Минимизация вероятности переобучения для композиций линейных классификаторов низкой размерности*

Соколов Е. А., Воронцов К. В.
sokolov.evg@gmail.com, vokov@forecsys.ru
Москва, ВМК МГУ

Предлагается улучшенная комбинаторная оценка вероятности переобучения, учитывающая эффекты расслоения и связности в семействах алгоритмов классификации, а также способ её эффективного вычисления на основе метода случайных блужданий. Использование этой оценки в качестве критерия отбора признаков повышает обобщающую способность простого голосования линейных классификаторов, одновременно снижая их размерность и сокращая их число в композиции.

Оценки обобщающей способности алгоритмов классификации часто используются для отбора признаков, оптимизации сложности и структуры модели. Большинство из известных оценок довольно сильно завышены, что может приводить к неоптимальному выбору структурных параметров. Комбинаторная теория переобучения приводит к более точной оценке благодаря учёту свойств расслоения и связности в семействах алгоритмов [1]. В данной работе предлагается ещё более точная комбинаторная оценка и метод её эффективного вычисления. Полученная оценка применяется в качестве критерия отбора признаков в линейных классификаторах, которые затем объединяются в композицию путём простого голосования. Каждый базовый алгоритм обучается методом SVM по подвыборке объектов, формируемой методом комитетного бустинга ComBoost [2]. Получаемая композиция может рассматриваться как разреженная (неполносвязная) двухслойная нейронная сеть. При этом высокой обобщающей способностью обладает как сама композиция, так и отдельные базовые классификаторы (нейроны первого слоя) при весьма скромном их количестве (3–6).

Комбинаторная теория переобучения

Пусть $\mathbb{X} = \{x_1, \ldots, x_L\}$ — конечная *генеральная совокупность* объектов, $\mathcal{A} = \{a_1, \ldots, a_D\}$ — конечное множество *алгоритмов* с попарно различными бинарными векторами ошибок $\tilde{a} = \big(I(a, x_i)\big)_{i=1}^{L}$, где $I(a, x) \in \{0, 1\}$ — индикатор ошибки алгоритма $a \in \mathcal{A}$ на объекте $x \in \mathbb{X}$.

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

Число ошибок алгоритма a на выборке $X \in \mathbb{X}$ определяется как $n(a, X) = \sum\limits_{x \in X} I(a, x)$.

Частота ошибок алгоритма a на выборке X определяется как $\nu(a, X) = n(a, X)/|X|$.

Переобученностью $\delta(\mu, X, \bar{X})$ метода μ при разбиении $\mathbb{X} = X \sqcup \bar{X}$ называется величина отклонения частоты ошибок на контроле и обучении

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

Предполагается, что все разбиения множества объектов $\mathbb{X} = X \sqcup \bar{X}$ на две выборки — наблюдаемую обучающую X длины ℓ и скрытую контрольную $\bar{X}$ длины $k = L - \ell$, равновероятны.

Основной задачей является получение верхних оценок *вероятности переобучения*

$$Q_\varepsilon(\mu, \mathbb{X}) = \mathsf{P}\big[\delta(\mu, X, \bar{X}) \geqslant \varepsilon\big], \quad \varepsilon \in (0, 1).$$

В данной работе рассматриваются методы μ, *минимизирующие эмпирический риск* (МЭР):

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

Для получения верхних оценок Q_ε вводится метод *пессимистичной* МЭР:

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

Введём на множестве бинарных векторов ошибок естественное отношение (частичного) порядка $a \leqslant b$, хэммингово расстояние $\rho(a, b)$ и отношение предшествования: $a \prec b \Leftrightarrow (a \leqslant b) \wedge (\rho(a, b) = 1)$, т. е. *$a$ предшествует b*, если b ошибается на тех же объектах, что и a, и ещё на одном объекте.

Графом расслоения-связности множества $\mathcal{A}$ называется ориентированный граф $G = (\mathcal{A}, E)$ с множеством ребер $E = \big\{(a, b) \mid a \prec b\big\}$.

Верхней связностью $u(a)$ алгоритма a называется число таких алгоритмов b, что $a \prec b$.

Неполноценностью $q(a)$ алгоритма a называется число объектов $x \in \mathbb{X}$, на которых a ошибается, при том, что существует алгоритм $b \leqslant a$, не ошибающийся на x.

