Верхние оценки переобученности и профили разнообразия логических закономерностей

Ивахненко А. А., Воронцов К. В.
andrej_iv@mail.ru, voron@ccas.ru
Москва, Вычислительный Центр РАН

Логические алгоритмы классификации представляют собой композиции элементарных классификаторов, называемых также закономерностями. Существуют два противоположных подхода к повышению качества (обобщающей способности) таких алгоритмов: либо увеличение числа закономерностей в композиции [1], либо повышение качества закономерностей. Качество алгоритма в обоих случаях может оказаться сопоставимым, однако при втором подходе получаются более простые, легко интерпретируемые алгоритмы. В данной работе понятие обобщающей способности, которое обычно определяется для алгоритмов, распространяется на случай закономерностей. В рамках комбинаторного подхода [2] выводятся сложностные оценки качества закономерностей. Предлагается методика эмпирического измерения завышенности получаемых оценок, основанная на скользящем контроле.

Основные определения

Рассмотрим стандартную постановку задачи классификации. Задано множество допустимых объектов X, конечное множество имён классов Y и обучающая выборка $X^\ell = (x_i, y_i)_{i=1}^\ell \subset X \times Y$. Предполагается, что $y_i = y^*(x_i)$, где $y^*\colon X \to Y$ — неизвестная целевая зависимость. Требуется построить алгоритм $a\colon X \to Y$, приближающий y^* на всём X.

Закономерностью называется предикат $\varphi_y\colon X \to \{0,1\}$, выделяющий достаточно много объектов класса y и достаточно мало объектов всех остальных классов. Предикат φ_y *выделяет* объект x, если $\varphi_y(x) = 1$.

Логические алгоритмы представляются в виде линейных композиций вида $a(x) = \arg\max\limits_{y\in Y} \sum_{t=1}^{T_y} w_y^t \varphi_y^t(x)$, где φ_y^t — закономерности, w_y^t — веса закономерностей, T_y — число закономерностей класса y.

Методом обучения называется отображение μ, которое по выборке X^ℓ строит набор закономерностей $\mu X^\ell \equiv \mu(X^\ell) = \left\{\varphi_y^t(x)\right\}_{y\in Y}^{t=1,T_y}$.

Частота ошибок закономерности φ_y на выборке $U \subset X$ есть

$$\nu(\varphi_y, U) = \frac{1}{|U|} \sum_{x\in U} \big[\varphi_y(x) \neq [y^*(x) = y]\big].$$

Переобученностью закономерности $\varphi_y \in \mu X^\ell$ при заданной контрольной выборке X^k называется разность частот её ошибок на контроле и на обучении $\delta(\varphi_y, X^\ell, X^k) = \nu(\varphi_y, X^k) - \nu(\varphi_y, X^\ell)$.

Рассмотрим множество всех разбиений полной выборки $X^L = X_n^\ell \cup X_n^k$ на две подвыборки — обучающую длины ℓ и контрольную длины k, где $\ell + k = L$, индекс n пробегает множество всех разбиений $N = \{1, \ldots, C_L^\ell\}$.

Введём функционал *полного скользящего контроля* $Q_\varepsilon(\mu, X^L)$ как долю переобученных закономерностей среди закономерностей, построенных методом μ по всевозможным подвыборкам $X_n^\ell \subset X^L$:

$$Q_\varepsilon(\mu, X^L) = \frac{1}{|N|} \sum_{n \in N} \frac{1}{|\mu X_n^\ell|} \sum_{\varphi \in \mu X_n^\ell} \big[\delta(\varphi, X_n^\ell, X_n^k) > \varepsilon\big].$$

где $\varepsilon \in [0, 1)$ — *порог переобученности*. Аналогичный функционал, его верхние оценки и связь с теорией Вапника-Червоненкиса рассматривались в [2] для алгоритмов классификации и регрессии.

Коэффициенты и профили разнообразия

Назовем предикаты $\varphi, \varphi' \colon X \to \{0, 1\}$ *неразличимыми* или эквивалентными на выборке X^L, если $\varphi(x) = \varphi'(x)$ для всех $x \in X^L$. *Коэффициентом разнообразия* (shatter coefficient) $\Delta(\Phi, X^L)$ множества предикатов Φ на выборке X^L называется максимальное число попарно неразличимых предикатов из Φ, оно же число классов эквивалентности на Φ. Коэффициент разнообразия характеризует сложность множества предикатов Φ относительно заданной выборки X^L.

