


КЛАСТЕРИЗАЦИЯ

Науки делятся на две группы – на физику и собирание марок.
Э. Резерфорд

Технологии анализа данных

Содержание

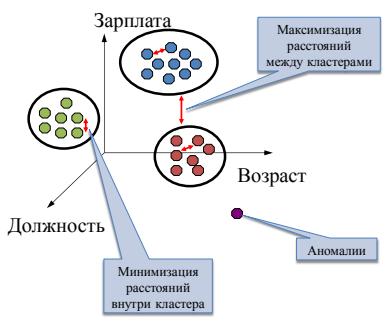
2

- Понятие кластеризации
- Виды данных для кластеризации
- Основные методы кластеризации

Кластеризация

3

- Разбиение заданного множества элементов на непересекающиеся подмножества (*кластеры*) таким образом, чтобы каждый кластер состоял из элементов, близких друг другу, а элементы разных кластеров существенно различались в смысле некоторой функции расстояния.
- Семантика кластеров заранее не известна.



Кластеризация

4

- Машинное обучение (Machine learning)
  - Обучение без учителя (unsupervised learning)
  - Классификация без учителя (unsupervised classification)
- Статистика
  - Классификация
- Маркетинг
  - Сегментирование

Кластеризация: примеры

5

- Уменьшение количества размеров военной формы, хранимой на складе, с сохранением возможности попадания в размер.
- Традиционная система размеров
  - упорядоченное множество градуированных размеров, в котором все измерения увеличиваются одновременно.
- Система размеров после кластеризации
  - размеры, соответствующие типам тела
    - например, один размер для женщин с ногами небольшой длины, руками средней длины, небольшим объемом талии, большим объемом бедер, большой длиной корпуса, широкими плечами и небольшой окружностью шеи.

Кластеризация: примеры

6

- Маркетинг
  - Определение различных групп покупателей для разработки соответствующих программ лояльности
- Биология
  - автоматизированный вывод таксономии животных
  - нахождение генов со сходными функциями
- Землепользование
  - определение территорий со схожими характеристиками землепользования в базе данных наблюдений о Земле
- Страхование
  - выявление групп держателей полисов автострахования с высокой средней стоимостью претензии
- Городское планирование
  - определение групп домов со сходными характеристиками (стоимость, тип, географическое положение и др.)

Требования к кластеризации

7

- Масштабируемость
- Возможность работы с атрибутами различных типов
- Нахождение кластеров произвольной формы
- Минимальные требования к знаниям о предметной области
- Устойчивость к шумам и аномалиям
- Нечувствительность к порядку входных данных
- Обработка данных, имеющих большую размерность
- Внедрение ограничений, определяемых пользователем
- Интерпретируемость и используемость

Виды атрибутов

8

- Интервальный
  - Вес, температура, доход
- Бинарный
  - Продукт высоких технологий
- Категория
  - Цвет={красный, желтый, зеленый}, пол={муж, жен}
- Порядковый
  - Ранг
- Соотношение (ratio-scaled)
  - Имеющие область значений вида $Ae^{\pm Bt}$ ($A,B>0$, t – обычно время)
- Смешанный
  - Комбинация вышеперечисленных

Представление данных

9

- Матрица данных

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

- Матрица сходства

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \ldots & \ldots & 0 \end{bmatrix}$$

Метрика

10

- Свойства
  - $d(i,j) \geq 0$
  - $d(i,i) = 0$
  - $d(i,j) = d(j,i)$
  - $d(i,j) \leq d(i,k) + d(k,j)$
- Разновидности
  - Евклидово расстояние
    $$d(i,j)=\sqrt{(|x_{i_1}-x_{j_1}|^2+|x_{i_2}-x_{j_2}|^2+...+|x_{i_p}-x_{j_p}|^2)}$$
  - Манхэттенское расстояние
    $$d(i,j)=|x_{i_1}-x_{j_1}|+|x_{i_2}-x_{j_2}|+...+|x_{i_p}-x_{j_p}|$$
  - Чебышёвское расстояние
    $$d(i,j)=\max\{|x_{i_1}-x_{j_1}|,|x_{i_2}-x_{j_2}|,...,|x_{i_p}-x_{j_p}|\}$$
  - Расстояние Минковского
    $$d(i,j)=\sqrt[q]{(|x_{i_1}-x_{j_1}|^q+|x_{i_2}-x_{j_2}|^q+...+|x_{i_p}-x_{j_p}|^q)}$$
  - Взвешенное расстояние
    $$d(i,j)=\sqrt[q]{w_1(|x_{i_1}-x_{j_1}|^q+w_2|x_{i_2}-x_{j_2}|^q+...+w_p|x_{i_p}-x_{j_p}|^q)}$$

