МИНИСТЕРСТВО ОБРАЗОВАНИЯ АЗЕРБАЙДЖАНСКОЙ РЕСПУБЛИКИ

АЗЕРБАЙДЖАНСКИЙ ГОСУДАРСТВЕННЫЙ ЭКОНОМИЧЕСКИЙ УНИВЕРСИТЕТ

«ЦЕНТР МАГИСТРАТУРЫ»

На правах рукописи

МАМЕДОВА ЭЛЬМИРА МАХИР

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

НА ТЕМУ:

«ИССЛЕДОВАНИЕ DATA MINING - ТЕХНОЛОГИЙ, ОСНОВАННЫХ НА КЛАСТЕРНОМ АНАЛИЗЕ»

Наименование и шифр специальности: 060509 «Компьютерные науки»

Наименование и шифр специализации: İİM020004 «Информационные технологии управления»

Научный руководитель: **к.ф.-м.н., доц. АБДУЛЛАЕВ А.Х.**

Руководитель

магистерской программы: **к.ф.-м.н., доц. АЛИЕВА Т.А.**

Заведующий кафедрой: **акад. АББАСОВ А.М.**

БАКУ – 2018

Содержание

Введение

Актуальность темы. Кластеризация - стандартная процедура многомерного анализа данных, предназначенная для изучения естественной структуры объектов данных, где объекты в одном кластере максимально похожи, а объекты в разных кластерах настолько разнообразны, насколько это возможно. Выявленные в ходе анализа кластеры, являются средством обобщения объектов данных и их особенностей. Кластеризация часто является одним из первых шагов Data Mining. Методы кластеризации применяются во многих областях, таких как медицина, психология, социология, экономика, распознавание образов, и т.д. Люди часто выполняют задачу кластеризации бессознательно; например, при просмотре двумерной карты автоматически распознаются разные области в зависимости от того, насколько близко к каждому расположены другие места, независимо от того, разделены ли места реками, озерами или морем и т. д. Однако, если описание объектов по их характеристикам достигает более высоких измерений, интуитивные суждения менее легко получить и оправдать. Термин кластеризация часто путают с классификацией или дискриминантным анализом. Однако, эти три вида анализа данных имеют существенные отличия. Так, классификация присваивает объекты уже определенным классам, тогда как для кластеризации нет необходимости в априорных знаниях об объектных классах и их членах. В свою очередь, дискриминантный анализ направлен на улучшение уже предоставленной классификации за счет усиления демаркаций классов, тогда как кластерный анализ должен сначала установить структуру класса.

Предмет и объект исследования. Предметом исследования является математические методы, модели и алгоритмы кластерного анализа. Объектом исследования выступает применение методов кластерного анализа к геопространственным данным.

Основная цель и задачи исследования. Целью данной работы является: выявление основных алгоритмов кластерного анализа, как инструмента Data Mining –технологий, исследование механизма применения кластеризации.

Научная новизна. Научная новизна исследования заключается в научном обосновании роли кластеризации при интеллектуальном анализе данных, в обосновании теоретического и практического применения кластерного анализа как начальный шаг при анализе больших данных.

Теоретико-методологической основой исследования является геокластерный анализ более 100 тысячи объектов, применение которого способно улучшить работу предприятий во многих сферах деятельности, так или иначе имеющие в информацию об географическом положении объектов.

Практическая значимость данного исследования состоит в том, что результаты работы могут быть использованы аналитиками в процессе создания структуры анализа данных, и его дальнейшего осуществления. Полученные результаты позволяют сделать вывод о наличии перспектив применения кластерного анализа. Выдвинутые предложения и рекомендации могут быть полезны при усовершенствовании алгоритмов кластеризации.

Объем и структура исследования. Работа состоит из введения, трех глав, 10 подглав, 18 рисунков, предложения и рекомендация, списка использованной литературы и интернет-ссылок, заключения на азербайджанском и английских языках.

В I главе рассматриваются этапы интеллектуального анализа данных, его архитектура, методы и задачи. Устанавливается значимость кластеризации, как задачи Data Mining-технологий. Рассматриваются инструменты Data Mining-технологий, особый акцент делается на инструменты, в основе которых лежит кластерный анализ [3,4,6].

Во II главе рассматривается сущность кластерного анализа. Определение меры сходства, позволяющей более точно выбрать алгоритм

кластеризации. Типы и методы кластеризации. Указываются преимущества использования определенных методов в заданной сфере [1,2,5].

В III главе на конкретном примере показана возможность применения кластерного анализа к большому массиву данных. Выявляются проблемы, возникающие при использовании кластеризации. Требования к кластеризации, способные предотвратить данные проблемы. Рассматриваются причины практического применения кластерного анализа [8,9,10,11,12,21].

I ГЛАВА. ЗАДАЧИ, СТАДИИ И МЕТОДЫ DATA MİNİNG-ТЕХНОЛОГИЙ

1.1. Этапы интеллектуального анализа данных

Достижения в области технологий распознавания и хранения данных, а также значительный рост таких технологий, как поиск в Интернете, цифровое изображение и видеонаблюдение, создали множество высокопроизводительных наборов данных с большими объемами. Большая часть данных хранится в цифровом виде на электронных носителях, что дает огромный потенциал для разработки методов автоматического анализа, классификации и поиска данных. В дополнение к росту объема данных также увеличилось разнообразие доступных данных (текст, изображение и видео). Недорогие цифровые и видеокамеры сделали доступными огромные архивы изображений и видео. Распространенность RFID-меток или транспондеров из-за их низкой стоимости и небольших размеров привела к развертыванию миллионов датчиков, способных передавать данные в режиме реального времени. Электронные письма, блоги, данные транзакций и миллиарды веб-страниц каждый день создают терабайты новых данных. Многие из этих потоков данных неструктурированы, что затрудняет их анализ.

Полученное за последнее десятилетие внушительное увеличение мощности и скорости обработки данных, позволило науке перейти от ручных, трудоемких и рутинных действий к быстрому, легкому и автоматизированному анализу данных. В то время как основная проблема технологов баз данных заключалась в том, чтобы найти эффективные способы хранения, извлечения и обработки данных, основная проблема сообщества машинного обучения заключалась в разработке методов обучения знаний из данных. Чем сложнее и обширнее собираемые массивы данных, тем больше возможностей для извлечения интересующих сведений. В связи с чем, широкое применение получили технологии Data Mining. В

процессе перехода от аналогового к цифровому, большие наборы данных были сгенерированы, собраны и сохранены, открывая статистические шаблоны, тенденции и скрытую в данных информацию, способные помочь при построении прогностических шаблонов. Исследования показывают, что интеллектуальный анализ данных быстрее и гораздо более интуитивно понятен, чем традиционный анализ данных. История показывает, что мы являемся свидетелями революционных изменений в исследованиях. Сбор данных полезен для очистки данных, предварительной обработки данных и интеграции баз данных. Исследователи могут найти любые аналогичные данные из базы данных, которые могут привести к любым изменениям в исследовании. Идентификация любых совпадающих последовательностей и корреляция между любыми действиями могут быть известны. Визуализация данных и интеллектуальный анализ данных дают нам четкое представление о данных.

Data Mining- это междисциплинарная область, возникновение и развитие которой произошло на базе таких наук как прикладная математики и статистика, распознавание образов, ИИ, теория баз данных и др., посвященная научным методам, процессам и системам, направленным на извлечение знаний или сведений из данных, представленных в различных структурированных или неструктурированных формах [1] (Рис. 1).

Data Mining - это набор методологий, используемых при анализе данных из разных измерений и перспектив, поиска ранее неизвестных скрытых шаблонов, классификации и группировки данных и суммирования идентифицированных отношений. Сегодня Data Mining используется компаниями с сильной ориентацией на потребителя, такими как розничные, финансовые, коммуникационные и маркетинговые организации. Добыча данных позволяет этим компаниям определять отношения между «внутренними» факторами, такими как цена, позиционирование продукта или навыки персонала, и «внешними», такими как экономические показатели, конкуренция и демографические данные клиентов. Это позволяет

им определить, какое влияние эти отношения могут оказать на продажи, удовлетворенность клиентов и корпоративную прибыль. Наконец, эти технологии позволяют им «развернуть» сводную информацию для просмотра подробных транзакционных данных и поиска способов применения этих знаний для улучшения бизнеса [2].

Рис.1. Связь Data Mining с другими областями

С помощью интеллектуального анализа данных розничный торговец может использовать отчеты о покупках в торговых точках для отправки целевых рекламных акций на основе истории покупок отдельных лиц. При разработке демографических данных из комментариев или гарантийных карточек розничные торговцы могут разрабатывать продукты и рекламные акции для обращения к конкретным сегментам клиентов. В последние годы интеллектуальная обработка данных широко используется в областях науки и техники, таких как биоинформатика, генетика, медицина, образование и электроэнергетика. При изучении генетики человека, Data Mining помогает решить важную задачу понимания отношения индивидуальных вариаций последовательности ДНК человека и восприимчивости к болезням. Один из методов интеллектуального анализа данных, который используется для

выполнения этой задачи, известен как многофакторное понижение размерности.

В целом, Data Mining технологии имеют большой потенциал для улучшения системы здравоохранения. Они используют данные и аналитику для выявления лучших практик, способных улучшить уход и снизить затраты. Исследователи используют различные подходы к интеллектуальному анализу данных, такие как многомерные базы данных, машинное обучение, компьютерные вычисления, визуализация данных и статистика. Анализ данных может использоваться для прогнозирования объема пациентов в каждой категории. Разрабатываются процессы, которые гарантируют, что пациенты получат надлежащую помощь в нужном месте и в нужное время. Также данные технологии могут помочь страховщикам здравоохранения выявлять мошенничество [3]. Миллиарды долларов были потеряны в результате мошенничества, естественно не только в среде здравоохранения. Традиционные методы обнаружения мошенничества являются трудоемкими и сложными. Сбор данных помогает в предоставлении значимых шаблонов и превращении данных в информацию. Любая достоверная и полезная информация - это знания. Совершенная система обнаружения мошенничества должна защищать информацию всех пользователей. Контролируемый метод включает сбор образцов записей, которые классифицируются как мошеннические или немощные. Модель построена с использованием этих данных, и алгоритм делается для определения того, является ли запись мошеннической или нет.

В области электроэнергетики методы интеллектуального анализа широко используются для мониторинга состояния высоковольтного электрооборудования. Целью мониторинга состояния является получение ценной информации, например, о состоянии изоляции (или других важных параметрах, связанных с безопасностью). В образовательных исследованиях интеллектуальный анализ данных использовался для изучения факторов, побуждающих учащихся принимать участие в действиях, отрицательно

влияющих на их обучение, и понимания факторы, влияющие на удержание студентов в университетах. Аналогичным примером социального применения интеллектуального анализа данных является его использование в системах поиска экспертных знаний. В результате чего, для облегчения поиска экспертам, особенно в научной и технических областях, извлекаются, нормализуются и классифицируются дескрипторы человеческого опыта. Таким образом, интеллектуальный анализ данных может облегчить и институциональную память. В связи с чем, можно сделать вывод, что, охват областей, в которых можно успешно применить технологии Data Mining, очень широк [20].

Аналитические методы, используемые при интеллектуальном анализе данных, часто являются известными математическими алгоритмами и методами. Однако, новаторство заключается именно в применении этих методов для общих бизнес-задач, что стало возможным благодаря увеличению доступности данных, их недорого хранения и обработки. Кроме того, использование графических интерфейсов привело к тому, что инструменты стали более понятны и просты, в следствии чего бизнес-эксперты могут легко их использовать.

Data Mining - это пятиступенчатый процесс:

- Идентификация исходной информации
- Выборка данных, которые необходимо проанализировать
- Извлечение соответствующей информации из данных
- Идентификация значений ключа из выделенного набора данных
- Интерпретация и отчетность результатов

Рассмотрим выделенные ступени более подробно. Определение исходной информации-один из ключевых моментов. Необходимо проверить разные наборы данных и различные коллекции информации, а затем объединить их, чтобы создать реальную картину того, что именно вы хотите. Существует несколько стандартных наборов данных, к которым приходится

возвращать повторно. Различные наборы данных, как правило, раскрывают новые проблемы и задачи, в связи с чем, желательно иметь в виду множество проблем при рассмотрении методов обучения.

Выборка точек данных помогает идентифицировать данные, подлежащие анализу. Байесовские методы, основанные на создании массива данных, определяют вероятность того, насколько данные напрямую связаны с информацией, которую вы извлекли. В зависимости от сложности данных и информации, с которой вы работаете, извлечение этой информации и вычисление требуемой вероятности могут быть простыми или сложными, но их легко определить, вычислив частоту, или же основываясь на прошлом анализе аналогичных источников данных. Как только процесс извлечения и идентификации данных подходит к концу, следует превратить полученную информацию и структуру в результат.

Извлечение и идентификация ключевых значений помогают фильтровать полученную информацию. Методы обучения более сложны, и они полагаются на текущие и прошлые данные, чтобы создать структуру прошлого, действительного опыта, который в конечном итоге можно сравнить с новой информацией, а затем интерпретировать и извлечь. Эти шаги помогают как с извлечением, так и с идентификацией извлеченной информации. Далее необходимо интерпретировать результаты этой сортировки. Существует много разных подходов к этому вопросу, но все они основываются на предыдущих шагах, используя дополнительную валидацию и квалификацию информации для выбора необходимых данных ключа.

Интерпретация и отчетность являются заключительными этапами пятиэтапного процесса интеллектуального анализа данных, и включают в себя разрешение информации на более равные допустимые значения, такие как использование базовых численных отсчетов, сравнение прямых значений или групповое сравнение для выделения конкретных элементов. Данные, извлеченные на ранних этапах, можно объединить в конечный результат.

1.2. Архитектура, основные методы и задачи Data Mining - технологий

Data Mining описывается как процесс обнаружения или извлечения интересных знаний из больших объемов данных, хранящихся в нескольких источниках данных, таких как файловые системы, базы данных, хранилища данных ... и т. д. Эти знания приносят много преимуществ бизнес-стратегиям, научным, медицинским исследованиям, правительствам и отдельным лицам. Бизнес-данные собираются ежеминутно посредством бизнес-операций и хранятся в реляционных системах баз данных. Для обеспечения понимания бизнес-процессов были созданы системы хранилищ данных для предоставления аналитических отчетов, которые помогают бизнес-пользователям принимать решения.

Теперь данные хранятся в базах данных и / или хранилищах данных, поэтому встает вопрос о разработки системы интеллектуального анализа данных, которая отделяет или соединяется с базами данных и системами хранилищ данных. Этот вопрос приводит к четырем возможным возможностям архитектуры интеллектуального анализа данных:

- Отсутствие связи: в этой архитектуре система интеллектуального анализа данных не использует никаких функций базы данных или системы хранилища данных. Система интеллектуального анализа данных без соединения извлекает данные из определенного источника данных, такого как файловая система, обрабатывает данные с использованием основных алгоритмов интеллектуального анализа данных и сохраняет результаты в файловой системе. Архитектура интеллектуального анализа данных без соединения не имеет преимуществ для базы данных или хранилища данных, но эффективна при организации, хранении, доступе и извлечении данных. Архитектура без соединения считается плохой архитектурой для

системы интеллектуального анализа данных, однако используется для простых процессов интеллектуального анализа данных.

- Неплотная связь: в этой архитектуре система интеллектуального анализа данных использует базу данных или хранилище данных для извлечения данных. В архитектуре интеллектуального анализа данных с ослабленным соединением система интеллектуального анализа данных извлекает данные из базы данных или хранилища данных, обрабатывает данные с использованием алгоритмов интеллектуального анализа данных и сохраняет результат в этих системах. Эта архитектура в основном предназначена для системы интеллектуального анализа данных на основе памяти, которая не требует высокой масштабируемости и высокой производительности.
- Полуплотная связь: в полужесткой архитектуре интеллектуального анализа соединений, помимо связывания с базой данных или системой хранилища данных, система интеллектуального анализа данных использует несколько функций систем хранения данных или данных для выполнения некоторых задач интеллектуального анализа данных, включая сортировку, индексирование, агрегацию ... и т. д. В этой архитектуре некоторые промежуточные результаты могут быть сохранены в системе баз данных или хранилища данных для лучшей производительности.
- Плотная связь: в архитектуре интеллектуального анализа данных с жесткой связью база данных или хранилища данных рассматриваются как компонент поиска информации системы интеллектуального анализа данных с использованием интеграции. Все функции базы данных или хранилища данных используются для выполнения задач интеллектуального анализа данных. Эта архитектура обеспечивает масштабируемость системы, высокую производительность и интегрированную информацию.

В архитектуре интеллектуального анализа данных с жесткой связью имеется три уровня (рис. 2):

1. Уровень данных: как уже упоминалось выше, уровень данных может быть базой данных и / или системами хранения данных. Этот уровень является интерфейсом для всех источников данных. Результаты интеллектуального анализа данных хранятся в слое данных, поэтому его можно представить конечному пользователю в виде отчетов или другого вида визуализации.
2. Уровень приложения интеллектуального анализа данных используется для извлечения данных из базы данных. Для преобразования данных в желаемый формат может быть выполнена некоторая процедура преобразования. Затем данные обрабатываются с использованием различных алгоритмов интеллектуального анализа данных.
3. Front-end слой обеспечивает интуитивно понятный и удобный пользовательский интерфейс для взаимодействия конечного пользователя с системой интеллектуального анализа данных. Результат интеллектуального анализа данных, представленный в форме визуализации для пользователя в интерфейсном слое.

