Data Mining
Laboratorium 4: Grupowanie

dr inż. Robert Bembenik, dr inż. Grzegorz Protaziuk
Politechnika Warszawska
Instytut Informatyki

Pojęcia podstawowe

- Obiekt – opisany przez zbiór atrybutów.
  - atrybuty o wartościach nominalnych
  - atrybuty o wartościach liczbowych
- Baza danych (BD) – zbiór obiektów.

Grupowanie

Celem grupowania jest podział zbioru obiektów na klasy (grupy) podobnych obiektów (mających podobne wartości atrybutów).

W zależności od metody grupowania liczba grup jest albo nie jest określona jako parametr wejściowy metody.

Cechą dobrego grupowania jest wysokie podobieństwo obiektów w ramach tej samej grupy, natomiast niskie podobieństwo obiektów z różnych grup.

Podobieństwo często jest określane jako pewna miara odległości między dwoma obiektami.

Algorytmy partycjonujące: podstawowe koncepcje

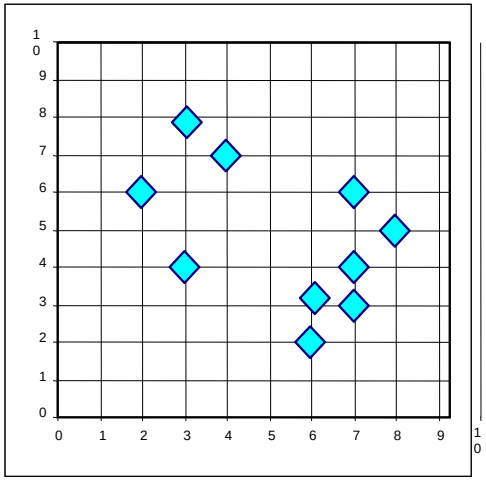
- Metoda partycjonowania: Stwórz podział bazy danych *D* złożonej z *n* obiektów na *k* grup, min. sumę odległości w kwadracie między punktem *p* a środkiem c_i grupy C_i

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

- Mając dane *k*, znajdź taki podział na *k* grup, który optymalizuje wybrane kryterium podziału
  - metody heurystyczne: algorytmy *k-średnich* i *k-środków*
    - *k-średnich* (MacQueen'67): każda grupa jest reprezentowana przez środek grupy
    - *k-środków* lub PAM (Partition around medoids) (Kaufman & Rousseeuw'87): każda grupa jest reprezentowana przez jeden z obiektów w grupie

