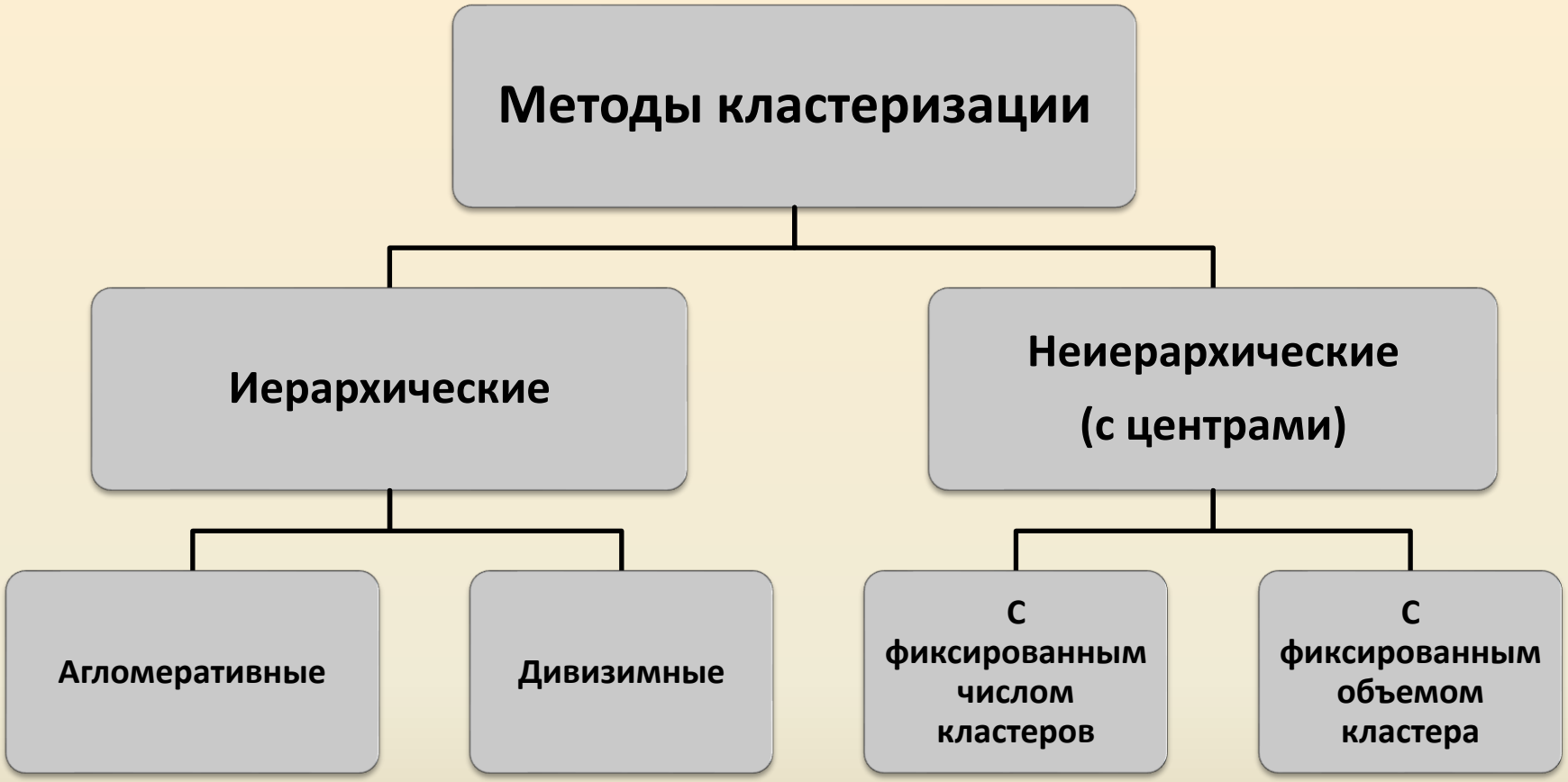
Кластеризация

- Кластеризация – выделение групп схожих объектов
- Применение кластеризации:

  - группировка результатов поиска;

  - сокращение выборки за счет выбора представителей кластеров

  - поиск схожих характеристик объектов
- Алгоритмы кластеризации строятся на сравнении объектов между собой в соответствии с некоторой мерой близости
- Проблемы кластеризации: выбор меры близости, оценка качества кластеризации, обоснованность результатов, разнотипные данные



Single-link, Complete-link и др.

MST-алгоритм

K-medians, K-medoids fuzzy K-means

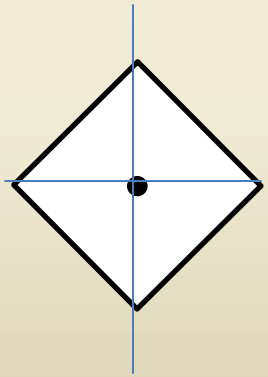
QT-алгоритм, ФОРЕЛЬ

Расстояния между объектами

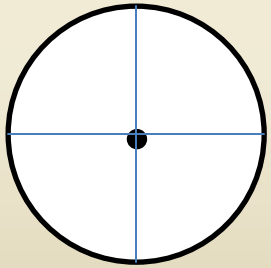
Расстояния: семейство Minkovsky

$$d_k(x, y) = \left(\sum_{i=1}^{n} (x_i - y_i)^k \right)^{1/k}$$

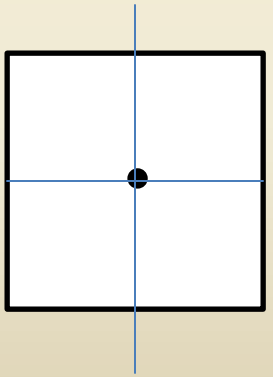
k = 1
Manhattan



k = 2
Euclidean



k $\rightarrow \infty$
Maximal



Меры близости

- Расстояние Евклида

$$d_E(X,Y) = \sqrt{\sum_{i=1}^{M}(x_i - y_i)^2}$$

- Расстояние манхэттана

$$d_M(X,Y) = \sum_{i=1}^{M}|x_i - y_i|$$

- Расстояние Чебышева

$$d_{\max}(X,Y) = \max_{i=1..M}|x_i - y_i|$$

Расстояния

- Расстояние Canberra:

