Введение в кластерный анализ, алгоритм k-means

Кластеризация — одна из задач Data Mining, кластер – группа похожих объектов.

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 є X ставит в соответствие метку кластера y є 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.

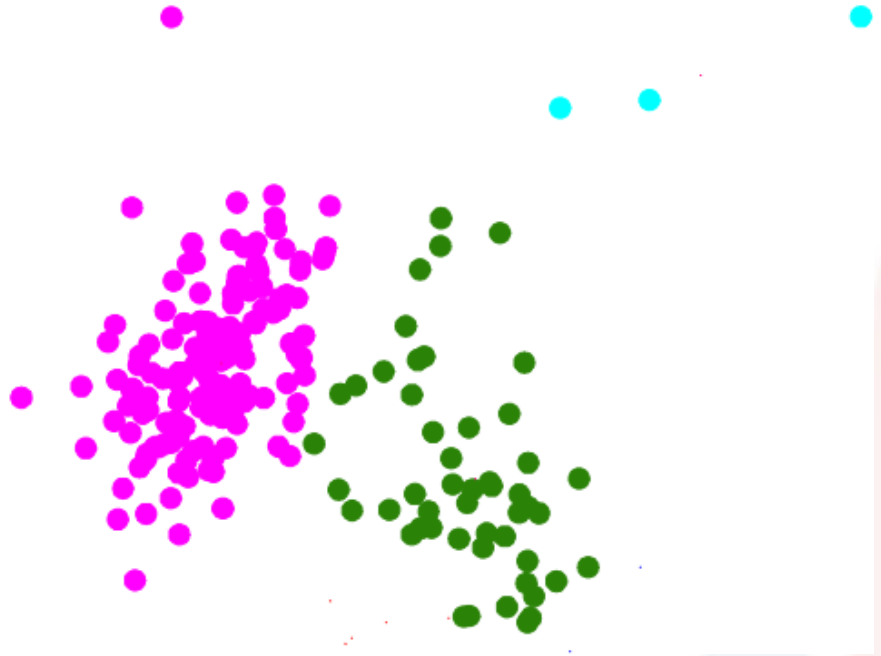
Кластеризация — одна из задач Data Mining, кластер – группа похожих объектов.

Определение:

Кластеризация — 1) группировка объектов на основ близости их свойств; каждый кластер состоит из схо объектов, а объекты разных кластеров существенно отличаются; 2) процедура, которая любому объекту X ставит в соответствие метку кластера y ∈ 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 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. Естественно, приведенный вариант разбиения не является единственным

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.

№	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.

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.**

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.

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 прежде всего алгоритм **k-means**

Метрики в кластерном анализе

- ***Меры, основанные на расстоянии:*** Евклидово (Euclidian), Манхэттена и Канберра (Manhattan & Canberra), Махаланобиса (Mahalanobis)
- ***Меры, основанные на корреляции:*** Коэффициент корреляции Пирсона (Pearson product-moment correlation), Коэффициент ранговой корреляции Спирмена (Spearman rank correlation)
- ***Информационно-теоретические:*** Расстояние Хэмминга (Hamming distance) для категориальных данных

Метрики в кластерном анализе

Евклидово расстояние

По умолчанию R измеряет расстояния Евклида между всеми строками объектов. Для двух образцов X и Y с количеством в N i-х признаков:

$$d_{Euc}(X,Y) = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2}$$

Расстояние Манхэттена

Расстояние Манхэттена (или «расстояние городских кварталов») похоже на перемещение по координатной сетке:

