ВОЛГОГРАДСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

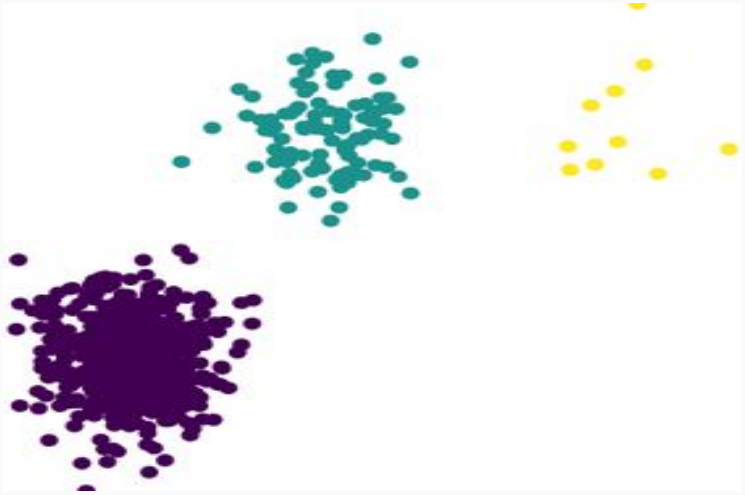
Кафедра государственного и муниципального управления

# Тема №8. Область применения многомерного метода анализа данных: кластеризация.

Выполнил:
Студент ГМУм-201
Мелихов Денис Олегович
Проверил:
Доцент, к.э.н.
Калинина Вера Владимировна

Волгоград - 2021

Кластеризация (или кластерный анализ) - это задача разбиения множества объектов на группы, называемые кластерами. Внутри каждой группы должны оказаться «похожие» объекты, а объекты разных группы должны быть как можно более отличны.



Главное отличие кластеризации от классификации состоит в том, что перечень групп чётко не задан и определяется в процессе работы алгоритма.

Задача кластеризации относится к статистической обработке, а также к широкому классу задач обучения без учителя (один из способов машинного обучения, при котором испытуемая система спонтанно обучается выполнять поставленную задачу без вмешательства со стороны экспериментатора).

Спектр применений кластерного анализа очень широк: его используют в археологии, медицине, психологии, химии, биологии, государственном управлении, филологии, антропологии, маркетинге, социологии, геологии и других дисциплинах.

Однако универсальность применения привела к появлению большого количества несовместимых терминов, методов и подходов, затрудняющих однозначное использование и непротиворечивую интерпретацию кластерного анализа.

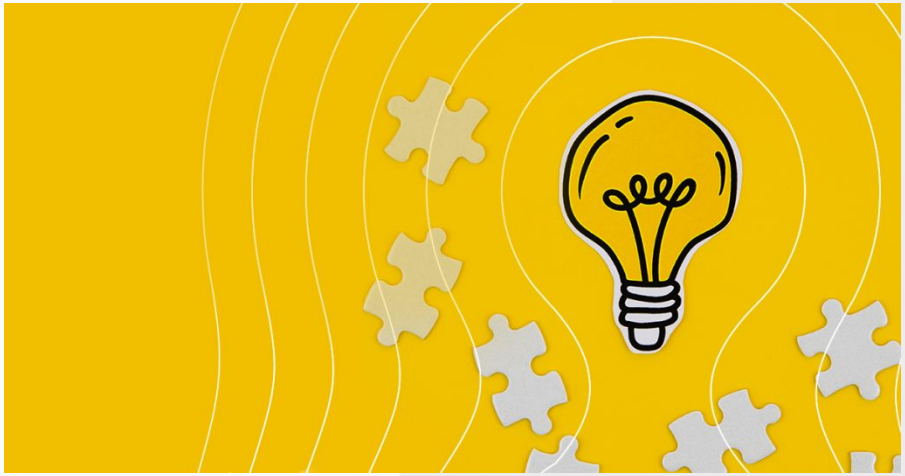
Кластерный анализ выполняет следующие основные задачи:

- 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.

Можно встретить описание двух фундаментальных требований, предъявляемых к данным - однородность и полнота. Однородность требует, чтобы все кластеризуемые сущности были одной природы, описывались сходным набором характеристик.



