МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ УКРАИНЫ
Одесский национальный университет имени И. И. Мечникова
Институт математики, экономики и механики
Кафедра математического обеспечения компьютерных систем

РЕФЕРАТ

На тему:
«Методы крастеризации и их основные виды»

Выполнил студент
6 курса ФИТ,
Царюк А.О.

Преподаватель
Петрушина Т. И.

Одесса – 2018

Введение

Кластерный анализ (англ. Data clustering) — задача разбиения заданной выборки объектов (ситуаций) на подмножества, называемые кластерами, так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались. Задача кластеризации относится к статистической обработке, а также к широкому классу задач обучения без учителя. **Кластерный анализ** — это многомерная статистическая процедура, выполняющая сбор данных, содержащих информацию о выборке объектов, и затем упорядочивающая объекты в сравнительно однородные группы (кластеры)(Q-кластеризация, или Q-техника, собственно кластерный анализ). **Кластер** — группа элементов, характеризуемых общим свойством, главная цель кластерного анализа — нахождение групп схожих объектов в выборке [9. с-3].

Кластерный анализ применяют в различных областях человеческой деятельности: медицина, химия, психология, управление и во многом другом. Кластерный анализ выполняет следующие *основные задачи*:

Разработка типологии или классификации; исследование полезных концептуальных схем группирования объектов; порождение гипотез на основе исследования данных, проверка гипотез или исследования для определения, действительно ли типы (группы), выделенные тем или иным способом, присутствуют в имеющихся данных.

Глава 1. Кластерный анализ

Термин кластерный анализ, впервые введенный Трионом (Tryon) в 1939 году, включает в себя более 100 различных алгоритмов.

В отличие от задач классификации, кластерный анализ не требует априорных предположений о наборе данных, не накладывает ограничения на представление исследуемых объектов, позволяет анализировать показатели различных типов данных (интервальным данным, частотам, бинарным данным). При этом необходимо помнить, что переменные должны измеряться в сравнимых шкалах. [1. с-4]

Работа кластерного анализа опирается на два предположения. Первое предположение - рассматриваемые признаки объекта в принципе допускают желательное разбиение пула (совокупности) объектов на кластеры. Второе предположение - правильность выбора масштаба или единиц измерения признаков.

Методы кластерного анализа можно разделить на две группы:

1. иерархические;
2. неиерархические.

Каждая из групп включает множество подходов и алгоритмов. Используя различные методы кластерного анализа, аналитик может получить различные решения для одних и тех же данных. Это считается нормальным явлением.

Рассмотрим иерархические и неиерархические методы подробно.

Алгоритм кластерного анализа k-средних (k-means)

Наиболее распространен среди неиерархических методов алгоритм k-средних, также называемый **быстрым кластерным анализом**. Полное описание алгоритма можно найти в работе Хартигана и Вонга (Hartigan and Wong, 1978). В отличие от иерархических методов, которые не требуют

предварительных предположений относительно числа кластеров, для возможности использования этого метода необходимо иметь гипотезу о наиболее вероятном количестве кластеров.

Алгоритм k-средних строит k кластеров, расположенных на возможно больших расстояниях друг от друга. Основной тип задач, которые решает алгоритм k-средних, - наличие предположений (гипотез) относительно числа кластеров, при этом они должны быть различны настолько, насколько это возможно. Выбор числа k может базироваться на результатах предшествующих исследований, теоретических соображениях или интуиции.

Общая идея алгоритма: заданное фиксированное число k кластеров наблюдения сопоставляются кластерам так, что средние в кластере (для всех переменных) максимально возможно отличаются друг от друга. [5. с-68-73]

1.1.1.Описание алгоритма

1. *Первоначальное распределение объектов по кластерам*.

Выбирается число, именуемое k, и эти точки считаются "центрами" кластеров. Каждому кластеру соответствует один центр.

Выбор начальных центров осуществляется следующим образом:

1. выбор k-наблюдений для максимизации начального расстояния;

2. случайный выбор k-наблюдений;

3. выбор первых k-наблюдений.

2. *Итеративный процесс*

Вычисляются центры кластеров, которыми затем считаются покоординатные средние кластеров. Объекты перераспределяются.

Процесс вычисления центров и перераспределения объектов продолжается до тех пор, пока не выполнено одно из условий:

1. кластерные центры стабилизировались, т.е. все наблюдения принадлежат кластеру, которому принадлежали до текущей итерации;

2. число итераций равно максимальному числу итераций.

На рис. 1 приведен пример работы алгоритма k-средних для k, равного двум.

