АКАДЕМИЯ НАУК СССР
СИБИРСКОЕ ОТДЕЛЕНИЕ
АВТОМЕТРИЯ

№ 2 1978

# ОБРАБОТКА ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

УДК 519.95 : 681.3.06

**Ю. П. ДРОБЫШЕВ, В. В. ПУХОВ**

*(Новосибирск)*

## ЭКВИВАЛЕНТНЫЕ ПРЕОБРАЗОВАНИЯ ТАБЛИЦ ЭМПИРИЧЕСКИХ ДАННЫХ

В большом числе случаев данные эксперимента можно рассматривать как совокупность $p$-мерных наблюдений (или измерений) над одним или многими объектами. Такие данные удобно представлять таблицей $T$ типа «объект-признак», где имеются $m$ строк (объектов) и $p$ столбцов (признаков). Весьма часто обработка подобных данных может быть сведена к преобразованию начальной таблицы $T$. Например, при анализе данных часто предпочитают перейти от $T$ к матрице главных компонент $Y$ с помощью специального преобразования $c: T=Yc^*$, где * — знак транспонирования. Используют также и факторизацию более общего вида:

$$T=AU+V.$$

Здесь $U$ — таблица преобразованных данных (новых координат). Очевидно, что при различных преобразованиях таблиц данных некоторые их свойства должны оставаться инвариантными. Так, в методе главных компонент сохраняются эвклидовы расстояния.

Приведенный пример относится к количественным данным (признакам в метрических шкалах). В более общих случаях, когда признаки могут быть как количественными, так и качественными или когда они измерены в разнородных шкалах (метрических, порядковых, номинальных и др.), преобразования таблиц менее разработаны. Также мало исследованы разложения таблиц на составляющие (декомпозиции таблиц), что имеет смысл, если обработка вторичных таблиц окажется проще, а хранение их — более удобным.

В данной работе рассматривается задача преобразования таблиц, структура признаков в которой задана через меры сходства градаций признаков. Преобразование заключается в переходе в новое признаковое пространство, причем меры сходства для градаций новых признаков являются линейными формами от прежних мер сходства, а степень влияния каждого начального признака на новые учитывается с помощью некоторого присвоенного этому признаку положительного числа, условно называемого информационным весом признака.

**1. Таблицы. Преобразования таблиц.**

1.1. Под таблицей $T$ будем понимать в дальнейшем конструкцию следующего вида:

$$T=\langle M;\ \{\mathscr{X}_i\};\ f_T\rangle,\ i\in I.$$

Здесь $\mathscr{X}_i = (X_i;\ \mu_i;\ \rho_i)$ — $i$-й признак; $I$ — конечное множество индексов признаков; $X_i$ — конечное множество градаций признака; $\mu_i$ — мера сходства * градаций признаков в $X_i \times X_i$ ($\mu_i : X_i \times X_i \longrightarrow [0, 1]$) и характеризуется свойствами:

$$\text{а)}\ \mu_i(x, y) = \mu_i(y, x),\ \forall\ x, y \in X_i;$$
$$\text{б)}\ \mu_i(x, x) = 1,\ \forall\ x \in X_i;$$

$\rho_i$ — информационный вес признака $\mathscr{H}_i$ в общей системе признаков $\{\mathscr{H}_i\}_{i \in I}$:

$$\rho_i > 0,\ \sum_{i \in I} \rho_i = 1.$$

Функция $f$ предназначена для выделения множества $\Omega_T$ — области допустимых значений таблицы $T$:

$$M \subseteq \Omega_T \subseteq \prod_{i \in I} X_i;$$
$$f : \prod X_i \to \{0,1\};$$
$$f(x) = \begin{cases} 1, & x \in \Omega_T; \\ 0, & x \overline{\in} \Omega_T; \end{cases}$$

$M$ — множество конкретных реализаций объектов или строк таблицы.

