***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*

Методы кластеризации в исследовании экспрессии генов

Богатырев М.Ю.

Тульский государственный университет

okkambo@mail.ru

Доклад является изложением обзора методов кластеризации, применяемых в исследовании экспрессии генов. Обзор содержит описание классических методов «плоской» и иерархической кластеризации, а также методов, использующих эволюционные вычисления и анализ формальных понятий. Применение эволюционных вычислений позволяет эффективно использовать многокритериальную оптимизацию в решении задач кластеризации. Анализ формальных понятий даёт возможность исследовать иерархии кластеров как результат точной кластеризации данных экспрессии генов. Обзор имеет целью привлечь внимание специалистов в области молекулярной биологии и в области компьютерных наук к совместным исследованиям.

Ключевые слова: *экспрессия генов, кластеризация, анализ формальных понятий, эволюционные вычисления.*

Clustering Methods in Gene Expression Research

Bogatyrev M.Y.

Tula State University

The report provides an overview of clustering methods used in gene expression research. The review describes the classic methods of "flat" and hierarchical clustering, and also methods based on Evolutionary Computation and Formal Concept Analysis. The use of Evolutionary Computation in clustering makes it possible to effectively use multi-criteria optimization in solving clustering problems. The analysis of formal concepts makes it possible to investigate cluster hierarchies as a result of precise clustering of gene expression data. The review aims to draw an attention of specialists in the field of molecular biology and computer science to joint research.

Key words: gene expression, clustering, evolutionary computation, formal concept analysis.

1. Введение

В последнее время достигнут значительный прогресс в исследовании экспрессии генов – одной из главных задач биоинформатики. Созданы и совершенствуются как технологии получения данных экспрессии на биоматериалах, так и методы их обработки. Среди методов обработки данных экспрессии генов методы кластеризации занимают центральное место. Кластеризация – это разделение множества объектов на непересекающиеся подмножества – кластеры, так, что в каждом из подмножеств объекты максимально близки друг другу согласно некоторому критерию, а их межкластерная близость минимальна. Особенности структур данных, варианты критериев близости объектов и способов нахождения оптимальных количества и состава кластеров порождают огромное разнообразие методов кластеризации. Целью данного обзора является систематизация методов кластеризации, применяемых в исследовании экспрессии генов, как с точки зрения упомянутых особенностей методов кластеризации, так и в соответствии с задачами анализа экспрессии генов, для решения которых важна интерпретация кластеров.

Данная тема освещена в обзорах [1–3], составленных в разное время. В данном обзоре рассмотренные в прошлых обзорах методы кластеризации дополнены новыми методами, появление которых связано с развитием теории машинного обучения и интеллектуального анализа данных. Это относится к анализу формальных понятий [4] – направлению в анализе данных, позволяющему обобщить методы би- и трикластеризации, а также к эволюционным вычислениям [5].

В обзоре приводятся описания и псевдокоды основных алгоритмов, применяемых в исследовании экспрессии генов, на которые в данном тексте даны ссылки.

Список источников в обзоре включает 46 наименований и в данном докладе приведены лишь основные работы из этого списка.

***Proceedings of the International Conference "Mathematical Biology and Bioinformatics".** Ed. V.D. Lakhno. Vol. 8. Pushchino: IMPB RAS, 2020. Paper No. e21. doi: 10.17537/icmbb20.27*

2. Задача исследования экспрессии генов

Методы кластеризации работают на данных экспрессии генов, которые представлены в виде матриц или тензоров, когда размерность представления больше двух. Для получения этих данных применяются специальные технологии, среди которых технология микрочипов является основной.

2.1. Технология микрочипов

Технологии микрочипов появилась достаточно давно [6] и описана в большом числе работ. Микрочип изготавливается из стекла с химическим покрытием в виде нейлоновой мембраны или кремниевого слоя, на которых размещаются десятки тысяч молекул ДНК в фиксированных местах-ячейках сетки в виде пятен размером в несколько микрон. Каждая ячейка сетки содержит фрагмент-последовательность ДНК длиной в десятки нуклеотидов.

Анализ экспрессии генов выполняется путём гибридизации различных ДНК – той, которая нанесена на микрочип, и содержащейся в образцах, дополнительно наносимых на микрочип. Эти образцы представляют собой исследуемый биоматериал, который содержит флуоресцентные метки. Далее микрочип сканируют конфокальным лазером и фиксируют величины флуоресцентных сигналов. Технология позволяет вводить эти данные сразу в компьютер для дальнейшей обработки.

