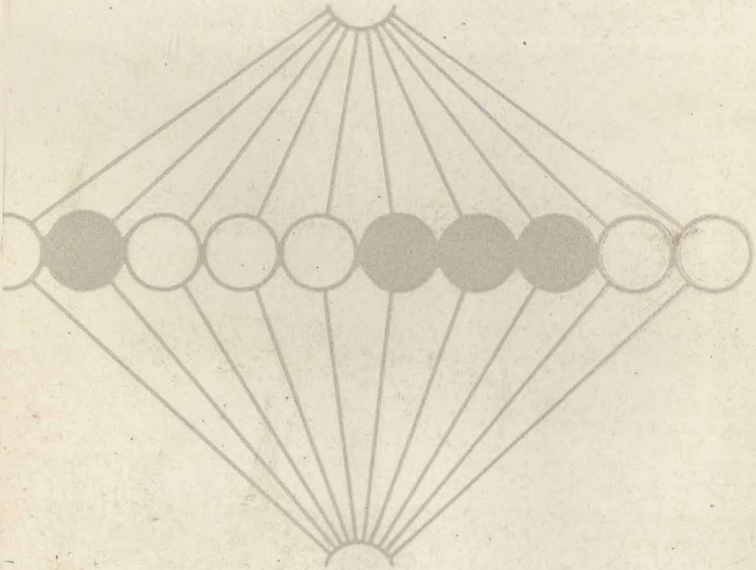
МЕТОДЫ АНАЛИЗА МНОГОМЕРНОЙ ЭКОНОМИЧЕСКОЙ ИНФОРМАЦИИ



ИЗДАТЕЛЬСТВО «НАУКА»
СИБИРСКОЕ ОТДЕЛЕНИЕ

АКАДЕМИЯ НАУК СССР
СИБИРСКОЕ ОТДЕЛЕНИЕ
ИНСТИТУТ ЭКОНОМИКИ И ОРГАНИЗАЦИИ
ПРОМЫШЛЕННОГО ПРОИЗВОДСТВА

МЕТОДЫ АНАЛИЗА МНОГОМЕРНОЙ ЭКОНОМИЧЕСКОЙ ИНФОРМАЦИИ

Ответственный редактор
канд. физ.-мат. наук Б. Г. МИРКИН

ИЗДАТЕЛЬСТВО «НАУКА»
СИБИРСКОЕ ОТДЕЛЕНИЕ
Новосибирск · 1981

Методы анализа многомерной экономической информации.— Новосибирск: Наука, 1981.

В сборнике рассматриваются методы решения задач качественного факторного анализа и их экспериментальное обоснование, модели анализа сложных, в том числе качественных, признаков, построение типологий и классификаций, программное обеспечение обработки многомерной социально-экономической информации.

Книга рассчитана на специалистов в области статистики, прикладной математики и математического моделирования в социально-экономических исследованиях.

М $\frac{10803-854}{042(02)-81}$ 300.81.0604020102.

Б. Г. МИРКИН, И. Б. МУЧНИК

ГЕОМЕТРИЧЕСКАЯ ИНТЕРПРЕТАЦИЯ ПОКАЗАТЕЛЕЙ КАЧЕСТВА КЛАССИФИКАЦИИ

В последнее время для предварительного анализа данных широко используются методы автоматической классификации [1], называемые также методами таксономии [2], кластерного анализа [3], диагонализации матриц связи [4] и т. п. Суть их состоит в том, что рассматриваемая совокупность объектов разбивается на классы, состоящие из объектов, в том или ином смысле похожих друг на друга. Поэтому формальные модели классификации основаны на использовании того или иного уточнения концепции связи или расстояния между объектами для характеристики степени их похожести. Наличие такого уточнения позволяет формулировать задачу классификации как задачу оптимизации некоторого показателя качества разбиения, в котором учитываются характеристики связи объектов внутри классов, а также между классами [1—5].

В то же время представляется естественным трактовать проблему классификации как проблему аппроксимации имеющихся данных в классе наиболее просто устроенных, номинальных, признаков, характеризуемых разбиениями множества объектов. Такой взгляд на проблему классификации объектов по качественным признакам был использован в [5]. При этом для задания как исходных данных, так и результирующего разбиения использовались матрицы связи между объектами, порождаемые соответствующими бинарными отношениями.