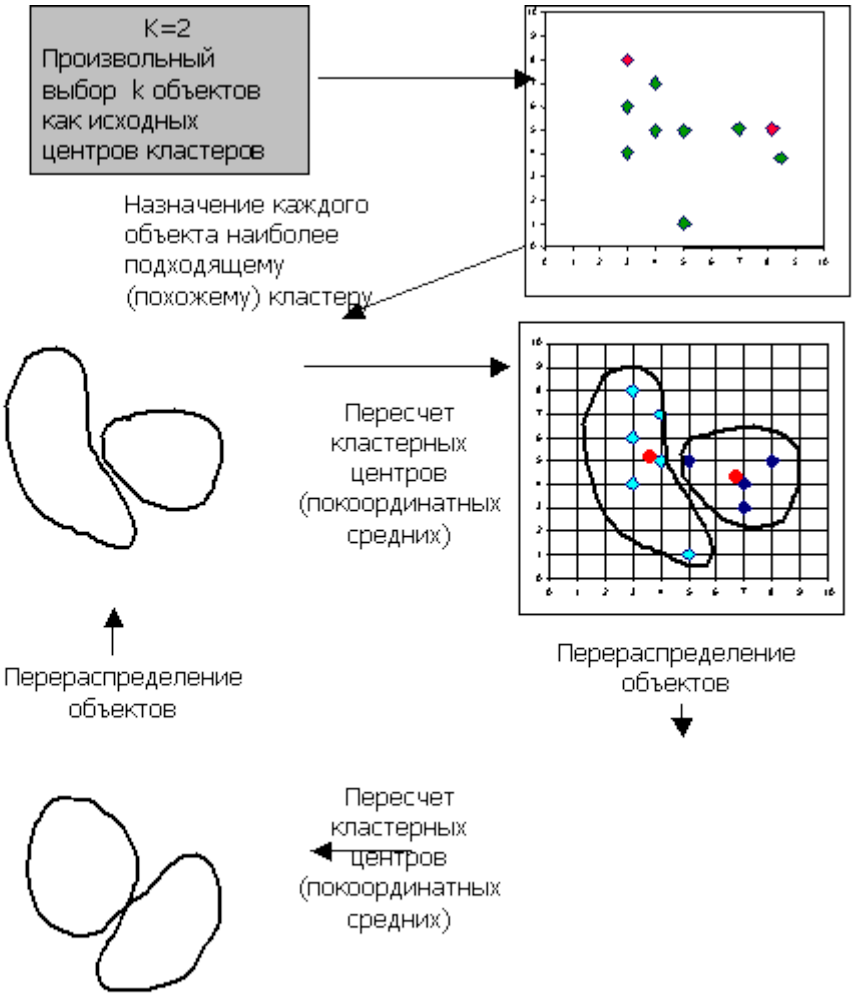

Рис. 1. Пример работы алгоритма k-средних (k=2) [5. с-68-73]

1.1.2. Проверка качества кластеризации

После получений результатов кластерного анализа методом k-средних следует проверить правильность кластеризации (т.е. оценить, насколько кластеры отличаются друг от друга). Для этого рассчитываются средние значения для каждого кластера. При хорошей кластеризации должны быть получены сильно отличающиеся средние для всех измерений или хотя бы большей их части.

***Достоинства алгоритма k-средних**:*

1. простота использования;

2. быстрота использования;

3. понятность и прозрачность алгоритма.

***Недостатки алгоритма k-средних**:*

1. алгоритм слишком чувствителен к выбросам, которые могут искажать среднее. Возможным решением этой проблемы является использование модификации алгоритма - алгоритм k-медианы;

2. алгоритм может медленно работать на больших базах данных.

Возможным решением данной проблемы является использование выборки данных. [8. с-6]

Алгоритм PAM (partitioning around Medoids)

PAM является модификацией алгоритма k-средних, алгоритмом k-медианы (k-medoids). Алгоритм менее чувствителен к шумам и выбросам данных, чем алгоритм k-means, поскольку медиана меньше подвержена влияниям выбросов. PAM эффективен для небольших баз данных, но его не следует использовать для больших наборов данных.

1.1. Сложности, возникающие при кластерном анализе

Как и любые другие методы, методы кластерного анализа имеют определенные слабые стороны, т.е. некоторые сложности, проблемы и ограничения.

При проведении кластерного анализа следует учитывать, что результаты кластеризации зависят от критериев разбиения совокупности исходных данных. При понижении размерности данных могут возникнуть

определенные искажения, за счет обобщений могут потеряться некоторые индивидуальные характеристики объектов.

