Проблемы алгоритмов кластеризации [М.128]

Ранее уже отмечалось, что одно и то же множество объектов можно разбить на *k* кластеров по-разному. Это привело к изобилию алгоритмов кластеризации. Пожалуй, ни одна из других задач Data Mining не имеет в своем арсенале столько алгоритмов и методов решения.

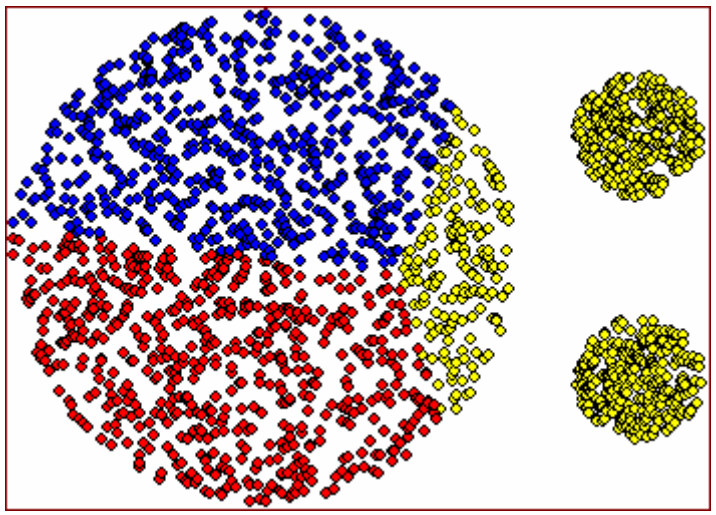
Причинами этого является несколько факторов, имеющих общее объяснение – не существует одного универсального алгоритмы кластеризации. Перечислим эти факторы и остановимся на каждом детальнее.

Неопределенность выбора критерия качества кластеризации.

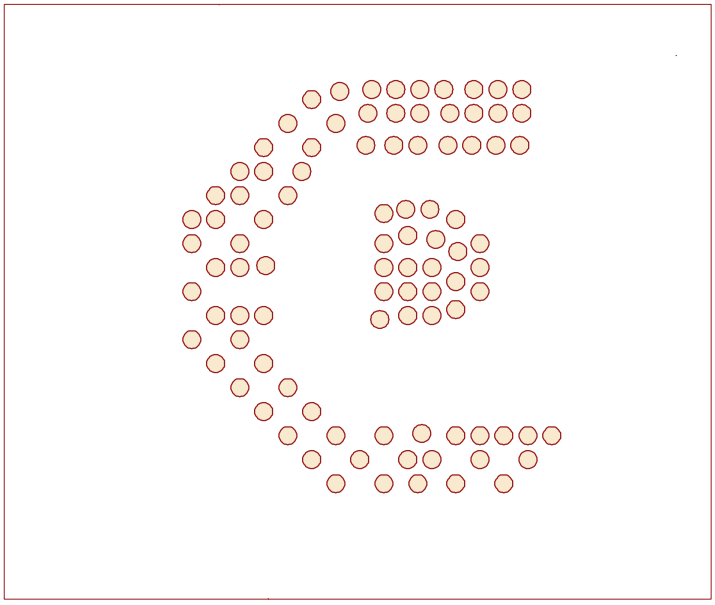
В Data Mining популярность при решении задач кластеризации имеют алгоритмы, которые ищут оптимальное разбиение множества данных на группы. Критерий оптимальности задается в целевой функции. Она полностью определяет результат кластеризации.

Например, семейство алгоритмов *k-means* показывает хорошие результаты, когда данные в пространстве образуют компактные сгустки, четко отличимые друг от друга. Поэтому и критерий качества основан на вычислении расстояний точек до центров кластера.

На рис. 1 приведен примеры неуспехов алгоритма *k-means*. На первом рисунке демонстрируется так называемый «эффект расщепления» – один кластер значительно больше остальных, и они находятся близко друг от друга.



а



б

Рис. 1. Слабые стороны алгоритма k-means:

а – эффект расщепления, б – вложенные кластеры

При использовании евклидовой метрики кластеры в *k-means* имеют форму сферы. Поэтому такой алгоритм никогда не разделит на кластеры случай вложенных друг в друга множеств объектов (рис. 1*б*).

Главная трудность выбора критерия качества кластеризации заключается в том, что *на практике в условиях, когда объекты описываются десятками и сотнями свойств, становится сложно оценить их взаимное расположение и подобрать адекватный алгоритм*.

Трудность выбора меры близости, обусловленная различной природой данных.

Особенность бизнес-данных такова, что в таблицах, описывающих свойства объектов, часто присутствуют различные типы данных[1]. В задаче кластеризации чаще всего это числовые и строковые данные. Строковые типы, в свою очередь, делятся на упорядоченные и категориальные.

Присутствия тех или иных типов данных в наборе определяют его природу. Назовем набор данных *числовым*, если он состоит только из целых и вещественных признаков. К вычислению расстояний между объектами таких наборов чаще всего подходит популярная метрика в виде евклидового расстояния.

Назовем набор данных **строковым**, если он состоит из упорядоченных и категориальных признаков (сюда же отнести логический признак). Для упорядоченных можно снова использовать евклидово расстояние, закодировав значения признака целыми числами. А вот к категориальным типам эта мера не подходит. Здесь нужно применять специальную меру расстояния, например, функцию отличия (англ.: *different function*), которая задается следующим образом:

$$d(x,y)=\begin{cases}0, & \text{если } x=y\\ 1, & \text{в остальных случаях}\end{cases},$$

где *x* и *y* – категориальные значения.

