УДК 519.24

ЭНТРОПИЙНЫЙ СПОСОБ ДЕКОМПОЗИЦИИ МАТРИЦЫ КАЧЕСТВЕННЫХ ОЦЕНОК ПО НОМИНАЛЬНОЙ ШКАЛЕ

А.Б. Архипова, В.М. Белов

В статье рассмотрены основные подходы декомпозиции матриц. Приведено обоснование энтропийного способа декомпозиции матрицы экспертных оценок в социологических исследованиях. Разработанный алгоритм апробирован на тестовом примере.

Ключевые слова: декомпозиция, энтропия, показатель информативности.

Введение

В социологических исследованиях большое значение имеют качественные оценки по номинальной шкале. На основании ответов экспертов, используя стандартные статистические преобразования, формируют корреляционные матрицы, отражающие тесноту взаимосвязи между конкретными показателями. Пример такой матрицы приведен на рисунке 1, где 'н', 'с', 'в' означают низкий, средний, высокий уровень зависимости.

Для более корректной обработки данных в матрицах такого вида применяют процедуры декомпозиции, а именно разбиение на блоки.

В общем виде разбиение m×n матрицы A на блоки выглядит следующим образом:

$$\begin{array}{cc} A=\begin{bmatrix} A_{11} & \dots & A_{1q} \\ \dots & A_{ij} & \dots \\ A_{p1} & \dots & A_{pq} \end{bmatrix} & m_i \\ n_j & \end{array}.$$

Здесь Aij – блок (подматрица) размером mi×nj. Тогда A=(Aij) –блочная матрица размером p×q.

Операции с блочными матрицами идентичны обычным при условии соразмерности блоков. Существует несколько методов приведения матриц к блочному виду. Наиболее известные из них описаны Х. Д. Икрамовым, Р. Тьюарсоном, О. Эстербю, З. Златевым [1-4].

Однако, несмотря на наличие работ в области формирования блочных матриц, методы декомпозиции, основанные на использовании энтропии [5-8] в литературе отмечены не были. Поэтому на сегодняшний день актуально адаптировать существующий энтропийный подход к задаче формирования блочных матриц.

Алгоритм

Алгоритм формирования блочных матриц, основанный на энтропийном подходе, состоит в декомпозиции исходной матрицы по информативности столбцов с максимально различающей способностью.

Под ведущими столбцами будем понимать столбцы с наибольшей информативностью.

Пусть необходимо привести коррелляционную матрицу размером 30×30 к блочному виду с использованием энтропийного подхода путем формирования однородных классов. Выберем, например, уровень однородности 45%.

Обозначим число значений ‘н’ по столбцу j через nj(н), ‘с’ – через nj(с), ‘в’ – через nj(в).

Первый этап, заключающийся в отборе ведущих столбцов, реализуют посредством выполнения последовательных шагов путем вычисления энтропии каждого столбца, суммарной общей и условной энтропии, коэффициента влияния.

Энтропию каждого столбца (H_j) вычисляют следующим образом:

$$H_j=\frac{1}{n}\begin{bmatrix} n\log_3 n-n_j(н)\log_3 n_j(н)- \\ -n_j(с)\log_3 n_j(с)-n_j(в)\log_3 n_j(в) \end{bmatrix},\quad(1)$$

где j=1,…,30.

Причем выбор основания логарифма обусловлен наличием трехуровневых оценок (‘н’, ‘с’, ‘в’), распределенных в исходной матрице.

Полученные результаты позволяют вычислить суммарную общую энтропию ($\sum_{j\neq i} H_{ij}$):

$$\sum_{j\neq i} H_{ij}=\frac{1}{n}\sum_{j\neq i}\left[n\log_3 n-\sum_{k,t=н,с,в} n_{ij}(kt)\log_3 n_{ij}(kt)\right].\quad(2)$$

