МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение

высшего образования

**«КУБАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»**

**(ФГБОУ ВО «КубГУ»)**

**Кафедра прикладной математики**

**КУРСОВАЯ РАБОТА**

**МАТЕМАТИЧЕСКИЕ МОДЕЛИ КЛАСТЕРНОГО АНАЛИЗА**

Работу выполнил \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_В.В.Меликян

(подпись, дата) (инициалы,фамилия)

Факультет компьютерных технологий и прикладной математики курс 3

Направление 02.03.03. Математическое обеспечение и администрирование информационных систем

Научный руководитель, д.т.н.

профессор \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_А.А.Халафян

(подпись, дата) (инициалы, фамилия)

Нормоконтролер,

к.ф.м.н.\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Г.В. Калайдина

(подпись, дата) (инициалы, фамилия)

Краснодар 2017

СОДЕРЖАНИЕ

Введение 3

1 Метод кластеризации 4

2 Классификация задач кластеризации 6

2.1 Меры расстояний 6

2.2 Основные цели кластерного анализа 8

2.3 Формальная постановка задачи кластерного анализа 9

3 Методы анализа данных 11

3.1 Алгоритм k-means(k-средних) 11

3.2 Алгоритм иерархической кластеризации 17

Заключение 24

Список использованных источников 26

ВВЕДЕНИЕ

Настоящая работа посвящена изучению методов и алгоритмов кластеризации, широко используемых в системах интеллектуального анализа данных.

Интеллектуальный анализ данных связан с поиском во множестве информации скрытых нетривиальных закономерностей, позволяющих получить новые знания об исследуемых объектах.

Особенный интерес к методам анализа возник в связи с развитием средств сбора и хранения данных, позволившим накапливать большие объемы информации. Известные статистические методы покрывают лишь часть нужд по обработке собираемых данных, и для их использования необходимо иметь четкое представление об искомых закономерностях.

В такой ситуации методы интеллектуального анализа данных приобретают особую актуальность. Их основная особенность заключается в установлении наличия и характера скрытых закономерностей, тогда как традиционные методы занимаются главным образом параметрической оценкой уже установленных правил.

Среди методов интеллектуального анализа данных особое место занимает кластеризация. Кластеризация, основываясь на установленном отношении схожести элементов, устанавливает подмножества (кластеры), в которые группируются входные данные.

Цель: изучить возможности практического применения обобщенных методов кластерного анализа.

Задачи: на примере конкретных задач изучить основные методы кластерного анализа – для алгоритма k-means – задача про исследование цветков Ириса, для иерархической модели кластеризации – исследование анкетирования сотрудников фирмы с целью определения, каким образом можно наиболее эффективно управлять персоналом.

1 Метод кластеризации

Как и многие понятия в науке, кластеризация имеет немало различных определений и толкований. Многие специалисты дают ему такое определение: "Кластерный анализ (Data clustering) – задача разбиения заданного множества объектов (ситуаций) на непересекающиеся подмножества, называемые кластерами, так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались".

Применение кластерного анализа в общем виде сводится к следующим этапам:

* отбор выборки объектов для кластеризации;
* определение множества переменных, по которым будут оцениваться объекты в выборке. При необходимости – нормализация значений переменных;
* вычисление значений меры сходства между объектами;
* применение метода кластерного анализа для создания групп сходных объектов (кластеров);
* представление результатов анализа.

Где применяется кластеризация? В маркетинге – это сегментация конкурентов и потребителей. В менеджменте – разбиение персонала на различные группы по уровню мотивации, классификация поставщиков, выявление схожих производственных ситуаций, при которых возникает брак. В медицине – классификация симптомов, пациентов, препаратов. В социологии – разбиение респондентов на однородные группы. Нетрудно заметить, что кластерный анализ хорошо зарекомендовал себя во всех сферах жизнедеятельности человека.

Следует отметить, что в результате применения различных методов кластерного анализа могут быть получены кластеры различной формы. Например, возможны кластеры "цепочного" типа, когда кластеры представлены длинными "цепочками", кластеры удлиненной формы и т.д., а некоторые методы могут создавать кластеры произвольной формы.

Различные методы могут стремиться создавать кластеры определенных размеров (например, малых или крупных) либо предполагать в наборе данных наличие кластеров различного размера.

В результате применения различных методов кластеризации могут быть получены неодинаковые результаты, это нормально и является особенностью работы того или иного алгоритма. Такие особенности следует учитывать при выборе метода кластеризации.