Представляется естественным рассмотреть возможности использования аппроксимационного подхода к проблеме классификации непосредственно в терминах исходной таблицы «объект — признак». Далее формулируются и исследуются две аппроксимационные моде-

ли, которые, как оказалось, эквивалентны использованию введенных ранее критериев качества классификации по средней суммарной внутренней связи [4] и суммарной внутренней связи с учетом порога существенности [5, 6]. При этом уточняются оценки связи между объектами и признаками, оценки значимости тех или иных исходных признаков или их значений, а также статистическая и геометрическая природа этих критериев.

Введем соответствующую терминологию.

Пусть имеющиеся данные сведены в таблицу $N \times n$ объект-признак, где N — число объектов, а n — число признаков, так что (i, k)-й элемент таблицы представляет собой значение k-го признака на i-м объекте $(k = 1, \ldots, n;\ i = 1, \ldots, N)$. Будем считать, что среди признаков могут быть как количественные, так и номинальные [5]. Количественный признак может рассматриваться как элемент N-мерного векторного пространства, задаваемый k-м столбцом таблицы объект-признак. Номинальные признаки могут количественно задаваться с помощью булевских матриц. Каждому s-му значению номинального признака x^k сопоставляется булевский столбец таблицы объект-признак, i-й элемент которого равен единице, если значение x^k на i-м объекте равно s, и нулю — в противном случае. Таким образом, номинальный признак с m значениями характеризуется булевской $N \times m$ матрицей, (i, s)-й элемент которой равен единице тогда и только тогда, когда признак x^k принимает на i-м объекте свое s-е значение.

Такое количественное представление номинального признака вполне адекватно в следующем смысле. Пусть φ — произвольное числовое преобразование значений данного номинального признака, сопоставляющее каждому s величину $\varphi(s) = a_s$. Вектор новых значений для данных N объектов, легко видеть, равен Za, где Z — булевская матрица $N \times m$ данного признака, а $a = (a_1, \ldots, a_m)$. Это означает, что множество всех преобразований номинальной шкалы может быть получено в результате умножения матрицы Z на произвольные вектора $a = (a_1, \ldots, a_m)$ и совпадает, по сути дела, с линейным подпространством $L(Z) = \{z / z = Za,$ a — любое$\}$, порожденным столбцами матрицы Z.

С другой стороны, номинальный признак и соответствующая булевская $N \times m$ матрица Z взаимо-однозначно задают разбиение $R = \{R_1, \ldots, R_m\}$ множества рассматриваемых объектов на m классов R_s, состоящих из тех и только тех объектов, на которых признак имеет s-е значение ($s = 1, \ldots, m$).

Таким образом, исходные данные могут быть представлены количественной матрицей X, состоящей из n подматриц X_k размерности $N \times m_k$, $k = 1, \ldots, n$, причем для количественных признаков $m_k = 1$. Проблема состоит в том, чтобы аппроксимировать эту матрицу с помощью булевской $N \times m$ матрицы номинального признака Z, задающей произвольное разбиение $R = \{R_1, \ldots, R_m\}$ множества объектов. Для уточнения проблемы необходимо уточнить способ измерения близости между X и Z.

В качестве меры близости X и Z естественно принять сумму «расстояний» $\alpha(X_k, Z)$ между X_k и Z по $k = 1, \ldots, n$. Расстояние же $\alpha(X_k, Z)$ удобно измерять любым из двух нижеследующих симметричных способов: во-первых, как расстояние от Z (как $(N \times m)$-мерного вектора) до его проекции на подпространство $L(X_k)$, порожденное столбцами матрицы X_k; во-вторых, расстояние от X_k как $N \times m_k$-мерного вектора до его проекции на подпространство $L(Z)$, порожденное столбцами матрицы Z. Напомним, что оператор проектирования на подпространство $L(Y)$ задается формулой

$$P_Y = Y(Y^T Y)^{-1} Y^T$$

для любой матрицы Y, причем при ее вырожденности вместо обращения матриц рассматривается псевдообращение.

