Математические методы группирования данных для принятия управленческих решений в задачах планирования

Mathematical Methods of Grouping Data for Making Managerial Solutions in the Tasks of Planning

Чугунов В.Р. (Chugunov V.R.)[1], Жукова Л.В.(Zhukova L.V.)[2],
Ковальчук И.М. (Kovalchuk I.M.)[3], Ковалева А.С. (Kovaleva A.S.)[4]

[1]ЗАО «ЕС-лизинг» (JSC "EC-leasing"), руководитель отдела (head of department),
vchugunov@ec-leasing.ru
[2]ЗАО «ЕС-лизинг» (JSC "EC-leasing"), специалист (expert),
lvzhukova@mail.ru
[3]ЗАО «ЕС-лизинг» (JSC "EC-leasing"), ассистент (assistant),
ikovalchuk@ec-leasing.ru
[4]ЗАО «ЕС-лизинг» (JSC "EC-leasing"), ассистент (assistant),
princess.com@inbox.ru

Аннотация. Исключительно важная роль информации в современном мире привела к выявлению информации как собственного ресурса, столь же важного и необходимого, как энергетические, финансовые, сырьевые ресурсы. Потребности общества в сборе, хранении и обработке информации как товара создали новый спектр услуг – рынок информационных технологий. Объемы информации стремительно растут, предлагаются к анализу данные более терабайт, называемые Big Data.

Для решения задач управления на основе анализа таких данных необходимо учитывать их разнородность, высокую степень вариации. Поэтому систематизация и группировка полученной информации позволяет повысить качество принимаемых решений в задачах планирования и управления производством. При выборе методов группирования следует учитывать помимо типа поставленной задачи и большую размерность данных, которая сказывается на времени обработки информации.

В работе представлены результаты исследования методов группирования данных для некоторого диапазона практических задач по обработке больших данных, а также представлены результаты решения различных практических управленческих задач с использованием различных методов.

Abstract. The extremely important role of information in the modern world has led to the identification of information as an own resource, as important and necessary as energy, financial, raw materials. The needs of society in the collection, storage and processing of information as a commodity have created a new range of services – the information technology market. The volumes of information are growing rapidly, such kind of data volume is called "Big Data», and has been offered for analysis. In order to solve management problems based on

the analysis of such data, it is necessary to take into account their heterogeneity, high degree of variation. Therefore, the systematization and grouping of the information obtained makes it possible to improve the quality of the decisions made in the planning and production management tasks. In the process of choosing the grouping methods, the greater dimensionality of the data that affects the processing time of information should be taken into account besides the type of the task in hand.

This work presents the results of research of the methods of grouping data for a certain range of practical problems in the processing of large data, as well as the results of solving various practical management problems using various methods.

Ключевые слова: кластеризация, кластерный анализ, систематизация
Keywords: cluster analysis, cauterization, systematization

1 Введение

Вместе со стремительным накоплением информации быстрыми темпами развиваются и технологии анализа данных. Если еще несколько лет назад было возможно, скажем, лишь сегментировать клиентов на группы со схожими предпочтениями, то теперь возможно строить модели для каждого клиента в режиме реального времени, анализируя, например, его перемещение по сети Интернет для поиска конкретного товара. Интересы потребителя могут быть проанализированы в различных аспектах – наборы товаров, типичность поведения, реакция на новинки, отличие активных покупателей от не активных, и в соответствии с построенной моделью выведена подходящая реклама или конкретные предложения для каждого конкретного потребителя с учетом его особенностей на основании анализа большого объема информации обо всех клиентах. Модель должна учитывать как общие характеристики, так и индивидуальные для обособленной группы людей. В связи со стремительным развитием компьютерных технологий все чаще предлагается модифицировать и управлять моделью для тысяч клиентов или сотрудников, имеющих общие черты, но и некоторые существенные отличия, влияющие на их потребление.