$$d(x,y)=\sum_{i=1}^{M}\frac{|x_i-y_i|}{|x_i|+|y_i|}$$

- Косинусное ~~расстояние~~:
  (косинус угла между векторами)

$$d(x,y)=\frac{\vec{x}\cdot\vec{y}}{\|\vec{x}\|\times\|\vec{y}\|}$$

- Расстояние по Хэммингу:
  (число различных признаков)

$$d(x,y)=\sum_{i=1}^{M}[x_i\neq y_i]$$

- Коэффициент Жаккарда:
  (для номинальных пок-лей)

$$J(x,y)=\frac{\sum_{i=1}^{M}[x_i=1\wedge y_i=1]}{\sum_{i=1}^{M}[x_i=1\vee y_i=1]}$$

Иерархическая кластеризация

- В результате иерархической кластеризации получается структура вложенных кластеров
- Диаграмма объединения (или разделения) кластеров называется ***дендрограммой***
- *Агломеративная кластеризация* (AGNES) последовательно объединяет ближайшие кластеры, начиная с кластеров, содержащих по одному объекту и заканчивая одним кластером, содержащим все объекты
- *Дивизимная кластеризация* (DIANA) последовательно разделяет удаленные подмножества кластера

Агломеративная кластеризация

$AGLOMERATIVE\_CLUSTERING(D)$

$R = DistanceMatrix(D)$

$K = AssignClusters(D)$

while $|K| > 1$

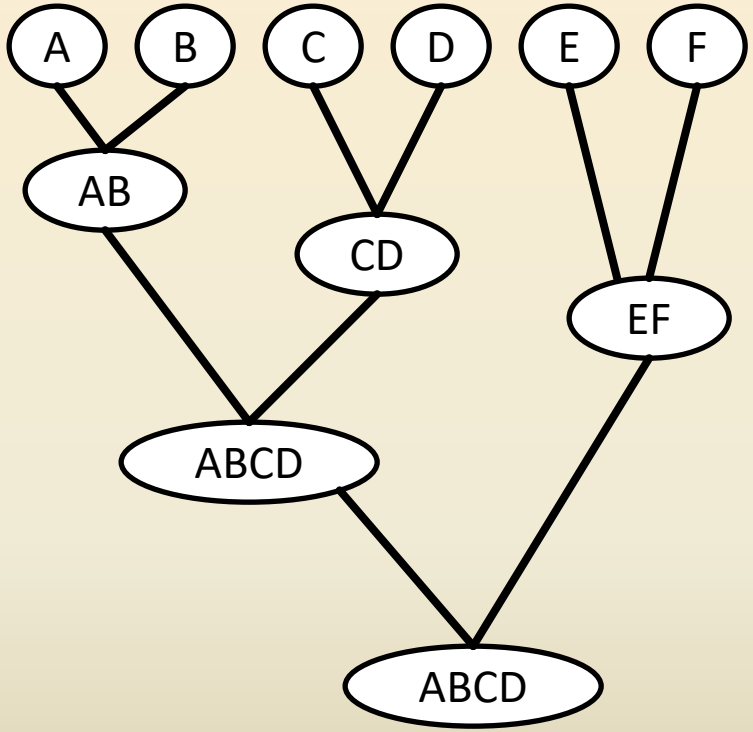
$(i, j) = findClosestClusters(R, K)$

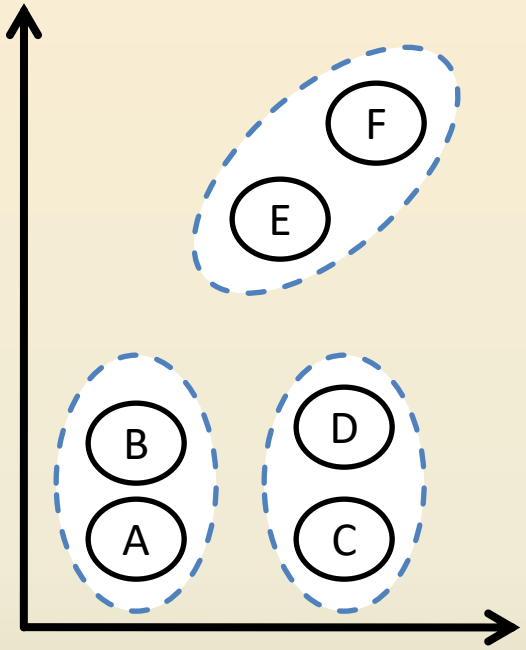
$K_i = K_i \cup K_j$

$K = K \setminus K_j$

Параметры алгоритма: мера близости для вычисления расстояния между парой объектов (Евклидово расстояние, Манхэттана и др), метод вычисления расстояния между кластерами (метод одиночной связи, метод полной связи, средней связи).

Иерархическая кластеризация





Вычисление расстояния между кластерами

- Метод одиночной связи (single-link)