Типы атрибутов кластеризуемых данных

11

- Бинарные
  - Пол, ПациентЖив
- Интервальные
  - Вес, Температура, Доход
- Номинальные категории
  - Цвет
- Порядковые категории
  - ВоинскоеЗвание
- Записи (комбинации вышеперечисленных)

Расстояние между бинарными атрибутами

12

- Виды бинарных атрибутов
  - Симметричные – оба значения одинаково важны: Пол.
  - Асимметричные – одно из значений более важно: ПациентЖив.
- Таблица сопряженности

		j = 1	j = 0	SUM
i	1	q	r	$q+r$
	0	s	t	$s+t$
	SUM	$q+s$	$r+t$	p

- Расстояния
  - Симметричные атрибуты
    $$d(i,j)=\frac{r+s}{q+r+s+t}$$
  - Асимметричные атрибуты
    $$d(i,j)=\frac{r+s}{q+r+s}$$

Пример

13

ФИО	Пол	Жар	Кашель	Тест-1	Тест-2	Тест-3	Тест-4
Иванов	М	Да	Нет	+	-	-	-
Петрова	Ж	Да	Нет	+	-	+	-
Сидоров	М	Да	Да	-	-	-	-

- Симметричные атрибуты:
  - Пол
- Несимметричные атрибуты
  - Жар, Кашель, Тест-1, …-4

$$d(Иванов, Петрова) = \frac{0+1}{2+0+1} = 0.33$$

$$d(Иванов, Сидоров) = \frac{1+1}{1+1+1} = 0.67$$

$$d(Сидоров, Петрова) = \frac{1+2}{1+1+2} = 0.75$$

Расстояние между номинальными атрибутами

14

- Способ 1:
  - Заменить НомАтр={Зн1,Зн2,…,ЗнN} на БинАтр1, БинАтр2,…,БинАтрN
- Способ 2:
  - $d(i,j) = \frac{p-m}{p}$

  p – количество номинальных атрибутов, m – количество совпадений (атрибутов, в которых объекты i и j имеют одно и то же состояние)

ОИД	Атр Категория
1	A
2	B
3	C
4	A

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

Расстояние между порядковыми атрибутами

15

- Заменить каждое значение порядкового атрибута x_{if} объекта i его номером $r_{if} \in \{1,\ldots,M_f\}$
- Нормализовать значение r_{if} (привести к [0;1])

  $z_{if} = \frac{r_{if}-1}{M_f-1}$

- Вычислить расстояние, используя z_{if} и техники для интервальных атрибутов

ОИД	Атр Порядковый
1	Отл.
2	Удовл.
3	Хор.
4	Отл.

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1 & 0.5 & 0 \end{bmatrix}$$

Расстояние между атрибутами-соотношениями

16

- Обращаться так же, как с интервальными атрибутами – не лучший выход, поскольку высока вероятность искажений.
- Способ 1:
  - Логарифмическая трансформация $y_{if} = log(x_{if})$
- Способ 2:
  - Рассматривать значения как непрерывные порядковые и брать порядковый номер

ОИД	Атр Соотношение	*log(Amp)*
1	445	2.65
2	22	1.34
3	164	2.21
4	1210	3.08

$$\begin{bmatrix} 0 & & & \\ 1.31 & 0 & & \\ 0.44 & 0.87 & 0 & \\ 0.43 & 1.74 & 0.87 & 0 \end{bmatrix}$$

Расстояние между смешанными атрибутами

17

- Взвешенная формула, учитывающая влияние каждого атрибута:

$$d(i,j) = \frac{\sum_{f=1}^{P} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{P} \delta_{ij}^{(f)}}$$

- Вычисление $\delta_{ij}^{(f)}$
  - 0, если
    - либо $x_{if}=x_{jf}$ и f – ассиметричный бинарный атрибут
    - либо x_{if} или x_{jf} отсутствует;
  - 1, иначе
- Вычисление $d_{ij}^{(f)}$ для различных видов f
  - интервальный:

$$d^{(f)}_{ij} = \frac{|x_{if} - x_{jf}|}{max_h\{x_{hf}\} - min_h\{x_{hf}\}}$$

    - индекс h пробегает все присутствующие объекты с атрибутом f
  - бинарный или номинальный: $d_{ij}^{(f)}=0$, если $x_{if}=x_{jf}$, иначе $d_{ij}^{(f)}=1$
  - порядковый: вычислить r_{if} и z_{if} , далее брать z_{if} и обращаться как с интервальным
  - соотношение:
    - логарифмическая трансформация, далее обращаться как с интервальным
    - обращаться как с непрерывным порядковым атрибутом: вычислить r_{if} и z_{if} , далее брать z_{if} и обращаться как с интервальным

