Мультимодальная кластеризация в анализе данных осложнений инфаркта миокарда

Богатырев М.Ю.[1], Шестака Т.В.[1]

[1]*Тульский государственный университет*

okkambo@mail.ru

Работа посвящена кластеризации биомедицинских данных. На данных осложнений инфаркта миокарда выполняется мультимодальная кластеризация методами Анализа формальных понятий. Для этого строятся формальные контексты и соответствующие им решётки формальных понятий. Формальное понятие представляют собой максимально возможный плотный многомерный гиперкуб, полностью заполненный элементами из множеств, образующих формальный контекст, и является мультимодальным кластером, представляющим собой сочетания данных, относящихся к различным аспектам диагностики и терапии инфаркта миокарда: анамнезу, анализам, применению лекарственных препаратов и осложнениям. Мультимодальная кластеризация позволяет выполнить фенотипирование заболеваний путем выполнения запросов к базе данных, содержащей решётку формальных понятий. Каждый такой запрос возвращает соответствующий ему набор формальных понятий – узлов решётки. Узлы образуют иерархию, что позволяет различать менее и более общие понятия и понятия одного уровня. Реализованный способ навигации в решётке понятий даёт возможность находить и анализировать узлы, соседние к заданному узлу. Результаты применения описанного метода к данным осложнений инфаркта миокарда сравниваются с результатами применения графовых моделей на тех же данных. Показано, что сравниваемые результаты дополняют друг друга.

Ключевые слова: *мультимодальная кластеризация, инфаркт миокарда, формальный контекст, решётка понятий, фенотипирование заболевания.*

Multimodal Clustering in the Data Analysis of Myocardial Infarction Complications

Bogatyrev M.Y.[1], Shestaka T.V.[1]

[1]*Tula State University*

This work is devoted to the biomedical data clustering. On the data of myocardial infarction complications, multimodal clustering is performed using the methods of Formal Concept Analysis. To do this, formal contexts and the lattices of formal concepts corresponding to them are constructed. The formal concept is the maximum possible dense multidimensional hypercube, completely filled with elements from the sets that form the formal context, and is a multimodal cluster, which is a combination of data related to various aspects of the diagnosis and treatment of myocardial infarction: anamnesis, analyzes, drug use and complications. Multimodal clustering allows one to perform phenotyping of a myocardial infarction by issuing queries to a database containing a conceptual lattice. Each such query returns a corresponding set of formal concepts – lattice nodes. Nodes form a hierarchy, which makes it possible to distinguish between less and more general concepts and concepts of the same level. The implemented method of navigation in the conceptual lattice makes it possible to find and analyze nodes adjacent to a given node. The results of applying this method to the data of myocardial infarction complications are compared with the results of using graph models on the same data. It is shown that the compared results complement each other.

Key words: multimodal clustering, myocardial infarction, formal context, conceptual lattice, disease phenotyping.

1. Введение

В анализе биомедицинских данных кластеризация [1] играет значительную роль, возрастающую с появлением феномена больших данных [2].

Известной проблемой в кластерном анализе является проблема интерпретации полученных кластеров. Любой алгоритм кластеризации использует некоторую меру близости, определенную на множестве кластеризуемых объектов. Поэтому «смысл» получаемых кластеров определяется, прежде всего, используемой мерой близости: объекты входят в один кластер потому, что они близки друг другу согласно выбранной числовой мере близости.

Мультимодальная кластеризация является одним из вариантов кластеризации, в котором проблема интерпретации кластеров может быть решена на новом уровне. Здесь кластеризация одновременно выполняется не на одном, а на двух, трёх и, в общем случае, на произвольном числе множеств: это *бикластеризация*, *трикластеризация* и собственно *мультимодальная кластеризация*.

Бикластеризация достаточно давно применяется в исследовании экспрессии генов [3], где одновременно обрабатываются множество генов и множество их состояний в образцах. Трикластеризация используется, когда данные экспрессии генов имеют третье измерение в виде временного ряда [4]. Применение би- и трикластеризации в исследовании экспрессии генов позволяет более точно выявлять ключевые гены биологических процессов.

