Всероссийское СМИ
«Академия педагогических идей «НОВАЦИЯ»
Свидетельство о регистрации ЭЛ №ФС 77-62011 от 05.06.2015 г.
(выдано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций)

Сайт: akademnova.ru
e-mail: akademnova@mail.ru

Цынко Д.О., Мамаев Х.Ш. Кластеризация. Методы кластеризации // Академия педагогических идей «Новация». Серия: Научный поиск. – 2018. – №10 (декабрь). – АРТ 52-эл. – 0,2 п.л. - URL: http://akademnova.ru/series-scientific-search

УДК 004

Цынко Дарья Олеговна,

Мамаев Хаджимурад Шамильевич

Магистры 2-го курса, отдел магистратуры

Донской государственный технический университет,

г. Ростов-на-Дону, Российская Федерация

e-mail: masyanya.7777@yandex.ru

КЛАСТЕРИЗАЦИЯ. МЕТОДЫ КЛАСТЕРИЗАЦИИ

Аннотация: В статье рассмотрено понятие кластеризации, методы кластеризации, этапы кластеризации.

Ключевые слова: кластеризация, методы, этапы.

Tsynko D.O.,

Mamaev H.Sh,

Masters of the 2nd course

Don State Technical University,

Rostov-on-Don, Russian Federation

Всероссийское СМИ
«Академия педагогических идей «НОВАЦИЯ»
Свидетельство о регистрации ЭЛ №ФС 77-62011 от 05.06.2015 г.
(выдано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций)

Сайт: akademnova.ru
e-mail: akademnova@mail.ru

CLUSTERING. CLUSTERING METHODS

Abstract: The article discusses the concept of clustering, clustering methods, stages of clustering.

Keywords: clustering, methods, stages.

Методы кластеризации широко используются при анализе больших наборов данных для группировки образцов с похожими свойствами. Существует много алгоритмов для выполнения кластеризации, и результаты могут существенно различаться. В частности, количество групп, присутствующих в наборе данных, часто неизвестно, а количество кластеров, идентифицированных алгоритмом, может изменяться в зависимости от используемых параметров. На данный момент число методов разбиения групп объектов на кластеры достаточно велико – несколько десятков алгоритмов и еще больше их модификаций.

Областей применения кластерного анализа достаточно много: сегментация изображений, маркетинг, прогнозирование, анализ текстов и многое другое.

Эта задача в последнее время имеет очень большое значение связи с большим количеством данных и тем, что их надо каким-то образом структурировать. Зачастую, кластеризация выступает самым первым этапом для разделения на группы при анализе данных. После выделения групп, для каждой группы строится отдельная модель.

Основные этапы кластерного анализа:

- отбор объектов для кластеризации;
- определение критериев, по которым будут оцениваться объекты;
- вычисление меры однородности между объектами;

Всероссийское СМИ
«Академия педагогических идей «НОВАЦИЯ»
Свидетельство о регистрации ЭЛ №ФС 77-62011 от 05.06.2015 г.
(выдано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций)

Сайт: akademnova.ru
e-mail: akademnova@mail.ru

- применение метода кластеризации для создания групподнородных объектов (кластеров);
- вывод результатов анализа.

Для начала необходимо составить список характеристик для каждого из объектов — как правило, это числовые значения, например, возраст и успеваемость ученика. Кроме того существуют алгоритмы, которые работают с качественными характеристиками.

Следующим этапом после определения вектора характеристик, следует выделить нормализацию, после которой все объекты выборки дают одинаковый вклад при расчете «расстояния». В данном этапе все значения приводятся к некоторому диапазону значений.

Далее, для каждой следующей пары объектов измеряется «расстояние» между ними (степень схожести). Вот некоторые из метрик измерения расстояний:

Евклидово расстояние.

Наиболее общий тип расстояния, представляющий собой геометрическое расстоянием в многомерном пространстве:

$$\boldsymbol{p(x,y)=\sqrt{\sum(xi-yi)2ni}}$$

Квадрат евклидова расстояния.

Зачастую такой тип измерения расстояния требуется при достаточно большом расстоянии между объектами.

$$\boldsymbol{p(x,y)=\sum(xi-yi)2ni}$$

Манхэттенское расстояние.

Сравнивая данный метод вычисления расстояния с Евклидовым, стоит отметить, что меры больших разностей между объектами оказывают меньшее влияние.

$$\boldsymbol{p(x,y)=\sum|xi-yi|\ ni}$$

Всероссийское СМИ
«Академия педагогических идей «НОВАЦИЯ»
Свидетельство о регистрации ЭЛ №ФС 77-62011 от 05.06.2015 г.
(выдано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций)

Сайт: akademnova.ru
e-mail: akademnova@mail.ru

Расстояние Чебышева.

В основном применяется в случае, когда необходимо определить оба объекта , как различные друг от друга, если они различны хотя бы по одной координате.

$$p(x,y)=max(|xi-yi|)$$

Процент несогласия.

Данный метод вычисления расстояния используется в основном при наличии категориальных данных у объектов.

$$p(x,y)=count\ (xi\neq yi)i/$$

Задачей исследователя на данном этапе служит выбор наиболее подходящего, правильного метода, так как полученные данные в ходе решения задачи могут существенно отличаться от его выбора.