Имеются различные методы кластеризации для решения поставленной задачи. Сложность заключается в отсутствии на момент начала анализа какой-либо дополнительной информации о данных. В связи с этим возможное множество решений по мощности сопоставимо с входным множеством, что на практике может вызвать трудности.

Решение поставленныхзадач методом кластеризации получило широкое распространение, поэтому на данный момент существует более 100 различных алгоритмов кластеризации, однако, в данной работе рассмотрены два наиболее часто используемых метода – иерархический кластерный анализ и кластеризация методом k-средних.

2 Классификация задач кластеризации

Кластерный анализ выполняет следующие основные задачи:

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

Независимо от предмета изучения применение кластерного анализа предполагает следующие этапы:

* отбор выборки для кластеризации. Подразумевается, что имеет смысл кластеризовать только количественные данные;
* определение множества переменных, по которым будут оцениваться объекты в выборке, то есть признакового пространства;
* вычисление значений той или иной меры сходства (или различия) между объектами;
* применение метода кластерного анализа для создания групп сходных объектов;
* проверка достоверности результатов кластерного решения.

После получения и анализа результатов возможна корректировка выбранной метрики и метода кластеризации до получения оптимального результата.

2.1 Меры расстояний

Как же определять «похожесть» объектов? Для начала нужно составить вектор характеристик для каждого объекта – как правило, это набор числовых значений, например, рост или вес человека. Однако существуют также алгоритмы, работающие с качественными (категорийными) характеристиками.

После того, как мы определили вектор характеристик, можно провести нормализацию, чтобы все компоненты давали одинаковый вклад при расчете «расстояния». В процессе нормализации все значения приводятся к некоторому диапазону, например, [-1,-1] или [0,1].

Наконец, для каждой пары объектов измеряется «расстояние» между ними – степень похожести. Существует множество метрик, вот лишь основные из них:

Евклидово расстояние. Наиболее распространенная функция расстояния. Представляет собой геометрическим расстоянием в многомерном пространстве:

Квадрат евклидова расстояния. Применяется для придания большего веса более отдаленным друг от друга объектам. Это расстояние вычисляется следующим образом:

Расстояние городских кварталов (манхэттенское расстояние). Это расстояние является средним разностей по координатам. В большинстве случаев эта мера расстояния приводит к таким же результатам, как и для обычного расстояния Евклида. Однако для этой меры влияние отдельных больших разностей (выбросов) уменьшается (т.к. они не возводятся в квадрат). Формула для расчета манхэттенского расстояния:

Расстояние Чебышева – это расстояние может оказаться полезным, когда нужно определить два объекта как «различные», если они различаются по какой-либо одной координате. Расстояние Чебышева вычисляется по формуле:

Степенное расстояние. Применяется в случае, когда необходимо увеличить или уменьшить вес, относящийся к размерности, для которой соответствующие объекты сильно отличаются. Степенное расстояние вычисляется по следующей формуле:

,

где r и p – параметры, определяемые пользователем. Параметр p ответственен за постепенное взвешивание разностей по отдельным координатам, параметр r ответственен за прогрессивное взвешивание больших расстояний между объектами. Если оба параметра – r и p — равны двум, то это расстояние совпадает с расстоянием Евклида.

Выбор метрики полностью лежит на исследователе, поскольку результаты кластеризации могут существенно отличаться при использовании разных мер.

2.2 Основные цели кластерного анализа:

1) Понимание данных путём выявления кластерной структуры. Разбиение выборки на группы схожих объектов позволяет упростить дальнейшую обработку данных и принятия решений, применяя к каждому кластеру свой метод анализа.

2) Сжатие данных. Если исходная выборка избыточно большая, то можно сократить её, оставив по одному наиболее типичному представителю от каждого кластера.

3) Обнаружение новизны (англ. Novelty detection). Выделяются нетипичные объекты, которые не удаётся присоединить ни к одному из кластеров.

В первом случае число кластеров стараются сделать поменьше. Во втором случае важнее обеспечить высокую степень сходства объектов внутри каждого кластера, а кластеров может быть сколько угодно. В третьем случае наибольший интерес представляют отдельные объекты, не вписывающиеся ни в один из кластеров.

Во всех этих случаях может применяться иерархическая кластеризация, когда крупные кластеры дробятся на более мелкие, т.е. в свою очередь дробятся ещё мельче, и т. д. Такие задачи называются задачами таксономии. Результатом таксономии является древообразная иерархическая структура. При этом каждый объект характеризуется перечислением всех кластеров, которым он принадлежит, обычно от крупного к мелкому.