Кроме анализа экспрессии генов, бикластеризация также применяется в других направлениях анализа биомедицинских данных [5].

В Анализе формальных понятий (АФП) [6] используется иной подход к мультимодальной кластеризации. Алгоритмы АФП работают на тензорных представлениях многомерных данных и строят на них решётки понятий или мультимодальные кластеры. При этом меры близости кластеризуемых объектов непосредственно не применяются [7, 8].

В мультимодальном кластере сам факт сочетания определенных данных может иметь важное значение с точки зрения пользователя и нести новую информацию. В методах АФП интерпретация кластеров также напрямую никак не связана с мерой близости объектов.

В настоящей работе АФП-метод мультимодальной кластеризации применяется к данным осложнений инфаркта миокарда, заданных в виде объектно-признакового представления. На этих данных строится решётка понятий как основной информационный ресурс, на котором реализуются алгоритмы мультимодальной кластеризации. Результаты кластеризации сравниваются с аналогичными результатами, полученными другим методом.

2. Мультимодальная кластеризация

Мультимодальная кластеризация выполняется на данных в виде объектно-признакового представления. В наиболее общем виде такие данные могут быть заданы как многомерный формальный контекст:

$$\mathbb{K} = <K_1, K_2, ..., K_n, R>, \quad (1)$$

который определяется отношением $R \subseteq D_1 \times D_2 \times ... \times D_n$ на доменах данных $D_1, D_2, ..., D_n$, $K_i \subseteq D_i$.

Мультимодальные кластеры на контексте (1) строятся в виде

$$\mathbb{C} = < X_1, X_2, ..., X_m >, \quad (2)$$

$X_i \subseteq K_i$ и обладают следующим свойством замыкания [9]:

$$\forall u = (x_1, x_2, ..., x_m) \in X_1, X_2, ..., X_m, u \in R, \quad (3)$$

при этом

$$\forall j = 1, 2, ..., n, \forall x_j \in D_j \setminus X_j < X_1, ..., X_j \cup \{x_j\}, ..., X_m >$$

не удовлетворяет условию (3).

Мультимодальный кластер - это подмножество в виде комбинаций элементов из разных наборов K_i. Он также определяется как замкнутое *m*-множество, поскольку свойство замыкания (3) обеспечивает его «самодостаточность»: кластер не может быть увеличен без нарушения условия замыкания.

Модальность кластера – это количество подмножеств его образующих, $m \le n$.

Размерность кластера – это количество объектов в кластере, как числовых, так и нечисловых.

Формальное понятие – это такой мультимодальный кластер, в котором для всех его элементов выполняется условие

$$u = (x_1, x_2, ..., x_m) \in X_1, X_2, ..., X_m, u \in R \quad (4)$$

Формальное понятие представляют собой максимально возможный *m*-мерный гиперкуб, полностью заполненные единицами. В АФП введено понятие *плотности мультимодального кластера* и формальные понятия интерпретируются как абсолютно плотные кластеры [10].

2.1. Методы мультимодальной кластеризации

Существует несколько подходов к формированию алгоритмов мультимодальной кластеризации [10–12], в том числе и для биомедицинских данных [13].

Классические АФП-алгоритмы строят решётки понятий на формальных контекстах [14]. Если целью кластеризации являются неплотные кластеры, то они строятся алгоритмами либо ориентированными на конкретную организацию данных [11, 12], либо более общими алгоритмами, например, использующими глобальную оптимизацию и эволюционные вычисления [15].

Выбор алгоритма кластеризации определяется следующими условиями.

2.1.1 *Организация данных*.

Для алгоритмов, работающих с формальными контекстами, необходимо сформировать такие контексты на исследуемых данных, например, хранящихся в базах данных. Это требует реализации ряда программных решений, включая, нормализацию, обработку пропущенных значений и т.п.

