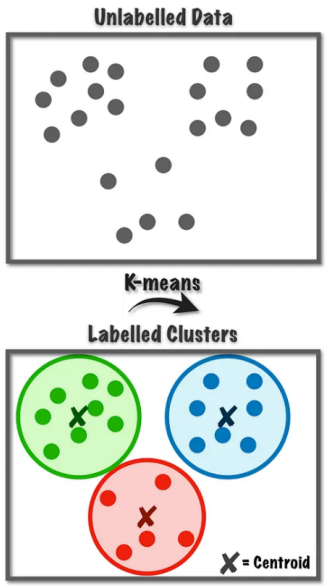
Продвинутая кластеризация

Виктор Китов

victorkitov.github.io



Курс поддержан фондом 'Интеллект'

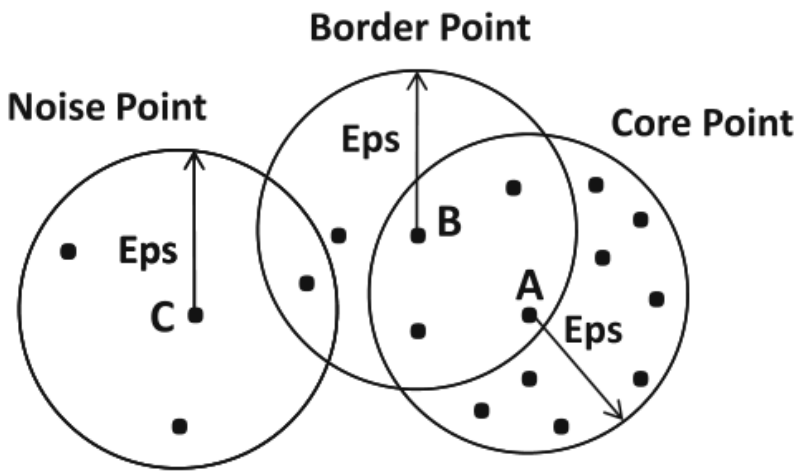
Победитель конкурса VK среди курсов по IT

Содержание

DBScan

k, ε - параметры метода.
Разделим множество объектов на 3 категории:

- основные точки: имеющие$\geq k$ точек внутри ε-окрестности
- пограничные точки: не основные, но содержащие хотя бы одну основную внутри ε-окрестности
- шумовые точки: не основные и не пограничные



