GRUPIRANJE PODATAKA

(engl. Cluster analysis,
engl. Taxonomy analysis – posebno u bio znanostima)

Cilj: Pridružiti objekte u grupe na temelju *sličnosti* objekata.

Sličnost je predefinirani kriterij koji se računa iz opažanja (mjerenja) svojstava na objektima.

Pitanja:

- Koju mjeru sličnosti ili različitosti (engl. similarity, dissimilarity) koristiti ?

- Koji algoritam grupiranja koristiti?

Za grupiranje objekata koristi se **metrika**, za grupiranje varijabli – **korelacijski koeficijenti**

Metrika

Bilo koji skup X se kaže da je metrički prostor ako za bilo koje dvije točke a, b $\in$ X postoji funkcija sa d: X x X -> R za koju vrijede sljedeća svojstva:

(1) d(a, b) > 0 ako a≠b, d(a, a)=0 ; (*pozitivna definitnost*)
(2) d(a, b) = d(b, a) ; (*simetričnost*)
(3) d(a, c) $\leq$ d(a, b) + d(b, c), (*pravilo trokuta*)

Svaka funkcijna s navedenim svojstvima zaziva se **funkcija udaljenosti ili metrika**.

Mjera udaljenosti (engl. *dissimilarity measure*) je mjera različitosti podataka.
Primjer mjere udaljenosti ili metrike jest tzv. **Minkovski metrika** (L_n) koja je definirana sa:

$$d(a,b) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^n} \qquad (1)$$

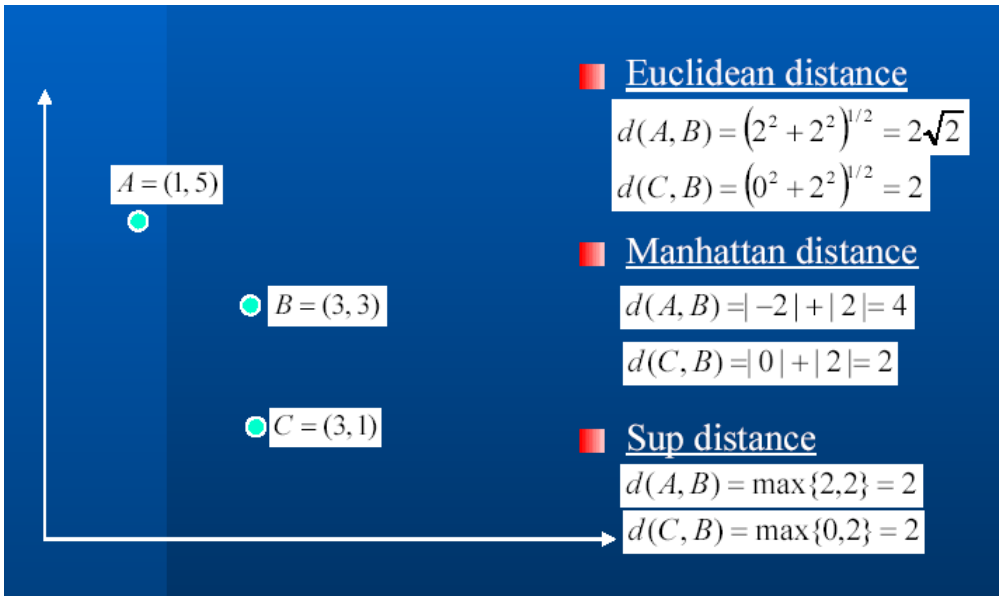
Specijalni slučaj Minkovski metrike za n=2 zove se **Euklidska metrika (L_2)** i najpoznatija je metrika.
Primjer: Skup točaka u 2-dim prostoru koji je od neke čvrste točke, središta, udaljen za odabranu konstantnu vrijednost r je kružnica u L_2 metrici.

Specijalni slučaj za Minkovski metrike za n=1 jest **Manhattan ili Cityblock metrika**. Za binarne vektore ta se metrika zove **Hammingova metrika**.
Primjer: Skup točaka u 2-dim prostoru koji je od neke čvrste točke, središta, udaljen za odabranu konstantnu vrijednost r je kružnica u L_1 metrici dan na slici.
n
Kada n$\rightarrow \infty$, formula (1) naziva se **Čebiševljeva** udaljenost (L_∞)

$$d(a, b) = \max_{1 \leq i \leq n} \{|a_i - b_i|\}$$

Primjer: skup točaka u 2-dim prostoru koji je od neke čvrste točke, središta, udaljen za odabranu konstantnu vrijednost *r* je kvadrat.