Следующий шаг реализации алгоритма – это вычисление суммарной условной энтропии ($\sum_{j\neq i} h_{ij}$):

$$\sum_{j\neq i} h_{ij}=\sum_{j\neq i}\left(H_i+H_j-H_{ij}\right).\quad(3)$$

в	с	в	в	в	н	в	в	с	с	с	с	в	с	с	в	в	с	с	в	в	с	в	н	с	н	в	в	в	с
с	в	в	в	н	в	с	н	в	с	н	в	с	с	н	с	н	в	в	с	н	с	в	в	с	н	н	в	в	в
в	в	в	в	н	с	с	в	в	в	н	с	с	н	н	с	н	н	в	н	н	н	н	с	с	с	в	н	н	в
в	в	в	в	в	в	в	н	в	с	н	в	в	н	н	в	с	в	в	в	с	в	в	н	н	с	н	н	н	в
в	н	н	в	в	в	в	в	в	в	с	с	н	н	в	в	с	с	с	н	с	с	в	в	н	в	в	н	н	с
н	в	с	в	в	в	н	н	с	с	с	с	н	н	в	в	с	в	н	в	в	в	в	с	в	в	с	н	н	с
в	с	с	в	в	н	в	с	с	н	с	с	н	н	н	н	с	в	с	в	с	в	с	с	в	н	н	с	с	в
в	н	в	н	в	н	с	в	с	с	в	н	в	в	с	с	н	в	с	с	с	с	с	в	н	с	н	н	в	в
с	в	в	в	в	с	с	с	в	в	с	в	в	в	с	н	в	с	н	с	н	в	в	в	н	в	в	в	в	в
с	с	в	с	в	с	н	с	в	в	в	н	в	н	н	н	н	в	с	с	в	н	в	в	в	с	н	в	с	с
с	н	н	н	с	с	с	в	с	в	в	в	н	в	с	в	в	в	н	с	в	н	н	н	в	с	н	н	с	с
с	в	с	в	с	с	с	н	в	н	в	в	в	в	н	н	в	н	с	с	н	с	н	н	в	в	в	н	с	с
в	с	с	в	н	н	н	в	в	в	н	в	в	в	с	н	с	н	с	с	н	с	н	н	н	в	с	с	с	с
с	с	н	н	н	н	н	в	в	н	в	в	в	в	в	н	с	в	н	с	с	н	в	с	н	с	с	в	с	с
с	н	н	н	в	в	н	с	с	н	с	н	с	в	в	в	н	н	в	в	в	н	в	с	с	с	с	в	н	в
в	с	с	в	в	в	н	с	н	н	в	н	н	н	в	в	н	н	с	н	с	с	н	с	в	в	с	н	н	с
в	н	н	с	с	с	с	н	в	н	в	в	с	с	н	н	в	с	н	в	в	с	с	с	с	н	с	с	н	н
с	в	н	в	с	в	в	в	с	в	в	н	н	в	н	н	с	в	в	н	с	н	с	н	н	с	с	с	н	н
с	в	в	в	с	н	с	с	н	с	н	с	с	н	в	с	н	в	в	с	в	н	с	н	н	в	в	в	н	н
в	с	н	в	н	в	в	с	с	с	с	с	с	с	в	н	в	н	с	в	с	н	в	н	н	с	в	в	н	н
в	н	н	с	с	в	с	с	н	в	в	н	н	с	в	с	в	с	в	с	в	в	в	н	н	с	в	н	в	в
с	с	н	в	с	в	в	с	в	н	н	с	с	н	н	с	с	н	н	н	в	в	в	н	н	с	в	н	с	н
в	в	н	в	в	в	с	с	в	в	н	н	н	в	в	н	с	с	с	в	в	в	в	н	н	н	н	н	в	н
н	в	с	н	в	с	с	в	в	в	н	н	н	с	с	с	с	н	н	н	н	н	н	в	н	н	н	н	н	с
с	с	с	н	н	в	в	н	н	в	в	в	н	н	с	в	с	н	н	н	н	н	н	н	в	н	н	н	н	с
н	н	с	с	в	в	н	с	в	с	с	в	в	с	с	в	н	с	в	с	с	с	н	н	н	в	н	н	н	с
в	н	в	н	в	с	н	н	в	н	н	в	с	с	с	с	с	с	в	в	в	в	н	н	н	н	в	в	в	в
в	в	н	н	н	н	с	н	в	в	н	н	с	в	в	н	с	с	в	н	н	н	н	н	н	н	в	в	с	н
в	в	н	н	н	н	с	в	в	с	с	с	с	с	н	н	н	н	н	в	в	с	в	н	н	н	в	с	в	н
с	в	в	в	с	с	в	в	в	с	с	с	с	с	в	с	н	н	н	в	в	н	н	с	с	с	в	н	н	в

Рисунок 1 – Исходная корреляционная матрица.