2.1.2. *Характер решаемых задач*.

Методами кластеризации решаются задачи установления зависимостей между данными, извлечения фактов, прогноз значений переменных.

Соответственно, выбираются и алгоритмы кластеризации.

2.1.3. *Особенности интерпретации результатов кластеризации*.

Здесь принципиально важно, достаточно ли использовать полученные кластеры непосредственно для получения нужной информации для решения задачи. Если недостаточно, то необходимо строить специальные пользовательские интерфейсы для интерпретации кластеров. При этом модальность кластеров имеет значение: кластеры с более высокой модальностью, например, при m >2 в выражении (2) могут быть в определенных задачах менее интерпретируемы, чем кластеры с меньшей модальностью.

3. Анализ данных осложнений инфаркта миокарда

Инфаркт миокарда имеет ряд осложнений, часто являющихся причиной летального исхода заболевания. Поэтому сбор и исследование данных, относящихся ко всем аспектам этого заболевания очень важны. Такой подход соответствует современному научному направлению «Вычислительная биомедицина» [2, 11, 12].

3.1. Исследуемые данные

Мы используем набор данных осложнений инфаркта миокарда [16], содержащий информацию о пациентах, страдающих этим заболеванием. Объектно-признаковое представление набора содержит 1700 объектов и 123 атрибута, собранных в многозначном формальном контексте. Среди атрибутов есть сведения о пациентах (все пациенты являются анонимными и имеют идентификационные номера), их анамнезе, методах лечения и осложнениях после лечения. Атрибут может быть бинарным или иметь числовое значение.

3.2. Задачи анализа данных

На данных об инфаркте миокарда была решена задача фенотипирования этого заболевания. Фенотипирование относится к определению формы заболевания на основе клинического профиля. *Клинический профиль* – это кластер, который может включать в себя различные данные, описывающие как само заболевание, так и методы его лечения, а также состояние пациентов. Поэтому нас интересовали различные сочетания наборов атрибутов из подмножеств атрибутов "пациент", "лечение", "результаты лечения". Такие сочетания получаются в результате мультимодальной кластеризации данных.

3.3. Результаты моделирования

На данных осложнений инфаркта миокарда были построены формальные контексты модальности 2, соответствующие подмножествам атрибутов из доменов "пациент", "лечение", "результаты лечения". Также на всех данных набора был построен общий формальный контекст. Состав формальных контекстов показан в таблице 1.

Таблица 1. Состав формальных контекстов

Контекст	**Объекты**	**Атрибуты**
Анамнез	1700	33
Терапия	1700	24
Анализы	1700	19
Инфаркт	1700	6
ЭКГ	1700	27
Осложнения	1700	14
Все данные	1700	123

Мультимодальная кластеризация выполнялась путем построения решётки понятий на общем формальном контексте. Решётка имеет 3683 узла, являющихся формальными понятиями, и хранится в базе данных.

Фенотипирование инфаркта осуществляется путем формирования запросов к базе данных. Каждый такой запрос возвращает соответствующий ему набор формальных понятий – узлов решётки. Узлы образуют иерархию, что позволяет различать менее и более общие понятия, а также понятия одного уровня. Реализованный способ навигации в решётке понятий даёт возможность находить и анализировать узлы, соседние к заданному узлу.

Мультимодальная кластеризация позволяет выполнять фенотипирование в виде установления зависимостей между данными и извлечения фактов.

3.3.1. *Установление зависимостей между данными*

В кластерах-формальных понятиях каждый объект связан со всеми атрибутами и наоборот – каждый атрибут связан со всеми объектами. Поэтому построив кластер для некоторого множества пациентов-объектов, можно фиксировать соответствующие им подмножества атрибутов, связанных друг с другом. В этом случае нас интересуют кластеры с большим числом пациентов. Тогда наборы атрибутов отражают существующие на атрибутах закономерности, например, стандартный процесс терапии инфаркта миокарда, характерный для большинства пациентов.