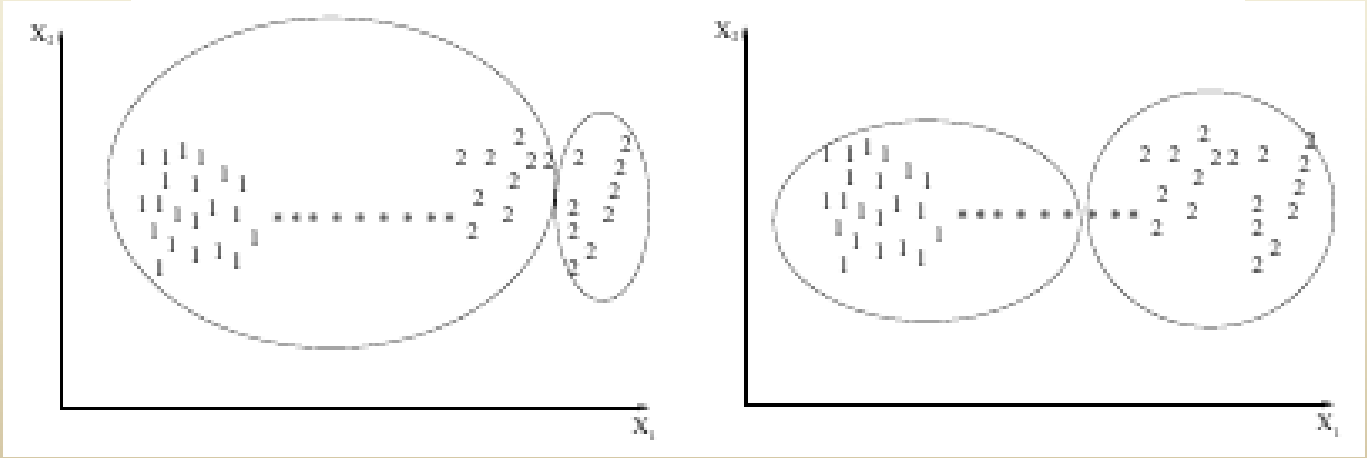
$$D(K_i, K_j) = min\{\, d(x,y) \mid x \in K_i, y \in K_j \}$$

- Метод полной связи (complete-link)

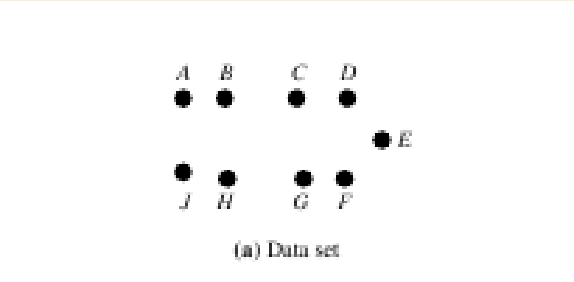
$$D(K_i, K_j) = max\{\, d(x,y) \mid x \in K_i, y \in K_j \}$$

- Метод средней связи (average-link)

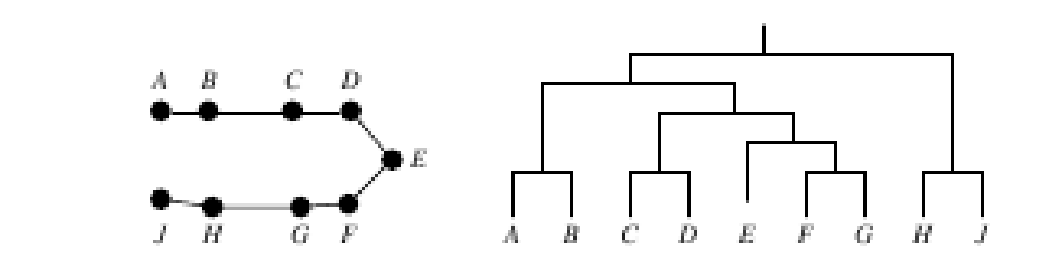
$$D(K_i, K_j) = avg\{\, d(x,y) \mid x \in K_i, y \in K_j \}$$



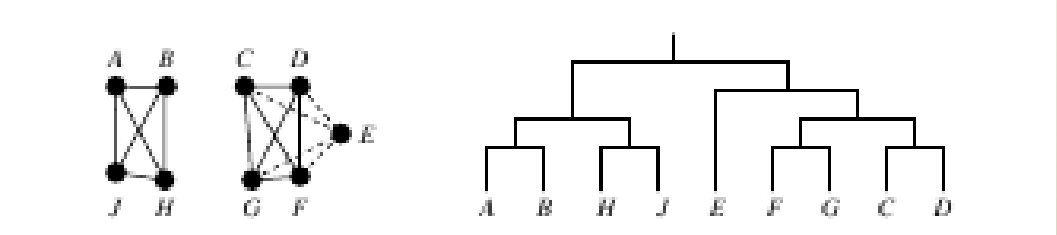
Complete-link VS. Single-link



Single-link



Complete-link



Иерархическая кластеризация Уорда

Для каждого кластера вычисляется сумма квадратов отклонений признаков от средних значений

$$V_k = \sum_{i=1}^{n_k} \sum_{j=1}^{m} (x_{ij} - x^*_{jk})^2$$

k - номер кластера, m – число признаков,

n_k - число объектов в k-кластере

x^*_{jk} - среднее значение j-признака в k-кластере

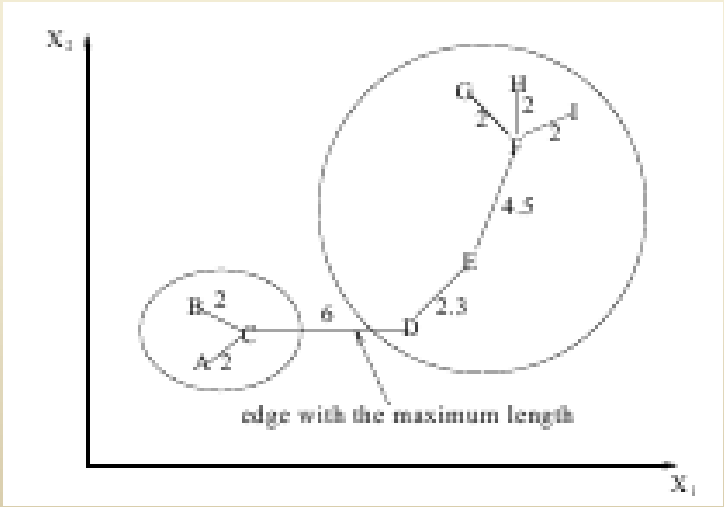
Алгоритм: последовательно объединяются ближайшие кластеры, которые обеспечивают наименьшие значения V_k

(дивизимный) MST-алгоритм

- Алгоритм использует представление данных в виде графа – каждый объект соответствует вершине в многомерном пространстве.
- В результате строится иерархическая кластеризация (дивизимная)

Алгоритм:

