Сегментиране

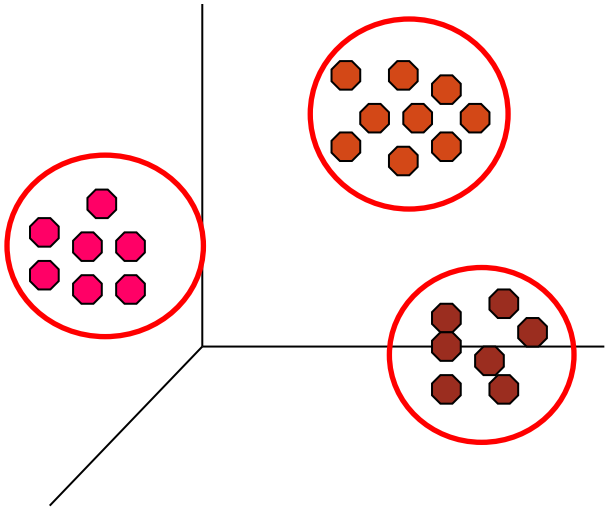
Клъстеризация

Клъстеризация

- Процес на групиране на данните в групи, клъстери така, че
  - обектите в една група да имат високо подобие помежду си, и
  - големи разлики с обектите от друга група
- Приликите и разликите се оценяват по стойностите на атрибутите, описващи обектите

Клъстери

- Групи по подобие
- Групите не са известни предварително
- Критерии
  - минимално разстояние между екземплярите в група
  - максимално разстояние между групите



Клъстерен анализ

- Cluster analysis (*clustering*, *data segmentation*, …)
  - Намиране на подобия между данните, по отношение на открити в тях характеристики и разделяне според подобията
  - подобията се измерват чрез функции за разстояние, напр. $d(i, j)$
  - мерките са различни според типа на данните, които се сегментират
- Цел – намаляване на обема на данните за изследване
- Unsupervised learning – описанията на групите не са известни предварително - *learning by observations* , както са при класификацията - *learning by examples* (supervised)

Приложения

- От най-ранно детство хората се научават да различават котка от куче, животно от растение
- Форми на употреба
  - Като самостоятелно средство за изследване
  - Като предварителна стъпка за прилагане на други методи и алгоритми за изследване
- Според целите
  - за сегментиране на данните
  - за откриване на частните случаи, отдалечени от останалите данни
- В различни сфери на бизнеса и научните изследвания, като
  - биологията – таксономии
  - управление на информация - класифициране на документи
  - в медицината - групиране на гени и протеини с подобно поведение
  - в маркетинга - групиране на клиенти и стоки
  - в икономиката, науките за земята и др.

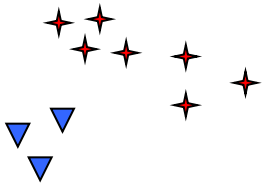
Примери

Групи, класове на подобие

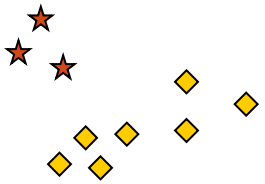
Колко клъстера са показани?

6

2



4



Разновидности на клъстерите

- Non-exclusive
  - един обект може да принадлежи на повече от един клъстер
  - гранични обекти
- Fuzzy
  - един обект принадлежи на всеки клъстер със степен на принадлежност между 0 и 1
  - Сумата на теглата е 1
- Partial
  - в някои случаи се клъстеризира само част от данните
- Heterogeneous
  - клъстери с различни размери, форми и плътности

Методи за клъстеризация - свойства

- Scalability – възможност за успешно приложение както върху малки множества данни, така и върху много големи
- възможност да се прилагат върху различни типове данни – бинарни, номинални, ординални и др.
- различни форми на намерените клъстери, не само сферични, разположени около определен център
- наличие на необходимост от входни данни, напр. предварително известен брой на клъстерите или опраничителни условия
- чувствителност към шум в данните
- чуствителност към реда на подаване на данните
- количество на атрибутите, с които оперират
  - човек може да наблюдава до 3 атрибута едновременно
- възможност за интерпретиране и прилагане на резултатите

