Методы кластеризации

Лекция 16

План лекции

- Введение
- Формальная постановка задачи
- Метод k-средних
- Метод ISODATA
- Агломеративный метод
- Дивизимный метод

Введение

- Задача кластеризации состоит в разделении исследуемого множества объектов на группы «похожих» объектов, называемых *кластерами*
- Решение задачи кластеризации называют *кластерным анализом*

Введение

- Кластеризация отличается от классификации тем, что этап обучения на примерах отсутствует
- В задачах классификации множество классов заранее известно, в кластеризации классы определяются в процессе анализа
- Поэтому кластеризация относится к задачам *обучения без учителя* (unsupervised learning)

Введение

- Эта задача решается на начальных этапах исследования, когда о данных мало что известно
- Ее решение помогает лучше понять данные
- После определения кластеров применяются другие методы Data Mining, чтобы попытаться установить, что означает такое разбиение

Введение

- Кластерный анализ позволяет рассматривать достаточно большой объем информации и сжимать большие массивы информации,
делать их компактными и наглядными

Формальная постановка задачи

- Дано множество данных, состоящее из *N* объектов (векторов):

$$S_1, S_2, \ldots, S_N$$

- Каждый объект описывается набором признаков:

$$x_1, x_2, \ldots, x_m,$$

где *m* – размерность пространства признаков

Формальная постановка задачи

- Таким образом, *i*-й объект можно записать в виде:

$$S_i = (x_{i1}, x_{i2}, \ldots, x_{im})$$

- Класс для каждого объекта неизвестен

Формальная постановка задачи

Требуется:

- найти способ сравнения $d(S_p, S_q)$ объектов между собой (меру сходства, функцию расстояния)
- определить множество кластеров

$$C_1, C_2, \ldots, C_r$$

причем количество кластеров r – неизвестно

- разбить данные по кластерам

Формальная постановка задачи

В качестве меры сходства используются:

- евклидово расстояние
- квадрат евклидова расстояния
- расстояние Хэмминга
- расстояние Чебышева

Формальная постановка задачи

Методы кластерного анализа можно разделить на две группы:

- неиерархические
- иерархические

Метод k-средних

- Неиерархическим методом кластеризации является метод k-средних (k-means)
- Предварительно необходимо выбрать вероятное число кластеров *k*

Метод k-средних

1. Выбирается *k* произвольных исходных центров кластеров – обычно выбираются *k* объектов
2. Все объекты разбиваются на *k* групп, наиболее близких к одному из центров
3. Вычисляются новые центры кластеров
4. Проводится новое разбиение всех объектов на основании близости к новым центрам

Шаги 3 и 4 повторяются до тех пор, пока центры кластеров не перестанут меняться или пока не достигнуто максимальное число итераций

Метод k-средних

- Выбор числа кластеров является сложным вопросом
- Если нет предположений относительно этого числа, рекомендуют создать 2 кластера, затем 3, 4, 5 и т. д., сравнивая полученные результаты

Метод k-средних

Начальный выбор центров кластеров осуществляется следующим образом:

- выбор *k* объектов для максимизации начального расстояния
- случайный выбор *k* объектов
- выбор первых *k* объектов

Метод k-средних

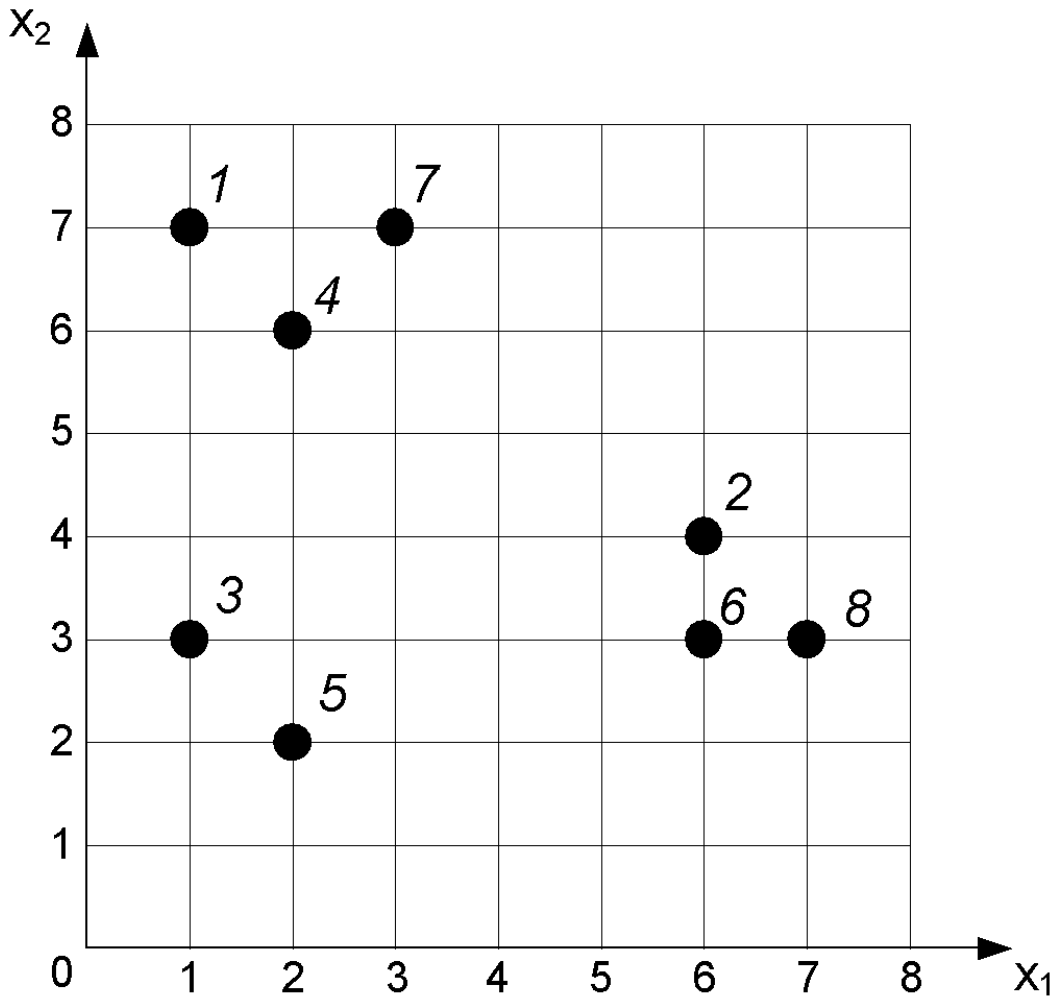
- Центры кластеров вычисляются по формулам:

$$x_{1C} = \frac{\sum_{i=1}^{N_C} x_{i1}}{N_C} \qquad x_{2C} = \frac{\sum_{i=1}^{N_C} x_{i2}}{N_C} \qquad \ldots \qquad x_{mC} = \frac{\sum_{i=1}^{N_C} x_{im}}{N_C}$$

где N_C – количество объектов, входящих в кластер C

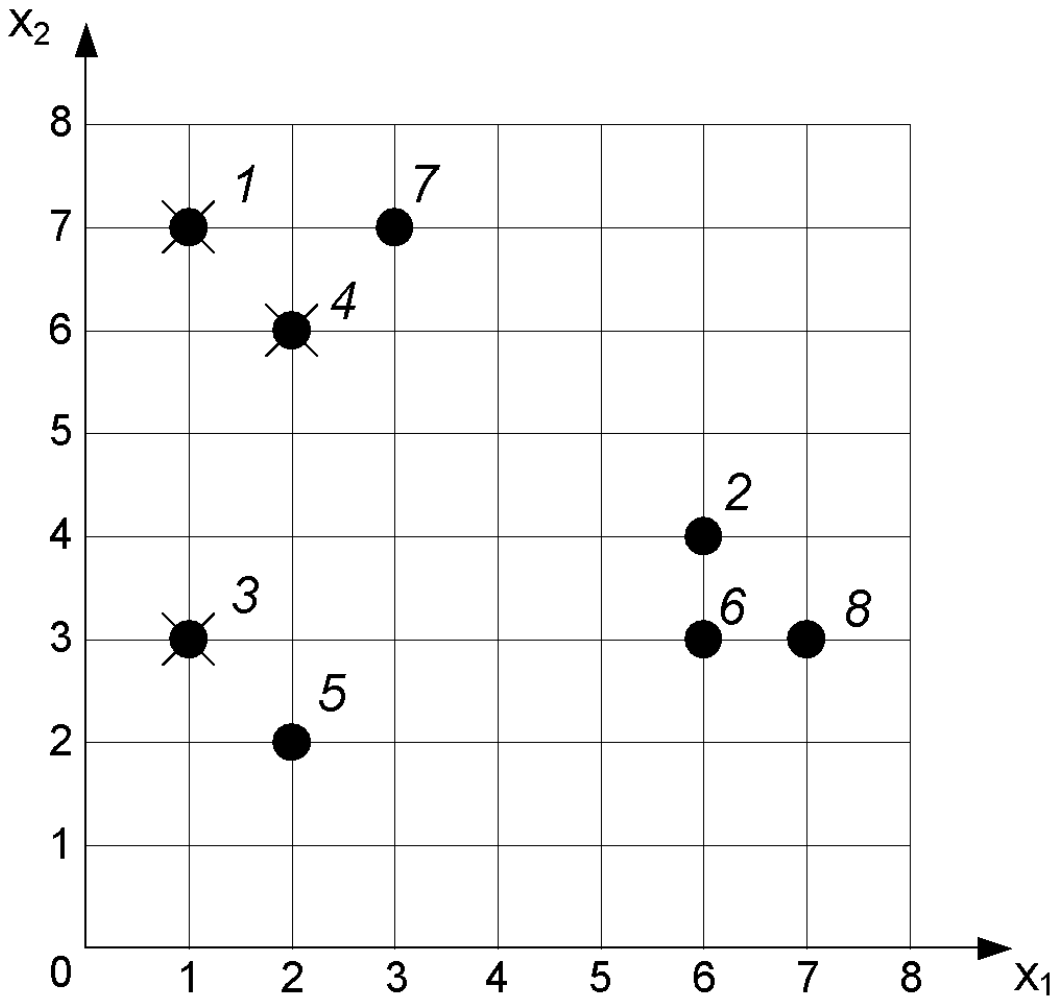
Метод k-средних

- ***Пример***.
- Примем $k = 3$
- Начальные центры – объекты 1, 3, 4



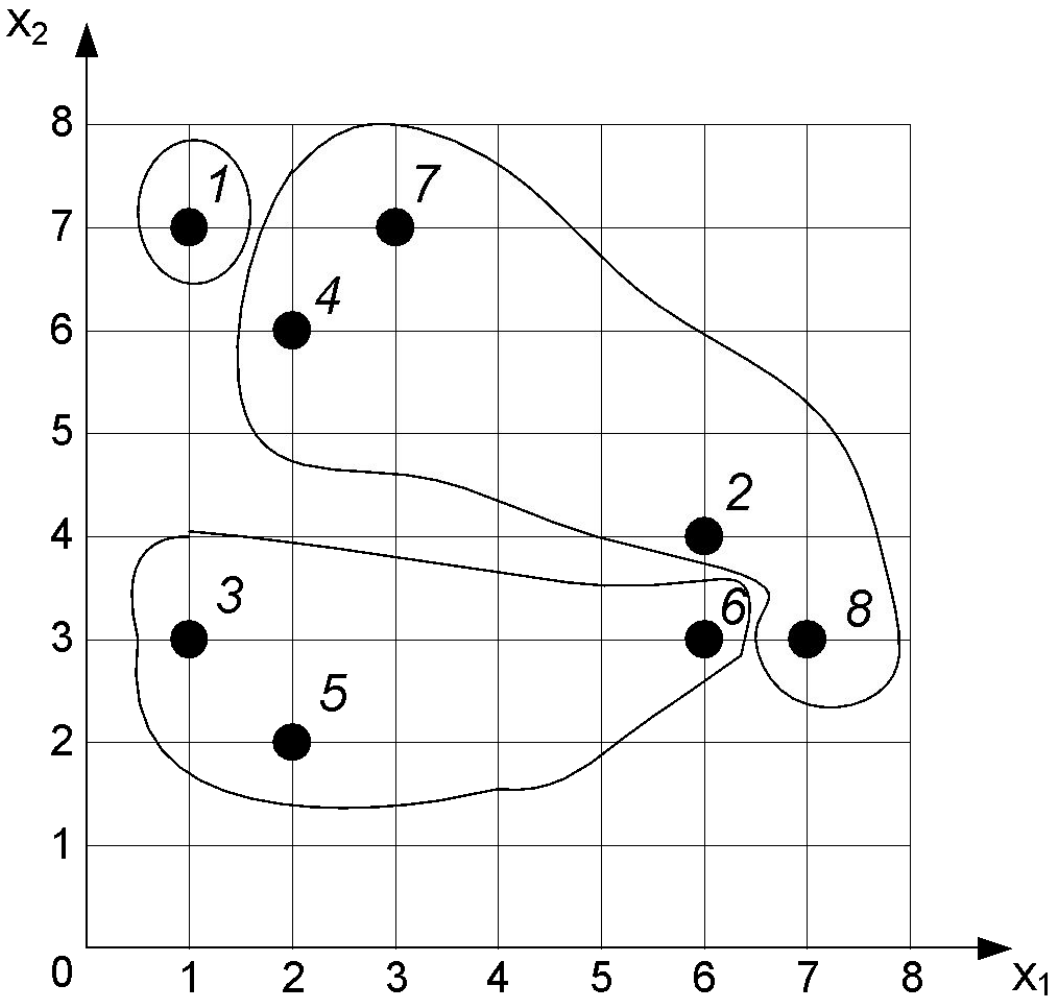
Метод k-средних

- ***Пример***.
- Примем $k = 3$
- Начальные центры – объекты 1, 3, 4
- Разобьем все объекты по кластерам