Пример

18

ОИД	Атр1 Категория	Атр2 Порядковый	Атр3 Соотношение	*log(Amp)*
1	A	Отл.	445	2.65
2	B	Удовл.	22	1.34=min_h
3	C	Хор.	164	2.21
4	A	Отл.	1210	3.08=max_h

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1 & 0.5 & 0 \end{bmatrix} \begin{bmatrix} 0 & & & \\ 1.31 & 0 & & \\ 0.44 & 0.87 & 0 & \\ 0.43 & 1.74 & 0.87 & 0 \end{bmatrix} \begin{bmatrix} 0 & & & \\ 0.75 & 0 & & \\ 0.25 & 0.50 & 0 & \\ 0.25 & 1.00 & 0.50 & 0 \end{bmatrix} \Rightarrow \begin{bmatrix} 0 & & & \\ 0.92 & 0 & & \\ 0.58 & 0.67 & 0 & \\ 0.08 & 1.00 & 0.67 & 0 \end{bmatrix}$$

Основные методы кластеризации

19

- *Разделительные (partitioning)*
  - Разбиение исходного множества векторов на кластеры (в каждом кластере имеется, по крайней мере, один объект и каждый объект принадлежит в точности одному кластеру);
    итеративное перемещение объектов между кластерами с целью улучшить начальное разбиение (чтобы объекты из одного кластера были более "близкими", а из разных кластеров – более "далекими" друг другу)
  - k-Means, k-Medoids, k-Median, k-Mode

Основные методы кластеризации

20

- *Иерархические*
  - Последовательное иерархическое разбиение исходного множества объектов
    - *Агломеративный (снизу-вверх)* подход
      - предполагается, что каждый исходный объект образует отдельный кластер, и затем выполняется слияние близких друг к другу объектов или кластеров до тех пор, пока не будет получен единственный кластер или не будет выполнено условие завершения слияния
      - Представитель: AGNES
    - *Дивизимный (снизу вверх)* подход
      - предполагается, что все исходные объекты входят в один кластер, и затем итеративно выполняется его разбиение на менее мощные кластеры до тех пор, пока не будут получены кластеры-синглтоны или не будет выполнено условие завершения слияния
      - Представитель: DIANA

Основные методы кластеризации

21

- *Плотностные (density-based)*
  - Добавление объектов (*точек*) в кластер до тех пор, пока *плотность* (количество) соседних точек не превысит некоторого заданного порога концентрации. В окрестности каждой точки кластера должно находиться некоторое минимальное количество других точек.
  - DBSCAN, OPTICS, DENCLUE
- *Решеточные (grid-based)*
  - Разбиение пространства исходных данных на конечное число ячеек, формирующих решеточную структуру, над которой выполняются операции, необходимые для кластеризации.
  - STING, WaveCluster

Основные методы кластеризации

22

- *Модельные*
  - Основываются на предположении, что данные генерируются с помощью некоторого набора функций распределения вероятностей.
  - EM
- *Нечеткие*
  - Вектор может одновременно принадлежать всем кластерам, но в разной степени; сумма степеней принадлежности равна 1.
  - Fuzzy C-Means.

Разделительный подход: k-means

23

- Вход: D – множество из n объектов, $k>0$
- Выход: k кластеров
- Алгоритм:

взять k случайных объектов в качестве центров кластеров;
repeat
(пере)присвоить каждому объекту номер наиболее похожего кластера на основе среднего значения объектов в кластере;
обновить средние значения в кластерах;
until нет изменений;

Критерии останова k-means

24

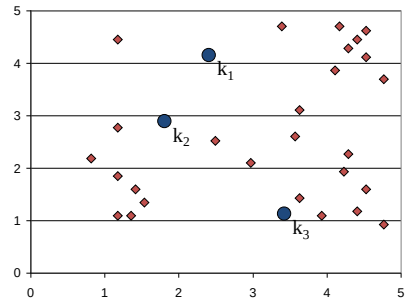
- Отсутствие или min значение переприсваиваний объектов кластерам
- Отсутствие или min изменение в центроиде
- Min уменьшение SSE (sum of squared error)

$$SSE = \sum_{j=1}^{k} \sum_{x \in C_j} dist(x, m_j)^2$$

  - C_j – j-й кластер
  - m_j – центроид кластера C_j (вектор со средними по всем объектам C_j координатами)
  - $dist(x, m_j)$ – расстояние между объектом x и центроидом m_j