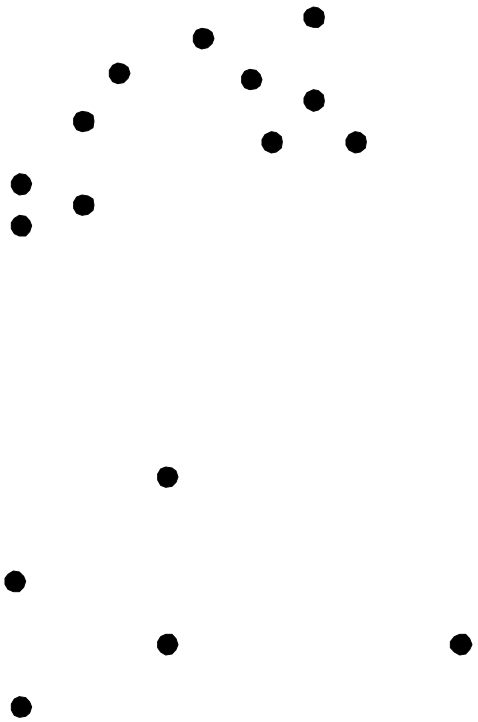
Методи за клъстеризация

- Типове методи
  - разделяне в групи
  - построяване на йерархии
- Критерии за избор
  - Локални или глобални условия
    - Алгоритмите за разделяне обикновено имат глобални цели
    - Йерархичните алгоритми за клъстеризация обикновено имат локални цели

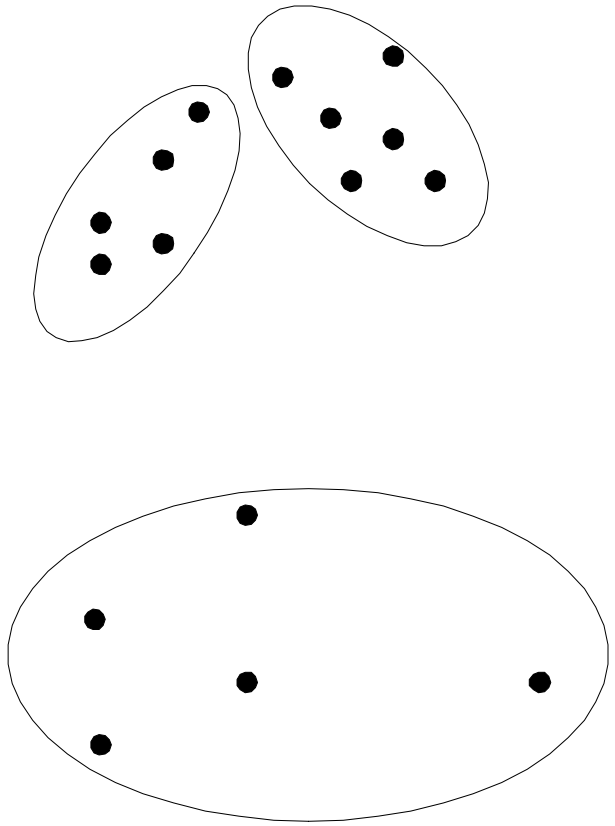
Типове методи (1)

- Partitional Clustering - Разделяне в групи
  - Разделяне на n обекти в k не-препокриващи се подмножества
  - Две условия
    - Всеки клъстер съдържа поне един обект
    - Всеки обект принадлежи на точно един клъстер
  - подход
    - прави се начално разделяне на групи
    - прилагат се итеративни реорганизиращи техники
    - до постигане на удовлетворителни резултати съобразно избрани критерии

