Обучение без учителя. Разделение смеси распределений. Кластеризация.

Федяев И. Понизова В.

Обучение без учителя (Unsupervised learning) — раздел машинного обучения, в котором изучается класс задач обработки данных, в которых известны только описания множества объектов (признаки объектов) из обучающей выборки, и требуется обнаружить внутренние зависимости, существующие между объектами.

Типы задач обучения без учителя:

- Кластеризация
- Поиск ассоциативных правил
- Заполнение пропущенных значений
- Сокращение размерности
- Визуализация данных

Пусть имеется подмножество $X \subset \mathbb{R}^n$, которое мы будем называть пространством объектов.
Некоторый конечный набор $X^m = \{x_1, \ldots, x_m\}$ — обучающая выборка.
$\rho : X \times X \to [0, \infty)$ — функция расстояния между объектами.
Необходимо найти множество кластеров Y и алгоритм кластеризации $a : X \to Y$ такие, что:

- каждый кластер состоит из близких объектов (относительно ρ);
- объекты разных кластеров различались существенно.

Некорректность

Решение задачи кластеризации принципиально неоднозначно:

- точной постановки задачи кластеризации нет;
- существует много критериев качества кластеризации;
- существует много эвристических методов кластеризации;
- число кластеров, как правило, не известно заранее;
- результат кластеризации сильно зависит от метрики ρ, выбор которой также не однозначен.

Цели кластеризации

- Упростить дальнейшую обработку данных
- Сократить объём хранимых данных
- Поиск выбросов
- Построить иерархию множества объектов

Алгоритмы кластеризации

- Статистические методы (model-based)
  - EM-алгоритм
  - k-means
- Эвристические методы
  - Иерархическая агломеративная кластеризация
  - FOREL
  - DBSCAN

Смесь распределений

Гипотеза: выборка X^m — случайна, независима и взята из смеси распределений, плотность которой $p(x) = \sum_{j=1}^{k} w_j p_j(x;\theta_j), \sum_{j=1}^{k} w_j = 1.$

$p_j(x;\theta_j)$ – плотность распределения j-го кластера с параметрами θ_j, w_j — априорная вероятность кластера j.

Предлагается, зная число кластеров k и вид плотностей p_j, оценить параметры w_j и θ_j, максимизируя логарифм функции правдоподобия

$$\ln \mathcal{L}(\{x_i\}; \{w_j\}; \{\theta_j\}) = \sum_{i=1}^{m} \ln \sum_{j=1}^{k} w_j p_j(x_i; \theta_j) \to \max_{\{w_j\},\{\theta_j\}}$$

при условии $\sum_{j=1}^{k} w_j = 1; w_j \geqslant 0$.

Шаг E (expectation)

Пусть $p(x, \theta_j)$ — плотность вероятности того, что объект x получен из j-ой компоненты смеси. По формуле условной вероятности:

$$p(x, \theta_j) = p(x)P(\theta_j \mid x) = w_j p_j(x; \theta_j).$$

Обозначим $g_{ij} = P(\theta_j \mid x)$. По формуле Байеса:

$$g_{ij} = \frac{w_j p_j(x_i; \theta_j)}{\sum_{s=1}^{k} w_s p_s(x_i; \theta_s)}.$$

Шаг M (maximization)

Максимизируем

$$\ln \mathcal{L}(\{x_i\};\{w_j\};\{\theta_j\}) = \sum_{i=1}^{m} \ln \sum_{j=1}^{k} w_j p_j(x_i;\theta_j) \to \max_{\{w_j\},\{\theta_j\}}$$

при условии $\sum_{j=1}^{k} w_j = 1$. Лагранжиан:

$$L(\{w_j\},\{\theta_j\};\{x_i\}) = \sum_{i=1}^{m} \ln \sum_{j=1}^{k} w_j p_j(x_i;\theta_j) - \lambda \left(\sum_{j=1}^{k} w_j - 1 \right).$$

Из равенства нулю производной по w_j следует

$$w_j = \frac{1}{m} \sum_{i=1}^{m} g_{ij}, \quad j = 1, \ldots, k.$$

Из равенства нулю производной по θ_j следует

$$\theta_j = \arg\max_{\theta} \sum_{i=1}^{m} g_{ij} \ln p(x_i; \theta), \quad j = 1, \ldots, k.$$