2.3 Формальная постановка задачи кластерного анализа.

Введем формальную постановку задачи кластерного анализа:

Пусть  – множество объектов,  – множество номеров (имён, меток) кластеров. Задана функция расстояния между объектами  . Имеется конечная обучающая выборка объектов  .



Требуется разбить выборку на непересекающиеся подмножества, называемые кластерами, так, чтобы каждый кластер состоял из объектов, близких по метрике  , а объекты разных кластеров существенно отличались. При этом каждому объекту  приписывается номер кластера  .



Алгоритм кластеризации – это функция  , которая любому объекту  ставит в соответствие номер кластера  . Множество  в некоторых случаях известно заранее, однако чаще ставится задача определить оптимальное число кластеров, с точки зрения того или иного критерия качества кластеризации.



Решение задачи кластеризации принципиально неоднозначно, и тому есть несколько причин:

1) Не существует лучшего критерия качества кластеризации. Известен целый ряд эвристических критериев, а также ряд алгоритмов, не имеющих чётко выраженного критерия, но осуществляющих достаточно разумную кластеризацию «по построению». Все они могут давать разные результаты. Следовательно, для определения качества кластеризации требуется эксперт предметной области, который бы мог оценить осмысленность выделения кластеров.

2) Число кластеров, как правило, неизвестно заранее и устанавливается в соответствии с некоторым субъективным критерием. Это справедливо только для методов дискриминации, так как в методах кластеризации выделение кластеров идёт за счёт формализованного подхода на основе мер близости.

3) Результат кластеризации существенно зависит от метрики, выбор которой, как правило, также субъективен и определяется экспертом. Но стоит отметить, что есть ряд рекомендаций к выбору мер близости для различных задач.

3 Методы анализа данных

3.1 Алгоритм k-means (k-средних)

Алгоритм k-means (k-средних) - доступный и широко используемый алгоритм кластеризации. При данном множестве объектов (записей) основная задача кластеризации состоит в том, чтобы разделить эти объекты на группы или «кластеры» такие, что объекты внутри группы имеют тенденцию быть более похожими друг на друга по сравнению с объектами, принадлежащими различным группам. Другими словами, алгоритмы объединения в кластеры размещают подобные точки в тот же кластер, в то время как непохожие точки помещают в различные кластеры.

В отличие от контролируемых задач, таких как регресс или классификация, где есть понятие целевого значения или метка класса, объекты, которые образуют входы в процедуру кластеризации, не идут со связанной целью. Следовательно, кластеризация часто упоминается как бесконтрольное обучение. Поскольку нет никакой необходимости в маркированных данных, алгоритмы обучения без учителя являются подходящими для многих приложений, где маркированные данные трудно получить. Бесконтрольные задачи, такие как кластеризация, так же часто используются, чтобы исследовать и охарактеризовать набор данных перед запуском контролируемой задачи обучения. Поскольку кластеризация производится, не используя метки класса, некоторое представление о сходстве должно быть определено на основании атрибутов объектов.

Описание сходства и метода, в котором точки сгруппированы, отличаются на основании применяемого алгоритма кластеризации. Таким образом, различные алгоритмы кластеризации подходят для различных типов набора данных и различных целей.  Выбор «лучшего» алгоритма кластеризации, следовательно, зависит от приложения. Это не редкость, использовать несколько различных алгоритмов и выбирать в зависимости от того, какой является более полезным.

Алгоритм k-means является простым повторяющимся алгоритмом кластеризации, который разделяет определенный набор данных на заданное пользователем число кластеров k. Алгоритм прост для реализации и запуска, относительно быстрый, легко адаптируется и распространен на практике. Это исторически один из самых важных алгоритмов интеллектуального анализа данных.

Алгоритм k-means применяется к объектам, которые представляются точками в d-мерном векторном пространстве. Таким образом, это кластеры набора d-мерных векторов, D = {**xi**|i= 1,..., N}, где **xi**∈ Rd обозначает i-ый объект или «точку данных». Как уже говорилось ранее, k-means является алгоритмом кластеризации, который разделяет D на k кластеров точек. То есть, алгоритм k-means объединяет все точки данных в D так, что каждая точка **xi**попадает в один и только один из k разделов. Можно отследить, какая точка находится в каком кластере, назначив каждой точке номер кластера. Точки с таким же номером кластера находятся в одном и том же кластере, в то время как точки с различными номерами кластера находятся в разных кластерах.  Это можно обозначить как кластерный составной вектор m длинной N, где mi является номером кластера xi.