Существует ряд сложностей, которые следует продумать перед проведением кластеризации.

1. Сложность выбора характеристик, на основе которых проводится кластеризация. Необдуманный выбор приводит к неадекватному разбиению на кластеры и, как следствие, - к неверному решению задачи.

2. Сложность выбора метода кластеризации. Этот выбор требует неплохого знания методов и предпосылок их использования. Чтобы проверить эффективность конкретного метода в определенной предметной области, целесообразно применить следующую процедуру: рассматривают несколько априори различных между собой групп и перемешивают их представителей между собой случайным образом. Далее проводится кластеризация для восстановления исходного разбиения на кластеры. Доля совпадений объектов в выявленных и исходных группах является показателем эффективности работы метода.

3. Проблема выбора числа кластеров. Если нет никаких сведений относительно возможного числа кластеров, необходимо провести ряд экспериментов и, в результате перебора различного числа кластеров, выбрать оптимальное их число.

4. Проблема интерпретации результатов кластеризации. Форма кластеров в большинстве случаев определяется выбором метода объединения. Однако следует учитывать, что конкретные методы стремятся создавать кластеры определенных форм, даже если в исследуемом наборе данных кластеров на самом деле нет. [5. с-68-73]

1.4. Сравнительный анализ иерархических и неиерархических методов кластеризации

Неиерархические методы выявляют более высокую устойчивость по отношению к шумам и выбросам, некорректному выбору метрики, включению незначимых переменных в набор, участвующий в кластеризации. Ценой, которую приходится платить за эти достоинства метода, является слово "априори". Аналитик должен заранее определить количество кластеров, количество итераций или правило остановки, а также некоторые другие параметры кластеризации. Это сложно начинающим специалистам.

Если нет предположений относительно числа кластеров, рекомендуют использовать иерархические алгоритмы кластерного анализа. Однако если объем выборки не позволяет это сделать, возможный путь - проведение ряда экспериментов с различным количеством кластеров, например, начать разбиение совокупности данных с двух групп и, постепенно увеличивая их количество, сравнивать результаты. За счет такого "варьирования" результатов достигается достаточно большая гибкость кластеризации.

Иерархические методы, в отличие от неиерархических, отказываются от определения числа кластеров, а строят полное дерево вложенных кластеров. Сложности иерархических методов кластеризации: ограничение объема набора данных; выбор меры близости; негибкость полученных классификаций. Преимущество этой группы методов в сравнении с неиерархическими методами - их наглядность и возможность получить детальное представление о структуре данных.

Глава 2. Алгоритмы кластерного анализа

В последнее время ведутся активные разработки новых алгоритмов кластеризации, способных обрабатывать сверхбольшие базы данных. В них основное внимание уделяется масштабируемости. К таким алгоритмам относятся обобщенное представление кластеров (summarized cluster representation), а также выборка и использование структур данных, поддерживаемых нижележащими СУБД. Разработаны алгоритмы кластерного анализа, в которых методы иерархической кластеризации интегрированы с другими методами. К таким алгоритмам относятся: BIRCH, CURE, CHAMELEON, ROCK. [5. с-68-73]

2.1. Алгоритм BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)

Алгоритм предложен Тьян Зангом и его коллегами. Благодаря обобщенным представлениям кластеров, скорость кластеризации увеличивается, алгоритм при этом обладает большим масштабированием. В этом алгоритме реализован двухэтапный процесс кластеризации.

В ходе первого этапа формируется предварительный набор кластеров. На втором этапе к выявленным кластерам применяются другие алгоритмы кластерного анализа - пригодные для работы в оперативной памяти.

Если каждый элемент данных представить себе как бусину, лежащую на поверхности стола, то кластеры бусин можно "заменить" теннисными шариками и перейти к более детальному изучению кластеров теннисных шариков. Число бусин может оказаться достаточно велико, однако диаметр теннисных шариков можно подобрать таким образом, чтобы на втором этапе можно было, применив традиционные алгоритмы кластерного анализа, определить действительную сложную форму кластеров.

2.2. Алгоритм WaveCluster

WaveCluster представляет собой алгоритм кластеризации на основе волновых преобразований. В начале работы алгоритма данные обобщаются

путем наложения на пространство данных многомерной решетки. На дальнейших шагах алгоритма анализируются не отдельные точки, а обобщенные характеристики точек, попавших в одну ячейку решетки. В результате такого обобщения необходимая информация умещается в оперативной памяти. На последующих шагах для определения кластеров алгоритм применяет волновое преобразование к обобщенным данным.

