Проблема переобучения функций близости при построении алгоритмов вычисления оценок

Воронцов К. В., Ульянов Ф. М.
voron@ccas.ru
Москва, Вычислительный центр РАН

Модель алгоритмов вычисления оценок (АВО) была предложена Ю. И. Журавлёвым в начале 70-х [1]. В данной работе предлагается новый метод обучения АВО, основанная на *принципе явной максимизации отступов* — direct optimization of margin [4]. Известно, что максимизация отступов повышает обобщающую способность линейных композиций классификаторов. АВО как раз и является такой композицией, причём роль базовых классификаторов в ней играют функции близости. Вводится понятие переобученности функций близости и предлагается эмпирическая методика подбора управляющих параметров, позволяющая снижать переобученность функций близости в процессе их построения.

Постановка задачи и используемый вариант АВО

Пусть X — пространство объектов, Y — множество имён классов, $y^*\colon X \to Y$ — целевая функция, значения которой известны только на объектах конечной обучающей выборки $X^\ell = (x_i, y_i)_{i=1}^\ell \subset X \times Y$, $y_i = y^*(x_i)$. Задача заключается в том, чтобы построить алгоритм классификации $a\colon X \to Y$, аппроксимирующий y^* на всём множестве X.

Пусть на множестве X заданы функции расстояния $r_j\colon X \times X \to \mathbb{R}_+$, $j = 1, \ldots, n$, не обязательно метрики. Когда объекты из X описываются n признаками $f_j\colon X \to \mathbb{R}$, можно положить $r_j(x, x') = \bigl|f_j(x) - f_j(x')\bigr|$.

Рассмотрим алгоритм вычисления оценок, имеющий вид

$$a(x) = \arg\max_{y \in Y} \Gamma_y(x); \qquad \Gamma_y(x) = \sum\nolimits_{i \in I_y} B_i(x);$$

где $\Gamma_y(x)$ — оценка объекта x за класс y; $I_y \subseteq \{1, \ldots, \ell\}$ — множество индексов «наиболее типичных» обучающих объектов класса y, называемых *эталонами*; $B_i(x)$ — функция близости, оценивающая сходство объекта x с эталоном x_i по *опорному множеству* $\omega_i \subseteq \{1, \ldots, n\}$. В данной работе функции близости задаются в виде

$$B_i(x) = \bigl[\, \rho_i(x, x_i) \leqslant R_i \,\bigr]; \qquad \rho_i(x, x_i) = \sum\nolimits_{j \in \omega_i} \tfrac{1}{\varepsilon_j} r_j(x, x_i);$$

где ε_j — нормировочные коэффициенты. Каждая функция близости $B_i(x)$ представляет собой *шар* с центром в эталонном объекте x_i и *радиусом* R_i относительно расстояния $\rho_i(x, x_i)$. Будем говорить, что шар $B_i(x)$ *выделяет* объект x, если $B_i(x) = 1$. Итак, алгоритм $a(x)$ задаётся набором параметров $\langle I_y, \omega_i, x_i, R_i, \varepsilon_j \rangle$, где $y \in Y$, $i \in I_y$, $j = 1, \ldots, n$.

Особенностью данного варианта АВО является то, что с каждым эталонным объектом x_i связывается своё (и ровно одно) опорное множество ω_i и свой радиус R_i, однозначно задающие шар $B_i(x)$.

Метод обучения АВО

Ставится задача построить надёжный, хорошо интерпретируемый алгоритм. Для этого шаров должно быть не слишком много; опорные множества ω_i должны иметь невысокую мощность; каждый шар $B_i(x)$ должен быть закономерностью класса y_i, т. е. выделять как можно больше объектов класса y_i и как можно меньше объектов остальных классов; наконец, шар должен обладать обобщающей способностью, т. е. оставаться закономерностью класса y_i на объектах, не вошедших в состав обучения.

Предлагается следующий метод обучения данного варианта АВО.

Сначала вычисляются нормировочные коэффициенты ε_j как среднее значение функции расстояния ρ_j по всем парам обучающих объектов.

Затем начинается поиск «хороших» шаров, реализуемый тремя вложенными циклами перебора. На внешнем цикле обучающие объекты по очереди рассматриваются как кандидаты в эталоны (центры шаров). Для каждого кандидата x_i методом случайного поиска с адаптацией [2] перебираются опорные множества ω_i. Для каждого опорного множества перебираются такие значения радиуса R_i, при которых шар $B_i(x)$ выделяет различные (по составу объектов) подвыборки. Из построенных шаров выбирается тот, который максимизирует критерий $W(B_i)$, определяемый ниже.

Отступом объекта $x \in X$ называется величина

$$M(x) = \Gamma_{y^*(x)}(x) - \max_{y \in Y \setminus \{y^*(x)\}} \Gamma_y(x).$$

Чем больше $M(x)$, тем надёжнее классифицируется объект x. Известно, что оптимальным с точки зрения понижения вероятности ошибки является такое распределение отступов, при котором все они принимают одинаковое и как можно большее значение [4].