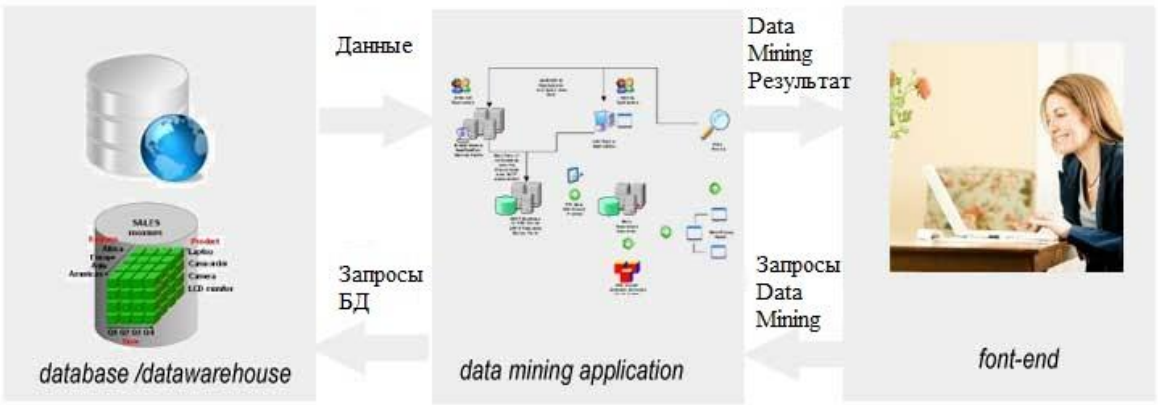


Рис.2. Архитектура Data Mining

Как можно заметить, добыча данных - это не простой процесс, и он основан на систематическом и математическом подходе к данным. Но он также зависит от гибкости и принятия данных, которые могут не обязательно вписываться в хорошо организованный и последовательный формат. Всевозможные методы моделирования, классификации, кластеризации и прогнозирования являются базисом при работе с Data Mining-технологиями.

Можно выделить два типа методов:

- Классические методы: статистика, линейная регрессия, байесовские сети, «соседи» и кластеризация
- Методы следующего поколения: деревья, нейронные сети

Рассмотрим выделенные методы более подробно. По сути, «статистика» или статистические методы не являются интеллектуальными данными. Данные методы использовались задолго до того, как термин «интеллектуальный анализ данных» был придуман для применения к бизнес-приложениям. Однако, статистические методы основаны на данных и используются для обнаружения шаблонов и построения прогностических моделей. Поэтому, при решении проблемы «интеллектуального анализа данных» встает вопрос относительно того, хотите ли вы атаковать его статистическими методами или другими методами интеллектуального анализа данных. По этой причине важно иметь некоторое представление о том, как работают статистические методы и как они могут применяться. Одним из мощных и широко используемых инструментов в статистике является регрессия. Простейшей формой регрессии является линейная регрессия, которая содержит только один предсказатель и предсказание. Простую модель линейной регрессии можно рассматривать как линию, которая минимизировала частоту ошибок между фактическим значением предсказания и предсказанием из модели. Графически это выглядело бы так, как показано на рис. 3.

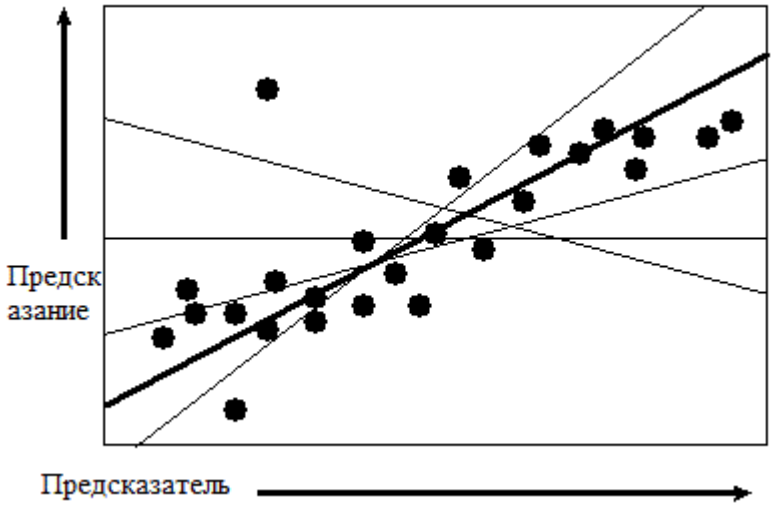


Рис.3. График Линейной Регрессии

Связь между ними может быть отображена в двумерном пространстве данными, построенными для значений предсказания вдоль оси Y и значений предсказателя вдоль оси X. Простейшая форма регрессии направлена на создание прогностической модели, которая представляет собой линию, отображающую между каждым значением предсказателя значение предсказания. Из множества возможных линий, которые можно было бы провести через данные, минимизирующей расстояние между линией и точками данных является выбранная для предсказательной модели линия. Прогностическая модель - это линия, показанная на рисунке. Линия примет заданное значение для предсказателя и отобразит его в заданное значение для предсказания. Фактическое уравнение будет выглядеть примерно так:

Prediction = a + b * Predictor.

Это просто уравнение для прямой Y = a + bX. Далее можно выделить байесовские сети, которые являются типом вероятностной графической модели, представляют собой набор переменных и их условные зависимости через ориентированный ациклический граф, и могут использоваться для построения моделей из данных и / или экспертных заключений. Они могут применяться для широкого круга задач, включая прогнозирование, обнаружение аномалий, диагностику, автоматическую проницательность,

рассуждение, прогнозирование временных рядов и принятие решений в условиях неопределенности.

Кластеризация и метод приближения ближайшего соседа относятся к самым старым методам, используемым для интеллектуального анализа данных.

Метод ближайшего соседа (рис. 4) - это метод предсказания, который очень похож на кластеризацию. Суть его в том, что для прогнозирования того, какое значение предсказания находится в одной записи, нужно искать записи со схожими значениями предсказателя в исторической базе данных и использовать значение предсказания из записи, находящейся «ближайшей» к неклассифицированной записи. Одним из усовершенствований, обычно применяемых к основному алгоритму ближайшего соседа, является нахождение ближайших «К» соседей. Нет особого правила, позволяющего определить преимущества конкретной техники над другой. Иногда эти решения принимаются относительно произвольно, исходя из наличия аналитиков интеллектуального анализа данных, которые наиболее опытные в одном методе чем в других. И даже выбор классических методов в большей степени зависит от наличия хороших инструментов и хороших аналитиков.

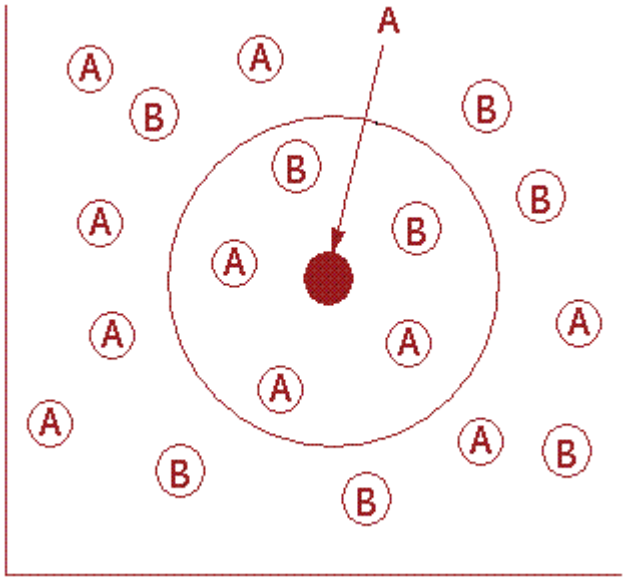


Рис.4. Визуализация метода ближайших К-соседей

Какие бы ни были выбраны методы, будь то классическое или следующее поколение, все выделенные здесь методы были доступны и опробованы на протяжении двух десятилетий. Поэтому даже техника следующего поколения является твердым гарантом успешной реализации. Дерево решений, относящееся к следующему поколению, - это предсказательная модель, которая, как следует из ее названия, может рассматриваться как дерево. В частности, каждая ветвь дерева является вопросом классификации, а листья дерева - это разделы набора данных с их классификацией. Например, если мы собираемся классифицировать клиентов, которые отказываются продлевать свои телефонные контракты в индустрии сотовых телефонов, дерево решений может выглядеть примерно так, как показано на рис. 5.

С точки зрения бизнес-решений деревья решений можно рассматривать как создание сегментации исходного набора данных (каждый сегмент будет одним из листьев дерева). Сегментация клиентов, продуктов и регионов продаж - это именно то, что менеджеры по маркетингу делают в течение многих лет.

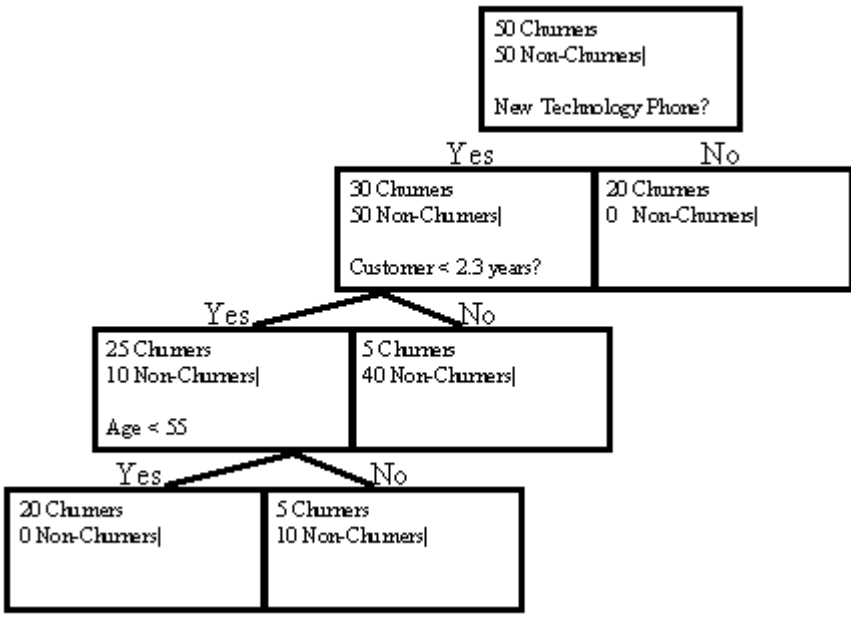


Рис.5. Дерево Решений

Технология дерева решений может использоваться для исследования набора данных и бизнес-задач. Это часто делается путем просмотра предсказателей и значений, которые выбираются для каждого разбиения дерева. Часто эти предсказатели обеспечивают полезную информацию или предлагают вопросы, на которые необходимо ответить.

Другим способом использования технологии дерева решений является предварительная обработка данных для других алгоритмов прогнозирования. Поскольку алгоритм довольно устойчив по отношению к множеству типов предсказателей (например, число, категориальное и т. д.), и может быть запущен относительно быстро, деревья решений используются на первом этапе запуска интеллектуального анализа данных для создания подмножеств, которые в свою очередь могут быть поданы в нейронные сети, или обработаны методами ближайшего соседа и нормальными статистическими процедурами.

Чтобы быть более точным с термином «нейронная сеть», лучше говорить именно об «искусственной нейронной сети». Истинными нейронными сетями являются биологические системы, которые обнаруживают закономерности, делают прогнозы и самообучаются. Искусственными являются компьютерные программы, реализующие сложные алгоритмы обнаружения и машинного обучения на компьютере для создания прогностических моделей из крупных исторических баз данных. Нейронные сети применяются в самых разных областях. Они используются во всех аспектах бизнеса, чтобы выявить мошенничество с использованием кредитных карт и прогнозирование кредитного риска для увеличения количества целевых рассылок. Они также имеют долгую историю применения в других областях, от автоматизированного вождения беспилотного транспортного средства со скоростью 30 миль в час на асфальтированных дорогах до биологических симуляций, таких как изучение правильного произношения английских слов из письменного текста.

Нейронная сеть состоит из двух основных структур:

- Узел, соответствующий нейрону в мозгу человека.
- Связь, соответствующая соединениям между нейронами (аксонами, дендритами и синапсами) в мозге человека.

На рис. 6. показана простая нейронная сеть.

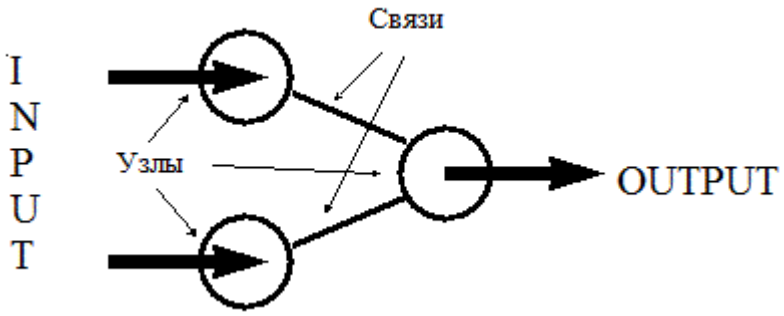


Рис. 6. Простая нейронная сеть

Круглые круги представляют узлы, а соединительные линии представляют собой связи. Нейронная сеть функционирует, принимая значения прогнозирования слева и выполняя вычисления этих значений для создания новых значений в узле в крайнем правом углу. Значение на этом узле представляет собой предсказание из модели нейронной сети. В этом случае сеть принимает значения для предикторов по указанным параметрам и прогнозирует возможное решение.

Очевидно, что одной из самых сложных задач при принятии решения о внедрении системы интеллектуального анализа данных является определение того, какой метод использовать и когда. Некоторые критерии, которые важны при определении используемого метода, определяются методом проб и ошибок. Существуют определенные различия в типах проблем, которые наиболее благоприятны для каждого метода, но реальность настоящих данных и динамический способ формирования рынков, клиентов и, следовательно, данных, которые представляют их, означает, что данные постоянно меняются. Эта динамика означает, что уже нет смысла строить «идеальную» модель по историческим данным, поскольку все, что было

известно в прошлом, не может адекватно предсказать будущее, потому что будущее настолько отличается от того, что было раньше.

Также, можно выделить следующие стадии Data Mining:

1. Свободный поиск- выявление скрытых закономерностей
2. Валидация-проверка достоверности выявленных закономерностей
3. Прогностическое моделирование- предсказание неизвестных значений, на основе полученных на первой стадии закономерностей
4. Анализ исключений-выявление аномалий найденных закономерностей

Решаемые при помощи технологии Data Mining задачи можно подразделить на описательные и предсказательные. К наиболее распространенным задачам можно причислить следующие:

1. Кластеризация- это процесс обнаружения групп и кластеров в данных таким образом, что степень ассоциации между двумя объектами является самой высокой, если они принадлежат к одной и той же группе и самым низким в противном случае. Фактически, кластер представляет собой набор схожих объектов. Это означает, что объекты похожи друг на друга в одной и той же группе, и, соответственно отличаются друг от друга, неодинаковы или не связаны с объектами в других группах или в других кластерах.
2. Классификация используется для извлечения важной и актуальной информации о данных и метаданных. Классификация похожа на кластеризацию тем, что она также сегментирует записи данных в разные сегменты, называемые классами. Но в отличие от кластеризации, здесь бизнес-аналитики будут знать какие именно классы должны получиться на выходе.
3. Ассоциация помогает смоделировать зависимости между различными переменными в больших базах данных. А именно, раскрыть скрытые шаблоны в данных, которые могут использоваться для идентификации переменных в данных и совпадения различных переменных, которые

очень часто появляются в наборе данных. Правила ассоциации полезны для изучения и прогнозирования поведения клиентов.

4. Анализ и обнаружение отклонений относится к наблюдению за элементами данных в наборе данных, которые не соответствуют ожидаемому шаблону или ожидаемому поведению. Отклонения также известны как выбросы, новинки, шумы, аномалии и исключения. Часто они предоставляют важную и действенную информацию. Отклонение - это элемент, который значительно отличается от общего среднего в наборе или комбинации данных.
5. Предсказание, как следует из названия, является одним из методов интеллектуального анализа данных, который обнаруживает взаимосвязь между только независимыми переменными или же между зависимыми и независимыми переменными.
6. Анализ последовательностей направлен на обнаружение или идентификацию аналогичных шаблонов, регулярных событий или тенденций в данных транзакциях за отчетный период.
7. Оценивание включает в себя предсказание непрерывных значений признака
8. Анализ связей находит зависимости в массиве данных
9. Визуализация представляет собой процесс графического представления анализируемых данных

Основными этапами решения этих задач являются следующие:

1. постановка задачи анализа;
2. сбор данных;
3. подготовка данных (фильтрация, дополнение, кодирование);
4. выбор модели (алгоритма анализа данных);
5. подбор параметров модели и алгоритма обучения;
6. обучение модели (автоматический поиск остальных параметров модели);

7. анализ качества обучения, если данный пункт не приносит желаемого результата, то возвращение к пункту 5 или 4;
8. анализ выявленных закономерностей, если неудовлетворительный, то переход к первому пункту.

1.3. Инструменты Data Mining- Технологий

Для быстрого анализа данных с использованием любой технологии интеллектуального анализа данных важно иметь представление о различных инструментах. Все упомянутые ниже инструменты имеют свою специфику с точки зрения реализации, и каждый из них имеет свои достоинства. Все это сводится к требованию задачи. Самое главное - знать, что существуют инструменты, которые могут значительно повысить эффективность ученого-исследователя данных или студента, работающего над каким-то проектом, предоставляющие возможность сосредоточиться на более важных действиях, а не получении полезных сведений и создание прогнозов. Также данные инструменты устраняют неудобства от реализации любого стандартного алгоритма с нуля, но в то же время дают право изменять код инструмента (с открытым исходным кодом) в соответствии с требованиями. Рассмотрим некоторые из этих инструментов.

Rapid Miner- это достаточно популярный инструмент, поскольку представляет из себя готовое программное обеспечение с открытым исходным кодом, не требующее кодирования, и предоставляющее расширенную аналитику. Написанный на Java, он включает в себя многогранные функции интеллектуального анализа данных, такие как предварительная обработка данных, визуализация, интеллектуальный анализ и может быть легко интегрирована с WEKA и R-инструментом, чтобы напрямую давать модели из сценариев, написанных в первых двух. Помимо стандартных функций интеллектуального анализа данных, таких как очистка

данных, фильтрация, кластеризация и т. д., программное обеспечение также имеет встроенные шаблоны, повторяемые рабочие потоки, профессиональную среду визуализации и бесшовную интеграцию с такими языками, как Python и R, в рабочие потоки, которые помогают быстрому макетированию.

Weka - это набор алгоритмов машинного обучения для задач интеллектуального анализа данных. Алгоритмы могут быть применены непосредственно к набору данных или вызваны из вашего собственного кода Java. Weka содержит инструменты для предварительной обработки данных, классификации, регрессии, кластеризации, правил ассоциации и визуализации. Он также хорошо подходит для разработки новых схем машинного обучения [6]. Пользователи Python зачастую используют инструмент Orange. Это библиотека Python, которая управляет сценариями Python с его богатой компиляцией алгоритмов интеллектуального анализа и машинного обучения для предварительной обработки данных, классификации, моделирования, регрессии, кластеризации и других различных функций. В Orange также есть среда визуального программирования, а ее рабочий стол состоит из инструментов для импорта данных и перетаскивания виджетами и ссылками для подключения разных виджетов для завершения рабочего процесса.