Показателем информативности столбца считают величину I_i, представляющую собой:

$$I_i = \frac{\sum_{j \neq i} h_{ij}}{\sum_{j \neq i} H_{ij}}. \qquad (4)$$

Столбец, имеющий максимальное значение I_i, является наиболее информативным по отношению к остальным и позволяет разбить исходную матрицу на классы принадлежности ведущему столбцу.

Приведем расчет первого цикла выявления наиболее информативных столбцов.

По формуле (1) рассчитаем H_j:

H_1 = 0,858672711; H_2 = 0,97949;
H_3 = 0,987781244; H_4 = 0,866277584;
H_5 =0,965401367; H_6=0,979448546;
H_7=0,979448546; H_8 = 0,990543356;
H_9=0,844353685; H_{10}=0,987781244;
H_{11}=1; H_{12}=0,996960792;
H_{13}=0,996960792; H_{14}=1;
H_{15}=0,996960792; H_{16}=0,991159471;
H_{17}=0,972267716; H_{18}=0,996960792;
H_{19} =1; H_{20}=0,990543356;
H_{21} =0,979448546; H_{22} = 0,987781244;
H_{23}=0,930418662; H_{24} =0,918990046;
H_{25} =0,895049359; H_{26}=0,987781244;
H_{27} =0,972267716; H_{28} =0,920619836;
H_{29} =0,965401367; H_{30} =0,987781244.

Далее рассчитаем $\sum_{j \neq i} H_{ij}$ в соответствии с формулой (2):

$\sum_{j=1}^{30} H_{1j}$ =51,0365695; $\sum_{j=1}^{30} H_{2j}$ =54,2863188;

$\sum_{j=1}^{30} H_{3j}$ =53,65196; $\sum_{j=1}^{30} H_{4j}$ =51,3698487;
$\sum_{j=1}^{30} H_{5j}$ = 53,8060192; $\sum_{j=1}^{30} H_{6j}$ = 54,4087631;
$\sum_{j=1}^{30} H_{7j}$ = 54,0650103; $\sum_{j=1}^{30} H_{8j}$ = 55,2017127;
$\sum_{j=1}^{30} H_{9j}$ =50,7472658; $\sum_{j=1}^{30} H_{10j}$ = 54,6761887;
$\sum_{j=1}^{30} H_{11j}$ =54,5811689; $\sum_{j=1}^{30} H_{12j}$ = 54,7437562;
$\sum_{j=1}^{30} H_{13j}$ =53,8541572; $\sum_{j=1}^{30} H_{14j}$ = 54,6627036;
$\sum_{j=1}^{30} H_{15j}$ = 54,8871295; $\sum_{j=1}^{30} H_{16j}$ = 54,209853;
$\sum_{j=1}^{30} H_{17j}$ = 54,1483419; $\sum_{j=1}^{30} H_{18j}$ = 54,926563;
$\sum_{j=1}^{30} H_{19j}$ =55,2574742; $\sum_{j=1}^{30} H_{20j}$ = 54,0227027;
$\sum_{j=1}^{30} H_{21j}$ = 54,060737; $\sum_{j=1}^{30} H_{22j}$ =54,6737598;
$\sum_{j=1}^{30} H_{23j}$ =53,0957704; $\sum_{j=1}^{30} H_{24j}$ =52,8506371;
$\sum_{j=1}^{30} H_{25j}$ =51,4442786; $\sum_{j=1}^{30} H_{26j}$ =54,5974545;
$\sum_{j=1}^{30} H_{27j}$ =54,0808864; $\sum_{j=1}^{30} H_{28j}$ =52,7710078;
$\sum_{j=1}^{30} H_{29j}$ =53,4004809; $\sum_{j=1}^{30} H_{30j}$ =53,8334532.

