Математические структуры и моделирование
2024. № 1 (69). С. 69–76

УДК 519.237.8
DOI 10.24147/2222-8772.2024.1.69-76

АНАЛИЗ МЕТОДОВ КЛАСТЕРИЗАЦИИ ДЛЯ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ В РОЗНИЧНОЙ ТОРГОВЛЕ

Т.С. Катермина
к.т.н., e-mail: nggu-lib@mail.ru
А.А. Макамбеджан
магистрант, e-mail: aygul-mak@mail.ru

Нижневартовский государственный университет, Нижневартовск, Россия

Аннотация. В статье рассматриваются теоретические основы интеллектуального анализа данных, изучаются методы кластерного анализа на примере выполнения задачи кластеризации данных о продажах от компании «1С». Рассмотренные методы применимы к задаче поддержки принятия решений в различных сферах, в том числе в сфере розничной торговли. В статье приводится обзор методов кластерного анализа, а также эксперимент с применением агломеративного метода кластерного анализа и метода k-средних. Выявлены преимущества и недостатки указанных методов, приведены результаты моделирования.

Ключевые слова: искусственный интеллект, кластерный анализ, интеллектуальный анализ данных.

Введение

Интеллектуальный анализ данных (далее – ИАД) – это процесс поддержки принятия решений, основанный на поиске в данных скрытых закономерностей (шаблонов информации) [1]. Для обнаружения скрытых знаний применяют специальные методы автоматического анализа – DataMining. К наиболее известным методам представления и получения знаний можно отнести классификацию, регрессию, кластеризацию, прогнозирование временных рядов и т. п. [2]. К задачам, с решением которых может помочь ИАД, относятся:

1) получение отчётов;
2) решение нестандартных вопросов (например, классификация);
3) построение моделей и прогнозирование.

Область применения ИАД обширна и включает в себя все сферы общества: научную, социальную, экономическую и политическую.

К одному из самых востребованных методов интеллектуального анализа данных относится кластеризация.

Для выполнения задачи разделения группы товаров на категории по соотношению цены и качества можно прибегнуть к использованию кластеризации. Под термином «кластеризация» понимается группировка объектов на основе близости их свойств. Группы, по которым распределяются данные, называются кластерами.

Каждый из них состоит из схожих объектов, но элементы разных кластеров должны отличаться друг от друга.

Иными словами, кластеризация – процедура, которая любому объекту $x \in \mathbb{X}$ ставит в соответствие метку кластера $y \in \mathbb{Y}$, из чего можно сделать вывод, что основной задачей кластеризации является разбиение векторов на подмножества, которое ориентируется на выявленные сходства и различия элементов [3].

Кластеризация обычно применяется при отсутствии априорной информации о группах, по которым будут классифицироваться данные, или когда кластеризируемых объектов слишком много для ручного анализа.

В задачах кластеризации нет необходимости в указании выходной переменной, а число кластеров может быть неизвестным. Стоит отметить, что кластеризация указывает только на схожесть объектов и не даёт готовый ответ в процессе решения задачи. Например, не указывает на закономерности, по которым были сформированы кластеры. Кластеризация участвует в решении следующих задач:

1. Изучение данных. Кластерный анализ данных помогает выявить структуру данных, сделать их наглядными и удобными для восприятия человека. При помощи кластерного анализа можно выявить необходимые параметры для дальнейшего моделирования сложных систем [4].
2. Облегчение анализа. Кластеризация упрощает последующую обработку данных и построение моделей, также появляется возможность создать индивидуальную модель каждого кластера.
3. Сжатие данных. Кластеризация может помочь сохранить объём хранимых данных, сохранив по одному наиболее приближённому к остальным представителю от каждой группы.
4. Прогнозирование. Кластерный анализ может определить, в какой кластер будет отнесён новый объект исходя из критериев кластеризации, и предсказать его поведение, опираясь на информацию о схожих элементах.
5. Обнаружение аномалий. Кластеризация способна определять нетипичные объекты, которые нельзя отнести ни к одному из кластеров.

Проводить кластеризацию можно по нескольким признакам, после чего будет достаточно просто определить, объектов с какими характеристиками больше, а каких – меньше, и сделать выводы из получившихся результатов. Кластеризация применяется в:

1) маркетинге;
2) сегментации изображений;
3) медицине;
4) археологии;
5) социологии;
6) государственном управлении и т. д.

Методы кластерного анализа

В связи с тем, что одно и то же множество объектов можно разделить на кластеры по-разному, появилось большое количество методов кластеризации. Не существует одного универсального метода кластеризации, но можно подобрать из них

наиболее подходящий под ту или иную ситуацию.