Таким образом, в качестве критерия аппроксимации исходных данных с помощью разбиения, соответствующего матрице Z, может рассматриваться любое из выражений

$$\sum_k \| X_k - P_Z X_k \|^2, \tag{1}$$

$$\sum_k \| Z - P_{X_k} Z \|^2, \tag{2}$$

где $\|Y\|$ — евклидовская норма, равная $\sqrt{\sum_{i,j} y_{ij}^2}$.

Задача состоит в том, чтобы минимизировать (1) или (2) в классе матриц Z, соответствующих разбиениям множества объектов. Очевидно, что этот класс состоит из всевозможных булевских матриц, имеющих N строк и попарно ортогональные столбцы.

Проанализируем сначала критерий (1). Прежде всего отметим, что

$$\sum_k \| X_k - P_Z X_k \|^2 = \| X - P_Z X \|^2, \qquad (3)$$

так как оба выражения представляют собой одну и ту же сумму величин $\|x^{ks} - P_Z x^{ks}\|$ для отдельных столбцов x^{ks} матриц X_k по всем s и k.

Формула (3) показывает аналогию между задачей минимизации (1) и задачей компонентного анализа [3]. Задача компонентного анализа состоит в минимизации критерия (3) по всевозможным матрицам Z заданной размерности $N \times m$ с ортонормированными столбцами. Как известно, ее решение дается первыми m собственными векторами матрицы XX^T. Рассматриваемая задача отличается областью минимизации критерия (3), которая состоит из всевозможных булевских матриц Z с ортогональными столбцами.

Показатель (3) может быть интерпретирован как в терминах признаков, так и в терминах объектов. Рассмотрим сначала смысл этого показателя в терминах признаков.

Прежде всего выясним вид матрицы оператора проектирования $P_Z = Z(Z^TZ)^{-1}Z^T$, учитывая специфику матрицы Z. Нетрудно видеть, P_Z матрица размерности $N \times N$, (i, j)-й элемент которой следующим образом характеризует связь объектов i и j: он равен нулю, если i и j находятся в разных классах разбиения R, задаваемого матрицей Z, и равен $1/N_s$, если $i, j \in R_s$ (здесь и далее N_s — число объектов в R_s). Это вытекает из того, что матрица Z^TZ — диагональная с элементами N_s $(s = 1, \ldots, m)$.

Отсюда ясно, что для произвольного столбца x^k матрицы X вектор $P_Z x^k$ имеет в качестве i-й компоненты величину, равную среднему значению $\bar{x}_s^k = \frac{1}{N_s} \sum_{i \in R_s} x_i^k$ признака x^k на объектах класса R_s. В частности, при булевском x^k эта величина равна условной вероятности

того, что объекты из R_s содержатся в множестве, выделяемом столбцом x^k.

Из сказанного вытекает следующее представление:

$$\|X - P_Z X\|^2 = \sum_{k=1}^{M} \sum_{s=1}^{m} \sum_{i \in R_s} (x_i^k - \bar{x}_s^k)^2.$$

Эта величина имеет смысл средневзвешенной дисперсии признаков в классах разбиения R. Действительно, дисперсия столбца x^k ($k = 1, \ldots, M$) в классе R_s определяется формулой

$$\sigma_{ks}^2 = \frac{1}{N_s} \sum_{i \in R_s} (x_i^k - \bar{x}_s^k)^2,$$

так что средняя дисперсия всех столбцов x^k в R_s равна

$$\sigma_s^2 = \frac{1}{MN_s} \sum_{i \in R_s} \sum_{k=1}^{M} (x_i^k - x_s^k)^2.$$

Отсюда следует, что

$$\|X - P_Z X\|^2 = \sum_{s=1}^{m} MN_s \sigma_s^2 = NM \sum_{s=1}^{m} p_s \sigma_s^2, \qquad (4)$$

где $p_s = N_s/N$ — доля объектов в s-м классе разбиения R.

Формула (4) показывает, что задача аппроксимации (1) эквивалентна классической для статистических исследований задаче о построении разбиения с минимальной средней внутриклассовой дисперсией признаков x^k.

С другой стороны, раскрывая скобки в выражении для σ_{ks}^2, получим, что минимизация (4) эквивалентна максимизации взвешенного среднего x^k по классам R_s:

$$\sum_s p_s \bar{x}_s, \qquad (5)$$

где $\bar{x}_s = \frac{1}{M} \sum_{k=1}^{M} \bar{x}_s^k$ — среднее суммарное значение признаков в R_s.