Оценки вероятности переобучения

Известна верхняя оценка вероятности переобучения, которая выражается через неполноценность и верхнюю связность всех алгоритмов множества $\mathcal{A}$ и справедлива для ПМЭР, значит, и для произвольного метода МЭР [1].

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

Теорема 1. *Для пессимистичного минимизатора эмпирического риска μ, любых $\mathbb{X}$, $\mathcal{A}$ и $\varepsilon \in (0,1)$*

$$Q_\varepsilon \leqslant \sum_{a\in\mathcal{A}} \frac{C^{\ell-u}_{L-u-q}}{C^\ell_L} \mathcal{H}^{\ell-u,\ m-q}_{L-u-q}\left(\tfrac{\ell}{L}(m-\varepsilon k)\right), \qquad (1)$$

где $\mathcal{H}^{\ell,m}_L(z) = \sum_{s=0}^{\lfloor z\rfloor} \frac{C^s_m C^{\ell-s}_{L-m}}{C^\ell_L}$ — функция гипергеометрического распределения, $u \equiv u(a)$, $q \equiv q(a)$, $m \equiv n(a,\mathbb{X})$.

Данная оценка является точной для некоторых нетривиальных модельных семейств алгоритмов [3], но на реальных семействах оказывается завышенной на 1–2 порядка. В данной работе вводятся характеристики попарного сходства алгоритмов, которые позволяют улучшить данную оценку.

Определим для произвольных двух алгоритмов a_i и a_j множество A_{ij} объектов, на которых a_i не допускает ошибку, а a_j допускает:

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

Лемма 2. *Пусть алгоритмы пронумерованы в порядке неубывания числа ошибок, μ — ПМЭР. Тогда для произвольной обучающей выборки $X \subset \mathbb{X}$*

$$[\mu X = a_i] = \left(\prod_{j=1}^{i-1}\Big[|X\cap A_{ji}| \leqslant |X\cap A_{ij}|\Big]\right)\times \times\left(\prod_{j=i+1}^{D}\Big[|X\cap A_{ji}| < |X\cap A_{ij}|\Big]\right).$$

Непосредственное использование данного выражения для вычисления вероятности переобучения не представляется возможным из-за вычислительной сложности. Но его можно превратить в необходимое условие, оставив для алгоритма a_i только множители, соответствующие предшествующему ему истоку в графе расслоения-связности G и алгоритмам из его верхней полуокрестности. Нетрудно показать, что тогда из необходимого условия будет следовать комбинаторная оценка теоремы 1. В данной работе предлагается учесть связь алгоритма a_i с произвольным алгоритмом a_s.

Лемма 3. *Пусть μ — ПМЭР, a_i и a_s — два произвольных алгоритма из $\mathcal{A}$. Тогда*

$$\mathsf{P}\big[\mu X = a_i\big]\big[\nu(a_i,\bar{X}) - \nu(a_i,X) \geqslant \varepsilon\big] \leqslant \leqslant \sum_{t=0}^{T_{is}} \frac{C^t_q C^{\ell-u-t}_{L-u-q}}{C^\ell_L} \mathcal{H}^{\ell-u-t,\,m-q}_{L-u-q}\left(\tfrac{\ell}{L}(m-\varepsilon k)-t\right),$$

где $u \equiv u(a_i)$, $q \equiv |A_{si}|$, $T_{is} = \min(|A_{is}|,|A_{si}|)$, $m \equiv n(a_i,\mathbb{X})$.

Распорядимся свободой выбора алгоритма a_s: для каждого a_i будем брать в качестве a_s тот из истоков графа расслоения-связности G, который даёт наименьший вклад в оценку.

Теорема 4. *Пусть μ — ПМЭР, S — множество всех истоков графа расслоения-связности. Тогда, в обозначениях леммы 3*

$$Q_\varepsilon \leqslant \sum_{i=1}^{D} \min_{s\in S}\left\{\sum_{t=0}^{T_{is}} \frac{C^t_q C^{\ell-u-t}_{L-u-q}}{C^\ell_L} \times \times \mathcal{H}^{\ell-u-t,\,m-q}_{L-u-q}\left(\tfrac{\ell}{L}(m-\varepsilon k)-t\right)\right\}. \qquad (2)$$

Непосредственное вычисление оценки (2) требует перебора всех алгоритмов семейства $\mathcal{A}$, что на практике неосуществимо. Значимый вклад в оценку вносят лишь алгоритмы из некоторого числа t нижних слоев графа G. Тем не менее, даже этих алгоритмов может быть слишком много.