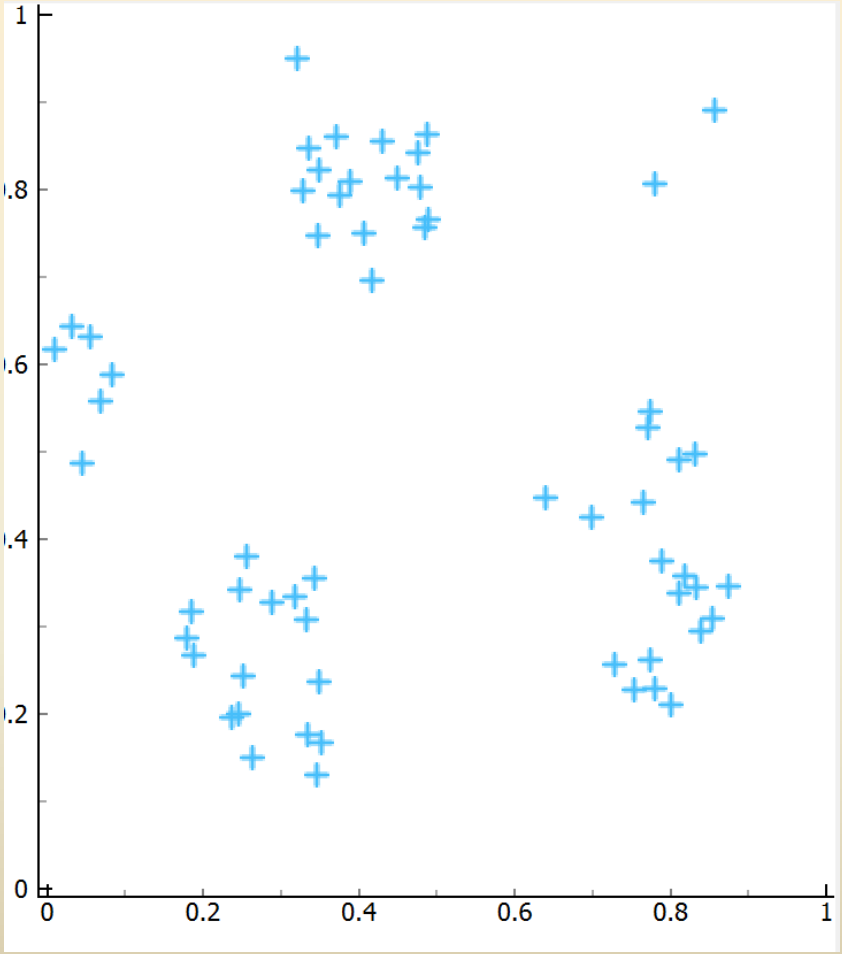
1. Построить минимальное охватывающее дерево по точкам данных (алгоритмы Прима, Борувка и др.)
2. Последовательно удалять самое длинное ребро дерева. При этом кластер разделяется на две части.



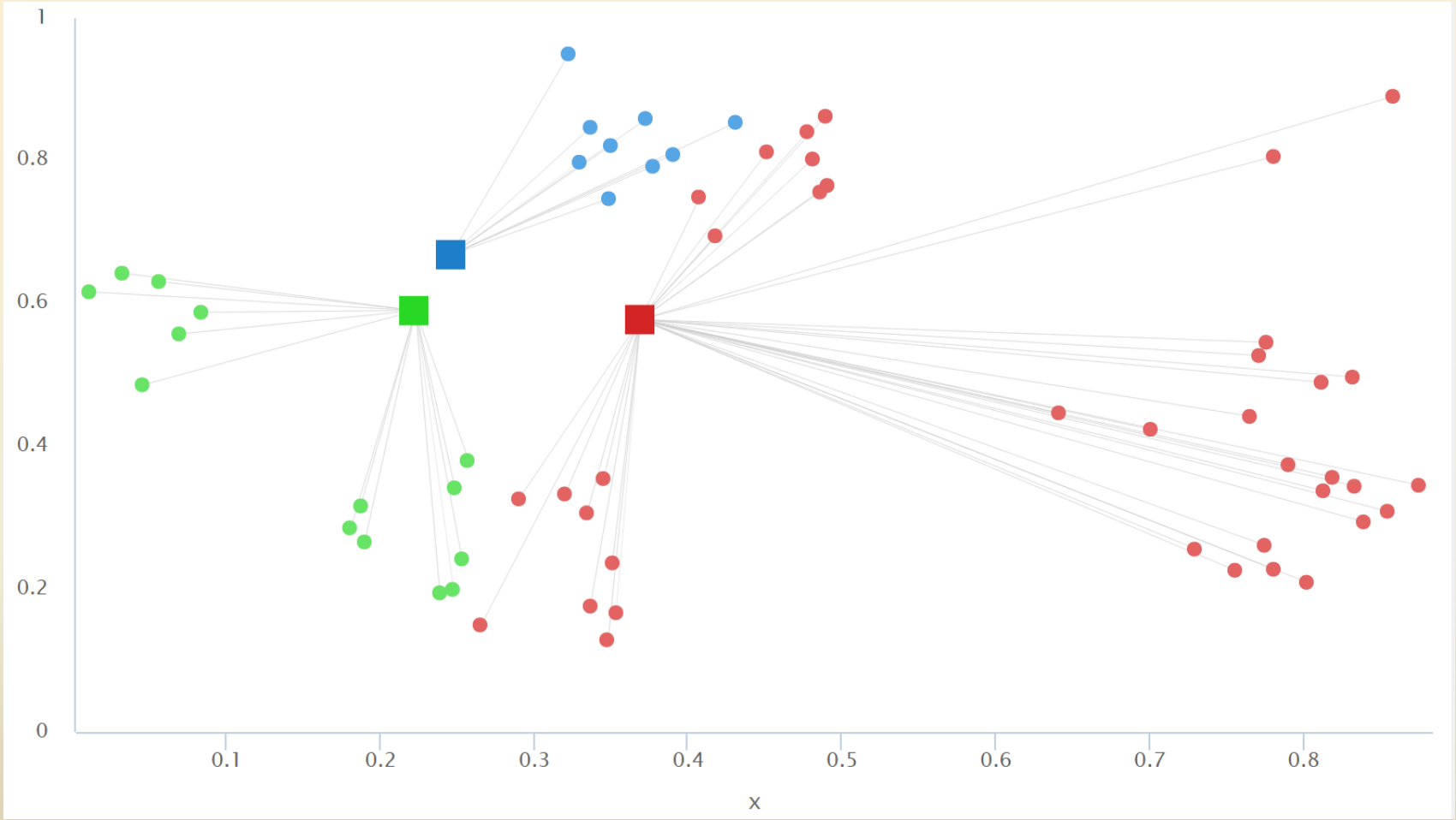
Алгоритм k-средних (k-means)