При решении задач управления процесс обработки информации происходит несколько этапов:

1. Постановка бизнес-задачи
2. Постановка задачи обработки данных
3. Процесс моделирования, включающий сегментацию и разработку нескольких моделей
4. Верификация моделей и ее адаптация

Таким образом, в процессе моделирования встают различные задачи группировки, сортировки, разбиения на однородные группы или выделение групп с особыми специфическими характеристиками. Для реализации этих задач используются различные методы сегментирования и классификации, которые

зависят как от поставленной задачи, так и от формата исходных данных. При выборе метода приходится сталкиваться с обработкой очень больших объемов информации, в том числе и Big Data, представленных как количественными, так и качественными показателями.

Задача данной статьи – продемонстрировать алгоритм выбора метода анализа больших данных в задачах группировки и анализ применимости разных методов к большим данным. Приводятся примеры решения задач кластеризации различными методами на реальных данных.

2 Проблемы современного анализа

Термин Big Data характеризует совокупности данных с возможным экспоненциальным ростом их объема, которые слишком велики, слишком не форматированы или не структурированы для анализа традиционными методами.

В современных обсуждениях понятие Big Data описывают как данные порядка терабайт информации. Признаки Big Data определяются как «три V»: **volume** – объем; **variety** – разнородность, множество; **velocity** – скорость (необходимость очень быстрой обработки).

Данные типа Big Data обычно хранятся и организуются в распределенных файловых системах, они хранится на нескольких (иногда тысячах) жестких дисках, на стандартных компьютерах может не хватить места для всего массива информации, отсюда возникает одна из задач – агрегация и группирование для последующего анализа.

Большая проблема состоит в том, чтобы сформировать одинаковую модель для всех объектов на основе обработки Больших Данных, поскольку слишком большое количество разнообразных факторов и особенностей влияют на результаты. Эти данные обычно имеют множество подгрупп и высокую внутреннюю неоднородность, что может затруднить получение единой модели для результатов. Группирование данных в однородные группы и построение нескольких моделей с учетом особенностей групп позволяет выстроить более гибкую систему обработки.

Задача анализа больших данных обладает несомненной актуальностью, так как обработка таких данных представляет отдельный процесс, связанный с применением новых технологий в хранении и обработки данных. Авторы статьи предлагают использовать некоторые методы кластеризации, выбранные с учетом ниже описанных критериев, для предварительной кластеризации больших данных, уделяя внимание методам кластеризации с учетом особенностей самих больших данных.

3 Задачи кластерного анализа

Одним из способов систематизации и классификации является кластерный анализ. Он представляет собой набор методов, используемых для группировки объектов или событий в относительно однородные группы, которые называ-

ют кластерами (clusters). Объекты в каждом кластере должны быть похожи между собой и отличаться от объектов в других кластерах.

Кластерный анализ позволяет открыть в данных ранее неизвестные закономерности, которые практически невозможно исследовать другими способами и представить их в удобной для пользователя форме. Методы кластерного анализа используются как самостоятельные инструменты исследований, так и в составе других средств Data Mining (например, нейросетей см. [1])

Кластерный анализ используют для решения нескольких задач: определить однородные (гомогенные) группы, рассортировать объекты по набору характеристик или выделить особенные специфические объекты. При кластеризации учитывают тип показателей группировки – количественные или качественные.

Сложный для решения вопрос кластерного анализа — это вопрос о количестве кластеров, которые следует выбрать. Здесь нет твердых правил, позволяющих быстро принять решение. Некоторые методы кластерного анализа требуют задания этого значения, тогда он выбирается на основе поставленной задачи, например, для удобства последующей интерпретации результата или исходя из экономических свойств данных, а некоторые методы позволяют определить наилучшее значение количества кластеров, при котором достигается высокая однородность, на основе оптимизации по какому-то критерию (например, критерий Акаике).

Таким образом, при задании количества кластеров до анализа есть возможность получить в результате группы с более высокой чем возможно неоднородностью, а при автоматической процедуре выбора числа кластеров – сложно интерпретируемый результат.

Результат кластеризации – получение групп с наличием общих черт. Описание и профилирование кластеров осуществляется на основе анализа кластерных центроидов. Центроиды представляют средние значения объектов, содержащиеся в кластере по каждой из переменных. Они позволяют выявлять индивидуальные характеристики кластера как группы объектов, присвоить ему номер или метку, сформировать описание и построить управленческую модель для полученных групп.