На основании значений H_j и H_{ij} рассчитаем $\sum_{j \neq i} h_{ij}$ в соответствии с формулой (3):

$\sum_{j=1}^{30} h_{1j}$ = 1,93277618; $\sum_{j=1}^{30} h_{2j}$ = 2,06475034;
$\sum_{j=1}^{30} h_{3j}$ = 2,93242462; $\sum_{j=1}^{30} h_{4j}$ = 1,81243344;
$\sum_{j=1}^{30} h_{5j}$ = 2,15172891; $\sum_{j=1}^{30} h_{6j}$ = 1,94230605;
$\sum_{j=1}^{30} h_{7j}$ = 2,28605877; $\sum_{j=1}^{30} h_{8j}$ = 1,46001107;
$\sum_{j=1}^{30} h_{9j}$ = 1,8211472; $\sum_{j=1}^{30} h_{10j}$ = 1,90819596;
$\sum_{j=1}^{30} h_{12j}$ = 2,09765583; $\sum_{j=1}^{30} h_{13j}$ = 2,98725483;
$\sum_{j=1}^{30} h_{14j}$ = 2,26380624; $\sum_{j=1}^{30} h_{15j}$ = 1,95428251;
$\sum_{j=1}^{30} h_{16j}$ = 2,46912196; $\sum_{j=1}^{30} h_{17j}$ = 2,00166396;
$\sum_{j=1}^{30} h_{18j}$ = 1,91484899; $\sum_{j=1}^{30} h_{19j}$ = 1,6690356;
$\sum_{j=1}^{30} h_{20j}$ = 2,6390211; $\sum_{j=1}^{30} h_{21j}$ = 2,29033209;
$\sum_{j=1}^{30} h_{22j}$ = 1,91062486; $\sum_{j=1}^{30} h_{23j}$ = 1,88246191;
$\sum_{j=1}^{30} h_{24j}$ = 1,80759403; $\sum_{j=1}^{30} h_{25j}$ = 2,5436133;
$\sum_{j=1}^{30} h_{26j}$ = 1,98693017; $\sum_{j=1}^{30} h_{27j}$ = 2,06911949;
$\sum_{j=1}^{30} h_{28j}$ = 1,93285744; $\sum_{j=1}^{30} h_{29j}$ = 2,55726723;
$\sum_{j=1}^{30} h_{30j}$ = 2,75093146.

По формуле (4) рассчитаем показатель I_i
I_1 = 0,037870417; I_2 = 0,038034451;
I_3 = 0,05465643; I_4 = 0,035282048;
I_5 = 0,039990487; I_6 = 0,035698405;
I_7 = 0,042283517; I_8 = 0,026448655;
I_9 = 0,035886607; I_{10} = 0,034899945;
I_{11} = 0,042969782; I_{12} = 0,038317718;
I_{13} = 0,055469345; I_{14} = 0,041414092;
I_{15} = 0,035605479; I_{16} = 0,045547476;
I_{17} = 0,036966302;I_{18} = 0,034861984;
I_{19} = 0,030204703; I_{20} = 0,048850223;
I_{21} = 0,042365906; I_{22} = 0,03494592;
I_{23} = 0,035454084;I_{24} = 0,034201935;
I_{25} = 0,049444046; I_{26} = 0,036392359;
I_{27} = 0,038259719; I_{28} = 0,03662726;
I_{29} =0,047888468;I_{30} =0,05110078.

Проанализировав рассчитанные значения показателя, можно сделать заключение, что ведущим является тринадцатый столбец. В результате можно построить три класса (блока) согласно 'н', 'с', 'в' 13-ого столбца соответственно.

Следовательно, этап выявления наиболее информативных столбцов можно считать реализованным.

Далее следует этап проверки классов на однородность.

Обозначим через $r_j(k)$, j=1,….,m число наиболее частых значений j-ого столбца класса k. Тогда вероятность ошибочной классификации оценивают следующим образом:

$$p_j = \frac{m - \sum_k r_j(k)}{m}.$$

Средняя вероятность ошибочной классификации по всему набору столбцов имеет вид:

$$p = \frac{\sum_j p_j}{m} = \frac{m^2 - \sum_j \sum_k r_j(k)}{m^2}. \quad (5)$$

По формуле (5) средняя вероятность ошибочной классификации составит 0,507.