Случай нормальных плотностей

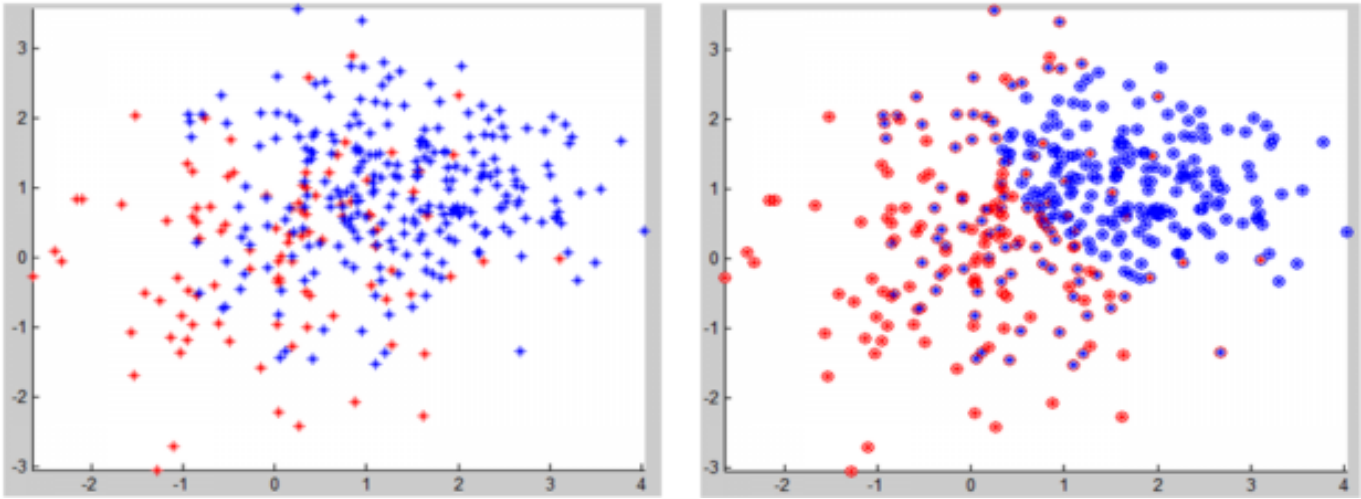
Пусть компоненты смеси имеют нормальные многомерные распределения со средними μ_j и матрицами ковариаций Σ_j, тогда имеем следующие оценки параметров

$$\mu_j = \frac{1}{mw_j} \sum_{i=1}^{m} g_{ij} x_i,$$

$$\Sigma_j = \frac{1}{mw_j} \sum_{i=1}^{m} g_{ij} (x_i - \mu_j)(x_i - \mu_j)^{\mathrm{T}}.$$

EM-алгоритм

1. Вычислить начальное приближение w_y, θ_y
2. **Повторять**
3. E-шаг: $g_{ij}^0 = g_{ij}$; $\quad g_{ij} = \frac{w_j p_j(x_i;\theta_j)}{\sum_{s=1}^{k} w_s p_s(x_i;\theta_s)}$.
4. M-шаг: $\theta_j = \arg\max_\theta \sum_{i=1}^{m} g_{ij} \ln p(x_i;\theta)$; $w_j = \frac{1}{m}\sum_{i=1}^{m} g_{ij}$.
5. **Пока** $\max_{i,j} |g_{ij} - g_{ij}^0| > \delta$.



(a) Реальные данные (b) Результат кластеризации

Рис.: Работа EM-алгоритма

Плюсы и минусы

Достоинства:

- Для этого алгоритма есть более-менее чётко поставленная задача;
- Нет необходимости масштабировать признаки.

Недостатки

- Алгоритм неустойчив по начальным данным (то есть тем, которые инициализируют вектора параметров w и θ на первой итерации);
- не позволяет определять количество k компонент смеси. Эта величина является структурным параметром алгоритма.

Алгоритм k-means

1. Сформировать начальное приближение центров кластеров;
2. **Повторять**
3. Отнести каждый объект к ближайшему центру (аналог E-шага);
4. Усреднить объекты в кластерах и получаем новое положение центров (аналог M-шага);
5. **Пока** состав кластеров не перестанет изменяться.

Алгоритм k-means можно получить из EM-алгоритма, если

- заменить подсчёт вероятностей g_{ij} принадлежности i-го объекта j-ому кластеру на жёсткое приписывание объекта к этому кластеру,
- ковариационные матрицы в нормальной модели ограничить только диагональными.