Primjer:



(Slika preuzeta od Jeong-Ho Chang, Seoul National University Seoul, Korea)

Statistička udaljenost:

Mahalonobisova udaljenost (1948.g.)

$$d(\mathbf{x},\mathbf{y}) = \text{sqrt}\ [\ (\mathbf{x}-\mathbf{y})'\Sigma^{-1}(\mathbf{x}-\mathbf{y})],$$

gdje je Σ^{-1} inverz matrice varijanci-kovarijanci.

Ta je **udaljenost pozitivno definitna kvadratna forma** oblika x'Ax, gdje je A = Σ^{-1} i poopćenje je euklidske udaljenosti ako varijable imaju različite standardne devijacije i korelirane su!
Na primjer ako se Mahalanobisova udaljenost koristi za računanje udaljenosti jedne multivarijatne opservacije od centra populacije:

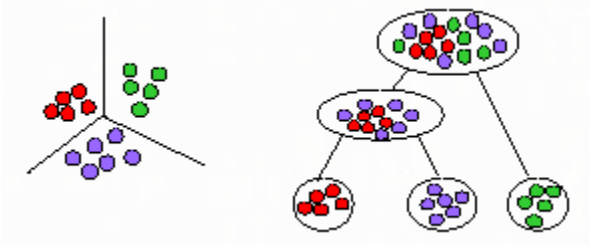
$$D^2 = \sum_{i=1}^{p}\sum_{j=1}^{p}(x_i - \overline{x_i})v_{ij}(x_j - \overline{x}_j)$$

gdje su (x_1, x_2, … , x_p) vrijednosti varijabli X_1, X_2, … , X_p, a v_{ij} je element u i-tom retku i j-tom stupcu inverzne matrice varijanci kovarijanci.

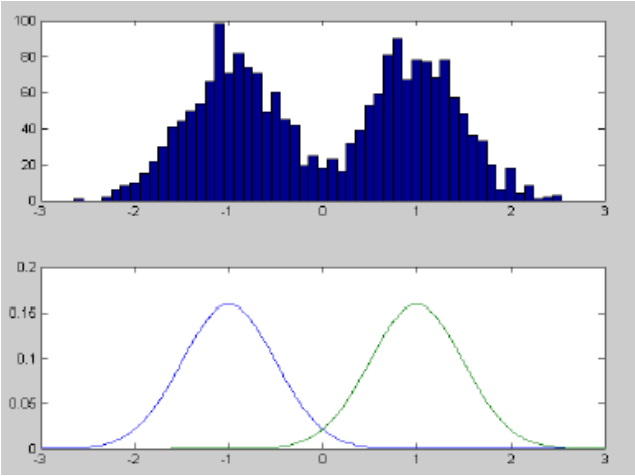
(Primjer: skup točaka u 2-dim prostoru koji je od neke čvrste točke, središta, udaljen za odabranu konstantnu vrijednost *r* je elipsa)

Vrste grupiranja:

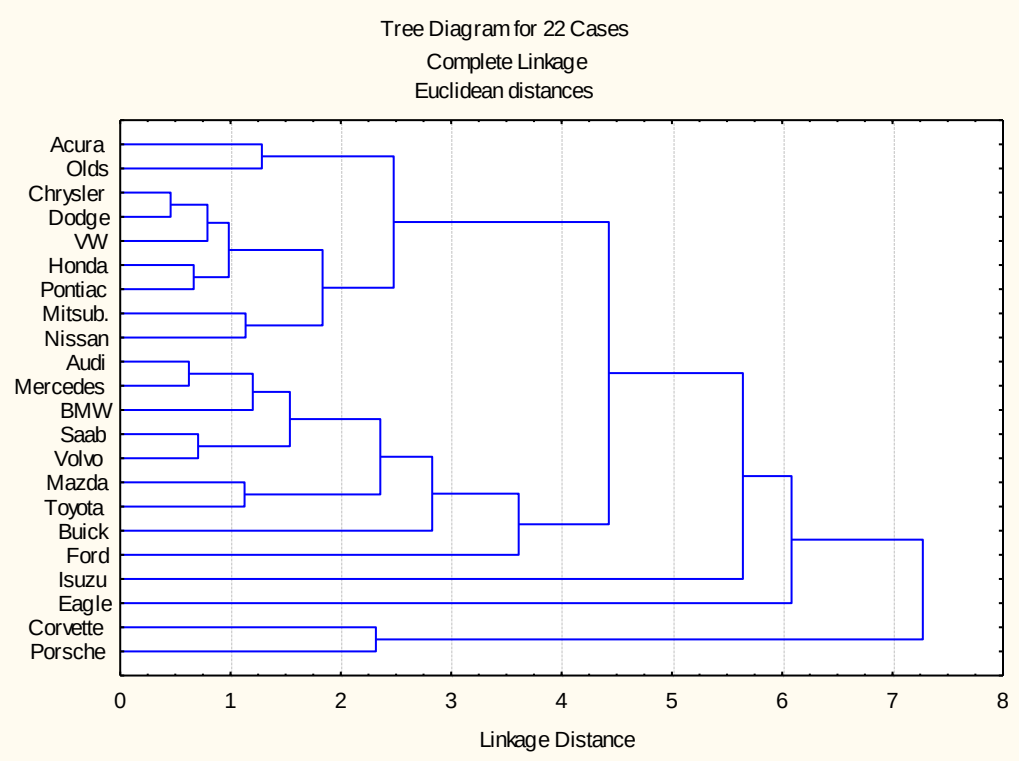
Particijska Hijerarhijska



Parametarska



Hijerarhijska grupiranja – rezultat grupiranja je grafički prikazan **dendogramom**



- Aglomerativna (*bottom-up*) (počinju individualnim objektom, inicijalno n objekata – n grupa, najsličniji objekti se grupiraju, grupe se stapaju u skladu s odabranim kriterijem)

- Divizivna (*top-down*) (rade suprotno, inicijalno svih n podataka je jedna grupa, koja se dijeli na podgrupe, podgrupe se dijele dalje u skladu s odabranim kriterijem)

Particijska grupiranja – nisu hijerarhijske (*engl. flat*)

- **K srednjih vrijednosti , (k –means)**
- SOM

Parametarski model

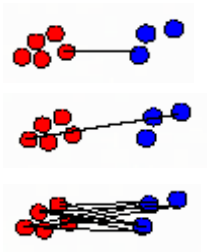
- EM algoritam

Aglomerativna hijerarhijska grupiranja

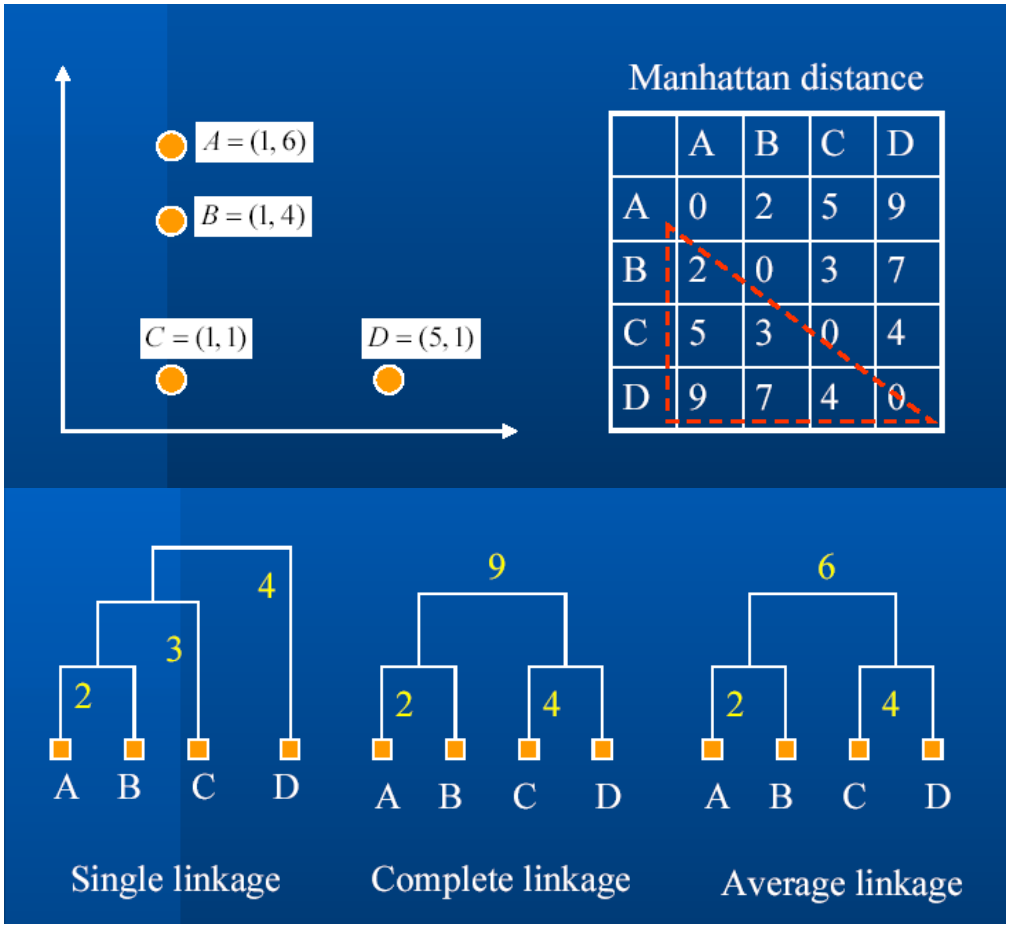
Metode povezivanja (linkage methods)