Наборы данных, содержащие признаки, к которым нельзя применять одну и ту же меру расстояний, называются **смешанными** (англ.: *mixed datasets*).

Проиллюстрируем вышесказанное на примере. Пусть требуется вычислить попарные расстояния между следующими объектами с атрибутами (*Возраст*, *Цвет глаз*, *Образование*):

(1) {23, карий, высшее},
(2) {25, зеленый, среднее},
(3) {26, серый, среднее}

Первый атрибут является числовым, остальные – строковыми. Причем признак *Цвет глаз* имеет категориальный тип, а *Образование* – упорядоченный. Если для вычисления расстояний мы выберем одну метрику – евклидову, то возникнут проблемы с признаком *Цвет глаз*. Для него подходит только функция отличия. А как учитывать при кластеризации, если категориальные атрибуты важнее числовых? На этот вопрос уже нет однозначного ответа.

Главная трудность выбора меры близости состоит в том, что *необходимость использования комбинации метрик ухудшает работу алгоритма, а эффективных алгоритмов кластеризации для смешанных наборов данных почти нет*.

Различные требуемые машинные ресурсы (память и время).

Как и любые алгоритмы, алгоритмы кластеризации имеют различную вычислительную сложность. Вопрос масштабируемости в кластеризации стоит особенно остро, так как эта задача Data Mining часто выступает первым шагом в анализе: после выделения схожих групп применяются другие методы, для каждой группы строится отдельная модель. В частности, именно по причине больших вычислительных затрат в Data Mining не получили распространение иерархические алгоритмы, которые строят полное дерево вложенных кластеров.

Для кластеризации больших массивов данных, содержащих миллионы строк, разработаны специальные алгоритмы, позволяющие добиваться приемлемого качества за несколько проходов по набору данных. Такие задачи, к примеру, актуальны при сегментации покупателей супермаркета по их чекам.

[1] *Типы данных обсуждались в модуле М.004 Структурированные данные*

Получение масштабируемых алгоритмов основано на идее отказа от *локальной* функции оптимизации. Парное сравнение объектов между собой в алгоритме *k-means* есть не что иное, как локальная оптимизация, т.к. на каждом шаге необходимо рассчитывать расстояние от центра кластера до каждого объекта. Это ведет к большим вычислительным затратам. При задании *глобальной* функции оптимизации добавление новой точки в кластер не требует больших вычислений: оно рассчитывается на основе старого значения, нового объекта и параметров кластера. К сожалению, ни *k-means*, ни сеть Кохонена не используют глобальную функцию оптимизации.

Выбор числа кластеров.

Редко, но такие случаи бывают, когда точно известно, сколько кластеров нужно выделить. Но чаще всего этот вопрос остается открытым перед процедурой кластеризации. Если алгоритм не поддерживает автоматическое определение оптимального количества кластеров, то здесь есть несколько эмпирических правил при условии, что каждый кластер будет в дальнейшем подвергаться содержательной интерпретации аналитиком.

- Два или три кластера, как правило, не достаточно – кластеризация будет слишком грубой, приводящей к потере информации об индивидуальных свойствах объектов.
- Больше десяти кластеров не укладывается в известное «числа Миллера 7÷2»: аналитику трудно держать в кратковременной памяти столько кластеров.
- Поэтому в подавляющем большинстве случаев число кластеров варьируется от 4 до 9.

Глядя на изобилие алгоритмов кластеризации, возникает вопрос, существует ли объективная, «естественная» кластеризация или она всегда носит субъективный характер? Не существует. Любая кластеризация субъективна, потому что она выполняется на основе конечного подмножества свойств объектов. Выбор этого подмножества – всегда субъективен, как и выбор критерия качества и меры близости.

Популярные алгоритмы *k-means* и сеть Кохонена изначально разрабатывались для числовых данных, и, хотя впоследствии появились их модификации применительно к смешанным наборам данных, они все равно лучше решают задачи кластеризации на числовых признаках.

Для **корректного** применения кластеризации и снижения риска получения результатов моделирования, не имеющего никакого отношения к реальной действительности, необходимо следовать следующим правилам.

Правило 1. Перед кластеризацией четко обозначьте цели ее проведения: облегчение дальнейшего анализа, сжатие данных и т.п. Кластеризация сама по себе не представляет особой ценности.

Правило 2. Выбирая алгоритм, убедитесь в том, что он корректно работает с теми данными, которыми вы располагаете для кластеризации. В частности, если присутствуют категориальные признаки, удостоверьтесь: умеет ли та реализация алгоритма, которую вы используете, правильно обрабатывать их. Это особенно актуально для алгоритмов *k-means* и сетей Кохонена (впрочем, и других, основанных на метриках), которые в большинстве случаев используют евклидову меру расстояния. Если алгоритм не умеет работать со смешанными наборами данных, постарайтесь сделать его однородным, т.е. *отказаться* от категориальных или числовых признаков.

Правило 3. После кластеризации обязательно проведите содержательную интерпретацию каждого кластера: постарайтесь понять, *почему* объекты были сгруппированы в каждый кластер, что их объединяет? Для этого можно использовать визуальный анализ, графики, кластерограммы, статистические характеристики кластеров, многомерные карты. Полезно каждому кластеру дать «емкое» название, состоящее из нескольких слов. Встречаются ситуации, когда алгоритм кластеризации не выделил никаких особых групп. Возможно, набор данных и до кластеризации был однороден, не расслаивался на изолированные подмножества, а кластеризация подтвердила эту гипотезу.

Таким образом, не существует единого универсального алгоритма кластеризации. Поэтому при использовании любого алгоритма важно понимать его достоинства, недостатки и ограничения. Только тогда кластеризация будет эффективным инструментом в руках аналитика.