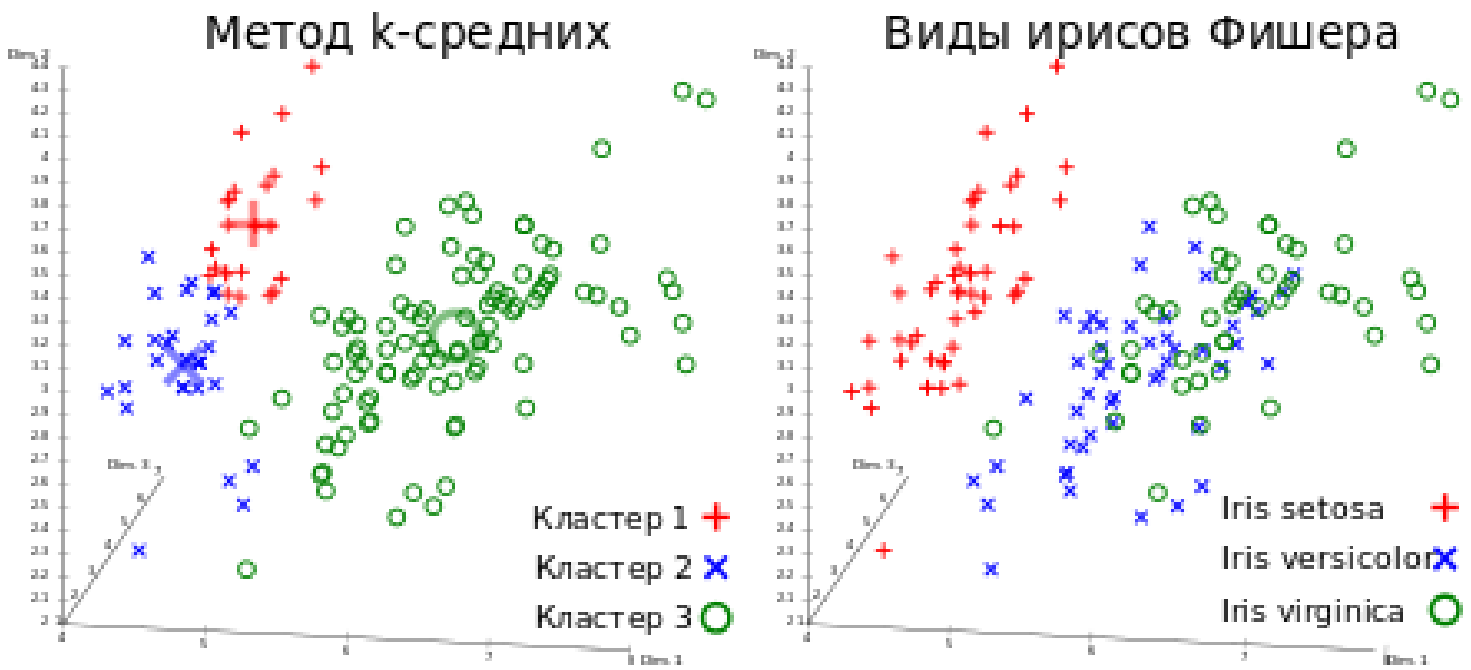


Рис.: Результат работы k-means

Достоинства:

- Алгоритм очень гибкий
- Простой

Недостатки:

- Кластеризация очень сильно зависит от начального приближения
- Кластеризация может быть неадекватной, если изначально было выбрано неверное число кластеров.
- Необходимость самостоятельно задавать число кластеров;
- Форма кластеров только сферическая.

Алгоритм Ланса-Уильямса

1. Инициализировать множество кластеров C_1:
   $t = 1; \quad C_t = \{\{x_1\}, \ldots \{x_m\}\}; R(\{x_i\}, \{x_j\}) = \rho(x_i, x_j);$
2. **Для всех** $t = 2, \ldots, m$
3. найти в C_{t-1} два ближайших кластера:
   $(U, V) = \arg\min_{U \neq V} R(U, V); R_t = R(U, V);$
4. слить их в один кластер:
   $W = U \cup V; C_t = C_{t-1} \setminus \{U, V\} \cup \{W\};$
5. **для всех** $S \in C_t$
6. вычислить расстояние $R(W, S)$
   по формуле Ланса-Уильямса.

Формула Ланса-Уильямса

$$R(W,S) = \alpha_U R(U,S) + \alpha_V R(V,S) + \beta R(U,V) + \gamma |R(U,S) - R(V,S)|,$$

где $\alpha_U, \alpha_V, \beta, \gamma$ — числовые параметры.

- Расстояние ближнего соседа:

$$R^{б}(W,S) = \min_{w\in W, s\in S} \rho(w,s);$$

$$\alpha_U = \alpha_V = 1/2, \quad \beta = 0, \quad \gamma = -1/2;$$

- Расстояние дальнего соседа:

$$R^{д}(W,S) = \max_{w\in W, s\in S} \rho(w,s);$$

$$\alpha_U = \alpha_V = 1/2, \quad \beta = 0, \quad \gamma = 1/2;$$

- Среднее расстояние:

$$R^{c}(W,S) = \frac{1}{|W||S|} \sum_{w\in W} \sum_{s\in S} \rho(w,s);$$

$$\alpha_U = \frac{|U|}{|W|}, \quad \alpha_V = \frac{|V|}{|W|}, \quad \beta = \gamma = 0;$$

- Расстояние между центрами:

$$R^{\text{ц}}(W,S) = \rho^2\left(\sum_{w\in W}\frac{w}{|W|}, \sum_{s\in S}\frac{s}{|S|}\right);$$

$$\alpha_U = \frac{|U|}{|W|},\ \ \alpha_V = \frac{|V|}{|W|},\ \ \beta = -\alpha_U\alpha_V,\ \ \gamma = 0;$$