1.2. Пусть имеются две таблицы $T$ и $T'$ с одинаковой мощностью множеств $M$ и $M'$. Будем считать, что таблицы $T$ и $T'$ (и соответственно множества $M$ и $M'$) неразличимы, если существует отображение

$$F : T \longrightarrow T', \tag{1}$$

удовлетворяющее следующим условиям (условиям неразличимости):

а) существует $F_1$, устанавливающее взаимно-однозначное соответствие между подмножествами $\{J_k\}$ и $\{J'_k\}$, $1 \leqslant k \leqslant s$, множеств $I$ и $I'$, $I = \bigcup_{k \leqslant s} J_k$, $I' = \bigcup_{k \leqslant s} J'_k$:

$$F_1 : J_k \to J'_k;$$

б) существует $F_2$:

$$F_2\ (\Omega_T)_k \to (\Omega_{T'})_k,$$

отображающее для каждого $k \leqslant s$ множество допустимых градаций $(\Omega_T)_k$ в аналогичное множество $(\Omega_{T'})_k$ таблицы $T'$, здесь $(\Omega_W)_k$ есть проекция множества $\Omega_W$ на пространство признаков $J_k$ таблицы $W$;

в) $\rho_k = \sum_{i \in J_k} \rho_i = \sum_{i \in J'_k} \rho'_i = \rho'_k,\ \forall k \leqslant s;$

г) $\sum_{i \in J_k} \rho_i \mu_i(x, y) = \sum_{i \in F_1(J_k)} \rho'_i \mu_i(F_2(x), F_2(y)),\ \forall x, y \in (\Omega_T)_k.$

Преобразование (1) при условиях а—г будем называть основным и обозначать $F = \langle F_1, F_2 \rangle$.

Заметим, что если для таблиц $T_1$ и $T_2$ существует основное преобразование $F : T_1 \longrightarrow T_2$, то и существует преобразование $G : T_2 \longrightarrow T_1$, являющееся основным и обратным к $F$. Тем самым на множестве всех

* Принят также термин «расплывчатая толерантность» [1] или более общий — «нечеткое отношение» [2].

таблиц можно ввести отношение неразличимости, называемое также отношением толерантности.

Ниже рассмотрим два частных случая основного преобразования: операции свертки и разложения признаков.

1.3. Пусть задано некоторое множество $I^0 \subseteq I$. Система признаков $\{\mathscr{X}_i\}$, $i \in I_0$, может пониматься как некоторый интегральный признак $\mathscr{X}_{I^0}$, имеющий множество градаций $X_{I^0}$, меру сходства $\mu_{I^0}$, информационный вес $\rho_{I^0} = \sum_{i \in I_0} \rho_i$. Здесь $X_{I^0} = (\Omega_T)_{I^0}$ — градации нового признака, куда входят только те элементы из $\prod_{i \in I_0} X_i$, которые имеют непустые проекции на $\Omega_T$. Мера сходства $\mu_{I^0}$ по подмножеству $I_0$ вводится аналогично мерам по всему множеству признаков [3]:

$$\mu_{I^0}(x, y) = \sum_{i \in I^0} \frac{\rho_i}{\rho_{I^0}} \mu_i((x)_i, (y)_i), \quad \forall x, y \in X_{I^0}.$$

Преобразования такого типа назовем свертками и обозначим $S$. Таблица $T$ при этом переходит в таблицу $T'$, и это преобразование удовлетворяет условиям неразличимости пп. а—г.

Пример 1.

$$T = \langle M; \{\mathscr{X}_i\}; f\rangle, \quad I = \{1, 2\},$$

где

$$\mathscr{X}_1 = \langle X_1, \mu_1, \rho_1\rangle; \quad X_1 = \{A, B, C\};$$

$$\mu_1 = \begin{matrix} A \\ B \\ C \end{matrix}\begin{bmatrix} 1 & 0.1 & 0 \\ 0.1 & 1 & 0.8 \\ 0 & 0.8 & 1 \end{bmatrix}; \quad \rho_1 = 0{,}3;$$

$$\mathscr{X}_2 = \langle X_2; \mu_2; \rho_2\rangle; \quad X_2 = \{a, b\};$$

$$\mu_2 = \begin{matrix} a \\ b \end{matrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}; \quad \rho_2 = 0{,}7; \quad f \equiv 1;$$

