Лекция 6

Методы кластеризации

1. Введение в кластеризацию

Кластеризация — это метод машинного обучения, применяемый для группировки похожих объектов в кластеры, основываясь на их характеристиках. Кластеризация относится к типу обучения без учителя, поскольку она не требует предварительно размеченных данных и стремится выявить внутреннюю структуру данных, обнаруживая скрытые связи. Основная цель кластеризации — максимизировать похожесть внутри каждого кластера и минимизировать похожесть между кластерами.

Методы кластеризации находят широкое применение в таких областях, как анализ рынка, биоинформатика, обработка изображений, обработка естественного языка и маркетинговая сегментация. Например, в маркетинге кластеризация используется для сегментации клиентов на основе их покупательских предпочтений, а в биоинформатике — для группировки генов или белков со схожими функциями.

2. Основные подходы к кластеризации

Существует несколько подходов к кластеризации, каждый из которых имеет свои особенности и применимость. Рассмотрим основные группы методов кластеризации:

- **Центроидные методы**: Базируются на нахождении центральных точек (центроидов) для каждого кластера. Примеры: K-means и K-медиоидная кластеризация.
- **Иерархические методы**: Постепенно формируют кластеры на основе структуры данных. Примеры: агломеративная и дивизивная иерархическая кластеризация.
- **Плотностные методы**: Ищут области с высокой плотностью точек, образуя кластеры. Пример: DBSCAN.
- **Модели на основе распределений**: Предполагают, что данные генерируются из определенного распределения. Пример: алгоритмы на основе смеси гауссиан.

3. Центроидные методы кластеризации

3.1 Метод K-means

Метод K-means — один из самых популярных и простых алгоритмов кластеризации. Он делит данные на KKK кластеров, определяя для каждого из них центроид. Алгоритм работает следующим образом:

1. Задается число кластеров KKK и случайным образом выбираются KKK центроидов.
2. Каждая точка данных назначается кластера, чей центроид находится к ней ближе всего.
3. Вычисляются новые центроиды для каждого кластера.
4. Повторяются шаги 2 и 3 до тех пор, пока центроиды не перестанут изменяться или пока не будет достигнуто заданное количество итераций.

Преимущества K-means включают простоту и эффективность, однако алгоритм требует заранее заданного количества кластеров и чувствителен к выбросам. Кроме того, он предполагает, что кластеры имеют сферическую форму, что не всегда соответствует реальным данным.

3.2 Метод K-медиоидной кластеризации (K-medoids)

K-медиоидная кластеризация похожа на K-means, но вместо центроидов используются медиоиды — реальные точки данных, которые минимизируют расстояние до остальных точек в кластере. Этот подход устойчив к выбросам, поскольку медиоиды не так сильно смещаются из-за аномальных значений. Однако K-медиоидная кластеризация может быть менее эффективной для больших наборов данных, чем K-means, так как требует больше вычислительных ресурсов.

4. Иерархические методы кластеризации

Иерархические методы кластеризации строят дерево кластеров (дендрограмму), которая визуализирует структуру данных. В зависимости от подхода иерархическая кластеризация может быть агломеративной или дивизивной.

4.1 Агломеративная кластеризация

Агломеративная кластеризация — это нисходящий метод, который начинает с того, что каждая точка данных считается отдельным кластером, и постепенно объединяет кластеры на основе расстояния между ними, пока не останется один большой кластер. Основные шаги:

1. Инициализация: каждая точка данных — отдельный кластер.
2. Поиск ближайших кластеров и их объединение.
3. Повторение шага 2 до тех пор, пока не останется один кластер или не будет достигнуто нужное количество кластеров.

4.2 Дивизивная кластеризация

Дивизивная кластеризация — это восходящий метод, который начинает с одного большого кластера, включающего все данные, и постепенно делит его на более мелкие кластеры, основываясь на расстоянии между точками. Хотя дивизивная кластеризация применяется реже, она может быть полезна в задачах, где требуется выделение четко разделенных групп данных.

Преимущества иерархических методов заключаются в том, что они не требуют заранее определенного количества кластеров и позволяют визуализировать структуру данных. Однако они вычислительно затратны, особенно на больших наборах данных.

