Кластеризация — задача группировки множества объектов на подмножества (кластеры) таким образом, чтобы объекты из одного кластера были более похожи друг на друга, чем на объекты из других кластеров по какому-либо критерию.

Задача кластеризации относится к классу задач обучения без учителя.

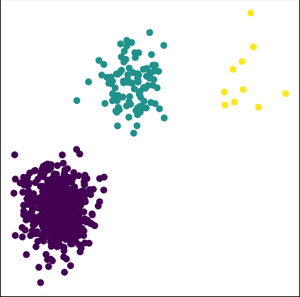


Рисунок 1. Пример кластеризации

Пусть X — множество объектов, Y — множество идентификаторов (меток) кластеров. На множестве X задана функция расстояния между объектами ρ(x,x′). Дана конечная обучающая выборка объектов Xm={x1,…,xm}⊂X. Необходимо разбить выборку на подмножества (кластеры), то есть каждому объекту xi∈Xm сопоставить метку yi∈Y, таким образом чтобы объекты внутри каждого кластера были близки относительно метрики ρ, а объекты из разных кластеров значительно различались.

Множество Y в некоторых случаях известно заранее, однако чаще ставится задача определить оптимальное число кластеров, с точки зрения того или иного критерия качества кластеризации.

Кластеризация (обучение без учителя) отличается от классификации (обучения с учителем) тем, что метки объектов из обучающей выборки yi изначально не заданы, и даже может быть неизвестно само множество Y.

Решение задачи кластеризации объективно неоднозначно по ряду причин:

Не существует однозначного критерия качества кластеризации. Известен ряд алгоритмов, осуществляющих разумную кластеризацию "по построению", однако все они могут давать разные результаты. Следовательно, для определения качества кластеризации и оценки выделенных кластеров необходим эксперт предметной области;

Число кластеров, как правило, заранее не известно и выбирается по субъективным критериям. Даже если алгоритм не требует изначального знания о числе классов, конкретные реализации зачастую требуют указать этот параметр;

Результат кластеризации существенно зависит от метрики. Однако существует ряд рекомендаций по выбору метрик для определенных классов задач.

Число кластеров фактически является гиперпараметром для алгоритмов кластеризации. Подробнее про другие гиперпараметры и их настройку можно прочитать в статье.

**Типы входных данных**

Признаковое описание объектов. Каждый объект описывается набором своих характеристик, называемых признаками. Признаки могут быть как числовыми, так и категориальными;

Матрица расстояний между объектами. Каждый объект описывается расстоянием до всех объектов из обучающей выборки.

Вычисление матрицы расстояний по признаковому описанию объектов может быть выполнено бесконечным числом способов в зависимости от определения метрики между объектами. Выбор метрики зависит от обучающей выборки и поставленной задачи.

**Цели кластеризации**

Классификация объектов. Попытка понять зависимости между объектами путем выявления их кластерной структуры. Разбиение выборки на группы схожих объектов упрощает дальнейшую обработку данных и принятие решений, позволяет применить к каждому кластеру свой метод анализа (стратегия «разделяй и властвуй»). В данном случае стремятся уменьшить число кластеров для выявления наиболее общих закономерностей;

Сжатие данных. Можно сократить размер исходной выборки, взяв один или несколько наиболее типичных представителей каждого кластера. Здесь важно наиболее точно очертить границы каждого кластера, их количество не является важным критерием;

Обнаружение новизны (обнаружение шума). Выделение объектов, которые не подходят по критериям ни в один кластер. Обнаруженные объекты в дальнейшем обрабатывают отдельно.

**Методы кластеризации**

Графовые алгоритмы кластеризации. Наиболее примитивный класс алгоритмов. В настоящее время практически не применяется на практике;

Вероятностные алгоритмы кластеризации. Каждый объект из обучающей выборки относится к каждому из кластеров с определенной степенью вероятности:

[EM-алгоритм](https://neerc.ifmo.ru/wiki/index.php?title=EM-%D0%B0%D0%BB%D0%B3%D0%BE%D1%80%D0%B8%D1%82%D0%BC);

[Иерархические алгоритмы кластеризации](https://neerc.ifmo.ru/wiki/index.php?title=%D0%98%D0%B5%D1%80%D0%B0%D1%80%D1%85%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B0%D1%8F_%D0%BA%D0%BB%D0%B0%D1%81%D1%82%D0%B5%D1%80%D0%B8%D0%B7%D0%B0%D1%86%D0%B8%D1%8F). Упорядочивание данных путем создания иерархии вложенных кластеров;

[Алгоритм k-средних](https://neerc.ifmo.ru/wiki/index.php?title=K-%D1%81%D1%80%D0%B5%D0%B4%D0%BD%D0%B8%D1%85&action=edit&redlink=1). Итеративный алгоритм, основанный на минимизации суммарного квадратичного отклонения точек кластеров от центров этих кластеров;

Распространение похожести. Распространяет сообщения о похожести между парами объектов для выбора типичных представителей каждого кластера;

Сдвиг среднего значения (англ. mean shift). Выбирает центроиды кластеров в областях с наибольшей плотностью;

Спектральная кластеризация (англ. spectral clustering). Использует собственные значения матрицы расстояний для понижения размерности перед использованием других методов кластеризации;

Основанная на плотности пространственная кластеризация для приложений с шумами (англ. Density-based spatial clustering of applications with noise, DBSCAN). Алгоритм группирует в один кластер точки в области с высокой плотностью. Одиноко расположенные точки помечает как шум

Алгоритм, который будет использоваться <https://algowiki-project.org/ru/%D0%A3%D1%87%D0%B0%D1%81%D1%82%D0%BD%D0%B8%D0%BA:Nikmedoed/%D0%9D%D0%B5%D1%87%D0%B5%D1%82%D0%BA%D0%B8%D0%B9_%D0%B0%D0%BB%D0%B3%D0%BE%D1%80%D0%B8%D1%82%D0%BC_%D0%A1-%D1%81%D1%80%D0%B5%D0%B4%D0%BD%D0%B8%D1%85_(Fuzzy_C-means)>