ОБЗОР СОВРЕМЕННЫХ АЛГОРИТМОВ КЛАСТЕРИЗАЦИИ ДАННЫХ

Мангутова Е.А.[1], Гончаров А.С.[2]
[1] ТПУ, ИШИТР, 8К02, e-mail: eam49@tpu.ru
[2] ТПУ, ИШИТР, ассистент ОИТ, e-mail: asg19@tpu.ru

Введение

Для анализа данных используются различные алгоритмы обработки информации, одним из которых является кластеризация данных. И в настоящее время начинающие программисты могут столкнуться с проблемой выбора метода кластеризации данных, поскольку каждая задача может требовать определенного алгоритма кластеризации для достижения наилучших результатов. Поэтому знание различных алгоритмов кластеризации данных может значительно облегчить решение поставленной задачи.

Кластеризация данных — это метод анализа данных, который позволяет разбить большой набор данных на множество меньших подгрупп, называемых кластерами, таким образом, чтобы объекты внутри каждого кластера были максимально похожими между собой, а объекты из разных кластеров - различными. Кластеризация данных может быть использована для обнаружения скрытых закономерностей в данных, для группировки данных в соответствии с определенными критериями и для упрощения сложных данных путем их структурирования в группы.

Кластеризация данных может быть выполнена различными способами, включая иерархическую кластеризацию, метод k-средних и алгоритмы, основанные на плотности данных. Выбор конкретного метода зависит от типа данных, которые необходимо анализировать, и от конкретной задачи, которую необходимо решить.

Алгоритмы кластеризации

1. Алгоритм К-средних является одним из наиболее распространенных методов кластеризации данных. Он основывается на разбиении данных на К кластеров, где К является заданным параметром. Алгоритм работает следующим образом: сначала случайным образом выбираются К центроидов. Затем каждая точка данных относится к ближайшему центроиду, и центроиды обновляются в соответствии с новым разбиением данных. Процесс повторяется до тех пор, пока не будет достигнута заданная точность;

2. Алгоритм DBSCAN основывается на плотности данных. Он может определять кластеры произвольной формы и обнаруживать выбросы. Алгоритм работает следующим образом: для каждой точки данных определяется, сколько точек находятся в заданном радиусе от нее. Если точка находится в плотной области данных (имеет достаточное количество соседей), то она считается частью кластера. Если же точка находится в области данных с малой плотностью или находится далеко от других точек, то она считается выбросом;

3. Алгоритм OPTICS является расширением DBSCAN. Он не только выделяет кластеры, но и упорядочивает их по убыванию плотности. OPTICS позволяет выявлять как плотные, так и разреженные кластеры, и работает с выбросами;

4. Алгоритм HDBSCAN является улучшенной версией DBSCAN. Он использует иерархическую структуру кластеров и позволяет выделять кластеры различных размеров и форм, и не требует задания числа кластеров заранее;

5. Аффинное распространение – это алгоритм, который определяет число кластеров автоматически, и позволяет выделять кластеры различной формы и размеров. Он основывается на распространении сообщений между точками, и использует матрицы схожести для выявления кластеров. Аффинное распространение может быть чувствительно к выбору начальных значений и может дать неравномерные кластеры в зависимости от выбора параметров;

Формула элементов матрицы схожести:

$$s_{j,k} = \sum_j (a_{ji} - a_{jk})^2 \quad (1)$$

где j, k – номер строки и столбца исходного набора данных.

6. Спектральная кластеризация - алгоритм использует матрицу сходства для разбиения данных на кластеры. Эта матрица также описывает полный граф с вершинами и рёбрами с весом, соответствующим степени похожести связанных вершин. Алгоритм преобразует матрицу в новое пространство и выполняет кластеризацию в этом пространстве;

7. Смешанная модель Гаусса - алгоритм использует статистический подход для моделирования распределения данных. Он моделирует каждый кластер как набор нормальных (Гауссовых) распределений и использует метод максимального правдоподобия, чтобы определить параметры каждого распределения. Затем он использует эти модели для присвоения точек входных данных к наиболее вероятному кластеру на основе оценки плотности вероятности. В отличие от методов, основанных на расстоянии, смешанная модель Гаусса может моделировать кластеры, которые имеют нетипичные формы, и может найти скрытые кластеры в данных. Однако этот метод требует настройки нескольких гиперпараметров, таких как количество кластеров и форма каждого распределения, что может быть сложно для подбора на практике.