Разделяне в групи



Обекти

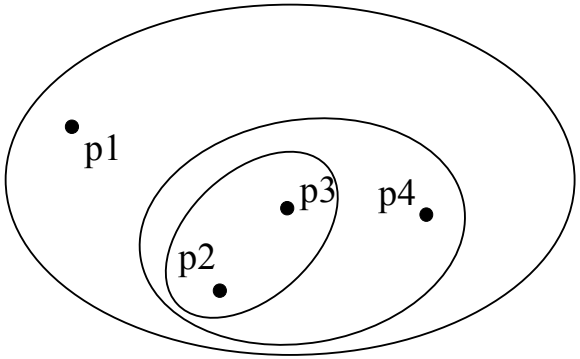


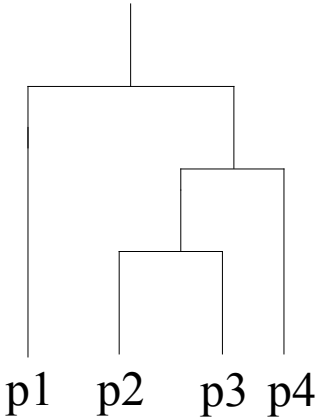
Групи

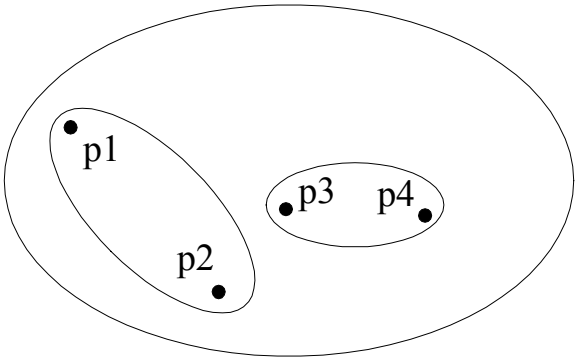
Типове методи (2)

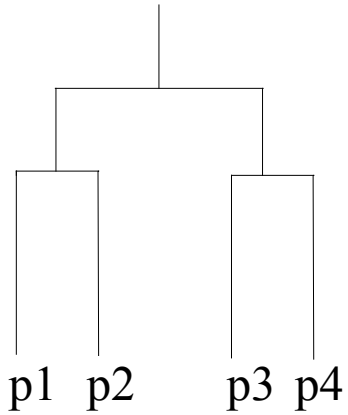
- Hierarchical Clustering - Изграждане на йерархии
  - Множество от вложени клъстери, организирани в йерархично дърво
  - йерархична декомпозиция на множеството от данни
    - agglomerative approach – от долу нагоре:
      - в началото всеки обект е група
      - групите се обединяват итеративно
    - divisible approach – от горе надолу
      - в началото всички обекти са в една група
      - итеративно разделяне на по-малки групи
    - критерии за край
      - удовлетворяване на условие
      - край на данните

Изграждане на йерархии





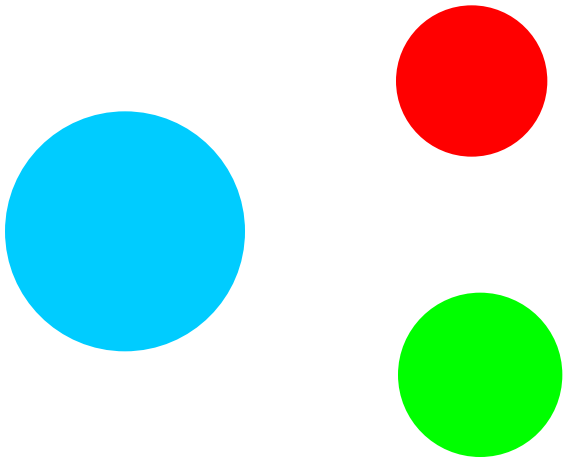




Дендограми

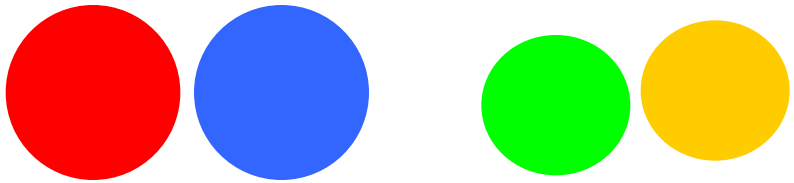
Правила за клъстеризация

- Добре различими клъстери
  - всеки обект е по-близо до всеки друг обект в клъстера, отколкото до обект извън него



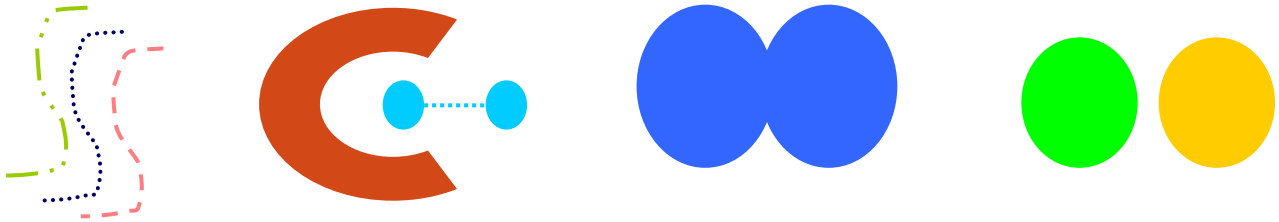
Правила за клъстеризация

- Централно-базирани тенденции
  - всеки обект е по-близо до центъра на клъстера, отколкото до центъра на друг клъстер
  - centroid – средният обект в клъстера
  - medoid – най-представителният обект