R – это свободная программная среда для статистических вычислений и графики, написанная на C ++. R Studio - это IDE, специально разработанная для языка R. Это один из ведущих инструментов, используемых для выполнения задач интеллектуального анализа данных, огромная поддержка сообщества, а также упаковка сотен библиотек, созданных специально для интеллектуального анализа данных.

Knime - это мощный инструмент с графическим интерфейсом, который показывает сеть узлов данных. Популярный среди аналитиков финансовых данных, он имеет модульную сборку узлов данных, используя возможности машинного обучения и концепции интеллектуального анализа данных для

создания отчетов бизнес-аналитики. Rattle, расширенный до «Аналитического инструмента для быстрого изучения», был разработан с использованием статистического языка программирования R. Программное обеспечение может работать в Linux, Mac OS и Windows, имеет функции статистики, кластеризации, моделирования и визуализации с вычислительной мощностью R. В настоящее время Rattle используется в бизнесе, а также в коммерческих и учебных целях в австралийских и американских университетах.

TANAGRA - бесплатное программное обеспечение для интеллектуального анализа данных с открытым исходным кодом для академических и исследовательских целей. Он предлагает несколько методов интеллектуального анализа данных из анализа исследовательских данных, статистического обучения, машинного обучения и базы данных. TANAGRA является более мощным инструментом, он содержит определенное контролируемое обучение, но также и другие парадигмы, такие как кластеризация, факторный анализ, параметрическая и непараметрическая статистика, правило ассоциации, выбор функций и алгоритмы построения. Основная цель проекта Tanagra - упростить исследователям и студентам разработку программного обеспечения в этой области, а также позволить проанализировать реальные или синтетические данные.

XLMiner является единственной всеобъемлющей надстройкой интеллектуального анализа данных для Excel с нейронными сетями, деревьями классификации и регрессии, логистической регрессией, линейной регрессией, классификатором Байеса, ближайшими соседями K, дискриминантным анализом, правилами ассоциации, кластеризацией, основными компонентами и т. д. XLMiner предоставляет все необходимое для отбора данных из многих источников - PowerPivot, баз данных Microsoft / IBM / Oracle или электронных таблиц, исследует и визуализирует ваши данные с помощью нескольких связанных диаграмм; очищает ваши данные, подгоняет модели интеллектуального анализа данных и оценивает

прогностическую способность ваших моделей. Недостатком XL Miner является то, что это платная добавка для excel. Однако, программное обеспечение обладает большими возможностями, и его интеграция в excel облегчает жизнь.

II ГЛАВА. КЛАСТЕРИЗАЦИЯ - ОДНА ИЗ ЗАДАЧ DATA MİNİNG

2.1. Основополагающие моменты кластерного анализа

Кластерный анализ - это исследовательский анализ, целью которого является идентификация структуры данных. Кластерный анализ также называется таксономическим и сегментационным анализом. Если говорить более конкретно, он пытается идентифицировать гомогенные группы объектов в том случае, если группировка ранее не известна. То есть, в отличии от классификации, при использовании кластерного анализа имеется возможность решить задачу без каких-либо предварительных знаний о данных. Поскольку кластеризация является исследовательским анализом, то применяя ее, не делается различий между зависимыми и независимыми переменными. Зачастую, кластерный анализ используется в сочетании с другими анализами, например, с дискриминантным. Различные методы кластерного анализа могут обрабатывать двоичные, номинальные, порядковые и масштабные (интервальные или относительные) данные. Разработка методологии кластеризации была по-настоящему междисциплинарной задачей. Таксономисты, социологи, психологи, биологи, статистики, математики, инженеры, ученые-компьютерщики, медицинские исследователи и другие, собирающие и обрабатывающие реальные данные, внесли свой вклад в методологию кластеризации.

Согласно JSTOR (2009), термин *«кластеризация данных»* впервые появился в названии статьи 1954 года, посвященной антропологическим данным [4]. Кластеризация данных также известна как Q-анализ, типология, clumping и таксономия, в зависимости от области, в которой она применяется. Существует несколько книг, посвященных кластеризации данных. Книги посвященные классическому кластерному анализу (Сокал и Сниф, (1963), Андерберг (1973), Хартиган (1975), Джайн и Дюбс (1988) и т.д.) Алгоритмы кластеризации также были широко изучены в литературе,

посвящённой области интеллектуального анализа данных (см. книги Хан и Камбер (2000) и Тан и др. (2005)) и машинному обучению (Бишоп, 2006).

Алгоритм решения задачи кластеризации выглядит следующим образом:

1. Построение гипотезы: самый важный шаг всего алгоритма. Необходимо определить все возможные выборки объектов для кластеризации
2. Выявление исходного списка переменных: определить множество переменных, по которым будет производиться дальнейшая оценка.
3. Визуализация данных: очень важно знать распределение по выбранной переменной перед началом любого анализа.
4. Очистка данных от шумов: кластерный анализ очень чувствителен к выбросам. Очень важно очищать данные по всем переменным, принятым во внимание.
5. Установка меры сходства анализируемых объектов. (о ней будет сказано ниже)
6. Применение одного из выбранных при помощи предыдущих шагов метода кластерного анализа
7. Выявление конвергенции (схожести) кластеров: хороший кластерный анализ имеет все кластеры с населенностью от 5 до 30% общей базы.
8. Профилирование кластеров: после проверки конвергенции кластерного анализа необходимо определить поведение каждого кластера
9. Предоставление результатов анализа

Вообще говоря, кластеризацию можно разделить на две подгруппы:

- Жесткая кластеризация: при жесткой кластеризации каждая точка данных либо принадлежит кластеру полностью, либо нет.
- Мягкая кластеризация: в мягкой кластеризации вместо того, чтобы помещать каждую точку данных в отдельный кластер, назначается вероятность того, что данные указывают на эти кластеры.

Поскольку задача кластеризации субъективна, средства, которые могут быть использованы для достижения этой цели, изобилуют. Каждая

методология придерживается другого набора правил для определения «подобия» между точками данных. На самом деле известно более 100 алгоритмов кластеризации [7]. Но немногие из алгоритмов используются повсеместно, давайте рассмотрим их подробно:

- Модели подключения. Как следует из названия, эти модели основаны на представлении о том, что данные, расположенные ближе в пространстве данных, имеют большее сходство друг с другом, чем точки данных, расположенные дальше. Эти модели могут следовать двум подходам. В первом подходе они начинаются с классификации всех точек данных в отдельные кластеры и затем агрегирования их по мере уменьшения расстояния. Во втором подходе все точки данных классифицируются как один кластер, а затем разбиваются по мере увеличения расстояния. Кроме того, выбор функции расстояния субъективен. Эти модели очень легко интерпретировать, но они не имеют масштабируемости для обработки больших наборов данных. Примерами этих моделей являются иерархический алгоритм кластеризации и его варианты.
- Центроидные модели. Это итеративные алгоритмы кластеризации, в которых понятие сходства определяется близостью точки данных к центроиду кластеров. Алгоритм кластеризации K-Means является популярным алгоритмом, который попадает в эту категорию. В этих моделях, данные за пределами кластеров, требуемых в конце, должны быть упомянуты заранее. В связи с чем необходимо иметь предварительное знание набора данных. Эти модели работают итеративно, чтобы найти локальные оптимумы.
- Модели распределения. Эти модели кластеризации основаны на представлении о том, насколько вероятным является то, что все точки данных в кластере принадлежат к одному и тому же распределению (например: Normal, Gaussian). Эти модели часто страдают от переобучения. Популярным примером этих моделей является алгоритм

максимизации ожиданий, который использует многомерные нормальные распределения.

- Модели плотности. Эти модели ищут пространство данных для областей с различной плотностью точек данных в пространстве данных. Он изолирует различные области плотности и назначает точки данных в этих регионах в одном кластере. Популярными примерами моделей плотности являются DBSCAN и OPTICS.

2.2. Определение меры сходства в кластерном анализе

Правильно подобранный метод кластеризации обеспечивает высокое качество кластеров с высоким внутриклассовым и низким межклассовым сходствами. В свою очередь, качество результата кластеризации зависит как от меры сходства, используемого методом, так и его реализации [9]. Показатель несходства/сходства выражается функцией расстояния $\boldsymbol{d(x, y)}$ – такой, что

1. $d(x, y) > 0$
2. $d(x, y) = 0, \text{если } x = y$
3. $d(x, y) = d(y, x)$
4. $d(x, y) < d(x, z) + d(z, y)$

Существует отдельная функция «качества», которая измеряет «правильность» кластера. Определения функций расстояния обычно очень отличаются для интервально-масштабированных, логических, категориальных и порядковых переменных.

Стоит отметить два основных класса измерения расстояния:

1. Евклидово, основанное на местоположении точек в пространстве
2. Неевклидово, основанное на свойстве точек, но не их местоположении в пространстве

Рассмотрим основные Евклидовы метрики более подробно:

- Евклидово расстояние само по себе является геометрическим расстоянием между точками x и y в многомерном n-пространстве и вычисляется следующим образом:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

- Квадрат евклидова расстояния придает больший вес отдаленным точкам

$$d(x,y) = \sum_{i=1}^{n}(x_i - y_i)^2$$

- Манхэттенское расстояние (расстояние городских кварталов)-сумма разности по координатам в различных измерениях

$$d(x,y) = \sum_{i=1}^{n}|x_i - y_i|$$

- Расстояние Чебышева

$$d(x,y) = \max(|x_i - y_i|)$$

- Степенное расстояние, применяется при необходимости изменения веса размерности, т.е. в случае большого различия соответствующих объектов,

$$d(x,y) = \sqrt[l]{\sum_{i=1}^{n}(x_i - y_i)^k}$$

Где l и k-задаваемые пользователем параметры измерения, причем если оба параметра равны двум, то данное расстояние эквивалентно Евклидовому.

Из неевклидовых можно выделить следующие:

- Мера Жаккара- бинарная мера сходства, причем, первый известный коэффициент сходства

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

- Коэффициент Отиаи

$$K = \frac{|A \cap B|}{\sqrt{|A| \cdot |B|}}$$

Где $|A|$ и $|B|$-мощности соответствующих множеств

Существуют другие методы несходства, широко используемые для анализа данных, такие как расстояния, основанные на корреляции. Расстояние на основе корреляции определяется путем вычитания коэффициента корреляции из 1. Можно использовать различные типы методов корреляции, такие как:

- Корреляционное расстояние Пирсона

$$d_{cor}(x,y) = 1 - \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- Коэффициент корреляции Эйзен-косинуса является особым случаем корреляции Пирсона, где $\bar{x}$ и $\bar{y}$ оба заменены на ноль:

$$d_{eisen}(x,y) = 1 - \frac{\sum_{i=1}^{n}|x_i y_i|}{\sqrt{\sum_{i=1}^{n} {x_i}^2 \sum_{i=1}^{n} {y_i}^2}}$$

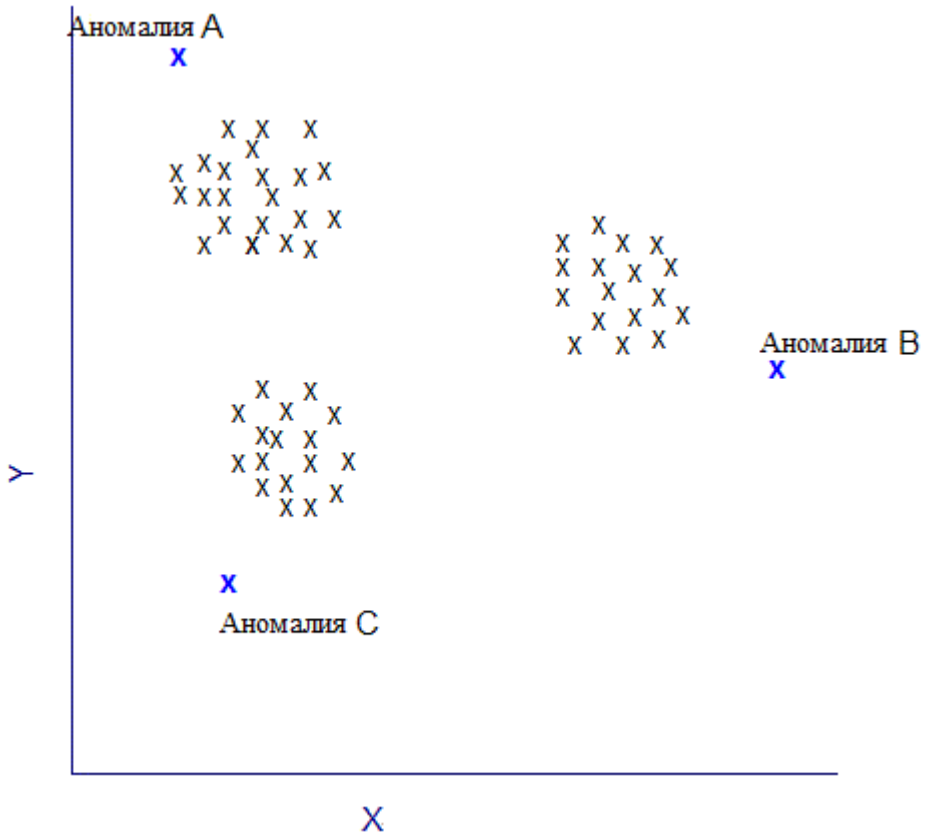
- Корреляционное расстояние Спирмана вычисляет корреляцию между рангами переменных x и y:

$$d_{spear}(x,y) = 1 - \frac{\sum_{i=1}^{n}(x_i' - \bar{x'})(y_i' - \bar{y'})}{\sqrt{\sum_{i=1}^{n}(x_i' - \bar{x'})^2 \sum_{i=1}^{n}(y_i' - \bar{y'})^2}}$$

где $x_i' = rank(x_i)$ и $y_i' = rank(y_i)$

Выбор меры расстояния очень важен, так как он оказывает сильное влияние на результаты кластеризации. Для большинства распространенных программ кластеризации значение расстояния по умолчанию - это евклидово расстояние. В зависимости от типа данных и вопросов исследователя могут быть предпочтительными другие меры несходства. Например, расстояние, основанное на корреляции, часто используется в анализе данных экспрессии

генов. Расстояние на основе корреляции считает два объекта одинаковыми, если их особенности сильно коррелированы, хотя наблюдаемые значения могут быть далеко друг от друга в терминах евклидова расстояния. Расстояние между двумя объектами равно 0, когда они отлично коррелированы. В свою очередь, качество метода кластеризации измеряется его способностью обнаруживать некоторые или все скрытые шаблоны. Кластеризация также может использоваться для обнаружения аномалий (рис. 7). Как только данные будут сегментированы в кластеры, можно обнаружить, что некоторые случаи плохо вписываются в какие-либо кластеры. Эти случаи и являются аномалиями или выбросами. Корреляция Пирсона довольно



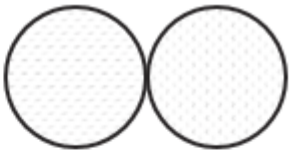
чувствительна к выбросам. Данную проблему можно устранить, используя корреляцию Спирмена вместо корреляции Пирсона.

Рис. 7. Аномалии

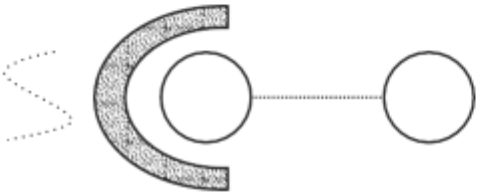
2.3. Типы и методы кластеризации

Кластеризация направлена на поиск полезных групп объектов (кластеров), где полезность определяется целью анализа данных. Неудивительно, что существует несколько различных понятий кластера, которые на практике полезны. Чтобы визуально проиллюстрировать различия между этими типами кластеров, используем двумерные точки, как

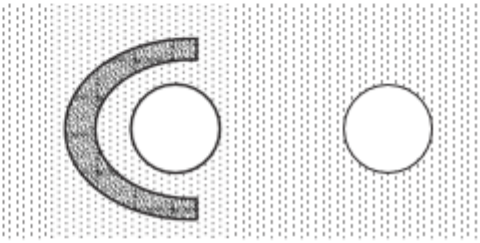
(a) Хорошо разделенные кластеры. Каждая точка ближе ко всем точкам в кластере, чем к любой точке другого кластера.



(a) Центровые кластеры. Каждая точка ближе к центру своего кластера, чем к центру любого другого кластера.



(b) Кластеры, основанные на скоплении. Каждая точка ближе к по меньшей мере одной точке своего кластера, чем к любой точке другого кластера.



(c) Кластеры на основе плотности. Кластеры представляют собой области с высокой плотностью, разделенные областями с низкой плотность

(d) Концептуальные кластеры.Точки в кластере имеют общее общее свойство, которое происходит от всего набора точек. (Точки в пересечении кругов принадлежат обоим.)

объекты данных (рис.8).

Рис.8. Различные типы кластеров, показанные множествами двумерных точек

Однако стоит подчеркнуть, что типы кластеров, описанные здесь, одинаково справедливы для других видов данных. Рассмотрим более подробно эти типы [10, 11].

Хорошо Разделенный Кластер представляет собой набор объектов, в которых каждый объект находится ближе (более схож) к каждому другому объекту в кластере, чем к любому объекту, не входящему в кластер. Иногда пороговое значение используется для указания того, что все объекты в кластере должны быть достаточно близкими (или похожими) друг к другу. Это идеалистическое определение кластера удовлетворяется только тогда, когда данные содержат естественные кластеры, которые довольно далеки друг от друга. На рис. 8(а) приведен пример хорошо разделенных кластеров, состоящий из двух групп точек в двумерном пространстве. Расстояние между любыми двумя точками в разных группах больше, чем расстояние между любыми двумя точками внутри группы. Хорошо разделенные кластеры не обязательно должны быть шаровидными, т.е. могут иметь любую форму.