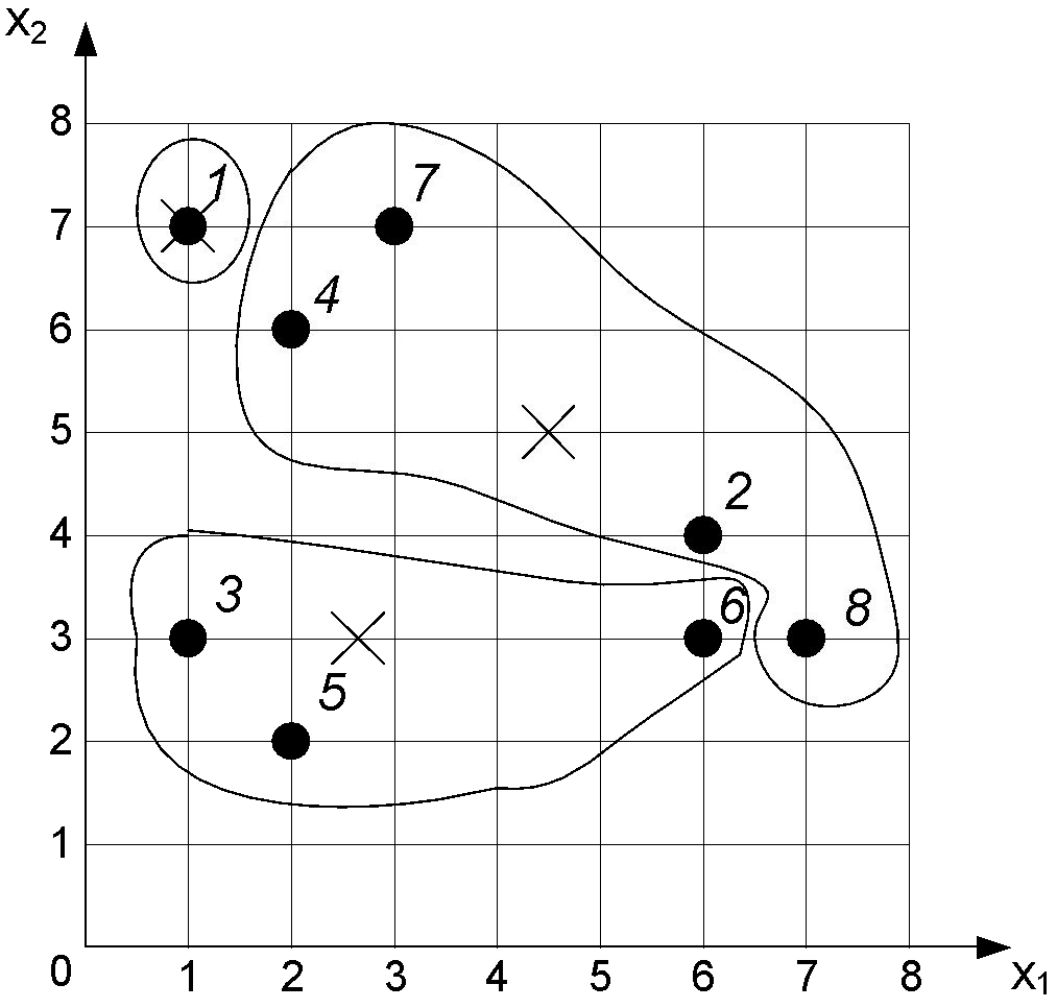
Метод k-средних

- Найдем новые центры кластеров



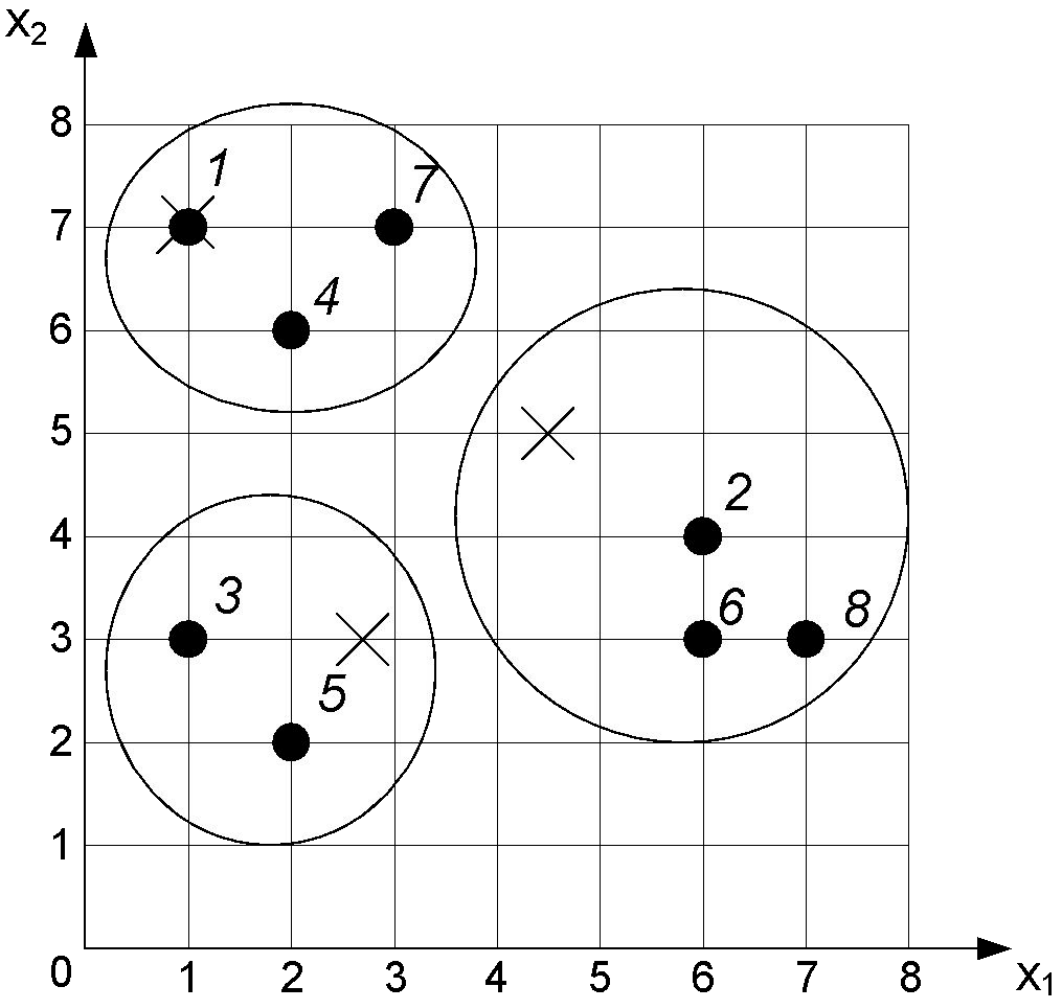
Метод k-средних

- Найдем новые центры кластеров
- Разобьем все объекты по новым кластерам



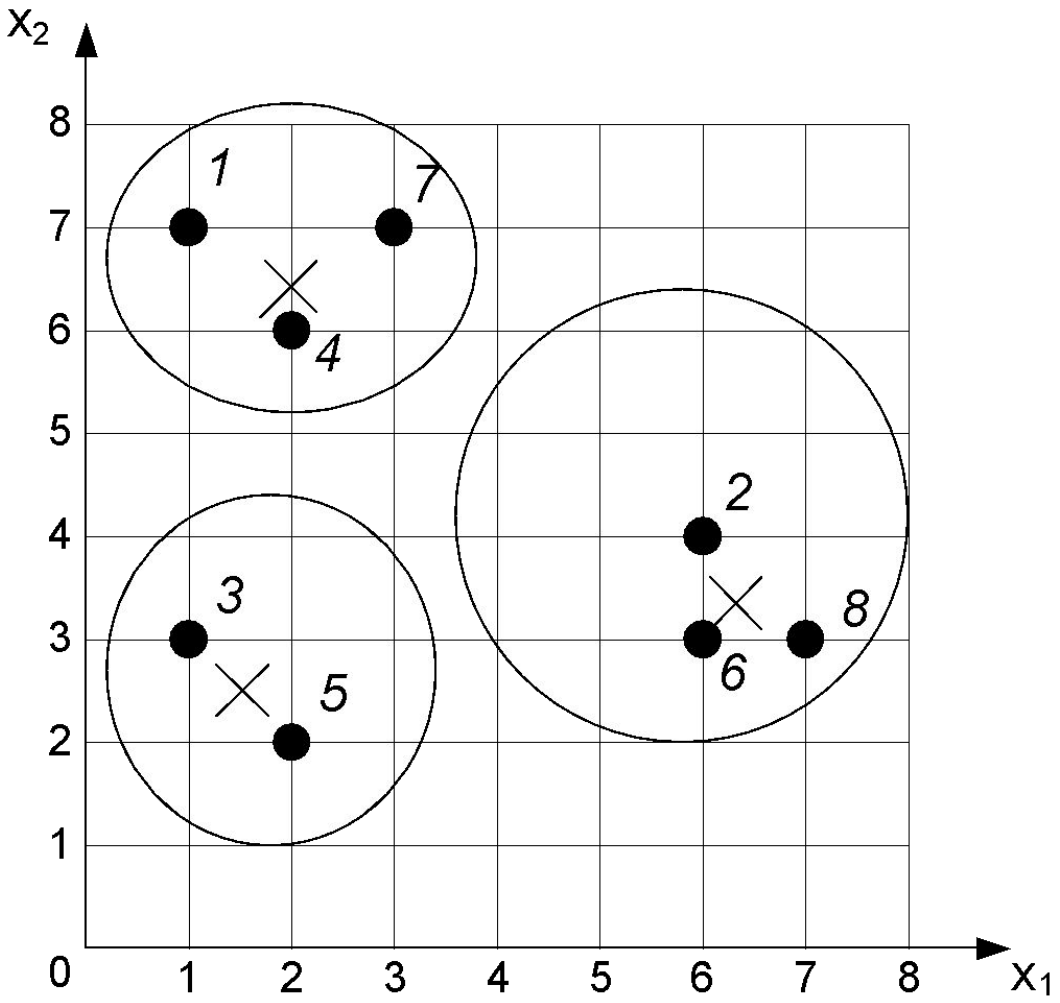
Метод k-средних

- Пересчитаем центры кластеров



Метод k-средних

- Разбивка объектов по новым кластерам не меняет расположение центров



Метод ISODATA

- ISODATA – Iterative Self-Organizing Data Analysis Techniques – итеративный самоорганизующийся метод анализа данных
- Более сложный алгоритм, чем k-means, дополненный несколькими эвристиками