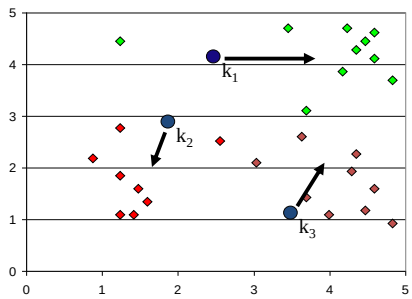
Пример: 3-means, шаг 1

25



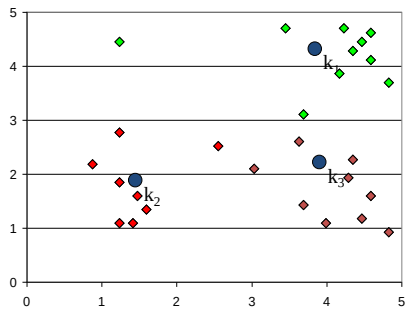
Пример: 3-means, шаг 2

26



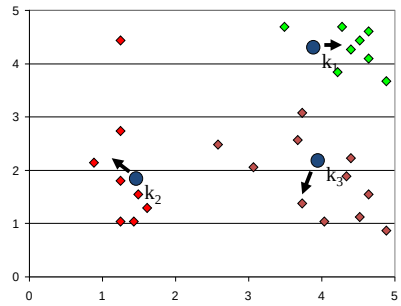
Пример: 3-means, шаг 3

27



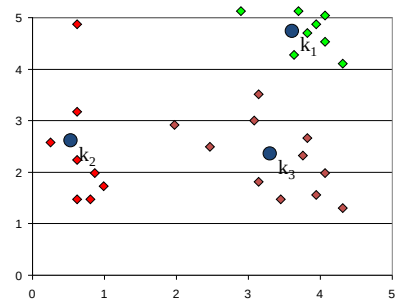
Пример: 3-means, шаг 4

28



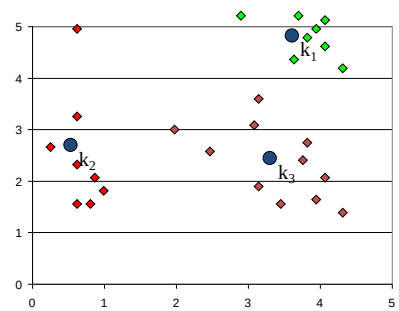
Пример: 3-means, шаг 5

29



Пример: 3-means, шаг 6

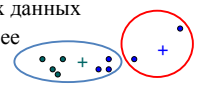
30



k-means: плюсы и минусы

31

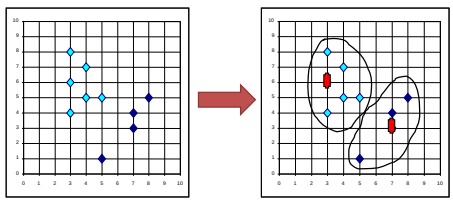
- Плюсы
  - Простота реализации
  - Относительная эффективность: *O(tkn)*, где *n* – количество объектов, *k* – количество кластеров, *t* – число итераций. Обычно *k* и *t* << *n*.
  - Часто завершается нахождением лок. оптимума
- Минусы
  - Применим только для интервальных данных
  - Значение *k* необходимо задать заранее
  - Не обрабатывает шумы и аномалии
  - Не способен находить невыпуклые кластеры



Методы k-medoid

32

- *Медоид* – объект исходного множества, который может представлять кластер.



- PAM (Partitioning Around Medoids), CLARA (Clustering LARge Applications)

Алгоритм PAM

33

взять *k* случайных объектов в качестве медоидов;

repeat

(пере)назначить каждому объекту кластер с ближайшим медоидом;

случайным образом выбрать объект не медоид;

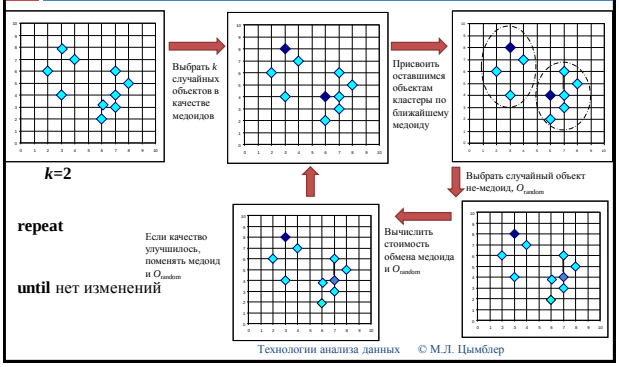
подсчитать стоимость обмена местами объекта и медоида;

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} d(p, o_i)^2$$

if $E<0$ **then** обменять местами объект и медоид

until нет изменений

Пример: PAM



***k*=2**

repeat

until нет изменений