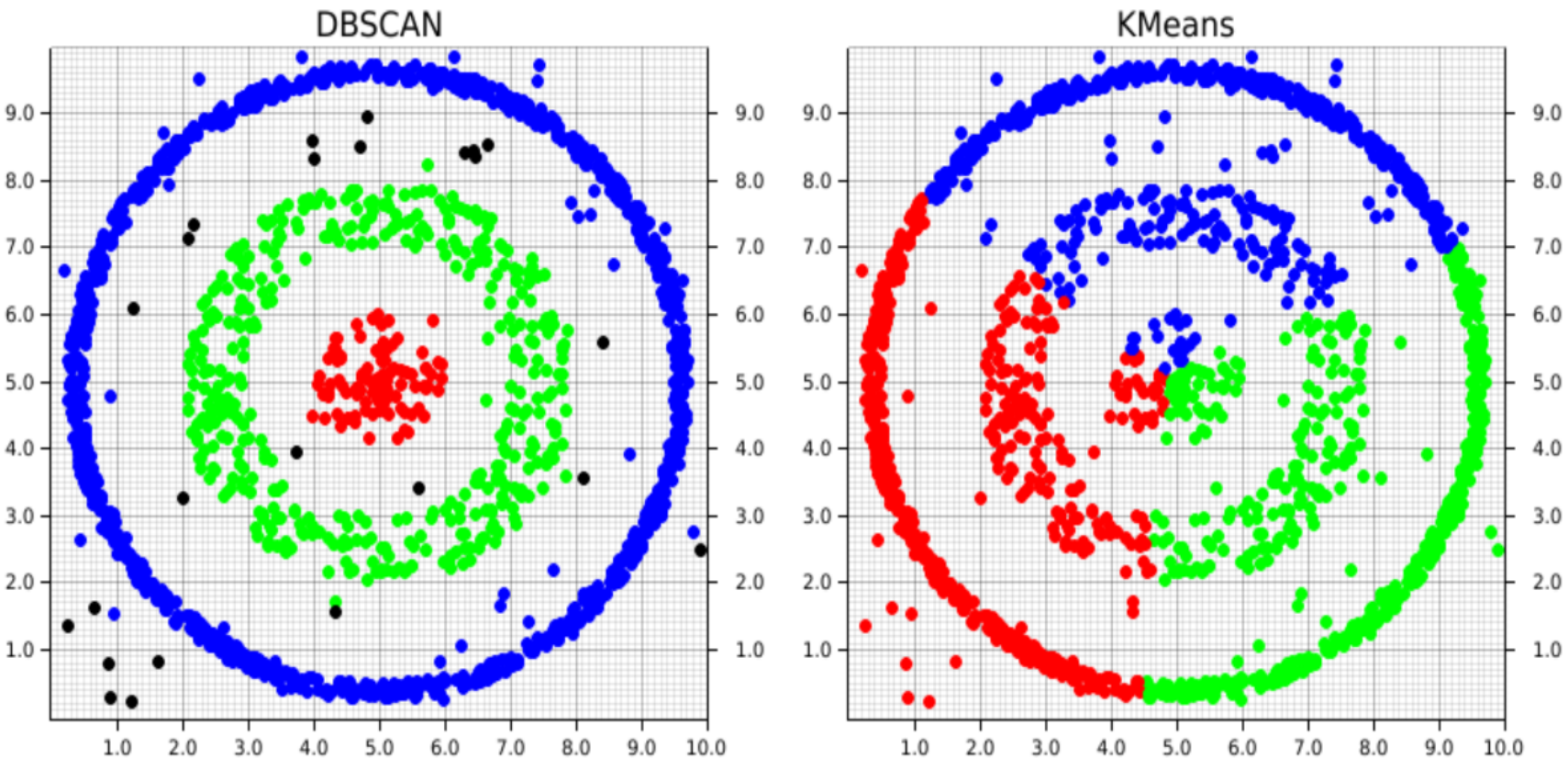
Алгоритм

ВХОД: выборка, параметры ε, k.

1) Определить основные/пограничные/шумовые точки, используя ε, k.
2) Создать граф: узлы-основные точки, связи - если точки на расстоянии $\leq \varepsilon$ друг от друга.
3) Определить компоненты связности в графе =кластеры (методом распространения).
4) Соотнести основные точки кластерам=компонентам связности, а пограничные-по основным в их ε окрестности.

ВЫХОД: разбиение на кластеры (основных и пограничных точек)

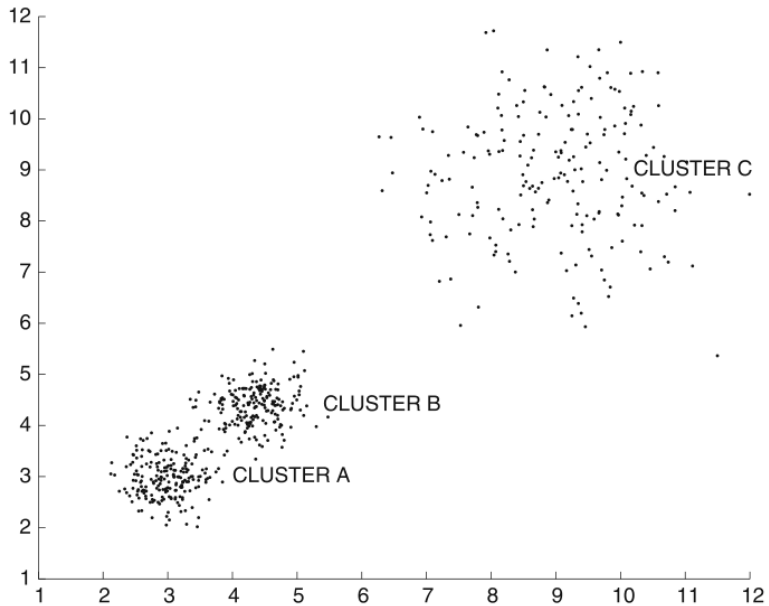
Пример работы DBScan[1]



[1] Источник иллюстрации.

