Кластеризация данных

Курс «Основы анализа текстовых данных»

Кафедра управления и интеллектуальных технологий

НИУ «МЭИ»

Весна 2023 г.

Что такое кластеризация?

Кластеризация - задача разбиения заданной выборки *объектов* на непересекающиеся подмножества, называемые кластерами, так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались.
Под схожестью обычно понимается близость друг к другу относительно выбранной метрики.

Задача кластеризации относится к разделу задач обучения без учителя.

Обучение без учителя (Unsupervised learning) — один из разделов машинного обучения. Изучает широкий класс задач обработки данных, в которых известны только описания множества объектов (обучающей выборки), и требуется обнаружить внутренние взаимосвязи, зависимости, закономерности, существующие между объектами.

Обучение без учителя часто противопоставляется обучению с учителем, когда для каждого обучающего объекта задаётся «правильный ответ», и требуется найти зависимость между объектами и ответами.

Курс лекций К.В. Воронцова: http://www.machinelearning.ru/wiki/images/2/28/Voron-ML-Clustering-slides.pdf

Постановка задачи кластеризации

Дано:

X – пространство объектов

$\vec{X}_l$ – обучающая выборка; l = 1…L

ρ - функция расстояния между объектами

Найти:

Y – множество кластеров и

a: **X** → **Y** – алгоритм кластеризации, такой, что:

- каждый кластер состоит из близких объектов
- объекты разных кластеров существенно различны

Особенности задачи кластеризации

Решение задачи кластеризации принципиально неоднозначно:

- Точной постановки задачи кластеризации нет
- Существует множество критериев качества кластеризации
- Число кластеров $|Y|$ заранее, как правило, не известно
- Результат кластеризации существенно зависит от метрики ρ, которую эксперт задает субъективно

Цели кластеризации

Понимание данных путём выявления кластерной структуры. Разбиение выборки на группы схожих объектов позволяет упростить дальнейшую обработку данных и принятия решений, применяя к каждому кластеру свой метод анализа (стратегия «разделяй и властвуй»).

Сжатие данных. Если исходная выборка избыточно большая, то можно сократить её, оставив по одному наиболее типичному представителю от каждого кластера.

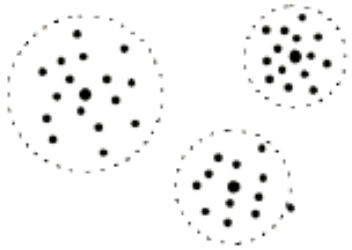
Обнаружение новизны (novelty detection). Выделяются нетипичные объекты, которые не удаётся присоединить ни к одному из кластеров.

Построение иерархии множества объектов (задача таксономии)

Примеры кластерных структур

внутрикластерные расстояния, как правило, меньше межкластерных

ленточные кластеры



кластеры с центром

Примеры кластерных структур

кластеры могут соединяться перемычками

кластеры могут накладываться на разрежен-
ный фон из редко расположенных объектов

кластеры могут перекрываться

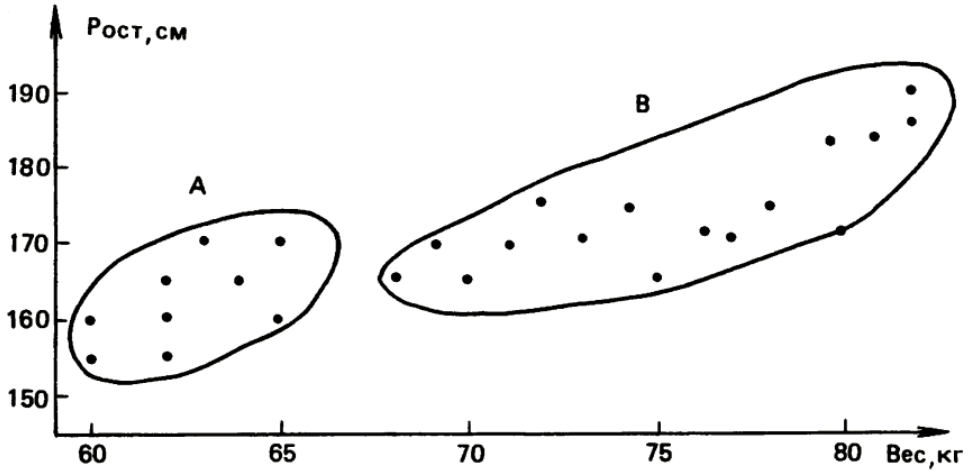
Примеры кластерных структур

кластеры могут образовываться не по сход-
ству, а по иным типам регулярностей

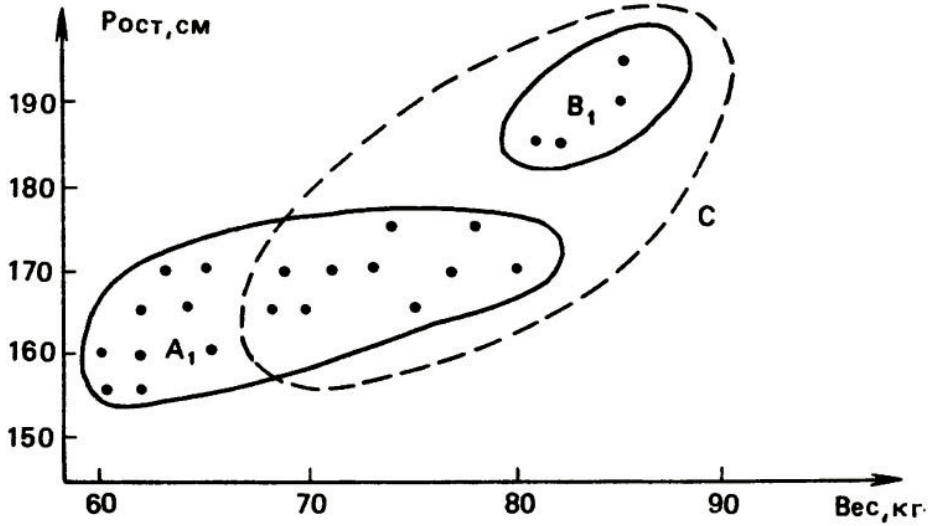
кластеры могут вообще отсутствовать

- Каждый метод кластеризации имеет свои ограничения и выделяет кластеры лишь некоторых типов
- Понятие «тип кластерной структуры» зависит от метода и не имеет формального определения

Проблема чувствительности к метрике



A – девушки
B – молодые люди



После перенормировки
(сжали ось «Вес» вдвое)

Качество кластеризации

Сумма средних внутрикластерных расстояний:

$$F_0 = \sum_{k \in Y} \frac{1}{N_k} \sum_{l=1}^{N_k} \rho(\vec{X}_l, \mu_k) \to \min$$

Сумма межкластерных расстояний:

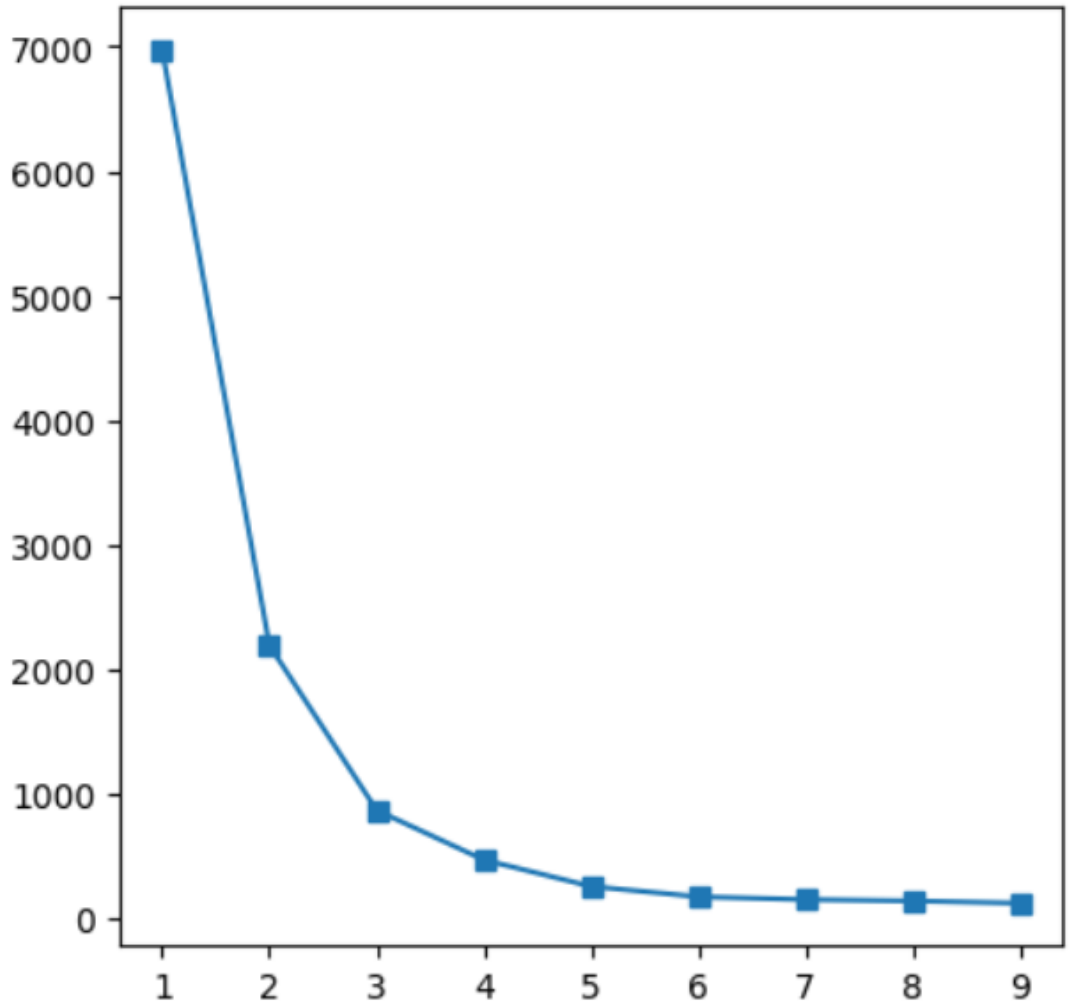
$$F_1 = \sum_{j,k \in Y} \rho(\mu_j, \mu_k) \to \max$$