$$M = \{(A, a), (A, b), (B, a), (C, b)\}.$$

Найдем свертку признаков $\mathscr{X}_1$ и $\mathscr{X}_2$. Таблица $T$ переходит в таблицу

$$T' = \langle M'; \{\mathscr{X}'\}; f'\rangle,$$

где градациями признака $\mathscr{H}'$ является множество

$$\{Aa, Ab, Ba, Bb, Ca, Cb\},$$

а мера сходства $\mu'$ задается матрицей

$$\mu' = \begin{matrix} Aa \\ Ab \\ Ba \\ Bb \\ Ca \\ Cb \end{matrix}\begin{bmatrix} 1 & 0{,}3 & 0{,}73 & 0{,}03 & 0{,}7 & 0 \\ & 1 & 0{,}03 & 0{,}73 & 0 & 0{,}7 \\ & & 1 & 0{,}3 & 0{,}94 & 0{,}24 \\ & & & 1 & 0{,}24 & 0{,}94 \\ & & & & 1 & 0{,}3 \\ & & & & & 1 \end{bmatrix}.$$

Здесь $M' = \{Aa, Ab, Ba, Cb\}$, $f' \equiv 1$.

1.4. Рассмотрим теперь операцию разложения $R$ признака как обратную к свертке. При этом какой-то из признаков исходной таблицы замещается группой вторичных признаков таким образом, что определенные свойства первичного признака: мера сходства, мощность градаций

и информационный вес — сохраняются. Первичный признак является интегральным по отношению ко вторичным и получается их них операцией свертки.

Будем считать, что признак $\mathscr{X}_{i_0}$ разлагается в произведение признаков $\{\mathscr{X}_j^{i_0}\}$, $j \in J$, если

$$\rho_{i_0} = \sum_{j \in J} \rho_j^{i_0};$$

для каждой градации в $X_{i_0}$ существует единственный набор градаций в $\prod_{j \in J} X_j^{i_0}$, и меры сходства в $\mathscr{X}_j^{i_0}$ связаны соотношениями

$$\mu_{i_0}(x, y) = \sum_{j \in J} \frac{\rho_j^{i_0}}{\rho_{i_0}} \mu_j^{i_0}((x)_j, (y)_j), \quad \forall x, y \in X_{i_0}.$$

Преобразование $R$ таблицы $T$ приводит к таблице

$$T = \langle M'; \{\mathscr{X}_i'\}_{i \in I'}; f' \rangle,$$

где $f'$ равна единице на наборе градаций из $\Omega_{T'}$, полученном из набора множества $\Omega_T$ заменой градации из $X_{i_0}$ соответствующей группой.

Признаки $\{\mathscr{X}_i'\}$ для $i \in I \setminus \{i_0\}$ идентичны, а признак $\mathscr{X}_{i_0}$ замещается совокупностью признаков $\{\mathscr{X}_j^{i_0}\}$, $j \in J$. Нетрудно видеть, что преобразование $R$ удовлетворяет условиям пп. а—г.

Пример 2. Пусть исходной таблицей для преобразования служит таблица $T'$ в примере 1:

$$T' = \langle M'; \mathscr{X}'; f' \rangle,$$

где

$$M' = \{Aa, Ab, Ba, Cb\};$$
$$\mathscr{X}' = (X', \mu'), \; X' = \{Aa, Ab, Ba, Bb, Ca, Cb\}; \; f' \equiv 1.$$