3.3.2. Извлечение фактов

Всякое нетипичное сочетание данных, которое привлекает внимание и отражает имеющие место событие или явление, будем называть *фактом*. На исследуемых данных кластеры – факты, как правило, содержат либо небольшое число пациентов (иногда это даже один пациент), либо небольшое число атрибутов.

На рис. 1 показан пример кластера, содержащего факт летального исхода для одного пациента, относительно молодой женщины.

Информация об узле:

Узел	2862
Атрибуты	Отек легких при поступлении в реанимацию; Ритм ЭКГ на момент поступления в стационар: синусовый (при ЧСС 60–90); Преждевременные сокращения желудочков на ЭКГ при поступлении в стационар; Фибринетлитическая терапия с помощью целиазы 500k МЕ; Наличие инфаркта миокарда правого желудочка; Использование опиоидных препаратов в отделении интенсивной терапии в первые часы госпитального периода; Использование опиоидных препаратов бригадой неотложной кардиологии; Летальный исход (причина); Кардиогенный шок при поступлении в реанимацию
Объекты	0.0588235% пациентов: 100.% женщин, 0.% мужчин. [1670] – женщина, возраст 46

Рис.1. Пример мультимодального кластера.

Интерпретация получаемых кластеров выполняется посредством пользовательского интерфейса, созданного для работы с базой данных и решёткой понятий. В запросах к решётке можно вводить ограничения на объекты и атрибуты, например, интересуясь только отдельными пациентами. Состав получаемых кластеров предъявляется пользователю, как это показано на рис. 1.

3.3.3. *Сравнение результатов моделирования*

Данные [16] были также использованы в работе [17], где к ним была применена структурная модель в виде *эластичных главных графов*, с помощью которой можно решать задачи кластеризации. Предложенный в работе [17] метод позволяет строить т.н. *клинические траектории лечения* и позиционировать пациента на определенной траектории.

Мультимодальные кластеры в виде формальных понятий в решётке понятий служат узлами графов – клинических траекторий и отражают сходные решения задачи кластеризации. В нашем случае – это детализация клинических траекторий. Отсюда следует, что сравниваемые результаты данной работы и работы [17] дополняют друг друга.

4. Заключение

В настоящей работе выполнен анализ данных осложнений инфаркта миокарда методом мультимодальной кластеризации. Выбранный метод основан на Анализе формальных понятий и не использует понятие меры близости объектов, что расширяет возможности интерпретации получаемых кластеров. Интерпретация кластеров позволяет выявлять зависимости между атрибутами и устанавливать их особые сочетания в виде фактов.

Метод легко реализуется в рамках стандартных технологий СУБД и имеющихся программных средств АФП.

Развитие данного метода применительно к данным в виде объектно-признакового представления предполагается в следующих направлениях.

1. Исследование вариантов мультимодальной кластеризации с модальностью $n > 2$. Это в ряде случаев позволит более детально анализировать данные, в которых имеются измерения в виде временного ряда и других координат.

2. Создание пользовательских интерфейсов, ориентированных конкретные задачи медицины, например, кардиологии, что позволит эффективно применять данный метод в системах поддержки принятия решений и вопросно-ответных информационных системах.

5. Благодарности

Работа выполнена при финансовой поддержке РФФИ: проекты № 19-07-01178, № 20-07-00055, а также РФФИ и Тульской области: проект № 19-47-710007.