В следующем разделе предлагается приближенный метод вычисления оценки на основе случайного блуждания, не требующий полного обхода графа. Он существенно более эффективен, чем метод оценивания профиля расслоения-связности по случайной выборке алгоритмов из равномерного распределения на $\mathcal{A}$, предложенный в [4].

Приближённое вычисление верхних оценок вероятности переобучения

Обозначим правую часть неравенства (2) через B_ε; отдельное слагаемое в (2), соответствующее алгоритму $a \in \mathcal{A}$ — через $b(a)$. Пусть имеется случайная независимая выборка алгоритмов $a_1,\ldots,a_n$ из первых t слоев семейства $\mathcal{A}$, причем вероятность выбрать алгоритм $a \in \mathcal{A}$ равна $p(a)$. Тогда оценка

$$\hat{B}_\varepsilon = \frac{1}{n}\sum_{i=1}^{n} \frac{b(a_i)}{p(a_i)} \qquad (3)$$

является несмещенной и состоятельной для B_ε [5].

Чтобы сгенерировать выборку $a_1,\ldots,a_n$, можно применить к подграфу $G_t = (V_t, E_t)$, образованному нижними t слоями графа расслоения-связности $G = (\mathcal{A}, E)$, технику ленивого случайного блуждания [6], которое сходится к стационарному распределению $\pi(a) = \frac{\deg(a)}{2|E_t|}$, где $\deg(a)$ — степень вершины a. Таким образом, отбросив некоторое число первых алгоритмов в выборке (так как на первых шагах распределение отличается от стационарного), и подставив алгоритмы и вероятности их получения $\pi(a)$ в (3), получим оценку $\hat{B}_\varepsilon \approx B_\varepsilon$.

Ниже будут приведены результаты эксперимента, показывающего, что вычисленная таким способом оценка является хорошим приближением лишь при очень больших n, значительно превосходящих число вершин в графе. Улучшить сходимость позволяет следующий прием. Преобразуем оценку вероятности переобучения (2):

$$Q_\varepsilon \leqslant B_\varepsilon = \sum_{i=1}^{D} b(a_i) = \sum_{m=0}^{L} |A_m|\left(\frac{1}{|A_m|}\sum_{a\in A_m} b(a)\right),$$

Алгоритм 1. Модификация Frontier Sampling.

Вход: Граф $G = (V, E)$; число итераций N;
набор стартовых вершин $P = (v^1, \ldots, v^s)$;
Выход: Выборка вершин графа $v_1, v_2, \ldots, v_N$;

для $i = 1, \ldots, N$
  выбрать $v \in P$ с вероятностью $\frac{\deg(v)}{\sum\limits_{u \in P} \deg(u)}$;
  с вероятностью $\frac{1}{2}$
    выбрать вершину v' из равномерного
    распределения на $\{v' \in V \mid (v, v') \in E\}$;
    $v_i := v'$;
  иначе
    $v' := v$; $v_i := v$;
  Заменить в P вершину v на v';

где $A_m = \{a \in \mathcal{A}\colon n(a, \mathbb{X}) = m\}$ есть m-й слой графа G; последовательность $|A_m|$, $m = 0, \ldots, L$, называется *профилем расслоения* семейства $\mathcal{A}$. Обозначив средний вклад в оценку алгоритмов из m-го слоя через $B_m = \frac{1}{|A_m|} \sum\limits_{a \in A_m} b(a)$, получим

$$Q_\varepsilon \leqslant B_\varepsilon = \sum_{m=0}^{L} |A_m| B_m. \tag{4}$$

Оценивание каждой величины B_m по отдельности по выборке $a_1, \ldots, a_n$ позволяет получать достаточно точные оценки $\hat{B}_\varepsilon$ даже при небольших n:

$$\hat{B}_m = \frac{\sum\limits_{i=1}^{n} \dfrac{[m(a_i) = m]}{\pi(a_i)} \dfrac{b(a_i)}{|V_t|}}{\sum\limits_{i=1}^{n} \dfrac{[m(a_i) = m]}{\pi(a_i)}}. \tag{5}$$

Данная оценка не является несмещенной, однако экспериментально было установлено, что в среднем она очень близка к истинному значению B_m, что также подтверждается следующим результатом.

