УДК 330.4

МАТРИЧНАЯ КЛАСТЕРИЗАЦИЯ КАК КЛАСТЕРИЗАЦИЯ МАТРИЦ ОДИНАКОВОЙ РАЗМЕРНОСТИ

MATRIX CLUSTERING AS A CLUSTERING MATRICES OF THE SAME DIMENSION

В.М. Московкин, Казимиру Эринелту
V.M. Moskovkin, Herinelto Casimiro

Белгородский государственный национальный исследовательский университет,
Россия, 308015, г. Белгород, ул. Победы, 85

Belgorod National Research University, 85 Pobeda st., Belgorod, 308015, Russia

E-mail: moskovkin@bsu.edu.ru; herineltocasimiro@hotmail.com

Аннотация
В работе приведен обзор исследований по матричной кластеризации и показано на отсутствие работ по кластеризации матриц одинаковой размерности. Под методами матричной кластеризации обычно понимают извлечение из матриц большой размерности субматриц меньшей размерности, обладающих определёнными свойствами. Для решения поставленной задачи предложено преобразовывать такие матрицы двумя способами в векторы одинаковой длины и проводить их кластеризации уже известными методами. Такую кластеризацию матричных объектов предлагается сопоставлять с их кластеризацией по методу естественных границ для скалярной характеристики матричных объектов. В простейшем случае она рассчитывается по формуле среднеарифметической из нормированных значений элементов исходной матрицы. Приведены примеры семи матричных объектов, которые можно кластеризировать, приводя их к векторам одинаковой размерности. Ввиду того, что кластеризация любых объектов существенно зависит от выбранных метрик и методов кластеризации, поэтому предложено проводить сценарные расчеты в количестве a×b, где a – количество метрик, b – количество методов кластеризации.

Abstract
The paper presents an overview research on matrix clustering and shows lack of works on clustering for matrices of the same dimension. In order to accomplish this problem, it is proposed to convert such matrices into vectors of the same length and carry out their already known clustering methods. Such clustering of matrix objects is proposed to be compared with their clustering according to the method of natural boundaries for the scalar characteristics of matrix objects. In the simplest case, it is calculated according to the formula of the arithmetical mean of the normalized values of the elements of the original matrix. There has been given seven examples of matrix objects which can be clustered leading them to the vectors of the same dimension. In view of the fact that the clustering of any object essentially depends on the selected metrics and clustering techniques, it is therefore suggested to carry out scenario calculations in the number of a×b, where a is the number of metrics, and b is the number of clustering methods.

Ключевые слова: метрики кластеризации, кластеризация матриц одинаковой размерности, методы кластеризации, алгоритмы кластеризации, матричные объекты, векторные объекты, скалярная интегральная характеристика.
Keywords: Cluster Metrics, clustering matrices of the same dimension, clustering methods, clustering algorithms, matrix objects, vector objects, scalar integral characteristics.

Введение

Обычно под методами матричной кластеризации понимают извлечение из матриц большой размерности субматриц меньшей размерности, обладающих определёнными свойствами. Хороший обзор по методам матричной кластеризации дан в работе [Чернышев Г., 2015] в контексте решения задач вертикального фрагментирования в реляционных СУБД. В нем под методами матричной кластеризации понимается выделение фрагментов по матрице использования атрибутов. Отмечается, что это исторически первый способ решения задачи вертикального фрагментирования. Здесь общая идея состоит в составлении, на основании нагрузки матрицы, близости атрибутов и ее кластеризации, то есть приведении ее к блочно-диагональному виду путем перестановки столбцов и строк (Bond Energy Algorithm, BEA). Полученные блоки и будут фрагментами в задаче вертикального фрагментирования. Самая ранняя работа данного типа была опубликована работе [Hoffer J.A., Severance D.G., 1975]. Сам алгоритм BEA был предложен в работе [McCormick W. T., et al., 1972], и он был использован в качестве первого шага в известной работе [Shamkant N. et al., 1984]. Вместе него используются и другие алгоритмы [Stagle J. R. et al., 1975; Bhat M. V., Haupt A., 1974 и др.]. В работе [Gorla N., Boe Boe W. J., 1990] предложена метрика близости атрибутов и новый алгоритм кластеризации. В ней удалось превзойти на отдельных наборах данных стоимостный метод кластеризации (создается модель системы, формируется некоторая стоимостная функция, которая минимизируется) [Hammer, Niamir, 1979].