- Расстояние Уорда:

$$R^{\text{ц}}(W,S) = \frac{|S||W|}{|S|+|W|}\rho^2\left(\sum_{w\in W}\frac{w}{|W|}, \sum_{s\in S}\frac{s}{|S|}\right);$$

$$\alpha_U = \frac{|S|+|U|}{|S|+|W|},\ \ \alpha_V = \frac{|S|+|V|}{|S|+|W|},\ \ \beta = -\frac{-|S|}{|S|+|W|},\ \ \gamma = 0;$$

- Гибкое расстояние:

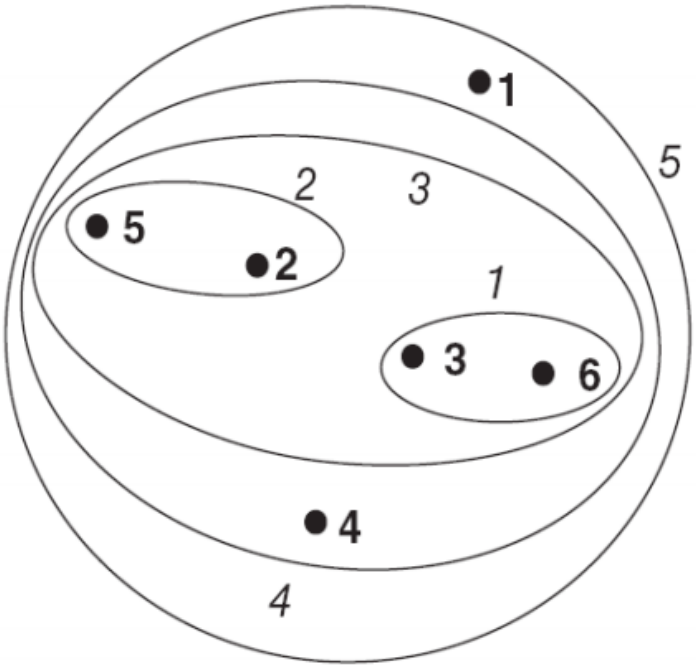
$$\alpha_U = \alpha_V = \frac{1-\beta}{2},\ \ \beta < 1\ (-0.25),\ \ \gamma = 0;$$

Свойства кластеризаций

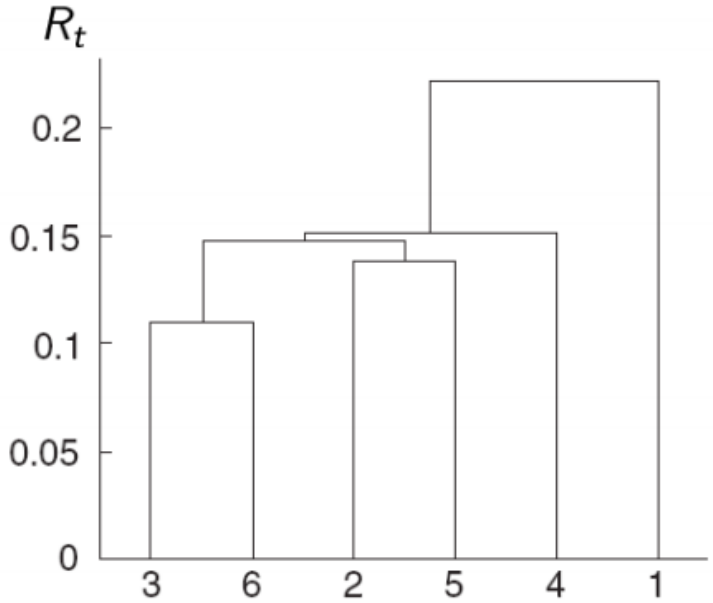
- Монотонность: кластеризация монотонна, если при каждом объединении расстояния между объединяемыми кластерами только увеличивается.
- Кластеризация сжимающая, если $R_t \leqslant \rho(\mu_U, \mu_V), \forall t$.
- Кластеризация растягивающая, если $R_t \geqslant \rho(\mu_U, \mu_V), \forall t$.

1. Расстояние ближнего соседа:

Диаграмма вложения

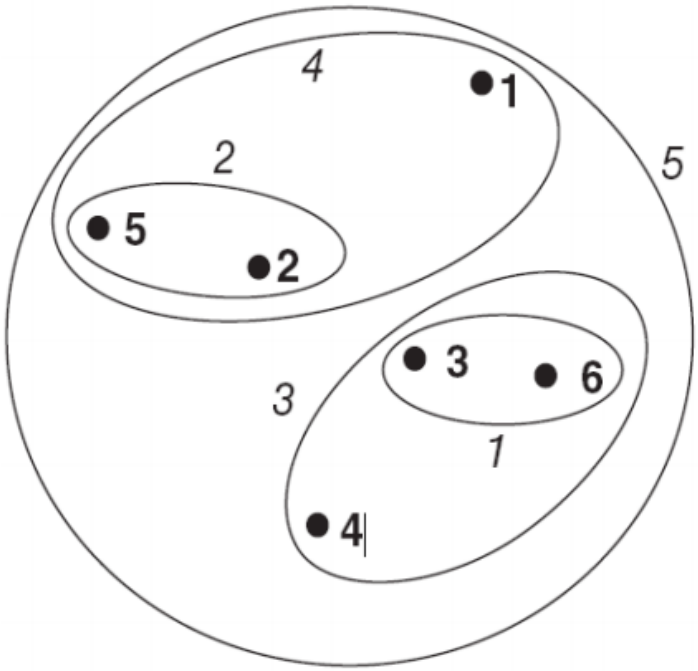


Дендрограмма

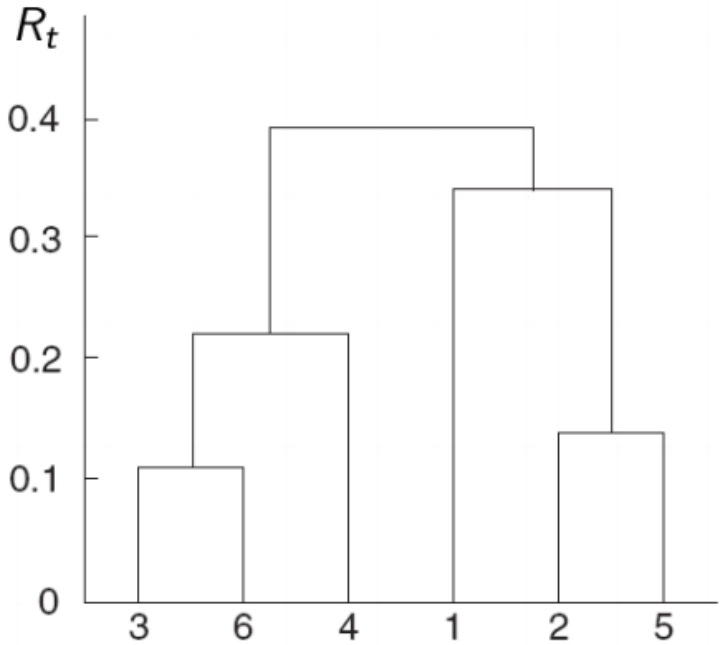


2. Расстояние дальнего соседа:

Диаграмма вложения

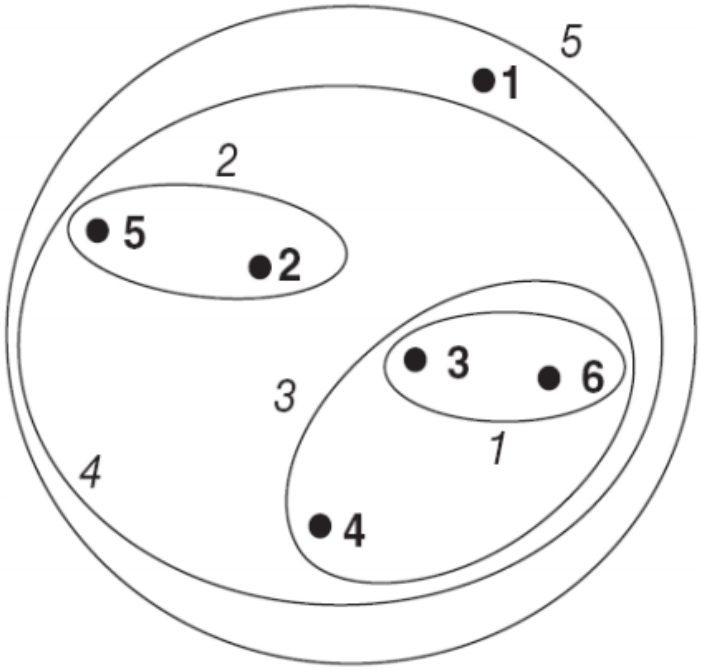


Дендрограмма

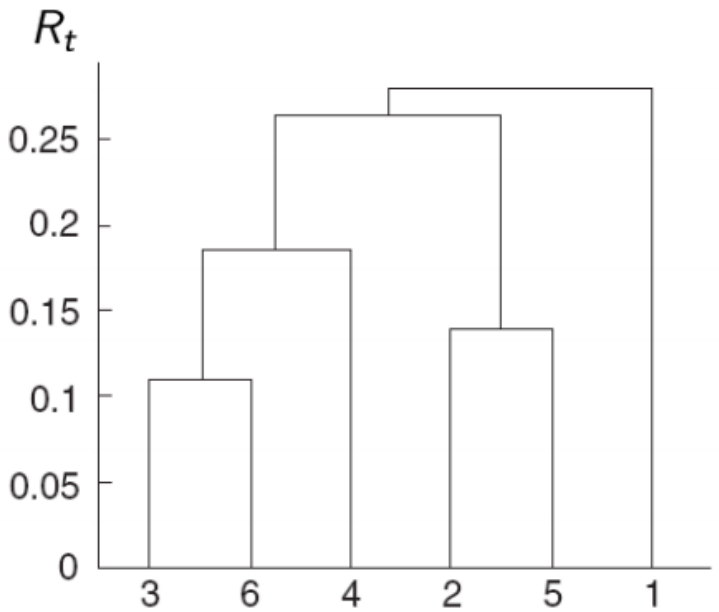


3. Групповое среднее расстояние:

Диаграмма вложения

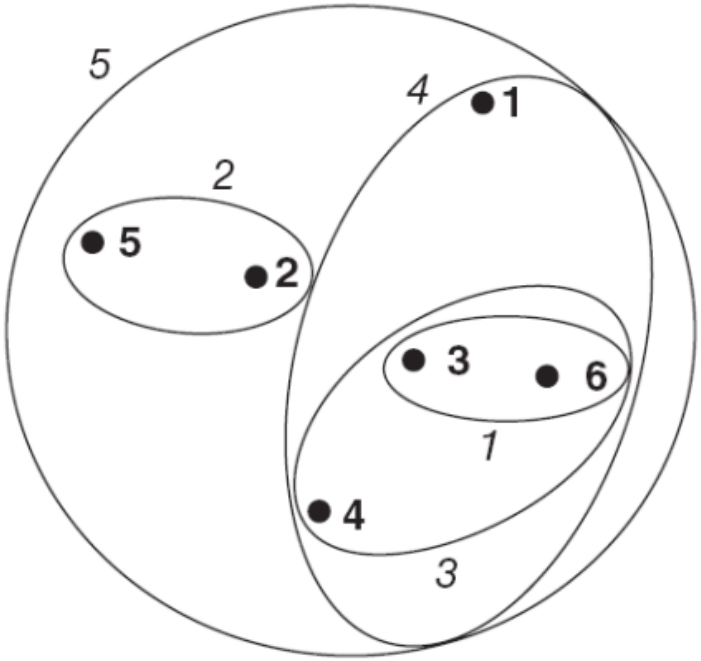


Дендрограмма

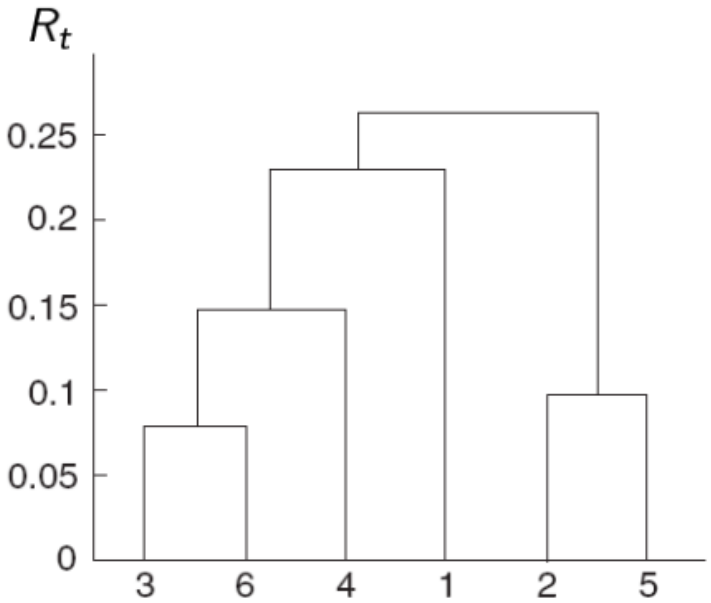


5. Расстояние Уорда:

Диаграмма вложения



Дендрограмма



Плюсы и минусы

Достоинства:

- В качестве результата можно получить дендрограмму.
- Форма кластеров может быть произвольной.
- Количество кластеров можно определить по дендрограмме.

Недостатки:

- Необходимость подбирать одно из множества различных расстояний.
- Отсутствие модели в задаче не позволяет однозначно предпочесть одно разделение на кластеры другому.

Функционалы качества

Задачу кластеризации можно ставить следующим образом: необходимо приписать номера кластеров объектам так, чтобы значение выбранного функционала качества было минимальным или максимальным.

Некоторые функционалы качества:

- Среднее внутрикластерное расстояние $F_0 = \frac{\sum_{i<j}[y_i=y_j]\rho(x_i,x_j)}{\sum_{i<j}[y_i=y_j]}$
- Среднее межкластерное расстояние $F_1 = \frac{\sum_{i<j}[y_i\neq y_j]\rho(x_i,x_j)}{\sum_{i<j}[y_i\neq y_j]}$
- Силуэт
- Индекс Данна и т.д.

Имеет смысл также вычислять отношение пары функционалов, чтобы учесть как внутрикластерные, так и межкластерные расстояния: $F_0/F_1 \to min$.

Силуэт

1. Принадлежность объекта своему кластеру
   $c(x_i) = \frac{1}{|K_i|-1} \sum_{x_j \in K_i, i \neq j} \rho(x_i, x_j)$
2. Принадлежность объекта другому кластеру
   $b(x_i) = \min_{l \neq i} \frac{1}{|K_l|} \sum_{x_j \in K_l} \rho(x_i, x_l)$
3. Определим силуэт объекта как $s(x_i) = \frac{b(x_i) - c(x_i)}{\max\{c(x_i), b(x_i)\}}$, если $|K_i| > 1$ и $s(x_i) = 0$, если $|K_i| = 1$.
4. $S = \frac{1}{m} \sum_{i=1}^{m} s(x_i)$ — силуэт кластеризации.

