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

УДК 519.248

МЕТОД ПРОГНОСТИЧЕСКИХ ПЕРЕМЕННЫХ

С.У.Жанатауов

Для анализа данных с пропусками в последнее время начали применяться линейные модели методов многомерного анализа данных: множественная линейная регрессия и метод главных компонент [3,4]. В частности, для этой цели использовалась обратная модель метода главных компонент [2,6], которая идейно близка к известному методу ZET-75 (см. [5]). В данной работе предлагается метод оценки пропусков,основанный на модели канонических корреляций [7], когда множество признаков разбито на два множества, причем пропуски имеются только у признаков одного из них. Многочисленные эксперименты показали, что непосредственное использование канонических переменных для исходных данных не дает хороших результатов.Поэтому необходимо предварительное преобразование исходных переменных.

Оказалось, что для этой цели удобно применять метод избыточных переменных [I], согласно которому каждое из двух множеств исходных переменных преобразуется так, чтобы максимизировать связь с другим исходным множеством. Далее излагается соответствующий двухступенчатый метод, который естественно назвать м е т о д о м п р о г н о с т и ч е с к и х п е р е м е н н ы х анализа неполных данных. Каждая пара прогностических переменных дает свою оценку пропущенному значению исходной переменной.Предлагаются два критерия выбора искомой оценки.

Пусть имеем $m \times n$-матрицу центрированных и стандартизованных данных, каждый столбец которой состоит из m значений переменной, а строка - это одно наблюдение над n признаками объекта. Такая матрица имеет вид $Z_{mn} = [Z_1 \mid Z_2]$. Для Z_1 и Z_2 существуют матрицы факторов $U_{mp} = Z_1 A$, $V_{mp} = Z_2 B$ и факторных нагрузок A_{qp},

B_{pp}. Если U - избыточные переменные, т.е. косоугольные факторы, то нагрузка a_{ij} интерпретируется как i-й регрессионный коэффициент при j-м факторе. Если же U - канонические переменные, т.е. ортогональные факторы, то a_{ij} имеет вид коэффициента корреляции между i-й исходной переменной и j-м фактором u_j. Аналогично для V и B. Чем больше b_{kl}^2, тем лучше прогноз k-й исходной переменной через l-й фактор. Ортогональные факторы принимают определенные значения только для определенной выборочной корреляционной матрицы. Набору дисперсий $\Lambda = \mathrm{diag}(\lambda_1,\ldots,\lambda_n)$ ортогональных факторов соответствует бесконечное число корреляционных матриц $R_{nn}^{(l)}$, матриц нагрузок $C_{nn}^{(l)}$, факторов $Y_{mn}^{(t)}$ с дисперсиями Λ, коррелированных переменных $Z_{mn}^{(t,l)} = Y^{(t)}C^{(l)T}$ (см. [2]), причем выполняются соотношения:

$$RC = C\Lambda,\quad C^TC = I,\quad \frac{1}{m}Z^TZ = R,\quad \frac{1}{m}Y^TY = \Lambda,$$
$$1 \le \mathrm{rk}(R) = \mathrm{rk}(Z) = \mathrm{rk}(Y) = \mathrm{rk}(\Lambda) \le n,\ n > 2. \qquad (1)$$

Пусть нам известна оценка корреляционной матрицы R для неполных данных, в первых q столбцах Z_1 которых нет пропусков, а в последних p столбцах имеются пропуски: $q + p = n > 2$. Тогда для $Z = [Z_1 \mid Z_2]$ реализуема матричная модель метода избыточных переменных [I]:

$$[Z_1 \mid Z_2] \Rightarrow (R, A^*, B^*, U^*, V^*, \Phi), \qquad (2)$$

где

$$U^* = Z_1A^*,\ V^* = Z_2B^*,\quad \Phi = \frac{1}{m}U^{*T}V^* = \Phi^T,$$

$$\frac{1}{m}U^{*T}U^* = I_{pp},\quad \frac{1}{m}V^{*T}V^* = I_{pp}.$$

Далее биортогонализируем два набора переменных U^* и V^* с помощью метода канонических переменных, т.е. реализуем матричную модель этого метода:

$$[U^* \mid V^*] \Rightarrow (\Phi, A, B, U, V, \Lambda), \qquad (3)$$

$$U = U^*A,\quad V = V^*B,\ \frac{1}{m}U^TU = I_{pp},\ \frac{1}{m}V^TV = I_{pp},\ \frac{1}{m}U^TV = \Lambda.$$

В методе избыточных переменных максимизируют три функции:

$$\varphi_3(r^2_{uv}) = r^2_{uv},$$

$$\varphi_4(r^2_{Yu}) = \frac{1}{p}\sum_{j=1}^{p} r^2_{y_j u},$$

$$\varphi_5(r^2_{Xv}) = \frac{1}{q}\sum_{i=1}^{q} r^2_{x_i v}$$

при условиях (3), исключая условие $\frac{1}{m}U^TV = \Lambda$. Здесь

$$r_{Yu} = (r_{y_1 u},\dots,r_{y_q u}),\quad r_{Xv} = (r_{x_1 v},\dots,r_{x_q v}),$$

$$u = u_j,\quad v = v_j,\quad y_j = z_{j+q},$$

$$A^* = [a^*_{\cdot 1}|\dots|a^*_{\cdot p}],\quad B^* = [b^*_{\cdot 1}|\dots|b^*_{\cdot p}],$$

$$a^*_{\cdot j} = (a^*_{1j},\dots,a^*_{qj}),\quad b^*_{\cdot j} = (b^*_{1j},\dots,b^*_{pj}).$$