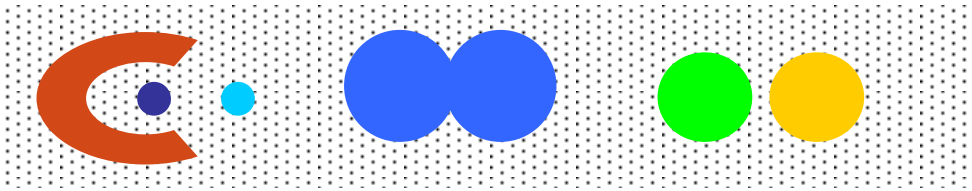
Правила за клъстеризация

- Contiguity-Based - Транзитивни
  - всеки обект е по-близо до един или повече други обекти в клъстера, отколкото до обект извън него



Правила за клъстеризация

- Density-based - Основани на плътност – брой обекти
  - цели се в определен радиус от всяка точка от клъстера да има зададен минимален брой съседи
  - подход - увеличаване на един клъстер, докато условието е все още изпълнено
  - резултат - плътни области от обекти, разделени от други плътни области чрез области с ниска плътност
  - За нерегулярни клъстери, шум и крайни случаи



Правила за клъстеризация

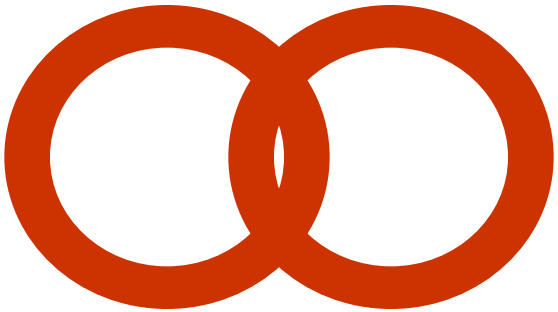
- Grid-based
  - пространството на данните се разделя на краен брой клетки на мрежа
  - операциите по разделяне се извършват в мерно пространство
  - по-бързо разделяна

Правила за клъстеризация

- Model-based
  - предварително се създават хопотези - модели на клъстерите
  - търси се най-голяма близостс модела
- Дефинирани чрез функция
  - обекти, които минимизират или максимизират определена функция за клъстеризиране

Правила за клъстеризация

- Концептуални
  - обекти с общи свойства



Проблемни области

- High-Dimensional Data
- Многомерни данни
  - голям брой атрибути
- Подходи
  - разделяне на пространството на атрибутите на под-множества
  - търсене на повтарящи се шаблони в пространството на атрибутите

Проблемни области

- Constrain-Based Clustering
- При наложени специфични ограничения
  - от потребителя
  - от приложението
- Предварително се задават свойства на бъдещите резултати
  - напр. определени двойки данни да попаднат в една група или не
- Процесът се следи итеративно

Алгоритми за клъстеризация

Избор на алгоритъм

- Според
  - типа на данните
  - целта на изследването
- Правило
  - да се пробват повече от един

Нотация

- D - множество от n обекти за клъстеризиране
- Един обект е представен чрез d променливи (атрибути или димeнсии)
- Броят на клъстерите е k

Partitioning

- за извадка от данни ***D*** от ***n*** обекти, които се разделят в ***k*** клъстера, така, че сумата от квадратите на разстоянията между обектите да е минимална