В дальнейшем по этой тематике вышла серия работ Chun-Hung Cheng [Cheng C. H., 1995; Cheng C. H., Motwani J. 2009; Cheng C. H. et al, 2011], а также работы [Gorla N. 2007; Jindal A., Dittrich J.2012; Jindal A. et al.,2013].

Существует большой кластер работ, в которых аналогичная идеология (выделение блочно-диагональных фрагментов матрицы на основе перестановки ее строк и столбцов) применяется для бинарных матриц [Oyanagi S., Kubota K., Nakase A., 2001a, 2006, 2003; Kuo J.J., Zhang Y. J., 2012; Zhang Y. J., et al., 2010; Nagaraj G. et al., 2015; Qyelade J. et al., 2016]. Здесь под матричной кластеризацией понимается метод извлечения плотных субматриц из базовой разряженной бинарной матрицы с помощью перестановки строк и столбцов. Для такой кластеризации в работе [Oyanagi S., Kubota K., Nakase A.,2003] разработан быстрый Ping-pong алгоритм, который использовался в работе [Kuo J. J., Zhang Y. J., 2012] для задач библиотечного обслуживания, а в работе [Nagaraj G. et al., 2015] использованы Rank Order Clustering алгоритмы (ROC и ROC-2) для моделирования клеточных производственных систем (Cellular manufacturing systems). В наиболее крупном последнем обзоре по алгоритмам кластеризации в приложении к данным экспрессии генов (Gene Expression Data), насчитывающим 173 литературных источника [Qyelade J. et al., 2016] рассмотрен метод бинарной матричной факторизации (Binary matrix factorization), предложенный в работе [Zhang Y. J., et al., 2010]. Но следует отметить, что основная область приложения таких алгоритмов связана с задачами WWW-анализа, Web Usage Mining и анализа изображений.

Из этого обзора работ по матричной кластеризации мы видим, что отсутствуют работы по матричной кластеризации в смысле кластеризации матриц одинаковой размерности. Рассмотрим концептуально эту задачу.

Основная часть

Пусть имеются K матриц одинаковой размерности $m \times n$,

$$\left(a_{ij}^{k}\right) = \begin{vmatrix} a_{11}^{k} & a_{12}^{k} & \dots & a_{1i}^{k} & \dots & a_{1n}^{k} \\ a_{21}^{k} & a_{22}^{k} & \dots & a_{2i}^{k} & \dots & a_{2n}^{k} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ a_{j1}^{k} & a_{j2}^{k} & \dots & a_{ji}^{k} & \dots & a_{jn}^{k} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ a_{m1}^{k} & a_{m2}^{k} & \dots & a_{mi}^{k} & \dots & a_{mn}^{k} \end{vmatrix},$$

где $1 \le k \le K, 1 \le i \le n, 1 \le j \le m.$

Эти матрицы лежат в К-мерном матричном пространстве (или в К –мерном пространстве матриц размерности $m \times n$). Евклидово расстояние между двумя матрицами размерности $m \times n$ $\left(a_{ij}^{k}\right)$ и $\left(a_{ij}^{c}\right)$, где $1 \le k, c \le K$, равняется:

$$d = \left(\left(a_{ij}^{k}\right), \left(a_{ij}^{c}\right)\right) = \left[\frac{1}{m\,n}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(\left(a_{ij}^{k}\right) - \left(a_{ij}^{c}\right)\right)^{2}\right]^{\frac{1}{2}}. \quad (1)$$

Помимо евклидова расстояния можно использовать и множество других метрик. Например, в Mat Lab запрограммировано 9 метрик и 5 алгоритмов кластеризации иерархического вида [Demirel M. C., Kahya E, 2007].

Чтобы кластеризовать K матриц одинаковой размерности, мы предлагаем представить матрицу $\left(a_{ij}^{k}\right)$ в виде вектора длины $m \times n$. Это можно сделать двумя способами, формируя вектор последовательно по строкам матрацы или по ее столбцам.