Таблица

$$T'' = \langle M'', \{\mathscr{X}_1'', \mathscr{X}_2''\}, f'' \rangle$$

получена из таблицы $T'$ операцией $R$, где

$$\mathscr{X}_1'' = (X_1'', \mu_1'', \rho_1''),$$
$$X_1'' = \{p_1, p_2\}; \; \rho_1'' = 0{,}7;$$
$$\mu_1'' = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix};$$
$$X_2'' = \{q_1, q_2, q_3, q_4, q_5, q_6\}; \quad \rho_2'' = 0{,}3;$$
$$\mu_2'' = \begin{array}{c} \\ q_1 \\ q_2 \\ q_3 \\ q_4 \\ q_5 \\ q_6 \end{array} \begin{array}{c} \begin{array}{cccccc} q_1 & q_2 & q_3 & q_4 & q_5 & q_6 \end{array} \\ \begin{bmatrix} 1 & 0{,}1 & 0 & 1 & 0{,}1 & 0 \\ & 1 & 0{,}8 & 0{,}1 & 1 & 0{,}8 \\ & & 1 & 0 & 0{,}8 & 1 \\ & & & 1 & 0{,}1 & 0 \\ & & & & 1 & 0{,}8 \\ & & & & & 1 \end{bmatrix} \end{array};$$
$$f'' = \begin{cases} 1, & \mathscr{X} \in \Omega_{T''}; \\ 0, & \mathscr{X} \overline{\in} \Omega_{T''}, \end{cases}$$

где

$$\Omega_{T''} = \{p_1q_1, p_1q_2, p_1q_3, p_2q_4, p_2q_5, p_2q_6\};$$
$$M'' = \{(p_1, q_1), (p_2, q_4), (p_1, q_2), (p_2, q_6)\}.$$

Отображение $F_2$ задается соответствиями:

$$Aa \longrightarrow (p_1, q_1); \qquad Bb \longrightarrow (p_2, q_5);$$
$$Ab \longrightarrow (p_2, q_4); \qquad Ca \longrightarrow (p_1, q_3);$$
$$Ba \longrightarrow (p_1, q_2); \qquad Cb \longrightarrow (p_2, q_6).$$

**2. Эквивалентные преобразования таблиц.** Будем считать, что таблицы $T$ и $T'$ находятся в отношении $\omega$, если существует последовательность таблиц $T_k (1 \leqslant k \leqslant m)$:

$$T = T_1 \xrightarrow{\alpha_1} T_2 \xrightarrow{\alpha_2} \ldots \xrightarrow{\alpha_{m-1}} T_m = T', \tag{2}$$

в которой для любого $k$, $\alpha_k$ есть $R$ или $S$.

2.1. Легко показать, что для любой таблицы $T$ верно $T\omega T$.

2.2. Пусть справедливо $T\omega T'$, тогда верно $T'\omega T$. В самом деле, пусть дано (2). Рассмотрим цепочку преобразований

$$T' = T_m \xrightarrow{\beta_{m-1}} T_{m-1} \xrightarrow{\beta_{m-2}} \ldots \xrightarrow{\beta_1} T_1 = T,$$

где $\beta_k$ определена как обратная операция к операции $\alpha_k$ и, следовательно, $\beta_k$ есть $R$ или $S$.

2.3. Если $T\omega T'$ и $T'\omega T''$, то $T\omega T''$.

В самом деле, если верны цепочки

$$T = T_1 \xrightarrow{\alpha_1} T_2 \xrightarrow{\alpha_2} T_3 \xrightarrow{\alpha_3} \ldots \xrightarrow{\alpha_{m_1-1}} T_{m_1} = T'$$

и

$$T' = T'_1 \xrightarrow{\beta_1} T'_2 \xrightarrow{\beta_2} \ldots \xrightarrow{\beta_{m_2-1}} T'_{m_2} = T'',$$

то верна и цепочка

$$T = T_1 \xrightarrow{\alpha_1} T_2 \xrightarrow{\alpha_2} \ldots \xrightarrow{\alpha_{m_1-1}} T_{m_1} \xrightarrow{\beta_1} T'_2 \xrightarrow{\beta_2} \ldots \xrightarrow{\beta_{m_2-1}} T'_{m_2} = T'',$$

где $\alpha_i, \beta_j \in \{R, S\}$ и, значит, $T\omega T''$.

На основании пп. 2.1, 2.2, 2.3 заключаем, что отношение $\omega$ есть отношение эквивалентности и тем самым с каждой таблицей данных $T$ связан класс таблиц $[T]$, эквивалентных данных. Общая задача преобразования исходной таблицы $T$ ставится следующим образом.