Алгоритмом кластеризации является функция a: X→Y , которая любому объекту x множества X соотносит метку кластера y множества Y. В некоторых методах, количество меток Y известно заранее.

Решение задачи кластеризации не однозначно, по некоторым причинам. Во-первых, нет общего критерия качества кластерного анализа. Однако существуют некоторый разумный ряд критериев.

Во-вторых, зачастую, число кластеров заранее неизвестно и устанавливается относительно достаточного субъективного критерия.

В-третьих, результаты зависят от первоначально заданной метрики, которая весьма субъективна и выбирается исследователем.

Рассмотрим 3 алгоритма:

Алгоритм k-means (k-средних).

Наиболее распространенный и простой метод кластеризации объектов, но в то же время не достаточно точный. Для него количество кластеров (k) должно быть известно заранее. Идея алгоритма такова, что для каждой

Всероссийское СМИ
«Академия педагогических идей «НОВАЦИЯ»
Свидетельство о регистрации ЭЛ №ФС 77-62011 от 05.06.2015 г.
(выдано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций)

Сайт: akademnova.ru
e-mail: akademnova@mail.ru

итерации при которой объекты распределяются по кластерам, выполняется обновление центра масс всех кластеров. Затем процесс распределения объектов по кластерам повторяется уже между обновленными центрами масс. Так продолжается, пока при очередном обновлении, ни один из центра масс не изменится.

Недостатки данного алгоритма:

- количество кластеров необходимо знать заранее.

- результат сильно зависит от первоначального выбора центров масс.

Алгоритм кластеризации c-means.

Данный алгоритм находит в себе решение этой проблемы алгоритма k-means. Вместо утверждения того, что объект принадлежит конкретному кластеру, он определяет вероятность “схожести” объекта к каждому кластеру. Получается, можно сказать, что «Объект X принадлежит к кластеру 1 с вероятностью 80%, а к кластеру 2 с вероятностью 20%» будет более удобным.

Метод ближайшего соседа.

Данный метод наименее оптимальный из перечисленных, но в то же время и самый простой. Ход его действий следующий:

Переходим по каждому объекту, находящемуся вне кластера и относим его к кластеру расстояние между центром масс которого меньше заранее заданного порога. Результаты рассматриваются и при необходимости порог увеличивается, если например объект относится к большому числу кластеров.

Минимальное покрывающее дерево.

Данный алгоритм относится к алгоритмам иерархического типа «сверху-вниз». Это достаточно большой класс алгоритмов кластеризации

Всероссийское СМИ
«Академия педагогических идей «НОВАЦИЯ»
Свидетельство о регистрации ЭЛ №ФС 77-62011 от 05.06.2015 г.
(выдано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций)

Сайт: akademnova.ru
e-mail: akademnova@mail.ru

который базируется на выборке в виде графа. Вершинами графа приходятся объекты, а ребрами - расстояния между объектами.

Плюсами алгоритмов основанных на графах можно считать наглядность и возможность вносить некоторые улучшения.

Недостатками является то, что алгоритм находит применение в основном для кластеров типа “сгущений” или типа “лент”. Также недостатком является высокая трудоемкость, т.е. для нахождения минимального пути требуется O(N^3).

Первоначально вся структура объектов помещается в один общий кластер. На каждом последующем шаге, один из кластеров разбивается на два таким образом, чтобы расстояние между ними было максимально. Пример представлен на рисунке ниже.

Алгоритм выделения связных компонентов.

Данный метод, как и «минимальное покрывающее дерево» также основывается на теории графов. Вводится параметр R, после чего в графе удаляются все ребра (i,j), при при которых pij > R . В итоге, пары объектов, которые наиболее близки остаются соединенными. Алгоритм продолжает выполняться до тех пор, пока граф не будет иметь несколько связных компонент, которые и будут кластерами.

Для кластеризации данных большого объема часто выбирают метод k-Means, а также его модификации.

Существуют проблема нехватки памяти для объектов. Решить эту проблему можно путем разбиения всего множества объектов на подмножества и проведения кластеризации непосредственно внутри них, далее работая уже с одним представителем от каждого кластера.

Всероссийское СМИ
«Академия педагогических идей «НОВАЦИЯ»
Свидетельство о регистрации ЭЛ №ФС 77-62011 от 05.06.2015 г.
(выдано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций)

Сайт: akademnova.ru
e-mail: akademnova@mail.ru

В ходе написания данной статьи было рассмотрено понятие кластерного анализа и основные области его применения. Также были рассмотрены некоторые из мер расстояний и представлены определенные алгоритмы кластеризации.

Список использованной литературы:

1. Котов А., Красильников Н. Кластеризация данных. 2006.
2. Чубукова И.А. Курс лекций «Data Mining», Интернет-университет информационных технологий [Электронный ресурс]: URL: www.intuit.ru/department/database/datamining/ (дата обращения: 01.10.18)
3. Seagate [Электронный ресурс] URL: http://www.seagate.com/ru/ru/our-story/data-age-2025/ (дата обращения: 18.10.18)
4. BaseGroup Labs [Электронный ресурс] URL: https://basegroup.ru/community/articles/datamining/ (дата обращения: 20.10.18)

Дата поступления в редакцию: 18.12.2018 г.
Опубликовано: 24.12.2018 г.