Проанализируем теперь смысл показателя (3) в терминах объектов и связей между ними. Для этого

представим квадрат нормы (3) через операцию умножения матриц. Как известно,

$$\|X - P_Z X\|^2 = T_r[(X - P_Z X)^T (X - P_Z X)], \qquad (6)$$

где через $T_r(A)$ обозначена сумма диагональных элементов — след A.

Раскрывая в (6) скобки, получим, что минимизация (6) эквивалентна максимизации

$$T_r(X^T P_Z X). \qquad (7)$$

Очевидно, что

$$T_r\left(X^T P_Z X\right) = T_r\left(P_Z X X^T\right) = \sum_{ij} p_{ij} \sum_k x_i^k x_j^k,$$

где p_{ij} — элемент матрицы P_Z, равный $1/N_s$ при $i, j \in R_s$ $(s = 1, \ldots, m)$ и нулю при $i \in R_s$, $j \in R_t$ для $s \neq t$. Величина $a_{ij} = \sum_k x_i^k x_j^k$ — элемент матрицы XX^T, который может рассматриваться как характеристика связи объектов i и j по исходным данным. При булевских столбцах x^k, очевидно, a_{ij} есть количество признаков, имеющих одинаковые значения на объектах i и j. Эта мера связи часто используется из эвристических соображений.

Таким образом, минимизация (1) эквивалентна задаче построения такого разбиения R, которое максимизирует среднюю суммарную связь в классах разбиения R:

$$f(R) = T_r\left(X^T P_Z X\right) = \sum_{s=1}^{m} \frac{1}{N_s} \sum_{i,j \in R_s} a_{ij}. \qquad (8)$$

Этот показатель был предложен для произвольных матриц связи в работе [4] из эвристических соображений. Эффективность данного критерия в задаче классификации подтверждена экспериментально [4].

По аналогии с задачей компонентного анализа ясно, что величина (8) играет роль суммы собственных чисел матрицы XX^T и дает оценку той части суммарной дисперсии исходных признаков, которая учтена в классификации R.

Проанализируем теперь показатель (2) аналогичным образом. Рассмотрим сначала случай, когда все

признаки x^k количественные, так что матрица $P_{X_k} = X_k (X_k^T X_k)^{-1} X_k^T$ имеет вид $\frac{1}{(x^k, x^k)} (x_i^k x_j^k)$. Это означает, что суммарное отклонение (2) может быть переписано через элементы z_i^s матрицы Z следующим образом:

$$\sum_k \| Z - P_{X_k} Z \|^2 = \sum_k T_r \left[(Z - P_{X_k} Z)^T (Z - P_{X_k} Z) \right] =$$

$$= \sum_k T_r \left[Z^T Z - Z^T P_{X_k} Z \right] = nN - T_r \left(Z^T \sum_k P_{X_k} Z \right) =$$

$$= nN - \sum_{k=1}^{n} \sum_{s=1}^{m} \sum_{i=1}^{N} Z_i^s \sum_{j=1}^{N} Z_j^s \frac{x_i^k x_j^k}{(x^k, x^k)}.$$

Это означает, что минимизация (2) эквивалентна максимизации величины

$$\sum_{k=1}^{n} \frac{1}{(x^k, x^k)} \sum_{s=1}^{m} \left(\sum_{i \in R_s} x_i^k \right)^2,$$

которая имеет следующее представление через средние значения $\bar{x}_s^k = \frac{1}{N_s} \sum_{i \in R_s} x_i^k$, $\bar{x}_s = \frac{1}{n} \sum_{k=1}^{n} \bar{x}_s^k$ признаков X^k в классах R_s:

$$\sum_{s=1}^{m} p_s^2 \sum_{k=1}^{n} \frac{1}{(x^k, x^k)} (\bar{x}_s^k)^2 = \sum_{s=1}^{m} p_s^2 \bar{x}_s^2 \qquad (9)$$

(при нормированных x^k, $(x^k, x^k) = 1$), что аналогично показателю (5), только внутриклассовые средние $\bar{x}_s$ взвешиваются здесь величинами p_s^2, а не p_s. Критерий (9) характеризует смысл показателя (2) в терминах признаков x^k.