В классе $[T]$ найти таблицу $T'$ более простой структуры, чем таблица $T$. Здесь мы не беремся сформулировать критерий простоты и будем считать только, что он существует, по крайней мере, в большинстве конкретных случаев и сравнение таблиц по этому критерию может быть выполнено. В следующем разделе рассмотрен именно такой случай.

Пример 3. Таблицы $T$, $T'$ и $T''$ из примеров 1 и 2 эквивалентны.

**3. Разложение признаков в произведение номинальных, связь с задачей группирования.** Для некоторых задач обработки данных наиболее простым представлением может считаться такое, где признаки являются номинальными, или, другими словами, их матрицы сходства единичны, т. е.

$$\mu(x, x) = 1 \text{ и } \mu(x, y) = 0, \; x \neq y,$$

где $x$, $y$ — градации признака.

Представление признака $\mathscr{X}$ в виде произведения двух признаков $\mathscr{Y}$ и $\mathscr{Z}$ будем записывать как

$$\mathscr{X} \overset{\tau}{=} \mathscr{Y} \times \mathscr{Z}.$$

Здесь $\tau$ — функция соответствия градаций признака $\mathscr{X}$ и пар градаций признаков $\mathscr{Y}$ и $\mathscr{Z}$.

Признак $\mathcal{Y}$ будем искать в классе номинальных и называть аппроксимирующим. В основе выделения аппроксимирующего простого признака лежит группировка градаций признака $\mathcal{X}$. Группировка задается функцией $K(x)$, ставящей в соответствие каждой градации $x$ номер содержащей ее группы. Множество значений функции $K(x)$ есть градации признака $\mathcal{Y}$.

Назовем группировку допустимой, если:

а) $\mu_x(x, y) \neq 1$ для $x$, $y$ из разных групп;

б) $\mu_x(x, y) \neq 0$ для $x$, $y$ из одной группы.

С введенным понятием допустимой группировки тесно связано понятие нетранзитивного контура, т. е. будем считать, что градации $a_{i_1}, a_{i_2}, \ldots, a_{i_p}$ составляют нетранзитивный контур, если $\mu_{\mathcal{X}}(a_{i_1}, a_{i_2}) = \mu_{\mathcal{X}}(a_{i_2}, a_{i_3}) = \ldots = \mu_{\mathcal{X}}(a_{i_{p-1}}, a_{i_p}) = 1$ и $\mu_{\mathcal{X}}(a_{i_1}, a_{i_p}) = 0$. Очевидно, что функция $K(x)$ только тогда будет задавать допустимую группировку, когда на множестве градаций признака $\mathcal{X}$ нет нетранзитивного контура.

Сформулируем теперь более точно задачу выделения аппроксимирующего номинального признака.

Требуется найти два неотрицательных числа $c_1$ и $c_2$, $c_1 + c_2 = 1$, определить признак $\mathcal{Z}$ и функцию соответствия $\tau : X \longrightarrow Y \times Z$, такие, что

$$\mu_{\mathcal{X}}(x, y) = \underset{\forall x, y \in X}{c_1 \mu_{\mathcal{Y}}} ((\tau x)_Y, (\tau y)_Y) + c_2 \mu_{\mathcal{Z}}((\tau x)_z, (\tau y)_z) \qquad (3)$$

и $\mu_{\mathcal{Y}}$ выражено единичной матрицей.

Решение этой задачи может быть получено на основе следующего утверждения.

Теорема 1. Если множество градаций признака $\mathcal{X}$ не содержит нетранзитивного контура, то разложение (3) существует.

