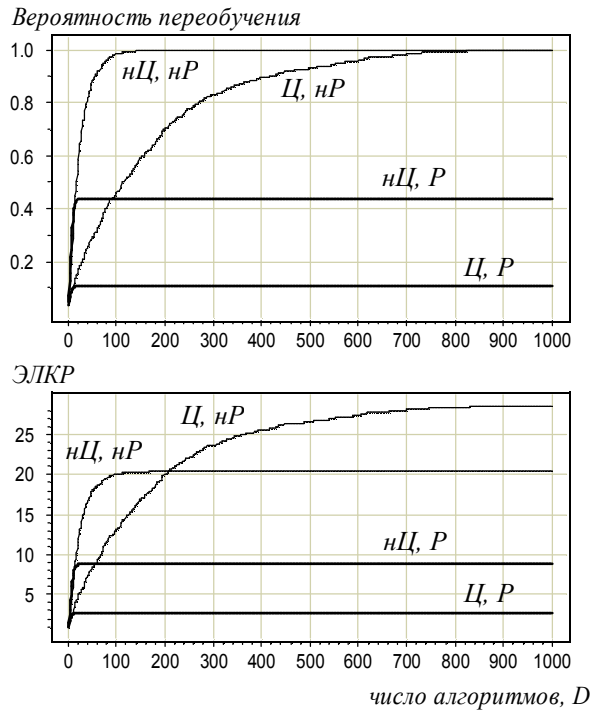

Рис. 2. Зависимость вероятности переобучения Q_ε и верхней оценки ЭЛКР $\overline{\Delta}_L^\ell$ от числа алгоритмов D, для четырёх типов семейств, при $\ell = k = 100$, $\varepsilon = 0.05$, $m = 10$. Оценка Методом Монте-Карло по 1000 случайных разбиений. Максимальное число алгоритмов 1000. Условные обозначения: *Ц* — цепочка, *нЦ* — не-цепочка, *Р* — расслоение, *нР* — без расслоения.

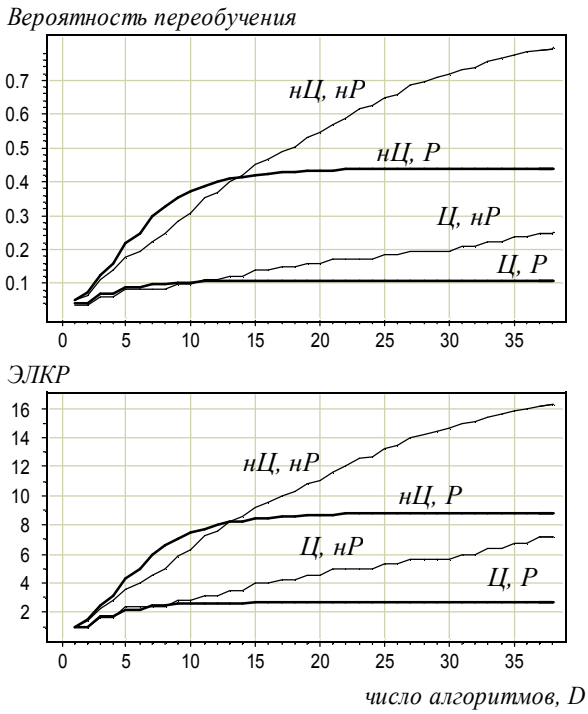

Рис. 3. Начальный участок графиков Рис. 2.

Выводы

1. Зависимость $\overline{\Delta}_L^\ell$ от D всегда «выходит на насыщение», тогда как $\overline{\Delta}_L^\ell = D$, согласно теории [2]. Оценка Вапника-Червоненкиса достигается только для не-цепочек и только при малых D (не более 10 в условиях данного эксперимента), рис. 3.
2. При наличии цепочки зависимость $\overline{\Delta}_L^\ell$ от D растёт медленнее, рис. 3.
3. При наличии расслоения вероятность переобучения Q_ε может не достигать 1 даже при очень больших D. В цепочках без расслоения вероятность переобучения достигает 1 при D порядка сотен. Таким образом, в условиях данного эксперимента наличие расслоения важнее, чем наличие цепочки.
4. Нетрудно показать, что функционал равномерной сходимости [2] совпадает с функционалом Q_ε только при отсутствии расслоения, иначе он может существенно превосходить Q_ε. Это означает, что функционал равномерной сходимости в принципе не способен объяснить, почему Q_ε не достигает 1 при наличии расслоения.
5. При больших D только одновременное наличие и цепочки, и расслоения позволяет избежать сильного переобучения (самые нижние кривые на графиках рис. 2, рис. 3). Внушает оптимизм тот факт, что именно этот случай чаще всего встречается на практике. Однако получение теоретических оценок, учитывающих оба эти явления, пока остаётся открытой проблемой.

Работа выполнена при поддержке РФФИ, проект №08-07-00422, и программы ОМН РАН «Алгебраические и комбинаторные методы математической кибернетики».

Литература

1. Vorontsov K. V. Combinatorial probability and the tightness of generalization bounds // Pattern Recognition and Image Analysis. — 2008. — Vol. 18, no. 2. — Pp. 243–259.
2. Вапник В. Н., Червоненкис А. Я. Теория распознавания образов. М.: Наука, 1974.
3. Langford J. Quantitatively tight sample complexity bounds. — 2002. — Carnegie Mellon Thesis.
4. Bax E. T. Similar classifiers and VC error bounds: Tech. Rep. CalTech-CS-TR97-14:6 1997.
5. Sill J. Monotonicity and connectedness in learning systems: Ph.D. thesis. — California Institute of Technology, 1998.

[1] Работа выполнена при поддержке РФФИ, проект 08-07-00422.