Кластер На Основе Прототипов представляет собой набор объектов, в которых каждый объект ближе (более похож) на прототип, определяющий кластер, чем на прототип любого другого кластера. Для данных с непрерывными атрибутами прототипом кластера часто является центроид, т.е. Среднее (среднее) всех точек в кластере. Когда центроид не имеет смысла, например, когда данные имеют категориальные атрибуты, прототип часто является средним, т. е. наиболее представительной точкой кластера. Для многих типов данных прототип можно рассматривать как самую центральную точку, и в таких случаях мы обычно ссылаемся на кластеры на основе прототипов как на кластеры на основе центра. Неудивительно, что

такие кластеры имеют тенденцию быть глобулярными. На рис. 8(б) показан пример центровых кластеров.

Кластер На основе Графов является таковым, если данные представлены в виде графов, где узлы являются объектами, а ссылки представляют собой соединения между объектами. Такой кластер может быть определен как подключенный компонент, то есть как группа объектов, которые связаны друг с другом, но которые не имеют связи с объектами вне группы. Важным примером кластеров на основе графов являются кластеры на основе соприкосновения, где два объекта связаны только в том случае, если они находятся на определенном расстоянии друг от друга. Это означает, что каждый объект в кластере на основе соприкосновения ближе к какому-либо другому объекту в кластере, чем к любой точке в отдельном кластере. На рис. 8(c) показан пример таких кластеров для двумерных точек. Это определение кластера полезно, когда кластеры нерегулярны или переплетаются, но могут иметь проблемы с присутствием шума, поскольку, как показано двумя сферическими кластерами на рис. 8(c), небольшой мост из точек может объединять два отдельных кластера. Возможны и другие типы кластеров на основе графов. Один такой подход определяет кластер как клику; то есть набор узлов в графе, которые полностью связаны друг с другом. В частности, если мы добавляем связь между объектами в порядке их расстояния друг от друга, кластер формируется, когда набор объектов формирует клику. Подобно кластерам на основе прототипов, такие кластеры имеют тенденцию быть глобулярными.

Кластер на основе плотности - это плотная область объектов, окруженная областью с низкой плотностью. На рисунке 8(d) показаны некоторые кластеры на основе плотности для данных, созданных путем добавления шума к данным на рис. 8 (c). Два круглых кластера не объединены, как на рис. 8 (c), потому что мост между ними исчезает в шуме. Аналогично, кривая, которая представлена на рис. 8(c), также исчезает в шуме и не образует кластер на рис. 8(d). Определение кластеров на основе

плотности часто используется, когда кластеры нерегулярны или переплетаются, а также при наличии шума и выбросов. Напротив, определение кластеров на основе соприкосновения не будет хорошо работать для данных на рис. 8(d)., поскольку шум будет иметь тенденцию образовывать мосты между кластерами.

Концептуальные кластеры берут за основу совместное свойство. В более общем плане мы можем определить кластер как набор объектов, которые имеют некоторое свойство. Это определение охватывает все предыдущие определения кластера; например, объекты в центральном кластере разделяют свойство, а именно, все они наиболее близки к одному и тому же центру или среднему. Однако подход с общими свойствами также включает новые типы кластеров. Рассмотрим кластеры, показанные на рисунке 8(e). Треугольная область (кластер) смежна с прямоугольной, и есть два переплетенных кружка (кластеры). В обоих случаях алгоритму кластеризации потребуется очень специфическая концепция кластера для успешного обнаружения этих кластеров. Процесс определения таких кластеров называется концептуальной кластеризацией. Кластеризация - это задача, для которой предложено множество алгоритмов. Ни один метод кластеризации не применяется повсеместно, и разные методы подходят для различных целей кластеризации. Поэтому для решения проблемы кластеризации требуется применение подходящего метода для данной задачи.

Существует несколько различных подходов к вычислению кластеров. Ниже перечислены ортогональные аспекты, с которыми можно сравнить методы кластеризации:

Критерии разделения. В некоторых методах все объекты разделены так, что между кластерами не существует иерархии. То есть все кластеры на одном уровне концептуально. Такой метод полезен, например, для разделения клиентов на группы, чтобы каждая группа имела своего собственного менеджера. В качестве альтернативы другие методы разделяют

объекты данных иерархически, где кластеры могут быть сформированы на разных семантических уровнях. Например, в области интеллектуального анализа текста мы можем захотеть организовать состав документов по нескольким общим темам, таким как «политика» и «спорт», каждый из которых может иметь подтемы, например «футбол», «баскетбол», бейсбол "и" хоккей "могут существовать как подтемы« спорта ». Последние четыре подтемы находятся на более низком уровне в иерархии, чем« спорт ».

Разделение кластеров. Некоторые методы разделяют объекты данных во взаимоисключающие кластеры. При кластеризации клиентов в группы так, чтобы о каждой группе заботился только один менеджер, каждый клиент может принадлежать только одной группе. В некоторых других ситуациях кластеры могут быть не эксклюзивными, то есть объект данных может принадлежать более чем одному кластеру. Например, при кластеризации документов в темы, документ может быть связан с несколькими темами. Таким образом, темы как кластеры могут быть не эксклюзивными.

Мера сходства. Некоторые методы определяют сходство между двумя объектами по расстоянию между ними. Такое расстояние может быть определено в евклидовом пространстве, дорожной сети, векторном пространстве или в любом другом пространстве. В других методах сходство может быть определено связностью, основанной на плотности или смежности, и может не полагаться на абсолютное расстояние между двумя объектами. Меры сходства играют основополагающую роль в разработке методов кластеризации. В то время как методы на основе расстояния могут использовать преимущества методов оптимизации, методы, основанные на плотности и непрерывности, могут часто находить кластеры произвольной формы.

Кластерное пространство: многие методы кластеризации ищут кластеры во всем заданном пространстве данных. Эти методы полезны для наборов данных с низкой размерностью. Однако при использовании высокоразмерных данных может быть много нерелевантных атрибутов,

которые могут сделать измерения подобия ненадежными. Следовательно, кластеры, найденные в полном объеме, часто бессмысленны. Часто лучше искать кластеры в разных подпространствах одного и того же набора данных. Кластеризация подпространств обнаруживает кластеры и подпространства (часто с низкой размерностью), которые демонстрируют сходство объектов.

Рассмотрим некоторые критерии для классификации методов кластеризации. Можно выделить следующие алгоритмы, подразделенные по способу обработки данных:

1. Иеархические методы
2. Неиеархические (плоские) методы
3. Методы на основе плотности
4. Спектральные методы
5. Методы на основе сетки
6. Методы на основе разбиения

По способу анализа данных:

1. Четкие
2. Нечеткие

По количеству применений алгоритмов:

1. одноэтапные
2. многоэтапные

По возможности расширения объема обрабатываемых данных:

1. масштабируемые
2. немасштабируемые

Рассмотрим подробнее первую классификацию. Иерархический метод группирует объекты данных в иерархию кластеров. Иерархию можно сформировать сверху вниз или снизу-вверх, таким образом получается система вложенных разбиений выборки на непересекающиеся кластеры. Иерархию кластеров обычно рассматривают как дерево, где наименьшие кластеры объединяются, чтобы создать следующий самый высокий уровень кластеров, а те, что на этом уровне, сливаются вместе, чтобы создать

следующий самый высокий уровень кластеров. На рисунке (идет рисунок) показано, как несколько кластеров могут образовывать иерархию. Когда создается иерархия кластеров, подобная этой, пользователь может определить, какое правильное количество кластеров адекватно суммирует данные, сохраняя при этом полезную информацию (с другой стороны, один кластер, содержащий все записи, представляет собой большое обобщение, но не содержит достаточно конкретной информации, которая будет полезна).
Иерархические алгоритмы полагаются на функцию расстояния для измерения сходства между кластерами. В свою очередь их можно подразделить на агломеративные и дивизионные алгоритмы. Методы агломеративной кластеризации начинаются с большого количества кластеров, так как есть записи, в которых каждый кластер содержит только одну запись. Кластеры, которые являются ближайшими друг к другу, объединяются вместе, образуя следующий по величине кластер. Это слияние продолжается до тех пор, пока иерархия кластеров не будет построена только с одним кластером, содержащим все записи в верхней части иерархии.

Построение иерархической агломерационной классификации может быть достигнуто с помощью следующего общего алгоритма.

1. Найти 2 ближайших объекта и объединить их в кластер
2. Найти и объединить следующие две ближайшие точки, где точка представляет собой либо отдельный объект, либо кластер объектов.
3. Если осталось несколько кластеров, вернуться к шагу 2.

Дивизимные алгоритмы-это методы делительной кластеризации, использующие противоположный методам агломерации подход. Эти методы начинают анализ со всех записей в одном кластере, а затем пытаются разбить этот кластер на более мелкие части, а тот, в свою очередь, разбить на еще меньшие части.Из этих двух типов агломерационные методы наиболее часто используются для кластеризации и имеют больше ориентированных на них алгоритмов.Примерами этих алгоритмов являются алгоритмы LEGCLUST, BRICH, (Баланс итеративного сокращения и кластеризации с использованием

иерархии), CURE (кластер, с использованием представителей, и Chemeleon. Преимущества иерархической кластеризации включают в себя встроенную гибкость в отношении уровня детализации и простоту обработки любых форм сходства или расстояния. Следовательно, его применимость к любым типам атрибутов и его логической структуре, делает его легким к чтению и интерпретации. Недостатки иерархической кластеризации связаны с неопределенностью критериев завершения, и тем фактом, что большинство иерархических алгоритмов не пересматривают некогда построенные (промежуточные) кластеры с целью их улучшения. А также, они относительно нестабильны и ненадежны, т. е. первая комбинация или разделение объектов, которые могут быть основаны на небольшой разнице в критерии, будет ограничивать остальную часть анализа.

Неиерархические методы, или плоские, напротив, строят одно разбиение объектов на кластеры. И в целом быстрее создаются из исторической базы данных, но требуют, чтобы пользователь принял какое-то решение о количестве желаемых кластеров или минимальной «близости», требуемом для двух записей, в пределах одного кластера. Неиерархический алгоритм зачастую цикличен, а именно, начиная с какой-либо произвольной или даже случайной кластеризации, затем итеративно улучшает кластеризацию путем перетасовки некоторых записей. Их подразделение идет исходя из применяемого метода-итеративный, плотностный, модельный, концептуальный, сетевой. Существуют два основных неиерархических метода кластеризации. Оба они очень быстро вычисляются в базе данных, но имеют некоторые недостатки. Первый - это методы с одним проходом. Они получают свое имя из того факта, что информация из база данных должна быть получена только один раз для создания кластеров (т.е. каждая запись считывается из базы данных единожды). Другой класс методов называется перераспределением методов. Они получили свое имя от движения или «перераспределения» записей из одного кластера в другой, чтобы создать лучшие кластеры. Методы перераспределения используют

несколько проходов через базу данных, но относительно быстры по сравнению с иерархическими. Одним из наиболее популярных неиеархических алгоритмов является алгоритм К-средних [19].

К-средних (рис. 9) - это итеративный алгоритм кластеризации, целью которого является найти локальные максимумы на каждой итерации. Этот алгоритм состоит из следующих 5 шагов:

1. Указать желаемое количество кластеров К
2. Случайным образом привязать точки к какому-либо кластеру
3. Вычислить центроиды кластеров
4. Повторно назначить каждую точку ближайшему центроиду
5. Повторно вычислить центроиды кластеров: затем, идет перерасчет центроидов для ранее вычисленных кластеров.

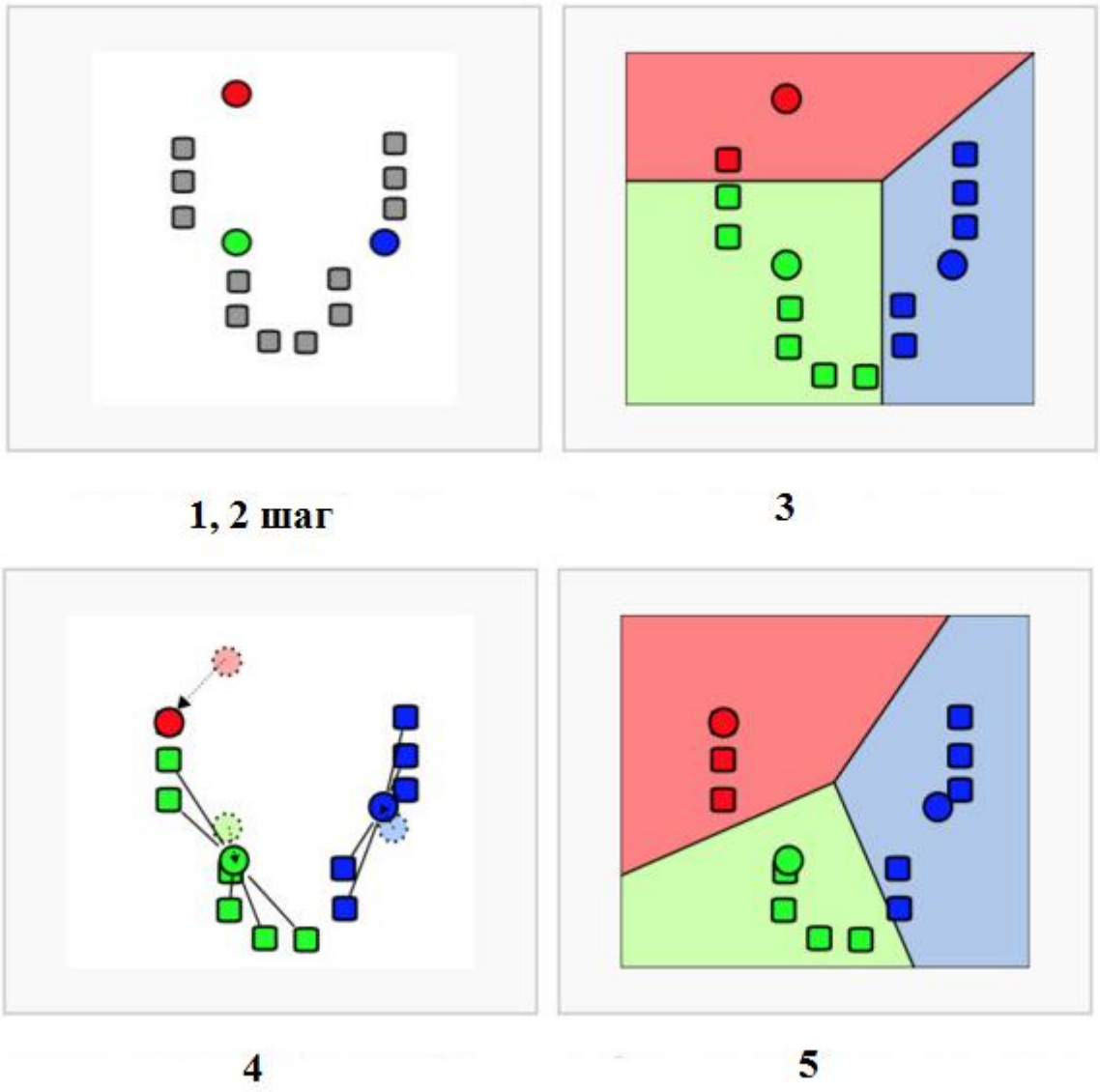


Рис. 9. Визуализация К-средних

Иерархическая кластеризация (рис. 10) имеет преимущество перед неиерархическими методами в том, что кластеры определяются исключительно данными (а не пользователями, предопределяющими количество кластеров), и что количество кластеров может быть увеличено или уменьшено простым перемещением вверх и вниз по иерархии. Иерархия создается либо путем запуска сверху (один кластер, который включает в себя все записи), и разбиением (разделяющая кластеризация), либо же началом сверху с количеством кластеров, равным количеству записей и слияния (агломерационной кластеризация). Обычно слияние и разделение выполняются одновременно по двум кластерам.

Рис. 10. Иерархическая кластеризация

Однако, в отличии от кластеризации K-средних, иерархическая кластеризация не может обрабатывать большие данные. Это связано с тем, что вычислительная сложность K-средних линейна, а именно, $O(n)$, а иерархической кластеризации квадратична, т.е. $O(n^2)$. Также, в кластеризации K-средних, поскольку мы начинаем со случайного выбора кластеров, результаты, полученные при запуске алгоритма несколько раз, могут отличаться. K-средних работает хорошо, когда форма кластеров является гиперсферической (например, круг в 2D, сфера в 3D), в отличии от иерархических алгоритмов.

Основной алгоритм K-средних был расширен многими способами. Некоторые из этих расширений основаны на дополнительных эвристиках, связанных с минимальным размером кластера, слиянием и разбиением кластеров. Двумя хорошо известными вариантами K-средних в литературе по распознаванию образов являются ISODATA Ball and Hall (1965) и FORGY Forgy (1965). В K-средних каждая точка данных назначается одному кластеру (называемому жестким назначением). Нечеткие C-средних, предложенные Данном (1973), а затем улучшенные по Бездеку (1981), являются расширением K-средних, где каждая точка данных может быть членом множества кластеров со значением принадлежности (мягкое присвоение).

Следующим алгоритмом, на который стоит обратить внимание является кластеризация среднего сдвига (Mean shift clustering). Это алгоритм, основанный на скользящем окне, который пытается найти плотные области данных. С другой стороны, данный алгоритм основан на центроидах, и его цель состоит в том, чтобы найти центральные точки каждой группы / класса. Кандидаты на центральные точки выявляются из среднего значения точек в скользящем окне. Затем эти окна-кандидаты фильтруются на этапе последующей обработки, чтобы устранить дубликаты, образуя окончательный набор центральных точек и их соответствующих групп. Пусть кандидат для центроида будет x_i, при итерации t, тогда кандидат обновляется в соответствии со следующим уравнением

$$x_i^{t+1} = x_i^t + m(x_i^t)$$

Где $N(x_i)$, - окрестность выборок на заданном расстоянии вокруг x_i, а m - средний вектор сдвига, который вычисляется для каждого центроида и указывает на область максимального увеличения плотности точек. Данный сдвиг вычисляется с использованием следующего уравнения, эффективно обновляющего центроид:

$$m(x_i) = \frac{\sum_{x_j \in N(x_i)} K(x_j - x_i) x_j}{\sum_{x_j \in N(x_i)} K(x_j - x_i)}$$

Данный алгоритм автоматически устанавливает количество кластеров, вместо того, чтобы полагаться на пропускную способность параметра, определяющего размер области для поиска. Алгоритм среднего сдвига не обладает высокой масштабируемостью, поскольку для выполнения алгоритма требуется несколько поисков ближайшего соседа. Алгоритм гарантированно сходится, однако прекращает итерацию, если изменений в центроидах мало. Маркировка нового образца выполняется путем нахождения ближайшего центроида для данного образца.