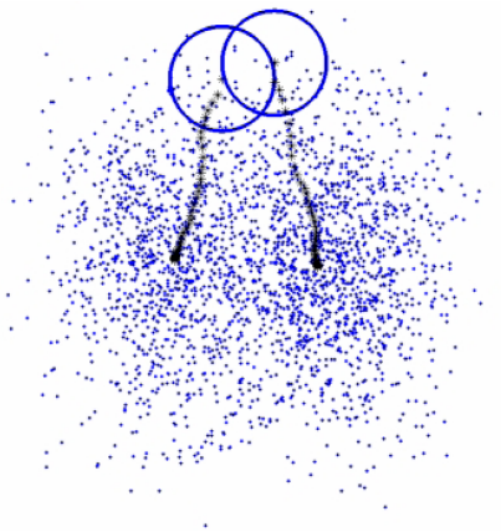
Комментарии

- Соединение основных точек - метод одиночной связи в агломеративной кластеризации с остановкой $\rho > \varepsilon$.
- Преимущества: автоматически определяется # кластеров, устойчиво к выбросам.
- Недостаток: не работает с кластерами разной плотности
  - высокое k-пропустим C; низкое k-A и B объединяться:



Кластеризация сдвигом среднего значения

Кластеризация сдвигом среднего значения (mean shift): точки итеративно сдвигаются в направлении локального увеличения плотности по правилу



Пример сходимости для top-hat ядра $K = \mathbb{I}\left[\frac{\rho(z,x)}{h} \leq 1\right]$

Кластер - итоговый локальный максимум плотности (отбрасываем максимумы с $p(x) < \tau$).

Комментарии

- Правило сдвига:

$$z_0 = x_n, \quad z = \frac{\sum_{k=1}^{N} K(\rho(z_i, x_k)/h) x_k}{\sum_{k=1}^{N} K(\rho(z, x_k)/h)}$$

- Ядро $K(\cdot)$ - некоторая $\downarrow$ ф-ция (ядро).
- Пример: Гауссово ядро

$$K(\rho(x, x')/h) = e^{-\rho(x,x')^2/h^2}$$

- Преимущества:
  - автоматически определяется #кластеров, кластеры могут быть произвольной формы
- Недостаток: вычислительная сложность, нет фильтрации выбросов

Кластеризация mean shift

ВХОД: выборка $x_1, ...x_N$, ядро $K(\cdot)$, ширина окна h.

ДЛЯ $n = 1, ...N$:

$z_n := x_n$

ПОВТОРЯТЬ до сходимости:

$$z_n := \frac{\sum_{k=1}^{N} K(\rho(z_n, x_k)/h) x_k}{\sum_{k=1}^{N} K(\rho(z, x_k)/h)}$$

ассоциировать x_n пику z_n

Объединить почти одинаковые расположения пиков $z_1, ...z_N$.

ВЕРНУТЬ кластеры точек, отнесенных одинаковым пикам плотности.

Содержание

Мотивация иерархической кластеризации

- #кластеров K заранее неизвестно.
- Кластеризация обычно не плоская, а иерархическая с разными уровнями детализации:
  - сайты в интернете
  - книги в библиотеке
  - животные в природе
- Подходы к иерархической кластеризации:
  - сверху вниз
    - более естественное для людей
  - снизу вверху (аггломеративная кластеризация)

Алгоритм

ВХОД:
выборка объектов, алгоритм плоской кластеризации A,
правила выбора листа и остановки

инициализировать дерево корнем, содержащим все объекты

ПОВТОРЯТЬ
    выбрать лист L по правилу выбора листа
    используя A разбить L на кластеры $L_1, ...L_K$
    добавить листы к T, соответствующие $L_1, ...L_K$
ПОКА выполнено условие остановки