Dunn index

$$D = \frac{\min_{1 \leqslant i < j < k} \delta(K_i, K_j)}{\max_{1 \leqslant s \leqslant k} \Delta(K_s)},$$

где $\delta(K_i, K_j)$ — расстояние между кластерами, $\Delta(K_s)$ — диаметр кластера.

Если доступна внешняя информация о разделении на классы априори, то можно воспользоваться следующими метриками:

- Rand Index

$$Rand = \frac{TP+FN}{TP+TN+FP+FN}$$

- Jaccard Index

$$Jaccard = \frac{TP}{TP+TN+FP}$$

- Minkowski Score
- Folkes and Mallows Index и т.д.

Кратчайший незамкнутый путь (КНП)

1. Найти пару вершин $(x_i, x_j) \in X^m$ с наименьшим $\rho(x_i, x_j)$ и соединить их ребром;
2. **Пока** в выборке остаются изолированные точки
3. найти изолированную точку, ближайшую к некоторой неизолированной;
4. соединить эти две точки ребром;
5. удалить $k - 1$ самых длинных рёбер;

FOREL

1. Пусть $U = X^m$
2. **Пока** есть некластеризованные точки, т.е. $U \neq \varnothing$;
3. взять случайную точку $x_0 \in U$;
4. **Повторять**
5. образовать кластер с центром в x_0 и радиусом R: $K_0 = \{x_i \in U \mid \rho(x_i, x_0) \leqslant R\}$;
6. переместить центр x_0 в центр масс кластера: $x_0 = \frac{1}{|K_0|} \sum_{x_i \in K_0} x_i$;
7. **Пока** состав кластера K_0 не стабилизируется;
8. $U = U \setminus K_0$;
9. применить алгоритм КНП к множеству центров кластеров;
10. каждый $x_i \in X^m$ приписать кластеру с ближайшим центром;

Плюсы и минусы

Достоинства:

- Получаем двухуровневую систему кластеров;
- Кластеры могут быть произвольной формы;
- варьируя R можно управлять детальностью кластеризации.

Недостатки:

- алгоритм очень чувствителен к R и к начальному выбору точки x_0

Density-Based Spatial Clustering of Aplications with Noise (DBSCAN)

Объект $x \in U$, его ε-окрестность
$U_\varepsilon(x) = \{u \in U : \rho(x,u) \leqslant \varepsilon\}$ Каждый объект может быть одного из трёх типов:

- корневой: имеет плотную окрестность $|U_\varepsilon(x)| \geqslant m$
- граничный: не корневой, но находится в окрестности корневого
- выброс: не корневой и не граничный.

1. $U = X^m$, $N = \varnothing$, $z = 0$;
2. **Пока** есть некластеризованные точки, т.е. $U \neq \varnothing$;
3. взять случайную точку $x \in U$;
4. **если** $|U_\varepsilon(x)| < m$, **то**
5. пометить x как шумовой;
6. **иначе**
7. создать новый кластер: $K = U_\varepsilon(x)$; $z = z + 1$;
8. **для всех** $x' \in K$
9. **если** $|U_\varepsilon(x')| \geqslant m$ **то** $K = K \cup U_\varepsilon(x')$;
10. **иначе** пометить x' как граничный элемент K;
11. $a(x_i) = z$ для всех $x' \in K$;
12. $U = U \setminus K$;