Рассмотрим множество закономерностей, получаемых методом μ по всевозможным обучающим подвыборкам: $\Phi_L^\ell \equiv \Phi_L^\ell(\mu, X^L) = \bigcup_{n=1}^N \mu X_n^\ell$. Его коэффициент разнообразия $\Delta_L^\ell \equiv \Delta_L^\ell(\mu, X^L) = \Delta(\Phi_L^\ell, X^L)$ назовём *локальным коэффициентом разнообразия* метода μ на выборке X^L.

Разобьём множество Φ_L^ℓ на $L + 1$ подмножеств, состоящих из закономерностей с фиксированным числом ошибок m на полной выборке X^L: $\Phi_m \equiv \Phi_m(\mu, X^L) = \big\{\varphi \in \Phi_L^\ell \colon \nu(\varphi, X^L) = \frac{m}{L}\big\},\ m = 0, \ldots, L$.

Локальным профилем разнообразия метода μ на выборке X^L назовём последовательность коэффициентов разнообразия $D_m \equiv D_m(\mu, X^L) = \Delta(\Phi_m, X^L),\ m = 0, \ldots, L$. Очевидно, что $\Delta_L^\ell = D_0 + \cdots + D_L$.

Наряду с функционалом Q_ε определим функционал $Q_{\varepsilon,m}$ как долю переобученных закономерностей, допускающих m ошибок на X^L:

$$Q_{\varepsilon,m}(\mu, X^L) = \frac{1}{|N|} \sum_{n \in N} \frac{1}{|\mu X_n^\ell|} \sum_{\varphi \in \mu X_n^\ell} \big[\delta(\varphi, X_n^\ell, X_n^k) > \varepsilon\big] \big[\nu(\varphi, X^L) = \tfrac{m}{L}\big].$$

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

$$Q_{\varepsilon,m}(\mu, X^L) \leqslant D_m H\tbinom{m\ s_1}{L\ \ell}, \quad m = 0, \ldots, L, \tag{1}$$

где $H\binom{m\ s_1}{L\ \ell} = \sum_{s=s_0}^{s_1} C_m^s C_{L-m}^{\ell-s} / C_L^\ell$ — хвост гипергеометрического распределения, $s_0 = \max\{0, m - k\}$ и $s_1 = \big\lfloor \frac{\ell}{L}(m - \varepsilon k) \big\rfloor$.

Задача	L	Глобальный	Локальный	Эффективный
crx	690	$2.8 \cdot 10^8$	$3.5 \cdot 10^4$	21 ± 11
german	1000	$5.2 \cdot 10^8$	$3.1 \cdot 10^4$	47 ± 38
hepatitis	155	$5.5 \cdot 10^6$	$1.8 \cdot 10^4$	58 ± 46
horse-colic	300	$1.9 \cdot 10^6$	$1.3 \cdot 10^4$	5 ± 3
hypothyroid	3163	$5.3 \cdot 10^8$	$2.2 \cdot 10^4$	43 ± 28
liver	345	$1.5 \cdot 10^7$	$2.9 \cdot 10^4$	12 ± 8
promoters	106	$4.4 \cdot 10^9$	$5.3 \cdot 10^4$	13 ± 4

Таблица 1. Коэффициенты разнообразия на 7 задачах классификации из репозитория UCI. Выборка разбивалась 20 раз случайным образом на равные части $\ell = k$ со стратификацией классов; $\varepsilon = 0.05$.

Теорема 2. *Для любых μ, X^L и порога переобученности $\varepsilon \in [0,1)$*

$$Q_\varepsilon(\mu, X^L) \leqslant \sum_{m=0}^{L} D_m H\tbinom{m\ s_1}{L\ \ell}.$$

Определим *эффективный локальный профиль разнообразия* $\widehat{D}_m$ как гипотетическое значение локального профиля D_m, при котором оценка (1) не является завышенной, т. е. неравенство обращается в равенство:

$$\widehat{D}_m = Q_{\varepsilon,m}(\mu, X^L) \;/\; H\tbinom{m\ s_1}{L\ \ell}, \quad m = 0, \ldots, L.$$

Эту величину легко измерить эмпирически, если в функционале $Q_{\varepsilon,m}$ заменить сумму по всем разбиениям N суммой по некоторому подмножеству $N' \subset N$ (в методе Монте-Карло N' — случайное подмножество).