Если кластерному анализу предшествует факторный анализ, то выборка не нуждается в «ремонте» - изложенные требования выполняются автоматически самой процедурой факторного моделирования. В противном случае выборку нужно корректировать.

# Применение метода кластеризации:

1. Биология и биоинформатика (в области экологии кластеризация используется для выделения пространственных и временных сообществ организмов в однородных условиях).
2. Медицина (используется в позитронно-эмиссионной томографии для автоматического выделения различных типов тканей на трехмерном изображении).
3. Маркетинг (кластеризация широко используется при изучении рынка для обработки данных, полученных из различных опросов).
4. Интернет (выделение групп людей на основе графа связей в социальных сетях).

Кластеризация – объединение в группы схожих объектов – является одной из фундаментальных задач в области анализа данных и Data Mining.

Кластеризация в Data Mining приобретает ценность тогда, когда она выступает одним из этапов анализа данных, построения законченного аналитического решения. Аналитику часто легче выделить группы схожих объектов, изучить их особенности и построить для каждой группы отдельную модель, чем создавать одну общую модель на всех данных.

Таким приёмом постоянно пользуются в маркетинге, выделяя группы клиентов, покупателей, товаров и разрабатывая для каждой из них отдельную стратегию.

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

- высокая размерность (тысячи полей) и большой объём (сотни тысяч и миллионы записей) таблиц баз данных и хранилищ данных (сверхбольшие базы данных);
- наборы данных содержат большое количество числовых и категорийных атрибутов.

Все атрибуты или признаки объектов делятся на числовые и категорийные. Числовые атрибуты – это такие, которые могут быть упорядочены в пространстве, соответственно категорийные – которое не могут быть упорядочены.

Например, атрибут «возраст» – числовой, а «цвет» – категорийный. Приписывание атрибутам значений происходит во время измерений выбранным типом шкалы, а это, представляет собой отдельную задачу.

Большинство алгоритмов кластеризации предполагают сравнение объектов между собой на основе некоторой меры близости (сходства). Мерой близости называется величина, имеющая предел и возрастающая с увеличением близости объектов. Меры сходства «изобретаются» по специальным правилам, а выбор конкретных мер зависит от задачи, а также от шкалы измерений.

Потребность в обработке больших массивов данных в Data Mining привела к формулированию требований, которым, по возможности, должен удовлетворять алгоритм кластеризации. К таким требованиям относятся:

- минимально возможное количество проходов по базе данных;
- работа в ограниченном объеме оперативной памяти компьютера;
- работу алгоритма можно прервать с сохранением промежуточных результатов, чтобы продолжить вычисления позже;
- алгоритм должен работать, когда объекты из базы данных могут извлекаться только в режиме однонаправленного курсора (т.е. в режиме навигации по записям).

Алгоритм, удовлетворяющий данным требованиям (особенно второму), называется масштабируемым. Масштабируемость – важнейшее свойство алгоритма, зависящее от его вычислительной сложности и программной реализации.

Трудно соблюсти баланс между высоким качеством кластеризации и масштабируемостью. Поэтому в идеале в арсенале Data Mining должны присутствовать как эффективные алгоритмы кластеризации микромассивов, так и масштабируемые для обработки сверхбольших баз данных.

Таким образом, не существует единого универсального алгоритма кластеризации. При использовании любого алгоритма важно понимать его достоинства и недостатки, учитывать природу данных, с которыми он лучше работает и способность к масштабируемости.

# Список использованной литературы


1. Барсегян и др. Методы и модели анализа данных: OLAP и Data Mining. - СПб., 2004.
2. Жамбю М. Иерархический кластер-анализ и соответствия. - М.: Финансы и статистика, 1988. - 345 с.
3. Хайдуков Д. С. Применение кластерного анализа в государственном управлении// Философия математики: актуальные проблемы. - М.: МАКС Пресс, 2009. - 287 с.
4. Обзор алгоритмов кластеризации данных [Электронный ресурс]. – Режим доступа: https://habr.com/ru/post/101338/, свободный.

ВОЛГОГРАДСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

# БЛАГОДАРЮ ЗА ВНИМАНИЕ!