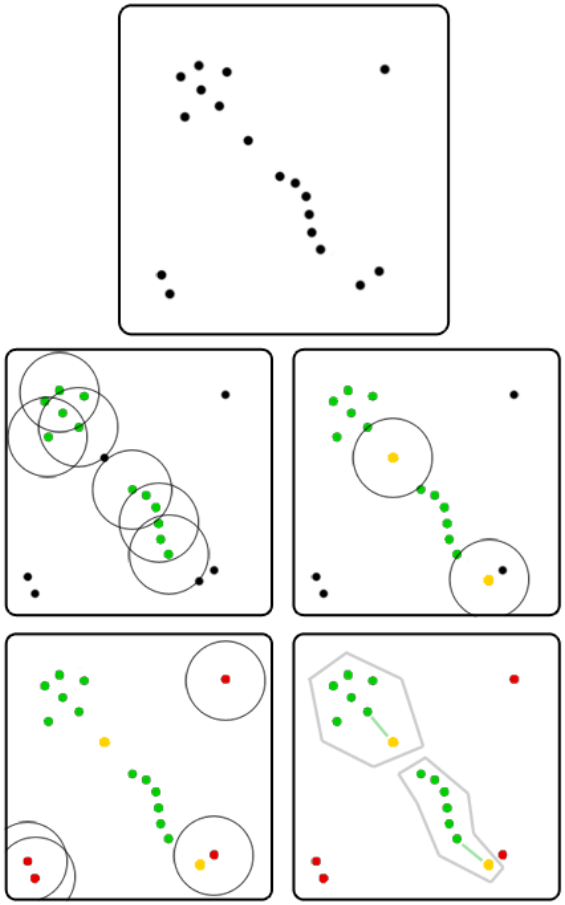


Рис.: Иллюстрация к алгоритму DBSCAN

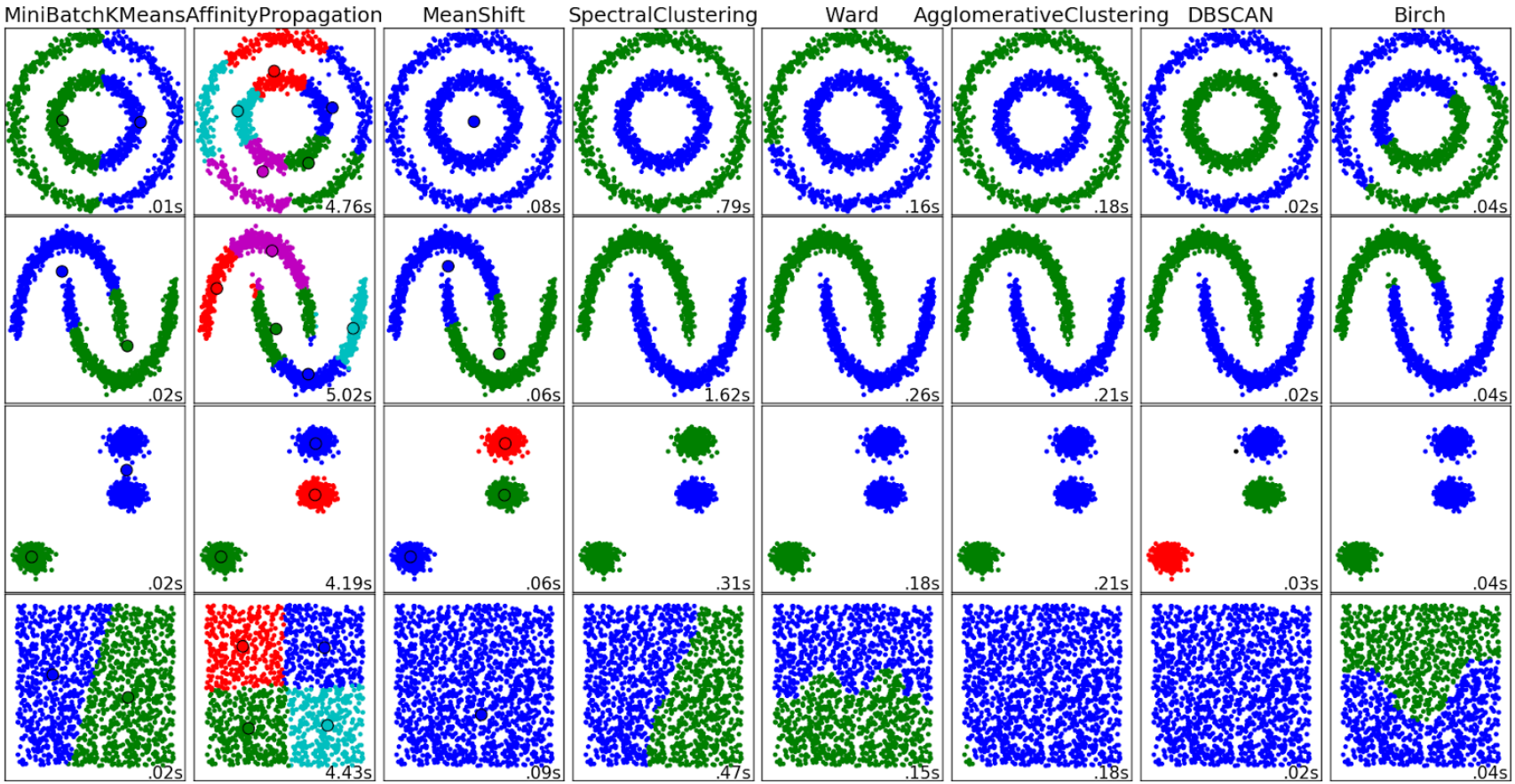


Рис.: Пример работы DBSCAN (второй столбец справа)

Плюсы и минусы

Достоинства:

- Быстрая кластеризация больших данных (от $O(m \ln m)$ до $O(m^2)$ в зависимости от реализации);
- Кластеры произвольной формы;
- Явная разметка шумовых объектов;
- Хорошо поддаётся модифицированию (существуют реализации, скрещенные с k-means и даже с GMM).

Недостатки:

- Алгоритм может неадекватно обрабатывать сильные вариации плотности данных внутри кластера, проёмы и шумовые мосты между кластерами.