Значение k является основным из входных данных алгоритма. Как правило, значение k основано на критериях, таких как предварительное знание о том, сколько на самом деле кластеров появится в D, как много кластеров требуется для текущего приложения, или типы кластеров, найденные путем изучения/экспериментирования   с различными значениями k. Неважно, каким образом выбран k, для того чтобы понять, как k-means разделяет набор данных D, и мы обсудим, как выбирать k, когда он не задан, в следующем разделе.

В k-means, каждый из k кластеров   представлен одной точкой в Rd. Обозначим этот набор представителей кластера как множество С = {сj|j=1,.,k}. Эти представители k кластеров также называются состоянием кластера или центройды кластера.

В алгоритмах кластеризации точки группируются некоторым понятием «близости» или «подобия». В k-means мера близости по умолчанию Евклидово расстояние. В частности, можно с готовностью показать, что k-means пытается минимизировать следующую неотрицательную функцию стоимости

Сost = (1)

Другими словами, k-means пытается минимизировать итоговый квадрат Евклидова расстояния между каждой точкой xi и ее самым близким представителем кластера cj. Уравнение 1 часто упоминается как целевая функция k-means.

Алгоритм k-means разделяет D итеративным способом, чередующимся между 2 шагами:

1. переприсваивание номера кластера всем точкам в D;

2) обновление представителей кластера, основанных на точках данных в каждом кластере. Алгоритм работает следующим образом: сначала представители кластера инициализируются, выбирая k из Rd.  Методы для выбора этих начальных источников включают случайную выборку из набора данных, устанавливая их как решение кластеризации маленького подмножества данных, или нарушая глобальное среднее значение данных k раз. В Алгоритме мы инициализируем случайно выбранные k точек.  Затем выполняется алгоритм итерации до сходимости за 2 шага.

Шаг 1. Присваивание данных: каждой точке данных присваивается ее самый близкий представитель при том, что связи нарушаются произвольно. Это приводит к разделению данных.

Шаг 2. Перемещение «средних»: каждый представитель кластера перемещается к центру (т. е. среднее арифметическое всех точек данных, присвоенных ему). Объяснение этого шага основано на наблюдении, что данное множество точек, единственный лучший представитель для этого множества (в смысле минимизации суммы квадрата Евклидова расстояния между каждой точкой и представителем) не что иное, как срединная точка данных. Именно поэтому представитель кластера часто взаимозаменяемо называют срединным элементом кластера или центроидом кластера, отсюда и название этого алгоритма.

Алгоритм сходится, когда присвоение (следовательно, и значения Сj) больше не изменяются. Можно показать, что целевая функция k-means, определенная в уравнении 1, будет уменьшаться всякий раз, когда есть изменения в присвоении или шагах измерения, и сближение (сходимость в одной точке) гарантировано за конечное число итераций.

Как упомянуто, выбор оптимального значения k может быть трудным. Если мы имеем какую-то информацию о наборе данных, например, число разделов, которые естественно включают набор данных, тогда эти значения могут быть использованы при выборе k. Иначе, необходимо использовать другие критерии выбора k, таким образом, решая проблему выбора модели. Самое простое решение попробовать несколько различных значений k и выбрать кластеризацию, которая минимизирует целевую функцию k-means.

К сожалению, значение целевой функции не столь же информативно, как можно было бы надеяться в этом случае. Для примера, стоимость оптимального решения сокращается с увеличением k до тех пор, пока она не достигнет 0, когда число кластеров равняется числу отличных точек данных. Это делает его более трудным для использования объективной функции (а) непосредственно сравнивая решения с различным числом кластеров и (в) найти оптимальное значение k. Таким образом, если требуемое k не будет известно заранее, как правило, k-means будет запущен с различными значениями k, а затем использовать другие, более подходящие критерии для выбора одного из результатов. В качестве альтернативы, можно прогрессивно увеличивать число кластеров в соединении с подходящим критерием остановки. Рассмотрим этапы реализации данного метода:

Этап 1. Первоначальное распределение объектов по кластерам

1) Выбор случайным образом k точек данных из D как начальное множество представителей кластера C

2) Распределение объектов по кластерам.