$$d_{Manh}(X,Y) = \sum_{i=1}^{N}|x_i - y_i|$$

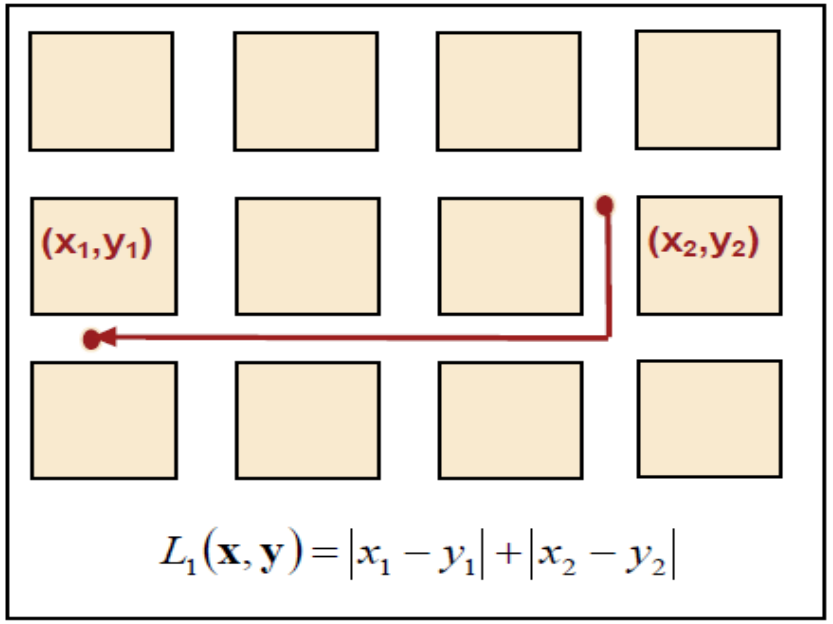
Расстояние Канберра

Расстояние Канберра – это расстояние Манхэттена, масштабированное по величинам X и Y (используется не часто, не входит в пакет *bioDist*):

$$d_{Cunb}(X,Y) = \sum_{i=1}^{N}\frac{|x_i - y_i|}{|x_i| - |y_i|}$$

Метрики в кластерном анализе

- Множество точек, равноудаленных от некоторого центра при использовании евклидовой метрики будет образовывать сферу (или круг в двумерном случае), и кластеры, полученные с использованием **евклидова расстояния**, также будут иметь форму, близкую к **сферической**.
- Преимущество метрики ***Расстояния Манхэттена*** заключается в том, что ее использование позволяет снизить влияние аномальных значений на работу алгоритмов. Кластеры, построенные на основе расстояния Манхэттена, стремятся к **кубической форме**.



Одной из широко используемых методик кластеризации является разделительная кластеризация (англ.: partitioning clustering) в соответствии с которой для выборки данных, содержащей n записей (объектов), задается число кластеров k , которое должно быть сформировано. Затем алгоритм разбивает все объекты выборки на k разделов (k < n), которые и представляют собой кластеры.

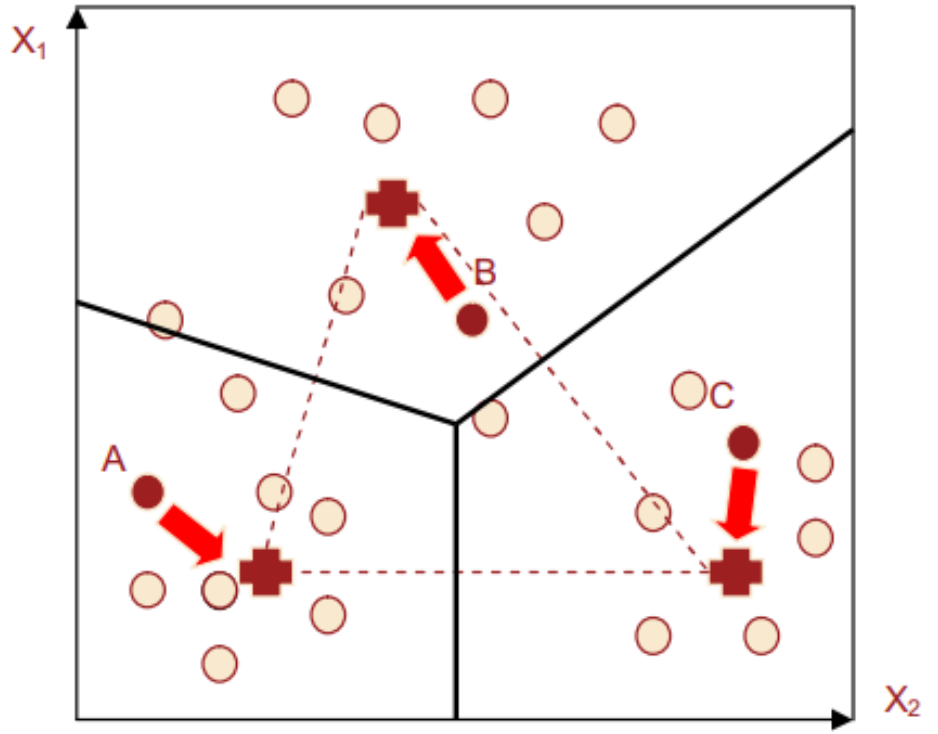
Одним из наиболее простых и эффективных алгоритмов кластеризации является алгоритм k-means (Mac-Queen, 1967) или в русскоязычном варианте k-средних (от англ. mean – среднее значение). Он состоит из четырех шагов.