Обобщенный функционал:

$$F_2 = \frac{F_0}{F_1} \to \min$$

μ_k - Центр масс кластера k

N_k - Размер кластера k



Качество кластеризации

Коэффициент силуэта

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

Среднее расстояние между i-м и остальными объектами кластера - **cohesion**

$$a_i = \frac{1}{|C_I| - 1} \sum_{j \in C_I, i \neq j} d(i,j)$$

Минимальное расстояние между i-м и остальными объектами других кластеров – **separation**

$$b_i = \min_{I \neq J} \frac{1}{|C_J|} \sum_{j \in C_J} d(i,j)$$

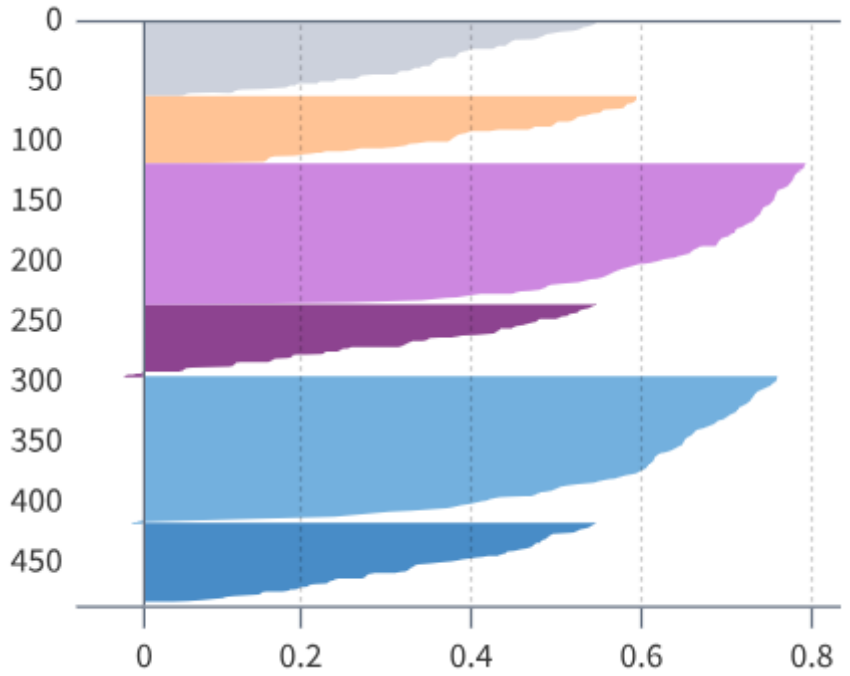
Для каждого объекта рассчитываются значения si.

si → 1 – высокая уверенность в принадлежности объекта кластеру

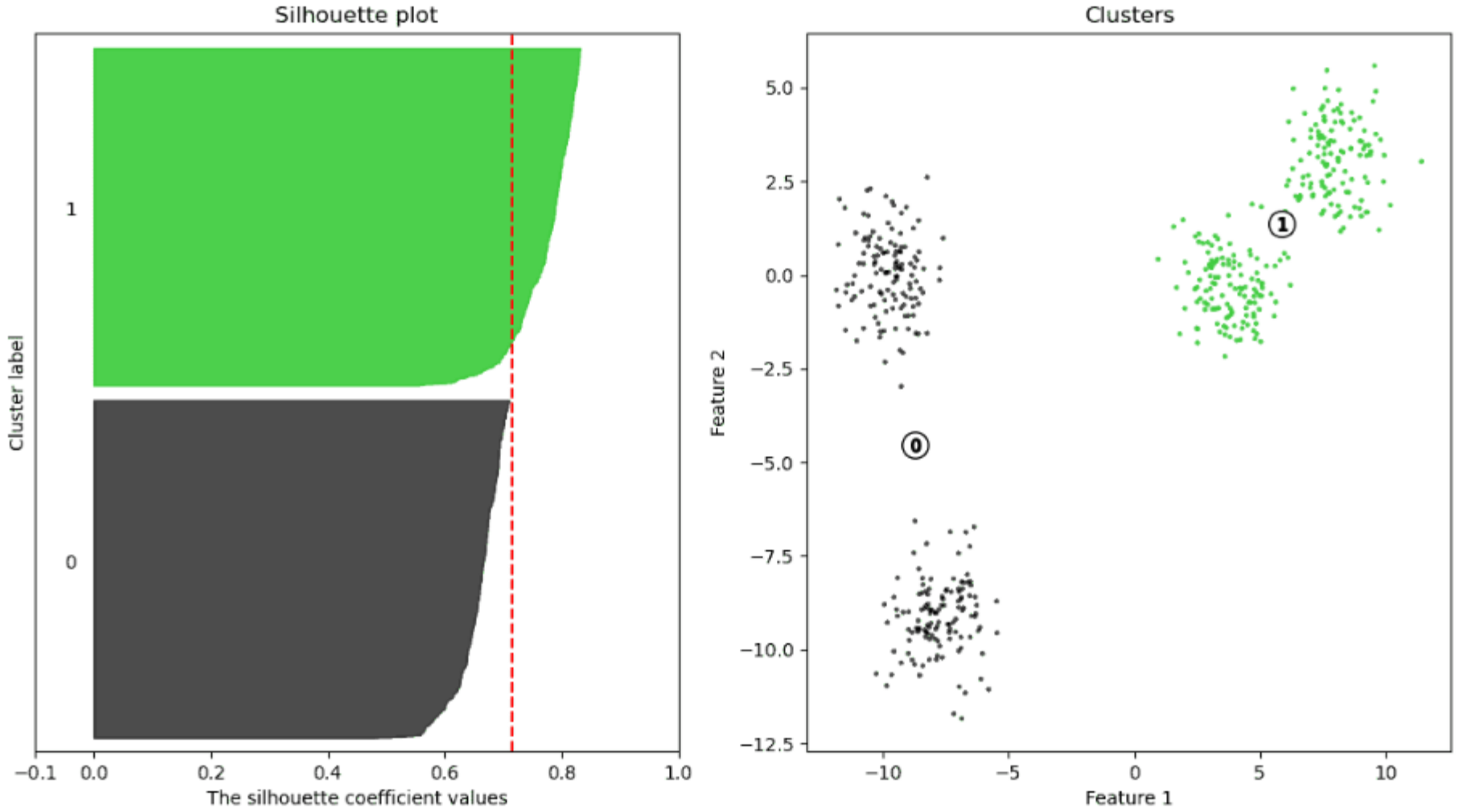
si → -1 – объект больше подходит другому кластеру.

Среднее si характеризует кластеризацию в целом.

Графическое отображение:

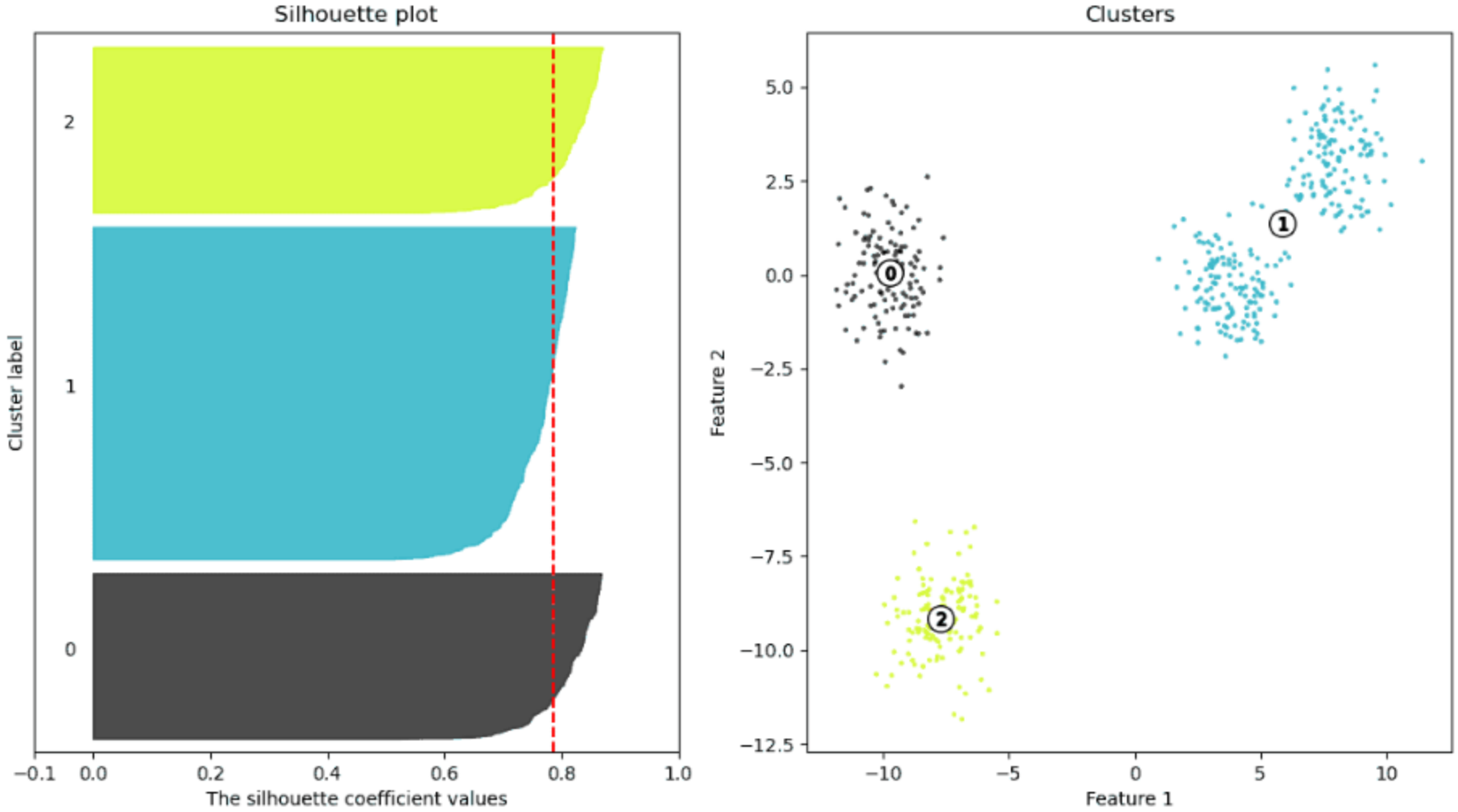


Качество кластеризации

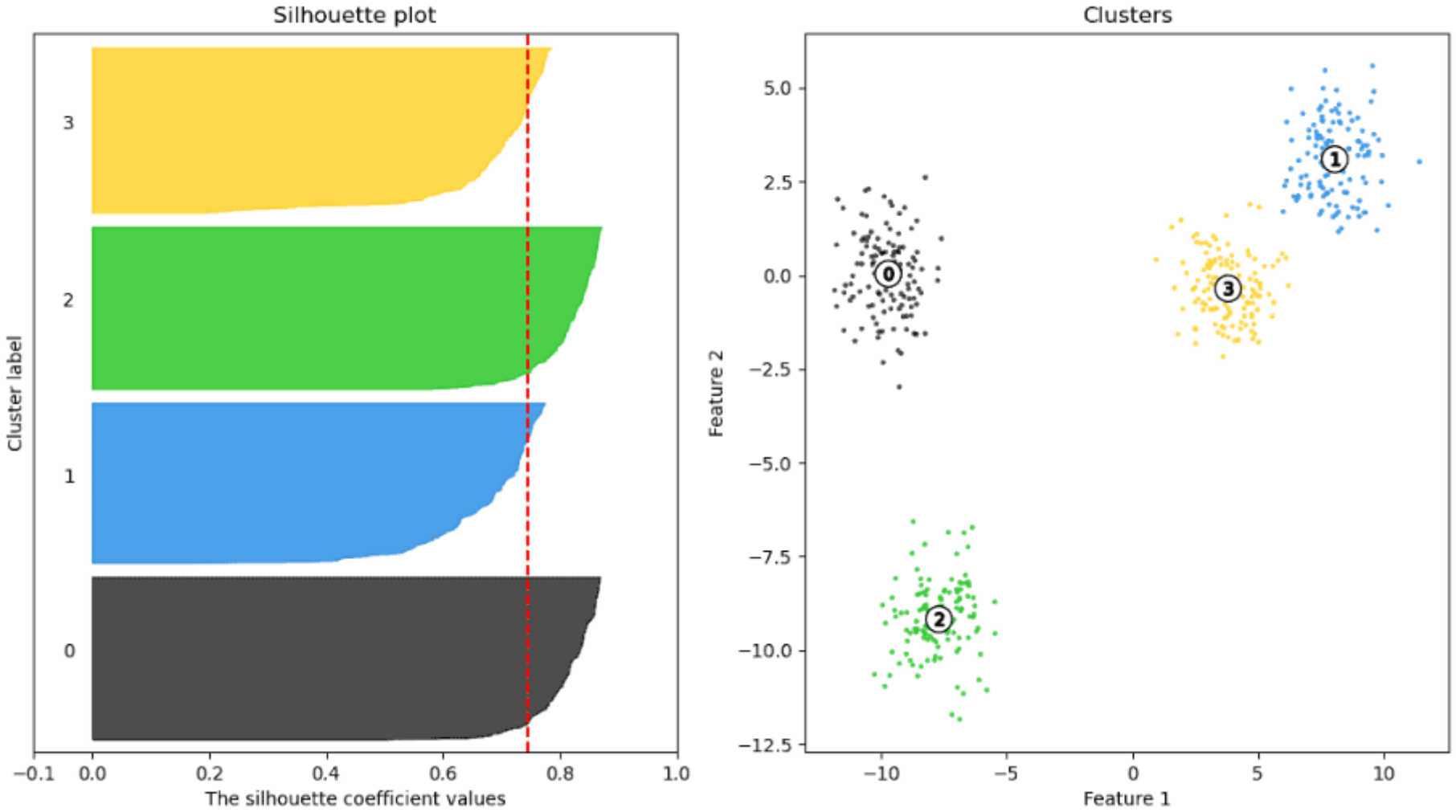


https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

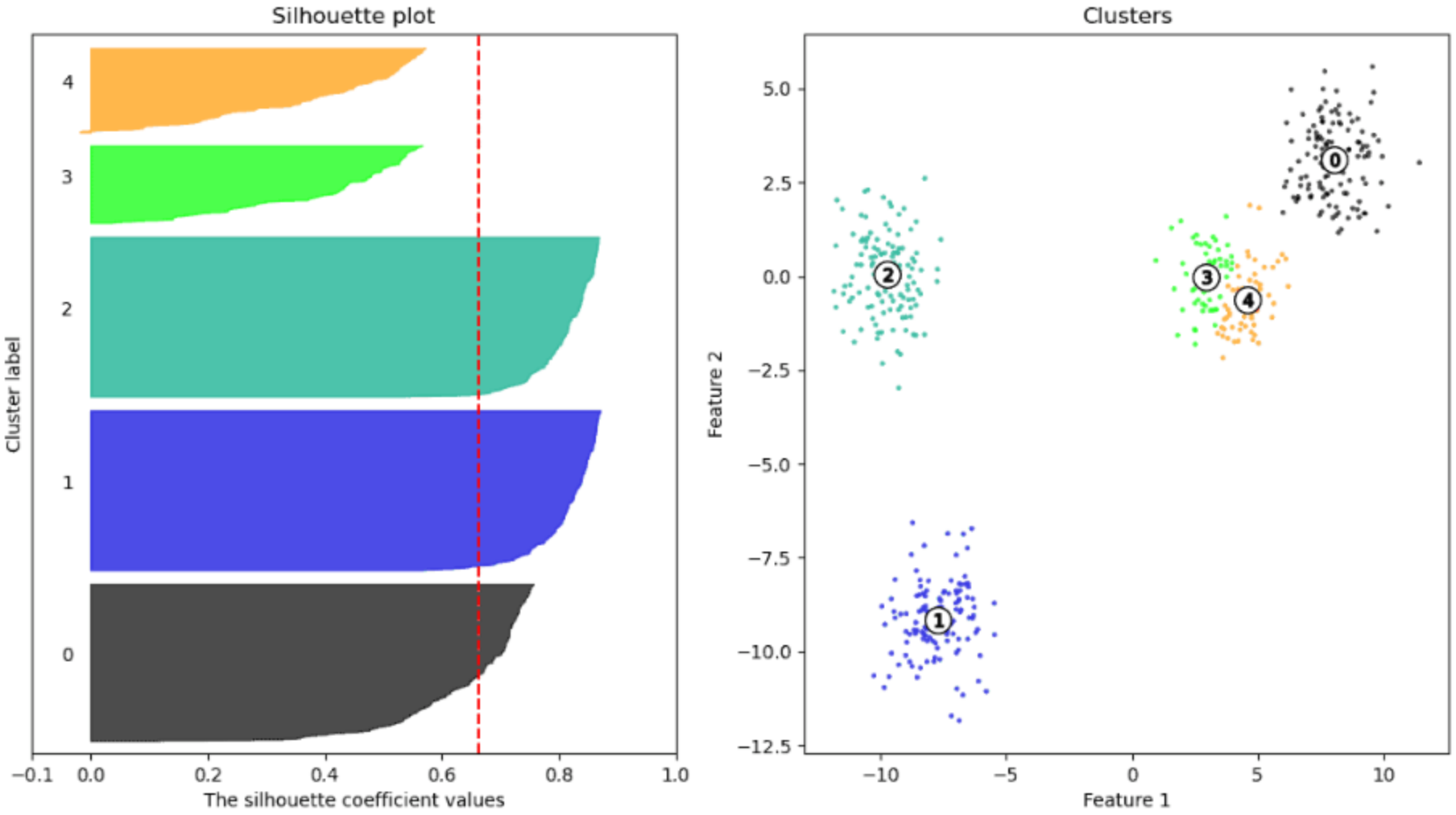
Качество кластеризации



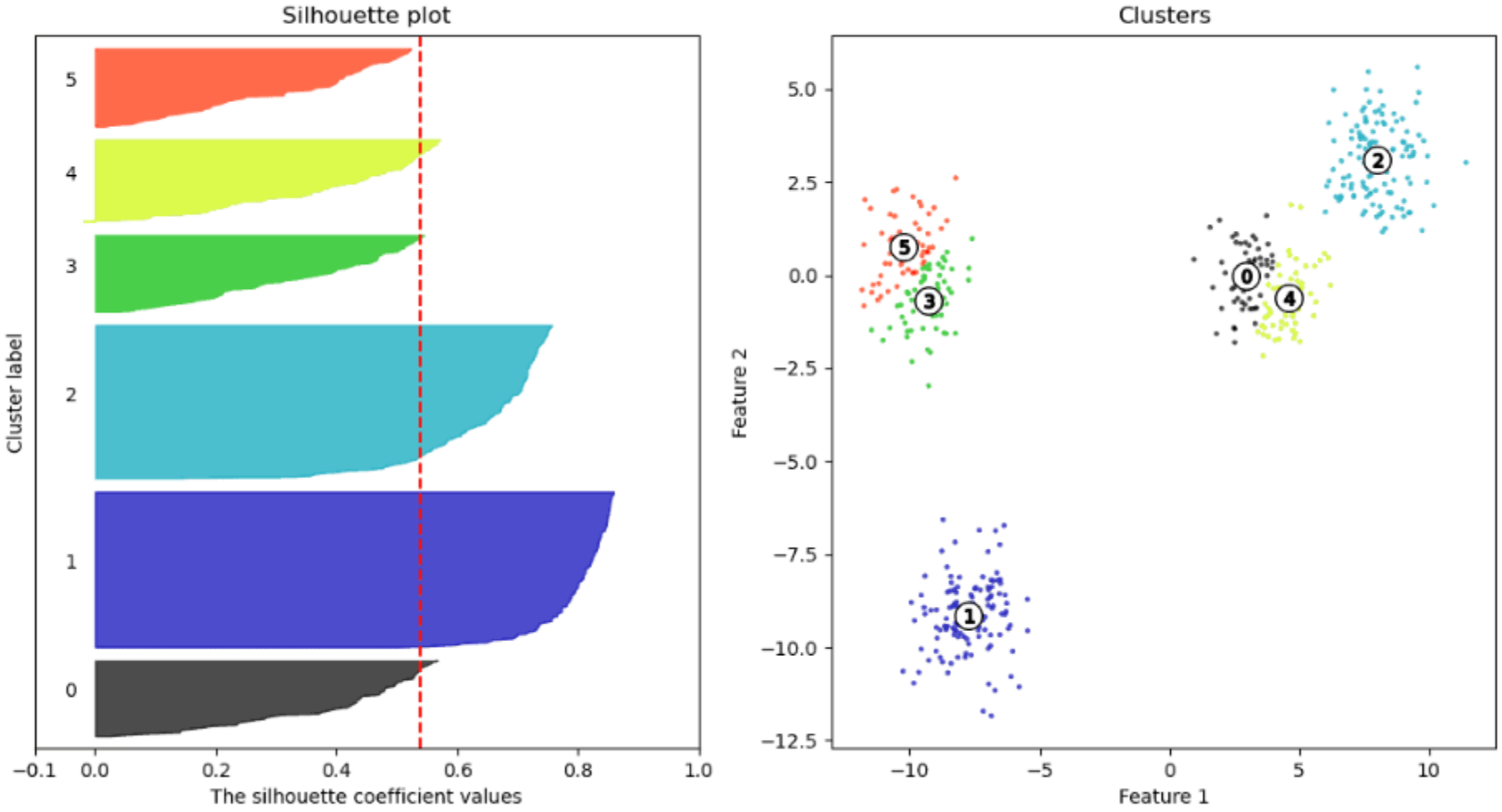
Качество кластеризации



Качество кластеризации

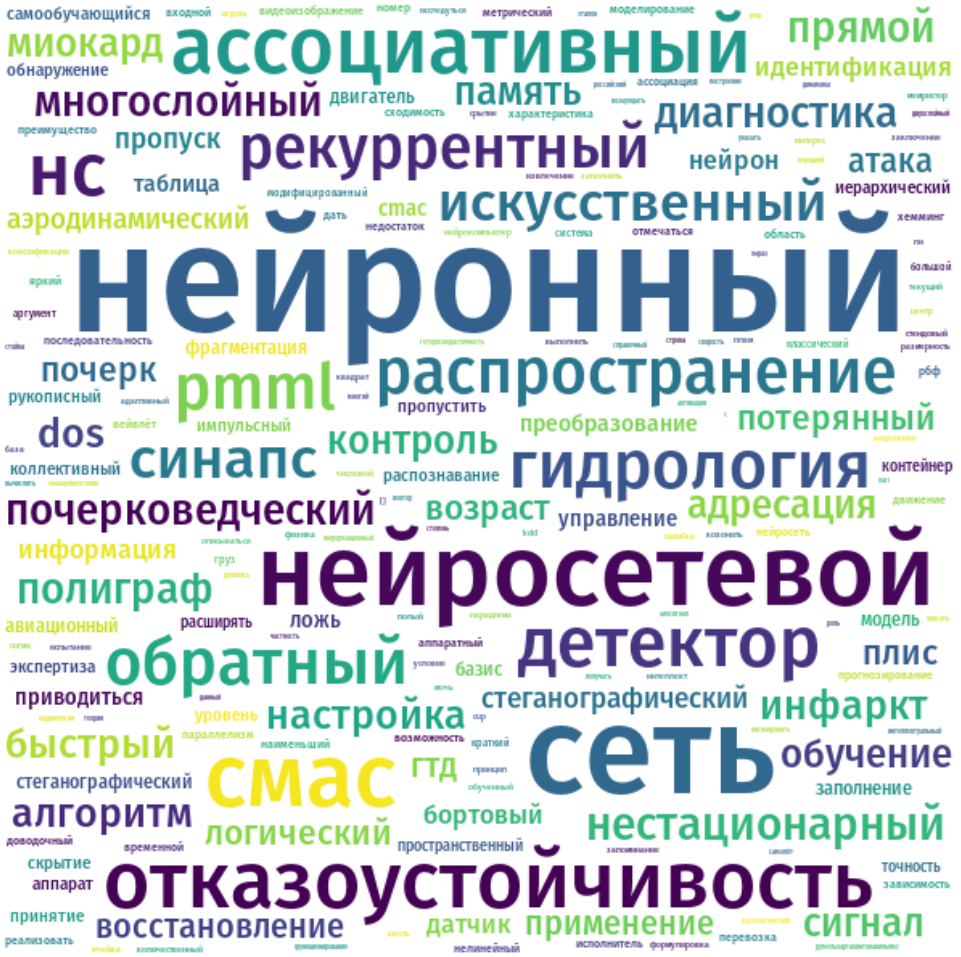


Качество кластеризации



Качество кластеризации

В задачах кластеризации текстов качество кластеризации можем косвенно оценить по наиболее частотным терминам, встречающимся в классе («Облако слов»). Т.е. мы могли бы дать название каждому кластеру исходя из наиболее частотных терминов :



Алгоритмы кластеризации

Иерархические

- Агломеративная кластеризация
- Дивизимная кластеризация

Статистические

- EM-алгоритмы
- К-средних (k-means)
- Алгоритм FOREL

Самоорганизующиеся карты - сети Кохонена

Иерархические алгоритмы кластеризации

Среди алгоритмов иерархической кластеризации различаются два основных типа. Дивизимные или нисходящие алгоритмы разбивают выборку на всё более и более мелкие кластеры. Более распространены агломеративные или восходящие алгоритмы, в которых объекты объединяются во всё более и более крупные кластеры

Сначала каждый объект считается отдельным кластером. Для одноэлементных кластеров естественным образом определяется функция расстояния $\rho(x_j, x_k)$

Затем запускается процесс слияний. На каждой итерации вместо пары самых близких кластеров U и V образуется новый кластер $W = U \cup V$