Для анализа (2) в терминах объектов рассмотрим суммарную матрицу $P = \sum_k P_{X_k}$ связей между объектами. Ее смысл особенно явно раскрывается в случае, когда все исходные признаки — номинальные. Тогда матрица P_{X_k} характеризует разбиение, задаваемое k-м признаком, так что p_{ij}^k равно 0, если x^k принимает разные значения на i и j, и равно $1/N_t^k$, где N_t^k —

частота t-го значения x^k, если значение x^k на i и j равно t. Это означает, что величина $p_{ij} = \sum_k p_{ij}^k$ равна сумме величин, обратных частотам тех значений признаков, по которым i и j одинаковы. Подобный способ измерения связи между объектами предлагался в рамках простой вероятностной модели в [7]. Он несколько более изощрен, чем полученный на основе показателя (1), поскольку учитывает признаки не равноправно, а с весами, обратными частотам соответствующих значений.

Таким образом, задача минимизации (2) эквивалентна задаче максимизации

$$T_r\left(Z^T P Z\right) = T_r\left(P Z Z^T\right) = \sum_{i,j=1}^{N} p_{ij} r_{ij},$$

где r_{ij} — элемент матрицы ZZ^T разбиения R, равный единице или нулю в зависимости от того, принадлежат ли i и j одному классу R или нет, так что

$$T_r\left(Z^T P Z\right) = \sum_{s=1}^{m} \sum_{i,j \in R_s} p_{ij}. \qquad (10)$$

Таким образом, минимизация (2) эквивалентна построению разбиения с максимальной суммой внутренних связей p_{ij}.

В данном случае задача максимизации (10) решается тривиальным образом: в силу неотрицательности p_{ij} максимум (10) достигается, если объединить все объекты в одном классе. Однако это решение носит «тривиальный», «паразитный» характер. Дело в том, что задача максимизации величины $T_r(Z^T P Z)$ при отсутствии ограничений на Z, как уже упоминалось, имеет решением собственные векторы матрицы P. В данном случае при номинальных x^k каждая матрица P_{X_k} имеет тривиальный собственный вектор $U = (1, 1, \ldots, 1)$, поскольку, очевидно, $U \in L(X_k)$ и, следовательно, проекция U на $L(X_k)$ совпадает с U: $P_{X_k} U = U$. Принято считать, что этот вектор не должен рассматриваться в содержательном анализе, поскольку определяется чисто формальными особенностями представления номинальных признаков, при котором сумма столбцов X_k всегда дает U [8].

В данном случае вектор U является собственным для P и в то же время удовлетворяет ограничениям за-

дачи, характеризуя тривиальное разбиение, состоящее из единственного класса, содержащего все объекты. В силу сказанного это решение порождается не особенностями данных, а свойствами их формального представления и должно быть элиминировано. Как и в методе главных компонент, следует перейти к рассмотрению «собственных векторов», соответствующих меньшим собственным значениям. Для этого необходимо рассмотреть матрицу остаточных связей, полученную вычитанием из P матрицы $\lambda \tilde{U}\tilde{U}^T$, где λ — максимально собственное значение P, равное n, а $\tilde{U} = \left(\frac{1}{\sqrt{N}}, \ldots, \frac{1}{\sqrt{N}}\right) = \frac{1}{\sqrt{N}} U$ — собственный вектор U после нормировки $((\tilde{U}, \tilde{U}) = 1)$. Очевидно, что вычитаемые «паразитные» связи одинаковы и равны величине $a = \frac{n}{N} = \frac{1}{N^2} \sum_{i,j} p_{ij}$ - средней связи между объектами.

Таким образом, после элиминирования «паразитной» средней связи мы приходим к задаче об отыскании разбиения R, максимизирующего сумму внутренних связей с вычтенным пороговым значением a:

$$g(R) = \sum_{s=1}^{m} \sum_{i,j \in R_s} (p_{ij} - a). \qquad (11)$$

Смысл критерия (11) как критерия кластерного анализа подробно обсуждается в [5, 6].

