Введение в кластеризацию [M.120]

Кластеризация — одна из задач Data Mining, а кластер – группа похожих объектов (постановка задачи кластеризации рассматривалась в разделе BG.001 «Технологии анализа данных»). Существует много определений кластеризации, поэтому приведем несколько.

Определение

Кластеризация — 1) группировка объектов на основе близости их свойств; каждый кластер состоит из схожих объектов, а объекты разных кластеров существенно отличаются; 2) процедура, которая любому объекту $x \in X$ ставит в соответствие метку кластера $y \in Y$.

Кластеризацию используют, когда отсутствуют априорные сведения относительно классов, к которым можно отнести объекты исследуемого набора данных, либо когда число объектов велико, что затрудняет их ручной анализ.

Постановка задачи кластеризации сложна и неоднозначна, так как:

- оптимальное количество кластеров в общем случае неизвестно;
- выбор меры «похожести» или близости свойств объектов между собой, как и критерия качества кластеризации, часто носит субъективный характер.

На рисунке 1 показан пример кластеризации объектов, которые описываются некоторыми двумя числовыми признаками, поэтому объекты легко изобразить на плоскости. К сожалению, в реальных приложениях количество признаков объектов измеряется десятками, и такой способ их представления не подходит.

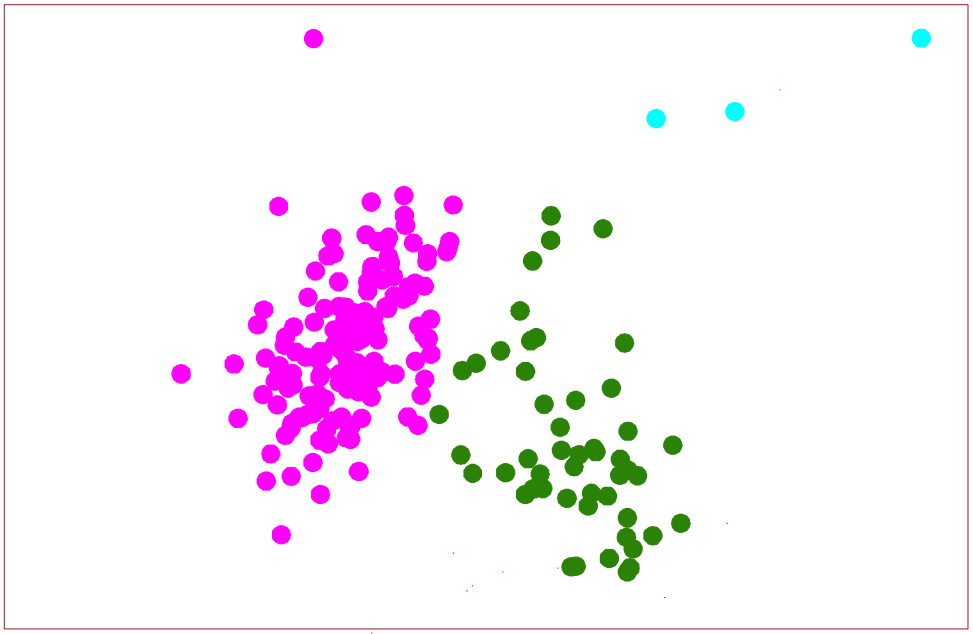

Рисунок 1 – Иллюстрация к задаче кластеризации

Естественно, приведенный вариант разбиения не является единственным.

Замечание

Задача кластеризации известна давно, и специалисты в различных областях оперируют рядом других терминов — таксономия, сегментация, группировка, самоорганизация. В Data Mining употребляется термин «кластеризация».

Остановимся на целях кластеризации и на ее применении в бизнес-аналитике.

Цели кластеризации в Data Mining могут быть различными и зависят от конкретной решаемой задачи. Рассмотрим эти задачи.

- **Изучение данных.** Разбиение множества объектов на группы помогает выявить внутренние закономерности, увеличить наглядность представления данных, выдвинуть новые гипотезы, понять, насколько информативны свойства объектов.
- **Облегчение анализа.** При помощи кластеризации можно упростить дальнейшую обработку данных и построение моделей: каждый кластер обрабатывается индивидуально, и модель создается для каждого кластера в отдельности. В этом смысле кластеризация может рассматриваться как подготовительный этап перед решением других задач Data Mining: классификации, регрессии, ассоциации, последовательных шаблонов.
- **Сжатие данных.** В случае, когда данные имеют большой объем, кластеризация позволяет сократить объем хранимых данных, оставив по одному наиболее типичному представителю от каждого кластера.
- **Прогнозирование.** Кластеры используются не только для компактного представления имеющихся объектов, но и для распознавания новых. Каждый новый объект относится к тому кластеру, присоединение к которому наилучшим образом удовлетворяет критерию качества кластеризации. Значит, можно прогнозировать поведение объекта, предположив, что оно будет схожим с поведением других объектов кластера.
- **Обнаружение аномалий.** Кластеризация применяется для выделения нетипичных объектов. Эту задачу также называют обнаружением аномалий (outlier detection). Интерес здесь представляют кластеры (группы), в которые попадает крайне мало, скажем один-три, объектов.