Алгоритм кластеризации k-means

1. Задается число кластеров k, которое должно быть сформировано из объектов исходной выборки.
2. Случайным образом выбирается k записей исходной выборки, которые будут служить начальными центрами кластеров. Такие начальные точки, из которых потом «вырастает» кластер, часто называют «семенами» (от англ. seeds – семена, посевы). Каждая такая запись будет представлять собой своего рода «эмбрион» кластера, состоящий только из одного элемента.
3. Для каждой записи исходной выборки определяется ближайший к ней центр кластера.
4. Производится вычисление центроидов – центров тяжести кластеров. Это делается путем простого определения среднего для значений каждого признака для всех записей в кластере.

Например, если в кластер вошли три записи с наборами признаков , то координаты его центроида будут рассчитываться следующим образом: $(x_1, y_1), (x_2, y_2), (x_3, y_3)$

$$(x, y) = \left(\frac{(x_1 + x_2 + x_3)}{3}, \frac{(y_1 + y_2 + y_3)}{3} \right)$$



Затем старый центр кластера смещается в его центроид. На рисунке центры тяжести кластеров представлены в виде крестиков.

Таким образом, центроиды становятся новыми центрами кластеров для следующей итерации алгоритма. Шаги 3 и 4 повторяются до тех пор, пока выполнение алгоритма не будет прервано либо не будет выполнено условие в соответствии с некоторым критерием сходимости. Остановка алгоритма производится тогда, когда границы кластеров и расположения центроидов не перестанут изменяться от итерации к итерации, т.е. на каждой итерации в каждом кластере будет оставаться один и тот же набор записей. На практике алгоритм k-means обычно находит набор стабильных кластеров за несколько десятков итераций.

Что касается критерия сходимости, то чаще всего используется критерий суммы квадратов ошибок между центроидом кластера и всеми вошедшими в него записями, т.е.

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} (p - m_i)^2$$

где $p \in C_i$ – произвольная точка данных, принадлежащая кластеру C_i, m_i – центроид данного кластера. Иными словами, алгоритм остановится тогда, когда ошибка E достигнет достаточно малого значения.

Ключевым моментом алгоритма ***k-means*** является вычисление на каждой итерации расстояния между записями и центрами кластеров, что необходимо для определения, к какому из кластеров принадлежит данная запись. Правило, по которому производится вычисление расстояния в многомерном пространстве признаков, называется метрикой. В k-means применяется Евклидово расстояние (метрика L2). Использует для вычисления расстояний следующее правило:

$$d_E(x,y) = \sqrt{\sum_i (x_i - y_i)^2},$$

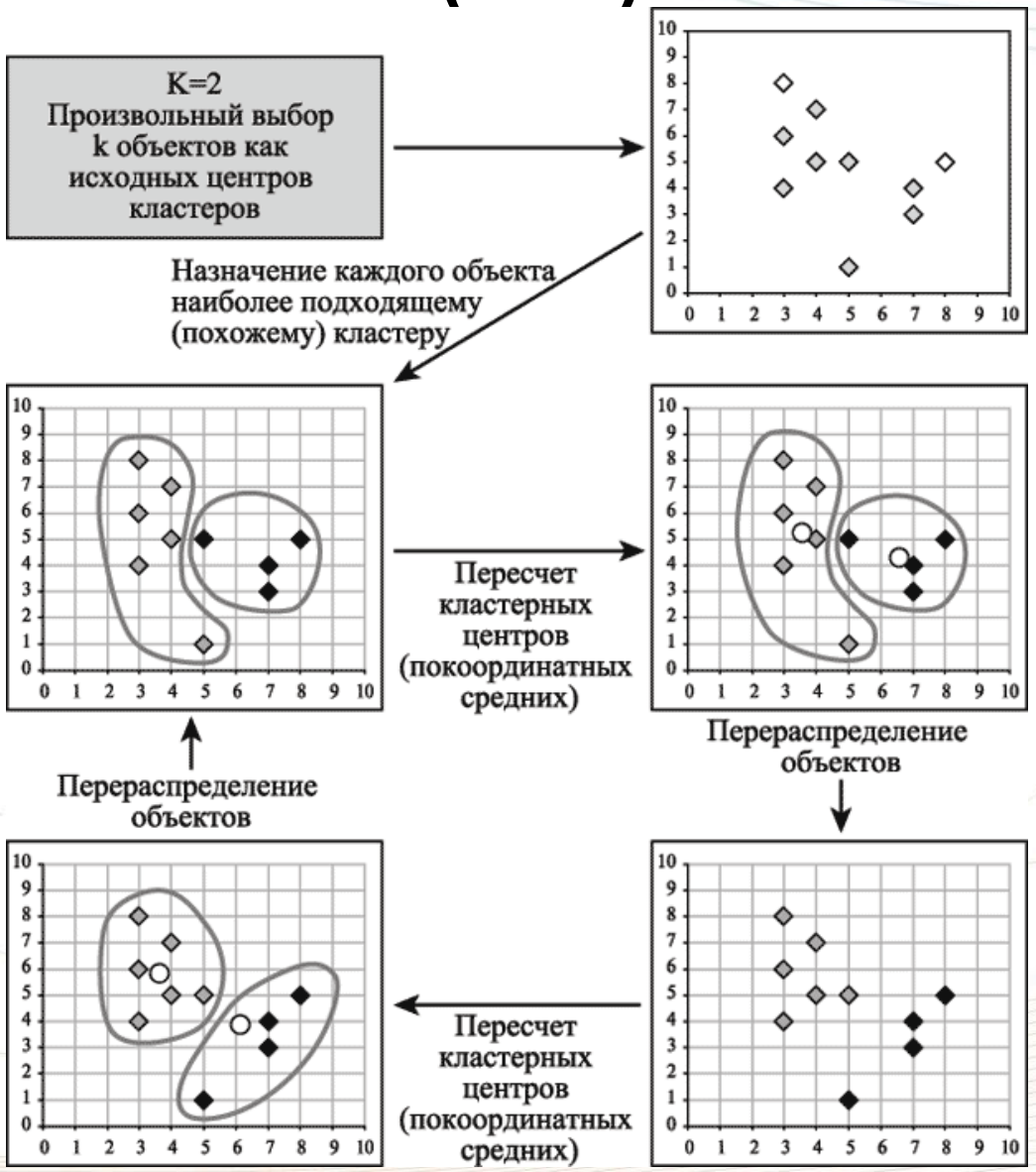
где $\mathbf{x} = (x_1, x_2, \ldots, x_m)$, $\mathbf{y} = (y_1, y_2, \ldots, y_m)$ – наборы (векторы) значений признаков двух записеи. Поскольку множество точек, равноудаленных от некоторого центра при использовании евклидовой метрики будет образовывать сферу (или круг в двумерном случае), то и кластеры, полученные с использованием евклидова расстояния, также будут иметь форму, близкую к сферической.

Свою популярность алгоритм **k-means** приобрел благодаря следующим свойствам. Во-первых, это умеренные вычислительные затраты, которые растут линейно с увеличением числа записей исходной выборки данных. Вычислительная сложность алгоритма определяется как **k·n·l** , где *k* – число кластеров, *n* – число записей и *l* – число итераций. Поскольку k и l заданы, то вычислительные затраты возрастают пропорционально числу записей исходного множества.

