Реферат

# Алгоритмы кластеризации. Метод k-средних. Метод k-медиан

**Введение**

Кластеризация — это процесс разбиения множества объектов на группы (кластеры), где объекты внутри одного кластера схожи между собой, а между кластерами — различимы. Этот метод широко применяется в машинном обучении, анализе данных и экономике для выявления закономерностей и структурирования информации. В данном реферате рассматриваются два популярных алгоритма кластеризации: метод k-средних и метод k-медиан, их особенности, метрики и применение.

1. **Понятие кластера и кластеризации**

Кластер — это группа объектов, обладающих схожими характеристиками. Кластеризация позволяет автоматически группировать данные без предварительных знаний о количестве или свойствах классов. Основная задача — минимизировать расстояния внутри кластеров и максимизировать расстояния между ними.

Формула расстояния между точками в многомерном пространстве:

где и — векторы признаков, — весовые коэффициенты, учитывающие значимость признаков.

Для работы с разнотипными данными (например, метры и килограммы) применяется нормировка, чтобы привести все признаки к сопоставимым масштабам.

1. **Метод k-средних**

Метод k-средних – это метод кластерного анализа, цель которого является разделение множества из N объектов (наблюдений) из пространства Rn на k кластеров, при этом каждый объект относится к тому кластеру, к центру которого оно ближе всего

Алгоритм:

1. Инициализация центроидов: выбираются k точек, которые становятся начальными центрами кластеров. Это можно сделать случайным образом или с использованием специальных методов, например, максимизации расстояния между центрами.
2. Распределение объектов по кластерам: для каждого объекта вычисляется расстояние до каждого центра кластера. Объект относится к тому кластеру, расстояние до центра которого минимально.
3. Обновление центроидов: для каждого кластера вычисляется новый центр как среднее арифметическое координат всех его объектов .
4. Проверка сходимости: если новые центры совпадают со старыми, алгоритм завершается. В противном случае шаги 2–4 повторяются.
5. **Метод k-медиан**

Метод k-медиан является модификацией метода k-средних, где вместо среднего значения для определения центра кластера используется медиана. Этот подход минимизирует ошибку в метрике L1(Манхеттенское расстояние), в отличие от L2 (Евклидово расстояние), применяемого в методе k-средних.

Количество кластеров k задаётся заранее и может быть определено на основе экспертных знаний или эмпирических данных.

1. **Модельные алгоритмы**

Модельные алгоритмы кластеризации предполагают наличие заранее заданной модели структуры кластера, и их цель — максимизировать соответствие данных этой модели.

Одним из популярных подходов является кластеризация на основе плотности, например, алгоритм DBSCAN (Density-Based Spatial Clustering of Applications with Noise). Его основная идея — выделение областей с высокой концентрацией точек, разделённых областями с низкой плотностью. Точки в областях с низкой плотностью считаются шумом.

Некоторые алгоритмы кластеризации на основе плотности:

DBSCAN. Группирует точки данных, плотно расположенные друг к другу, и помечает выбросы как шум на основе их плотности в пространстве объектов. DBSCAN может обнаруживать кластеры произвольной формы и не требует предварительного задания количества кластеров.

OPTICS. Сочетает элементы иерархической и плотностной кластеризации. В отличие от DBSCAN, OPTICS не требует заранее заданных параметров радиуса окрестности и минимального числа точек — вместо этого алгоритм строит граф достижимости, который учитывает расстояния между точками и плотность данных.

HDBSCAN. Создаёт иерархическую структуру кластеров и требует только одного параметра — минимального числа точек в кластере.

Выбор метрики:

* Для метода k-средних используется Евклидова метрика (L2).
* Для метода k-медиан применяется Манхеттенская метрика (L1).
* Для текстовых данных часто используется косинусная близость:

1. **Алгоритм EM**

Алгоритм использует широко известный метод максимизации ожиданий (максимума правдоподобия – Expectation Maximization). В наиболее простом случае предполагается, что кластер – результаты наблюдений, распределенные нормально. Тогда для их характеристики можно применять многомерную функцию Гаусса (многомерное распределение Гаусса)

где — математическое ожидание, а — стандартное отклонение.

Многомерное нормальное распределение для q-мерного пространства является обобщением предыдущего выражения. Многомерная функция плотности вероятности для случайного имеет вид

Здесь q – размерность пространства; – вектор признаков, – математическое ожидание – матрица ковариации.

В качестве расстояния используем – квадратичное расстояние Махаланобиса, представляющая собой меру расстояния между векторами случайных величин, обобщающую понятие евклидова расстояния, учитывающую корреляции между переменными.

– одно из распределений. И тогда основная задача – определить, к какому из распределений принадлежит каждая конкретная точка, оценив параметры этих распределений исходя из реального распределения точек.

Для оценки условной вероятностей будем использовать формула Байеса:

где

— априорная вероятность гипотезы A;

— вероятность гипотезы при наступлении события B (апостериорная вероятность);

— вероятность наступления события B при истинности гипотезы A;

— полная вероятность наступления события B.

Алгоритм кластеризации представим в виде

1. Инициализируем: – среднее отклонение распределений относительно начала координат (т.е. центры кластеров) и вероятности этих распределений для каждой точки. K - число кластеров - задаем.
2. E-шаг (Expectation): Оцениваются вероятности принадлежности каждой точки к каждому кластеру.

,

где – функция плотности распределения.

1. М-шаг (Maximization): Обновляются параметры распределений на основе текущих вероятностей.

;

;

.

1. Вычисление

и сравнение с , если да, то стоп - найден локальный максимум. Если нет, то переходим снова к шагу 1.

И еще одно замечание. Определение расстояния между точкой и множеством могут определяться различными способами. Например, как расстояние от данной точки, до некоторого центра множества; расстояние от данной точки, до ближайшей точки множества; расстояние от данной точки, до самой дальней точки множества.

Таким образом результаты кластеризации существенно зависит от применяемых методов кластеризации, в первую очередь, от выбранной метрики.