Расстояние от нового кластера W до любого другого кластера S вычисляется по расстояниям R(U, V), R(U, S) и R(V, S):

$$R(U \cup V, S) = \alpha_u R(U, S) + \alpha_v R(V, S) + \beta R(U, V) + \gamma |R(U, S) - R(V, S)|$$

где $\alpha_u, \alpha_v, \beta, \gamma$- числовые параметры

Эта универсальная формула обобщает практически все разумные способы определить расстояние между кластерами. Она была предложена Лансом и Уильямсом в 1967 году.

Иерархические алгоритмы кластеризации

На практике используются следующие способы вычисления расстояний R(W, S) между кластерами W и S. Для каждого из них доказано соответствие формуле Ланса-Вильямса при определённых сочетаниях параметров:

Расстояние ближнего соседа (single linkage):

$$R^{б}(W,S) = \min_{w\in W, s\in S} \rho(w,s); \quad \alpha_u = \alpha_v = \frac{1}{2}, \quad \beta = 0, \quad \gamma = -\frac{1}{2}$$

Расстояние дальнего соседа (complete linkage):

$$R^{д}(W,S) = \max_{w\in W, s\in S} \rho(w,s); \quad \alpha_u = \alpha_v = \frac{1}{2}, \quad \beta = 0, \quad \gamma = \frac{1}{2}$$

Расстояние до центра:

$$R^{ц}(W,S) = \rho^2\left(\sum_{w\in W}\frac{w}{|W|}, \sum_{s\in S}\frac{s}{|S|}\right); \quad \alpha_u = \frac{|U|}{|W|}, \quad \alpha_v = \frac{|V|}{|W|}, \quad \beta = -\alpha_u\alpha_v, \quad \gamma = 0$$

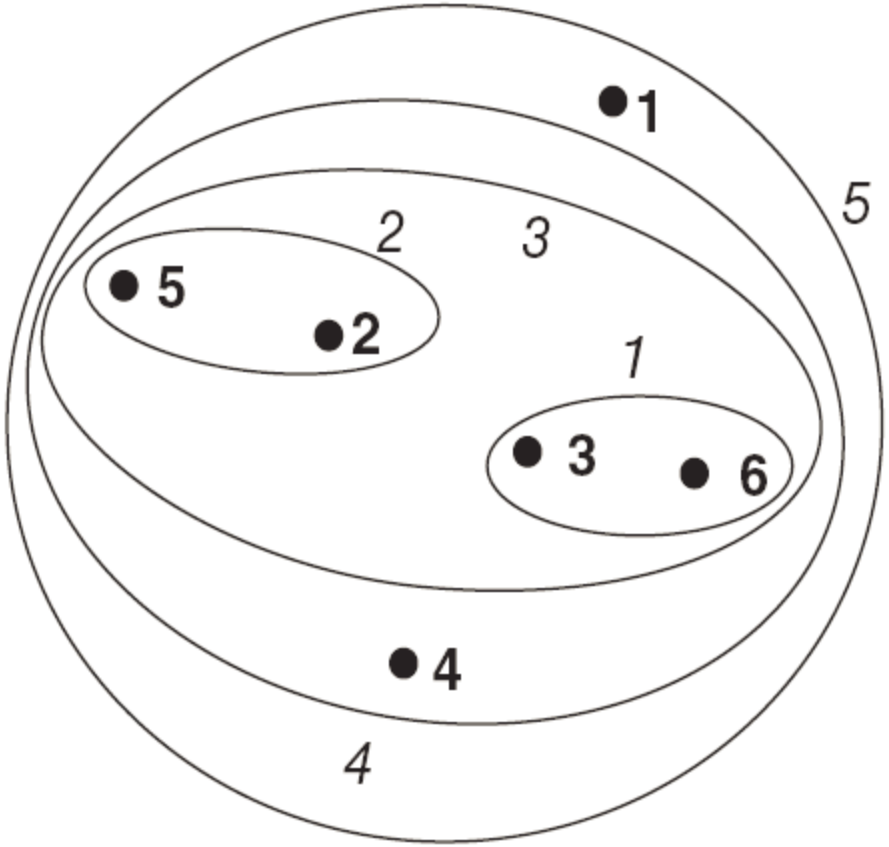
Расстояние Уорда (Варда, Ward):

$$R^{у}(W,S) = \frac{|S||W|}{|S|+|W|}\rho^2\left(\sum_{w\in W}\frac{w}{|W|}, \sum_{s\in S}\frac{s}{|S|}\right); \quad \alpha_u = \frac{|S|+|U|}{|S|+|W|} \quad \alpha_v = \frac{|S|+|V|}{|S|+|W|}, \quad \beta = -\frac{|S|}{|S|+|W|}, \quad \gamma = 0$$