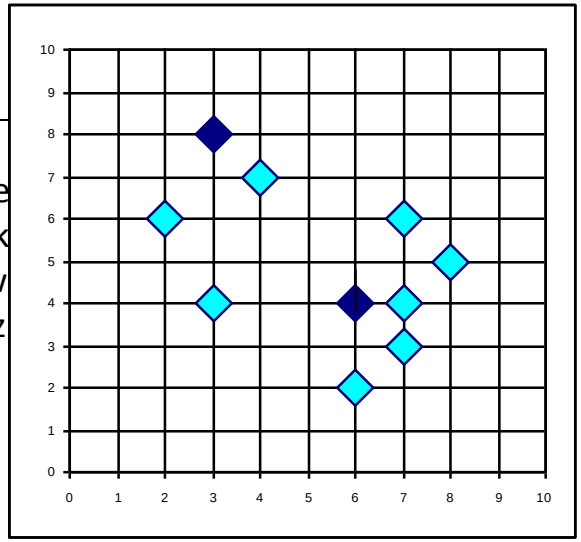
POLITECHNIKA WARSZAWSKA

Typowy algorytm *k-środków*



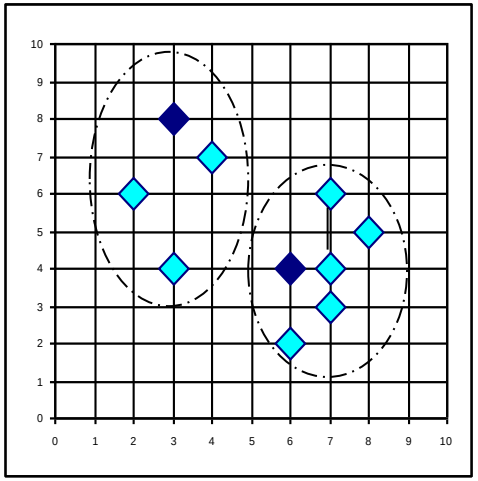
K=2

Dowolnie wybierz k obiektów jako pocz środki

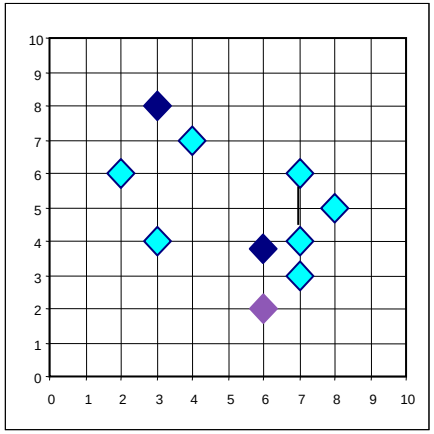


Przypisz każdy pozostał y obiekt do najbliższ ego środka

Całkowity koszt = 20

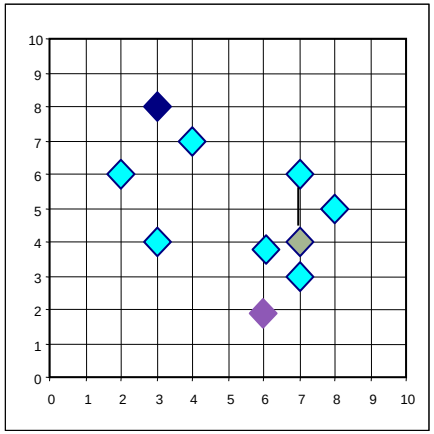


Losowo wybierz obiekt nie będ. środkiem, O_{ramdom}



Oblicz całkowity koszt zamiany

Całkowity koszt = 26



Zamiana O i O_{ramdom}

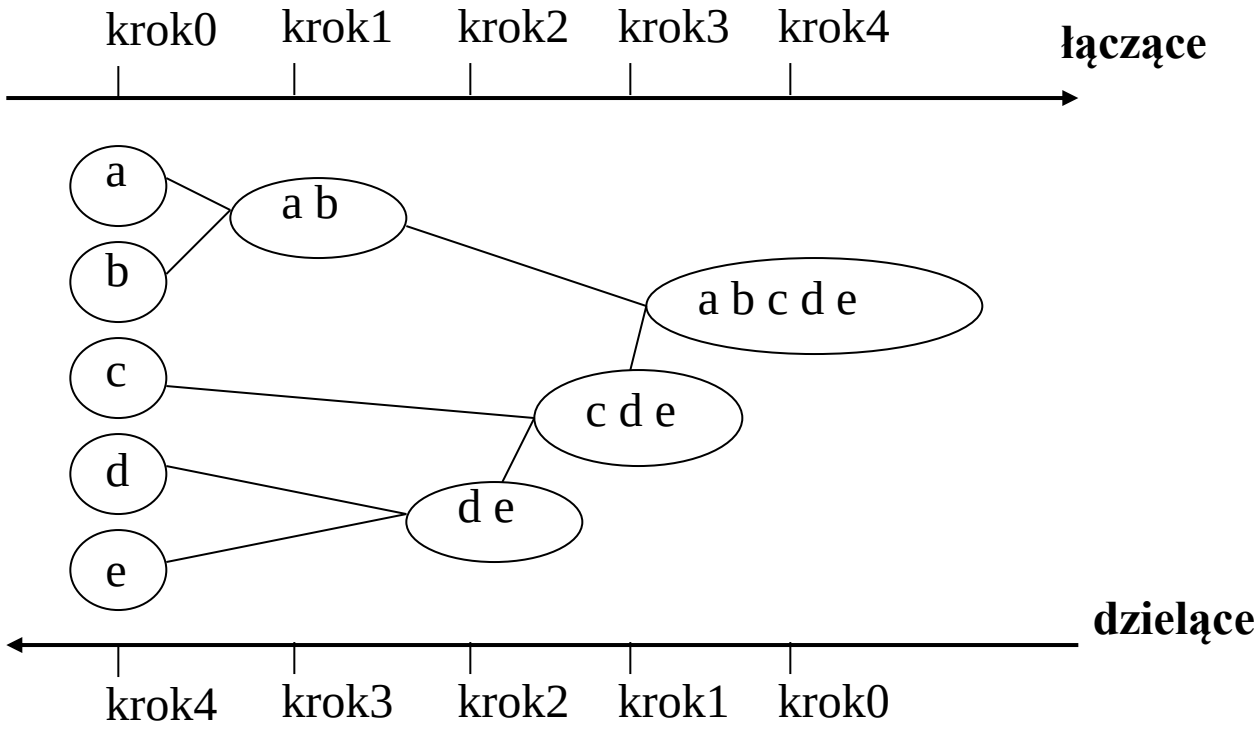
Jeżeli poprawa jakości

