Ученые записки Таврического национального университета им. В.И. Вернадского

Серия "Математика. Механика. Информатика и Кибернетика" N 2 (2002) 128 – 132

УДК 519.9

Донской В.И., Ильченко А.В.

СТАТИСТИЧЕСКОЕ ОБОСНОВАНИЕ ВОЗМОЖНОСТИ ВЫДЕЛЕНИЯ КЛАССОВ ОБЪЕКТОВ В ЭМПИРИЧЕСКИХ ВЫБОРКАХ

ВВЕДЕНИЕ

В задаче автоматической классификации (таксономии) [1] требуется представить эмпирическую выборку $\tilde{X}_q = (X_1, X_2, \ldots, X_q)$ объектов-векторов (точек) в виде совокупности непустых подмножеств-классов $K_1, K_2, \ldots, K_l$ так, что $K_1 \cup K_2 \cup \cdots \cup K_l = \tilde{X}_q$, и классы, вообще говоря, могут пересекаться.

Непосредственное применение к выборке $\tilde{X}_q$ известных алгоритмов кластеризации (классификации) всегда позволяет получить результат, который, к сожалению, может оказаться с содержательной точки зрения лишенным всякого смысла. Действительно, если априори известно, что число классов $l = 2$, то можно построить $2^q - 2$ классификаций, причем никакая из них не имеет содержательной ценности, если, например, выборка извлечена из однородной генеральной совокупности объектов.

Возникает вопрос: когда можно обоснованно приступать к построению классов в любой данной выборке $\tilde{X}_q$?

Выделение классов объектов или наблюдений в выборке, проводимое тем или иным алгоритмом, требует обоснования, проверки "реальности"классов [2]. В настоящей работе предлагается статистический подход к такому обоснованию, основанный на гипотезе классификации, введенной в [3].

Пусть каждое наблюдение в выборке - элементарное событие в выборочном вероятностном пространстве $< \Omega,\ F_1,\ P >$; $\Omega \subset R^n$; $r : \Omega \times \Omega \to R$ - метрика в R^n. В вероятностном пространстве $< \Omega \times \Omega, F_2, P >$ метрика r является измеримой функцией:

$$(\forall x \in R)\ \{(Y, Z) \mid Y \in \Omega,\ Z \in \Omega,\ r(Y, Z) < x\},$$

следовательно, случайная величина $r = r(Y, Z)$ имеет функцию распределения

$$F_r(x) = P\{(Y, Z) \mid r(Y, Z) < x\}.$$

Пусть $F_r(x)$ обладает свойством абсолютной непрерывности:

$$F_r(x) = \int_{-\infty}^{x} f(u)du$$

Тогда $f(r)$ - плотность распределения парных расстояний.

Большинство подходов к решению задачи классификации основано на объединении в классы "близких"(в метрике r) наблюдений. Например, в один класс включают пару наблюдений, если евклидово расстояние между ними не превышает некоторе пороговое значение d. Параметр d обычно является эвристикой и определяет максимально допустимый диаметр подмножества, образующего класс.

Используемая далее гипотеза классификации (в рамках вероятностной модели) гласит [3]: во множестве Ω с метрикой r можно выделить классы, если плотность распределения парных расстояний $f(r)$ имеет более одной моды.

Ниже рассматривается подход к проверке гипотезы классификации (существования более одной моды функции $f(r)$), использующий эмпирическую функцию распределения $\hat{F}_r(x)$ и оценивающий неслучайность обнаружения более одной моды эмпирической плотности $\hat{f}(r)$. Рассмотрение неслучайности как закономерности восходит к фундаментальным работам А.Н.Колмогорова [4] и лежит в основе предлагаемых нами принципов построения классификаторов и других алгоритмов анализа данных [5].

Имея выборку $\tilde{X}_q$, можно построить эмпирическую функцию распределения $\hat{F}_r(x)$ и оценить вероятность наличия более одной моды у плотности $f(r)$.

Множество $\Re = \{r_{jk} = r(X_j, X_k) \mid 1 \le j < k \le q;\ X_j, X_k \in \tilde{X}_q\}$ назовем порожденной выборкой. $\Re$ содержит $q(q-1)/2$ элементов, причем каждый из них независим с $(q^2-5q+6)/2$ элементами и зависим с $2(q-2)$ элементами. Поскольку $2(q-2) = o((q^2-5q+6)/2)$ при $q \to \infty$, то порожденная выборка с ростом числа q элементов в ней ведет себя почти как независимая.