Этап 2. Перераспределение срединных элементов

1) Вычисление центра для каждого кластера

2) Перераспределение объектов по кластерам

Из алгоритма видно, что каждая итерация нуждается в N\*k сравнений, которые определяют временную сложность одной итерации. Число итераций, требуемых для сходимости, изменяется и может зависеть от N. Соответственно, чем больше точек во множестве (N), тем дольше будет работать алгоритм.

Сократить время работы алгоритма можно путем распараллеливания этапа распределения точек по кластерам. Если у нас имеется Р процессоров, то мы можем создать Р потоков. В этом случае исходное множество данных можно разбить на Р частей и каждый поток будет работать только со своим объемом информации, независимо от остальных данных.

Потенциальная проблема алгоритма – проблема «пустых» кластеров. При запуске k-means, особенно с большим значением k и/или, когда данные находятся в большом размерном пространстве, возможно, что в какой-то момент исполнения, существует представитель кластера cj, такой, что все точки xi в D ближе к некоторому другому представителю кластера, который не является cj. Когда точки в D будут присвоены к ближайшему кластеру, j-му кластеру будут присвоены нулевые точки. Таким образом, кластер j является теперь пустым кластером. Стандартный алгоритм не принимает меры против пустых кластеров, но простые решения (такие как переинициализация представителя пустого кластера или «кража» некоторых точек из большого кластера) возможны.

Рассмотрим пример работы k-means на классическом тестовом множестве данных. В данном примере исследуются множество данных Ирис, который состоит из 150 точек данных из 3 классов. Каждый класс представляет различную разновидность цветков Ириса, и имеем по 50 точек в каждом классе. Несмотря на то, что есть 4 размерности (представляющие ширину чашелистика, длину чашелистики, ширину лепестков и длину лепестков), только 2 величины (ширина лепестка и длина лепестка) необходимы, чтобы отличать эти 3 класса. Набор данных Ирис изображен на рисунке 1 вдоль размерности ширины и длины лепестка.

На рисунке 1 изображено множество данных Ирис. На рисунке 2 показан пример работы алгоритма k-means на выборе данных Ирис с входными данными, при которых k=3, используя только атрибуты длины и ширины лепестка. Алгоритм k-means в состоянии разделить точки данных так, что каждый кластер состоит в основном из цветов одного вида.

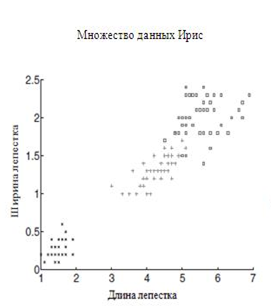


Рисунок 1 – Множество данных Ирис

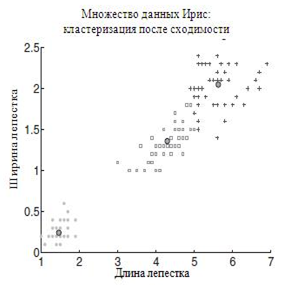


Рисунок 2 – Множество данных после кластеризации

3.2 Алгоритмы иерархической кластеризации

Иерархическая кластеризация – это метод или процедура для преобразования матрицы близости в последовательность вложенных разбиений, множество п объектов, подлежащих кластеризации, обозначается

A = {x1,x2,…,xn}, где xi— i-й объект.Разбиение S делит A на подмножества {S1 S2,..., Sm}, удовлетворяющие следующим условиям:

Si ∩Sj ≠0 для i и j от 1 до m, I ≠ j

S1 ∩ S2 ∩…∩ SM = A

Кластеризация – это разбиение. Компоненты разбиения называются кластерами.Разбиение В вкладывается (или является измельчением) в раз­биение С, если каждая компонента В есть собственное подмножество ком­поненты С. Другими словами, С образуется слиянием компонент В. Напри­мер, если кластеризация С делит выборку на три кластера, а кластеризация В делит выборку на пять кластеров, как это показано ниже, то В вкладывается в С. Как С, так и В кластеризуют множество объектов

{x1,x2…,x10}:

C= {(x1,x3,x5,x7),(x2,x4,x6,x8),(x9,x10)},

B = {(x1,x3),(x5,x7),(x2),(x4,x6,x8),(x9,x10)}.

Иерархическая кластеризация – это последовательность разбиений, в которой каждое разбиение вкладывается в следующее разбиение в последовательности.

Типичный агломеративный алгоритмиерархической кластеризации начинает работу с формирования несвязных кластеров, помещая каждый из n объектов в отдельный кластер. Используемый алгоритм кластеризации оп­ределяет, как следует интерпретировать матрицу близости, чтобы объеди­нить два или более таких тривиальных кластера в кластер следующего уров­ня. Процесс повторяется, чтобы сформировать последовательность вложен­ных кластеров, в которых число кластеров уменьшается, пока не будет сформирован один кластер, объединяющий все объекты.