Наконец, *эффективный локальный коэффициент разнообразия* определим как $\widehat{\Delta}_L^\ell = \widehat{D}_0 + \cdots + \widehat{D}_L$. Эта величина показывает, какое значение должен был бы принимать локальный коэффициент разнообразия, чтобы верхняя оценка не была завышенной. Данная методика измерения завышенности существенно уточняет методику, ранее предложенную в [3].

Измерение профилей разнообразия: эксперименты и выводы

Алгоритмы поиска закономерностей, основанные на непосредственном переборе предикатов, очень удобны для эмпирического исследования завышенности сложностных оценок, поскольку: (а) глобальный коэффициент разнообразия (*функция роста* по Вапнику) вычисляется по эффективной рекурсивной формуле [3]; (б) локальный коэффициент оценивается снизу числом различных закономерностей, найденных методом μ на подвыборках $\{X_n^\ell \colon n \in N'\}$. Результаты сравнения этих величин с эффективным локальным коэффициентом приведены в Таблице 1.

Эффективный локальный коэффициент всегда оказывался существенно меньшим длины выборки L. Это означает, что при фиксиро-

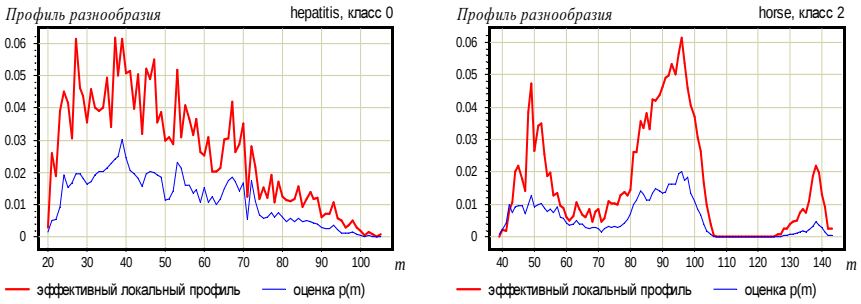

Рис. 1. Сравнение эффективного профиля $\widehat{D}_m$ и нормированного локального профиля $p(m)$ на двух задачах: hepatitis и horse.

ванных X^L, y^* и μ *ёмкость* (VC-dimension) эффективно используемого множества закономерностей никогда не превышает единицы.

Эмпирические нижние оценки локальных коэффициентов завышены, как минимум, на три порядка. Ни один из известных на сегодня подходов, включая наиболее точные [4], не способен дать оценки коэффициентов разнообразия порядка 10^1–10^2.

Интересные результаты дало сравнение эффективного профиля $\widehat{D}_m$ с нижней оценкой локального профиля $\tilde{D}_m$, подсчитанной как число различных закономерностей из $\{\mu X_n^\ell \colon n \in N'\}$, допускающих m ошибок на X^L. Практически во всех задачах оказалось, что нормированный локальный профиль $p(m) = \tilde{D}_m/(\tilde{D}_0 + \cdots + \tilde{D}_L)$ является лишь слегка заниженной оценкой эффективного профиля $\widehat{D}_m$ и, как правило, сильно с ним коррелирует (Рис. 1). Данному факту пока не найдено объяснения.

Работа выполнена при поддержке РФФИ, проекты №05-01-00877, №07-07-00181 и программы ОМН РАН «Алгебраические и комбинаторные методы математической кибернетики».

Литература

[1] *Cohen W. W., Singer Y.* A simple, fast and effective rule learner // Proc. of the 16 National Conference on Artificial Intelligence. — 1999. — Pp. 335–342.

[2] *Воронцов К. В.* Комбинаторный подход к оценке качества обучаемых алгоритмов // Мат. вопр. киберн. — М.: Физматлит, 2004. — Т. 13. — С. 5–36.

[3] *Воронцов К. В., Ивахненко А. А.* Эмпирические оценки локальной функции роста в задачах поиска логических закономерностей // *Искусственный Интеллект.* — Донецк, 2006. — С. 281–284.

[4] *Langford J.* Quantitatively tight sample complexity bounds. — 2002. — Carnegie Mellon Thesis.