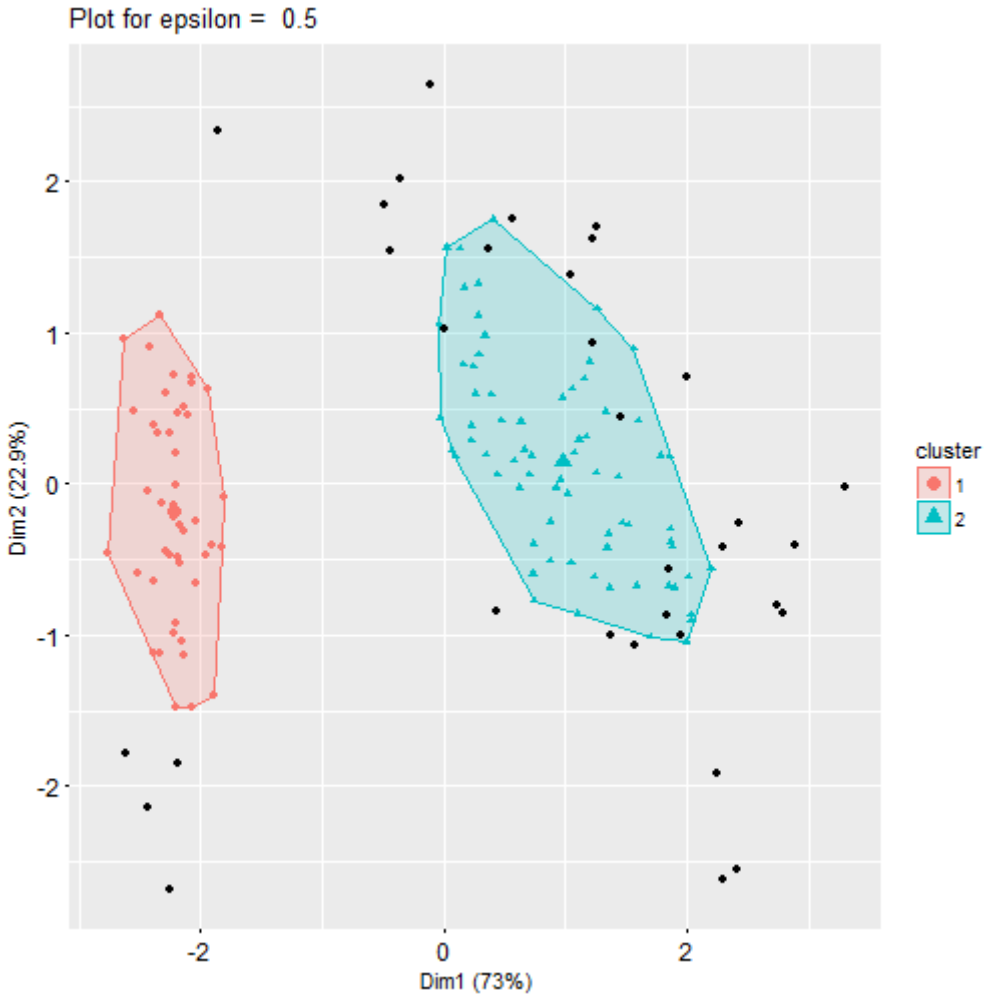
Алгоритм DBSCAN (рис.11) (Density based Clustering Algorithm-Алгоритм Кластеризации на Основе Плотности) рассматривает кластеры как области высокой плотности, разделенные областями с низкой плотностью. Из-за этого довольно общего вида кластеры, найденные DBSCAN, могут быть любой формы, в отличие от k-средних, которые предполагают, что кластеры выпуклые. Центральным компонентом DBSCAN является концепция образцов ядра, которые являются образцами, находящимися в областях с высокой плотностью. Таким образом, кластер представляет собой набор образцов ядра, каждый из которых близок друг к другу (измеряется с помощью некоторой меры измерения расстояния) и набор неосновных образцов, которые близки к образцу ядра (но сами не являются образцами ядра). Для алгоритма есть два параметра: min_samples и eps, которые формально определяют, что мы имеем в виду, когда называем область плотной. Более высокие min_samples или более низкие eps указывают на высокую плотность, необходимую для формирования кластера.

Выглядит данный алгоритм следующим образом:

1. Пометьте все точки как ядро, граница или шум.
2. Устранить помехи.
3. Поместите грань между всеми основными точками, находящимися внутри на расстоянии параметра Eps друг от друга.
4. Заключить каждую группу соединенных основных точек в отдельный кластер.
5. Назначить каждую граничную точку одному из кластеров опорных точек, связанных с данной граничной точкой.

Любая базовая выборка является частью кластера по определению. Любой образец, который не является образцом ядра, и существует, по меньшей мере, на расстоянии eps от любого образца ядра, считается аномалией алгоритма.

Алгоритм, основанный на плотности (Density based Clustering Algorithm), позволяет данному кластеру продолжать расти до тех пор, пока плотность в соседнем кластере не превышает определенный порог. Этот



алгоритм подходит для обработки шума в наборе данных.

Рис.11. Визуализация кластеров на основе алгоритма DBSCAN

В качестве особенностей этого алгоритма перечислены следующие моменты: он обрабатывает кластеры произвольной формы, обрабатывает шум, требуется только одно сканирование входного набора данных и параметры плотности, которые должны быть инициализированы. DBSCAN, DENCLUE и OPTICS являются примерами этого алгоритма.

Алгоритм DBSCAN является детерминированным, всегда генерируя одни и те же кластеры при предоставлении тех же данных в том же порядке.

Однако результаты могут отличаться, если данные предоставляются в другом порядке. Во-первых, несмотря на то, что выборки ядра всегда будут назначены одним и тем же кластерам, метки этих кластеров будут зависеть от порядка, в котором эти образцы встречаются в данных. Во-вторых, что более важно, кластеры, которым назначены неосновные образцы, могут различаться в зависимости от порядка данных. Это произойдет, когда непрозрачный образец будет иметь расстояние меньше, чем eps, до двух образцов центроида в разных кластерах. По треугольному неравенству эти два основных образца должны быть более отдаленными друг от друга, чем параметр eps или они будут в одном кластере. Неядерный образец присваивается кластеру, который генерируется первым в проходе через данные, и поэтому результаты будут зависеть от упорядочения данных.

Спектральная кластеризация относится к классу методов, опирающихся на собственную структуру матрицы подобия. Кластеры формируются путем разбиения точек данных с использованием матрицы подобия. Любой алгоритм спектральной кластеризации будет иметь три основных этапа: предварительную обработку, спектральное отображение и пост-картографирование. Предварительная обработка имеет дело с построением матрицы подобия. Спектральное картирование касается построения собственных векторов для матрицы подобия. Постобработка обрабатывает группировку точек данных. Алгоритм спектральной кластеризации объективно просто реализовать. Он не учитывает локальные оптимумы, является статистически согласованным, и имеет сравнительно высокую скорость. Основным, и можно сказать единственным, недостатком этого подхода является то, что он демонстрирует высокую вычислительную сложность. Для большого набора данных требуется O (n3), где n - количество точек данных. Примерами этого алгоритма являются алгоритм SM (Shi&Malik), KVV (Kannan, VempalaandVetta) и алгоритм NJW (Ng, Jordan и Weiss).

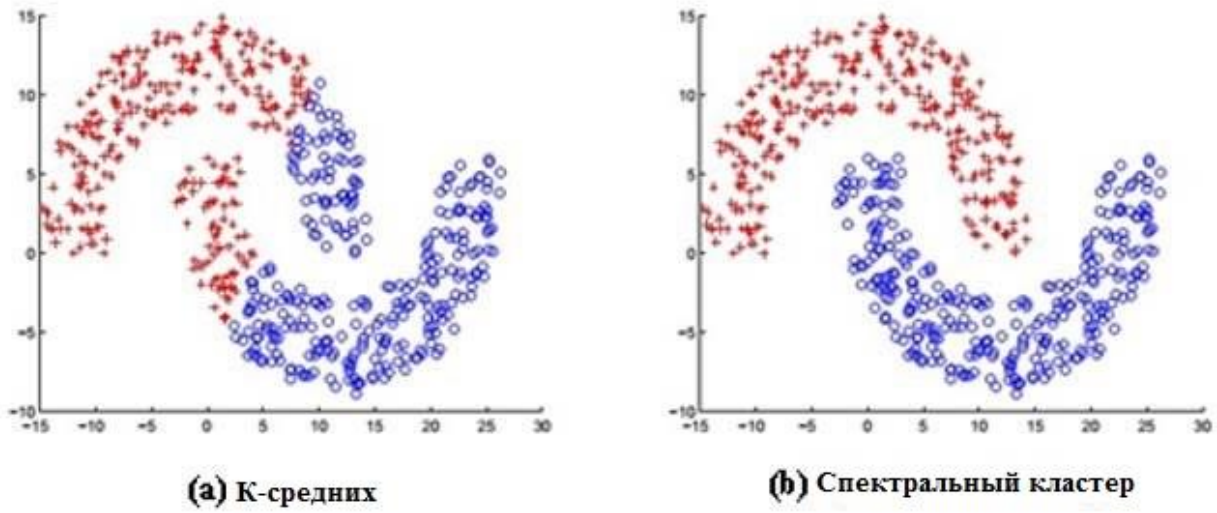


Рис. 12. Сравнение результов, полученных при использовании алгоритма К-средних и спектральной кластеризации

Алгоритм, основанный на сетке (рис. 13), (Grid based Clustering Algorithm) определяет размер пространства объекта в конечном числе ячеек, который формирует структуру сетки. Операции выполняются на этих сетках. Преимуществом этого метода является его меньшее время обработки. Сложность кластеризации основана на количестве заполненных ячеек сетки и не зависит от количества объектов в наборе данных. Основными особенностями этого алгоритма являются вычисления расстояний, и то, что кластеризация выполняется в суммированных точках данных. Формы ограничены объединением ячеек сетки, а сложность алгоритма обычно равна O (количество заполненных сетчатых ячеек). Алгоритм STING является примером этого алгоритма.

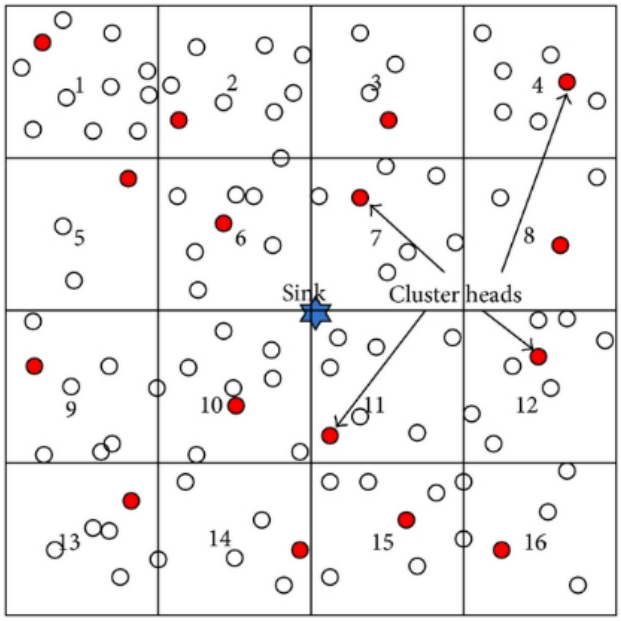


Рис. 13. Алгоритм, основанный на сетке

Следует также отметить алгоритм разбиения. Методы разбиения обычно приводят к множеству кластеров M, каждый из которых принадлежит одному кластеру. Каждый кластер может быть представлен центроидом или представителем кластера; это своего рода краткое описание всех объектов, содержащихся в кластере. Точная форма этого описания будет зависеть от типа объекта, который кластеризуется. В тех случаях, когда доступны вещественные данные, соответствующим представителем является среднее арифметическое векторов атрибутов для всех объектов в кластере. Альтернативные типы центроидов могут потребоваться в других случаях. Например, кластер документов может быть представлен списком тех ключевых слов, которые встречаются в некотором минимальном количестве документов внутри кластера. Если количество кластеров большое, центроиды могут быть дополнительно сгруппированы для создания иерархии в наборе данных.

Алгоритм кластеризации разбиения разбивает точки данных на k-разбиений, где каждое разбиение представляет кластер. Разбиение выполняется на основе определенных целевых функций. Одной из таких

критериальных функций является минимизация критерия квадратичной ошибки, который вычисляется как,

$$E = \sum\sum\|p - mi\|2$$

,где p- точка в кластере, а mi - среднее значение кластера. Причем, кластер должен обладать двумя свойствами:

1. каждая группа должна содержать хотя бы один объект
2. каждый объект должен принадлежать ровно одной группе

Основной недостаток этого алгоритма в том, что всякий раз, когда точка близка к центру другого кластера, он дает плохой результат из-за перекрытия точек данных.

Single Pass - очень простой метод разбиения, создающий разделенный набор данных в три этапа. В первую очередь, находится объект центроида для первого кластера. Для следующего объекта вычисляется сходство S с каждым существующим кластером, используя некоторый коэффициент подобия. Наконец, если высшая вычисленная S больше некоторого заданного порогового значения, то объект добавляется к соответствующему кластеру, а ценроид определяется заново. В противном случае данный объект используется для запуска нового кластера.Если какие-либо объекты не являются кластеризованными, следует вернуться к предыдущему шагу.Как следует из названия метода, для него требуется только один проход через набор данных. Требование времени обычно порядка $O(NlogN)$ для порядка $O(logN)$ кластеров, что делает это очень эффективным методом кластеризации для последовательного процессора. Недостатком является то, что результирующие кластеры не зависят от порядка обработки документов, причем первые кластеры обычно больше, чем созданные в ходе кластеризации.Выше упомянутый алгоритм кластеризации K-средних является одним из алгоритмов кластеризации на основе разбиения.

Теперь рассмотрим типы алгоритмов, классифицируемые по способу анализа данных. Нечеткая кластеризация обобщает методы кластеризации

(такие как k-средних и медоиды), позволяя объекту частично классифицироваться более чем один кластером. В Четких же (или непересекающиеся) алгоритмах каждый объект является членом одного единственного кластера.Предположим, мы имеем К-кластеров, и определяем набор переменных $m_{i1}, m_{i2}, \dots, m_{in}$, представляющих вероятность того, что объект i классифицируется в кластер К. В четких алгоритмах кластеризации одно из этих значений будет равно единице, а остальные нулю. Это означает, что эти алгоритмы классифицируют объект в один и только один кластер. В нечеткой кластеризации членство распространяется среди всех кластеров. Так, m_{ik} может быть от нуля до единицы, при условии, что сумма их значений равна единице. Это называется фазификацией конфигурации кластера. Преимущество состоит в том, что он не заставляет каждый объект принадлежать конкретный кластер. Однако, недостаток заключается в интерпретации гораздо большего объема информации.

Одним из наиболее распространённых нечетких алгоритмов был описан Кауфманом (1990). Он стремится минимизировать следующую объектную функцию C, состоящую из кластера членов и расстояния.

$$C = \sum_{k=1}^{K} \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} m_{ik}^2 m_{jk}^2 d_{ij}}{2 \sum_{j=1}^{N} m_{jk}^2}$$

,где m_{ik} представляет неизвестную принадлежность объекта i кластеру К, а d_{ij} несходство между объектами i и j. Принадлежность объекта ограничена тем, что она должна быть неотрицательной.Стоит отметить, что во многих ситуациях нечеткая кластеризация более естественна, чем четкая.

Хороший обзор кластеризации на основе нечетких множеств представлен Бейкером (1978). Сокращение данных путем замены представителей групп их центроидами до процесса кластеризации было использовано для ускорения К-средних и нечетких С-средних у Эшриха(2003)

Говоря про алгоритм К-средних, нельзя не отметить одну из его разновидностей- Генетический алгоритм К-средних (ГА К-средних).

К.Кришна и М.Нарасимха Мурти предложили данный новый гибридный генетический алгоритм (ГА), который находит глобально оптимальное разбиение данных на определенное количество кластеров. ГА, используемые ранее в кластеризации, используют либо дорогостоящий оператор кроссовера для генерации допустимых дочерних хромосом из родительских хромосом, либо дорогостоящую функцию фитнеса, либо и то, и другое. Чтобы обойти эти дорогостоящие операции, они гибридизировали ГА с классическим алгоритмом спуска градиента, используемым в кластеризации, а именно с алгоритмом K-средних. Они определили оператор K-средних, один шаг алгоритма K-средних и использовали его в ГА K-средних в качестве поискового оператора вместо кроссовера. Также, ими был определен предвзятый оператор мутации, специфичный для кластеризации, называемый мутацией на основе расстояния. Используя конечную цепь теории Маркова, они доказали, что ГА K-средних сходится к глобальному оптимуму. В симуляциях видно, что ГА K-средних сходится к наиболее известному оптимуму, соответствующему данным, в соответствии с результатом сходимости. Также наблюдается, что ГА K-средних ищет быстрее, чем некоторые другие эволюционные алгоритмы, используемые для кластеризации. Преимущество ГА K-средних заключается в том, что он быстрее, чем некоторые другие алгоритмы кластеризации.

III ГЛАВА. ПРАКТИЧЕСКОЕ ПРИМЕНЕНИЕ КЛАСТЕРНОГО АНАЛИЗА

3.1. Геокластерный анализ конкретных данных

В настоящее время мы обладаем огромным количеством данных, созданных как службами социальных сетей, так и полученных из реальной жизни, и все чаще эти данные содержат информацию о местоположении, которое дает нам широкий спектр возможностей для их анализа. Поскольку нас может заинтересовать не только сам контент, но и место, где было создано или запланировано его содержимое. Для точного и максимально успешного анализа геопространственных данных необходимо найти наилучший подход к геоклассификации. И на данный момент - это кластеризация массивных геоданных в реальном времени с высокой точностью[9]. Кластеризованные геоданные, основанные на их местоположении, улучшают их визуальный анализ и улучшают ситуационную осведомленность. Изучение геопространственных данных - это извлечение неявных знаний, геопространственных отношений и интересных характеристик и шаблонов, которые явно не представлены в геопространственной базе данных. В настоящее время основным применением геопространственных данных является анализ географических данных с целью извлечения и передачи геопространственной информации. Геопространственный анализ - это суть географических информационных систем, позволяющая получать скрытую информацию и знания из географических данных. Геопространственные данные могут использоваться в картографии и геодезических работах, при этом не ограничиваясь только этими рамками. Геопространственная кластеризация является важной областью исследований в области геопространственного сбора данных. Алгоритмы кластеризации играют ключевую роль в методах

пространственного анализа. Для повышения полезности кластеризации крайне важно изучить анализ геопространственных кластеров на основе ограничений. Концепция внедрения подхода, основанного на геопространственном фоне, в решении геопространственных задач не является новым процессом, о чем свидетельствуют многочисленные работы.