Однако, существуют сложности, с которыми предстоит столкнуться при использовании кластерного анализа:

1. Неопределённость в выборе критерия качества кластеризации. Проблема заключается в том, что на практике, когда объекты описываются десятками признаков, определение их расположения становится затруднительным.
2. Трудность выбора меры близости. Она обусловлена различной природой данных. Например, если для вычисления расстояния между числовыми данными можно использовать евклидово расстояние, то категориальным типам подобные меры не подойдут. Нужно будет прибегать к использованию специальной меры, которая, например, задаётся функцией отличия.
3. Различные требуемые машинные ресурсы. Обычно, чем точнее результат выдаёт кластер, тем больше он потребует затраченного времени и памяти. Поэтому в интеллектуальном анализе не получил распространение иерархический метод кластеризации, который строит полное древо вложенных кластеров.

Классификация методов кластерного анализа по способу обработки данных:

1. Иерархические методы:
   - агломеративный метод AGNES (Agglomerative Nesting);
   - дивизимный метод DIANA (Divisive Analysis).
2. Неиерархические методы:
   - метод k-средних (k-means);
   - метод разделения вокруг метоидов PAM (partitioning around medoids);
   - метод кластеризации категорийных данных CLOPE и т. д.

Также кластеризацию можно поделить на чёткую (например, метод k-средних) и нечёткую (например, метод c-средних).

При иерархической кластеризации выполняется последовательное объединение меньших кластеров в большие или разделение больших кластеров на меньшие. Она предоставляет достаточно точный результат, так как вырисовывает дерево вложенных кластеров, но при этом является более вычислительно дорогостоящей, так что её не рекомендуется использовать для кластеризации больших наборов данных. Пример дерева вложенных кластеров приведён на рис. 1.

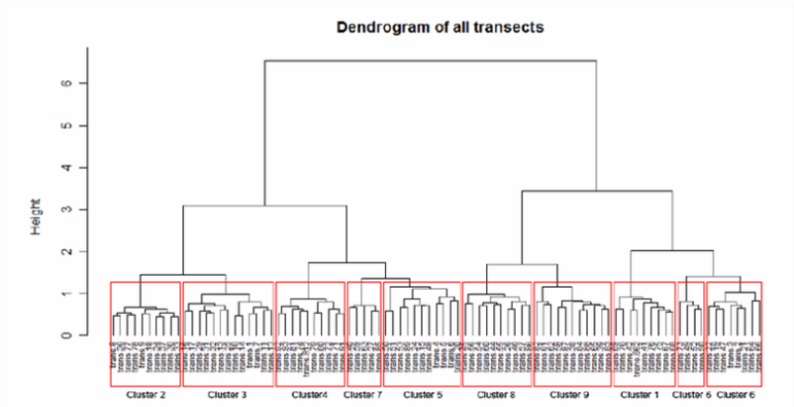

Рис. 1. Дендрограмма иерархической кластеризации [5]

Агломеративный иерархический метод характеризуется последовательным объединением исходных объектов и соответствующим уменьшением числа кластеров. На начальном этапе работы алгоритма все элементы являются кластерами. На первом шаге объекты, которые имеют схожие характеристики, объединяются в кластер. Далее объединение продолжается до тех пор, пока все объекты не будут составлять один кластер.

Дивизимный иерархический метод подразумевает собой последовательное разделение исходного кластера, состоящего из всех элементов, на кластеры меньшего размера, в результате чего образуется последовательность расщепляющих групп.

Метод k-средних – наиболее популярный метод кластерного анализа. Он стремится минимизировать суммарное квадратичное отклонение точек кластеров от центров этих кластеров. Данный метод разбивает множество элементов векторного пространства на заранее известное количество кластеров k. Суть метода k-средних такова: на каждой итерации перевычисляется центр масс для каждого кластера, который был получен на предыдущем шаге, затем векторы снова разбиваются на кластеры в соответствии с тем, какой из центров оказался ближе по выбранной метрике.

Пример результата кластеризации приведён на рис. 2.

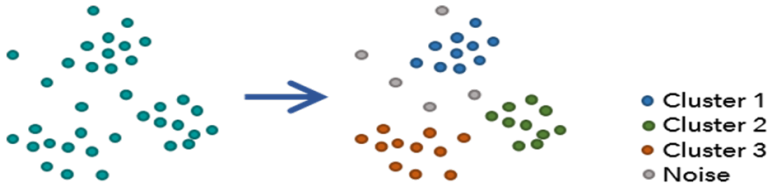

Рис. 2. Кластеризация на основе плотности [6]

Сравнительный анализ методов кластерного анализа будет проводиться на примере датасета Predict Future Sales (https://www.kaggle.com/competitions/competitive-data-science-predict-future-sales/data). Этот набор данных содержит ежедневные данные о продажах, предоставленные одной из крупнейших российских компаний – разработчиков программного обеспечения – фирмой «1С» – за период с 1 января 2013 г. по 31 октября 2015 г.