Полезным свойством **k-means** является то, что результаты его работы не зависят от порядка следования записей в исходной выборке, а определяются только выбором исходных точек. Одним из основных недостатков, присущих этому алгоритму, является ***отсутствие четких критериев выбора числа кластеров, целевой функции их инициализации и модификации.***

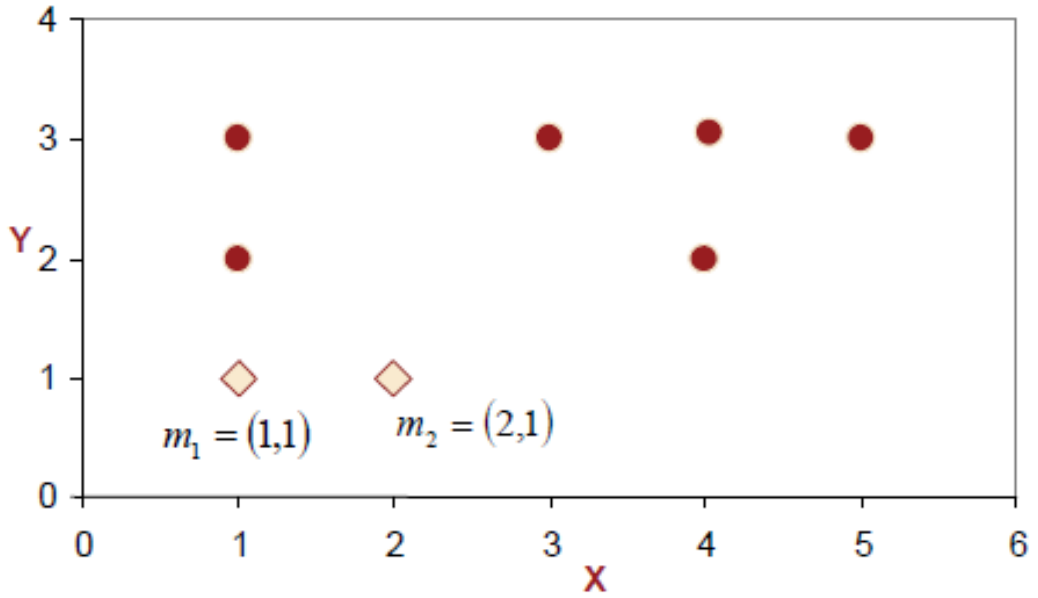
Кроме этого, он ***является очень чувствительным к шумам и аномальным значениям*** в данных, поскольку они способны значительно повлиять на среднее значение, используемое при вычислении положений центроидов. Чтобы снизить влияние таких факторов, как шумы и аномальные значения, иногда на каждой итерации используют не среднее значение признаков, а их медиану. Данная модификация алгоритма называется k-mediods (k-медиан).

Пример работы алгоритма k-средних (k=2)



Пример

A	B	C	D	E	F	G	H
(1,3)	(3,3)	(4,3)	(5,3)	(1,2)	(4,2)	(1,1)	(2,1)



Шаг 1. Определим число кластеров, на которое требуется разбить исходное множество k=2.

Шаг 2. Случайным образом выберем две точки, которые будут начальными центрами кластеров. Пусть это будут точки $m1$=(1;1) и $m2$=(2;1).

Шаг 3, проход 1. Для каждой точки определим к ней ближайший центр кластера с помощью расстояния Евклида.

Точка	Расстояние от m_1	Расстояние от m_2	Принадлежит кластеру
A	2,00	2,24	1
B	2,83	2,24	2
C	3,61	2,83	2
D	4,47	3,61	2
E	1,00	1,41	1
F	3,16	2,24	2
G	0,00	1,00	1
H	1,00	0,00	2

Таким образом, кластер 1 содержит точки A, E, G, а кластер 2 – точки B, C, D, F, H. Как только определятся члены кластеров, может быть рассчитана сумма квадратичных ошибок:

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} (p - m_i)^2 = 2^2 + 2{,}24^2 + 2{,}83^2 + 3{,}61^2 + 1^2 + 2{,}24^2 + 0^2 + 0^2 = 36.$$

Шаг 4, проход 1. Для каждого кластера вычисляется его центроид, и центр кластера перемещается в него. Центроид для первого кластера вычисляется как