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

където c_i - centroid или medoid на клъстера C_i

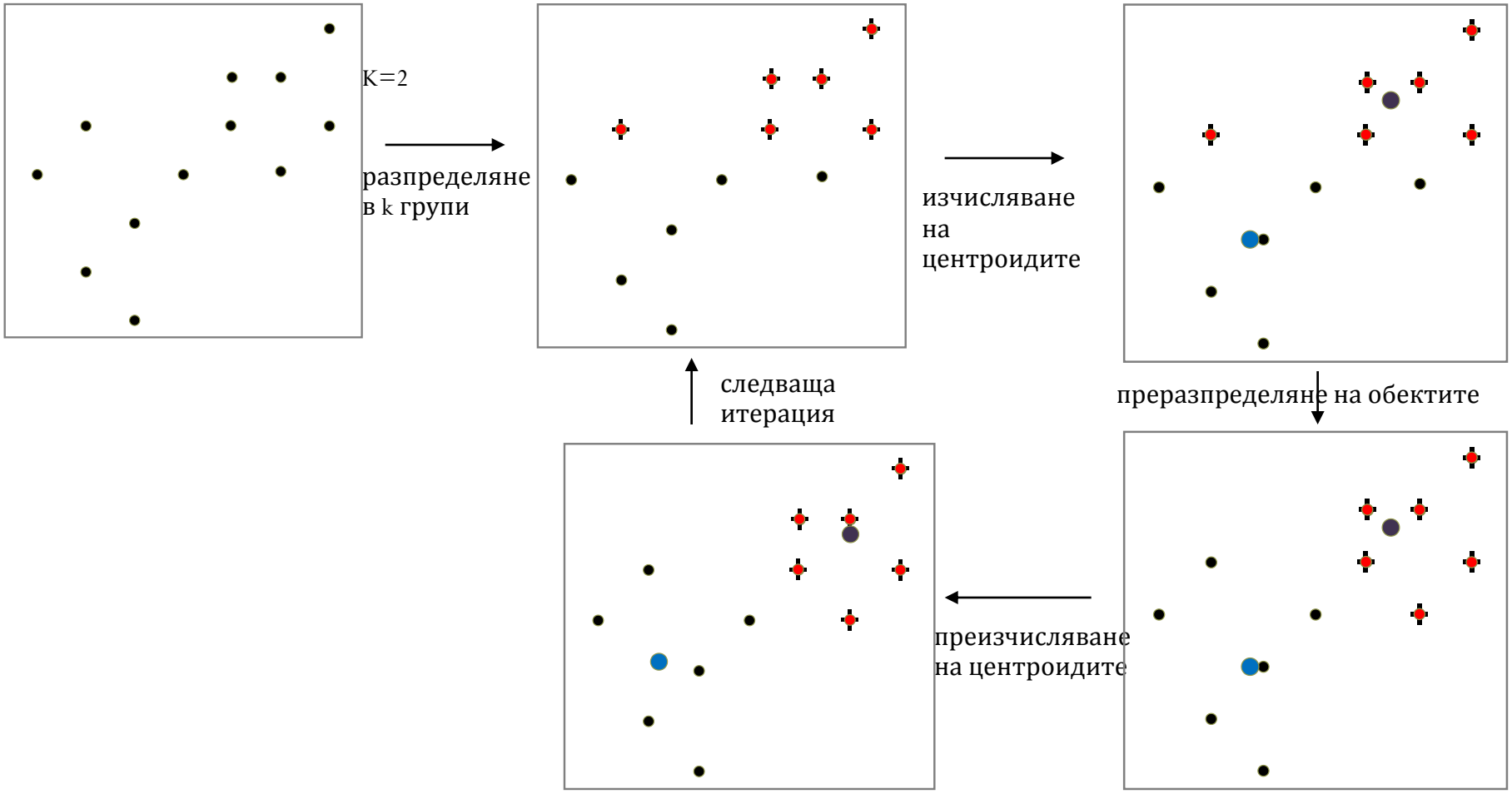
- *k-means* всеки клъстер се представя чрез центъра си
  - (MacQueen'67, Lloyd'57/'82)
- *k-medoids* PAM (Partition around medoids) Всеки клъстер се представя чрез един от обектите, разположен близо да центъра
  - (Kaufman & Rousseeuw'87)

Алгоритъм K-Means

Цел: определяне на клъстери, при които се минимизира функцията за средно-квадратична грешка

1. Задават се входен параметър *к* и *n* обекти, (k<=n)
2. Избират се *к* произволни точки за центроиди - центрове на привличане на група обекти
3. Обектите се разделят в *к* групи около центроидите според близостта си до тях
4. Изчисляват се нови центроиди за всеки клъстер, като средна стойност на данните, рапределени в клъстера при предишната итерация
5. Обектите се преразпределят, като всеки се причислява към групата, до чийто центроид сега е най-близо
6. Повтарят се стъпките, до достигане на устойчиво състояние

Пример


K=2
разпределяне в k групи
изчисляване на центроидите
преразпределяне на обектите
преизчисляване на центроидите
следваща итерация