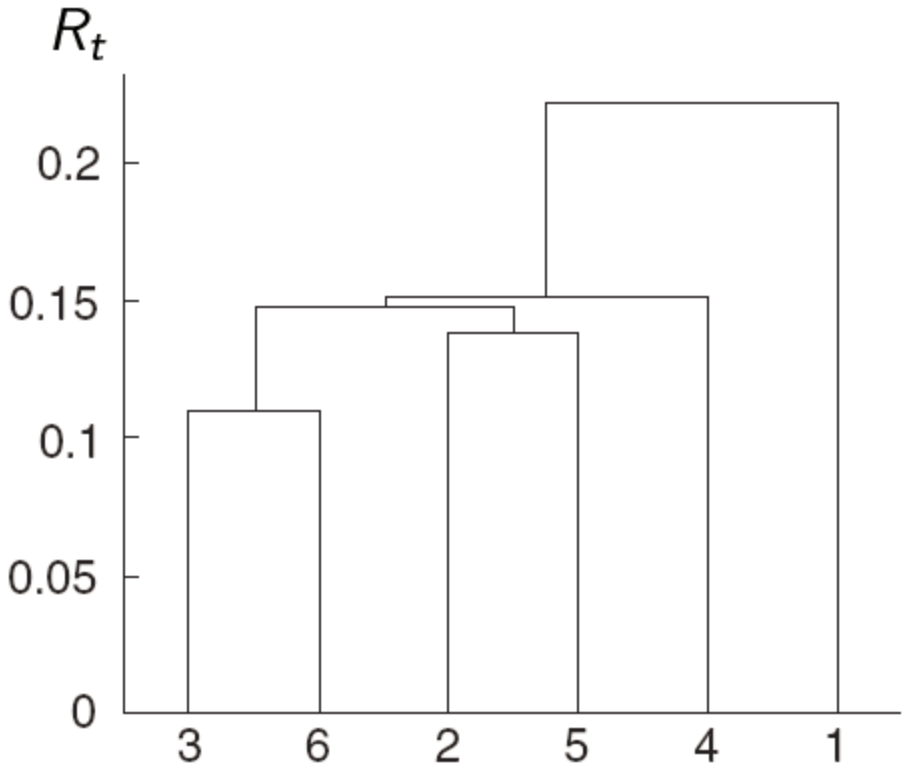
Иерархические алгоритмы кластеризации

Расстояние ближнего соседа

Диаграмма вложения



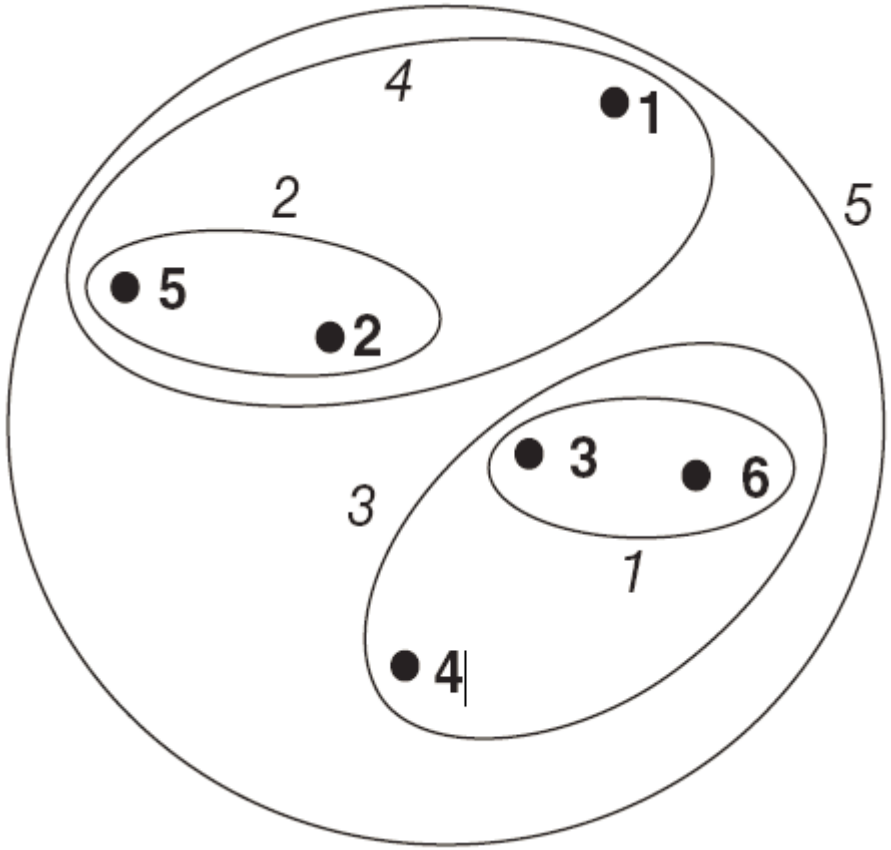
Дендрограмма



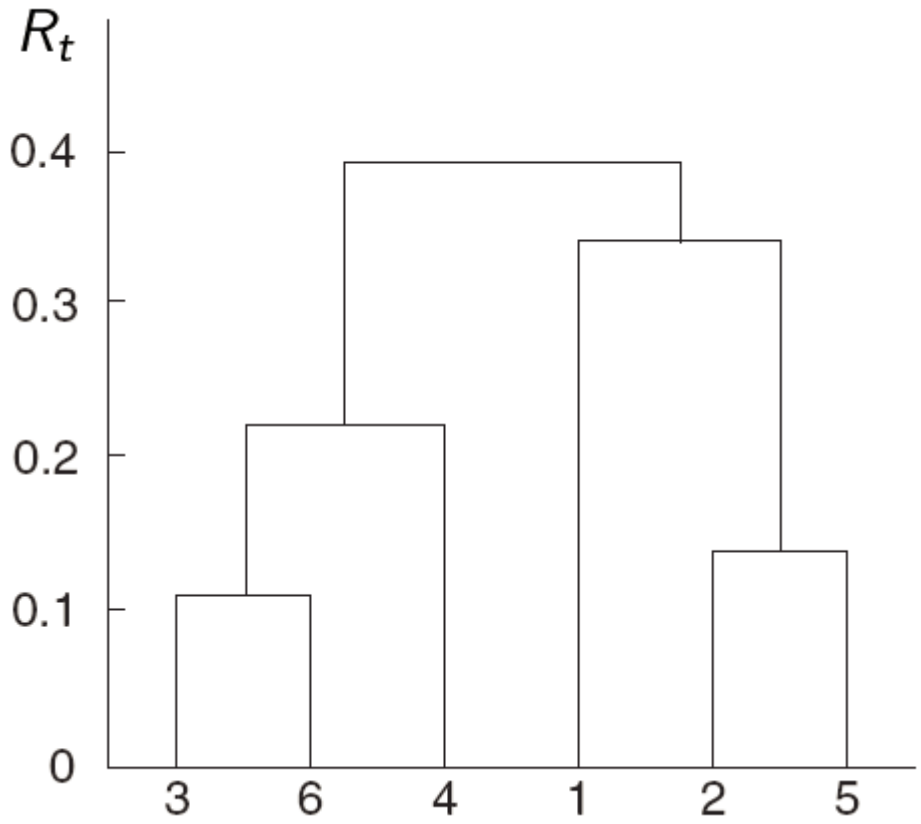
Иерархические алгоритмы кластеризации

Расстояние дальнего соседа

Диаграмма вложения



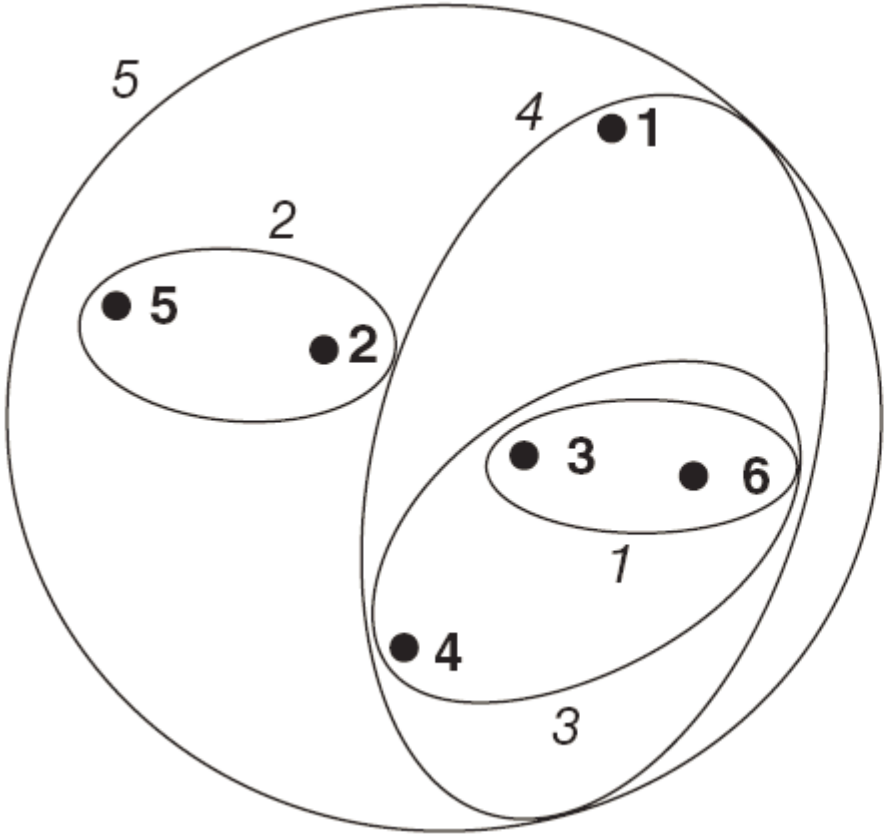
Дендрограмма



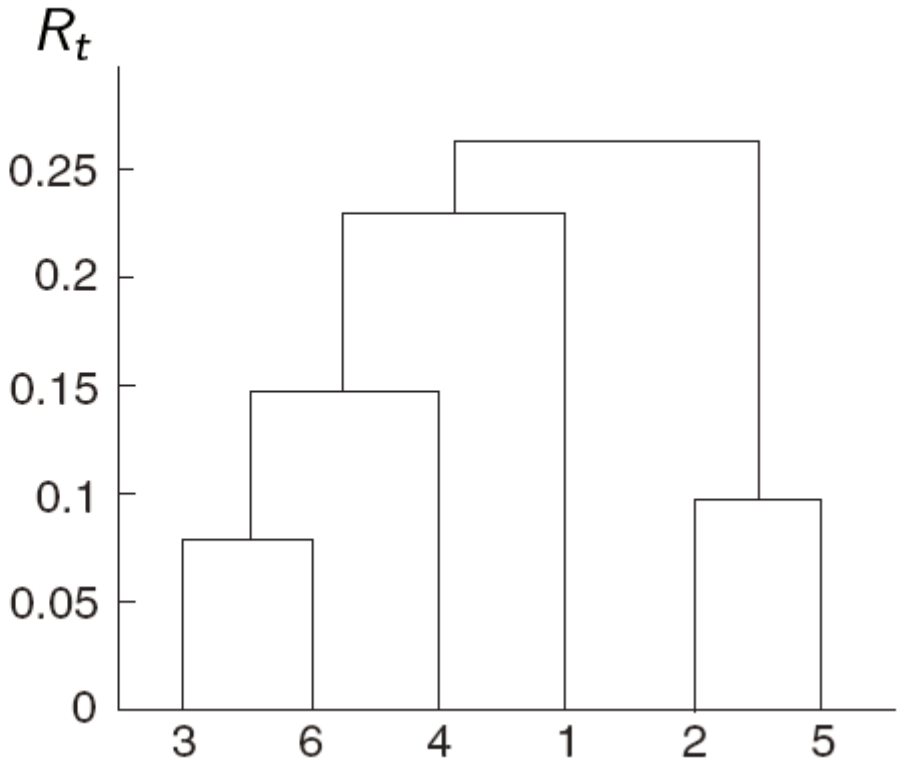
Иерархические алгоритмы кластеризации

Расстояние Уорда

Диаграмма вложения



Дендрограмма



Основные свойства иерархической кластеризации

Монотонность: дендрограмма не имеет самопересечений, при каждом слиянии расстояние между объединяемыми кластерами увеличивается: $R_2 \leq R_3 \leq R_4 \ldots$

$R^{ц}$ – не монотонна, $R^{б}$ $R^{д}$ $R^{у}$ – монотонны

Сжимаемость и растягиваемость:

$R_t \leq \rho(\mu_u, \mu_v), \forall t$ – сжимающее расстояние

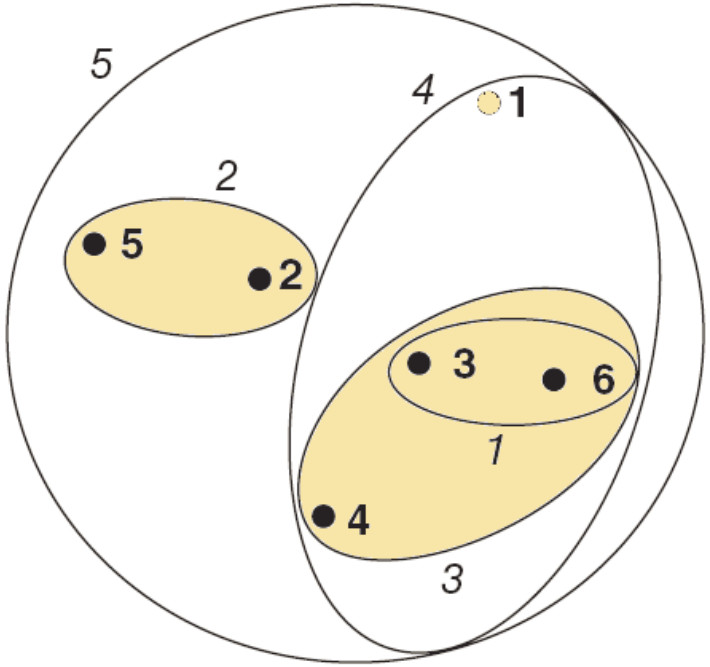
$R_t \geq \rho(\mu_u, \mu_v), \forall t$ – растягивающее расстояние