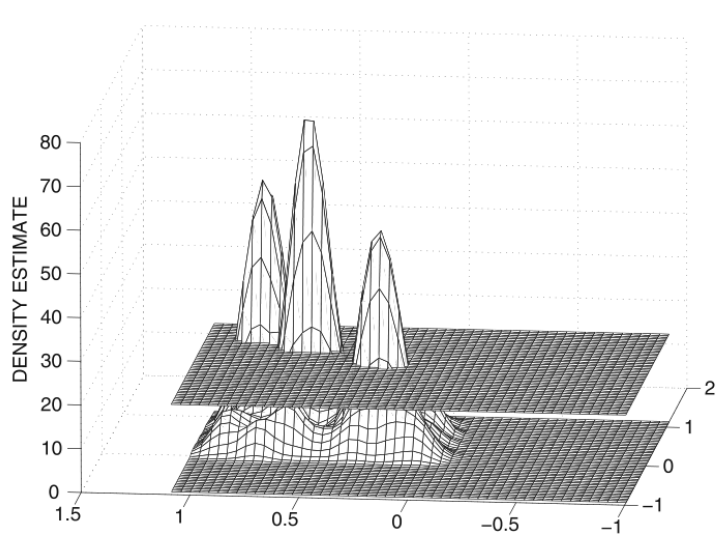
Комментарии

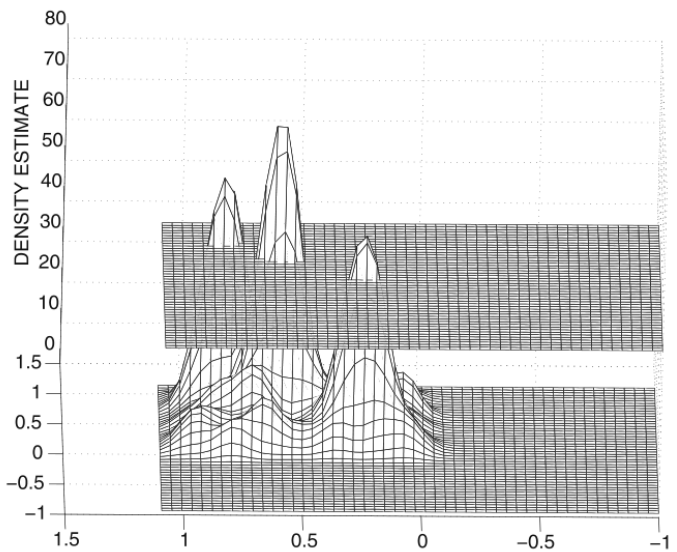
- Алгоритм выбора листа:
  - ближайший к корню
    => сбалансированное дерево по высоте
  - с максимальным числом элементов
    => сбалансированное дерево по #объектов в листах

2 Иерархическая кластеризация
- Иерархическая кластеризация сверху вниз
- Иерархическая кластеризация снизу вверх

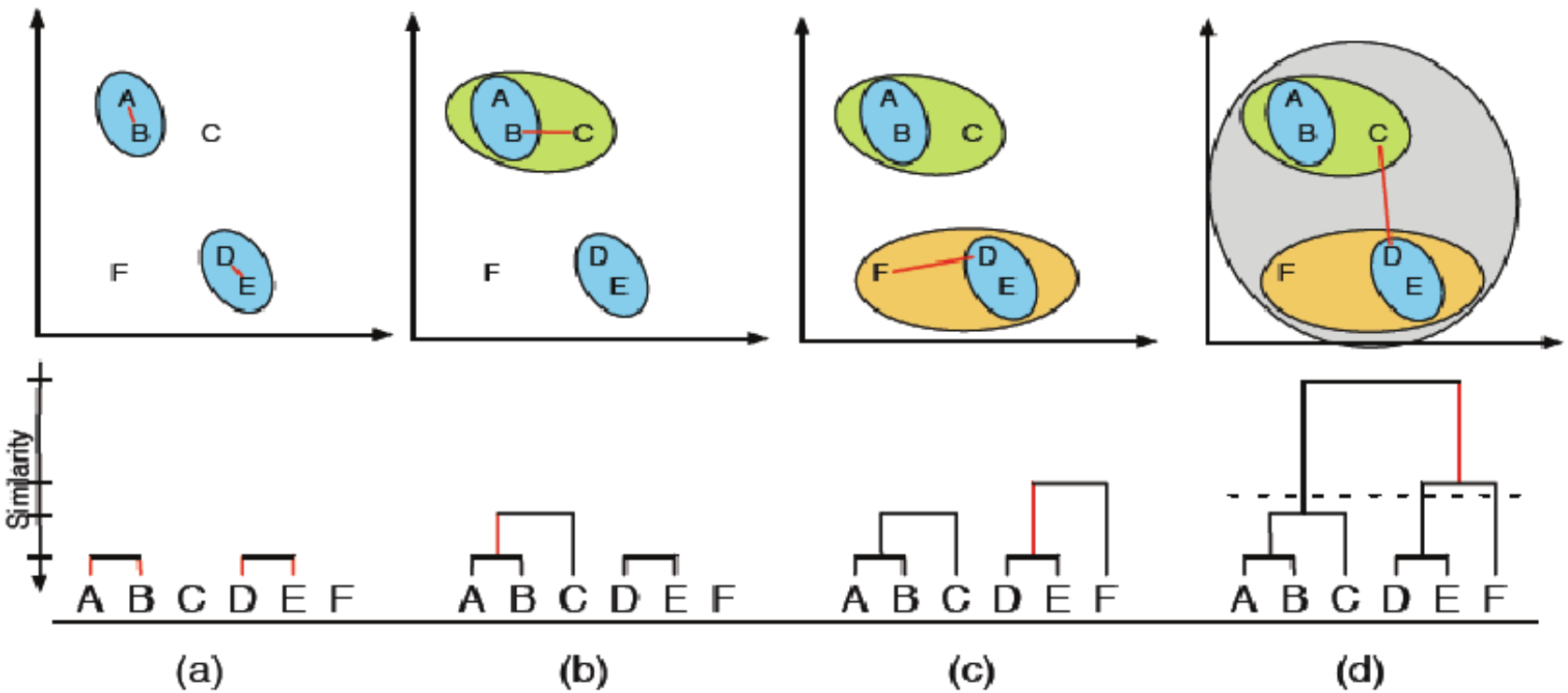
DENCLUE - иерахическое обобщение mean shift