ЛИТЕРАТУРА

1. **Аркадьев А. Г., Браверман Э. М.** Обучение машины классификации объектов. М.: Наука, 1971. 192 с.
2. **Загоруйко Н. Г.** Методы распознавания и их применение. М.: Сов. радио, 1972. 206 с.
3. **Айвазян С. А., Бежаева З. И., Староверов О. В.** Классификация многомерных наблюдений. М.: Статистика, 1974. 240 с.
4. **Браверман Э. М. и др.** Диагонализация матрицы связи и выявление скрытых факторов.— В кн.: Проблемы расширения возможностей автоматов. Вып. 1. М., 1971, с. 42—79.
5. **Миркин Б. Г.** Анализ качественных признаков. М.: Статистика, 1976. 166 с.
6. **Куперштох В. Л., Миркин Б. Г., Трофимов В. А.** Сумма внутренних связей как показатель качества разбиения.— Автоматика и телемеханика, 1976, № 3, с. 112—120.
7. **Смирнов Е. С.** Таксономический анализ. М.: Изд-во МГУ, 1969.
8. **Lebart L., Morineau A., Tabard N.** Techniques de la description statistique. Paris, Dunod, 1977. 351 p.

В. А. ТРОФИМОВ

КОНЕЧНЫЙ МЕТОД РЕШЕНИЯ ЗАДАЧИ КАЧЕСТВЕННОГО ФАКТОРНОГО АНАЛИЗА

Все большее распространение в анализе данных получают методы обработки многомерной статистической информации, в которых существенную роль играют качественные свойства [1].

К таким относятся методы решения задач качественного факторного анализа [2, 3], специфика которых состоит в том, что выявляемые закономерности в исходных данных характеризуются с помощью одного или нескольких взвешенных факторов, имеющих качественный характер.

В данной работе предлагается новый метод решения задачи качественного факторного анализа свойств системы взаимосвязанных объектов. Этот метод по сравнению с предложенными ранее [2, 3] учитывает взаимные корреляции факторов, что позволяет гарантировать его сходимость за конечное число шагов для некоторых широко распространенных классов качественных свойств системы взаимосвязанных объектов. Приведен пример решения задачи качественного факторного анализа пространственной неоднородности вариантов животного населения.

ЗАДАЧА КАЧЕСТВЕННОГО ФАКТОРНОГО АНАЛИЗА МАТРИЦЫ СВЯЗЕЙ

Многие свойства системы взаимосвязанных объектов могут быть описаны в терминах матриц связей между объектами (матрицы вида объект — объект [1]). Для этого на декартовом квадрате $\mathfrak{N} \times \mathfrak{N}$ фиксированного множества объектов $\mathfrak{N} = \{x_1, \ldots, x_n\}$ рассматривается некоторая числовая функция a: $\mathfrak{N} \times \mathfrak{N} \to \to R^1$, измеряющая связь любой пары объектов из $\mathfrak{N}$ с

точки зрения определенного свойства[1]. Результаты измерения сводятся в матрицу связей между объектами[2] $A = \| a_{ij} \|_{i,i=1}^{n}$ где $a_{ij} \equiv a(x_i, x_j)$, $i, j = 1, \ldots, n$.

Такой способ представления пригоден также и при изучении качественных свойств системы объектов, т. е. тех, которые обычно связывают с бинарными отношениями на объектах или с соответствующими булевскими матрицами [1]. Здесь числовая функция — индикатор соответствующего бинарного отношения. Если $\sigma \subseteq \mathfrak{N} \times \mathfrak{N}$ — некоторое бинарное отношение, то $S : \mathfrak{N} \times \mathfrak{N} \to \boldsymbol{R}^1$, где

$$S_{ij} \equiv S(x_i, x_j) = \begin{cases} 1, & (x_i, x_j) \in \sigma \\ 0, & (x_i, x_j) \notin \sigma \end{cases}$$

есть индикатор отношения σ, дающий представление соответствующего качественного свойства в виде матрицы[3] $S = \| S_{ij} \|_{i,j=1}^{n}$.