Подход к выбору метода кластеризации основан на анализе следующей информации:

- поставленная управленческая задача;
- тип анализируемых данных и показателей кластеризации;
- объем исходных данных и имеющихся ресурсов для их обработки;
- требования к интерпретируемости и визуализации результатов анализа.

4 Обзор методов кластеризации

Существует два типа методов кластеризации: **иерархические** и **неиерархические**.

Особенность **иерархической** кластеризации — объединение или декомпозиция (разъединение) множества объектов путем перебора всех возможных

значений. Иерархические методы кластеризации могут быть агломеративными (объединяющими) или дивизивными (разъединяющими).

Агломеративные методы используют различные меры расстояний и включают: метод одиночной связи, метод полной связи и метод средней связи. Агломеративная кластеризация берет начало с каждого объекта в отдельном кластере, затем объекты группируются во все более крупные кластеры. Этот процесс будет идти до тех пор, пока все объекты не станут членами одного единственного кластера. Следует выделить также дивизивную кластеризацию, которая берет начало со всех объектов, являющихся сгруппированными в единственном кластере. Кластеры будут делить пока каждый объект не окажется в отдельном кластере. Особенность этих методов – перебор всех объектов, что требует большого количества времени и ресурсов, они практически не применимы для Big Data.

Неиерархические методы кластеризации часто называют методами *k*-средних. Суть этой группы методов – определить центр кластера, а в следующую очередь сгруппировать все объекты в пределах заданного от центра порогового значения. Эти методы предполагают перебор всех значений для определения расстояния между объектами, они включают последовательный пороговый метод, параллельный пороговый метод и оптимизирующее распределение.

$$\mathrm{V} = \sum_{i=1}^{k} \sum_{x_j \in s_i} (x_j - \mu_i)^2 \qquad (7)$$

где k — число кластеров, s_i — полученные кластеры, i=1,2,…,k и μ_i — центры масс векторов $x_j \in s_i$.

Метод оптимизирующего распределения будет иметь отличия от остальных пороговых методов в том, что при отнесении объекта к одному кластеру при нахождении объекта в пределах порогового значения впоследствии возможно включить его в другой кластер (перераспределить) с учетом порогового значения в целях оптимизации суммарного критерия, которым является среднее внутрикластерное расстояние, установленное для данного числа кластеров.

Из иерархических методов стоит выделить несколько самых распространенных Birch и CURE. Еще один распространенный метод **– двухшаговый или метод BIRCH,** благодаря обобщенным представлениям кластеров скорость кластеризации увеличивается, алгоритм при этом обладает большим масштабированием и возможностью применять его для данных большой размерности.

В этом алгоритме реализован двухэтапный процесс кластеризации. Первый этап заключается в формировании предварительного набора кластеров, т.е. определении количества кластеров. Следующий этап заключается в применении к выявленным кластерам других алгоритмов кластеризации, которые были бы пригодны в работе с оперативной памятью.

Среди новых масштабируемых алгоритмов можно отметить также алгоритм **CURE –** алгоритм иерархической кластеризации, где понятие кластера формулируется с использованием концепции плотности.

Над масштабируемыми методами сейчас активно работают многие исследователи, основная задача которых – преодолеть недостатки алгоритмов, существующих на сегодняшний день.

В таблице 1 перечислены достоинства и недостатки этих методов.

Table 1. «Сравнительная характеристика методов кластеризации»

Метод	Достоинства	Недостатки
k-средних (неиерархический)	простота использования, скорость процедуры кластеризации, понятность и прозрачность алгоритма, широкое распространение, понятность результата	высокая чувствительность к выбросам, медленная работа на больших данных, количественные показатели для кластеризации, практически невозможно применить к Big Data из-за большой требуемой емкости ресурсов памяти, необходимо задавать количество кластеров
BIRCH (иерархический)	двухступенчатая кластеризация позволяет анализировать большие объемы данных, как количественных, так и качественных, работает на ограниченном объеме памяти, не требует задания количества кластеров	хорошо определяет только кластеры выпуклой формы, есть необходимость в задании пороговых значений
CURE (иерархический)	выполняет кластеризацию даже на высоком уровне даже при наличии выбросов, выделяет кластеры сложной формы и разных размеров, требует большой размер памяти для данных большой размерности	есть необходимость в задании количества кластеров и пороговых значений
Самоорганизующиеся карты Кохонена	используется уникальный апроксиматор – нейронная сеть, простота реализации, обучение сети	необходимо задавать количество кластеров, работа только с числовыми данными, сложность интерпретации результатов

