О завышенности оценок функции роста в задачах поиска логических закономерностей

К. В. Воронцов, А. А. Ивахненко
(Москва)

Широкий класс логических алгоритмов классификации основан на поиске закономерностей в данных. К этому типу алгоритмов относятся решающие деревья и списки, алгоритмы типа «Кора», и другие. Получение наиболее точных оценок обобщающей способности этих алгоритмов является актуальной открытой проблемой. Теория Вапника-Червоненкиса даёт сильно завышенные оценки частоты ошибок на контрольных данных. Основная причина завышенности — переоценка функции роста. В работе [1] показано, что функцию роста достаточно оценивать не по всему семейству алгоритмов, а только по локальной его части, реально используемой в конкретной задаче классификации. В данной работе предлагается методика эмпирического оценивания эффективной локальной функции роста для семейств предикатов, в рамках которых ведётся поиск закономерностей. Результаты экспериментов позволяют выдвинуть гипотезу, что в реальных задачах классификации эффективная локальная функция роста имеет порядок единицы.

Пусть X — множество объектов, Y — конечное множество возможных ответов, $X^l = (x_i, y_i)_{i=1}^l \subset X \times Y$ — обучающая выборка длины l, по которой строится алгоритм классификации $a : X \to Y$.

Закономерность класса $c \in Y$ — это предикат вида $\varphi_c : X \to \{0,1\}$. Если $\varphi_c(x) = 1$, то говорят, что закономерность φ_c относит объект x к классу c. Информативные закономерности, выделяющие значительную долю объектов только какого-то одного класса, являются исходным сырьём для построения логических алгоритмов классификации. Например, алгоритмы взвешенного голосования типа «Кора» конструируются в виде

$$a(x) = \arg\max_{c \in Y} \sum_{t=1}^{T_c} w_t \varphi_{ct}(x) ,$$

где $\varphi_{ct}(x)$ — закономерности класса c, w_t — неотрицательные веса, $t = 1,\ldots,T_c$. Как правило, закономерности строятся в виде конъюнкций $\beta_1(x) \wedge \ldots \wedge \beta_k(x)$, где термами $\beta_i(x)$ являются условия вида $[x^j = a]$ для номинальных признаков и $[x^j < a]$, $[x^j > a]$, $[a < x^j < b]$ для порядковых и количественных признаков, $(x^1,\ldots,x^n)$ — признаковое описание объекта x. Поиск закономерности по обучающей выборке сводится к оптимизации числа термов k, подмножества признаков и порогов a, b в каждом терме по критерию информативности.

Для оценивания качества (обобщающей способности) закономерностей введём функционал частоты ошибок предиката φ_c на выборке X^l:

$$\nu(\varphi_c, X^l) = \frac{1}{l}\sum_{i=1}^{l}\left[\varphi_c(x_i) \neq [y_i = c]\right].$$

Пусть Φ — семейство предикатов, из которого выбираются закономерности. Обозначим через φ^l закономерность, построенную (найденную) по обучающей выборке X^l. В комбинаторном варианте теории Вапника-Червоненкиса [1] выводится следующая верхняя оценка частоты ошибок на независимой контрольной выборке X^k:

$$P\left\{\nu(\varphi^l, X^k) > \nu(\varphi^l, X^l) + \varepsilon\right\} \leq \Delta(\Phi)\Gamma^l_{l+k}(\varepsilon), \qquad (1)$$

где $\Delta(\Phi)$ — локальная функция роста семейства Φ, $\Gamma^l_{l+k}(\varepsilon)$ — комбинаторный множитель, экспоненциально убывающий с ростом l и k. Заметим, что, в отличие от классического изложения теории [2], левая часть неравенства (1) не содержит супремума по семейству Φ, что позволяет эффективно оценивать её по эмпирическим данным.

Рассмотрим несколько способов оценить функцию роста.

1. *Теоретическая оценка.* Для конечных семейств отображений функцию роста принято оценивать мощностью семейства. Если длина конъюнкций ограничена числом K и j-й признак порождает N_j различных термов, то число различных конъюнкций, составленных из подмножества признаков $\Omega \subseteq \{1,\ldots,n\}$, не превосходит

$$D(\Omega) = H_1(\Omega) + \cdots + H_K(\Omega), \text{ где } H_k(\Omega) = \sum_{\omega\subseteq\Omega:|\omega|=k}\prod_{j\in\omega} N_j .$$

Вычисление числа всех конъюнкций $D(n) = D(\{1,\ldots,n\})$ легко организовать по рекуррентной формуле $H_k(\Omega\cup\{j\}) = H_k(\Omega) + N_j H_{k-1}(\Omega)$, полагая $H_0(\Omega) = 1$ и $H_k(\Omega) = 0$ при $k > |\Omega|$. Величина $D(n)$ представляется сильно завышенной оценкой локальной функции роста, поскольку в процессе решения конкретной задачи классификации никогда не задействуется всё семейство Φ, реально используется лишь некоторое локальное подсемейство, зависящее от исходных данных.