В первом случае будем иметь:

$$\left(a_{ij}^{k}\right) = \begin{vmatrix} a_{11}^{k} & a_{12}^{k} & \dots & a_{1n}^{k} \\ \vdots & \vdots & & \vdots \\ a_{i1}^{k} & a_{i2}^{k} & \dots & a_{in}^{k} \\ \vdots & \vdots & & \vdots \\ a_{m1}^{k} & a_{m2}^{k} & \dots & a_{mn}^{k} \end{vmatrix} \rightarrow \left(a_{ij}^{k}\right)_{1} =$$

$$= \left(a_{11}^{k}, a_{12}^{k}, \dots a_{1n}^{k}, \dots a_{i1}^{k}, a_{i2}^{k} \dots a_{in}^{k} \dots a_{m1}^{k}, a_{m2}^{k}, \dots a_{mn}^{k}\right),$$

а во втором:

$$\left(a_{ij}^{k}\right) \rightarrow \left(\overrightarrow{a_{ij}^{k}}\right)_{2} = \left(a_{11}^{k}, \dots a_{i1}^{k}, .. a_{m1}^{k}, \dots a_{12}^{k}, \dots a_{i2}^{k}, \dots a_{m2}^{k}, \dots a_{1n}^{k}, \dots a_{in}^{k}, \dots a_{mn}^{k}\right).$$

Очевидно, что евклидово расстояние между двумя матрицами $\left(a_{ij}^{k}\right)$ и $\left(a_{ij}^{c}\right)$ будет такое же, как и между соответствующими векторами:

$$d = \left(\left(a_{ij}^{k}\right), \left(a_{ij}^{c}\right)\right) = d\left(\left(\overrightarrow{a_{ij}^{k}}\right)_{1}, \left(\overrightarrow{a_{ij}^{c}}\right)_{1}\right) = d\left(\left(\overrightarrow{a_{ij}^{k}}\right)_{2}, \left(\overrightarrow{a_{ij}^{c}}\right)_{2}\right). \quad (2)$$

К полученным векторам одной длины m×n в количестве К можно применить уже известные алгоритмы кластеризации для векторных объектов.

Положение матричного объекта в К-мерном матричном пространстве можно определить скалярной интегральной характеристикой

$$I_a = \frac{1}{m\,n}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(\frac{a_{ij}^{k}}{max\left\{a_{ij}^{k}\right\}}\right), \quad (3)$$

значения которой лежат в единичном интервале. Если сделать распределение I_a. в порядке убывания его значений, то можно классифицировать (кластеризировать) матричные объекты с помощью метода естественных границ. Границы на кривой распределения I_a определяются по резким изменениям значений I_a. Важно сравнить такую простейшую кластеризацию с полноценной кластеризацией векторных объектов одинаковой длины.

В качестве функционала качества кластеризации обычно используют следующие положения [Воронцов К.В., 2010]: среднее внутрикластерное расстояние должно быть, как

можно меньше (F0→min), а среднее межкластерное расстояние – как можно больше (F1→max).

Если же алгоритм кластеризации вычисляет центры кластеров, то можно определить функционалы качества, вычислительно, более эффективным способом: сумма средних внутрикластерных расстояний должна быть как можно меньше ($\phi_0 \rightarrow$min), а сумма межкластерных расстояний – как можно больше ($\phi_1 \rightarrow$max).

Эти положения, на наш взгляд, позволяют ввести критерий качества кластеризации, который в работе [Грызлова Т.П., Балыкина А.С., 2011] назван критерием информативности признакового пространства X. В обозначениях работы [Воронцова К.В., 2010] он будет иметь вид:

$$I(X) = \frac{F_0}{F_1} \text{ или } I(X) = \frac{\phi_0}{\phi_1}. \qquad (4)$$

Все вышесказанное будет справедливо и при кластеризации матриц одинаковой размерности.