Методы геопространственной кластеризации в основном подразделяются на четыре типа: иерархические, разделенные, основанные на плотности и сетки [11]. Рассмотренный раннее алгоритм кластеризации DBSCAN имеет несколько недостатков. В частности, существует нехватка географических справочных знаний (ограничения геопространственной кластеризации), и возникают значительные трудности при приобретении и представлении отношений между геопространственными знаниями.

Далее рассмотрим подход к группированию методом кластеризации геоданных для онлайн-карт, а в данном случае азербайджанской карты GoMap (https://gomap.az). GoMap-это детальная карта Азербайджана, созданная на основании картографических данных, таблиц, схем и полевых исследований и разработанная компанией “SINAM”. Она включает в себя более 100 тысяч ПОИ, причем день за днем количество этих ПОИ неустанно растет, а, следовательно, появляется необходимость использовать разнообразные подходы для анализа имеющихся данных [21]. Далее покажем, как на реальном примере кластеризация геоданных упрощает анализ.

Для отображения данных динамической карты в веб-браузерах используется JavaScript библиотека с открытым исходным кодом, выпущенная под лицензией BSD 2-ой статьи (также известной как FreeBSD)-OpenLayers. Она может отображать тайлы карт, векторные данные и маркеры, загруженные из любого источника. OpenLayers была разработана для дальнейшего использования географической информации всех видов. Она предоставляет API для создания богатых веб-географических приложений, подобных картам Google, или, как в нашем случае GoMap [12].

Следует отметить, что OpenLayers поддерживает языки разметки GeoRSS, KML, GeoJSON, а также данные карты из любого источника, используя стандарты OGC в качестве службы веб-карт (WMS) или службы веб-функций (WFS). Стандартной проекцией для отображения карты, является проекция Web Mercator (EPSG: 3857). А именно, проецирование сферической поверхности на двумерную плоскость. Следует отметить, что данная проекция также положительно влияет при вычислении расстояния кластеров. Для базового создания карты нам понадобится библиотеки Jquery, OpenLayers, которые мы можем как загрузить, так и использовать онлайн. И простая html-страница, с элементом div, имеющего id=’map’. В первую очередь, процесс создания карты начинается с инициализации тайл-слоя. Слой карты GoMap.Az работает по подобию слоев OSM (OpenStreetMap). При этом для управления библиотекой OpenLayers передаются три параметра x, y, x, где z представляет собой уровень масштаба. Параметр lng задает язык надписей карты (az, en, ru). Параметр f указывает формат фрагментов карты.

```
openCycleMapLayer = new ol.layer.Tile({
  source: new ol.source.OSM({
    attributions: [
      new ol.Attribution({
        html: '© ' + ' <a href="https://gomap.az/">GoMap </a>'
      })
    ],
  crossOrigin: null,
   url: 'http://gomap.az/info/xyz.do?lng=ru&x={x}&y={y}&z={z}&f=jpg'
})
});
```

Затем инициализируем саму карту.

```
map = new ol.Map({
 layers: [openCycleMapLayer, vectorLayer],
 target: 'map',
 controls: ol.control.defaults({
   attributionOptions: ({
     collapsible: false})
   }),
 view: new ol.View({
 minZoom: 8,
 maxZoom: 19,
 center: [5551660,4921653],
 zoom: 15
 })
});
```

Здесь, в разделе layers указываются все используемые слои, а во view можно указать центр карты и уровень масштаба, которые позволят установить режим отображения карты при первом открытии страницы.

Данным образом динамическую карту Азербайджана можно считать созданной. В дальнейшем этот слой карты будет нам служит скорее визуальным помощником, для понимания зависимости объектов. Далее, попробуем добавить большой массив объектов на эту карту. Как было отмечено ранее, OpenLayers принимает несколько форматов данных, однако, в данном случае представим объекты в GeoJSON – формате, который используется для кодирования различных структур географических данных. GeoJSON поддерживает следующие типы геометрии: Point, LineString, Polygon, MultiPoint, MultiLineString и MultiPolygon. Геометрические объекты с дополнительными свойствами являются объектами Feature. Наборы функций содержатся объектами в FeatureCollection. В нашем примере

используется тип Point, так как в дальнейшем эти же данные будут кластеризированы, а кластеризация точек является наиболее типичной задачей в геопространственной кластеризации, и многие виды кластеризации пространственных объектов могут быть абстрагированы или преобразованы в кластеры точек. Во многих ситуациях геопространственные объекты представлены точками, такими как города, распространяющиеся в регионе, объекты в городе (ПОИ) и точки пространственной выборки по некоторым причинам исследования, например, образцы рудного сорта в анализе качества руды, образцы высот при анализе местности, образцы цен на землю при оценке земли и т. д. Геопространственная кластеризация является наиболее сложной среди всех методов интеллектуального анализа пространственных данных. Пусть объекты хранятся в файле markers.json. Это необязательный пункт, данные могут храниться как в базе данных, так и считываться из файлов. Более того, есть возможность преобразить объекты из обычного JSON-формата в GeoJSON-формат. Как видно, каждый feature состоит из geometry, в котором в обязательном порядке указываются координаты и тип. Далее инициализируем объекты, добавляя их на векторный слой.

```
var geojsonObject = " markers.json";
  var vectorSource = new ol.source.Vector({
      features: (new ol.format.GeoJSON()).readFeatures(geojsonObject)
    });
    var vectorLayer = new ol.layer.Vector({
      source: vectorSource,
      style: styleFunction
    });
```

```
{
 "type": "FeatureCollection",
 "features": [
  { "type": "Feature",
    "properties": { "id":1},
    "geometry": {
      "type": "Point",
      "coordinates": [ 49.405517578125, 40.371658891506094]
    }},
  {"type": "Feature",
    "properties": { "id":2},
    "geometry": {
      "type": "Point",
      "coordinates": [49.2791748046875, 40.30466538259176]
    }},
  { "type": "Feature",
    "properties": { "id":3},
    "geometry": {
      "type": "Point",
      "coordinates": [ 49.833984375, 40.44694705960048]
    }}
 ]}
```

Таким образом, запустив приложение, мы добавили 134 тысячи объектов на карту.

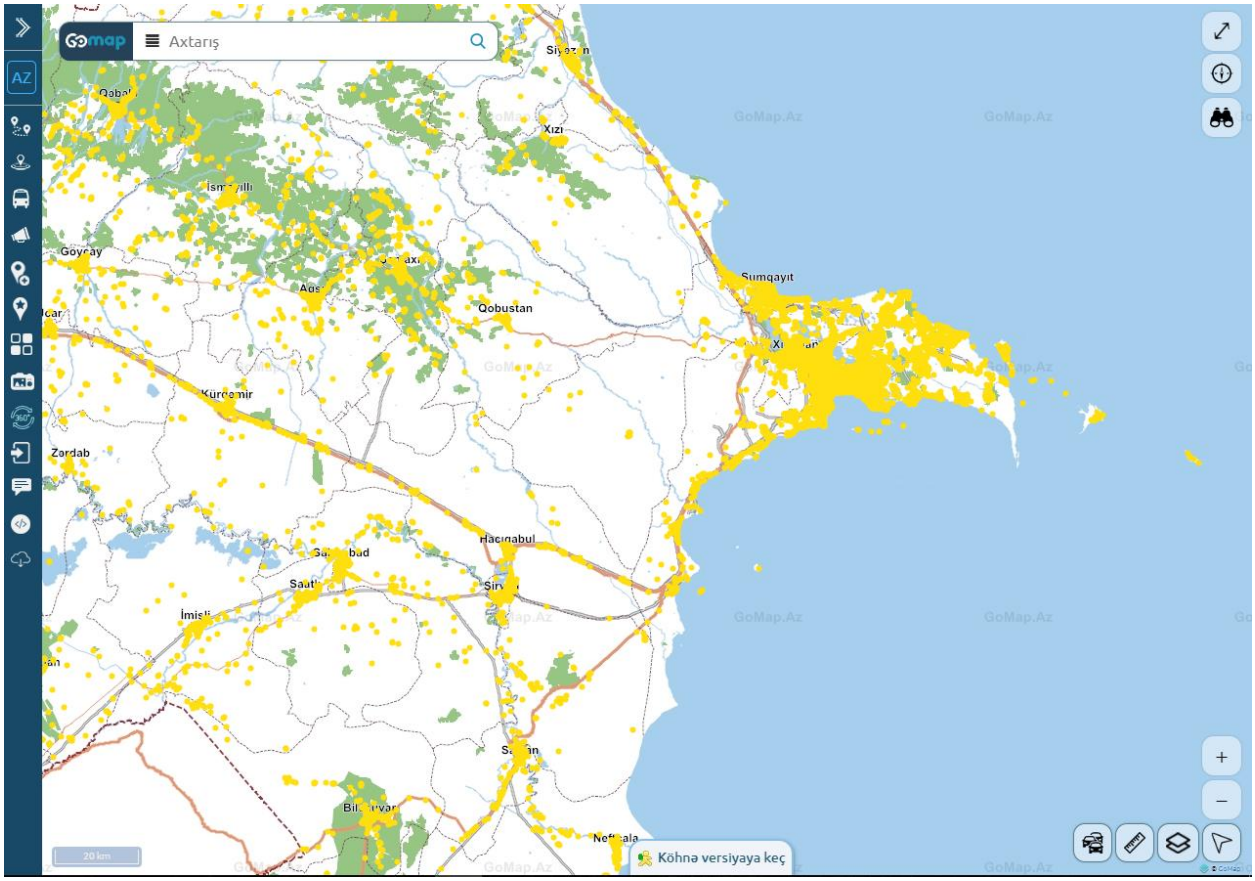


Рис. 14. Отображение 134 тысячи объектов на 9 zoom-е без использования кластеризации

Однако, как можно заметить на рис. 14, полученные маркеры находятся друг на друге, в связи с чем сложно не только хоть что-то проанализировать, но даже разобрать какой город или административный округ где находится. При приближении карты на 13 zoom, результат остается неизменным, и все объекты, ввиду большого количества, наложены друг на друга (рис.15). В связи с чем, уместнее всего будет применить метод кластеризации для анализа выделенного массива объектов. Для этого, вместо слоя vectorlayer создаем слой clusterlayer.

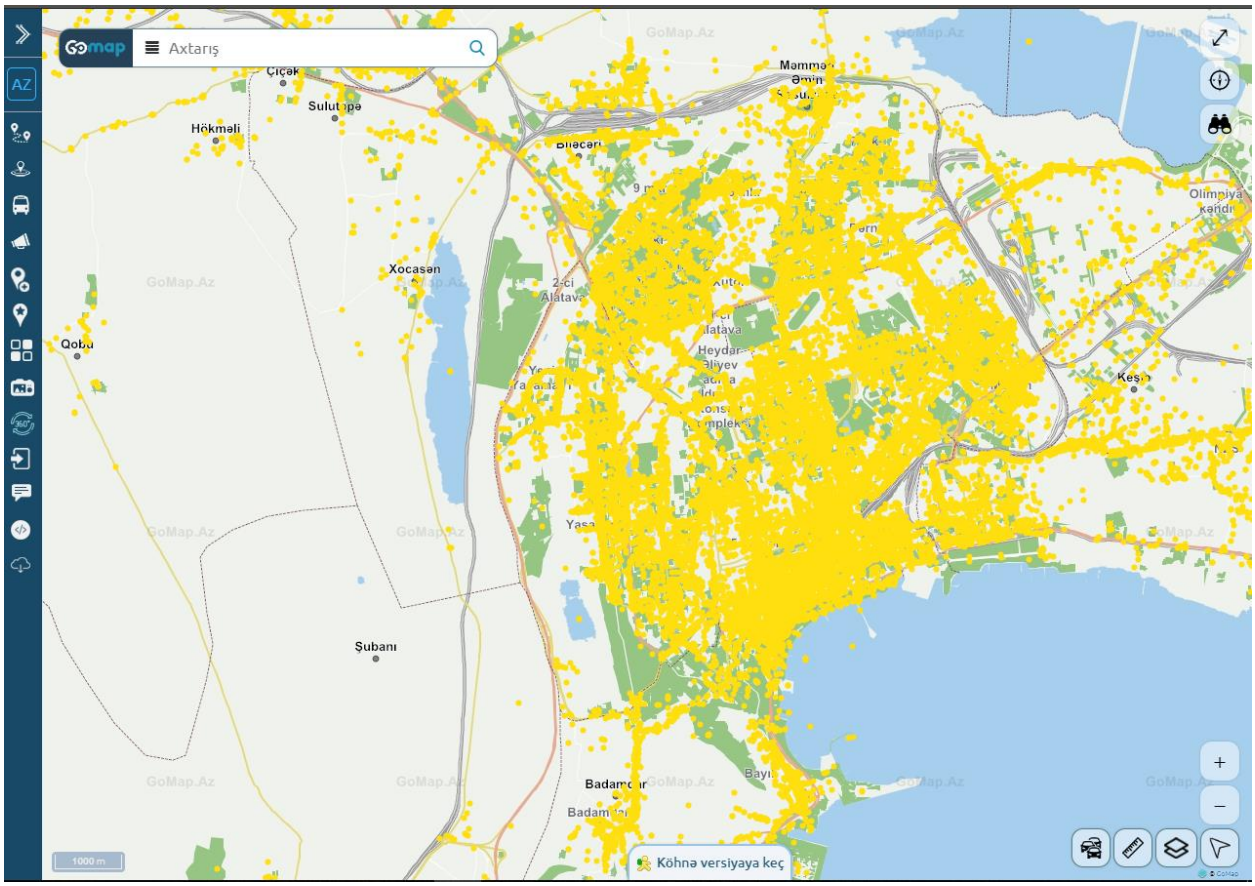


Рис. 15. Отображение 134 тысячи объектов на 13 zoom-е без использования кластеризации

```
var source = new ol.source.Vector({features: features});
    var clusterSource = new ol.source.Cluster({distance: n, source: source});
    var styleCache = {};
    var clusterlayer = new ol.layer.Vector({
      source: clusterSource,
      style: function(feature) {
        var size = feature.get('features').length;
        var style = styleCache[size];
        if (!style) {style = style}}});
```

И запускаем процесс, используя уже этот слой.

Как видно на рис. 16 кластеризация объектов прошла успешно. Даже на 7 zoom-е объекты выглядят аккуратно и понятно, уже можно сделать

определенные выводы о имеющихся данных. При помощи заданной функции style можно отобразить зависимость размера кластера от размера точки. Указанное на кластере число показывает, сколько объектов входит в данный кластер. На рис. 17, сделано приближение на пять уровней к центру Баку, в связи с чем можно заметить, что размер кластеров на выделенном участке уменьшился, а количество же напротив- увеличилось. При увеличении масштаба карты маркеры снова объединяются в кластеры.

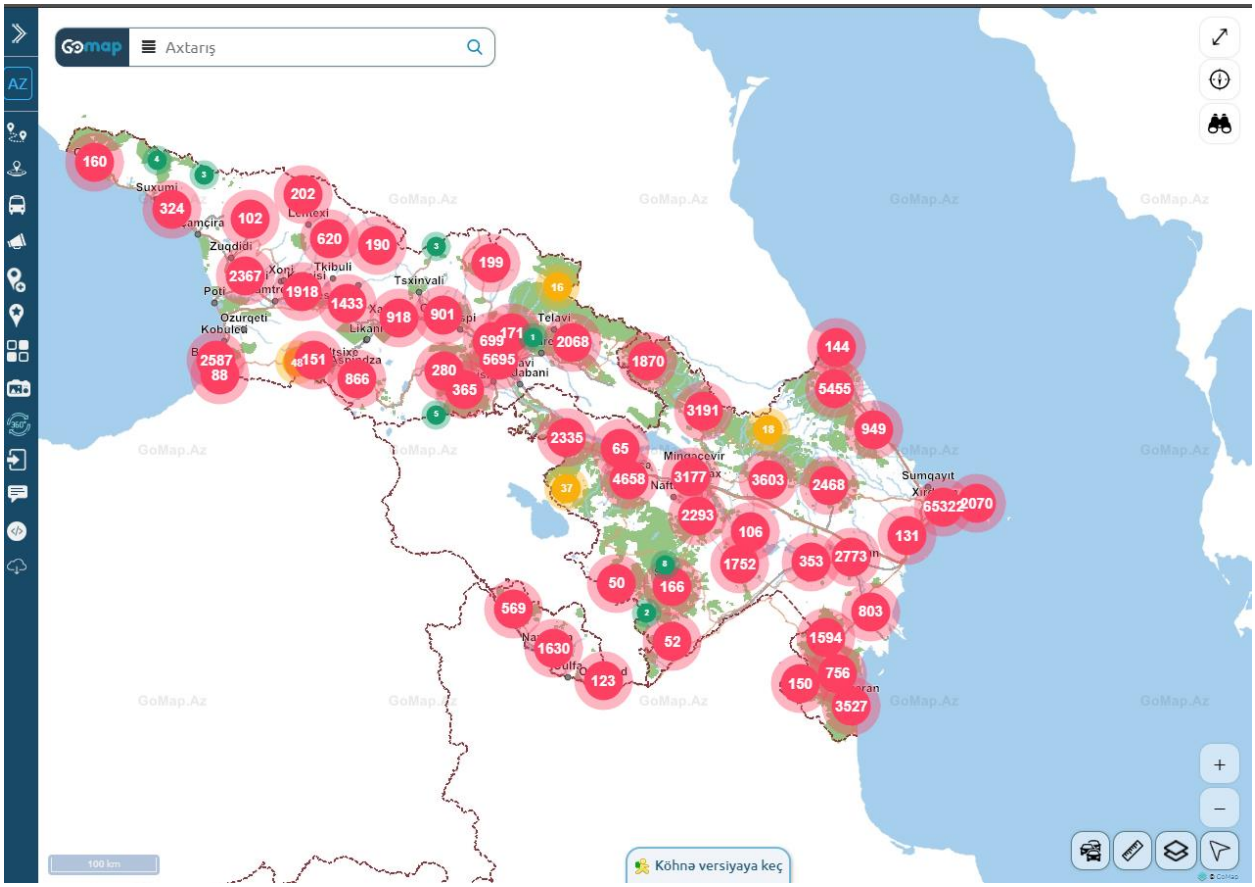


Рис. 16. Отображение 134 тысячи объектов на 7 zoom-е с использованием кластеризации

