ОБНАРУЖЕНИЕ ЭМПИРИЧЕСКИХ ЗАКОНОМЕРНОСТЕЙ С ПОМОЩЬЮ ЭВМ
(Вычислительные системы)

1984 год Выпуск 102

УДК 007:62-50

СОГЛАСОВАННАЯ ОЦЕНКА ВАЖНОСТИ ПРИЗНАКОВ И ОБЪЕКТОВ В ЗАДАЧАХ ТАКСОНОМИИ

Ф.Т.Адылова, М.М.Камилов

В методе вычисления оценок большое место занимает задача оценки важности признаков [1]. Один из подходов к ее определению базируется на вычислении изменения меры $\mu(S, K_q)$ принадлежности объекта S к классу K_q при отбрасывании оцениваемого признака.

Задачу оценки объектов по важности можно ставить и решать аналогично задаче оценки важности признаков. Эта аналогия опирается на возможности рассмотрения столбцов таблицы данных (признаков) аналогично строкам (объектам) в задаче оценки признаков. Для этого вводится в рассмотрение некоторое разбиение G множества P признаков на l групп и задается некоторая мера $\mu(P, G_s)$ оценки принадлежности p-го признака ($p \in P$) к S-й группе G_s.

После этого оценка важности объекта S определяется как среднее изменение критерия качества разбиения G признаков при исключении S из множества M объектов на множестве $M \setminus S$. Таким образом, все отличие задачи оценки важности объектов от задачи определения важности признаков целиком зависит от специфики введенного критерия разбиения G. Обычно предполагается, что такой критерий определяет меру "связности" признаков внутри групп: чем они более связаны в смысле заранее выбранного коэффициента связи, тем искомый критерий больше. В этой связи формулируется новая экстремизационная задача группировки признаков, которая отличается от известных тем, что в ней существенно использована специфика признаков как элементов группировки. Эта специфика состоит в инвариантности признака к инверсии его значений. Существо подхода состоит в следующем. Строится функционал $I(G)$, с помощью которого оценивается группировка $G = \{G_1, \dots, G_l\}$ множества P исходных признаков на заданное число l групп. Функционал $I(G)$ строится так, что его значение тем больше, чем более "тесными" в некотором смысле оказываются признаки, оказавшиеся внутри групп разбиения G. Далее строится приближенный алгоритм его экстремизации. Подход, таким образом, оказывается вполне аналогичным подходу к автоматической классификации объектов. Различие состоит в построении такого критерия, который учитывает отличие в представлениях "сходства" между объектами и признаками.

Рассмотрим построение критерия $I(G)$ для случая обработки булевой таблицы $T_{nm} = \|t_{ij}\|_m^n$ данных.

На множестве $P = \{t_1, t_2, \dots, t_n\}$ столбцов T_{nm} определим вектор $\sigma = \{\sigma_1, \dots, \sigma_n\}$ из булевых переменных. С помощью этого вектора зададим преобразование $T_j = \sigma_j t_{\cdot j}$ $(j = 1, n)$ инверсии признака:

$$\sigma_j t_j = \begin{cases} t_{\cdot j}, \text{ если } & \sigma_j = 0, \\ \overline{t}_{\cdot j}, \text{ если } & \sigma_j = 1, \end{cases} \tag{1}$$

где черта над вектором $t_{\cdot j}$ означает, что значения всех его компонент следует заменить на противоположные.

Пусть задано некоторое семейство Ω подмножеств объектов, каждому элементу которого $\omega \in \Omega$ поставим в соответствие характеристическую функцию сходства двух признаков:

$$r_\omega(t_{\cdot j}, t'_{\cdot j}) = \begin{cases} 1, \text{ если } T_{ij} = T'_{ij} \text{ для всех } i \in \omega, \\ 0 - \text{в противном случае.} \end{cases}$$

Функция $f(t_{\cdot j}, t'_{\cdot j})$ близости между двумя столбцами таблицы строится как сумма характеристических функций:

$$f(t_{\cdot j}, t'_{\cdot j}) = \sum_{\omega \in \Omega} r_\omega(t_{\cdot j}, t'_{\cdot j}). \tag{2}$$

Тогда искомый критерий $I(G)$ принимает вид:

$$I(G, \sigma) = \sum_{q=1}^{l} \sum_{j, j \in G_q} f(t_{\cdot j}, t'_{\cdot j}). \tag{3}$$

Задача группировки признаков формируется в данном случае как задача максимизации I одновременно по G и по σ. После группировки признаков по критерию (3) и определения способа вычисления $\mu(P, G_s, M)$, задача нахождения вектора оценок важности объектов в полной аналогии с тем, как находится важность признаков, сводится к подсчету величин:

$$\{\mu(P, G_s, M\setminus S);\ \forall S \in M,\ \ P \in K_s,\ \ s = 1,l\}. \qquad (4)$$

Тогда оценка β_j для j-го объекта определяется как величина:

$$\beta_j = \sum_{S=1}^{l} \sum_{P \in G_s} |\mu(P, G_s, M) - \mu(P, G_s, M\setminus S)|. \qquad (5)$$

Пусть на таблице $T = \|t_{ij}\|_m^n$ независимо получены оценки $\alpha=(\alpha_1,\dots$ $\dots, \alpha_n)$ и $\beta = (\beta_1,\dots,\beta_m)$ соответственно важности признаков и объектов. Каждый из этих рядов ассоциирует другие оценки:

$$\beta'_j = \sum_{i=1}^{m} t_{ij} \cdot \alpha_i; \qquad \alpha_i = \sum_{j=1}^{m} t_{ij} \cdot \beta_j. \qquad (6)$$

Возникает естественная задача - найти такую пару векторов (α, β), которая или в точности удовлетворяет соотношениям (6) или, если это невозможно, обеспечивает минимум невязки несоблюдения этих соотношений. Эту задачу мы будем называть задачей согласования оценок важности объектов и признаков. Алгоритм решения, который предлагается в настоящем докладе, состоит в построении следующей итерационной процедуры:

I) на основе матрицы T и разбиений K и G ее строк и столбцов находятся начальные векторы (α^0, β^0);

2) на t-м шаге итерации строятся две матрицы:

$$\left.\begin{aligned} T_\alpha^t &= \|t_{ij} \cdot \beta_j^{t-1}\|, \\ T_\beta^t &= \|t_{ij} \cdot \alpha_i^{t-1}\|. \end{aligned}\right\} \qquad (7)$$

По матрице T_α^t вычисляется вектор α^{t+1}, а по матрице T_β^t - вектор β^{t+1}. Алгоритм прекращает работу после выполнения заданного числа α итераций.

В качестве искомой пары (α^*, β^*) выбирается такая, которая обеспечивает минимум критерию невязок: $I(\alpha^*, \beta^*) = \min\limits_{\alpha,\beta} \{\sum_i (\alpha_i - -\sum_j t_{ij} \cdot \beta_j)^2 + \sum_j (\beta_j - \sum_i t_{ij} \alpha_i)^2\}$ из заданного множества пар $\{(\alpha^0, \beta^0),$ $(\alpha^1, \beta^1), \dots, (\alpha^\alpha, \beta^\alpha)\}$.

Литература

1. ЖУРАВЛЕВ Ю.И. и др. Алгоритмы вычисления оценок и их применение. -Ташкент: ФАН, 1974.

Поступила в ред.-изд.отд.
22 марта 1984 года