УСЛОВИЕ СУЩЕТСВОВАНИЯ БОЛЕЕ ЧЕМ ОДНОЙ МОДЫ ПЛОТНОСТИ РАСПРЕДЕЛЕНИЯ ПАРНЫХ РАССТОЯНИЙ

Теорема 1. *Пусть функция распределения* $F_r(x) = \int_{\infty}^{x} f(u)du$ *непрерывна, имеет производную* $f(x)$ *в каждой точке отрезка* $[a,b]$ *и заданы точки* $a < x_1 < x_2 < x_3 < b$ *так, что*
$x_2 - x_1 = x_1 - a;\ b - x_3 = x_3 - x_2$. *Тогда при выполнении неравенств*

$$2F_r(x_1) - F_r(a) - F_r(x_2) > 0 \tag{1}$$

$$F_r(b) - 2F_r(x_3) + F_r(x_2) > 0 \tag{2}$$

плотность $f(r)$ *в интервале* (a,b) *имеет локальный минимум.*

Доказательство. Обозначим $\Delta_1 = x_2 - x_1 = x_1 - a,\ \Delta_2 = b - x_3 = x_3 - x_2$. Перепишем неравенства ((1),(2)) в виде

$$F_r(x_1) - F_r(a) > F_r(x_2) - F_r(x_1),$$

$$F_r(b) - F_r(x_3) > F_r(x_3) - F_r(x_2).$$

По теореме Лагранжа

$$\exists \xi \in (a,\ x_1) : F_r(x_1) - F_r(a) = f(\xi)\Delta_1,$$

$$\exists \eta \in (x_1, x_2) : F_r(x_2) - F_r(x_1) = f(\eta)\Delta_1,$$

$$\exists \tau \in (x_2, x_3) : F_r(x_3) - F_r(x_2) = f(\tau)\Delta_2,$$

$$\exists\lambda \in (x_3,\ \ b) : F_r(b\ ) - F_r(x_3) = f(\lambda)\Delta_2.$$

Следовательно для точек $a < \xi < \eta < \tau < \lambda < b$ выполняются неравенства $f(\xi) > f(\eta)$ и $f(\tau) < f(\lambda)$, доказывающие теорему.

Следствие. *Если плотность $f(\xi)$ определена для всех $x \in R$ и выполнены условия теоремы 1, то она имеет более одной моды.*

Замечание. Неравенства в теореме 1 являются условиями, удобными для работы с выборками, и они специально выбраны для этой цели.

При обработке статистических наблюдений в неравенствах ((1),(2)) возможно использование только эмпирической функции распределения $\hat{F}_r(x)$. Обозначим

$$\delta = \sup_{-\infty<x<\infty} |F_r(x) - \hat{F}_r(x)|. \tag{3}$$

Теорема 2. *При выполнении неравенств*

$$2\hat{F}_r(x_1) - \hat{F}_r(a) - \hat{F}_r(x_2) > 4\delta, \tag{4}$$

$$\hat{F}_r(b) - 2\hat{F}_r(x_3) + \hat{F}_r(x_2) > 4\delta, \tag{5}$$

выполняются условия теоремы 1, и плотность $f(x)$ имеет более одной моды.

Доказательство. Неравенство (4) можно переписать в виде

$$2(\hat{F}_r(x_1) - \delta) - (\hat{F}_r(a) + \delta) - (\hat{F}_r(x_2) + \delta) > 0.$$

Из (3) следует, что $F_r(x_1) \geq \hat{F}_r(x_1) - \delta; \hat{F}_r(a) + \delta \leq F_r(a); \hat{F}_r(x_2) + \delta \leq F_r(x_2)$. Поэтому $2F_r(x_1) - F_r(a) - F_r(x_2) > 0$. Аналогично показывается справедливость неравенства $F_r(b) - 2F_r(x_3) + F_r(x_2) > 0$.

ПРОВЕРКА ГИПОТЕЗЫ КЛАССИФИКАЦИИ

Для проверки возможности выделения классов в выборке $\hat{X}_q$, предлагается выполнение следующих этапов.

1^0. Для выбора значений a, x_1, x_2, x_3, b, используемых в неравенствах ((4),(5)), необходимо по выборке $\hat{X}_q$ вычислить порожденную выборку $\Re$ длины $m = q(q-1)/2$ и построить гистограмму $\hat{f}(x)$. Значения a, x_1, x_2, x_3, b выбираются по гистограмме так, что точки x_1 и x_2 делят пополам соответственно отрезки $[a, x_2]$ и $[x_2, b]$.