Do loop

Until brak zmian

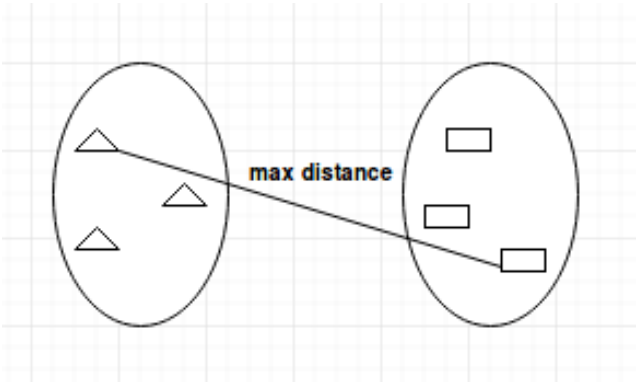
Grupowanie hierarchiczne

- Łączące (większość metod należy do tej kategorii)
- Dzielące (kończą działanie po napotkaniu kryt. stopu, np. ustalona liczba grup, osiągnięto ustaloną średnicę grupy)



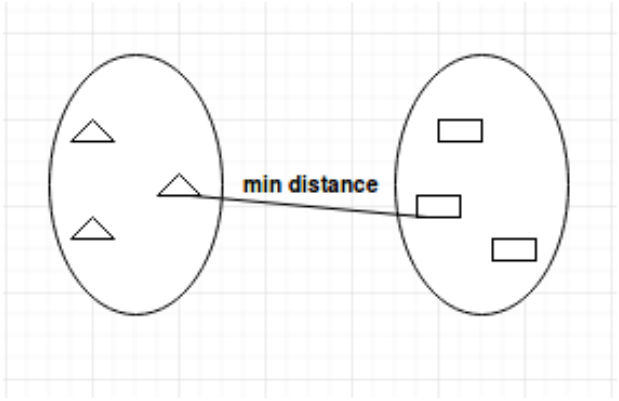
Grupowanie hierarchiczne: metody łączenia grup

- Maksymalna (kompletna): odległość między dwoma grupami to maksymalna wartość wszystkich par odległości między elementami w grupie 1 i w grupie 2. Tworzy grupy bardziej kompaktowe.