Лемма 5. *Оценка (5) является асимптотически несмещенной:* $\mathsf{E}\hat{B}_m \xrightarrow[n\to\infty]{} B_m$.

Чтобы узнать точное значение $|E_t|$ числа ребер в t нижних слоях, необходимо полностью обойти их, что крайне нежелательно. Предлагается вместо этого преобразовать величину $\pi(a)$:

$$\pi(a) = \frac{\deg(a)}{2|E_t|} = \frac{\deg(a)}{2|V_t|} \frac{|V_t|}{|E_t|}. \tag{6}$$

В экспериментах отношение $|V_t|/|E_t|$ достаточно точно оценивается по подграфу, полученному в процессе случайного блуждания, а величина $|V_t|$ — по случайной выборке алгоритмов из $\mathcal{A}$.

Известно, что обычное случайное блуждание может медленно сходиться к стационарному распределению, а получаемые с его помощью оценки имеют большую дисперсию, если граф разрежен [7]. Эти проблемы устраняются в методе Frontier Sampling [7], однако сходимость к стационарному распределению гарантируется только для графов, не являющихся двудольными, что не позволяет применять его к графу расслоения-связности. В данной работе предлагается модификация указанного метода (см. Алгоритм 1), гарантирующая сходимость для любых связных графов.

Лемма 6. *Пусть $p_i(a)$ — вероятность получить алгоритм a на шаге i случайного блуждания по графу G, осуществляемого по алгоритму 1. Тогда, если граф G является связным, то*

$$\|p_i(a) - \pi(a)\|_2 \xrightarrow[i\to\infty]{} 0.$$

Композиция линейных классификаторов с отбором признаков

Пусть известна некоторая оценка вероятности переобучения

$$\mathsf{P}[\nu(\mu X, \bar{X}) - \nu(\mu X, X) \geqslant \varepsilon] \leqslant \eta(\varepsilon).$$

Обратив ее, можно оценить частоту ошибок на контроле: с вероятностью не менее $1 - \eta$

$$\nu(\mu X, \bar{X}) \leqslant \nu(\mu X, X) + \varepsilon(\eta).$$

Величину в правой части неравенства можно использовать в качестве критерия при отборе признаков или при выборе модели. В данной работе он используется для отбора признаков при построении линейных классификаторов, объединяемых затем в композицию путём простого голосования:

$$a(x) = \operatorname{sign} \sum_{i=1}^{p} \operatorname{th}\langle w_i, x\rangle$$

Каждый из p базовых линейных классификаторов обучается по подвыборке объектов, отобранных методом ComBoost [2]: в подвыборку не включаются объекты, слишком хорошо и слишком плохо классифицируемые композицией предыдущих базовых классификаторов. Отбор признаков производится для каждого базового классификатора путём жадного добавления. Каждый набор признаков оценивается с помощью обращённой комбинаторной оценки вероятности переобучения, которая вычисляется следующим образом.

Сначала строится линейный классификатор методом SVM. Ему соответствует некоторая вершина в графе расслоения-связности G. Из этой вершины осуществляется спуск вниз по графу до истока. Затем из этого истока запускается обход всех слоев графа вплоть до $(m_0 + 3)$-го, где m_0 — число ошибок найденного истока, и фиксируются множество найденных истоков S. После этого генерируется 10000 случайных классификаторов, по которым оценивается профиль расслоения. Затем генерируется 2000 классификаторов из $t = m_0 + 20$

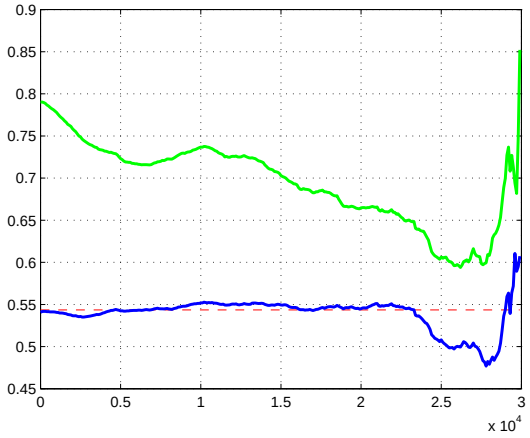

Рис. 1. Зависимость обращённых оценок вероятности переобучения от числа отброшенных первых сэмплов. Верхняя кривая соответствует оценке (3), нижняя (5). Пунктиром отмечено точное значение оценки.