Дивизимный алгоритм производит все действия в обратном порядке.

Дендрограмма– специальный вид структуры дерева – обеспечивает удобную форму представления иерархической кластеризации. Она состоит из слоев вершин, каждый из которых представляет кластер. Линии, соеди­няющие вершины, представляют кластеры, которые вложены один в дру­гой. Горизонтальный срез дендрограммы образует кластеризацию.

Два наиболее простых метода иерархической кластеризации получили наибольшее распространение - метод минимального расстояния и метод максимального расстояния. В обоих случаях исходные данные представлены в виде матрицы D = [d(ij)]. Мы считаем, что элементы матрицы соответствуют различию (расстоянию) между объектами, т.е. если d(l, 2) > d(l, 3) – это означает, что объекты 1 и 3 более близки, чем объекты 1 и 2.

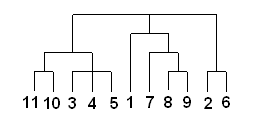


Рисунок 3 – Пример дендрограммы

Таблица 1 – Пример матрицы близости

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | X1 | X2 | X3 | X4 | X5 |
| X1 | 0 | 6 | 8 | 2 | 7 |
| X2 | 6 | 0 | 1 | 5 | 3 |
| X3 | 8 | 1 | 0 | 10 | 9 |
| X4 | 2 | 5 | 10 | 0 | 4 |
| X5 | 7 | 3 | 9 | 4 | 0 |

Различие между методами минимальной и максимальной связи можно пояснить следующим образом. Предположим, что кластеризация объек­тов {х1 х2,..., х} была получена слиянием кластеров Сm1 и Cm2 в кластеризации {Cm1,Cm2,…,Cm(n-m)}, гдеn– номер шага кластеризации,m– номер кластера.

При кластеризации по методу минимального расстояния мы объединяем кластеры, следуя правилу, при котором для каждой пары кластеров расстояние определяется по ближайшим точкам и выбирается пара кластеров, для которых это расстояние минимально. При кластеризации по методу максимальной связи правило меняется следующим образом – для каждой пары кластеров расстояние определяется по ближайшим точкам и выбирается пара кластеров, для которых это расстояние максимально.

Рассмотрим простейший алгоритм агломеративной кластеризации, взяв в качестве примера матрицу, приведенную на рисунке 3.

Агломеративный алгоритм по методу минимального расстояния.

Шаг 1. Начнем с несвязной кластеризации, вытекающей из порогового графа G(0), которая не содержит ребер и которая помещает каждый объект в свой кластер, и рассмотрим эту ситуацию как текущую кластеризацию. Положим к ← 1. Заметим, что число связанных компонент графа G(k) будет убывать с ростом к.

Шаг 2. Сформировать G(k). Если число компонент (максимально связан­ных подграфов) в G(k) меньше, чем число кластеров в текущей кластериза­ции (что естественно выполняется), переопределить текущую кластериза­цию, выбрав каждую компоненту G(k) в качестве кластера.

Шаг 3. Если G(k) состоит из единственного связного графа - стоп. В противном случае положить и вернуться к шагу 2.

Шаги работы алгоритма по формированию кластеров показаны на рисунке 4. Дендрограмма, иллюстрирующая результат работы алгоритма, пред­ставлена на рисунке 3.

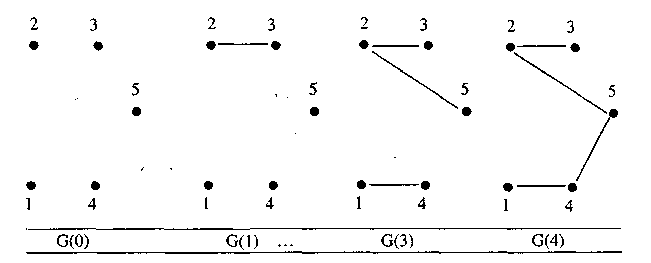
.

Рисунок 4 – Последовательные шаги работы алгоритма по формированию кластеров.

Рассмотрим пример работы алгоритмов ближней и дальней связи на при­мере матрицы D с использованием алгоритма редактирования матрицы рас­стояний.