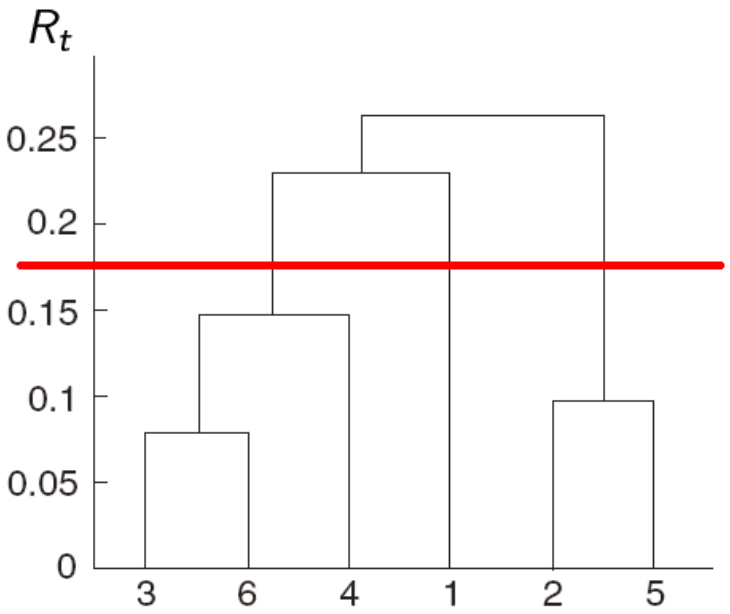
Свойство растяжения желательно, так как оно способствует более четкому отделению кластеров

$R^{б}$ –сильно сжимающее, $R^{д}$ $R^{у}$ – растягивающие, $R^{ц}$ –сохраняет метрику пространства

Выводы и рекомендации

- Рекомендуется пользоваться расстоянием Уорда.
- Обычно строят несколько вариантов и выбирают лучший визуально по дендрограмме.
- Определять число кластеров рекомендуется по максимальной высоте участка $|R_{t+1} - R_t|$ на дендрограмме.





EM-алгоритм. Предпосылки

Гипотеза о вероятностной природе данных:
Обучающая выборка X случайна и независима, состоит из смеси распределений

$$p(x) = \sum_{y \in Y} w_y p_y(x) \qquad \sum_{y \in Y} w_y = 1$$

$p_y(x)$ – функция плотности распределения кластера y,
w_y - априорная вероятность появления объектов из кластера y

Гипотеза о пространстве объектов и форме кластеров:
Кластеры n-мерные, гауссовские

$p_y(x) = (2\pi)^{-n\backslash 2} (\sigma_{y1} \cdots \sigma_{yn})^{-1} \exp(-1\backslash 2 \rho_y{}^2(x, \mu_y))$

$\mu_y = (\mu_{y1}, \ldots, \mu_{yn})$ – центр кластера y
$\Sigma_y = diag(\sigma_{y1}{}^2, \ldots, \sigma_{yn}{}^2)$ – диагональная матрица ковариаций => дисперсии

$$\rho_y{}^2(x, x') = \sum_{j=1}^{n} \sigma_{yj}{}^{-2} |x_j - x_{j'}|^2$$

EM-алгоритм

1: начальное приближение w_y, μ_y, Σ_y для всех $y \in Y$;
2: **повторять**
3: E-шаг (expectation):

$$g_{iy} := P(y|x_i) \equiv \frac{w_y p_y(x_i)}{\sum_{z \in Y} w_z p_z(x_i)}, \quad y \in Y, \quad i = 1, \ldots, \ell;$$

4: M-шаг (maximization):

$$w_y := \frac{1}{\ell} \sum_{i=1}^{\ell} g_{iy}, \quad y \in Y;$$

$$\mu_{yj} := \frac{1}{\ell w_y} \sum_{i=1}^{\ell} g_{iy} x_i \quad , \quad y \in Y, \quad j = 1, \ldots, n;$$

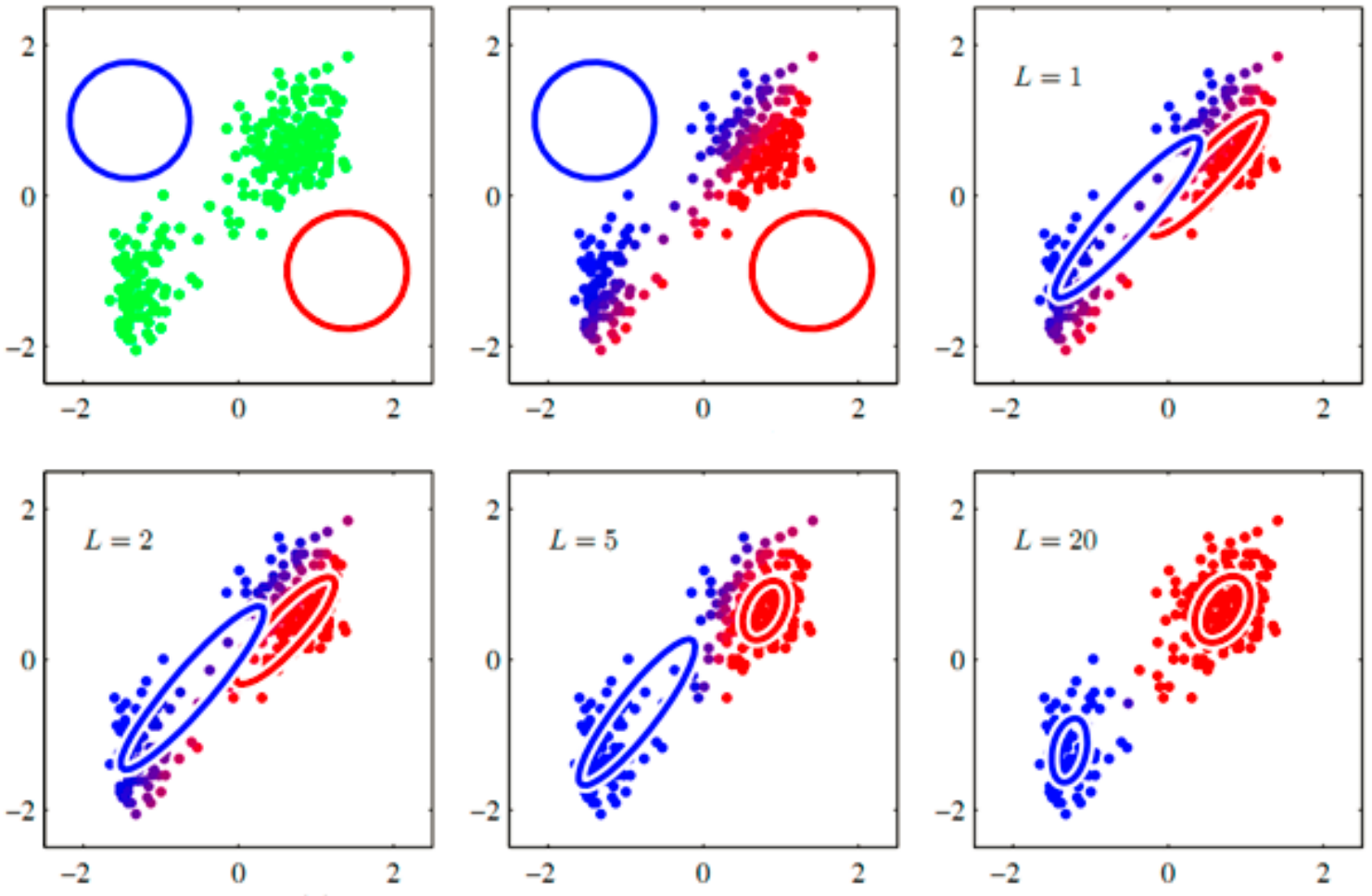
$$\sigma^2_{yj} := \frac{1}{\ell w_y} \sum_{i=1}^{\ell} g_{iy} (\ x_i \ - \mu_{yj} \ )^2, \quad y \in Y, \quad j = 1, \ldots, n;$$

5: $y_i := \arg\max_{y \in Y} g_{iy}, \quad i = 1, \ldots, \ell;$
6: **пока** y_i не перестанут изменяться;

На E-шаге по формуле Байеса расчитываются «скрытые переменные» g_{iy} – апостериорная вероятность того, что i-й объект принадлежит кластеру y

На M-шаге уточняются параметры каждого кластера используя скрытые переменные g_{iy}

EM-алгоритм



Метод к-средних (k-means)

Упрощенный аналог EM-алгоритма:
Жесткая кластеризация вместо мягкой

1. Начальное приближение центроидов $\mu_y, \quad y \in Y$
2. **Повторять:**
3. Аналог E-шага:
   отнести каждый x_i к ближайшему центру

$$y_i := \arg\min_{y \in Y} \rho(x_i, \mu_y), \quad i = 1, \ldots, \ell;$$

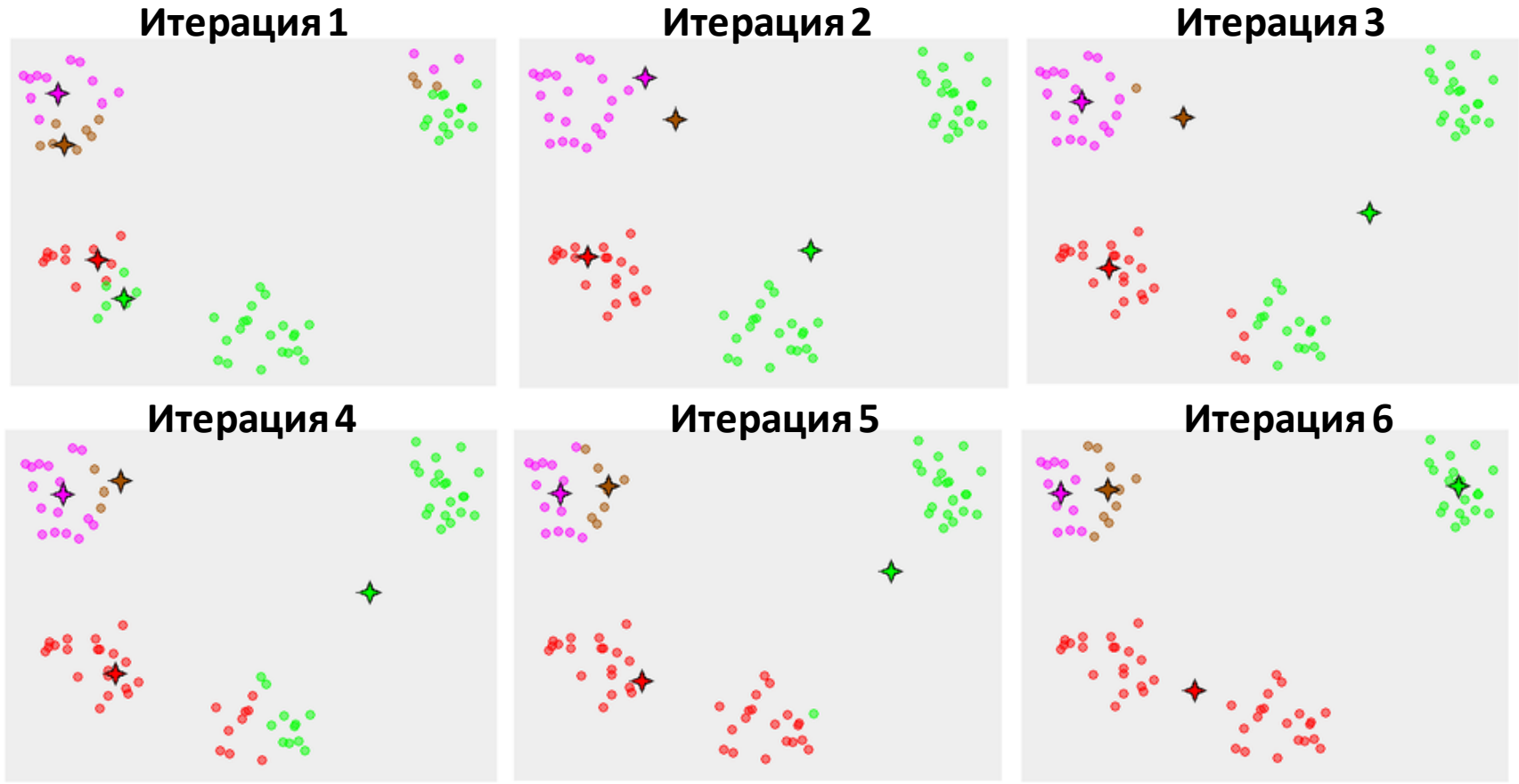
4. Аналог M-шага:
   вычислить новые положения центров:

$$\mu_{yd} := \frac{\sum_{i=1}^{\ell} [y_i = y] f_d(x_i)}{\sum_{i=1}^{\ell} [y_i = y]}, \quad y \in Y, \quad d = 1, \ldots, n;$$