Набор данных состоит из 6 csv-файлов:

- `item_categories.csv`, который содержит в себе информацию о категориях товаров;
- `items.csv`, который включает в себя сведения о самих товарах;
- `sales_train.csv`, для обучения модели;
- `test.csv`, который будет проверочным набором для модели;
- `sample_submission.csv`, который является шаблоном для вывода результата.

Содержимое файла `sales_train.csv`, над данными из которого проводился кластерный анализ, представлено в табл. 1.

В данной работе были рассмотрены следующие методы кластеризации: агломеративная иерархическая кластеризация и метод k-средних.

Таблица 1. Содержимое файла для обучения модели

$date$	$datenum$	$shopid$	$itemid$	$itemprice$	$itemcntday$
02.01.2013	0	59	22154	999	1
03.01.2013	0	25	2552	899	1
06.01.2013	0	25	2554	1709.05	1
15.01.2013	0	25	2555	1099	1
10.01.2013	0	25	2564	349	1
02.01.2013	0	25	2565	549	1
04.01.2013	0	25	2572	239	1
11.01.2013	0	25	2572	299	1

С одной стороны, иерархическая кластеризация способна продемонстрировать дерево кластеров, что значительно упростит пользователю выбор количества необходимых кластеров, так как визуализированную информацию человеку проще воспринять. Но, с другой стороны, данное преимущество имеет один фатальный недостаток – при работе с большим объёмом данных процесс кластеризации длится слишком долго. Метод k-средних считается одним из самых быстрых, доступных на данный момент и достаточно прост в реализации.

Класс KMeans реализован в библиотеке sklearn (scikit-learn.org). Визуализацию результатов кластеризации берёт на себя менеджер графического интерфейса фигур pyplot из библиотеки matplotlib (matplotlib.org). Он обеспечивает неявный способ построения графиков подобно MATLAB. Графический интерфейс pyplot и пример результата кластеризации проданной компанией «1С» продукции за 2 января 2013 г. продемонстрированы на рис. 3.

Класс AgglomerativeClustering, как и KMeans, находится в библиотеке sklearn (scikit-learn.org). Результат агломеративной иерархической кластеризации представлен на рис. 4.

Кластерный анализ проводился по столбцам itemprice (цена товара) и itemcntday (количество проданных товаров).

За 2 января 2013 г. компания «1С» продала 6 684 предмета, которые были распределены двумя методами кластерного анализа по 5 кластерам.

Метод k-средних распределяет товары из набора данных в среднем за 0,06383 секунды следующим образом:

- 1 кластер – 55 элементов;
- 2 кластер – 4 029 элементов;
- 3 кластер – 48 элементов;
- 4 кластер – 1 661 элемент;
- 5 кластер – 891 элемент.

С другой стороны, агломеративный иерархический метод затрачивает в среднем 0,73307 секунды и даёт следующие результаты:

Рис. 3. Результат кластеризации всей продукции компании «1С» за 2 января 2013 г. методом KMeans

- 1 кластер – 855 элементов;
- 2 кластер – 49 элементов;
- 3 кластер – 4 930 элементов;
- 4 кластер – 28 элементов;
- 5 кластер – 822 элемента.

Как видно, ни один кластер одного метода кластерного анализа не соответствует какому-либо кластеру другого метода.

Ориентируясь на полученные результаты, можно подчеркнуть, что самым прибыльным кластером, выделенным агломеративным методом, является 1, а самым неприбыльным – 4. В то же время самый прибыльный кластер, определённый методом k-средних, – 5, а тот, что принёс меньше всего прибыли, – 1.

Также можно сделать вывод, что разные методы кластеризации дают разные результаты. Однако расположение и содержание некоторых кластеров может примерно совпадать, как, например, у 2 кластера по методу агломеративной кластеризации и 3 кластера по методу k-средних. Но дать точный ответ на вопрос: «Какой метод кластерного анализа из представленных справился лучше?» – может только специалист в данной области применения анализа данных, а точнее в области розничной торговли, поэтому основные параметры, по которым будут сравниваться алгоритмы в данной работе, – это скорость и удобство использования.

Если ориентироваться на скорость выполнения, метод агломеративной иерархической кластеризации обработал 6 684 элементов в 11,5 раз медленнее, чем метод k-средних, что значительно повлияет на комфортность и эффективность работы менеджера магазина розничной торговли. Для специалиста очень важно быстро прове-