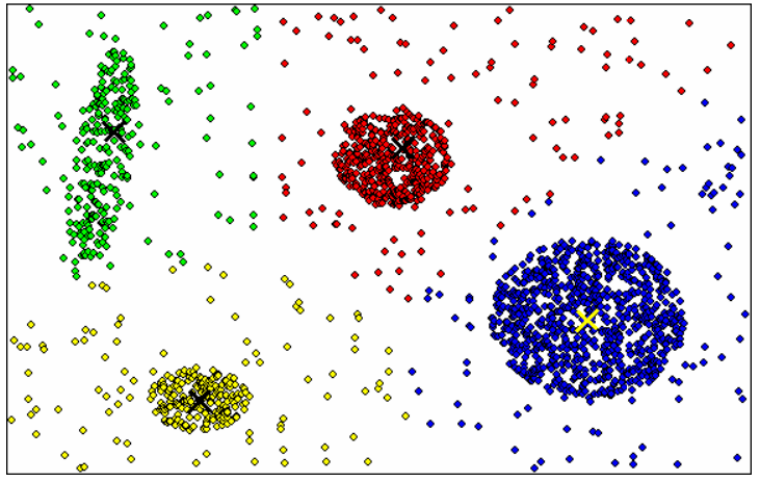
[(1+1+1)/3, (3+2+1)/3]=(1; 2).

Центроид для кластера 2 будет равен

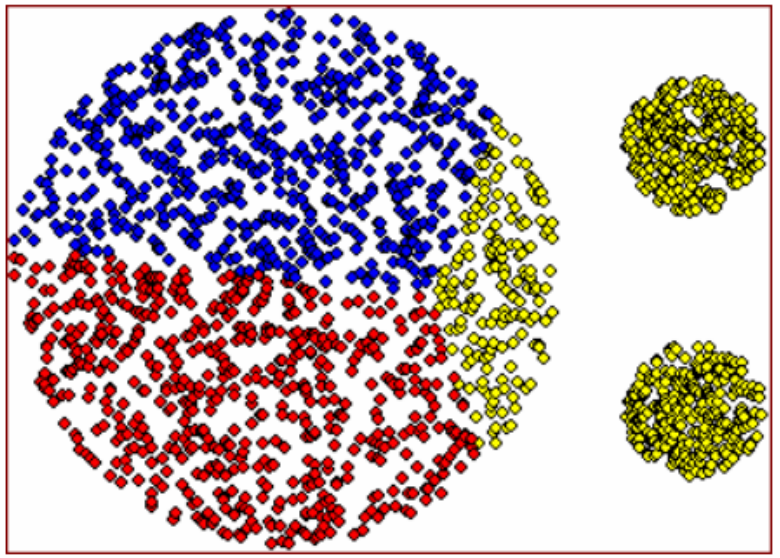
[(3 + 4 + 5 + 4 + 2)/5, (3 + 3 + 3 + 2 + 1)/5] = (3,6; 2,4).

Проблемы алгоритмов кластеризации

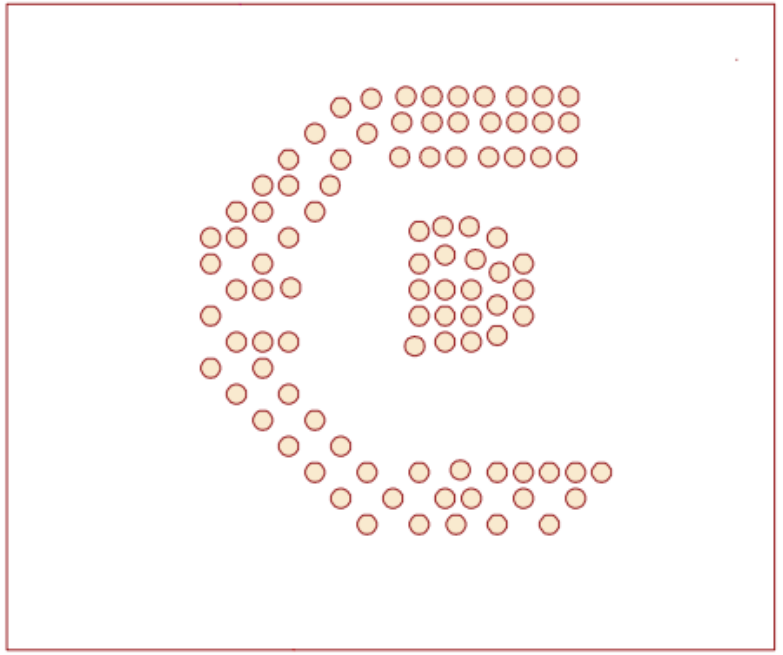
правильно



не правильно: а – эффект расщепления, б – вложенные кластеры



а



б

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. 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. б).

