МАШИННЫЕ МЕТОДЫ ОБНАРУЖЕНИЯ ЗАКОНОМЕРНОСТЕЙ
(Вычислительные системы)
1981 год Выпуск 88

Удк 519.95:681.3.06

ОБ ИСПОЛЬЗОВАНИИ МАТРИЦ СВЯЗИ ДЛЯ ОДНОВРЕМЕННОГО АНАЛИЗА НОМИНАЛЬНЫХ И КОЛИЧЕСТВЕННЫХ ПРИЗНАКОВ

Б.Г. Миркин

Для анализа качественных признаков автором было предложено [1-3] использовать язык булевских матриц между объектами, что вызвало появление новых методов [1] анализа данных, в том числе методов одновременного анализа признаков, измеренных в количественных и качественных шкалах [2]. В настоящей работе, следуя [3], рассмотрены некоторые теоретические аспекты данного направления.

На множестве N объектов, занумерованных индексами $1,2,\ldots,N$, рассмотрим совокупность n признаков $x^1,\ldots,x^n$, каждый из которых измерен в количественной или номинальной шкале (ранговые признаки сводят к количественным приписыванием стандартных рангов, соответствующих отдельным местам).

Каждый номинальный признак x охарактеризуем квадратной $N \times N$-матрицей $r = (r_{ij})$ с элементами, равными

$$r_{ij} = \begin{cases} 0, & \text{если } x(i) = x(j), \\ 1, & \text{если } x(i) \neq x(j), \end{cases}$$

где $x(i)$ - значение признака x на i-м объекте. Для количественного признака x рассмотрим квадратную $N \times N$-матрицу $x = (x_{ij})$, где $x_{ij} = (x_i - x_j)^2$. Мы рассматриваем не $x_i - x_j$, как в [4,5], а $(x_i - x_j)^2$, поскольку, во-первых, необходимо, как и в номинальном случае, иметь симметричную матрицу для равноправия обоих типов шкал и, во-вторых, квадраты разностей - стандартная характеристика, используемая для математического анализа моделей линейной статистики.

Используя матричный способ представления данных, можно рассматривать методы компонентного и регрессионного анализа, как модели аппроксимации, полностью аналогичные аппроксимационным пост-

роениям метода наименьших квадратов применительно к обычным количественным признакам. При этом основную роль играют скалярные произведения и коэффициенты корреляции матриц связи, рассматриваемых как вектора $N \times N$ -мерного пространства.

Для вычисления коэффициента корреляции матричного представления количественных признаков x и y обозначим через μ_{kl} обычные центральные смешанные моменты $\mu_{kl} = E((x - Ex)^k (y-Ey)^l)$, где E - операция взятия среднего значения.

Как известно, μ_{20} и μ_{02} - дисперсии x и y соответственно, μ_{11} - коэффициент ковариации, а $\mu_{11}/\sqrt{\mu_{20}\,\mu_{02}}$ - коэффициент корреляции.

Обозначим через M_{kl} центральные моменты матричных представлений x и y. Нетрудно видеть, что среднее значение матрицы связи равно удвоенной дисперсии:

$$\frac{1}{N^2}\sum_{i,j}(x_i - x_j)^2 = \frac{2}{N}\sum_i\Big(x_i - \frac{1}{N}\sum_j x_j\Big)^2 = 2\mu_{20}.$$

Аналогично дисперсии M_{20} и M_{02} матричных представлений могут быть выражены через моменты четвертого порядка:

$$M_{20} = 2(\mu_{40} + \mu_{20}^2),\quad M_{02} = 2(\mu_{04} + \mu_{02}^2).$$

Матричный коэффициент ковариации имеет вид

$$M_{11} = 2(\mu_{22} + 2\mu_{11}^2 - \mu_{20}\mu_{02}),$$

так что коэффициент корреляции матриц равен

$$\rho = \frac{M_{11}}{\sqrt{M_{20}\,M_{02}}} = \frac{\mu_{22} + 2\mu_{11}^2 - \mu_{20}\,\mu_{02}}{\sqrt{(\mu_{40} + \mu_{20})^2(\mu_{04} + \mu_{02}^2)}}\,. \qquad (1)$$

Можно проверить, что коэффициент (I) равен I тогда и только тогда, когда $y = \alpha x + \beta$.

Пусть теперь x и y - номинальные признаки с матрицами $r=(r_{ij})$ и $q = (q_{ij})$ соответственно. Среднее значение матрицы r выражается через распределение $(p_1,\dots,p_m)$ значений признака x в виде $\delta(x) = 1 - \Sigma p_s^2$, что равно качественной дисперсии номинального признака и равносильно вероятности ошибки пропорционального прогноза значений x (см. [I]).

Обозначим через $R = \{R_1,\dots,R_m\}$ разбиение множества объектов при признаку x. Нетрудно видеть, что скалярное произведение матриц r и q равно

$$(r,q) = \sum_{i,j} q_{ij} - \sum_{s=1}^{m} \sum_{i,j\in R_s} q_{ij} = N^2[\delta(y) - \sum_{s=1}^{m} p_s^2\delta(y/s)], \quad (2)$$

где $\delta(y/s) = 1 - \sum_t \left(\frac{p_{st}}{p_s}\right)^2$, причем p_{st} - доля объектов, имеющих s-ю градацию x и t-ю градацию y одновременно, так что $p_s = \sum_t p_{st}$.