Задача качественного факторного анализа[4] — представить наблюдаемое свойство (матрица A) в виде линейной комбинации качественных свойств (факторов). В основе этой задачи лежит следующая гипотеза: наблюдаемое свойство является слишком сложным для непосредственного анализа и дальнейшего прогноза, существуют простые (к таковым относятся качественные) свойства, которые в значительной мере определяют это свойство, позволяя не только содержательно интерпретировать его изменения в более простых терминах, но и осуществлять прогноз этих изменений по соответствующим изменениям простых свойств.

Итак, в задаче качественного факторного анализа ищется представление матрицы A (наблюдаемое сложное свойство системы взаимосвязанных объектов) в виде

$$A = \sum_{k=1}^{q} \lambda_k S^k, \qquad (1)$$

[1] Всюду в дальнейшем мы отождествляем свойство с его описанием в виде матрицы связи объект — объект.

[2] Диагональные элементы матрицы объект — объект нами не рассматриваются, там, где это не вызывает недоразумений, соответствующие условия на индексы опущены.

[3] Использование чисел 1, 0 для наших целей несущественно. Как будет видно из дальнейшего, допустимо использование любых чисел α, β, где $(\alpha > \beta)$.

[4] Более точно: линейного качественного факторного анализа.

где $S^1, \ldots, S^q$ — матрицы размера $n \times n$, представляющие качественные свойства (факторы) из некоторого возможного набора E (как правило, E фиксируется заранее), $\lambda_1, \ldots, \lambda_q$ — вещественные коэффициенты (веса факторов), а q — общее число этих факторов.

В работе [3] изложен метод решения задачи качественного факторного анализа, основанный на многократном использовании решения задачи линейной аппроксимации матрицы A представителями фиксированного класса [5] качественных свойств E.

Задача 1. Для матрицы A и фиксированного класса E найти $S \in E$ и λ, μ — вещественные, доставляющие минимум функционалу

$$\delta(A, S, \lambda, \mu) = (A - \lambda S - \mu \mathbf{1}, A - \lambda S - \mu \mathbf{1}). \qquad (2)$$

Здесь

$$(B, C) \equiv \sum_{i,j=1}^{n} b_{ij} c_{ij} / (n(n-1)) - \qquad (3)$$

скалярное произведение произвольных числовых матриц B, C размера $n \times n$ в $\boldsymbol{R}^{n(n-1)}$,

n — число объектов в $\mathfrak{N}$,

$\mathbf{1}$ — матрица размера $n \times n$, состоящая из одних единиц.

Метод, изложенный в [3], состоял в следующем.

Метод 1. Рассматривается последовательность матриц $A^1, A^2, \ldots, A^k, \ldots$ При этом $A^1 = A$ и для любого $k = 1, 2, \ldots$, $A^{k+1} = A^k - \lambda_k S^k - \mu_k \mathbf{1}$, где $S^k \in E$, λ_k, μ_k являются решением задачи 1 для матрицы A^k и класса E.

Представители $S^k \in E$ и λ_k (при $k = 1, 2, \ldots$) интерпретируются соответственно как факторы и их весовые коэффициенты.

Матрицы A^k при $k > 1$ интерпретируются как матрицы «остаточных» (неучтенных первыми $k-1$ факторами) связей.

В [3] показано, что метод 1 для некоторых A и E сходится (возможно за бесконечное число шагов), т. е. $(A^k, A^k) \to 0$ при $k \to \infty$ и матрица A, таким образом, представляется в этих случаях в виде $A = \sum_{k=1}^{\infty} (\lambda_k S^k + \mu_k \mathbf{1})$.

[5] В дальнейшем полагаем, что класс E формируют матрицы типа объект — объект, соответствующие качественным свойствам, а не сами свойства.

В данной работе мы изложим простую модификацию метода 1, которая позволит в ряде случаев гарантировать конечную сходимость, т. е. получать решение в виде (1), где $q < \infty$.

МОДИФИЦИРОВАННЫЙ МЕТОД

В модифицированном методе нами используются две задачи. Первая — задача 1 в более простой форме, вторая — задача качественной регрессии [2]. Сформулируем их.