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.

sklearn.cluster.KMeans

class `sklearn.cluster.`**`KMeans`**(*n_clusters=8, *, init='k-means++', n_init=10, max_iter=300, tol=0.0001, precompute_distances='deprecated', verbose=0, random_state=None, copy_x=True, n_jobs='deprecated', algorithm='auto'*) [source]

H20

```
prostate =
h2o.import_file("http://s3.amazonaws.com/h2o-
public-test-data/smalldata/prostate/prostate.csv")
predictors = ["AGE", "RACE", "DPROS", "DCAPS",
"PSA", "VOL", "GLEASON"]
train, valid = prostate.split_frame(ratios=[.8],
seed=1234)
encoding = "one_hot_explicit"
pros_km =
H2OKMeansEstimator(categorical_encoding
=encoding, seed=1234)
pros_km.train(x=predictors,
        training_frame=train,
        validation_frame=valid)
pros_km.scoring_history()
```

Encoding scheme for categorical features
One
of: `"auto"`, `"enum"`, `"one_hot_internal"`, `"one_hot_explicit"`, `"binary"`, `"eigen"`, `"label_encoder"`, `"sort_by_response"`, `"enum_limited"` (default: `"auto"`).

Knime Analytics Platform

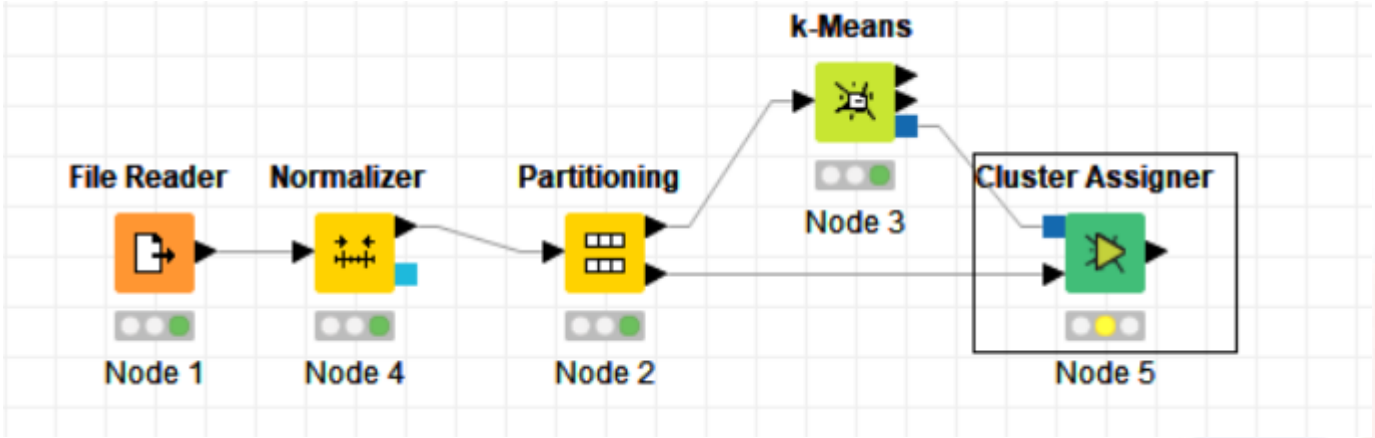


Table "default" - Rows: 8 | Spec - Columns: 6 | Properties | Flow Variables

Row ID	sepal_length	sepal_width	petal_length	petal_width	species	Cluster
Row8	0.028	0.375	0.068	0.042	setosa	cluster_1
Row30	0.139	0.458	0.102	0.042	setosa	cluster_1
Row35	0.194	0.5	0.034	0.042	setosa	cluster_1
Row79	0.389	0.25	0.424	0.375	versicolor	cluster_2
Row99	0.389	0.333	0.525	0.5	versicolor	cluster_2
Row108	0.667	0.208	0.814	0.708	virginica	cluster_2
Row137	0.583	0.458	0.763	0.708	virginica	cluster_2
Row146	0.556	0.208	0.678	0.75	virginica	cluster_2