Алгоритъм K-Modes

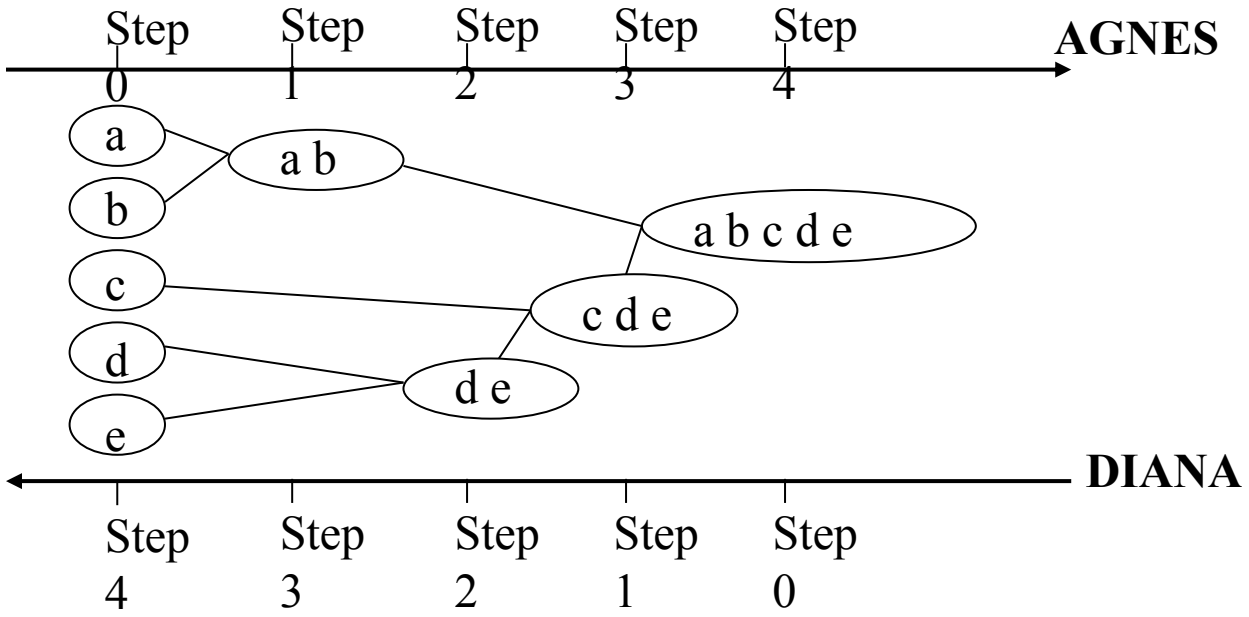
- Цел: определяне на клъстери, при които не може да се изчисли център, напр. за категорийни данни
  - вместо средна стойност, за клъстерите се изчислява мода
- Двата алгоритъма могат да се използват съвместно
- Предимства
  - ефективни
  - мащабируеми
- Недостатъци
  - не може да се приложи за клъстери с различна форма или разлика в размера

Алгоритъм K-Medoid

- Цел: определяне на клъстери, при големи разлики в стайностите на данните
- Вместо минимизиране на квадратичната грешка на отсточнието до центъра се минимизира сумата на абсолютните грешки на отстояние на всяка точка до центъра