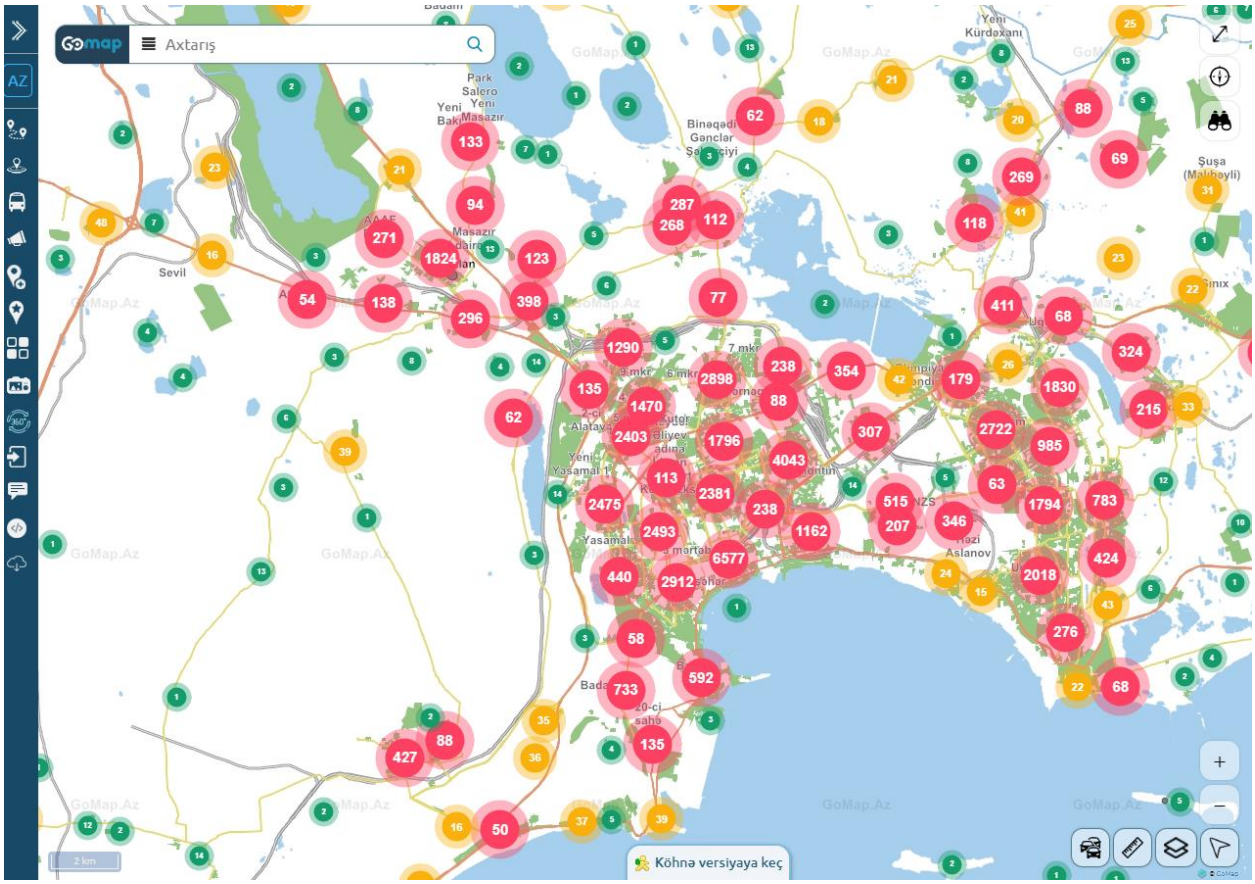


Рис. 17. Отображение 134 тысячи объектов на 12 zoom-е с использованием кластеризации

Библиотека OpenLayers использует метод кластеризации на основе сетки, который делит карту на квадраты определенного размера (размер изменяется на каждом уровне масштабирования) и группирует маркеры в каждую квадратную сетку. Он создает кластер на определенном центральном маркере и добавляет маркеры, находящиеся в границах кластера. Данный процесс повторяется, пока все маркеры не будут выделены ближайшим сетчатым маркерным кластерам на основе уровня масштабирования карты. Если маркеры находятся в границах более чем одного существующего кластера, определяется расстояние маркера от каждого кластера и добавляет его в ближайший кластер. Для настройки позиции кластера, позволяющего отрегулировать среднее расстояние между всеми маркерами, которые содержатся в нем, задается параметр distance.

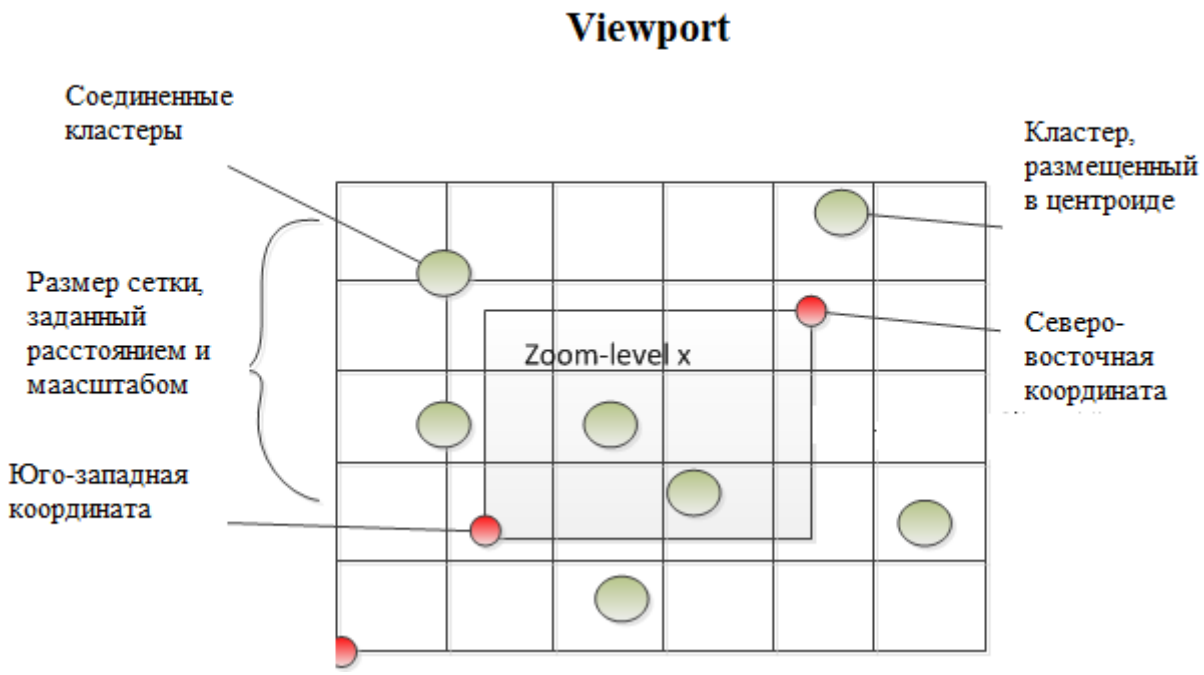


Рис. 18. Принцип алгоритма

Основная задача интеллектуального анализа геопространственных данных состоит в том, чтобы находить шаблоны данных в отношении его локального значения. Объем пространственной добычи увеличивается, поскольку увеличивается количество данных, полученных из различных источников. Многие сферы деятельности, таких как здравоохранение, маркетинг, природоохранные агентства, менеджмент используют геопространственную разработку для поиска информации, содержащейся в ней. Основные проблемы в области интеллектуального анализа геопространственных данных заключаются в том, что репозитории геопространственных данных имеют тенденцию быть очень большими, а диапазон и разнообразие представляют пространственные и не пространственные атрибуты данных в одном холсте. Хотя рассмотренный выше пример кластеризации направлен на решение такой проблемы, как масштабируемость и сложность, необходимо заметить, что идеальный алгоритм кластеризации, который понимает все проблемы с набором данных, является идеалистическим понятием. Текущие исследования продвигаются по более динамичным, адаптивным и инновационным методам, которые

расшифровывают значимые закономерности, способные эффективно удовлетворять требованиям обращения с огромными объемами данных более высоких размерностей, быть менее чувствительными к большим шумовым воздействиям, не зависеть от порядка ввода и не иметь предварительного знания области.

Как видно, Data Mining геопространственных данных- это поиск полезных ассоциаций и характеристик, которые могут неявно существовать в пространственных базах данных. Данный процесс концентрируется на автоматизации такого процесса, как обнаружение знаний. При этом, выполняются следующие функции

- Он играет важную роль в получении интересных пространственных моделей и характеристик;
- Отлавливает присущие ассоциаций между пространственными и не пространственными данными;
- Представляет достоверные данные на концептуальном уровне
- Оказывает помощь в реорганизации пространственных баз данных для использования семантики данных, в дополнение к повышению производительности

Пространственная кластеризация данных является важнейшей составляющей интеллектуального анализа пространственных данных и реализуется как таковая для извлечения шаблона из распределения объектов данных в конкретном наборе данных, и, как упоминалось ранее, оно имеет несколько направлений, таких как спутниковые снимки, географические информационные системы, и т.д.

3.2. Проблемы, возникающие в процессе кластеризации

Кластеризация сложна, потому что это неконтролируемая проблема обучения: нам предоставляется набор данных, и им предлагается определить

внутри него структуру (в данном случае - скрытые кластеры / категории в данных). Проблема заключается в том, что необязательно есть «правильное» решение, на которое можно ссылаться, если хотим проверить наши ответы. Это противоречит проблемам классификации, где правильный ответ заранее известен. Глубокие искусственные нейронные сети очень хороши в классификации, но кластеризация по-прежнему остается открытым вопросом. Например, рассмотрим применение кластеризации в здравоохранении. Основывая на классификацию, можно предсказать, имеет ли пациент общее заболевание, основанное на списке симптомов. Для этого можно опираться на прошлые клинические записи, или же собрать дополнительные данные (например, анализ крови), чтобы подтвердить полученное предсказание. Другими словами, предполагается, что существует самоочевидная основная истина (пациент либо имеет, либо не имеет заболевания X), которую можно наблюдать и проверить.

Для кластеризации же этой важной информации критически не хватает. Например, предположим, что испытуемому дали большое количество жуков и сказали группировать их в кластеры на основе их внешнего вида. Предполагая, что он не энтомолог, это будет связано с некоторыми суждениями и догадками. Если два испытуемых сортируют те же 100 жуков по 5 группам, то, вероятно, полученные ответы будут несколько отличаться. И - вот важная часть - на самом деле нет способа определить, кто из испытуемых «прав». Наиболее широко распространённая проблема кластеризации- это затруднение при определении количества кластеров в наборе данных. А именно, нет «истинного» количества кластеров (хотя некоторые цифры более близки к идеалу) и один и тот же набор данных надлежащим образом просматривается на разных уровнях определения в зависимости от целей анализа. Хотя эта проблема не может быть «решена» окончательно, существует довольно успешные способы борьбы с ней.

Иерархические кластерные методы обеспечивают кластерные назначения для всего возможного количества кластеров, позволяя аналитику

просматривать данные на разных уровнях детализации. Существуют также байесовские методы, которые адаптивно оценивают количество кластеров на основе гиперпараметра, настраивающего дисперсию. В ряде недавних работ основное внимание было уделено методам выпуклой кластеризации, которые объединяют кластерные центры вместе непрерывным образом вдоль пути регуляризации, при этом раскрывая иерархическую структуру для подхода к методу кластеризации, аналогичного k-means. Конечно, есть много других работ по этому вопросу. Наконец, еще одна проблема, связанная с кластеризацией, возникает из-за формы кластеров данных. Каждый алгоритм кластеризации делает структурные предположения о наборе данных, которые необходимо учитывать. Например, k-means работает, сводя к минимуму общее суммарное расстояние до центроидов кластера. Это может привести к нежелательным результатам, когда кластеры удлиняются в определенных направлениях, особенно если расстояние между кластерами меньше максимального расстояния внутри кластера. В отличие от этого, кластеризация с одной связью может хорошо работать в этих случаях, поскольку точки группируются вместе на основе их ближайшего соседа, что облегчает кластеризацию по «путям» в наборе данных.

На сегодняшний день, существует много эвристик, способных могут помочь преодолеть вышеупомянутые проблемы, однако, стоит подчеркнуть, что пока это всего лишь эвристика, а не гарантии. Одной из наиболее интересных эвристик, заслуживающих внимания, является ансамблевая кластеризация. Ее основная идея состоит в том, чтобы усреднить результаты нескольких методов кластеризации, или же применять различные случайные инициализации к одной и той же методике. По отдельности полученные кластеры могут быть не стабильны, однако среднее поведение ансамбля моделей будет наиболее удовлетворительным. Это решение называется усреднением ансамбля и успешно применяется к ряду проблем машинного обучения. На ряду с этим, многолетняя проблема кластерного анализа заключается в том, насколько серьезно относиться к полученным

кластерам. Этот вопрос, конечно, связан с выбором наилучшего количества кластеров. С одной стороны, можно рассматривать кластеры как чистые фикции, просто более или менее удобные способы суммирования некоторых частей данных без какого-либо другого смысла. С другой стороны, можно утверждать, что они отражают реальные разделения мира на разные типы. Это хороший показатель того, насколько серьезно нужно принимать во внимание теоретические конструкции и понимать, должна ли теория переходить в практику. Это особенно заманчиво, при использовании конкретных примеров, а именно, например, добавления значимых имен в кластеры (в отличии от простых названий «кластер 1», «кластер 2», ... «кластер k»)

3.3. Требования к успешному кластерному анализу

Для правильно понимания того, нужна ли кластеризация в каком-то конкретном случае, нужно учитывать следующие три правила «хорошего» кластера:

1. Хорошие кластеры должны хорошо обобщаться. Выше уже было сказано об этом, а если кратко, то кластеры должны продолжать описывать новые наблюдения по тем же признакам.
2. Хорошие кластеры должны обобщаться по новым, неявным показателям.
3. Хорошие кластеры должны вписываться в теорию, быть частью действительной системы обобщений, которая позволяет делать прогнозы о новых условиях и объясняет, почему все получается именно так, а не иначе.

Первый пункт является основным требованием, который, собственно, и определяет полезность кластеризации. Второй пункт значительно более строгий, однако его нужно рассматривать осторожно, поскольку для того,

чтобы быть действительными, кластеры не обязательно должны быть релевантными. Наконец, будучи частью хорошо устоявшейся теории, на практике в большинстве случаев, исследователи редко начинают выявлять кластеры в первую очередь. Не мало времени тратится на другие виды анализов, которые не выдают желаемого результата. Лишь после этого аналитики прибегают к кластерному анализу, на основе которого первичные анализы уже способны выявить нужные закономерности. Ниже приведены типичные требования кластеризации при интеллектуальном анализе данных:

- Масштабируемость. Многие алгоритмы кластеризации хорошо работают на небольших наборах данных, содержащих менее нескольких сотен объектов данных; однако большая база данных может содержать миллионы или даже миллиарды объектов, особенно в сценариях веб-поиска. Кластеризация только по образцу заданного большого набора данных может привести к предвзятым результатам. Поэтому необходимы масштабируемые алгоритмы кластеризации.
- Возможность иметь дело с различными типами атрибутов: многие алгоритмы предназначены для кластеризации числовых (интервальных) данных. Однако приложениям может потребоваться кластеризация других типов данных, таких как двоичные, номинальные (категориальные) и порядковые данные, или смеси этих типов данных. В последнее время все больше приложений нуждаются в методах кластеризации для сложных типов данных, таких как графики, последовательности, изображения и документы.
- Обнаружение кластеров с произвольной формой: многие алгоритмы кластеризации определяют кластеры на основе мер евклидовой или манхэттенской дистанции. Алгоритмы, основанные на таких дистанционных измерениях, имеют тенденцию находить сферические кластеры с одинаковым размером и плотностью. Однако кластер может иметь любую форму. Рассмотрим, например, датчики, которые часто развертываются для наблюдения за окружающей средой. Кластерный

анализ показаний датчиков может обнаруживать интересные явления. Мы можем захотеть использовать кластеризацию, чтобы найти границу идущего лесного пожара, которая часто не является сферической. Важно разработать алгоритмы, которые могут определять кластеры произвольной формы.

- Требования к области знания для определения входных параметров. Многие алгоритмы кластеризации требуют от пользователей предоставления области знаний в виде входных параметров, таких как требуемое количество кластеров. Следовательно, результаты кластеризации могут быть чувствительны к таким параметрам. Параметры часто трудно определить, особенно для наборов данных с высокой степенью значимости, и когда пользователям еще предстоит понять глубокое понимание их данных. Требование спецификации знаний домена не только обременяет пользователей, но и затрудняет контроль качества кластеризации.
- Возможность обработки шумовых данных. Большинство наборов данных реального мира содержат выбросы и / или отсутствующие, неизвестные или ошибочные данные. Например, показания датчиков часто бывают шумными - некоторые показания могут быть неточными из-за механизмов восприятия, и некоторые показания могут быть ошибочными из-за помех от окружающих переходных объектов. Алгоритмы кластеризации могут быть чувствительны к такому шуму и могут создавать некачественные кластеры. Поэтому нам нужны методы кластеризации, которые устойчивы к шуму.
- Инкрементальная кластеризация и нечувствительность к порядку ввода: во многих приложениях инкрементные обновления (представляющие более новые данные) могут появляться в любое время. Некоторые алгоритмы кластеризации не могут включать инкрементные обновления в существующую структуру кластеров, и вместо этого приходится перерассчитывать новую кластеризацию с

нуля. Алгоритмы кластеризации также могут быть чувствительны к порядку ввода данных. То есть, учитывая набор объектов данных, алгоритмы кластеризации могут возвращать резко разные кластеры в зависимости от порядка представления объектов. Необходимы инкрементные алгоритмы кластеризации и алгоритмы, которые нечувствительны к порядку ввода.