нижних слоев графа с помощью случайного блуждания. По ним вычисляется приближённая оценка (2), которая затем обращается при $\eta = \frac{1}{2}$ методом дихотомии. Наконец, комбинаторный критерий вычисляется по формуле

$$Q_c = \nu(a_0, X) + \varepsilon\left(\tfrac{1}{2}\right), \qquad (7)$$

где a_0 — лучший из найденных классификаторов.

Эксперименты

Сходимость случайного блуждания к стационарному распределению. Эксперименты проводились с семейством линейных классификаторов на модельной выборке из $L = 200$ объектов, при $\ell = 100$, $\varepsilon = 0{,}1$. Наилучший линейный классификатор на данной выборке допускал 8 ошибок. Случайное блуждание осуществлялось по первым $t = 30$ слоям графа расслоения-связности.

Рис. 1 показывает, что оценка, полученная по формуле (3), является смещенной и завышенной; сходимость метода крайней медленная и не достигается даже за 30000 итераций, что значительно превосходит число алгоритмов в нижних t слоях. Оценка, полученная по формуле (5), становится несмещенной после 5000 итераций.

Отбор признаков. Для проведения экспериментов с отбором признаков был выбран набор данных `Wine Quality` из репозитория UCI, образуемый 4898 объектами и 11 признаками. В обучающую выборку было отобрано 250 объектов, остальные составили тестовую выборку.

Сравнивалась доля корректно классифицированных объектов тестовой выборки для четырёх методов построения композиции, отличающихся только критерием отбора признаков в базовых классификаторах:

Оценка (7), в которой обращается эмпирическая оценка вероятности переобучения по 100 случайным разбиениям	0,66
Оценка полного скользящего контроля CCV [1], вычисленная с помощью случайных блужданий	0,67
Оценка (7), в которой обращается комбинаторная оценка (1), вычисленная с помощью случайных блужданий	0,69
Оценка скользящего контроля, вычисленная по 100 случайным разбиениям	0,71
Оценка (7), в которой обращается комбинаторная оценка (2), вычисленная с помощью случайных блужданий	0,74

Наилучшие результаты получились с использованием комбинаторного критерия, основанного на предложенной оценке (2) и модификации метода случайных блужданий по графу Frontier Sampling. Полученный классификатор состоял из 4-х базовых классификаторов, каждый из которых был построен по подпространству признаков размерности от двух до четырех.

Таким образом, тщательный контроль переобучения при отборе признаков в базовых классификаторах позволяет, в отличие от стандартных методов типа бустинга и бэггинга, обходиться малым числом базовых классификаторов. В терминах нейронных сетей данный метод строит сильно разреженную двухслойную нейронную сеть с автоматическим выбором числа нейронов в скрытом слое.

Литература

[1] *Воронцов К. В.* Комбинаторная теория переобучения: результаты, приложения и открытые проблемы // Всероссийская конференция ММРО-15. — М.: МАКС Пресс, 2011. — С. 40–43.

[2] *Маценов А. А.* Комитетный бустинг: минимизация числа базовых алгоритмов при простом голосовании // Всероссийская конференция ММРО-13. — М.: МАКС Пресс, 2007. — С. 180–183.

[3] *Botov P. V.* Exact bounds on probability of overfitting for multidimensional model sets of classifiers // Pattern Recognition and Image Analysis. — 2011. — Vol. 21., no. 1. — Pp. 52–65.

[4] *Kochedykov D. A.* A combinatorial approach to hypothesis similarity in generalization bounds // Pattern Recognition and Image Analysis. — 2011. — Vol. 21. — Pp. 616–629.

[5] *Avrachenkov K., Ribeiro B., Towsley D.* Improving random walk estimation accuracy with uniform restarts // Proc. of WAW 2010, December 2010.

[6] *Lovasz L.* Random walks on graphs: a survey // Combinatorics. — 1993. —no 2. — Pp. 1–46.

[7] *Ribeiro B., Towsley D.* Estimating ans sampling graphs with multidimensional random walks // 10th Conf. on Internet Measurement, 2010. — Pp. 390–403.