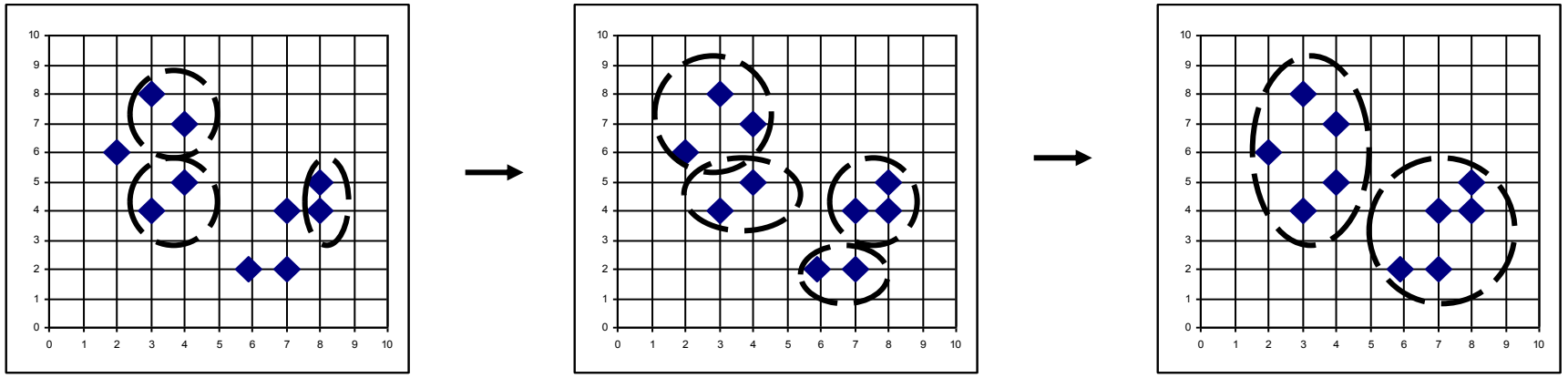
Йерархична клъстеризация

- Използва се матрица на дистанциите
  - Не е известен броя на клъстерите
  - Необходимо е условие за край
- Два алгоритъма
  - AGNES - agglomerative
  - DIANA - divisive



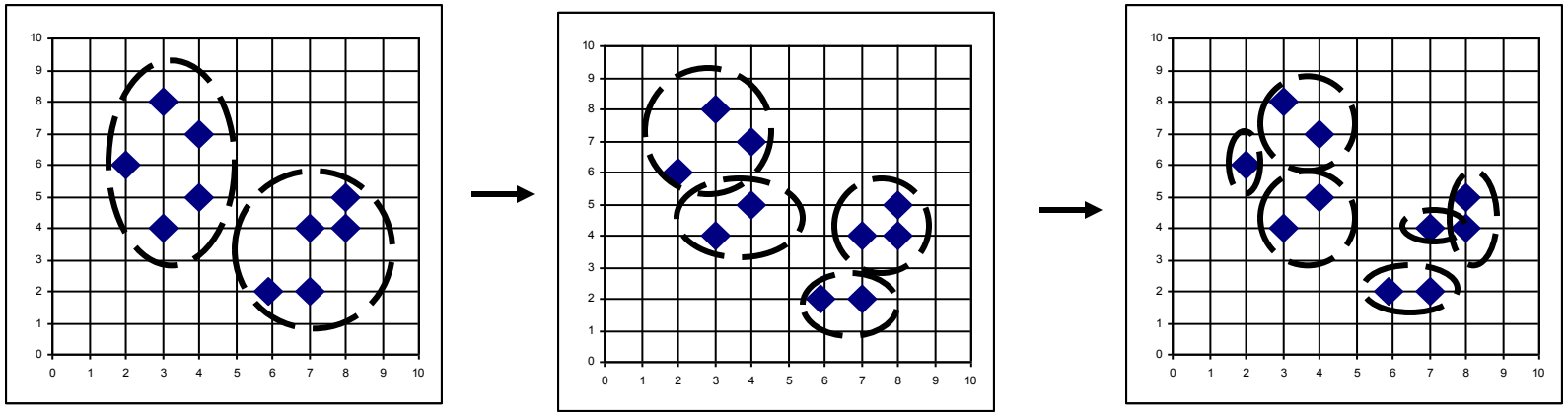
AGNES (Agglomerative Nesting)

- Kaufmann and Rousseeuw (1990)
- Използва се метод **single-link** и матрица на различията
- Смесват се обекти, които имат най-малки различия
- Продължава се в ненамаляващ ред
- Накрая всички обекти може да са в един клас



DIANA (Divisive Analysis)

- Kaufmann and Rousseeuw (1990)
- Ред, обратен на AGNES
- Накрая всеки обект може да е сам в клас



Мерки на клъстер

- Centroid: среда на клъстера

$$C_m = \frac{\sum_{i=1}^{N} (t_{ip})}{N}$$

- Radius: квадратен корен от средното разстояние от всяка точка до средата

$$R_m = \sqrt{\frac{\sum_{i=1}^{N} (t_{ip} - c_m)^2}{N}}$$

- Diameter: квадратен корен от средното разстояние между всяка двойка точки в клъстера

$$D_m = \sqrt{\frac{\sum_{i=1}^{N} \sum_{i=1}^{N} (t_{ip} - t_{iq})^2}{N(N-1)}}$$