2^0. Вычислить эмпиричекую функцию распределения $\hat{F}_r(x)$ и величину

$$d = \min\{2\hat{F}_r(x_1) - \hat{F}_r(a) - \hat{F}_r(x_2), \hat{F}_r(b) - 2\hat{F}_r(x_3) + \hat{F}_r(x_2)\}$$

Из теоремы 2 следует, что при выполнении условия

$$\sup_{-\infty<x<\infty} |F_r(x) - \hat{F}_r(x)| < d/4 \tag{6}$$

плотность рапределения парных расстояний имеет более одной моды, и проведение классификации возможно. Неравенство (6) может иметь место с некоторой вероятностью, которую можно оценить при помощи критерия А.Н.Колмогорова [6].

Критерий Колмогорова применяется для проверки непараметрической гипотезы, согласно которой независимые одинаково распределенные случайные величины

$X_1, X_2, \ldots\ X_m$ имеют заданную непрерывную функцию распределения $F(x)$. Согласно теореме Колмогорова

$$P(\sup_{-\infty<x<\infty} |F(x) - \hat{F}_m(x)| < \lambda/\sqrt{m}\ ) \to K(\lambda) \quad \text{при} \quad m \to \infty,$$

где

$$K(\lambda) = \sum_{i=-\infty}^{\infty} (-1)^i e^{-2i^2\lambda^2}.$$

Смысл использования критерия Колмогорова состоит в том, чтобы оценить, какова вероятность *случайного* обнаружения локального минимума в гистограмме плотности парных расстояний, иначе - случайности вычисления значения d. Если неравенство (6) достоверно при найденном значении d, то локальный минимум достоверно существует. Вероятность

$$P(\sup_{-\infty<x<\infty} |F_r(x) - \hat{F}_r(x)| < d/4\ )$$

оценивает меру случайности эмпирически обнаруженного условия достаточности существования локального минимума плотности $f(x)$.

3^0. Найти величину $\lambda = \frac{d\sqrt{m}}{4}$ (она определяется из соотношения $\frac{\lambda}{\sqrt{m}} = \frac{d}{4}$) и значение $K(\lambda)$. Поскольку длина порожденной выборки $m = q(q-1)/2$, то $\lambda = \frac{d\sqrt{2q(q-1)}}{8}$. Чем больше значение λ, тем больше вероятность выпонения гипотезы и существования классов в выборке.

В таблицу 1 сведены расчеты по указанной методике.

Таблица 1.

№ пп.	Минимальное отклонение d	Длина исходной выборки q	$\lambda = \frac{d\sqrt{2q(q-1)}}{8}$	Вероятность существования классов
1	0.04	250	1.7643	0.9960
2	0.05	190	1.6750	0.9925
3	0.10	60	1.0500	0.7400
4	0.10	70	1.2290	0.8980
5	0.10	80	1.4000	0.9600
6	0.10	90	1.5820	0.8960
7	0.10	95	1.6700	0.9924
8	0.10	100	1.7590	0.9958
9	0.20	40	1.3960	0.9603
10	0.20	50	1.7500	0.9956
11	0.30	32	1.6700	0.9924

Из таблицы 1, в частности видно, что обнаружение более чем одной моды в гистограмме плотности распределения парных расстояний при минимальном отклонении $d = 0.1$ позволяет с высокой вероятностью 0.9958 принять гипотезу классификации по выборке длины $q = 100$.

Список литературы

[1] Дюран Б., Оделл П. *Кластерный анализ.* – М.: Статистика, 1977, 128 с.

[2] Орлов А. И. *Некоторые вероятностные вопросы теории классификации.* // В кн. Прикладная статистика (ученые записки по статистике, т. 45) — М.: Наука, 1983, с. 166–179

[3] Донской В. И. *Бинарные отношения, порожденные распределением парных оценок близости, и классификация на их основе.* // Матем. методы распознавания образов. Тезисы всес. конф. ММРО-2 — Ереван: АН Арм.ССР, 1985, с. 61–63

[4] Колмогоров А. Н. *Теория информации и теория алгоритмов.* — М.: Наука, 1987, 304 с.

[5] Донской В. И., Ильченко А. В. *Алгоритмы поиска аналогий в булевых таблицах эмпирических данных* // Искусственный интеллект, вып.2, 2002, с. 99-107

[6] Абезгауз Г. Г., Тронь А. П., Копенкин Ю. Н., Коровина И. А. *Справочник по вероятностным расчетам.* — М.: Воениздат, 1970, 536с.