- Полное описание см.
  Ту Дж., Гонсалес Р. «Принципы распознавания образов», М.: Мир, 1978

Метод ISODATA

- Если в кластер входит менее заданного минимального числа объектов, кластер удаляется
- Если среднее расстояние между объектами кластера больше заданного максимального порога, кластер расщепляется на два новых кластера

Метод ISODATA

- Если расстояние между центрами двух кластеров меньше заданного минимального порога, кластеры сливаются
- В алгоритме ISODATA множество параметров, настройка которых представляет определенные трудности

Иерархические методы

К иерархическим методам кластеризации относятся:

- агломеративный алгоритм (Agglomerative Nesting, AGNES)
- дивизимный алгоритм (Divisive ANAlysis, DIANA)

Агломеративный метод

- В начале работы алгоритма все объекты являются отдельными кластерами
- На первом шаге наиболее похожие (близкие) два кластера объединяются в дин кластер
- На последующих шагах объединение продолжается до тех пор, пока все объекты не будут составлять один кластер
- На любом этапе объединение можно прервать, получив нужное число кластеров

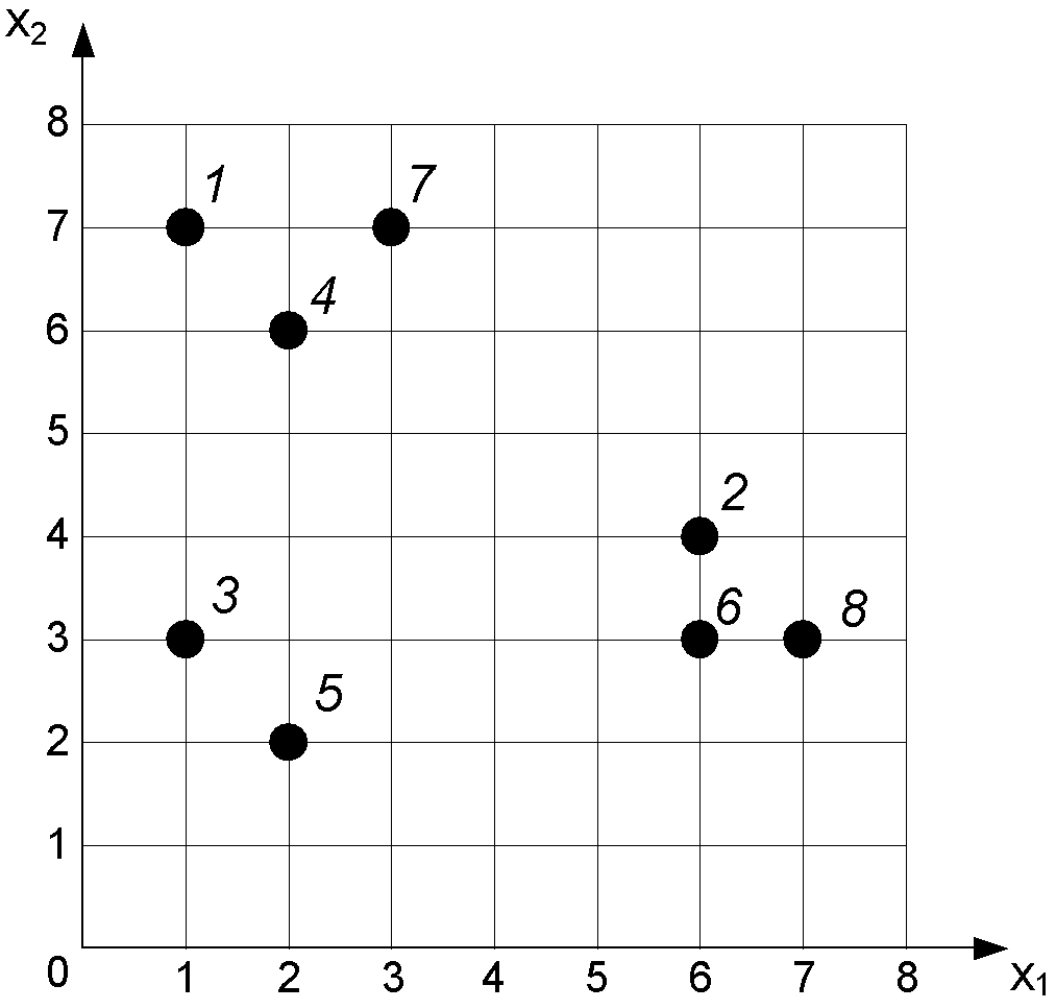
Агломеративный метод

Расстояние между кластерами можно определить различными способами:

- расстояние между центрами кластеров
- расстояние между двумя наиболее близкими объектами в кластерах
- расстояние между двумя наиболее дальними объектами в кластерах
- среднее расстояние между всеми парами объектов в них

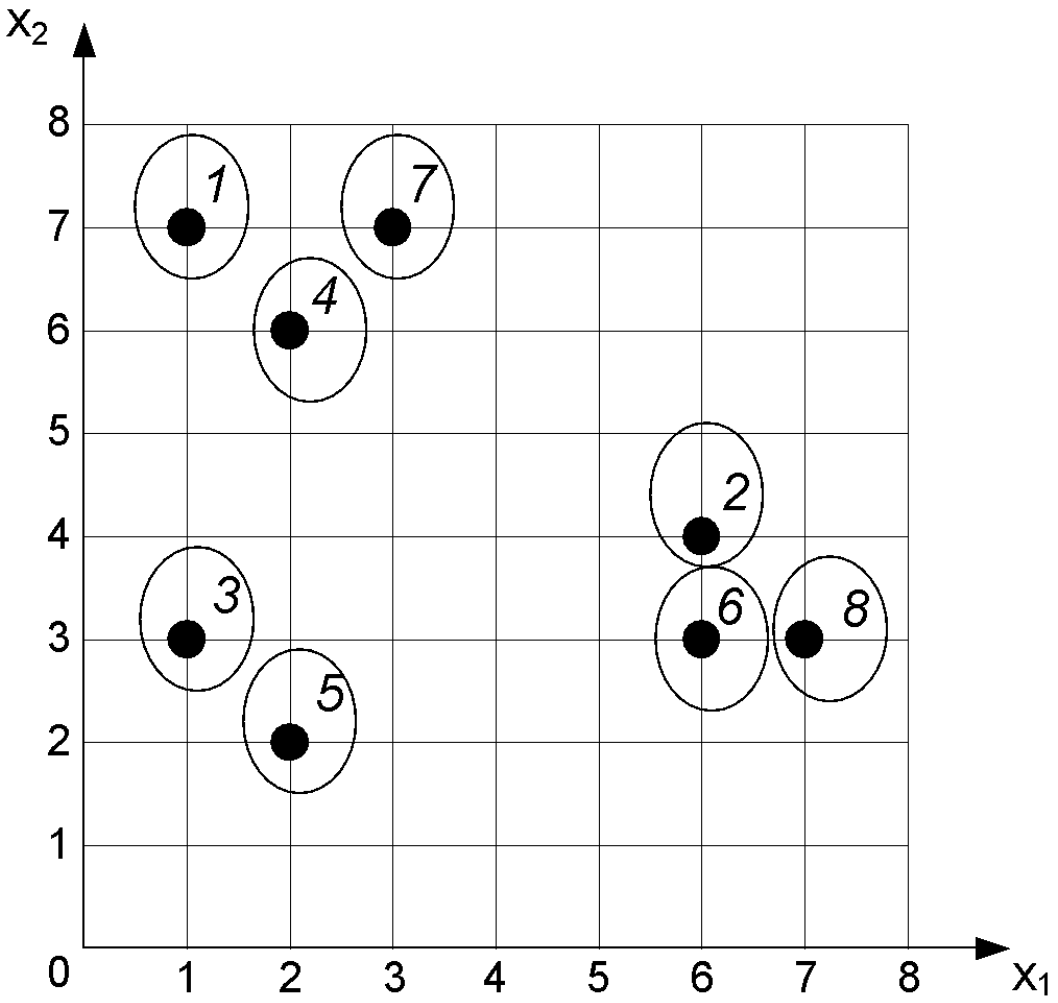
Агломеративный метод

- ***Пример***.
- Каждый объект формирует свой кластер



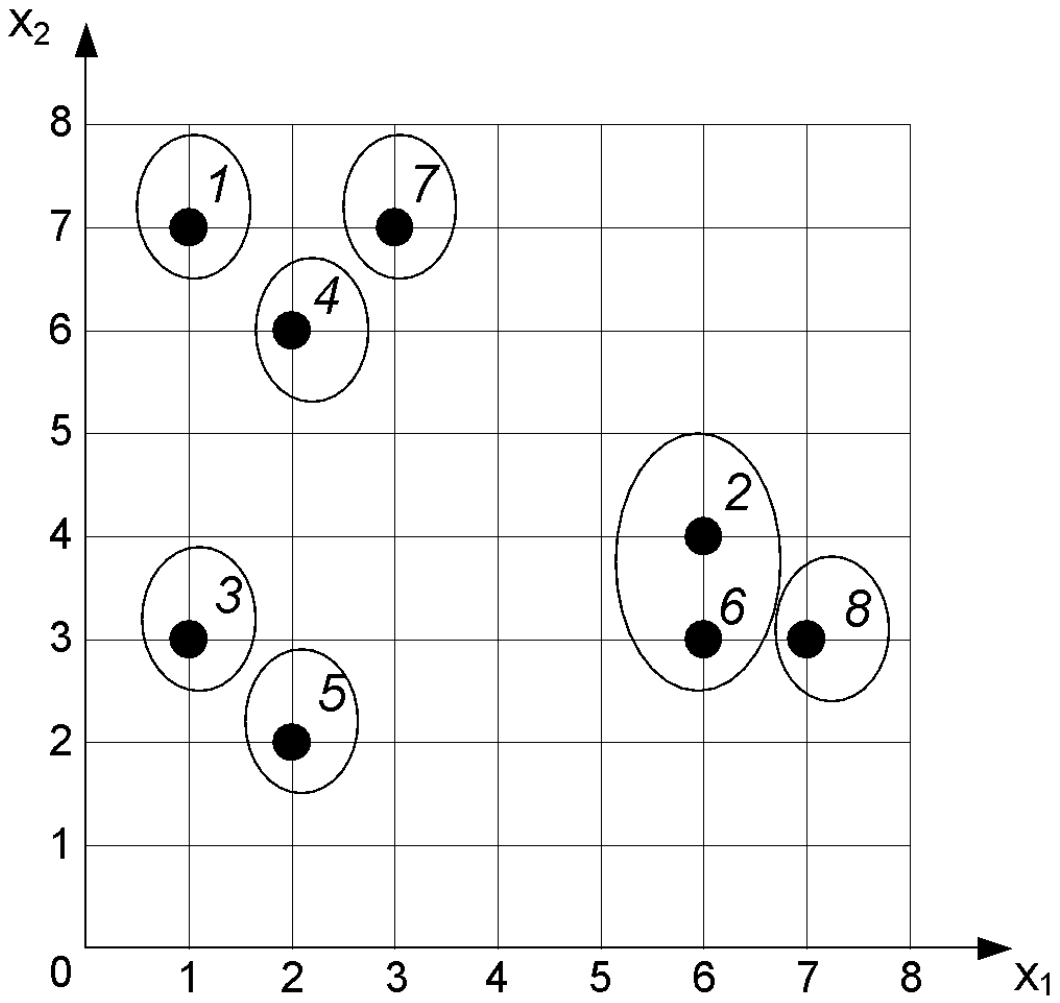
Агломеративный метод

- Выбираем и объединяем два наиболее близких кластера



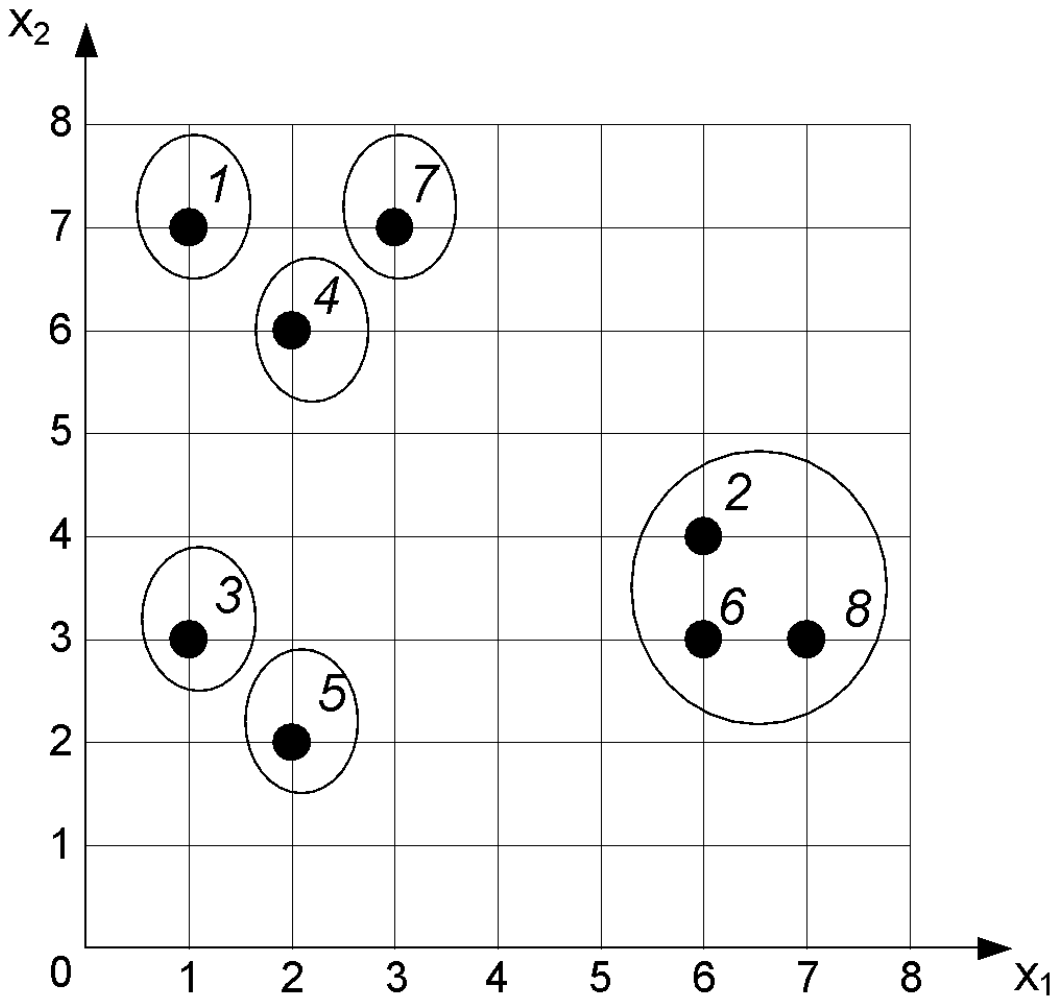
Агломеративный метод

- Выбираем и объединяем два наиболее близких кластера



Агломеративный метод

- Выбираем и объединяем два наиболее близких кластера



Дивизимный метод

- На первом шаге все объекты помещаются в один кластер C_1
- Выбирается объект, у которого среднее значение расстояния до других объектов в этом кластере наибольшее:

$$\bar{d}(S_p) = \frac{1}{N_C} \cdot \sum_{i=1}^{Nc} d(S_p, S_i)$$

Дивизимный метод

- Выбранный объект удаляется из кластера C_1 и формирует первый элемент второго кластера C_2
- На каждом последующем шаге объект в кластере C_1, для которого разность между средним расстоянием до объектов, находящихся в C_2 и средним расстоянием до объектов, остающихся в C_1, наибольшая, переносится в C_2

