Непараметрический иерархический классификатор для случая многих классов

Добротворский Д. И., Пестунов И. А., Синявский Ю. Н.
pestunov@ict.nsc.ru
Новосибирск, Институт вычислительных технологий СО РАН

В настоящее время проблема выбора информативных признаков в рамках параметрического подхода хорошо изучена, предложен ряд эффективных подходов к ее решению. Однако практически отсутствуют методы выбора признаков для непараметрических классификаторов [2]. В докладе представлен иерархический непараметрический классификатор на основе оценок Розенблатта-Парзена и связанный с ним метод выделения информативных признаков.

Традиционный подход к построению непараметрических правил классификации, основанных на оценках Розенблатта-Парзена, заключается в подстановке в байесовское решающее правило вместо неизвестных вероятностных характеристик классов соответствующих им оценок, полученных по обучающим выборкам [3]. Общий вид этих правил для $(0,1)$-матрицы потерь можно представить выражением

$$\hat{\delta}_0 = \hat{\delta}_0(x;V) = \underset{i\in\{1,\ldots,M\}}{\arg\max}\ q_i\hat{f}_i(x).$$

Здесь $x \in \mathbb{R}^k$; $V = \bigcup_{i=1}^{M} V^{(i)}$ — обучающая выборка объема $N = \sum_{i=1}^{M} N_i$, $V^{(i)} = \left\{x_j^{(i)} \in \mathbb{R}^k \text{ — наблюдение из } i\text{-го класса}\right\}$; q_i, $i = 1,\ldots,M$ — априорная вероятность i-го класса; $\hat{f}_i(x)$ — оценка условной плотности распределения i-го класса $f_i(x)$ в точке $x \in \mathbb{R}^k$, определяемая выражением

$$\hat{f}_i(x) = \frac{1}{N_i c^k}\sum_{j=1}^{N_i}\Phi\left(\frac{x - x_j^{(i)}}{c}\right),$$

где Φ — ядро, c — параметр сглаживания. Для случая двух классов Ω_1 и Ω_2 это правило можно переписать следующим образом:

$$\begin{cases} x \in \Omega_1, & \text{если } \hat{h}(x) = -\ln\frac{\hat{f}_1(x)}{\hat{f}_2(x)} < t, \\ x \in \Omega_2, & \text{в противном случае,} \end{cases}$$

где $\hat{h}(x)$ — непарамерическая оценка функции $h(x) = -\ln(f_1(x)/f_2(x))$, а $t = \ln(q_1/q_2)$ — решающий порог.

В соответствии с методом [2], выделение информативных признаков сводится к нахождению матрицы признаков решающей границы Σ_{DB}

и вычислению ее собственных векторов $(v_1, \ldots, v_k)$, задающих ортонормированный базис пространства признаков. Бóльшим собственным значениям соответствуют более информативные признаки.

Пусть $n(x)$ — единичный вектор нормали к решающей границе S в точке x. Тогда матрица Σ_{DB} определяется следующим образом:

$$\Sigma_{DB} = \int_S n(x)n^{\mathrm{T}}(x)f(x)dx \Big/ \int_S f(x)dx,$$

где $f(x)$ — плотность распределения вектора признаков в точке x.

Нахождение поверхности S и нормалей к ней осуществляется следующим образом. Пусть точки $x^{(1)}$ и $x^{(2)}$ правильно классифицированы и относятся к разным классам. Тогда отрезок, соединяющий эти точки, должен пересекать решающую границу. Поэтому, двигаясь вдоль этого отрезка, можно найти точку $x \in S$. В предлагаемом алгоритме поиск осуществляется методом деления отрезка пополам.

Уравнение байесовской решающей границы можно записать в виде $h(x) = t$. Поскольку функция $h(x)$ в непараметрическом случае неизвестна, вектор нормали к S в точке x приближенно выражается следующим образом:

$$\bigtriangledown h(x) = \frac{\partial h}{\partial x_1}x_1 + \frac{\partial h}{\partial x_2}x_2 + \cdots + \frac{\partial h}{\partial x_n}x_n \approx \frac{\Delta\hat{h}}{\Delta x_1}x_1 + \frac{\Delta\hat{h}}{\Delta x_2}x_2 + \cdots + \frac{\Delta\hat{h}}{\Delta x_n}x_n.$$

Алгоритм реализации описанного метода, предложенный в работе [2], является вычислительно сложным, что существенно ограничивает его применимость к выборкам большого объема. В работе [1] представлен быстрый алгоритм оценивания матрицы Σ_{DB}, быстродействие которого достигается за счет уменьшения объема выборки, участвующей в построении классификатора.

Выделение признаков в многоклассовом случае является достаточно сложной задачей. Для поиска информативных признаков в случае M классов ($M > 2$) в работе [2] предлагается традиционный способ сведения ее к решению нескольких двухклассовых задач. В этом случае матрица Σ_{DB} определяется по формуле

$$\Sigma_{DB} = \sum_{(\Omega_i, \Omega_j)} q_i q_j \Sigma_{DB}(\Omega_i, \Omega_j). \qquad (1)$$

Такой подход часто приводит к необоснованным вычислительным затратам и снижению качества выбираемых признаков (Рис. 1).

Представленный в докладе непараметрический иерархический классификатор сначала выделяет изолированные группы близких классов

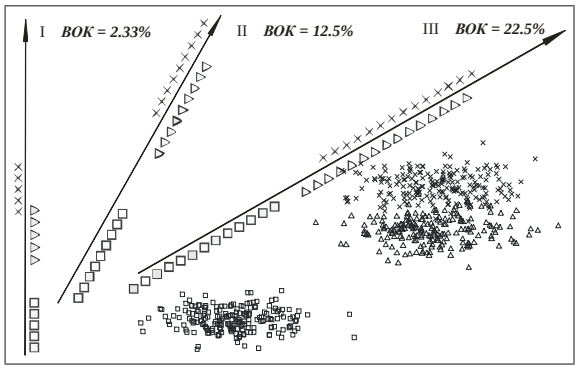

Рис. 1. Двумерная модель, состоящая из трех нормально распределенных классов. Стрелками показан оптимальный информативный признак (III) и признаки, выделенные по методу главных компонент (I) и по формуле (1) (II).

(на Рис. 1 обведены пунктирной линией), затем, при необходимости, разделяет их на более мелкие группы. На каждом уровне иерархии для классификации используется свой минимально достаточный набор информативных признаков, определяемый на основе матрицы (1). Это позволяет снизить трудоемкость алгоритма классификации без потери качества.

Статистическое моделирование на многочисленных модельных и реальных данных показывает, что предлагаемый метод построения непараметрического иерархического классификатора позволяет более чем на порядок сократить объем требуемых вычислений.

Работа выполнена в рамках интеграционного проекта СО РАН и ДВО РАН № 86.

Литература

[1] *Добротворский Д. И., Пестунов И. А.* Быстрый алгоритм извлечения признаков для непараметрического классификатора на основе решающей границы // Межд. конф. ВИТ-2006, Павлодар: ТОО НПФ «ЭКО», 2006. — Т. I. — С. 409–417.

[2] *Lee C., Landgrebe D. A.* Decision Boundary Feature Extraction for Non-Parametric Classification // IEEE Trans. on System, Man and Cybernetics. — 1993. — Vol. 23, N,2. — С. 433–444.

[3] *Харин Ю. С.* Робастность в статистическом распознавании образов. — Минск: Университетское, 1992. — 232 с.