Таблица 2 – Исходная матрица D

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | X1 | X2 | X3 | X4 | X5 |
| X1 | 0 | 2.3 | 3.4 | 1.2 | 3.7 |
| X2 | 2.3 | 0 | 2.6 | 1.8 | 4.6 |
| X3 | 3.4 | 2.6 | 0 | 4.2 | 0.7 |
| X4 | 1.2 | 1.8 | 4.2 | 0 | 4.4 |
| X5 | 3.7 | 4.6 | 0.7 | 4.4 | 0 |

Пусть дана n x n матрица расстояний D = [d(i,j)]. Кластеризация связывается с последовательностью номеров 0,l,...,(n-l). L(k) – уровень ограниче­ния расстояния, на котором происходит объединение точек в кластер. Кластер, сформированный при номере последовательности m обозначается (т). Расстояние между кластерами (г) и (s) обозначается d[(r),(s)].

Шаг 1. Начинаем с несвязной кластеризации, имеющей уровень L(0) = 0 и номер последовательности m=0.

Шаг 2. Найти наименее различающуюся пару кластеров в текущем раз­биении, например, пару {(r), (s)} в соответствии с критерием

d[(r),(s)] = min {d[(i),(j)]},

где минимум берется по всем парам кластеров в текущей кластеризации.

Шаг 3. Увеличить номер последовательности m m+1. Объединить кластеры (г) и (s) в единый кластер, чтобы сформировать разбиение. Положить уровень слияния кластеров для этого шага кластери­зации.

L(m) = d[(r), (s)]

Шаг 4. Откорректировать матрицу расстояний D, удаляя строки и столб­цы, соответствующие кластерам (г) и (s), и добавляя строки и столбцы, соответствующие вновь сформированным кластерам. Расстояние между новым (r,s) и старым (к) кластерами обозначается для метода минимальной связи

d[(k),(r,s)] = min {d[(k),(r)],d[(k),(s)]}

т.е. расстояние от кластера, полученного слиянием г и s, до ранее существовавшего кластера равно минимальному из расстояний. Для метода максимальной связи

d[(k), (r,s)] = max {d[(k),(r)],d[(k),(s)]}

Шаг 5. Если все объекты включены в один кластер ­­– стоп, в противном случае вернуться к шагу 2.

Рассмотрим конкретную задачу для систематизации знаний по иерархическому методу кластерного анализа. Предположим, компания провела анкетирование своих сотрудников и хочет определить, каким образом можно наиболее эффективно управлять персоналом. То есть планирует разделить сотрудников на группы и для каждой из них выделить наиболее эффективные рычаги управления. При этом различия между группами должны быть очевидными, а внутри группы респонденты должны быть максимально похожи.

Для решения задачи предлагается использовать иерархический кластерный анализ. В конечном результате мы должны определиться на сколько классов (кластеров) компания планирует разбить персонал. Предположим, что мы решили разбить персонал на три группы, тогда для изучения респондентов, попавших в каждый кластер, получим таблицу следующего содержания:

Таблица 3 – Входные данные для иерархической кластеризации

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Кластер | Муж | 30-50 лет | >50 лет | Рук. | Соцпакет | Льготы | з/п | Стаж | Образов. |
| 1 | 80% | 90% | 5% | 70% | 10% | 12% | 95% | 30% | 30% |
| 2 | 40% | 35% | 45% | 13% | 60% | 70% | 60% | 40% | 20% |
| 3 | 50% | 70% | 10% | 5% | 30% | 20% | 70% | 20% | 50% |

Проанализируем, как сформирована приведенная выше таблица 3:

В первом столбце расположен номер кластера – группы, данные по которой отражены в строке. Подробней о каждом из кластеров:

Первый кластер на 80% составляют мужчины. 90% первого кластера попадают в возрастную категорию от 30 до 50 лет, лишь 5% группы старше 50 лет, 70% занимают руководящие должности, только 10% группы нуждается в соцпакете, 12% респондентов считает, что льготы очень важны, размер заработной платы важен для 95% респондентов,30% уверены, что стаж имеет большое значение, 30% готовы заниматься обучение и повышением квалификации.

Второй кластер на 40% составляют мужчины, 35 % второго кластера попадают в возрастную категорию от 30 до 50 лет, 45% группы старше 50 лет, лишь 13% занимают руководящие должности, 60% группы нуждается в соцпакете, 70% респондентов считает, что льготы очень важны, размер заработной платы важен для 60% респондентов, 40% уверены, что стаж имеет большое значение, 20% готовы заниматься обучение и повышением квалификации.