Таким образом, следует обратить внимание на алгоритм BRICH и его модификации, хорошо себя зарекомендовавшие в различного типа задач. Метод k-средних имеет множество ограничений по объему данных и типам переменных и должен быть использован с осторожностью.

5 Примеры задач группировки, кластеризации и сегментации

Рассмотрим несколько примеров применения кластерного анализа в решении практических задач:

1. Торговые точки Заказчика. Конечная цель анализа торговых точек (ТТ) – определение потенциала отдельной ТТ. Однако более информативно по сравнению с прямым анализом финансового состояния оказалось предварительное разбиение всех ТТ на однородные группы для выявления ТТ со схожими показателями человеко-потока вокруг них. На основании 15 характеристик местоположения и окружающей среды (расстояние до метро, до ЖД станции, до мест притяжения населения, количество различных торговых точек в округе и т.п.) выявлялись ТТ со схожими показателями человеко-потока вокруг них, что напрямую влияет на финансовые показатели продаж. Была сформулирована задача кластерного анализа таким образом: все показатели количественные, объем данных – не относящихся к Big Data (36 650 строк), обязательное требование интерпретируемости результатов (количество кластеров до 20-ти), нет информации об оптимальном количестве кластеров.

В результате применения метода двухступенчатой кластеризации BIRCH были получены 13 кластеров, характеризующих особенности местоположения однородных групп точек (ТТ в торговых зонах на окраинах городов, ТТ в торговых зонах в центре городов, ТТ в удаленных районах и т.п.), описание которых по центроидам позволило дать характеристики местонахождения ТТ и определить торговую привлекательность по человеко-потоку согласно значению показателей, свидетельствующих об:

- экономической активности (количество и расстояние до ближайших мест притяжения населения);
- торговой активности и наличия торговых зон (количество и расстояние до ближайших ТТ ka-сетей и не ka-сетей);
- наличия транспортной инфраструктуры (количество и расстояние до ближайших объектов типа остановок, вокзалов, метро).

На основании полученной кластеризации были определены зоны привлекательности местоположения ТТ, финансовые показатели кластеров и их доходность, потенциалы продаж и выявлены группы ТТ с учетом их местоположения с крайне низкими доходами.

Для сравнения была проведена кластеризация методом самоорганизующихся карт Кохонена (раздел методологии нейронных сетей), в результате объекты были разделены на группы, интерпретация которых оказалась затруднена с точки зрения профиля кластера. В результате пришлось отказаться от результатов из-за их не интерепретируемости и перейти к BIRCH, результаты которого оказались экономически объяснимы (зоны ТТ), и востребованы заказчиком для дальнейшего управления.

2. Анализ ассортимента. Продаваемые товары сгруппированы обычно по товарным группам, которые однако не отражают ни частоту продаж, ни объем сбыта. Была поставлена задача выявления нескольких типов ассортимента с учетом объемов разовых продаж и частоты продаж в течении года по данным о продажах за год 40 различных подкатегорий товаров по ТТ. Она была сформулирована следующим образом: данные по ассортименту представляют собой

целые числа от 1 до 12 – количество месяцев, в которые товар продается. Всего 650 наименований товаров и примерно 36 000 строк – торговых точек. Такой объем данных (более 20 млн значений) представляет собой данные большой размерности.

Кластеризация двухшаговым методом позволила определить оптимальное количество кластеров, в результате выявилось 4 кластера ассортимента среди ТТ по набору товаров – ассортимент гипермаркетов, ассортимент супермаркетов, ассортимент небольших магазинов и ассортимент специализированных магазинов. Отличительная особенность результата – группировка товаров не по группам продаж или категориям, а по совокупности ассортимента.