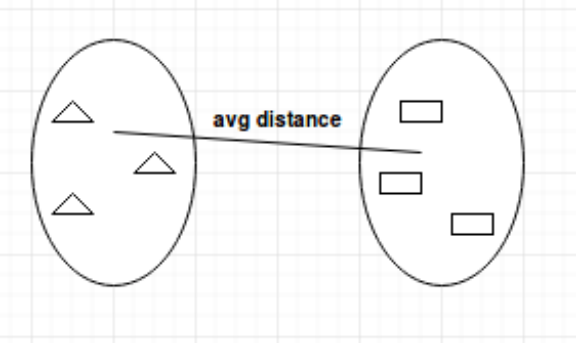
Grupowanie hierarchiczne: metody łączenia grup

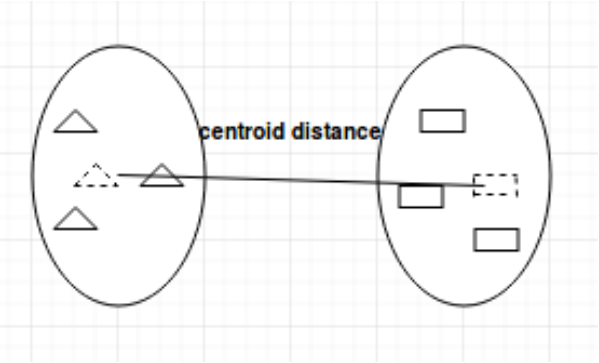
- Minimalna (pojedyncza): odległość między dwoma grupami to minimalna wartość wszystkich par odległości między elementami w grupie 1 i w grupie 2. Tworzy grupy bardziej „luźne”.



Grupowanie hierarchiczne: metody łączenia grup

- Średnia: odległość między dwoma grupami to średnia odległość między elementami w grupie 1 i w grupie 2.
- Łączenie środków ciężkości (centroidów) grup: odległość między dwoma grupami to odległość między centroidami dla grupy 1 i dla grupy 2.



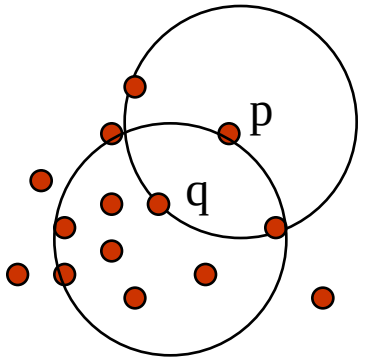


W każdym kroku grupowania łączone są dwie grupy mające najmniejszą odległość połączenia.

Grupowanie gęstościowe: DBSCAN podstawowe pojęcia

- Dwa parametry*:*
  - *Eps*: Maksymalny promień sąsiedztwa
  - *MinPts*: Minimalna liczba punktów w sąsiedztwie Eps tego punktu
- $N_{Eps}(p)$: *{q należy do D | dist(p,q) <= Eps}*
- Bespośrednio gęstościowo-osiągalny: Punkt *p* jest bezpośr. gęstościowo osiągalny z punktu *q* w odniesieniu do *Eps* i *MinPts* jeżeli
  - *p* należy do $N_{Eps}(q)$
  - warunek punktu głownego:

$$|N_{Eps}(q)| >= MinPts$$

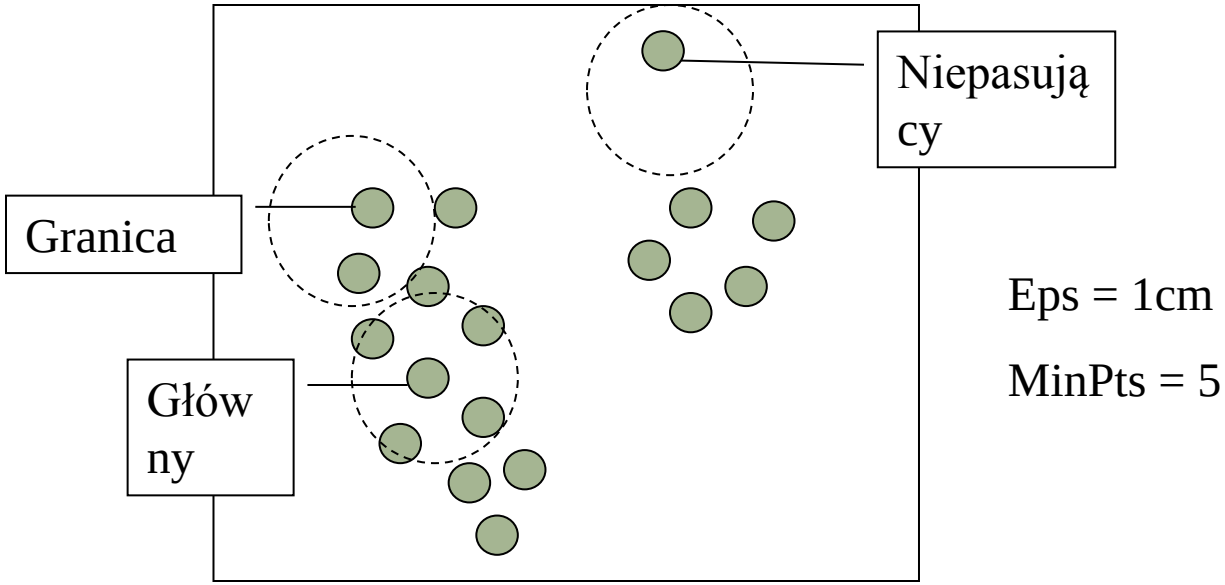


MinPts = 5

Eps = 1 cm

DBSCAN

Opiera się na pojęciu grupy opartym na gęstości: grupa jest zdefiniowana jako maksymalny zbiór gęstościowo połączonych punktów



Ocena jakości grupowania

Indeks Silhouette

$$Silhouette(x) = \frac{b(x) - a(x)}{\max(b(x), a(x))}$$

gdzie: a(x) – średnia odległość obiektu *x* do innych obiektów w grupie „x”

b(x) – minimalna odległość obiektu *x* od najbliższej grupy *G*, do której nie należy *x* (średnia odległość obiektu *x* od obiektów należących do *G*)

Indeks przyjmuje wartości <-1, 1>, gdzie 1 oznacza, że dany obiekt jest przydzielony do najlepszej z możliwych grup, 0 – obiekt znajduje się między dwoma grupami, -1 – zły przydział obiektu.

$$GSilhouette = \frac{1}{N}\sum_{i=1}^{N} Silhouette(x_i)$$

gdzie: N – liczba obiektów w zbiorze

Ocena jakości grupowania

Rand index

W – grupowanie wzorcowe, G - grupowanie oceniane

A – liczba par obiektów należących do tej samej grupy w grupowaniu W i G

B – liczba par obiektów należących do różnych grup w grupowaniu W i G

a – liczba par obiektów należących do tej samej grupy w grupowaniu W, ale należących do różnych grup w grupowaniu G

b – liczba par obiektów należących do różnych grup w grupowaniu W, ale do tych samych grup w grupowaniu G

n – liczba obiektów

$$R = \frac{A+B}{A+B+a+b} = \frac{A+B}{n(n-1)/2}$$

Grupowanie w R (1)

- Pakiety domyślne
  - scale() – centrowanie i/lub skalowanie danych numerycznych.
  - kmeans() - algorytm k-Means, funkcja zwraca obiekt kmeans z opisem utworzonych grup
  - hclust () - grupowanie hierarchiczne – buduje dendrogram
  - cutree() – utworzenie grup na podstawie dendrogramu
  - plot() – wizualizacja grupowania

Grupowanie w R (2)

- Pakiet fpc
  - dbscan () - algorytm DBScan
  - plotcluster() – wizualizacja grupowania.
- Pakiet cluster
  - zawiera implementacje kliku algorytmów grupowania