6. Список литературы

1. *Data Clustering: Algorithms and Applications*. Ed. Charu Aggarwal, Chandan Reddy. CRC Press, London, 2013. 652 p.
2. Röttger R. Clustering of Biological Datasets in the Era of Big Data. *Journal of Integrative Bioinformatics*. 2016. V. 13. № 1. P. 300. doi: 10.1515/jib-2016-300
3. Cheng Y., Church G. Biclustering of Expression Data. Proc. Int. Conf. on Intelligent Systems for *Molecular Biology*. 2000. P. 93-103.
4. Gutiérrez-Avilés D, Rubio-Escudero C. Mining 3D patterns from gene expression temporal data: a new tricluster evaluation measure. *Scientific World Journal*. 2014. Article ID 624371. doi: 10.1155/2014/624371
5. Madeira S.C., Oliveira A.L. Biclustering algorithms for biological data analysis: a survey. IEEE/ACM Trans. *Comput. Biol. Bioinform*. 2004. V.1. № 1. P. 24–45. doi: 10.1109/TCBB.2004.2
6. *Formal Concept Analysis: Foundations and Applications*: Lecture Notes in Artificial Intelligence. Eds. Ganter B., Stumme G., Wille R. No. 3626. Berlin: Springer-Verlag, 2005. doi: 10.1007/978-3-540-31881-1
7. Bogatyrev M.Y., Samodurov K.V. Conceptual Approach to Clustering in the Study of Gene Expression. In: *Proceedings of the International Conference "Mathematical Biology and Bioinformatics"*. Ed. V.D. Lakhno. Vol. 7. Pushchino: IMPB RAS, 2018. Paper No. e54. doi: 10.17537/icmbb18.81
8. Богатырев М.Ю. Методы кластеризации в исследовании экспрессии генов. В: *Доклады Международной конференции «Математическая биология и биоинформатика»*. Под ред. В.Д. Лахно. Том 8. Пущино: ИМПБ РАН, 2020. Статья № e21. doi: 10.17537/icmbb20.27
9. Cerf L., Besson J., Robardet C., Boulicaut J.F. Closed Patterns Meet *N*-ary Relations. In: *ACM Trans. Knowl. Discov. Data*. V.3. Issue1. 2009. P. 3–36.
10. Ignatov D.I., Gnatyshak D.V., Kuznetsov S.O., Mirkin B.G. Triadic Formal Concept Analysis and triclustering: searching for optimal patterns. *Mach. Learn*. 2015. V. 101. P. 271–302.

11. Zhao Q., Zong L., Zhang X., etc. A Multimodal Clustering Framework With Cross Reconstruction Autoencoders. *IEEE*. 2020. P. 218433–218443. doi: 10.1109/ACCESS.2020.3040644
12. Zhong G., Pun C.-M. Latent Low-rank Graph Learning for Multimodal Clustering. *2021 IEEE 37th International Conference on Data Engineering* (ICDE). 2021. P. 492–503. doi: 10.1109/ICDE51399.2021.00049
13. Horne E., Tibble H., Sheikh A., Tsanas A. Challenges of Clustering Multimodal Clinical Data: Review of Applications in Asthma Subtyping. *JMIR Med. Inform*. 2020. V. 8. № 5. Article No. e16452. doi: 10.2196/16452
14. Gnatyshak D., Ignatov D., Kuznetsov S. From Triadic FCA to Triclustering: Experimental Comparison of Some Triclustering Algorithms. In: *Proceedings of the Tenth International Conference on Concept Lattices and Their Applications* (*CLA'2013*). La Rochelle: Laboratory L3i, University of La Rochelle, 2013. P. 249–260.
15. Bogatyrev M., Orlov D., Shestaka T. On the Pareto-Optimal Solutions in the Multimodal Clustering Problem. In: *Recent Trends in Analysis of Images, Social Networks and Texts. AIST 2021. Communications in Computer and Information Science*, V. 1573. Springer, Cham. P. 179–194. 2022. doi: 10.1007/978-3-031-15168-2_15
16. *Myocardial infarction complications Data Set*. URL: http://archive.ics.uci.edu/ml/machine-learning-databases/00579/ (accessed 21.09.2022).
17. Golovenkin S.E., Bac J., Chervov A., Mirkes E.M., Orlova Y.V., Barillot E., Gorban A.N., Zinovyev A. Trajectories, bifurcations, and pseudo-time in large clinical datasets: applications to myocardial infarction and diabetes data. *Giga Science*. 2020. V. 9. № 11. doi: 10.1093/gigascience/giaa128