- Для работы алгоритма необходимо: задать число кластеров k, установить начальное положение центров кластеров
- Алгоритм итеративно распределяет точки по кластерам и смещает центры кластеров
- Проблемы алгоритма: обоснованность выбора числа k, начальных центров кластеров, аномальные выбросы

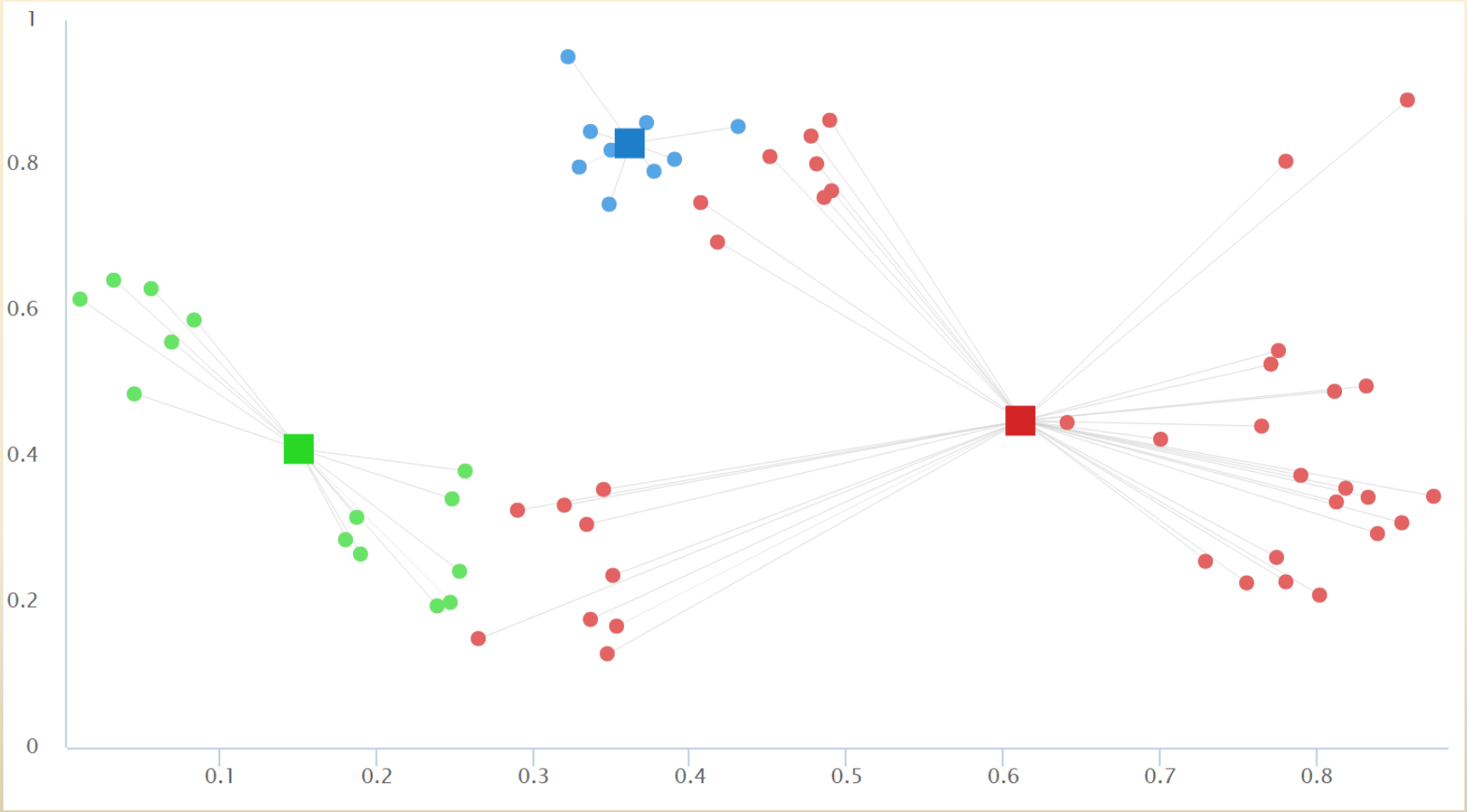
kMeans



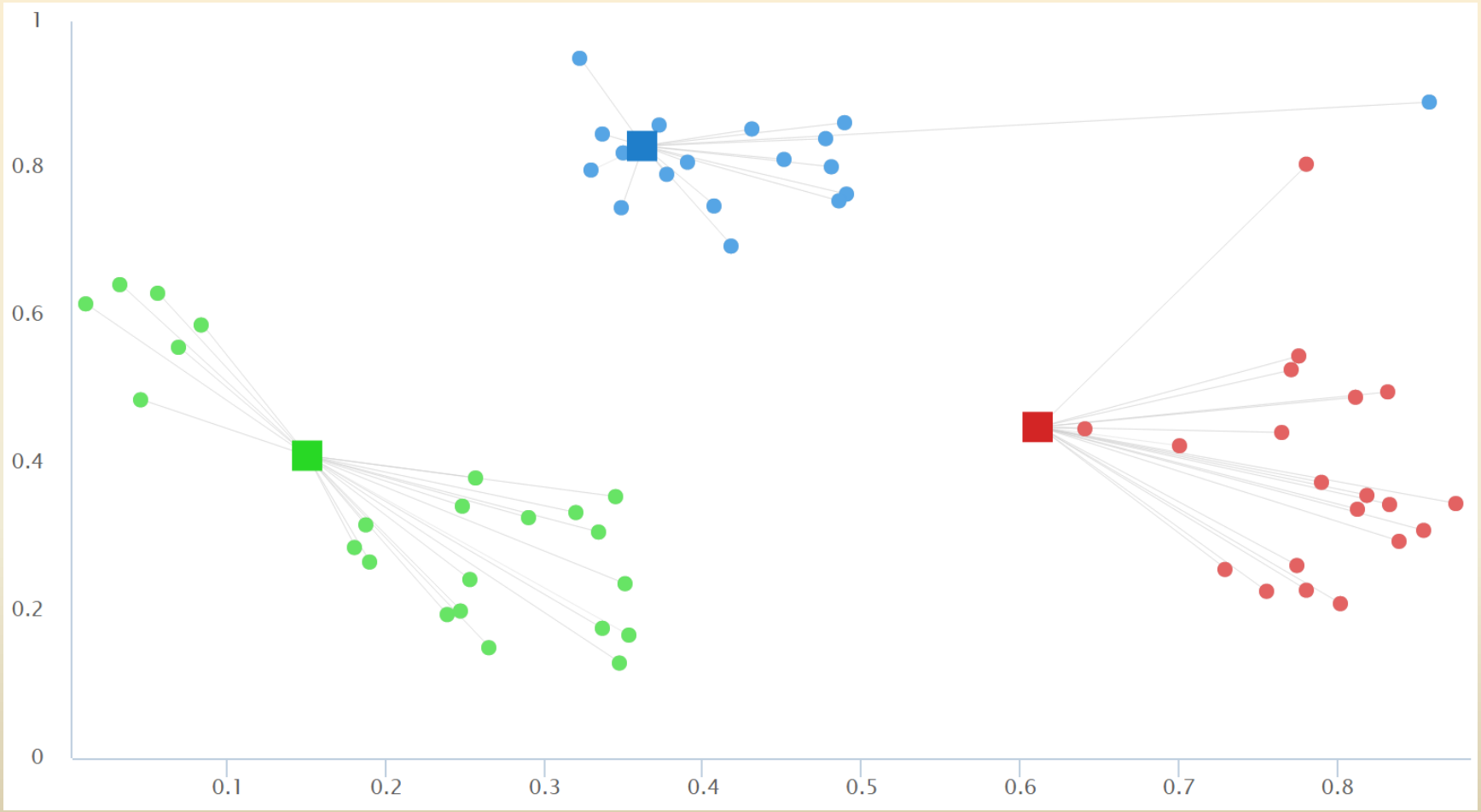
kMeans. Шаг 1-Распределение точек



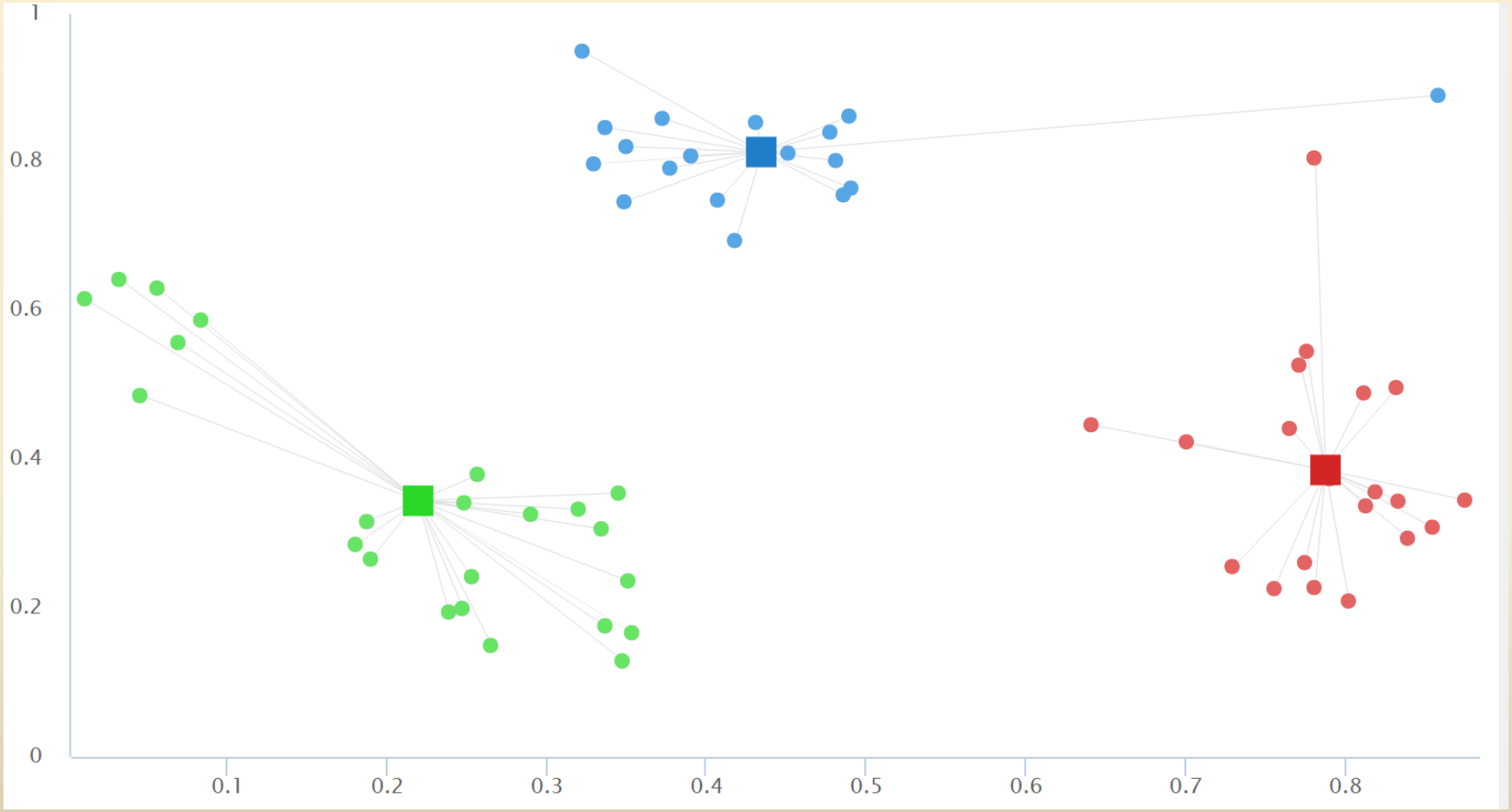
kMeans. Шаг 2-Смещение центров



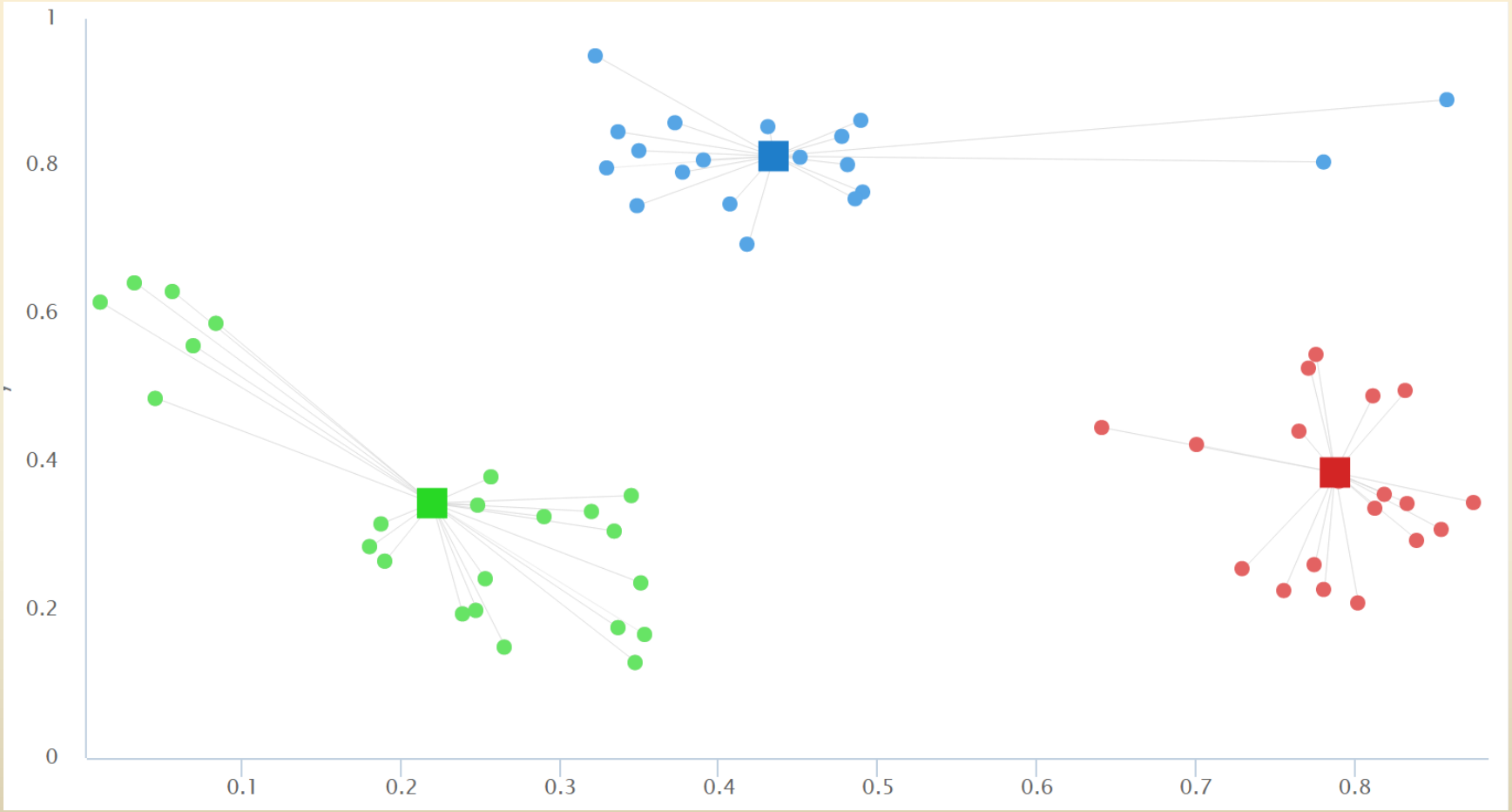
kMeans. Шаг 3-Распределение точек



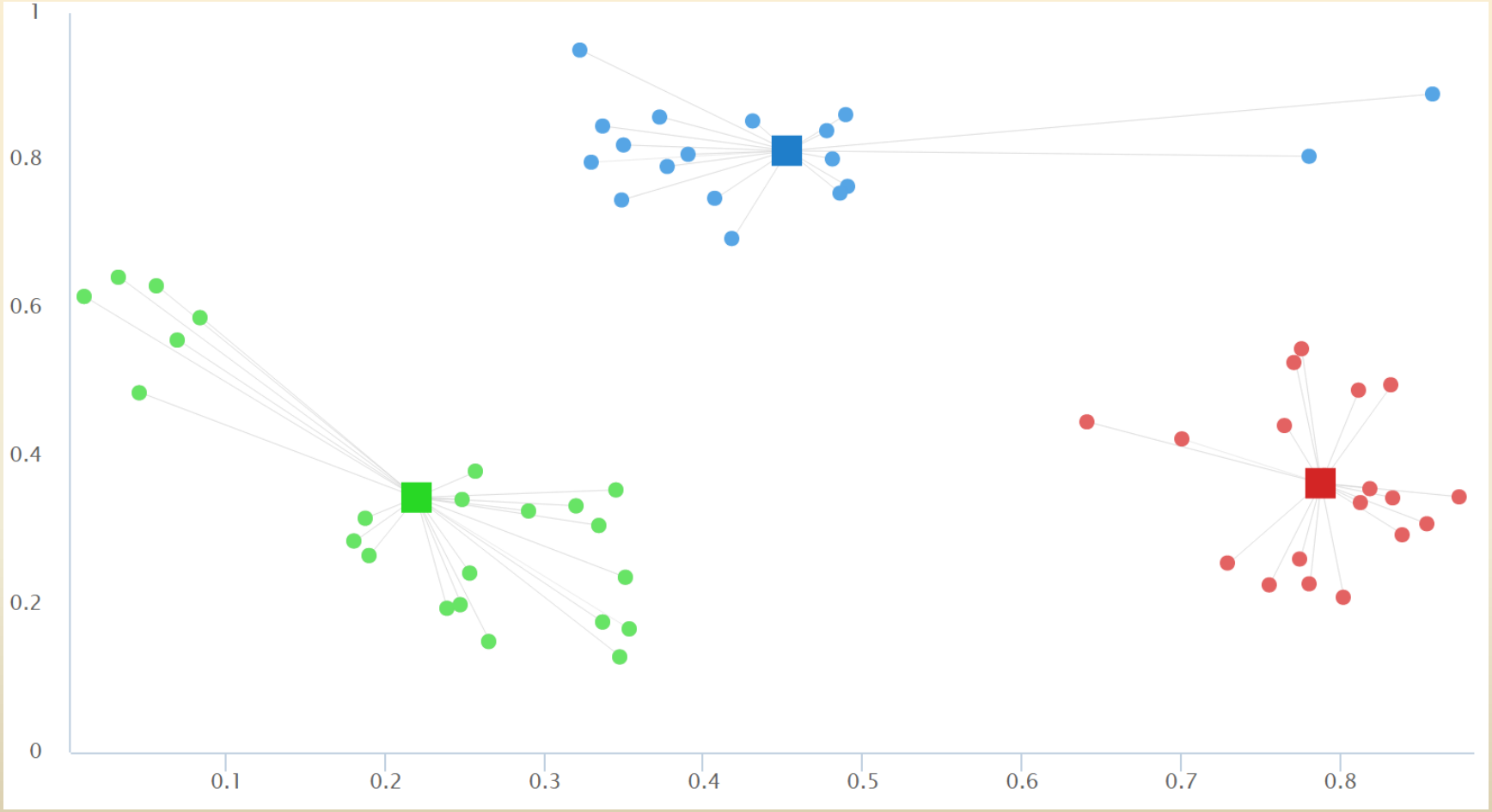
kMeans. Шаг 4-Смещение центров



kMeans. Шаг 5-Распределение точек