Дивизимный метод

- Переносы элементов из C_1 в C_2 продолжаются до тех пор, пока соответствующие разности средних не станут отрицательными,
то есть пока существуют элементы, расположенные к элементам кластера C_2 ближе, чем к элементам кластера C_1

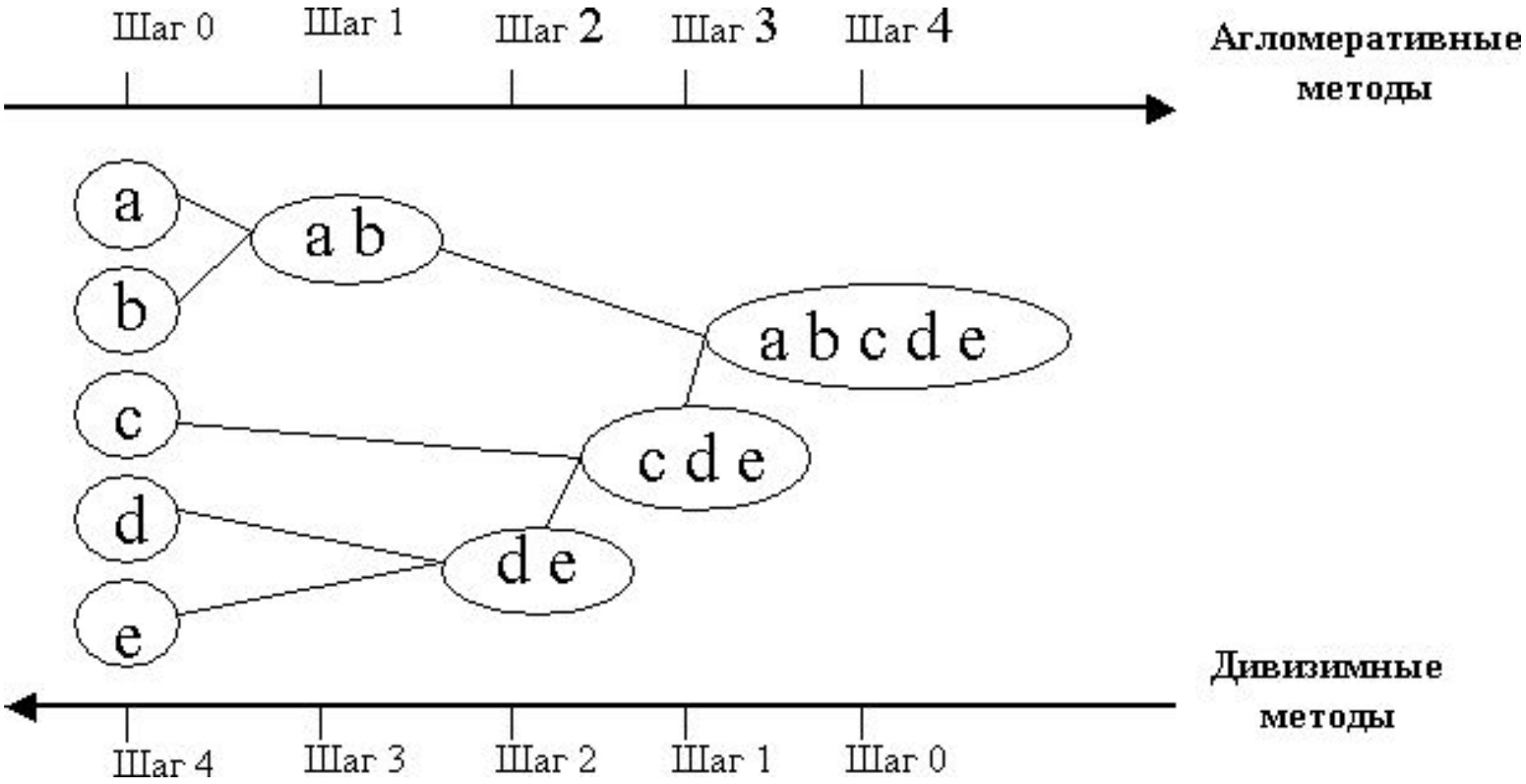
Дивизимный метод

- В результате один кластер делится на два дочерних, один из которых расщепляется на следующем уровне иерархии
- Каждый последующий уровень применяет процедуру разделения к одному из кластеров, полученных на предыдущем уровне

Дивизимный метод

- Кластер для расщепления выбирается, например, по наибольшему диаметру
- *Диаметр кластера* – расстояние между двумя наиболее дальними объектами кластера
- Рекурсивное разделение кластеров продолжается, пока хотя бы один кластер содержит более одного объекта

Иерархические методы



Иерархические методы

Проблема определения оптимального числа кластеров:

- иногда можно априорно определить число кластеров
- однако в большинстве случаев число кластеров определяется в процессе кластеризации

Иерархические методы

- В иерархических методах существует способ, позволяющий определить оптимальное число кластеров
- Процессу группировки объектов в иерархическом кластерном анализе соответствует постепенное возрастание коэффициента, называемого *критерием E*
- Критерий *E* на каждом шаге определяется как расстояние между ближайшими кластерами

Иерархические методы

- Скачкообразное увеличение критерия *E* говорит о переходе от сильно связанного к слабо связанному состоянию объектов
- Таким образом, нужно останавливать процесс разбиения, когда значение критерия E резко изменится