Следовательно, получили однородность классов – 0,493. Так как был выбран уровень однородности 45% (0,45), то классы являются однородными. В противном случае процесс выявления набора продолжают циклически с переходом на (j+1) столбец.

В общем виде при рассмотрении комбинации из s столбцов для вычисления уровня однородности используют следующие формулы:

$$h_{(i_1,\dots,i_s)g} = H_{i_1,\dots,i_s} + H_g - H_{i_1,\dots,i_s g};$$

$$I_{i_1 \dots i_s} = \frac{\sum\limits_{g \neq i_1 \dots i_s} h_{(i_1 \dots i_s)g}}{\sum\limits_{g \neq i_1 \dots i_s} H_{(i_1 \dots i_s)g}}.$$

Процедура классификации будет завершена, как только достигнут желаемый уровень однородности.

Таким образом, в статье предложен энтропийный способ декомпозиции матрицы качественных оценок по номинальной шкале.

СПИСОК ЛИТЕРАТУРЫ

1. Икрамов, Х. Д. Численные методы линейной алгебры / Х. Д. Икрамов. – М.: Знание, 1987. – 46 с.
2. Икрамов, Х. Д. Разреженные матрицы / Х. Д. Икрамов. – М.: Знание, 1989. – 48 с.
3. Шатихин, Л. Г. Структурные матрицы и их применение для исследования систем / Л. Г. Шатихин. – М.: Машиностроение, 1991. – 253 с.
4. Эстербю, О. Прямые методы для разреженных матриц / О. Эстербю, З. Златев. – М.: Мир, 1987. – 118 с.
5. Бешелев, С. Д. Математико-статистические методы экспертных оценок / С. Д. Бешелев, Ф. Г. Гурвич. – М.:Наука, 1980. – 180 с.
6. Мёллер, Ф. Роль энтропии в номинальной классификации / Ф. Мёллер // Математика (в)? социологии. Моделирование и обработка информации / Под ред. А. Аганбегяна, Х. Блейлока, Ф. Бородкина, Р. Будона, В. Капекки. - М.: Мир, 1977. – 385 с.
7. Панкова, Л. А. Организация экспертизы и анализ экспертной информации / Л. А. Панкова, А. М. Петровский, М. В. Шнейдерман. – М.: Наука, 1984. – 120 с.
8. Шибанов, Г. П. Информационные технологии // Порядок формирования экспертных групп и проведения коллективной экспертизы. – 2003 - №12. – С. 120-122.

*Аспирант **А.Б. Архипова** тел. 8-913-246-76-22, arhipova_ab@mail.ru - каф. Вычислительных систем и информационной безопасности АлтГТУ им. И. И. Ползунова. Профессор **В.М. Белов** тел. (383) 269-83-28, vmbelov@mail.ru - каф. БиУТ Сибирского государственного университета телекоммуникаций и информатики.*

УДК 57.087

ВЫБОР ЗНАЧИМЫХ ДЛЯ ПРОГНОЗИРОВАНИЯ ВРЕМЕНИ ВЫЖИВАНИЯ ПОКАЗАТЕЛЕЙ ПАЦИЕНТА

Д.Ю. Стрункин

В статье рассматривается задача прогнозирования времени выживания пациента. Для эффективного решения данной задачи необходимо среди всех показателей пациента отобрать те, что значимо влияют на время выживания. Приводится методика отбора, основанная на корреляционном анализе и экспертных оценках.

Ключевые слова: корреляционный анализ, экспертные оценки, согласованность мнений, прогнозирование времени выживания, значимость показателей.

Введение

Одной из важных задач в работе специализированных медицинских учреждений является задача оценки состояния пациента и прогнозирование его времени выживания: будет ли пациент жив через год, три, пять или более лет после лечения. Необходимость и важность прогноза времени выживания пациента обусловлена рядом причин. Во-первых, это необходимо для адекватного выбора программы лечения (особенно с точки зрения возможных побочных эффектов, связанных с применением рискованных схем лечения), а во-вторых – это