Приведем примеры матричных объектов, которые можно кластеризировать, приводя их к векторам одинаковой размерности:

Кластеризация стран мира на основе базы данных Trade Competitiveness Map, имеющих m экспортных секторов экономики, каждая из которых характеризуется n индикаторами;

Кластеризация предприятий, выпускающих (или экспортирующих) m видов однотипной продукции, каждая из которых характеризуется n индикаторами;

Кластеризация торговых сетей, продающих m видов однотипных товаров, каждый из которых характеризуется n индикаторами;

Кластеризация супермаркетов по покупательско-продуктовой структуре продаж (m – количество наиболее активных покупателей, n – количество видов продукта) на основе построения бинарных матриц (1 – покупатель купил продукт, 0 – покупатель не купил продукт);

Кластеризация стран мира на основе платформы Scimago по «скопусовской» публикационной активности, имеющих m предметных научных категорий, каждая из которых характеризуется n индикаторами публикационной активности и цитируемости;

Кластеризация студенческих групп по умению решать однотипные задачи (m – количество студентов в группе, n - количество однотипных задач) на основе построения бинарных матриц (1 – студент решил задачу, 0 – студент не решил задачу).

Кластеризация университетских подразделений на основе матрицы экспертных оценок (m – количество видов университетской деятельности, например, исследования, обучение, инновации и коммуникации; n – количество экспертов, оценивающих качественный уровень видов деятельности подразделений университета по десятибалльной или пятибалльной шкале).

В заключение отметим, что кластеризация любых объектов существенно зависит от выбранных метрик и методов кластеризации, поэтому целесообразно проводить сценарные расчеты, как это сделано в работе [Demirel M. C., Kahya E., 2007]. В ней для собранных по шести гидрологическим станциям данных были проведены на основе Mat Lab сценарные расчеты по 9 метрикам и 5 алгоритмам иерархической кластеризации (45 сценарных расчетов).

Заключение

Таким образом, в данной работе проведен обзор исследований по матричной кластеризации и показано на отсутствие работ по кластеризации матриц одинаковой размерности. Для решения этой задачи предложено преобразовывать матрицы одинаковой размерности ($m \times n$) в векторы одинаковой длины $m \times n$, и далее проводить их кластеризацию уже известными методами, проводя сценарные расчеты для различных методов кластеризации и различных метрик. Предложено также сравнивать такие сценарные расчеты с простейшей кластеризацией матричных объектов, проделанной с помощью метода естественных границ для скалярного интегрального показателя I_a.