Особенности WaveCluster:

сложность реализации

алгоритм может обнаруживать кластеры произвольных форм

алгоритм не чувствителен к шумам

алгоритм применим только к данным низкой размерности

2.3. Алгоритмы кластерного анализа Clarans, CURE, DBScan

Алгоритм Clarans (Clustering Large Applications based upon RANdomized Search) формулирует задачу кластеризации как случайный поиск в графе. В результате работы этого алгоритма совокупность узлов графа представляет собой разбиение множества данных на число кластеров, определенное пользователем. "Качество" полученных кластеров определяется при помощи критериальной функции. Алгоритм Clarans сортирует все возможные разбиения множества данных в поисках приемлемого решения. Поиск решения останавливается в том узле, где достигается минимум среди предопределенного числа локальных минимумов.

2.4. Алгоритм CLARA (Clustering LARge Applications)

Алгоритм CLARA был разработан Kaufmann и Rousseeuw в 1990 году для кластеризации данных в больших базах данных. Данный алгоритм строится в статистических аналитических пакетах, например, таких как S+.

Изложим кратко суть алгоритма. Алгоритм CLARA извлекает множество образцов из базы данных. Кластеризация применяется к каждому из

образцов, на выходе алгоритма предлагается лучшая кластеризация.

Для больших баз данных этот алгоритм эффективнее, чем алгоритм PAM. Эффективность алгоритма зависит от выбранного в качестве образца набора данных. Хорошая кластеризация на выбранном наборе может не дать хорошую кластеризацию на всем множестве данных. [9. с-3].

2.5. Итеративная кластеризация в SPSS

Обычно в статистических пакетах реализован широкий арсенал методов, что позволяет сначала провести сокращение размерности набора данных (например, при помощи факторного анализа), а затем уже собственно кластеризацию (например, методом быстрого кластерного анализа). Рассмотрим этот вариант проведения кластеризации в пакете SPSS.

Выбираем в меню: Analyze (Анализ) / Data Reduction (Преобразование данных) / Factor (Факторный анализ). При помощи кнопки Extraction (Отбор) следует выбрать метод отбора. Мы оставим выбранный по умолчанию анализ главных компонентов, который упоминался выше. Таже следует выбрать метод вращения - выберем один из наиболее популярных - метод варимакса. Для сохранения значений факторов в виде переменных в закладке "Значения" необходимо поставить отметку "Save as variables" (Сохранить как переменные).

В результате этой процедуры пользователь получает отчет "Объясненная суммарная дисперсия", по которой видно число отобранных факторов - это те компоненты, собственные значения которых превосходят единицу.

Полученные значения факторов, которым обычно присваиваются названия fact1_1, fact1_2 и т.д., используем для проведения кластерного анализа методом k-средних. Для проведения быстрого кластерного анализа выберем в меню Analyze (Анализ) / Classify (Классифицировать) / K-Means Cluster: (Кластерный анализ методом k-средних).

В диалоговом окне K Means Cluster Analysis (Кластерный анализ методом k-средних) необходимо поместить факторные переменные fact1_1,

fact1_2 и т.д. в поле тестируемых переменных. Здесь же необходимо указать количество кластеров и количество итераций.

В результате этой процедуры получаем отчет с выводом значений центров сформированных кластеров, количестве наблюдений в каждом кластере, а также с дополнительной информацией, заданной пользователем.

Алгоритм k-средних делит совокупность исходных данных на заданное количество кластеров. Для возможности визуализации полученных результатов следует воспользоваться одним из графиков, например, диаграммой рассеивания. Традиционная визуализация возможна для ограниченного количества измерений, ибо, как известно, человек может воспринимать только трехмерное пространство. Если мы анализируем более трех переменных, следует использовать специальные многомерные методы представления информации, о них будет рассказано в одной из последующих лекций курса.

Итеративные методы кластеризации различаются выбором параметров:

1. начальной точки
2. правилом формирования новых кластеров
3. правилом остановки

В пакете SPSS, например, при необходимости работы как с количественными (например, доход), так и с категориальными (например, семейное положение) переменными, а также если объем данных достаточно велик, используется метод Двухэтапного кластерного анализа, который представляет собой масштабируемую процедуру кластерного анализа, позволяющую работать с данными различных типов.

На первом этапе работы записи предварительно кластеризуются в большое количество суб-кластеров. На втором этапе полученные суб-кластеры группируются в необходимое количество. Если это количество неизвестно, процедура сама автоматически определяет его. [5. с-75-77]

В общем случае все этапы кластерного анализа взаимосвязаны, и решения, принятые на одном из них, определяют действия на последующих

этапах.