5. **Пока** y_i не перестанут изменяться

Недостатки метода (k-means)

- Не гарантируется достижение глобального минимума суммарного квадратичного отклонения *V*, а только одного из локальных минимумов.
- Результат зависит от выбора исходных центров кластеров, их оптимальный выбор неизвестен.
- Рекомендуется повторная прогонка алгоритма для избежания ситуации «плохой» кластеризации.
- Число кластеров надо знать заранее или перебирать в поисках оптимального.



Семейство алгоритмов FOREL (ФОРмальный ЭЛемент)

Алгоритм предложен Загоруйко Н. Г. и Ёлкиной В. Н. в 1967 году.

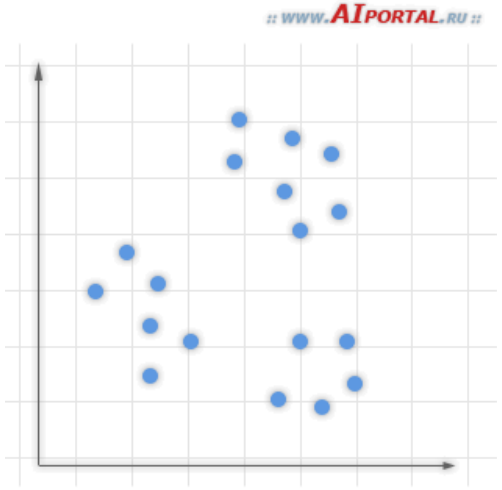
Задается параметр R – радиус поиска локальных сгущений.

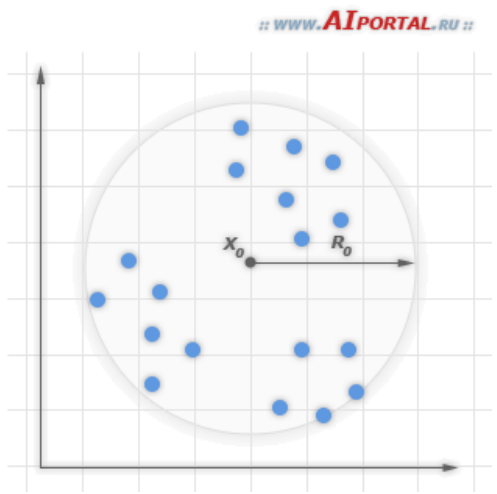
На каждом шаге мы
1. случайным образом выбираем объект из выборки,
2. раздуваем вокруг него сферу радиуса R,
3. внутри этой сферы выбираем центр тяжести и делаем его центром новой сферы.

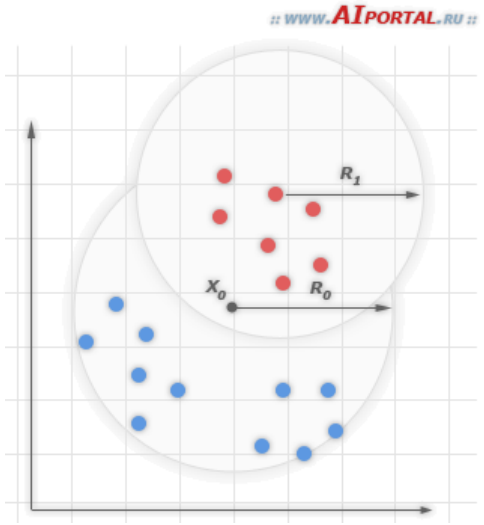
Таким образом, мы на каждом шаге двигаем сферу в сторону локального сгущения объектов выборки, т.е. стараемся захватить как можно больше объектов выборки сферой фиксированного радиуса.

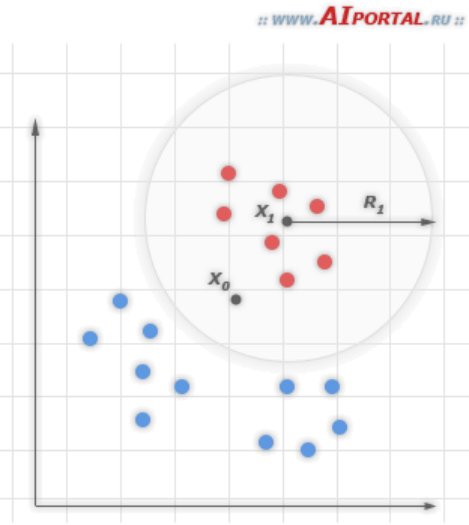
4. После того как центр сферы стабилизируется, все объекты внутри сферы с этим центром мы помечаем как кластеризованные и выкидываем их из выборки. Этот процесс мы повторяем до тех пор, пока вся выборка не будет кластеризована.

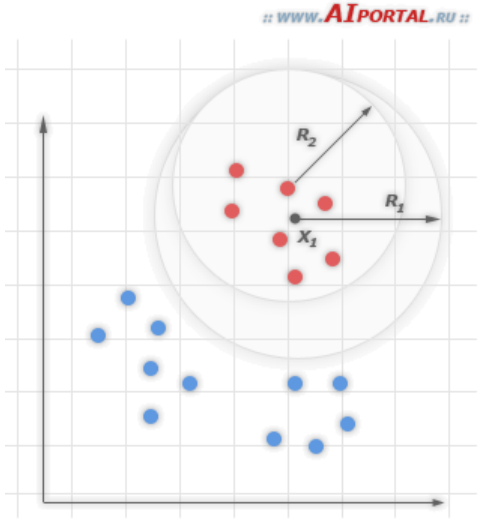
Визуализация алгоритма семейства FOREL

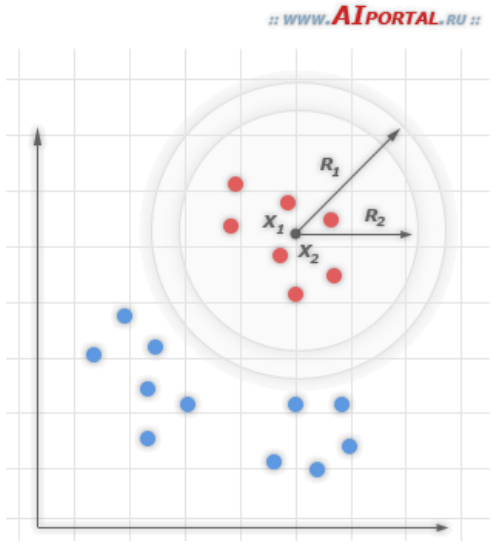


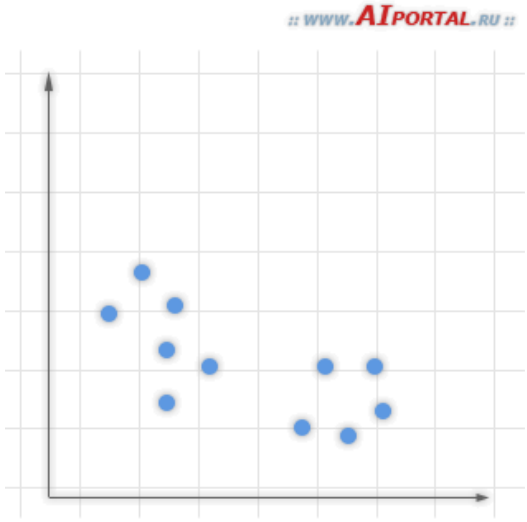












Свойства алгоритма FOREL

Преимущества:

- Точность минимизации функционала качества (при удачном подборе параметра R)
- Наглядность визуализации кластеризации
- Сходимость алгоритма
- Возможность подсчета промежуточных функционалов качества, например, длины цепочки локальных сгущений

Недостатки:

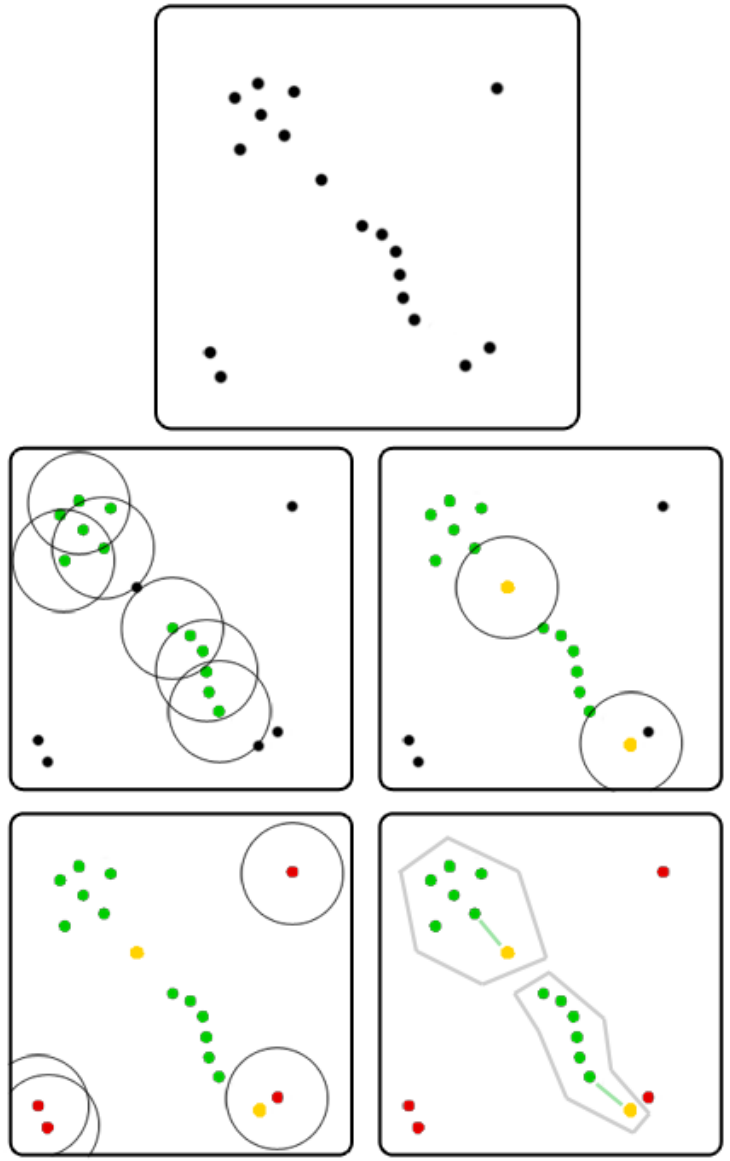
- Относительно низкая производительность
- Плохая применимость алгоритма при плохой разделимости выборки на кластеры
- Неустойчивость алгоритма (зависимость от выбора начального объекта)
- Произвольное по количеству разбиение на кластеры
- Необходимость априорных знаний о ширине (диаметре) кластеров

DBSCAN

Density-based spatial clustering of applications with noise, плотностной алгоритм пространственной кластеризации с присутствием шума

- p - «базовая» точка, если в *e*-окрестности находится minPts соседей.
- точка q прямо достижима из p, если находится на расстоянии не больше *e*
- точка r достижима из p, если к ней имеется путь прямо достижимых точек
- все точки, не достижимые из базовых являются выбросами
- все точки, достижимые из p, но не имеющие minPts соседей – края кластера

- Если точка p является базовой, то она формирует кластер со всеми достижимыми из нее точками.