Таким образом, стандартный микрочип формирует $n \times m$ матрицу данных $\mathbf{W} = \{w_{i,j}\}$, $i = \overline{1, n}$, $j = \overline{1, m}$ с двумя измерениями: множество генов $\boldsymbol{G}$ и множество образцов $\boldsymbol{P}$. На пересечениях измерений находятся действительные числа, соответствующие уровням экспрессии генов.

В последнее время всё чаще используются 3D микрочипы, позволяющие получить данные экспрессии генов в виде трёхмерных тензоров. Соответственно, на таких данных востребованы новые алгоритмы кластеризации, известные как алгоритмы трикластеризации [3].

По сравнению с традиционным подходом к геномным исследованиям, в котором основное внимание было уделено локальному исследованию и сбору данных об отдельных генах, технологии микрочипов позволили параллельно контролировать уровни экспрессии для десятков тысяч генов [1].

2.2. Кластеризация как метод исследования экспрессии генов

Кластеризация данных экспрессии генов позволяет решать следующие задачи.

1. Нахождение ранее неизвестных функций генов, проявляющих себя при исследовании экспрессии большого числа генов. Гены с аналогичными образцами экспрессии (совместно выраженные гены) могут быть сгруппированы в кластеры вместе с подобными клеточными функциями. Это позволяет выявлять функции многих генов, для которых информация ранее не была доступна.
2. Определение влияния генов на клеточные процессы. Совместно выраженные гены в одном и том же кластере, как правило, вовлечены в одни и те же клеточные процессы, и сильная корреляция структур экспрессии между этими генами указывает на совместную регуляцию.
3. Поиск общих последовательностей ДНК в зонах промотора генов внутри одного и того же кластера. Это позволяет определять регуляторные элементы, специфичные для каждого кластера генов. Изучение регуляции через кластеризацию данных экспрессии генов позволяет формировать гипотезы о механизмах работы транскрипционной регуляторной сети.
4. При формировании данных экспрессии генов в виде трёхмерных тензоров имеется возможность рассматривать в качестве третьего измерения различные множества данных, например, множество отсчётов времени, множество регуляторов экспрессии или иное множество данных, в том числе и гетерогенных, включающих тексты [8, 9].

3. Обзор методов кластеризации

Рассмотрим задачу кластеризации данных в общей постановке. Дано множество $\boldsymbol{A}$ элементов $a \in A$, каждый элемент множества обладает набором из k признаков, поэтому его можно представить в виде вектора размерности k: $a \to \bar{a}$. Вводится функция близости элементов множества: $f(\bar{a}_i, \bar{a}_j)$. Необходимо исходное множество $\boldsymbol{A}$ представить в виде непересекающихся подмножеств-кластеров $\boldsymbol{A}^{(i)}$, таких, что любые два кластера $\boldsymbol{A}^{(p)}$ и $\boldsymbol{A}^{(s)}$ содержит элементы, для которых для любых i, j, l, m выполняются неравенства:

$$\left.\begin{aligned} f(\bar{a}_i^{(p)}, \bar{a}_j^{(p)}) > f(\bar{a}_l^{(p)}, \bar{a}_m^{(s)}) \\ f(\bar{a}_i^{(s)}, \bar{a}_j^{(s)}) > f(\bar{a}_l^{(p)}, \bar{a}_m^{(s)}) \end{aligned}\right\} \qquad (1)$$

Здесь $\bar{a}_i^{(p)} \in A^{(p)}$ – элементы кластеров.

Заметим, что в ряде задач кластеризации требование непересекаемости кластеров может быть снято. Это делается, когда пересечения кластеров представляют особый интерес как возможные источники важной информации [10].

Если набор признаков объектов кластеризации составляют их *координаты*, то задача кластеризации имеет наглядную геометрическую интерпретацию. Зная координаты объектов, можно ввести понятие *расстояния* между ними – величину, обратную *близости*. Многие алгоритмы кластеризации, например, широко известный алгоритм *k*-средних, используют геометрическую интерпретацию и связанное с ней Евклидово расстояние между объектами. Если объекты кластеризации можно представить как точки в

n-мерном линейном векторном пространстве, то для любых двух точек **a**, **b** с координатами (a_1, a_2, …, a_n) и (b_1, b_2, …, b_n) Евклидово расстояние между объектами вычисляется по формуле:

$$d(\mathbf{a},\mathbf{b}) = \sqrt{\sum_{k=1}^{n}(a_k - b_k)^2} \qquad (2)$$

Существует несколько подходов к классификации методов кластеризации [1, 3, 4]. В данном обзоре для нас важно различать *обычную* и *мультимодальную* кластеризации, поскольку последняя объединяет методы, которые выводят исследование экспрессии генов на новый уровень. Под обычной подразумевают кластеризацию на одном множестве данных. Мультимодальная кластеризация выполняется сразу на нескольких множествах – на двух или трёх, как в случае би- и трикластеризации, применяемой в последнее время [3].