2. *Локальная оценка.* На практике конъюнкции строятся с помощью эффективных эвристических методов сокращённого перебора. В процессе их построения легко подсчитать число просмотренных предикатов, из которых были выбраны наиболее информативные конъюнкции. Однако эта оценка также может оказаться завышенной, поскольку большинство предикатов,

тестируемых в процессе поиска, обладают слишком низкой информативностью и не имеют шансов быть выбранными.

3. *Эмпирическая оценка*. Допустим, что по обучающей выборке X^l строится T закономерностей $\varphi_1^l,\ldots,\varphi_T^l$. Заранее выделив независимую контрольную выборку X^k, можно непосредственно измерить левую часть неравенства (1), следовательно, и функцию роста:

$$\Delta_\varepsilon(\Phi)=\frac{\frac{1}{T}\sum_{t=1}^{T}\left[\nu(\varphi_t^l,X^k)>\nu(\varphi_t^l,X^l)+\varepsilon\right]}{\Gamma_{l+k}^l(\varepsilon)}.$$

В качестве верхней оценки локальной эффективной функции роста принимается максимальное значение $\Delta_\varepsilon(\Phi)$ по всем допустимым ε (Рис. 1), при этом комбинаторный множитель $\Gamma_{l+k}^l(\varepsilon)$ оценивается снизу. Данный способ оценивания функции роста представляется наиболее точным.

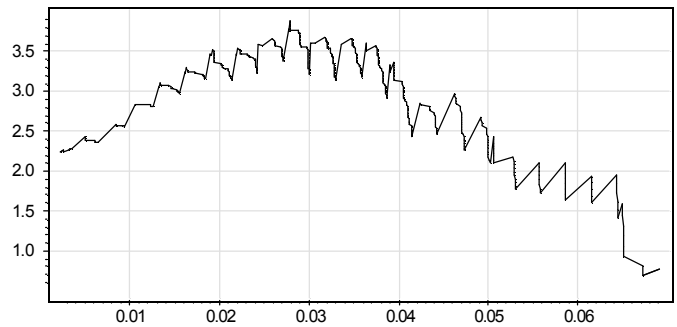

Рис. 1. График зависимости $\Delta_\varepsilon(\Phi)$ от ε, построенный по реальной задаче классификации `crx` из репозитория UCI.

4. *Тривиальная оптимистическая оценка* $\Delta(\Phi)=1$ равносильна гипотезе, что количество предикатов, из которых выбираются закономерности, не влияет на обобщающую способность выбранной закономерности. Согласно общепринятым представлениям теории Вапника-Червоненкиса, данная оценка должна быть сильно занижена.

Для сравнения точности всех четырёх оценок был проведён эксперимент на реальных задачах из репозитория UCI [3]. Результаты в трёх правых колонках таблицы убедительно показывают, что не только теоретические, но даже локальные оценки сильно завышены. Тривиальная оценка лишь слегка занижена и гораздо лучше согласуется с экспериментальными данными.

Интерпретация этого факта возможна следующая. Если логические закономерности объективно присутствуют в данных, то «достаточно разумные» эвристические алгоритмы именно их и находят. Количество объективных закономерностей, как правило, не велико, поэтому эмпирическая функция роста имеет порядок единицы, независимо от того, насколько богато семейство предикатов, в рамках которого ведётся поиск. Верно и обратное: если в процессе поиска эмпирические оценки функции роста существенно превышают единицу, значит, закономерности объективно отсутствуют в данных, либо применяемый алгоритм их не находит.

Задача	число признаков	число термов	объектов		Оценки функции роста		
			обуч.	тест	теоретич.	локальн.	эмпирич.
crx	15	1552	344	346	$1.1 \cdot 10^{11}$	$3.5 \cdot 10^{4}$	3.9
german	24	531	500	500	$5.7 \cdot 10^{9}$	$3.1 \cdot 10^{4}$	1.5
hepatits	19	134	77	78	$1.2 \cdot 10^{8}$	$1.8 \cdot 10^{4}$	2.6
liver	6	885	172	173	$7.9 \cdot 10^{10}$	$2.9 \cdot 10^{4}$	12.1

Обнаруженный эмпирический факт объясняет высокую обобщающую способность логических алгоритмов классификации, ёмкость которых выходит далеко за пределы применимости классических оценок Вапника-Червоненкиса. Обоснование алгоритмов взвешенного голосования типа «Кора» непосредственно вытекает из результатов [4], где показано, что сложность выпуклой комбинации классификаторов равна средней взвешенной сложности отдельных классификаторов. Таким образом, можно полагать, что эффективная локальная функция роста алгоритмов голосования по закономерностям также имеет порядок единицы.

Работа выполнена в рамках проекта РФФИ 05-01-00877 и программы ОМН РАН «Алгебраические и комбинаторные методы математической кибернетики».

Литература

1. Воронцов К. В. Комбинаторный подход к оценке качества обучаемых алгоритмов // Математические вопросы кибернетики. — 2004. — № 13. — С. 5–36.
2. Вапник В. Н., Червоненкис А. Я. Теория распознавания образов. — М.: Наука, 1974.
3. Blake C., Merz C. UCI repository of machine learning databases: Tech. rep.: University of California, Irvine, CA, 1998.
4. Bartlett P. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network // IEEE Transactions on Information Theory. — 1998. — Vol. 44, no. 2. — Pp. 525–536.