Доказательство. Рассмотрим какое-либо разбиение множества $X$, задаваемое функцией $K$. Такая функция существует, так как нетранзитивного контура нет. Матрица сходства признака $\mathcal{Y}$ единична, размерность ее положим равной мощности множества значений функции $K$. Размерность признака $\mathcal{Z}$ примем равной размерности признака $\mathcal{X}$. Функцию соответствия $\tau$ зададим следующим образом:

$$\tau(a_i) = \langle K(a_i), z_i \rangle, \quad a_i \in X.$$

После этого получаем систему уравнений (4), связывающих величины $c_1$, $c_2$, $\{\mu_{\mathcal{Z}}\}$:

$$\begin{cases} \mu_{\mathcal{X}}(a_i, a_j) = c_1 \delta(K(a_i), K(a_j)) + c_2 \mu_{\mathcal{Z}}(z_i, z_j); \\ c_1 + c_2 = 1; \\ c_1 \geqslant 0; \\ c_2 \geqslant 0, \end{cases} \qquad (4)$$

где

$$\delta(K(a_i), K(a_j)) = \begin{cases} 1, & K(a_i) = K(a_j); \\ 0, & K(a_i) \neq K(a_j). \end{cases}$$

Матрица $\{\delta\}$ есть матрица сходства $\mu_{\mathcal{Y}}$ номинального признака $\mathcal{Y}$. Система (4) имеет $r$ равенств и $r+1$ неизвестную величину:

$$r = m_{\mathcal{X}}(m_{\mathcal{X}} - 1)/2 + 1$$

($m_{\mathcal{X}}$ — мощность множества $X$). Поэтому если система (4) имеет решение, то одна из неизвестных величин может быть принята в качестве параметра.

Рассмотрим из (4) те равенства, где $\delta(K(a_i, K(a_j))=1$. Поскольку $c_2\mu_{\mathscr{Z}}(z_i, z_j)\geqslant 0$, то $c_1\leqslant\mu_{\mathscr{X}}(a_i, a_j)$.

Для уравнений, где $\delta(K(a_i), K(a_j))=0$ и $c_2\neq 0$, получаем

$$\mu_{\mathscr{X}}(a_i, a_j)/c_2=\mu_{\mathscr{X}}(a_i, a_j)/(1-c_1)\leqslant 1.$$

Отсюда следует $c_1\leqslant 1-\mu_{\mathscr{X}}(a_i, a_j)$ (случай $c_2=0$ соответствует номинальному признаку $\mathscr{Z}$). Общей областью определения $c_1$ служит интервал

$$0\leqslant c_1\leqslant c^*=\min_{(i,\,j)}\{\mu_{\mathscr{X}}(a_i, a_j),\ 1-\mu_{\mathscr{X}}(a_i, a_j)\},$$

где $\mu_{\mathscr{X}}(a_i, a_j)$ выбирается при $\delta(K(a_i), K(a_j))=1$ и $1-\mu_{\mathscr{X}}(a_i, a_j)$ выбирается при $\delta(K(a_i), K(a_j))=0$. В силу предположения о допустимости разбиения $K$ получаем, что $c^*>0$, при этом

$$\mu_{\mathscr{Z}}(z_i, z_j)=\begin{cases}\dfrac{\mu_{\mathscr{X}}(a_i, a_j)-c_1}{1-c_1} & \text{для}\quad \delta(K(a_i), K(a_j))=1;\\ \dfrac{\mu_{\mathscr{X}}(a_i, a_j)}{1-c_1} & \text{для}\quad \delta(K(a_i), K(a_j))=0.\end{cases}$$

Теорема доказана.

Таким образом, построение наилучшего аппроксимирующего признака $\mathscr{Y}$ сводится к выбору в качестве $c_1$ числа $c^*$, которое выражает собой степень сходства признаков $\mathscr{Y}$ и $\mathscr{X}$.

Пример 4. Подвергая дальнейшему разложению признак $\mathscr{X}'_2$ в примере 2, можно получить следующее разложение признака $\mathscr{X}'$:

$$\mathscr{X}'\overset{\tau}{=}\mathscr{Y}_1\times\mathscr{Y}_2\times\mathscr{Y}_3\times\mathscr{Y}_4,$$

где каждый из признаков номинален, при этом функция вложения задается следующим образом:

$$Aa\xrightarrow{\tau}p_1t_1s_1m_1;$$
$$Ab\xrightarrow{\tau}p_2t_1s_1m_1;$$
$$Ba\xrightarrow{\tau}p_1t_2s_1m_2;$$
$$Bb\xrightarrow{\tau}p_2t_2s_1m_2;$$
$$Ca\xrightarrow{\tau}p_1t_2s_2m_3;$$
$$Cb\xrightarrow{\tau}p_2t_2s_2m_3,$$