3.1. Одномерная кластеризация

Обычную кластеризацию часто называют одномерной, хотя это не совсем корректно: в случае векторной меры близости объектов $f(\bar{a}_i, \bar{a}_j)$ выполняется их многомерный анализ, а кластеризация выполняется на одном множестве, как правило, методами, которые в настоящее время можно считать классическими. Они делятся на два класса: *плоские* и *иерархические* методы.

Типичный представитель плоских методов, метод *k*-средних [11], применяется в исследовании экспрессии генов в двух вариантах: когда гены рассматриваются в качестве объектов, а образцы в качестве признаков и наоборот. В обоих случаях каждый объект кластеризации (ген или образец) задаётся вектором признаков, содержащим действительные числа из строки или, соответственно, столбца матрицы экспрессии. Евклидово расстояние (2) естественным образом применяется в данной модели.

Метод *k*-средних имеет ряд недостатков. Прежде всего – это необходимость задавать число кластеров *k*. Также этот алгоритм чувствителен к шуму в данных, который всегда присутствует в данных экспрессии генов.

Чтобы преодолеть эти недостатки, было предложено несколько новых алгоритмов кластеризации, например, [12, 13]. Эти алгоритмы используют глобальные параметры для управления качеством результирующих кластеров: максимальный радиус кластера и минимальное расстояние между кластерами.

В отличие от плоских алгоритмов кластеризации, которые однозначно представляют набор данных в виде набора непересекающихся кластеров, иерархические алгоритмы генерируют серию вложенных кластеров, которые могут быть графически представлены деревом, называемым *дендрограммой*. При этом они часто программируются так, что имеют визуальный интерактивный режим управления дендрограммой, что позволяет получать различные варианты кластеризации.

3.2. Мультимодальная кластеризация

Главной проблемой одномерной кластеризации данных экспрессии генов является искажения реальной картины экспрессии: предъявляемые непересекающиеся кластеры могут не соответствовать особенностям взаимосвязей данных.

Радикальное решение этой проблемы предлагается в методах мультимодальной кластеризации [2, 3, 7, 9, 14].

Рассмотрим основные подходы мультимодальной кластеризации. Первый алгоритм бикластеризации был предложен в работе [14]. Результатом такой кластеризации является подматрица матрицы экспрессии, содержащая значения экспрессии, удовлетворяющие определенным критериям. Это могут быть критерии, позволяющие найти гены, обладающие как одинаковыми значениями экспрессии по образцам (с определенной точностью), так и более сложные критерии, также привлекающие глобальные параметры. Применение бикластеризации позволило эффективно исследовать экспрессию генов на бикластерах, содержащих сочетания генов и образцов, соответствующих некоторым заболеваниям. Бикластеризация стандартно применяется в информационном хранилище GEO [16]. На рисунке 1 показан фрагмент бикластеризации матрицы экспрессии генов, выполненной в системе GEO. На рисунке представлены две дендрограммы – результат применения иерархического метода кластеризации. Цветом помечены бикластеры, имеющие сходные параметры по используемому критерию. Дендрограммы системы GEO не интерактивные, но позволяют визуально оценить связи между кластерами.

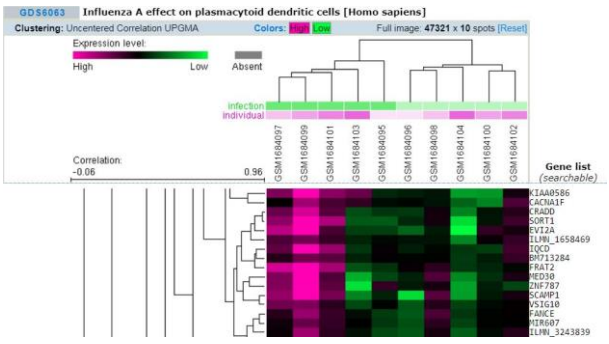

Рис. 1. Результат бикластеризации в системе GEO.

Трикластеризация расширяет метод бикластеризации на третье измерение. Как правило, таким измерением является время в виде дискретных интервалов – часов или дней.

Трёхмерные матрицы экспрессии с временно́й координатой генерируются на 3D микрочипах [7]. Кроме времени, используются и другие измерения, добавляемые к двумерным матрицам экспрессии [8, 9]. Вычислительная сложность методов трикластеризации [3] значительно возрастает по сравнению с методами бикластеризации. Другой особенностью этих методов является неоднозначность результатов кластеризации, выражающаяся в наличии нескольких вариантов структур кластеров для одних и тех же данных.