Величина $\delta(y/s)$ характеризует среднюю ошибку пропорционального прогноза значений y при заданном s (см. [1]).

Тогда коэффициент ковариации матричных представлений признаков x и y равен

$$\frac{1}{N^2}(r,q) - \delta(x)\cdot\delta(y) = \delta(y)(1 - \delta(x)) - \sum_{s=1}^{m} p_s^2\cdot\delta(y/s). \quad (3)$$

Следовательно, коэффициент корреляции матричных представлений номинальных признаков есть

$$\rho = \frac{\delta(y)(1-\delta(x)) - \sum_{s=1}^{m} p_s^2\delta(y/s)}{\sqrt{\delta(x)(1-\delta(y))\delta(y)(1-\delta(x))}}, \quad (4)$$

что близко к формуле, предложенной в [5].

Коэффициент ковариации количественного признака y с номинальным признаком x равен

$$\frac{1}{N^2}\sum_{i,j} r_{ij}\cdot y_{ij} - 2\delta(x)\sigma^2(y) = 2[\sigma^2(y)(1-\delta(x)) - \sum_{s=1}^{m} p_s^2\sigma_s^2(y)], \quad (5)$$

где σ_s^2 - дисперсия признака y в классе R_s, отвечающем s-му значению признака x.

Следовательно, коэффициент корреляции номинального x и количественного y равен

$$\rho = \frac{\mu_{02}(1-\delta(x)) - \sum_{s=1}^{m} p_s^2\sigma_s^2(y)}{\sqrt{\frac{1}{2}\delta(x)(1-\delta(x))(\mu_{04} + \mu_{02}^2)}}. \quad (6)$$

Формулы (I), (4) и (6) задают коэффициенты связи между номинальными и количественными признаками.

Рассмотрим теперь задачу конструирования номинального фактора: максимизировать суммарный коэффициент ковариации

$$\sum_{k=1}^{n} (a^k, r) = (a, r) \qquad (7)$$

по всем матрицам r, соответствующим разбиениям множества объектов. Здесь a^k - центрированная матрица признака x^k, r - центрированная матрица искомого разбиения $R=\{R_1,\dots,R_m\}$, а $a = \sum_k a^k$ - суммарная матрица связей между объектами.

УТВЕРЖДЕНИЕ I. Задача максимизации (7) эквивалентна задаче максимизации суммы связей внутри классов R за вычетом порога $\bar{a}$, равного в данном случае среднему значению связей a_{ij} (см.[3]), т.е.

$$f(R) = \sum_{s=1}^{m} \sum_{i,j \in R_s} (a_{ij} - \bar{a}). \qquad (8)$$

Методы решения задачи (8) описаны в [I].

Аналогично можно сформулировать задачу о конструировании количественного фактора x, максимизирующего функцию

$$\sum_{k=1}^{n} (a^k, x) = \sum_{i,j=1}^{N} a_{ij}(x_i - x_j)^2. \qquad (9)$$

Поскольку функция (9) не зависит от среднего значения x, будем считать его нулевым, т.е. вести максимизацию только по центрированным x. Для корректности задачи (9) необходимо также масштабирующее условие. Потребуем, чтобы дисперсия x была постоянна:

$$\sum_{i=1}^{N} x_i^2 = \text{const}. \qquad (10)$$

Дифференцируя функцию Лагранжа задачи (9)-(I0) по искомым x_k, получаем следующее

УТВЕРЖДЕНИЕ 2. Количественный фактор в смысле (9)-(I0) для произвольной матрицы $a=(a_{ij})$ связей между объектами

конструируется как собственный вектор соответствующие симметрированной и центрированнной матрицы А с элементами $a_{ij} + a_{ji}$ при $i \neq j$ и $-\sum_{k \neq i} (a_{ik} + a_{ki})$ при $i = j$, соответствующий ее максимальному собственному значению.

Это утверждение можно рассматривать как некоторое обоснование известных эвристических рекомендаций о нахождении "весов" объектов как компонент собственного вектора, соответствующего максимальному собственному числу матрицы связи.

В целом полученные результаты свидетельствуют о естественности предложенных процедур и возможности их использования.

Литература

1. МИРКИН Б.Г. Анализ качественных признаков. -М.: Статистика. 1976. - 166 с.

2. МИРКИН Б.Г., ВЫСОЦКАЯ Н.В. и др. Шкалы упорядочения. -В кн.: Моделирование в экономических исследованиях. Новосибирск, 1978, с. 109-119.

3. МИРКИН Б.Г. Об учете признаков разных типов шкал в линейных моделях анализа данных. -В кн.: Математические вопросы анализа данных. Новосибирск, 1980, с.20-31.

4. КЕНДЭЛ М.Дж. Ранговые корреляции. -М.: Статистика. 1975. - 214 с.

5. КУПЕРШТОХ В.Л., ПОЛИЩУК Л.И., ТРОФИМОВ В.А. Коэффициенты корреляций связей между объектами. -В кн.: Модели агрегирования социально-экономической информации. Новосибирск, 1978, с.17-34.

Поступила в ред.-изд.отд.
20 февраля 1981 года