Третий кластер на 50% составляют мужчины, 70 % третьего кластера попадают в возрастную категорию от 30 до 50 лет, 10% группы старше 50 лет, лишь 5% занимают руководящие должности, 30% группы нуждается в соцпакете, 20% респондентов считает, что льготы очень важны, размер заработной платы важен для 70% респондентов, 20% уверены, что стаж имеет большое значение, 50% готовы заниматься обучение и повышением квалификации.  
 Попытаемся составить портреты респондентов каждого кластера:

Первая группа - в основном мужчины зрелого возраста, занимающие руководящие позиции. Социальный пакет (Медицинское обслуживание, льготы и т.д.) их не интересует. Они предпочитают получать хорошую зарплату, а не помощь от работодателя.

Группа два наоборот отдает предпочтение социальному пакету. Состоит она, в основном, из людей "в возрасте", занимающих невысокие посты.   
 Третья группа наиболее "молодая". В отличие от предыдущих двух, очевиден интерес к возможностям обучения и профессионального роста. У этой категории сотрудников есть хороший шанс в скором времени пополнить первую группу. Зарплата для них безусловно важна, но есть и другие приоритеты.  
 Таким образом, планируя кампанию по внедрению эффективных методов управления персоналом, очевидно, что в нашей ситуации можно увеличить соцпакет у второй группы в ущерб, к примеру, зарплате. Если говорить о том, каких специалистов следует направлять на обучение, то можно однозначно рекомендовать обратить внимание на третью группу.

**ЗАКЛЮЧЕНИЕ**

В результате проведенной работы по изучению математических методов кластеризации я сделал следующие выводы:

Алгоритм k-means простой итеративный алгоритм кластеризации, разделяющий множество данных на k кластеров. По своей сути, алгоритм работает с помощью перебора в два этапа:

* кластеризация всех точек данных в зависимости от расстояния между точкой и ее ближайшим представителем кластера.
* переоценка представителей кластера. Ограничения алгоритма k-means включает чувствительность k-means к инициализации и определению значения k.

Несмотря на все недостатки, k-means остается наиболее широко используемым разделяющим алгоритмом кластеризации на практике. Алгоритм простой, понятный и достаточно масштабируемый и может быть легко модифицирован для решения различных задач, таких как частичное обучение с учителем или потоковых данных.

Общая идея методов иерархической кластеризации заключается в последовательной иерархической декомпозиции множества объектов. В зависимости от направления построения иерархии различают дивизимный и агломеративный методы.

В случае агломеративного метода (снизу вверх) процесс декомпозиции начитается с того, что каждый объект представляет собой самостоятельный кластер. Затем на каждой итерации пары близлежащих кластеров последовательно объединяются в общий кластер. Итерации продолжаются до тех пор, пока все объекты не будут объединены в один кластер или пока не выполнится некоторое условие остановки.

Дивизимный метод (сверху вниз) напротив, подразумевает, что на начальном этапе все объекты объединены в единый кластер. На каждой итерации он разделяется на более мелкие до тех пор, пока каждый объект не окажется в отдельном кластере или не будет выполнено условие остановки.

Очевидно, что кластерный анализ является мощным средством разведочного анализа данных и статистических исследований в любой предметной области, но для более точных результатов необходимо применять несколько алгоритмов кластерного анализа и делать выводы на основании общей оценки результатов работы алгоритмов.

Кластерный анализ можно считать успешным, если он выполнен разными способами, проведено сравнение результатов и найдены общие закономерности, а также найдены стабильные кластеры независимо от способа кластеризации. Кластерный анализ позволяет выявить проблемные ситуации и наметить пути их решения.

Следовательно, этот метод непараметрической статистики можно рассматривать как составную часть системного анализа.

# Список использованных источников

1. Халафян А.А. Статистический анализ данных. Краснодар, 2003, 192 с.
2. Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: Классификация и снижение размерности. — М.: Финансы и статистика, 1989, 607 с.
3. Барсегян А.А., Куприянов М.С. Технологии анализа данных. Data Mining, Visual Mining, Text Mining, OLAP. Изд.: БХВ-Петербург, 2007, 384 с.
4. Мандель И. Д*.* Кластерный анализ. — М.: Финансы и статистика, 1988, 176 с.
5. Загоруйко Н. Г. Прикладные методы анализа данных и знаний. — Новосибирск: ИМ СО РАН, 1999, 270 с.
6. Дюран Б., Оделл П*.* Кластерный анализ. — М.: Статистика, 1977, 128 с.