Аналитику следует решить, использовать ли все наблюдения либо же исключить некоторые данные или выборки из набора данных.

По мнению многих специалистов, выбор метода кластеризации является решающим при определении формы и специфики кластеров.

Анализ результатов кластеризации. Этот этап подразумевает решение таких вопросов: не является ли полученное разбиение на кластеры случайным; является ли разбиение надежным и стабильным на под выборках данных; существует ли взаимосвязь между результатами кластеризации и переменными, которые не участвовали в процессе кластеризации; можно ли интерпретировать полученные результаты кластеризации.

Проверка результатов кластеризации. Результаты кластеризации также должны быть проверены формальными и неформальными методами. Формальные методы зависят от того метода, который использовался для кластеризации.

Неформально включают следующие процедуры проверки качества кластеризации:

1. анализ результатов кластеризации, полученных на определенных выборках набора данных

2. кросс-проверка

3. проведение кластеризации при изменении порядка наблюдений в наборе данных

4. проведение кластеризации при удалении некоторых наблюдений

5. проведение кластеризации на небольших выборках

Один из вариантов проверки качества кластеризации - использование нескольких методов и сравнение полученных результатов. Отсутствие подобия не будет означать некорректность результатов, но присутствие похожих групп считается признаком качественной кластеризации. [2. с-2-3].

2.6. Кластеризация в Data Mining

Кластеризация в Data Mining приобретает ценность тогда, когда она выступает одним из этапов анализа данных, построения законченного аналитического решения. Аналитику часто легче выделить группы схожих объектов, изучить их особенности и построить для каждой группы отдельную модель, чем создавать одну общую модель на всех данных. Таким приемом постоянно пользуются в маркетинге, выделяя группы клиентов, покупателей, товаров и разрабатывая для каждой их них отдельную стратегию.

Очень часто данные, с которыми сталкивается технология Data Mining, имеют следующие важные особенности:

1. высокая размерность (тысячи полей) и большой объем (сотни тысяч и миллионы записей) таблиц баз данных и хранилищ данных (сверхбольшие базы данных)

2. наборы данных содержат большое количество *числовых* и *категорийных* атрибутов

Все атрибуты, или признаки объектов делятся на *числовые* (numerical) и *категорийные* (categorical). Числовые атрибуты – это такие, которые могут быть упорядочены в пространстве, соответственно категорийные – которое не могут быть упорядочены. Например, атрибут "возраст" – числовой, а "цвет" – категорийный. Приписывание атрибутам значений происходит во время измерений выбранным типом шкалы, а это, вообще говоря, представляет собой отдельную задачу.

Большинство алгоритмов кластеризации предполагают сравнение объектов между собой на основе некоторой меры близости (сходства). Мерой близости называется величина, имеющая предел и возрастающая с увеличением близости объектов. Меры сходства "изобретаются" по специальным правилам, а выбор конкретных мер зависит от задачи, а также от шкалы измерений. В качестве меры близости для числовых атрибутов очень часто используется *евклидово расстояние*, вычисляемое по формуле

$$D(x,y) = \sqrt{\sum_i (x_i - y_i)^2}$$

Для категорийных атрибутов распространена мера сходства *Чекановского-Серенсена и Жаккара*. Потребность в обработке больших массивов данных в Data Mining привела к формулированию требований, которым должен удовлетворять алгоритм кластеризации:

- Минимально возможное количество проходов по базе данных
- Работа в ограниченном объеме оперативной памяти компьютера
- Работу алгоритма можно прервать с сохранением промежуточных результатов, чтобы продолжить вычисления позже
- Алгоритм должен работать, когда объекты из базы данных могут извлекаться только в режиме однонаправленного курсора

Алгоритм, удовлетворяющий данным требованиям (особенно второму), будем называть *масштабируемым* (scalable). *Масштабируемость* – важнейшее свойство алгоритма, зависящее от его вычислительной сложности и программной реализации. Алгоритм называют масштабируемым, если при неизменной емкости оперативной памяти с увеличением числа записей в базе данных время его работы растет линейно. На заре становления теории кластерного анализа вопросам масштабируемости алгоритмов внимания практически не уделялось. Предполагалось, что все обрабатываемые данные будут умещаться в оперативной памяти, главный упор всегда делался на улучшение качества кластеризации. Трудно соблюсти баланс между высоким качеством кластеризации и масштабируемостью. Поэтому в идеале в арсенале Data Mining должны присутствовать как эффективные алгоритмы кластеризации микромассивов (microarrays), так и масштабируемые для обработки сверхбольших баз данных (large databases). [10. с-7].