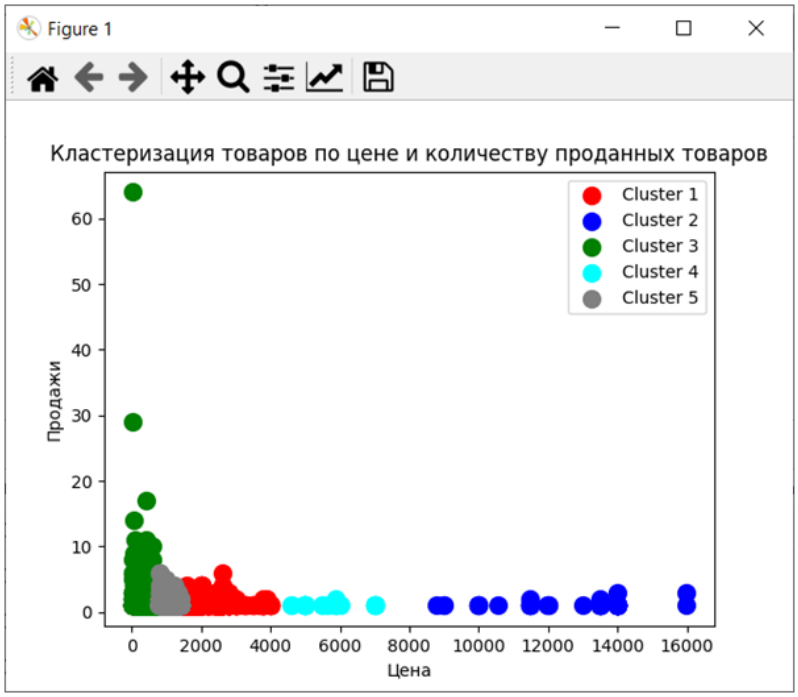

Рис. 4. Результат кластеризации всей продукции компании «1С» за 2 января 2013 г. методом AgglomerativeClustering

сти анализ сведений о проданной продукции и принять наиболее выгодное решение исходя из полученных результатов. Также в рамках анализа был проведён кластерный анализ 63 170 элементов (продажи компании «1С» за январь 2013 г.) методом k-средних. Кластеризация была проведена за 0,37588 секунды, что в 2 раза быстрее, чем кластеризация агломеративным методом набора данных из 6 684 элементов.

Заключение

Интеллектуальный анализ данных может быть использован в самых разных областях: от медицины и до экономики. Являясь одним из самых популярных методов ИАД, кластерный анализ способен решить задачу разделения данных на кластеры по некоторым характеристикам. Однако разные методы кластерного анализа обладают разной эффективностью и спецификой. Так метод k-средних в большей степени подходит для решения задачи кластеризации больших объёмов данных, с которыми иногда приходится иметь дело в решении задач в области розничной торговли. Он не требует такого большого количества временных ресурсов, как агломеративный иерархический метод, а значит, способен за то же время провести анализ большего количества данных, благодаря чему работа менеджера по продажам будет проходить гораздо эффективнее.

Литература

1. Остроух А.В., Николаев А.Б. Интеллектуальные информационные системы и технологии : моногр. СПб.: Лань, 2019.
2. Заболотникова В.С., Ромашкова О.Н. Анализ методов кластеризации для эффективного управления процессами в налоговой службе // Фундаментальные исследования. 2017. № 9-2. С. 303–307.
3. Ризаев И.С., Рахал Я. Интеллектуальный анализ данных для поддержки принятия решений. Казань: Школа, 2011.
4. Яковлева Е.А., Катермина Т.С., Шарич Э.Э., Яковлева Д.Д. Формирование потенциала финансовой системы для повышения инновационной активности // Вопросы инновационной экономики. 2019. Т. 9, № 2. С. 349–360.
5. Heiss M. The breeding bird communities of the Talish mountains (Azerbaijan) and their response to forest degradation: Diploma thesis. University of Greifswald, 2010. URL: `https://www.researchgate.net/figure/The-nine-breeding-bird-communities\\-of-the-Talish-mountains-represented-by-clusters-The\_fig2\_274533232` (дата обращения: 10.11.2023).
6. Find Point Clusters (GeoAnalytics). URL: `https://pro.arcgis.com/en/pro-app/latest/tool-reference/big-data-analytics/find-point-clusters.htm` (дата обращения: 10.11.2023).

ANALYSIS OF CLUSTERING TECHNIQUES TO SUPPORT RETAIL DECISION-MAKING

T.S. Katermina
Ph.D. (Techn.), e-mail: nggu-lib@mail.ru
A.A. Makambedzhan
Master's Degree, e-mail: aygul-mak@mail.ru

Nizhnevartovsk State University, Nizhnevartovsk, Russia

Abstract. The article considers theoretical bases of data mining, studies methods of cluster analysis on the example of the task of clustering data on sales from "1C". The methods considered apply to the task of supporting decision-making in various spheres, including retail trade. The article provides an overview of methods of cluster analysis, as well as an experiment with the application of agglomerative method of cluster analysis and k-means method. Advantages and disadvantages of these methods are revealed, results of modeling are given.

Keywords: artificial intelligence, cluster analysis, data mining.

Дата поступления в редакцию: 19.11.2023