Матрицы A^*_{qp}, B^*_{pp} находятся из решения обобщенной задачи на собственные числа и векторы (см. [1]): $(R_{12}R_{21} - \mu R_{11})A^* = 0$, $(R_{21}R_{12} - \nu R_{22})B^* = 0$, где $\mu = \mathrm{diag}(\mu_1,\dots,\mu_q)$, $\nu = \mathrm{diag}(\nu_1,\dots,\nu_p)$, $\mu \neq \nu$, $p \le q$. В методе канонических переменных максимизируют только φ_3 и находят [7] матрицы нагрузок $\tilde{A}_{pp}$ и $\tilde{B}_{pp}$ из решения другой обобщенной задачи $(\Phi_{12}\Phi_{22}^{-1}\Phi_{21} - \Lambda^2\Phi_{11})\tilde{A} = 0$, $\Phi_{22} = I_{pp}$, $\Phi_{11} = I_{pp}$, $(\Phi_{21}\Phi_{11}^{-1}\Phi_{12} - \Lambda^2\Phi_{22})\tilde{B} = 0$. Детали этих методов изложены в [1,7]. Обозначим $A_{qp} = A^*\tilde{A}$, $B_{pp} = B^*\tilde{B}$. Имея стандартизованные данные Z_1, Z_2 и биортогональные избыточные переменные U_{mp}, V_{mp}, выбираем прогностические переменные (u_j, v_j) такие, что $b_{\cdot j}$ и $a_{\cdot j}$ имеют значимые компоненты с относительно большими абсолютными величинами. Тогда оценка $(i, q+1)$-го элемента из Z при выбранном j имеет вид

$$\tilde{z}_{i,q+1} = \frac{m\lambda_j - s1}{u_{ij}\cdot b_{1j}} - \frac{s2}{b_{1j}}, \qquad (4)$$

где

$$\lambda_j = \frac{1}{m}\sum_{k=1}^{m} u_{kj}\vartheta_{kj},\quad u_{kj} = \sum_{i=1}^{q} z_{ki}a_{ij},\quad \vartheta_{kj} = \sum_{i=1+q}^{n} z_{ki}b_{i-q,j},$$

$$s1 = \sum_{k\neq i} u_{kj}\vartheta_{kj},\quad s2 = \sum_{i=q+2}^{n} z_{ii}b_{i-q,j},\quad j \in \{1,\ldots,p\}.$$

Далее оценки (4) суммируются по $j \in \mathcal{J} \subset \{1,\ldots,p\}$ ($\mathcal{J}$ - множество номеров признаков с пробелами) и усредняются. Этот первый режим требует принятия решений специалистом при выборе значимых нагрузок. При втором режиме выбор прогностических переменных (u_j, v_j) происходит по минимуму погрешности относительно средней по доступным элементам (q+1)-го столбца. Этот режим работает автономно. Метод реализован на языке Фортран-БЭСМ-6 и проверен на реальных и искусственных данных [6] по следующей схеме:

- оценка одиночного элемента z_{ij}, $j = q+1,\ldots,n$, $i=1,m$, без учета самого элемента z_{ij};

- оценка конечного числа "пропущенных" элементов j-го столбца;

- оценка элемента z_{ij} с учетом оценок предыдущих элементов. Для этих случаев имели оценку корреляционной матрицы R програм - мой REGS и по полным данным. Во всех случаях относительная $\delta_{ij} = |z_{ij} - \tilde{z}_{ij}|/|z_{ij}|$ и средняя $\delta_j = \frac{1}{m}\sum_{i=1}^{m}\delta_{ij}$ погрешности были удовлетворительными.

Литература

1. VOLLENBERG A.L.van der. Redundancy analysis - an alternative for canonical correlation analysis.-Psychometrika, 1977,v.42, N 2,p.207-219.

2. ЖАНАТАУОВ С.У. Метод анализа неполных данных. - Новосибирск, 1981. - 15 с. (Препринт/ВЦ СО АН СССР: 257).

3. BUCK S.F. A method of estimation of missing values in multivariate data suitable for use with an electronic computer. - J.Royal Stat.Soc.Ser.B, 1960,N 22,p.302-307.

4. DEAR R.E. A principal component missing data method for multiple regression models.- System.Developm.Corporation, Techn. Report SP-86, 1959.

5. ЗАГОРУЙКО Н.Г., ЕЛКИНА В.Н., ТИМЕРКАЕВ В.С. Алгоритм ZET-75 заполнения пропусков в эмпирических таблицах.-В кн.: Эмпирическое предсказание и распознавание образов (Вычислительные системы, вып. 67). Новосибирск, 1976, с.3-28.

6. ЖАНАТАУОВ С.У. Метод получения выборки с заданными собственными числами ее корреляционной матрицы. -В кн.: Математические вопросы анализа данных. Новосибирск, 1980, с. 62-76.

7. HOTELLING H. Relations between two sets of variates.-Biometrika,1936,v.28,p.321-377.

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