Первые две задачи наиболее популярны в бизнес-аналитике. В таблице 1 приведены практические примеры применения кластеризации в разных областях.

Таблица 1 – Примеры кластеризации в различных областях

№	Прикладная область	Цель кластеризации	Описание
1	Розничная торговля	Облегчение анализа	Построение в сети розничных магазинов ассоциативных правил, выявляющих совместно покупаемые продукты, приводило к громоздким результатам с большим числом правил. При помощи кластеризации все покупатели были разделены на несколько сегментов, и ассоциативные правила выявлялись в каждом сегменте отдельно. Это позволило разбить задачу над подзадачи и найти ассоциативные правила для каждого сегмента покупателей в отдельности
2	Банкинг	Изучение данных, облегчение анализа	Отдел продаж коммерческого банка, работающего на рынке розничного кредитования, задался целью изучить профили потенциальных клиентов, подающих заявки на потребительский кредит. Очень малочисленным оказался кластер под названием «молодежь» — работающие студенты и молодые люди в возрасте до 23 лет. Оказалось, что у банка не было точек продаж кредитов в тех районах города, где сосредоточены университеты и институты. Данный факт был учтен службой продаж и службой развития бизнеса банка
		Прогнозирование	Исследование клиентов банка, взявших автокредит, при помощи инструментов кластеризации выделило кластер, в который попали мужчины в возрасте от 23 до 28 лет, проживающие в Московской области менее года. Их объединяло то, что практически все они имели длительные просрочки по кредиту. Скорее всего, это молодые люди, переоценившие свои возможности, либо мошенники. Эта информация была учтена банком при разработке скоринговых карт
3	Теле-коммуникации	Изучение данных	Анализ базы данных клиентов крупной сети сотовой связи позволил выделить несколько кластеров. Самым малочисленным кластером оказались пожилые женщины, совершающие звонки в весенне-летнее время. С большой долей вероятности это пенсионерки-дачницы, значительную часть теплого времени года проживающие за городом. Для увеличения численности этого кластера был разработан соответствующий тарифный план, который наилучшим образом устраивал бы абонентов этой группы
4	Страхование	Обнаружение аномалий	Кластеризация клиентов страховой компании, застрахованных от несчастных случаев, выявила небольшой по объему кластер, в котором фигурировали одни и те же фамилии врачей; суммы страховых выплат тоже варьировались незначительно. Проверка показала, что в 90 % таких случаев имел место сговор с врачом
5	Государственные службы	Изучение данных	В ходе исследования базы данных миграционной службы, в которой содержалась информация о людях, переехавших из села в город (возраст, образование, семейное положение и т. д.), с помощью алгоритма кластеризации было выделено несколько сегментов, характеристики которых позволили дать им содержательную интерпретацию. Например, выделился кластер, куда вошли женщины в возрасте более 60 лет, дети которых живут в городе. С большой вероятностью это бабушки, которые едут в город нянчить своих внуков. Выделились также кластеры «демобилизованные солдаты», «невесты» и др.

Ранее в определении кластеризации встречалось словосочетание «похожесть свойств». Термины «похожесть», «близость» можно понимать по-разному, поэтому в зависимости от того, какой вариант оценки близости между свойствами объектов мы выберем, получим тот или иной вариант кластеризации.

В Data Mining распространенной мерой оценки близости между объектами является метрика, или способ задания расстояния. Проблема выбора той или иной метрики в моделях всегда остро стоит перед аналитиком. Наиболее популярные метрики — *евклидово расстояние* и *расстояние Манхэттена* — рассматриваются далее при изложении конкретных алгоритмов.

Важно понимать, что сама по себе кластеризация не приносит каких-либо результатов анализа. Для получения эффекта необходимо провести содержательную интерпретацию каждого кластера. Такая интерпретация предполагает присвоение каждому кластеру емкого названия, отражающего его суть, например «Разведенные женщины с детьми», «Дачники», «Работающие студенты» и т. д. Для интерпретации аналитик детально исследует каждый кластер: его статистические характеристики, распределение значений признаков объекта в кластере, оценивает мощность кластера — число объектов, попавших в него. Интерпретация значительно облегчается, если имеются способы представления результатов кластеризации в специализированном виде: *дендограммы*, *кластерограммы*, *карты*.

На сегодня предложено несколько десятков алгоритмов кластеризации и еще больше их разновидностей. Несмотря на это в Data Mining в первую очередь применяются алгоритмы, которые понятны и просты в использовании. Это алгоритм *k-means* и сети *Кохонена*. Они рассматриваются в следующих модулях.