Задача 1 может быть записана в более простом виде, при котором параметр μ отсутствует. Для этого мы используем формулы для оптимальных значений параметров λ, μ в случае, когда A и S фиксированы. Предварительно обозначим $m(B) = (B, \mathbf{1})$, $d^2(B) = (B - m(B)\mathbf{1}, B - m(B)\mathbf{1})$, соответственно среднее и дисперсия произвольной числовой матрицы B размера $n \times n$ для скалярного произведения (3) в $\boldsymbol{R}^{n(n-1)}$. Тогда, согласно обычным соотношениям метода наименьших квадратов [4], минимум (2) при A и S фиксированных достигается для

$$\lambda = (A - m(A)\mathbf{1},\ S - m(S)\mathbf{1})/d^2(S), \qquad \mu = m(A) - \lambda m(S). \tag{4}$$

Будем рассматривать задачу 1, предполагая, что $m(A) = 0$ (матрица A центрирована) и $m(S) = 0$, $d^2(S) = 1$ для любого $S \in E$ (матрица S нормирована [6]).

Задача 1 м (модифицированная). Для фиксированных $A(m(A) = 0)$ и E (класс нормированных матриц, представляющих качественные свойства) найти $S \in E$ и λ вещественное, доставляющие минимум функционалу

$$\delta_1(A, S, \lambda) = (A - \lambda S, A - \lambda S). \tag{5}$$

---

[6] Заметим, что по нормированной матрице S восстанавливается то же самое отношение σ, что и по исходной. В этом случае

$$S(x_i, x_j) \equiv S_{ij} \equiv \begin{cases} \alpha = \sqrt{(n(n-1) - |\sigma|)/|\sigma|}, & (x_i, x_j) \in \sigma, \\ \beta = -\sqrt{|\sigma|/(n(n-1) - |\sigma|)}, & (x_i, x_j) \notin \sigma. \end{cases}$$

Так что вместо чисел 1, 0 используется α, β $(\alpha > \beta)$.

Для фиксированных A и $S \in E$ минимум δ_1 достигается, как легко видеть, при

$$\lambda = (A, S). \tag{6}$$

Ввиду (4), (6), очевидно, минимумы (2), (5) достигаются одновременно. Следовательно, задачи 1 и 1м эквивалентны.

Задача качественной регрессии [2] используется нами в следующем виде.

Задача 2. Для фиксированных A и $S^1, S^2, \ldots, S^p \in E$ найти $\lambda_1, \lambda_2, \ldots, \lambda_p$ вещественные, минимизирующие

$$\Delta(A, S^1, S^2, \ldots, S^p, \lambda_1, \lambda_2, \ldots, \lambda_p) = \left(A - \sum_{k=1}^{p} \lambda_k S^k, A - \sum_{k=1}^{p} \lambda_k S^k\right). \tag{7}$$

Здесь также считаем, что A центрирована, а $S^1, S^2, \ldots, S^p$ нормированы и центрированы.

Перейдем теперь к описанию модифицированного метода.

Метод 2. Рассматривается последовательность матриц $A^1, A^2, \ldots, A^k, \ldots$, где $A^1 = A$ и для любого $k = 1, 2, \ldots, A^{k+1} = A - \sum_{l=1}^{k} \lambda_l^k S^l$. При этом фактор $S^l \in E$ ($l = 1, 2, \ldots, k$) — есть решение задачи 1м для A^l и E (получающийся попутно коэффициент λ_l игнорируется). Коэффициенты $\lambda_1^k, \lambda_2^k, \ldots, \lambda_k^k$ являются решением задачи 2 для A и полученного к данному моменту набора факторов $S^1, S^2, \ldots, S^k \in E$.

Матрицы A^k ($k > 1$) так же, как и в методе 1, интерпретируются как матрицы «остаточных» связей.

Метод 2 отличается от метода 1 тем, что весовые коэффициенты при факторах пересчитываются всякий раз, как только их уже имеющийся набор пополняется новым фактором.

Будем говорить, что метод 2 конечен для A и E, если для последовательности $A^1, A^2, \ldots, A^k, \ldots$ матриц «остаточных» связей существует $p < \infty$ такое, что $(A^k, A^k) = 0$ для всех $k > p$.

Сформулируем условие на A и E, выполнение которого гарантирует конечность метода 2 для A и E.

Назовем матрицу $A \neq \mathbf{0}$ коррелированной с классом E, если существует $S \in E$ так, что $(A, S) \neq \mathbf{0}$.