где
$$Y_1=\{p_1, p_2\};\quad \rho(Y_1)=0{,}7;$$
$$Y_2=\{t_1, t_2\};\quad \rho(Y_2)=0{,}24;$$
$$Y_3=\{s_1, s_2\};\quad \rho(Y_3)=0{,}03;$$
$$Y_4=\{m_1, m_2, m_3\};\quad \rho(Y_4)=0{,}03.$$

Рассмотренную процедуру можно продолжить, разлагая признак $\mathscr{Z}$ и так далее до появления на каком-то шаге номинального остаточного признака или признака с нетранзитивным контуром так, как это сделано в примере 4. Кроме того, можно считать, что чем ближе $c_1=c^*$ к единице, тем лучше признак $\mathscr{Y}$ аппроксимирует признак $\mathscr{X}$.

Аппроксимация в данном случае заключается в замене матрицы близости $\mu_{\mathscr{X}}$ матрицей $\mu'_{\mathscr{X}}$, состоящей только из нулей и единиц,

причем

$\mu'_{\mathscr{X}}(a_i, a_j) = 1$, если $a_i$ и $a_j$ лежат в одной группе,

$\mu'_{\mathscr{X}}(a_i, a_j) = 0$, если $a_i$ и $a_j$ принадлежат разным группам.

Перестановкой строк и столбцов $\mu'_{\mathscr{X}}$ приводится к блочно-диагональному виду, где внутри блоков стоят единицы, а вне их — нули. Матрица близости признака $\mathscr{Y}$ получается из матрицы $\mu'_{\mathscr{X}}$ факторизацией по отношению эквивалентности, задаваемой функцией $K(x)$. Каждой аппроксимации признака $\mathscr{X}$ соответствует своя блочно-диагональная матрица сходства $\mu_{\mathscr{Y}}$. Класс таких матриц обозначим через $D$.

Рассмотрим на множестве всех матриц размерности $m_{\mathscr{X}}$ метрику $d$, задаваемую соотношением

$$d(\mu, \mu') = \max_{ij} |\delta_{ij} - \delta'_{ij}|.$$

Свойства метрики проверяются тривиально. Между расстоянием $d$ и коэффициентом информативности $c_1 = c^*$ существует тесная связь, где $c^*$ и $d$ вычисляются для некоторой аппроксимирующей матрицы $\mu_{\mathscr{Y}}$ и соответствующего признака $\mathscr{Y}$ в теореме 1.

Следующая теорема устанавливает связь двух задач: поиска наилучшего аппроксимирующего признака $\mathscr{Y}$ для данного признака $\mathscr{X}$ и аппроксимации матрицы $\mu_{\mathscr{X}}$ соответствующей матрицей $\mu_{\mathscr{Y}}$. Отметим, что при построении матрицы $\mu_{\mathscr{Y}}$ функция группировки $K(x)$ выступает как параметр. Задача же поиска наилучшей аппроксимации по множеству допустимых разбиений здесь не рассматривается.

Теорема 2. Для каждой допустимой группировки, задаваемой функцией $K(x)$, имеет место равенство

$$c^*(\mathscr{X}, K) + d(\mu_{\mathscr{X}}, \mu(K)) = 1,$$

где $c^*(\mathscr{X}, K) = c_1 = c^*$ из теоремы 1, а $\mu(K)$ есть блочно-диагональная матрица для данной группировки $K$.

Доказательство.

Так как $\mu(K) \in D$, то

$$|\delta_{ij} - \delta'_{ij}| = \begin{cases} \delta_{ij}, & K(a_i) \neq K(a_j); \\ 1 - \delta_{ij}, & K(a_i) = K(a_j), \end{cases}$$

где $\delta_{ij} = \mu_{\mathscr{X}}(a_i, a_j)$, а $\delta'_{ij}$ — элемент матрицы $\mu(K)$.