- Возможность кластеризации высокоразмерных данных: набор данных может содержать множество измерений или атрибутов. Например, при кластеризации документов каждое ключевое слово можно рассматривать как измерение, и, зачастую есть тысячи ключевых слов. Большинство алгоритмов кластеризации хороши при обработке низкоразмерных данных, таких как наборы данных, содержащие только два или три измерения. Поиск кластеров объектов данных в высокоразмерном пространстве является сложным, особенно учитывая, что такие данные могут быть очень скудными и сильно искаженными.
- Кластеризация на основе ограничений: приложениям в реальном мире, возможно, потребуется выполнить кластеризацию при различных ограничениях. Предположим, что ваша задача - выбрать места для определенного количества новых банкоматов (банкоматов) в городе. Чтобы принять это решение, вы можете группировать домохозяйства при рассмотрении таких ограничений, как речные и автомобильные сети города, а также типы и количество клиентов на кластер. Задача состоит в том, чтобы найти группы данных с хорошим характером кластеризации, которые удовлетворяют заданным ограничениям.
- Интерпретация и удобство использования: пользователи хотят, чтобы результаты кластеризации были интерпретируемыми, понятными и пригодными для использования. То есть для кластеризации может потребоваться наличие связи с определенными семантическими интерпретациями и приложениями. Важно изучить, как цель приложения может влиять на выбор функций и методов кластеризации.

3.4 Причины использования кластерного анализа

Почему же стоит выбрать кластеризацию? Кластерный анализ распространен в любой дисциплине, которая включает в себя анализ многомерных данных. Поиск через Google Scholar (2018) обнаружил 46.400 записей в результате поиска выражения «кластерный анализ» и около 4.060.000 записей с использованием «cluster analysis». Такие обширные результаты говорят о важности кластеризации при анализе данных. Трудно перечислить все многочисленные научные области и приложения, в которых использовались методы кластеризации, а также тысячи опубликованных алгоритмов. Так, например,

- Сегментация изображений, важная проблема в компьютерном зрении, может быть сформулирована как проблема кластеризации.
- Документы могут быть кластеризованы по тематическим иерархиям для эффективного доступа к информации или поиску.
- Кластеризация также используется для группировки клиентов по разным типам для эффективного маркетинга. Она помогает маркетологам обнаружить различные кластеры в их базы клиентов, а затем использовать эти знания для маркетинговых программ.
- Кластеризация пригодна для управления и планирования рабочей силы
- Кластеризация используется для изучения данных генома в биологии.
- Идентификация районов аналогичного земледелия в базе данных возможно при помощи кластеризации
- Кластеризация помогает определить группы владельцев страховых полисов с высокой средней стоимостью заявки
- Кластеризация способна подразделить группы домов в соответствии с их типом, стоимостью и географическим местоположение

- Наблюдаемые эпицентры землетрясения должны быть кластеризованы вдоль разломов континентов

Организация данных в разумные группировки - один из самых фундаментальных способов понимания и обучения. В качестве примера можно привести общую схему научной классификации, которая ставит организмы в систему ранговых таксонов: домен, королевство, тип, класс и т.д. Кластерный анализ - это формальное исследование методов и алгоритмов группировки или кластеризации объектов в соответствии с измеренными или воспринимаемыми внутренними характеристиками или сходством. В кластерном анализе не используются метки категорий, которые тегируют объекты с предшествующими идентификаторами, т. е. метки классов. Отсутствие информации о категории отличает кластеризацию данных (неконтролируемое обучение) от классификации или дискриминантного анализа (контролируемое обучение). Целью кластеризации является поиск структуры данных и, следовательно, она обладает исследовательским характером. Кластеризация имеет долгую и богатую историю в различных научных областях. Один из самых популярных и простых алгоритмов кластеризации, K-средних, был впервые опубликован в 1955 году. Несмотря на то, что K- средних был предложен более 50 лет назад, и с тех пор изданы тысячи алгоритмов кластеризации, он до сих пор актуален и широко применяется. Это говорит о сложности разработки алгоритма кластеризации общего назначения и некорректной постановки задачи кластеризации. Кластеризация идентифицирует группы родственных записей, которые могут использоваться в качестве отправной точки для изучения дальнейших отношений. Этот метод поддерживает сегментационные модели. Дополнительные анализы с использованием стандартных аналитических и прочих методов интеллектуального анализа данных могут определять характеристики этих сегментов в отношении какого-либо желаемого результата. Например, покупательские привычки нескольких сегментов

населения можно сравнить, чтобы определить, какие именно сегменты будут нацелены на новую рекламную кампанию.

Кластерный анализ делит данные на группы (кластеры), которые имеют смысл,или же полезны, иногда соблюдаются оба этих пункта. Если целью является многозначимая группировка, то кластеры должны фиксировать естественную структуру данных. Однако в некоторых случаях кластерный анализ является лишь полезной отправной точкой для других целей, таких как обобщение данных. Кластерный анализ, как инструмент для понимания или выявления полезности, уже давно играет важную роль в самых разных областях: психологии и других социальных науках, биологии, статистике, распознавании образов, поиске информации, машинном обучении и интеллектуальном анализе данных. Можно выделить огромное количество применений кластерного анализа для практических задач. Ниже будут приведены некоторые конкретные примеры, организованные тем, является ли целью кластеризации понимание или полезность.

Кластеризация для понимания классов или концептуально значимых групп объектов, которые имеют общие характеристики, играет важную роль в том, как люди анализируют и описывают мир. Действительно, люди умеют делить объекты на группы (кластеризация) и назначать конкретные объекты этим группам (классификация). Например, даже относительно маленькие дети могут быстро маркировать объекты на фотографии в виде зданий, транспортных средств, людей, животных, растений и т. д. В контексте понимания данных кластеры представляют собой потенциальные классы, а кластерный анализ - это изучение методов для автоматически определяющие классы. Ниже приведены некоторые примеры:

Биология. Биологи потратили много лет, создавая таксономию (иерархическую классификацию) всех живых существ: царство, тип, класс, порядок, семью, род и виды. Таким образом, не удивительно, что большая часть ранней работы по кластерному анализу стремилась создать дисциплину математической таксономии, которая могла бы автоматически находить

такие структуры классификации. В последнее время биологи применяют кластеризацию для анализа большого количества генетической информации, которая теперь доступна. Например, кластеризация была использована для определения групп генов, которые имеют сходные функции.

Поиск информации. Всемирная паутина состоит из миллиардов веб-страниц, а результаты запроса к поисковой системе могут возвращать тысячи страниц. Кластеризация может использоваться для группировки этих результатов поиска в небольшое количество кластеров, каждый из которых фиксирует конкретный аспект запроса. Например, запрос «фильм» может возвращать веб-страницы, сгруппированные по категориям, таким как обзоры, трейлеры, артисты и т.д. Каждая категория (кластер) может быть разбита на подкатегории (подкластеры), создавая иерархическую структуру, которая в дальнейшем помогает поиску результатов запроса пользователем.

Климат. Понимание климата Земли требует определения закономерностей в атмосфере и океане. С этой целью кластерный анализ был применен для определения закономерностей атмосферного давления полярных регионов и районов океана, которые оказывают значительное влияние на земельный климат.

Психология и медицина. Болезнь или общее состояние здоровья часто имеют ряд вариаций, и кластерный анализ может быть использован для идентификации этих различных подкатегорий. Например, кластеризация была использована для определения различных типов депрессии. Кластерный анализ также может быть использован для обнаружения закономерностей пространственного или временного распределения болезни.

Бизнес. Предприятия собирают большое количество информации о текущих и потенциальных клиентах. Кластеризация может использоваться для сегментации клиентов в небольшое количество групп для дополнительного анализа и маркетинговой деятельности.Кластеризация для анализа полезности кластеров обеспечивает абстрагирование от отдельных объектов данных до кластеров, в которых находятся эти объекты данных.

Кроме того, некоторые методы кластеризации характеризуют каждый кластер в терминах прототипа кластера; то есть объект данных, который является репрезентативным для других объектов в кластере. Эти прототипы кластера могут использоваться в качестве основы для ряда методов анализа данных или обработки данных. Поэтому в контексте полезности кластерный анализ представляет собой изучение методов определения наиболее представительных кластерных прототипов.

Суммирование. Многие методы анализа данных, такие как регрессия, имеют сложность времени или пространства O(m2) или выше (где m - количество объектов) и, следовательно, не подходят для больших наборов данных. Однако вместо применения алгоритма ко всему набору данных его можно применить к сокращенному набору данных, состоящему только из кластерных прототипов. В зависимости от типа анализа, количества прототипов и точности, с которой прототипы представляют данные, результаты могут быть сопоставимы с результатами, которые были бы получены, если бы все данные могли быть использованы.

Сжатие. Кластерные прототипы также могут использоваться для сжатия данных. В частности, создается таблица, которая состоит из прототипов для каждого кластера; то есть каждому прототипу присваивается целочисленное значение, которое является его позицией (индексом) в таблице. Каждый объект представлен индексом прототипа, связанного с его кластером. Этот тип сжатия известен как векторное квантование и часто применяется к изображениям, звуковым и видеоданным, где многие объекты данных очень похожи друг на друга, допустима некоторая потеря информации и требуется существенное уменьшение размера данных.

Удобный поиск ближайших соседей. Поиск ближайших соседей может потребовать вычисления попарного расстояния между всеми точками. Часто кластерами и их кластерными прототипами можно получить результат гораздо эффективнее. Если объекты относительно близки к прототипу их кластера, то мы можем использовать прототипы для уменьшения количества

вычислений расстояний, необходимых для нахождения ближайших соседей объекта. Интуитивно, если два прототипа кластера находятся далеко друг от друга, то объекты в соответствующих кластерах не могут быть ближайшими соседями друг от друга. Следовательно, чтобы найти ближайших соседей объекта, нужно только вычислить расстояние до объектов в соседних кластерах, где близость двух кластеров измеряется расстоянием между их прототипами.

ПРЕДЛОЖЕНИЯ И РЕКОМЕНДАЦИИ

Организация данных в логичные группировки естественно возникает во многих научных областях. Поэтому неудивительно видеть постоянную популярность кластеризации данных. Важно помнить, что кластерный анализ является исследовательским инструментом, в связи с чем, вывод алгоритмов кластеризации предполагает только гипотезы. В то время как многочисленные алгоритмы кластеризации были опубликованы, а новые продолжают появляться, нет единого алгоритма кластеризации, который, как было показано, превзошел бы другие алгоритмы во всех аспектах. Большинство алгоритмов, включая простые K-средних, являются допустимыми алгоритмами. С появлением новых инструментов становится все более очевидным, что задача поиска наилучшего принципа кластеризации действительно может оказаться бесполезной.Использование кластеризации при анализе данных повлекло за собой многочисленные истории успеха. Несмотря на это, сообществам по машинному обучению и распознаванию образов необходимо решить ряд вышеперечисленных проблем, чтобы улучшить общее понимание кластеризации данных. Ниже приведен список проблем и направлений исследований, на которых стоит сосредоточить внимание :

- Необходим набор базовых данных (с обоснованной истиной), доступных исследовательскому сообществу для тестирования и оценки методов кластеризации. Тест должен включать в себя наборы данных разных типов (документы, изображения, временные ряды, транзакции клиентов, биологические последовательности, социальные сети и т. Д.). Они также должны включать в себя как статические, так и динамические данные (последние были бы полезны при анализе кластеров, меняющихся со временем), количественные и / или качественные атрибуты, связанные и не связанные объекты и т. Д. Хотя идея предоставления контрольных данных не нова (например, UCI ML

и KDD-репозиторий), текущие тесты ограничены небольшими статическими наборами данных.

- Необходимо добиться более тесной интеграции между алгоритмами кластеризации и потребностями приложения. Например, некоторым приложениям может потребоваться генерация только нескольких когезионных кластеров, в то время как другие могут потребовать более четких раздел всех данных. В большинстве приложений это не обязательно лучший алгоритм кластеризации, который действительно имеет значение. Скорее, более важно выбрать правильный метод извлечения объектов, который идентифицирует базовую структуру кластеризации данных.
- Независимо от принципа (или цели), большинство методов кластеризации в конечном итоге сводятся к задачам комбинаторной оптимизации, которые направлены на поиск разбиения данных, которые оптимизируют цель. В результате вычислительная проблема становится критической, если приложение включает крупномасштабные данные. Например, найти глобальное оптимальное решение для K-средних N-сложно. Следовательно, важно выбрать принципы кластеризации, которые приводят к эффективным вычислениям.
- Основной проблемой, связанной с кластеризацией, является ее стабильность или согласованность. Хороший принцип кластеризации должен привести к разделению данных, которое является устойчивым по отношению к искажениям данных. Необходимо разработать методы кластеризации, которые приводят к устойчивым решениям.
- Учитывая присущую сложность кластеризации, имеет смысл развивать методы полуквалифицированной кластеризации, в которых маркированные данные и парные ограничения могут использоваться для определения как представления данных, так и соответствующей целевой функции для кластеризации данных.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Котов А., Красильников Н. Кластеризация данных. 2006.
2. Нейский И.М. Характеристика технологий и процессов интеллектуального Анализа данных
3. “A Tutorial on Spectral Clustering” Ulrike von Luxburg, 2007
4. Berkhin, P. Survey of Clustering Data Mining Techniques / P. Berkhin. - USA: Accrue Software, 2002
5. Bradley, P., Fayyad, U., Reina, C. Scaling Clustering Algorithms to Large Databases, Proc. 4th Int'l Conf. Knowledge Discovery and Data Mining, AAAI Press, Menlo Park, Calif., 1998.
6. Cluster Analysis: Basic Concepts and algorithms. https://www-users.cs.umn.edu/~kumar001/dmbook/ch8.pdf
7. Ester, Martin, Alexander Frommelt, Hans-Peter Kriegel, and Jörg Sander, “Spatial Data Mining: Database Primitives, Algorithms and Efficient DBMS Support”, Data Mining and Knowledge Discovery, Vol. 4, Pp. 193-216, 2000.Dibbell, Julian (January 2003). «The Unreal Estate Boom». Wired (11.01).
8. Hemlata Sahu, Shalini Shrma, Seema Gondhalakar A Brief Overview on Data Mining Survey, International Journal of Computer Technology and Electronics Engineering (IJCTEE) Volume 1, Issue 3
9. Mr.Subu Surendran, Neethu C V Review of Spatial Clustering Methods, International Journal of Information Technology Infrastructure , 2(3), May – June 20
10. P. IndiraPriya, Dr. D.K. Ghosh A Survey on Different Clustering Algorithms in Data Mining Technique, International Journal of Modern Engineering Research (IJMER) www.ijmer.com Vol.3, Issue.1, Jan-Feb. 2013 pp-267-274

11. Prof. M. A. Deshmukh, Prof. R. A. Gulhane Importance of Clustering in Data Mining, International Journal of Scientific & Engineering Research, Volume 7, Issue 2, February-2016
12. “Web Scale K-Means clustering” D. Sculley, Proceedings of the 19th international conference on World wide web(2010)
13. Информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных. Режим доступа: - www.machinelearning.ru.
14. Информационный ресурс, курс онлайн лекций. Режим доступа:- https://www.intuit.ru/studies/courses/6/6/lecture
15. http://scikit-learn.org/stable/modules/clustering.html#spectral-clustering
16. http://www.zentut.com/data-mining/data-mining-architecture/
17. https://www-users.cs.umn.edu/~kumar001/dmbook/slides/chap7_basic_cluster_analysis.pdf
18. http://www.ims.uni-stuttgart.de/institut/mitarbeiter/schulte/theses/phd/algorithm.pdf
19. https://openlayers.org/en/latest/doc/
20. https://images-na.ssl-images-amazon.com/images/G/01/books/stech-ems/DataMining-ch-9780123814791._V155175544_.pdf
21. https://gomap.az/

XÜLASƏ

İşdə Data Mining texnologiyasının tədqiqi üçün alət olaraq Klaster analizinin mahiyyəti açıqlanmışdır. Verilənlərin intelektual analizinin əsas mərhələləri ayrılmışdır. Data Mining arxitekturası göstərilmiş, əsas üsullar verilmişdir. Verilənlərin intelektual analizinin əsas məsələlərinə diqqət yönəldilmişdir.

Data Mining texnologiyasının əsas alətlərinin qiymətləndirilməsi aparılmışdır. Data Mining aləti olaraq Klaster analizindən istifadənin əsas halları formalaşdırılmışdır. Doğru klasterizasiya alqoritminin qurulmasına imkan verən uyğunluq ölçüsünə baxılmışdır. Klasterizasiyanın əsas üsullarının analizi və formalaşdırılması aparılmış, onların klassifikasiyası göstərilmişdir. Müxtəlif növ klasterlər və onların müqayisəsi göstərilmişdir. Geofəza(geoməkan) verilənlərinin analizi aparılmış, klasterizasiyadan sonra verilənlərin növbəti analizinin mümkünlüyü göstərilmişdir. Klaster analizi prosesi zamanı meydana gələn problemlər göstərilmişdir. Klasterizasiyanın tətbiq sahəsinə baxılmışdır. Uğurlu nəticə əldə etmək üçün klasterlərin tətbiqinin səbəbi və onlara olan tələb aşkarlanmışdır.

SUMMARY

The work reveals the essence of cluster analysis as a tool for researching Data Mining technologies. The main stages of data mining are identified. The architecture of Data Mining is shown, the main methods are stated. Attention was given to the main tasks of Data Mining. The main tools of Data Mining technologies were evaluated. The main points of using cluster analysis as a tool of Data Mining are formulated. Similarity measures, which make it possible to construct the correct clustering algorithm, are considered. The formulation and analysis of the main methods of clustering is carried out, their classification is shown. Different types of clusters are shown and compared. An analysis of geospatial data has been done. The possibility of further data analysis after clustering is shown. The problems arising in the process of cluster analysis are indicated. Areas of application of clustering are considered. The reasons of application and requirements to clusters for successful result are revealed.