Добавление шара $B_i(x)$ изменяет значение отступа $M(x)$ на величину $m(x) = B_i(x)\big(2\,[y_i{=}y^*(x)] - 1\big) \in \{\pm 1, 0\}$. Это изменение поощряется или наказывается путём назначения объекту x веса $w(M(x), m(x))$, где функция $w(M, m)$ задаётся из следующих эвристических соображений:

— увеличение малых (близких к нулю) значений отступа $M(x)$ должно поощряться, уменьшение — наказываться;
— для больших положительных значений отступа, наоборот, уменьшение должно поощряться, увеличение — наказываться, так как это способствует выравниванию распределения отступов;

— для наименьших отрицательных значений отступа увеличение должно поощряться, если ставится задача безошибочно классифицировать обучающую выборку; в противном случае объект может считаться шумовым выбросом, тогда поощряться должно уменьшение отступа.

Критерием качества шара является суммарный вес покрытых им объектов обучающей выборки: $W(B_i) = \sum_{i=1}^{\ell} w(M(x_i), m(x_i))$. Максимизация данного критерия приводит к тому, что каждый следующий шар стремится допустить как можно меньше ошибок, и при этом покрыть объекты, неуверенно классифицируемые композицией всех предыдущих шаров. Одновременно выделяются объекты-выбросы.

Эмпирическое оценивание переобученности шаров

Качество шара $B_i(x)$ на конечной выборке $U \subset X$ характеризуется частотой его ошибок $\nu(B_i, U) = \frac{1}{|U|} \sum_{x \in U} \big[B_i(x) \neq [y^*(x) = y_i]\big]$.

Допустим, что шар $B_i(x)$ был построен по обучающей выборке X^ℓ и имеется непересекающаяся с ней контрольная выборка X^k. *Переобученностью* шара $B_i(x)$ называется разность частоты его ошибок на контроле и на обучении: $\delta(B_i, X^\ell, X^k) = \nu(B_i, X^k) - \nu(B_i, X^\ell)$.

Предлагается методика эмпирического оценивания переобученности, основанная на скользящем контроле. Фиксируется множество разбиений *полной* выборки X^L на обучающую и контрольную, $X^L = X^\ell_n \cup X^k_n$, $L = \ell + k$, $n = 1, \ldots, N$. По каждой обучающей выборке строится АВО. Для каждого шара, полученного в процессе поиска, вычисляются следующие характеристики, зависящие только от обучающей выборки: количество покрытых объектов; количество ошибок; мощность опорного множества; вес шара $W(B_i)$, и др. Исследуется зависимость средней переобученности шаров от этих характеристик с целью понять причины переобучения и скорректировать управляющие параметры алгоритма поиска шаров. Практически во всех экспериментах наблюдались следующие закономерности. Шары, выделяющие меньше объектов, более переобучены. С увеличением числа перебираемых шаров переобученность сначала уменьшается, затем возрастает, однако оптимальная «глубина перебора» в каждой задаче своя. Мощность опорного множества практически не влияет на переобученность.

Результаты экспериментов

В экспериментах на реальных задачах из репозитория UCI качество предложенного алгоритма оказалось сопоставимым с лучшими из известных логических алгоритмов классификации [3]. В задаче распознавания участков генных последовательностей (promoters) точность классификации у АВО примерно втрое лучше, чем у конкурентов, см. Таблицу 1.

Задача	C4.5 Trees	C4.5 Rules	C5.0 Rules	RIP-PER	SLIP-PER	ABO
german	27.5	27.0	28.3	28.6	27.2	**25.5**
australian	18.8	18.8	20.1	**15.2**	15.7	15.8
ionosphere	10.3	10.3	12.3	10.3	7.7	**6.7**
liver	37.5	37.5	31.9	**31.3**	32.3	32.3
promoters	22.7	18.1	22.7	18.1	18.9	**5.5**
breast-cancer	6.6	5.2	5.0	3.7	4.2	**3.6**
hepatitis	20.8	20.0	21.8	19.7	17.4	**16.6**

Таблица 1. Процент ошибочных классификаций при 10-кратном скользящем контроле на 7 реальных задачах из репозитория UCI.

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

Литература

[1] *Журавлёв Ю. И.* Об алгебраическом подходе к решению задач распознавания или классификации // Пробл. кибернетики. — 1978. — Т. 33. — С. 5–68.

[2] *Лбов Г. С.* Методы обработки разнотипных экспериментальных данных. — Новосибирск: Наука, 1981.

[3] *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.

[4] *Mason L., Bartlett P., Baxter J.* Direct optimization of margins improves generalization in combined classifiers: Tech. rep.: Australian National Univ., 1998.