3.3. Эволюционные алгоритмы кластеризации

Классические и многие методы мультимодальной кластеризации в процессе обработки данных решают задачу оптимизации. В результате конфигурация кластеров, генерируемая алгоритмом, соответствует экстремальному значению определённого критерия качества. В методе *k*-средних это минимум среднеквадратического отклонения расстояния каждого элемента кластера до его центра. В других методах применяются иные критерии.

Функции, моделирующие зависимость критерия качества от определённых переменных, например, числа кластеров, могут оказаться негладкими и мультимодальными, т. е. имеющими несколько экстремумов. Эволюционные методы оптимизации эффективны при нахождении экстремумов именно таких сложных функций. Они также применяются в кластеризации [5].

Эволюционные алгоритмы основаны на *генетическом алгоритме* оптимизации [17], и в настоящее время представлены большим разнообразием вариантов. Все они используют понятие *эволюции* оптимизируемых объектов, а также понятия *ген*, *хромосома*, *популяция особей*, *мутация*, *кроссинговер*. Несмотря на то, что аналогия с биологическим содержанием этих понятий весьма отдалённая, эффективность этих алгоритмов имеет определённые элементы генетической природы [18].

Эволюционные алгоритмы применяются в кластеризации данных экспрессии генов как в варианте би-, так и трикластеризации [19, 20]. В большинстве алгоритмов в виде хромосом кодируется конфигурация кластеров. «Ген» такой хромосомы соответствует гену в данных экспрессии. Это приводит к необходимости обработки двоичных или символьных строк очень большой длины, для чего создаются специальные методы. В вариантах би-, и трикластеризации такие хромосомы имеют, соответственно, два и три отдельных участка.

Эволюционные алгоритмы эффективны в задачах многокритериальной оптимизации. При исследовании экспрессии генов такая оптимизация применяется для критериев размера кластера и его когерентности [21]. В этой работе также нашёл применение специально созданный алгоритм многокритериальной оптимизации NSGA-II.

3.4. Применение Анализа формальных понятий в кластеризации

Анализ формальных понятий (АФП) [4] – это направление в современном анализе данных, которое исследует данные, связанные отношением «объект – атрибут», и образующие *формальный контекст*.

Формальный контекст $\mathbf{K} = (G, M, I)$ задает связь между множествами объектов $\boldsymbol{G}$ и принадлежащих им атрибутов $\boldsymbol{M}$. Связь определяется отношением $I \subseteq G \times M$, которое задается матрицей контекста $\mathbf{K} = \{k_{i,j}\}$, ненулевые элементы которой фиксируют факты принадлежности атрибута $m_j \in M$ объекту $g_i \in G$. В классическом анализе формальных понятий используются «плоские» формальные контексты в виде бинарных матриц отношений «объект – атрибут». Современный АФП использует небинарные матрицы, а также многомерные формальные контексты в виде тензоров.

Любые данные экспрессии генов могут быть представлены в виде формальных контекстов. На формальных контекстах строятся *решётки понятий*, представляющие собой решения задач би- [22] и трикластеризации [23], а также мультимодальной кластеризации любой размерности. При этом формально не используется оптимизация и кластеризацию можно считать точной.

5. Заключение

Анализ методов кластеризации в исследовании экспрессии генов позволяет сделать следующие выводы.

1. Существует несколько направлений применения кластеризации, которые используют особенности данных экспрессии и поставленных конкретных задач её исследования. Это обычная и мультимодальная кластеризация, эволюционная кластеризация и кластеризация с применением Анализа формальных понятий.

2. Тенденция к унификации хранения данных и программных решений по их обработке обусловливает появление интегрированных хранилищ данных типа GEO, открывающих новые возможности в исследовании экспрессии генов. К таким возможностям относится, например, применение машинного обучения на данных хранилища с последующим формирование прогноза их состояния. Современные технологии нейронных сетей позволяют решать подобные задачи на больших данных экспрессии.

6. Благодарности

Работа выполнена при финансовой поддержке РФФИ (проект № 19-07-01178), а также РФФИ и Тульской области (проект № 19-47-710007).