DBSCAN

Гиперпараметры:

- e – радиус поиска локальных сгущений (соседства)
- minPts – минимальное количество соседей для образования кластера

Достоинства

- Хорошо справляется с большим объемом данных
- Хорошо работает с кластерами произвольной формы и размера
- Позволяет находить выбросы
- Количество кластеров определяется в процессе алгоритма

Недостатки

- Кластеры должны быть отделены друг от друга или, как минимум, иметь разреженное межкластерное пространство

Самоорганизующиеся карты Кохонена

Самоорганизующаяся карта Кохонена (Self-Organizing Map, SOM) – модель искусственной нейронной сети, способной обучаться без учителя. Предложена в 1984 году Теуво Кохоненом.
Целью применения данной сети является поиск скрытых закономерностей в данных, основываясь на снижении размерности исходного пространства в пространство меньшей размерности (на практике чаще всего используется двумерное, по причине, в частности, удобной визуализации). При этом топология исходного пространства остается той же самой. В результате обучения данной модели получается решетка, состоящая из обученных нейронов, она же и называется "картой" исходного пространства.

Самоорганизующиеся карты Кохонена. Алгоритм

Дано:

Выборка **X**, состоящая из объектов $X_i = \{x_1,....x_d\}$, i = 1…N, d – размерность данных. Размерность **X** = [d*N]

W – матрица весов нейронов размерностью [d*M]. $W_j = \{w_1,...w_d\}$ M – количество нейронов.

D (M*M) – матрица расстояний между нейронами в слое (топология нейронов). При этом, эта матрица – не то же самое, что $|W_i - W_j|$.

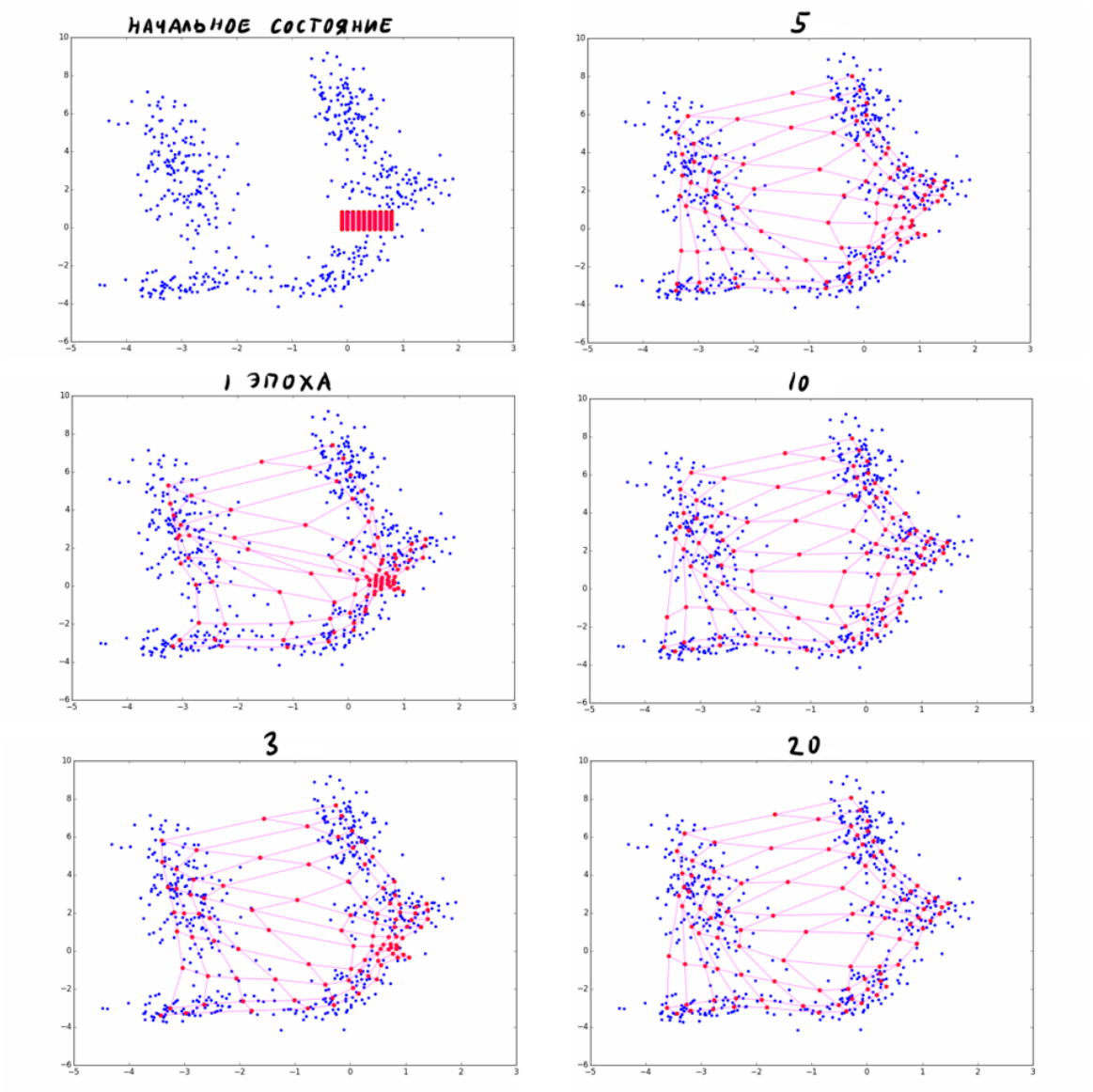
α, γ – показатели кооперации и затухания скорости обучения соответственно.

E – количество эпох.

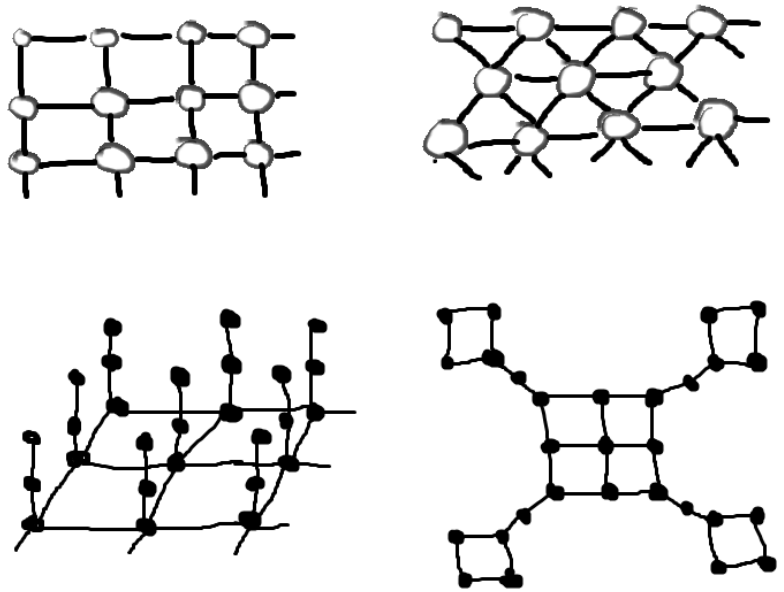
Алгоритм:

1. Инициализировать *W*
2. *E* раз:
    1. Для каждого объекта X_i из **X:**
        1. p = arg min $||Xi - Wi||$ - находим ближайший нейрон к объекту
        2. $\forall j: w_j = w_j + \mu h_{p,j}(x - w_j)$
        3. $\forall j: h_{p,j} = e^{D_{p,j}/\sigma\ (t)}$ - функция «растяжимости». t – зависимость от эпохи, например $\sigma\ (t) = (\sigma^{-bt})$
    2. $\mu = \mu * \gamma$
    3. $\sigma = \sigma * \alpha$

Самоорганизующиеся карты Кохонена. Топология сети

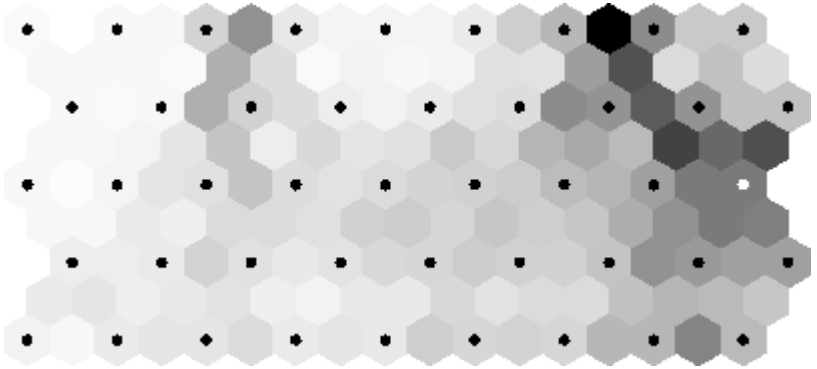


D – топология нейронов.
Не обязательно сетка, в общем случае – ограничивается фантазией, но в случае знания структуры выборки, должно выбираться удобным для визуализации.



U-matrix

Для решения задачи кластеризации строится U-матрица – унифицированная матрица расстояний (unified distance matrix). Это разновидность визуализации карт Кохонена, где вместе со значениями нейронов показывается расстояние между ними. Таким образом можно легко можно определить, где кластеры соприкасаются, а где проходит граница.



Здесь черные точки – нейроны. Чем темнее ячейка, тем больше расстояние между данными нейронами в матрице W и соответственно, разрывами между объектами выборки X
Таким образом, светлые области можно рассматривать как кластеры, темные – как межкластерное пространство.