Список литературы
References

1. Воронцов К.В., 2010. Лекции по алгоритмам кластеризации и многомерного шкалирования. URL: http://bmedicine.ru/lekcii-po-algoritmam-klasterizacii-i-mnogomernogo-shkalirovaniya. (дата обращения: 24 июня 2010).
Vorontsov K.V., 2010. Lektsii po algoritmam klasterizatsii i mnogomernogo shkalirovaniya [Lecture on algorithms of clustering and multidimensional scaling]. Available at: http://bmedicine.ru/lekcii-po-algoritmam-klasterizacii-i-mnogomernogo-shkalirovaniya.
(accessed 24 June 2010). (in Russian)
2. Грызлова Т. П., Балыкина А.С., 2011. Система оценки информативности диагностических признаков и признаковых пространств. Авиационно-космическая техника и технология, 9 (86): 148-154.
Gryzlova T. P, Balykina A. S., 2011. A system for assessing the information content of diagnostic features and feature spaces. Aviatsionno-kosmicheskaya tekhnika i tekhnologiya [Aerospace engineering and technology] 9 (86): 148-154. (in Russian)
3. Чернышев Г., 2015. Вертикальное фрагментирование в реляционных СУБД. URL: http://synthesis.ipi.ac.ru/sigmod/seminar/s20150430. (дата обращения: 17 июня 2015).
Chernyshev G., 2015. Vertikal'noye fragmentirovaniye v relyatsionnykh SUBD [Vertical fragmentation in relational Database management system]. Available at: http://synthesis.ipi.ac.ru/sigmod/seminar/s20150430. (accessed 17 June 2017). (in Russian)
4. Bhat M. V., Haupt A., 1976. An efficient clustering algorithm. IEEE Transactions on Systems, Man, and Cybernetics, 6: 61-64.
5. Cheng C., 1994. Algorithm for vertical partitioning in database physical design. Available at: http: // www.sciencedirect.com/science/article/pii/0305048394900426. (accessed 4 December 1993).
6. Cheng C. H., Motwani J., 2009. An examination of cluster identification-based algorithm for vertical partitions. Available at: http: // dx.doi.org/10.1504/IJBIS. 2009.026695. (accessed 25 January 2014).
7. Cheng C. H., Weng K. F., 2011. An improved branch-and-bound clustering approach for data partitioning. Available at: http// dx.doi.org/10.1111/j.1475-3995.2010.00781. x. (accessed 4 October 2010).
8. Gorla N., Boe W. J., 1990. Database opering efficiency in fragmented databases in mainframe, mini, and micro system environments. Available at: http:// www.sciencedirect.com/science/article/pii/016023X9090030H. (accessed March 1990).
9. Demirel M. C., Kahya E., 2007. Hydrological determination of hierarchical clustering scheme by using small experimental matrix. Conference Paper March 2016. Conference: AGU Hydrology Day. 161-168.
10. Gorla N., 2007. A methodology for vertical partitioning in a multi-relation database environment // Journal of computer Science & Technology, 7 (3): 217-227.
11. Hoffer J. A., Severance D. G., 1975. The use of cluster analysis in physical data base design. Available at: http://doi.acm.org/10.1145/1232480.1282486. (accessed 22 September 1975).
12. Hammer M., Niamir B., 1979. A heuristic approach to attribute partitioning. Available at: http://doi.acm.org/10.1145/582095.582110. (accessed 20 May 1979).
13. Jindal A., Dittrich J., 2012. Relax and let the dabase do the partitioning online. In. Castellano M., Dayal U., Lehner W. (eds) Enabling Real-Time Business Intelligence, Lecture Notes in Business Information Processing, 126: 65-80.
14. Jindal A., Palatinus E., Pavlov V., Dittrich J., 2013. A comparison of knives for bread slicing. Available at: http://dl.acm.org/citation.cfm?id=2536336.2536338. (accessed 6 April 2013).
15. Kuo J. J., Zhang Y. J., 2012. A Library Recommender System Using Interest Change over Time and Matrix Clustering. Taipei, Taiwan. 259 - 268.
16. McCormick W. T., Scheweitzer P. j., and White., 1972. W.W. Problem decomposition and data reorganization by a clustering technique. Oper. Resp. 20, 5 (Sept. 1972): 993-1009.

17. Nagaraj G., Sheik Syed Abuthahir S., Manimaran A., Paramasamy S., 2015. Comparison of Matrix Clustering Methods to Design Cell Formation. International Journal of Applied Engineering Research, ISSN 0973-4562, 10 (28): 21900-21904.
18. Shamkant N., Stefano C., Gio W., Jingle D. Vertical partitioning algorithm for database design., 1984. Available at: http: // doi.acm.org/10.1145/1994.2209. (accessed 4 December 1984)
19. Stagle J. R., Chang C. L., and Heller S. R., 1975. A clustering and data-reorganization algorithm. IEEE Trans. Syst., Man, Cybern., 5: 125-128.
20. Oyanagi S., Kubota K., and Nakase A., 2001. Matrix Clustering: A new Data Mining Method for CRM. Trans.IPSJ, 42 (8): 2156-2166.
21. Oyanagi S., Kubota K., and Nakase A., 2001. Application of matrix clustering to web log analysis and access prediction. WEBKDD 2001–Mining Web Log Data Across All Customers Touch Points, Third International Workshop. 13-21.
22. Oyanagi S., Kubota K., and Nakase A., 2003. Mining WWW Access Sequence by Matrix Clustering. WEBKDD 2002, LNAI 2703. 119–136.
23. Oyelade J., Isewon I., Oladipupo F., Aromolaran O., Uwoghiren E., Ameh F., Achas M., and Adebiyi E., 2016. Clustering Algorithms: Their Application to Gene Expression Data. Bioinform Biol Insights. 10: 237–253.
24. Zhang ZY, Li T., Ding C., Ren XW., Zhang S., 2010. Binary matrix factorization for analyzing gene expression data. Data Min K and Discover, 20 (1): 28–52.