7. Список литературы

1. Jiang D., Tang C., Zhang A. Cluster analysis for gene expression data: a survey. *IEEE Transactions on Knowledge and Data Engineering*. 2004. V. 16. P. 1370–1386. doi: 10.1109/TKDE.2004.68.
2. Madeira S., Oliveira A. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 2004. V. 1. P. 24–45. doi: 10.1109/TCBB.2004.2.
3. Henriques R., Madeira S. Triclustering Algorithms for Three-Dimensional Data Analysis: A Comprehensive Survey. *ACM Comput. Surv*. 2019. V. 51. № 5. P. 1–43. doi: 10.1145/3195833.
4. *Formal Concept Analysis: Foundations and Applications*: Lecture Notes in Artificial Intelligence, No. 3626. Eds. Ganter B., Stumme G., Wille R. Berlin: Springer-Verlag, 2003.
5. Hruschka E., Campello R., Freitas A., de Carvalho A. A Survey of Evolutionary Algorithms for Clustering. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*. 2009. V. 39. P. 133–155. doi: 10.1109/TSMCC.2008.2007252.
6. Schena M., Shalon D., Davis R., Brown P. Quantitative monitoring of gene expression patterns with a compolementatry DNA microarray. *Science*. 1995. V. 270. P. 467–470.
7. Zhao L., Zaki M.J. Tricluster: An effective algorithm for mining coherent clusters in 3d microarray data. In: *ACM SIGMOD International Conf. on Management of Data*. ACM, 2005. P. 694–705.
8. Alqadah F., Bhatnagar R. An effective algorithm for mining 3-clusters in vertically partitioned data. In: *Int. Conf. on Information and Knowledge Management*. ACM, 2008. P. 1103–1112.
9. Li A., Tuck D. An effective tri-clustering algorithm combining expression data with gene regulation information. *Gene Regulation and Systems Biology*. 2009. V. 3. P. 49.
10. Jiang D., Pei J., Zhang A. Interactive Exploration of Coherent Patterns in Time-Series Gene Expression Data. In: *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (*SIGKDD'03*). Washington, DC, USA, 2003.
11. McQueen J.B. Some methods for classification and analysis of multivariate observations. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Vol. 1. Univ. of Calif. Press, 1967. P. 281–297.
12. Heyer LJ., Kruglyak S., Yooseph S. Exploring expression data: identification and analysis of coexpressed genes. *Genome Res*. 1999. V. 9. № 11. P. 1106–1115.
13. De Smet F., Mathys J., Marchal K., Thijs G., De Moor B., Moreau. Y. Adaptive quality-based clustering of gene expression profiles. *Bioinformatics*. 2002. V. 18. P. 735–746. doi: 10.1093/bioinformatics/18.5.735.
14. Cheng Y., Church GM. Biclustering of expression data. In: *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology* (*ISMB*). 2000. V. 8. P. 93–103,
15. Li L., Guo Y., Wu W., Shi Y., Cheng J., Tao S. A comparison and evaluation of five biclustering algorithms by quantifying goodness of biclusters for gene expression data. *BioData Mining*. 2012. V. 5. Article No. 8. doi: 10.1186/1756-0381-5-8.
16. *GEO: Gene Expression Omnibus*: Public functional genomics data repository. URL: https://www.ncbi.nlm.nih.gov/geo/ (accessed 28.09.2020).
17. Holland J., Goldberg D. Genetic algorithms in search, optimization and machine learning. MA: Addison-Wesley, 1989.
18. Богатырев М.Ю. Генетическая природа эффективности эволюционных алгоритмов. В: *Математическая биология и биоинформатика*: доклады II Международной конференции (7–13 сентября 2008 г., Пущино). М.: МАКС Пресс. 2008. С. 64–65.
19. Ma P., Chan K., Yao X., Chiu D. An evolutionary clustering algorithm for gene expression microarray data analysis. *IEEE Transactions on Evolutionary Computation*. 2006. V. 10. P. 296–314. doi: 10.1109/TEVC.2005.859371.
20. Gutiérrez-Avilés D., Rubio-Escudero C., Martínez-Álvarez F. TriGen: A genetic algorithm to mine triclusters in temporal gene expression data. *Neurocomputing*. 2014. V. 132. P. 42–53. doi: 10.1016/j.neucom.2013.03.061.
21. Mitra S., Banka H. Multi-objective evolutionary biclustering of gene expression data. *Pattern Recognit*. 2006. V. 39. P. 2464–2477. doi: 10.1016/j.patcog.2006.03.003.
22. Kaytoue M., Kuznetsov S.O., Napoli A., Duplessis S. Mining gene expression data with pattern structures in formal concept analysis. *Information Sciences*. 2011. V. 181. № 10. P. 1989–2001.
23. 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.
24. Raza K. Formal Concept Analysis for Knowledge Discovery from Biological Data. *International Journal of Data Mining and Bioinformatics*. 2017. V. 18. № 4. P. 281–300.