В результате кластеризации были выстроены новые стратегии работы с ассортиментом товара в зависимости от его направленности, результаты были учтены при определении потенциала продаж ТТ с учетом особенностей типа ассортимента.

Кластеризация методом k-средних оказалась невозможной в связи с большим требуемым ресурсом памяти рабочего персонального компьютера. Также стоит отметить, что k-средние предполагают использование количественных характеристик, поэтому этот метод не соответствует типу исходных данных (целые значения от 1 до 12 (частота продаж) скорее относятся к типу порядковых значений и не могут быть классифицированы как шкалы) и поставленной задаче разбиения на однородные группы с неизвестным количеством кластеров.

При выборе типа метода и метода кластеризации следует учитывать несколько параметров:

- тип переменных для классификации (количественные или качественные данные – от этого зависит, какие методы могут не подходить, как например *k*-средние только в случае количественных характеристик),
- известна ли информация о количестве кластеров (если нет – то двухшаговый BIRCH может определить это значение),
- количество классифицируемых объектов и, соответственно, сложность вычислительной процедуры с точки зрения время затрат (при большой размерности стоит отдавать предпочтение неиерархическим методам),
- интерпретируемость получаемого результата (при достижении высокой однородности в группе объектов число таких групп может превышать 20, что значительно затрудняет ее интерпретацию)
- возможность настройки кластеризации (в случае необходимости задания вычислительной процедуры в виде машинного кода следует учитывать простоту (как например *k*-средние) или сложность (например, карты Кохонена) алгоритма для автоматизации и вычисления.

Таким образом, в решении практических задач можно встретить несколько подходов: для кластеризации небольшого массива количественных данных при известном значении количества кластеров (или хотя бы диапазона количества) можно использовать алгоритм к-средних или его модификации.

Для решения задачи кластеризации по количественным и качественным показателям большой размерности двухшаговый метод BIRCH позволит определить оптимальное количество кластеров и разгруппирует на однородные группы.

При решении задач в рамках автоматизации процесса, когда вывод и интерпретация результатов не так важна, и есть доступ к специальным программным продуктам, могут применяться самоорганизующиеся карты Кохонена.

6 Выводы

Кластеризация как метод группирования данных может применяться в решении широкого класса задач:

1. как промежуточный этап моделирования для построения более адекватной и гибкой модели (моделей) за счет повышения однородности данных, путем разбиения их на группы;
2. как первичный способ отбора однородных данных для последующего анализа в случае Big Data, например, отбор активных ТТ, занимающихся прямыми продажами;
3. как метод решения задач сегментации при задаче разбиения, выявления типов со схожими характеристиками, например, определение групп ТТ по ассортименту, или групп пользователей по модели поведения на сайте;
4. как метод выделения из общей массы отдельной, более важной с точки зрения бизнеса не похожей на остальные группы (например, группа особо активных подписчиков).

В зависимости от параметров и характеристик исследуемых признаков применяются различные методы кластеризации (К-средних, BIRCH и другие).

Литература

1. Н. Паклин «Алгоритмы кластеризации на службе Data Mining», BaseGroup Labs, URL: http://www.basegroup.ru/clusterization/datamining.htm (дата обращения: 17.07.2017).
2. Олендерфер М. С., Блэшфилд Р. К. Кластерный анализ / Факторный, дискриминантный и кластерный анализ: пер. с англ.; Под. ред. И. С. Енюкова. — М.: Финансы и статистика, 1989
3. Tian Zhang, Raghu Ramakrishnan, Miron Livny BIRCH: An Efficient Data Clustering Method for Very Large Databases // ACM SIGMOD International Conference on Management of Data. 1996 [PDF] (http://citeseer.ist.psu.edu/zhang96birch.html)
4. Майер-Шенбергер В, Кукьер К. «Большие данные: революция, которая изменит нашу жизнь, работу и мысли» И: Манн, Иванов и Фербер, 2013
5. James Manyika et al. Big data: The next frontier for innovation, competition, and productivity (англ.). McKinsey Global Institute, June, 2011.