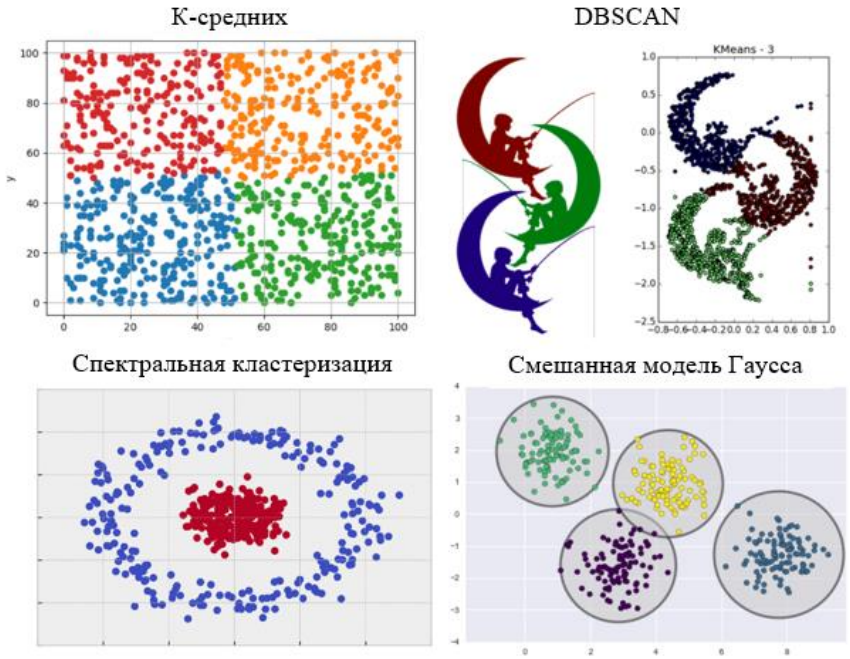

Рис. 1. Результаты работы методов

Заключение

Алгоритмы кластеризации данных являются эффективным инструментом для анализа и обработки больших объемов информации. Каждый из этих алгоритмов имеет свои преимущества и недостатки, и может быть результативным для решения конкретной задачи. Выбор оптимального алгоритма кластеризации данных зависит от типа данных, количества объектов, требуемой точности, и других факторов.

Список литературы

1. Кластеризация / [Электронный ресурс] // SciKit-learn: [сайт]. — URL: https://scikit-learn.ru/clustering/ (дата обращения: 27.02.2023).
2. Affinity Propagation Algorithm Explained / [Электронный ресурс] // Towards datascience: [сайт]. — URL: https://towardsdatascience.com/unsupervised-machine-learning-affinity-propagation-algorithm-explained-d1fef85f22c8 (дата обращения: 27.02.2023).
3. Реализация кластеризации методом k-средних / [Электронный ресурс] // Habr: [сайт]. — URL: https://habr.com/ru/post/585034/ (дата обращения: 28.02.2023).
4. Affinity Propagation / [Электронный ресурс] // SciKit-learn: [сайт]. — URL: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AffinityPropagation.html (дата обращения: 28.02.2023).
5. Анализ данных — основы и терминология / [Электронный ресурс] // Habr: [сайт]. — URL: https://habr.com/ru/post/352812/ (дата обращения: 28.02.2023).
6. Процесс анализа данных / [Электронный ресурс] // PythonRU: [сайт]. — URL: https://habr.com/ru/post/352812/ (дата обращения: 28.02.2023).