5. Плотностные методы кластеризации

5.1 DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

DBSCAN — это популярный алгоритм кластеризации, который образует кластеры на основе плотности точек в пространстве. В отличие от K-means, DBSCAN не требует предварительного задания числа кластеров и может обнаруживать кластеры произвольной формы. Основные шаги алгоритма:

1. Для каждой точки определяются соседние точки на основе радиуса ε\varepsilonε.
2. Если точка содержит достаточно соседей (заданное значение MinPts), она становится «ядром» кластера, и все соседние точки добавляются в этот кластер.
3. Процесс повторяется для всех точек, пока не будут определены все кластеры.
4. Точки, которые не входят в кластер, считаются шумом.

Преимущества DBSCAN включают способность обнаруживать кластеры произвольной формы и устойчивость к шуму. Однако алгоритм зависит от выбора параметров ε\varepsilonε и MinPts, и может не работать хорошо при наличии кластеров различной плотности.

5.2 OPTICS (Ordering Points to Identify the Clustering Structure)

OPTICS — это расширение DBSCAN, которое позволяет обнаруживать кластеры различной плотности. OPTICS упорядочивает точки данных таким образом, чтобы кластеры можно было выделить на основе плотности, а затем строит дендрограмму плотностей, которая помогает понять структуру данных. OPTICS требует меньше жестких параметров, чем DBSCAN, и подходит для данных с переменной плотностью.

6. Модельные методы кластеризации

Модельные методы предполагают, что данные генерируются из смеси распределений и стремятся найти параметры этих распределений для разделения данных на кластеры. Одним из самых известных методов является кластеризация на основе смеси гауссиан (Gaussian Mixture Models, GMM).

6.1 Модель смеси гауссиан (Gaussian Mixture Model, GMM)

GMM использует вероятностный подход для кластеризации, предполагая, что данные распределены по нескольким гауссовым распределениям с разными параметрами. Модель строит несколько гауссиан, каждая из которых описывает отдельный кластер, и оценивает параметры с помощью алгоритма ожидания-максимизации (Expectation-Maximization, EM).

Основные шаги GMM:

1. Задается количество гауссиан и их параметры (среднее, дисперсия).
2. Алгоритм EM чередует два шага:
    - **Ожидание (E-step)**: находит вероятности принадлежности каждой точки к каждому распределению.
    - **Максимизация (M-step)**: обновляет параметры распределений, чтобы максимизировать правдоподобие.

GMM позволяет создавать кластеры произвольной формы и учитывать вероятности принадлежности кластерам, что дает более гибкий подход, чем K-means. Однако GMM требует предварительного задания количества кластеров и чувствителен к выбросам.

7. Оценка качества кластеризации

Оценка кластеризации — сложная задача, поскольку в обучении без учителя отсутствуют метки для сравнения. Однако существует несколько метрик, которые помогают оценить качество кластеров:

- **Внутрикластерное и межкластерное расстояние**: Чем меньше расстояние между точками внутри кластера и больше расстояние между кластерами, тем лучше разделение.
- **Силуэтный коэффициент (Silhouette Score)**: Среднее расстояние между точкой и точками её кластера по сравнению с ближайшим кластером. Значение варьируется от -1 до 1, где большее значение указывает на лучшую кластеризацию.

Silhouette Score=b−amax⁡(a,b)\text{Silhouette Score} = \frac{b - a}{\max(a, b)}Silhouette Score=max(a,b)b−a

где aaa — среднее внутрикластерное расстояние, bbb — среднее расстояние до соседнего кластера.

- **Коэффициент Дэвиса-Болдина**: Среднее соотношение внутрикластерного расстояния к межкластерному расстоянию. Чем ниже значение, тем лучше качество кластеризации.

8. Применение методов кластеризации

Методы кластеризации находят применение в различных областях:

- **Маркетинг и анализ клиентов**: Сегментация клиентов для целевого маркетинга.
- **Биология и биоинформатика**: Группировка генов или белков по их функциям и характеристикам.
- **Обработка изображений**: Сегментация изображений для выделения объектов или областей.
- **Обработка естественного языка**: Кластеризация текстов или документов по тематике.

9. Заключение

Кластеризация — это мощный инструмент для анализа неразмеченных данных и выявления скрытых закономерностей. Различные методы кластеризации подходят для разных типов данных и задач, поэтому важно понимать их преимущества и ограничения.