kMeans. Шаг 6-Смещение центров



Алгоритм K-means

$KMEANS(k, D)$

ВХОД: k – *число кластеров,* D – *множество данных*

ВЫХОД: $\{K_i\}$

$\{Center_i\} = InitCenters(k, D)$

do

foreach $x \in D$

foreach $c \in C$

$d_c = Distance(Center_c, x)$

$c^* = \arg\min \{d_c \mid c \in C\}$

$class(x) = c^*$

$\{Center_i\} = UpdateCenters(D, bCentersChanged)$

loop while $bCentersChanged$

for $i = 1$ to k

$K_i = \{x \mid class(x) = i\}$

Качество кластеризации k-means

Качество кластеризации методом k-means можно оценить по формуле:

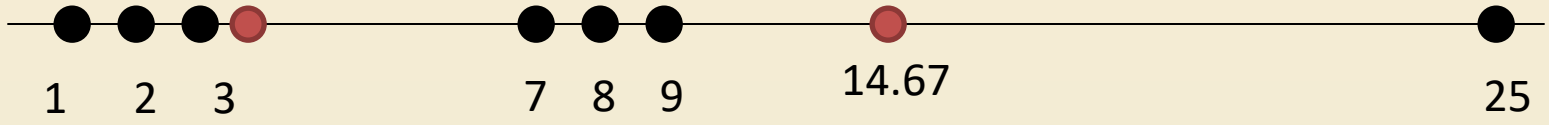
$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(p, c_i)^2$$