Положим

$$I_{ij} = \begin{cases} 1, & \text{если } K(a_i) = K(a_j); \\ 0, & \text{если } K(a_i) \neq K(a_j). \end{cases}$$

Тогда имеем

$$|\delta_{ij} - \delta'_{ij}| = (1 - I_{ij})\delta_{ij} + I_{ij}(1 - \delta_{ij}),$$

$$d(\mu_{\mathscr{X}}, \mu(K)) = \max_{(ij)} \{(1 - I_{ij})\delta_{ij} + I_{ij}(1 - \delta_{ij})\}.$$

Запишем теперь величину $c_1 = c^*$ из теоремы 1 в новых обозначениях и преобразуем ее:

$$c^*(\mathscr{X}, K) = \min_{(ij)} \left\{ \begin{matrix} \delta_{ij}, & 1 - \delta_{ij} \\ K(a_i) = K(a_j), & K(a_i) \neq K(a_j) \end{matrix} \right\} =$$

$$= \min_{(ij)} \{I_{ij}\delta_{ij} + (1 - I_{ij})(1 - \delta_{ij})\} = \min_{(ij)} \{1 + I_{ij}\delta_{ij} - I_{ij} -$$

$$- \delta_{ij} + I_{ij}\delta_{ij}\} = 1 + \min_{(ij)}\{-(1 - I_{ij})\,\delta_{ij} - I_{ij}(1 - \delta_{ij})\} =$$

$$= 1 - \max_{(ij)}\{(1 - I_{ij})\,\delta_{ij} + I_{ij}(1 - \delta_{ij})\} = 1 - d(\mu_{\mathscr{X}}, \mu(K)).$$

Следовательно, теорема доказана.

Пусть теперь имеется некоторый признак $\mathscr{X}$ и его разложение в произведение номинальных признаков $\{\mathscr{Y}_i\}$. Имеется также функция соответствия $\tau$, осуществляющая вложение множества градаций признака $\mathscr{X}$ в $\prod_{p=1}^{p} \mathscr{Y}_p$.

Если теперь расположить признаки по убыванию их информационных весов $\rho$, то, отбрасывая признаки с конца, будем получать аппроксимации исходного признака $\mathscr{X}$ через номинальные с различной степенью точности.

Пример 5. Рассмотрим разложение признака $\mathscr{X}'$ в примере 4.

| | | | | |
|---|---|---|---|---|
| *Aa* | $p_1$ | $t_1$ | $s_1$ | $m_1$ |
| *Ab* | $p_2$ | $t_1$ | $s_1$ | $m_1$ |
| *Ba* | $p_1$ | $t_2$ | $s_1$ | $m_2$ |
| *Bb* | $p_2$ | $t_2$ | $s_1$ | $m_2$ |
| *Ca* | $p_1$ | $t_2$ | $s_2$ | $m_3$ |
| *Cb* | $p_2$ | $t_2$ | $s_2$ | $m_3$ |
| $\rho_i$ | 0,7 | 0,24 | 0,03 | 0,03 |

Если необходимо сократить число признаков, то лучше всего удалить $\mathscr{Y}_4$. При этом степень точности аппроксимации характеризуется числом 0,97. Отбрасывая далее признак $\mathscr{Y}_3$, получаем разложение на уровне $\rho = \rho_1 + \rho_2 = 0{,}94$, причем стали неразличимыми градации *Ba* и *Ca*, а также *Bb* и *Cb*. На уровне $\rho = 0{,}7$, что соответствует только одному вторичному признаку, неразличимы градации *Aa*, *Ba*, *Ca* и градации *Ab*, *Bb*, *Cb*.

## ЛИТЕРАТУРА


1. **Орлов А. И.** Проблемы устойчивости и обоснованности решений в теории экспертных оценок.— В кн.: Статистические методы анализа экспертных оценок. М., «Наука», 1977.
2. **Заде Л. А.** Понятие лингвистической переменной и его применение к принятию приближенных решений. М., «Мир», 1976.
3. **Воронин Ю. А.** Введение мер сходства и связи для решения геолого-геофизических задач.— «ДАН», 1971, т. 199, № 5, с. 1011—1014.


*Поступила в редакцию 9 ноября 1977 г.*