1. Производим кластеризацию методом mean shift.
2. Объединяем кластеры с пиками, соединяемые цепочкой высоко вероятных значений плотности $p(x_{i(k)}) \geq h$.
   - варьируя h получаем иерархическую кластеризацию





Аггломеративная кластеризация - идея



Аггломеративная кластеризация - алгоритм

инициализировать матрицу попарных расстояний $M \in \mathbb{R}^{NxN}$ между кластерами из отдельных объектов $\{x_1\}, ...\{x_N\}$

ПОВТОРЯТЬ:

1) выбрать ближайшие кластеры i и j
2) объединить $i, j \to \{i+j\}$
3) удалить строки/столбцы i, j из M
4) добавить строку/столбец для нового $\{i+j\}$

ПОКА не выполнено условие остановки

ВЕРНУТЬ иерархическую кластеризацию

- Условие остановки:
  - Остался 1 кластер либо осталось $\leq K$ кластеров
  - расстояние между ближайшими кластерами$\geq$порога.
- Частичное обучение: если часть классов известна - объединяем i и j, только если там представители одного класса.

Расстояние между кластерами

- Расстояние между объектами => расстояние между кластерами:
  - Метод одиночной связи (single linkage)

$$\rho(A,B) = \min_{a\in A, b\in B} \rho(a,b)$$

  - Метод полной связи (complete linkage)

$$\rho(A,B) = \max_{a\in A, b\in B} \rho(a,b)$$

  - Метод средней связи (group average link)

$$\rho(A,B) = \text{mean}_{\,a\in A, b\in B}\rho(a,b)$$

  - Центроидный метод (pair-group method using the centroid average)

$$\rho(A,B) = \rho(\mu_A, \mu_B)$$

где $\mu_U = \frac{1}{|U|}\sum_{x\in U} x$ или $m_U = median_{x\in U}\{x\}$

Свойства межкластерных расстояний[3]

- Метод одиночной связи
  - извлекает кластеры произвольной формы
  - может случайно объединить разные кластеры цепочкой выбросов
  - $M_{(i\cup j)k} = \min\{M_{ik}, M_{jk}\}$
- Метод полной связи
  - создает компактные кластеры
  - $M_{(i\cup j)k} = \max\{M_{ik}, M_{jk}\}$
- Метод средней связи[2] и центроидный метод-компромисс между одиночной и полной связью.

[2] Как $M_{(i\cup j)k}$ будет пересчитываться для него?

[3] Пусть мы модифицируем $\rho(x, x')$ монотонным преобразованием F: $\rho'(x, x') = F(\rho(x, x'))$. При каких межкластерных расстояниях результат не изменится?

Свойства межкластерных расстояний

Метод средней связи предпочтительнее центроидного, поскольку

- центроидный метод может приводить к немонотонной последовательности расстояний дендрограммы.
  - методы одиночной, полной и средней связи дают монотонную последовательность
- представление кластера его центром не учитывает структуру кластера
- центроидный метод предпочитает более крупные кластера, для которых центроиды получаются в среднем ближе

Сложность аггломеративной кластеризации

- Сложность кластеризации K объектов: $O\left(K^3\right)$
  - K^2 для поиска ближайших K раз.
  - $O(K^2 \ln K)$ через алгоритм кучи
- Для снижения вычислений:
  1. применим K средних к N объектам (сложность $O(N)$)
  2. применим аггломеративную кластеризацию к найденным K кластерам
     - она позволяет выделять невыпуклые кластера

Заключение

- Плоская кластеризация:
  - K представителей
    - μ_k - вычисляемый (среднее: K-means [доступно ядерное обобщение], медиана: K medians)
    - μ_k - существующий объект
  - Основанная на плотности
    - DB-scan, mean-shift, DENCLUE
- Иерархическая кластеризация
  - сверху-вниз: рекурсивная плоская кластеризация
  - снизу-вверх (агломеративная)