Для каждой точки данных *p* оцениваем её расстояние до центра кластера.

Итоговая ошибка вычисляется как сумма квадратов расстояний по всем кластерам

Недостатки K-means

- Один из недостатков метода k-means: оперирование средними значениями. Это приводит к чувствительности к аномальным выбросам.



Разбиение 1: {1, 2, 3}, {7,8,9,25}; E = 196

Разбиение 2: {1, 2, 3, 7}, {8, 9, 25}; E = 189.67

- Центр кластера является «фиктивным» объектом.

k-medoids

- В модификации k-medoid центр кластера – это центрально-расположенная точка данных
- Для поиска центров (medoids) применяются:

  - выбор центрально-расположенного объекта

  - перебор точек в качестве центров
- Центр кластера – такая точка кластера, которая обеспечивает минимум суммы отклонений по точкам кластера

$$\sum_{p_i \in C} d(p_i, center_C)$$

QT-кластеризация

- Quality threshold

$$\text{QT\_Algorithm}(D, R)$$

$$\text{K} = \varnothing$$

$$\text{while } D \neq \varnothing$$

$$\text{foreach } p \in D$$

$$C_p = \left\{ q \in D \,\middle|\, dist(p, q) \leq R \right\}$$

$$C^* = \max_p \left| C_p \right|$$

$$K = K \bigcup C^*$$

$$D = D \setminus C^*$$

Форель-кластеризация

$Форель(D, R)$

$K = \varnothing$

while $D \neq \varnothing$

$p = SelectRandomPoint(D)$

$center = p$

do

$C^* = \{q \in D \mid dist(center, q) \leq R\}$

$center = UpdateCenter(C^*)$

while $Changed$

$K = K \cup C^*$

$D = D \setminus C^*$