- pogodne za varijable i objekte

1. single linkage (najbliži)
2. complete linkage (najudaljeniji)
3. average linkage (srednje vrijednosti)

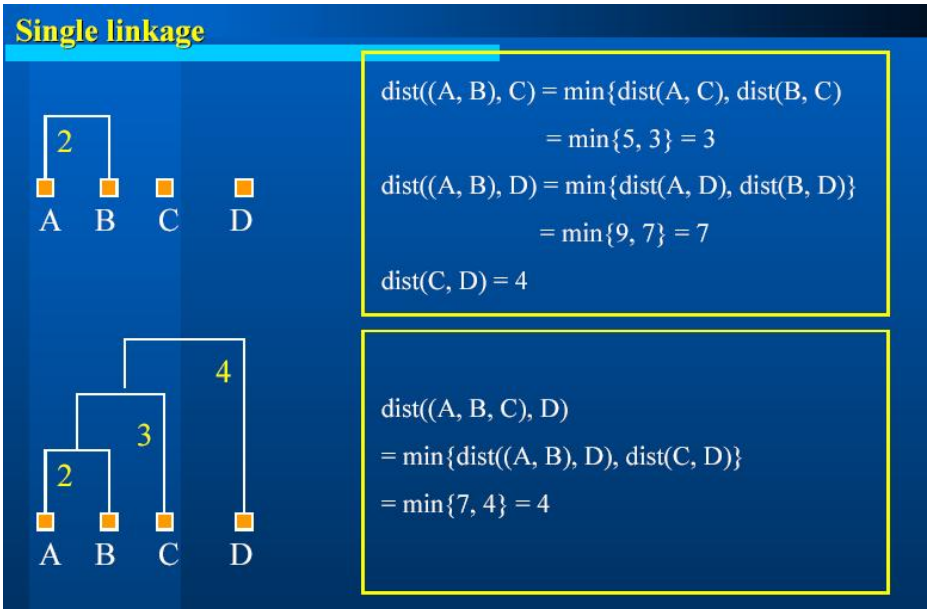


Primjer: Grupiranje 4 podataka u 2-dim prostoru



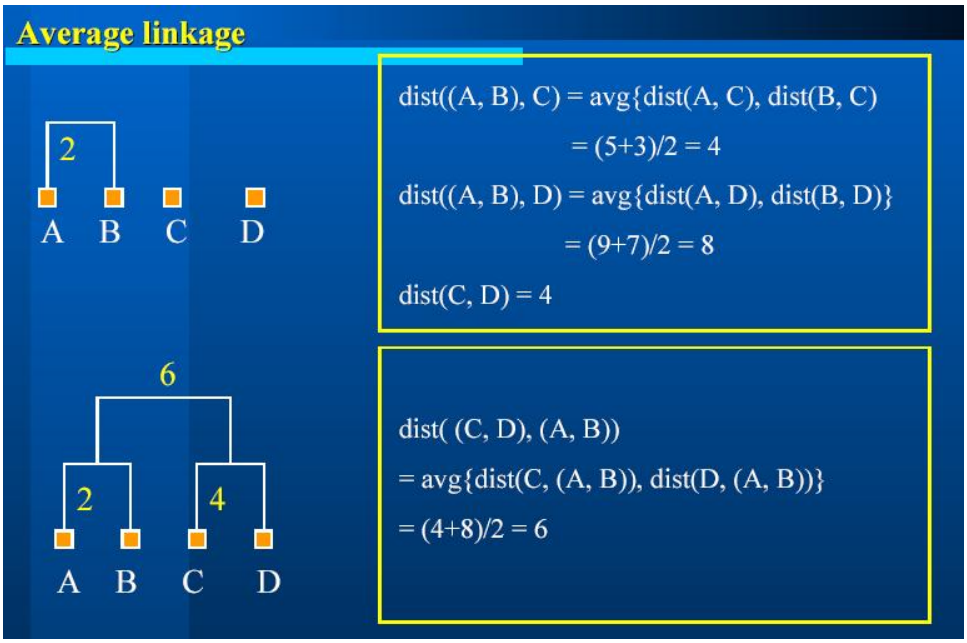
(Slika preuzeta od Jeong-Ho Chang, Seoul National University Seoul, Korea)

Single linkage – **Povezivanje na temelju minimalne udaljenosti**
ili povezivanje najbližeg susjeda
Podaci u proceduri mogu biti udaljenosti ili sličnosti između objekata. Najbliži susjed određuje najmanju udaljenost ili najveću sličnost između podataka.
Zbog načina spajanja ne može razlikovati slabo odjeljive grupe, ali može odijeliti ne-elipsoidalne grupe.
Ima tendenciju stvaranja duljih lanaca na čijim se krajevima jedinke mogu bitno razlikovati.



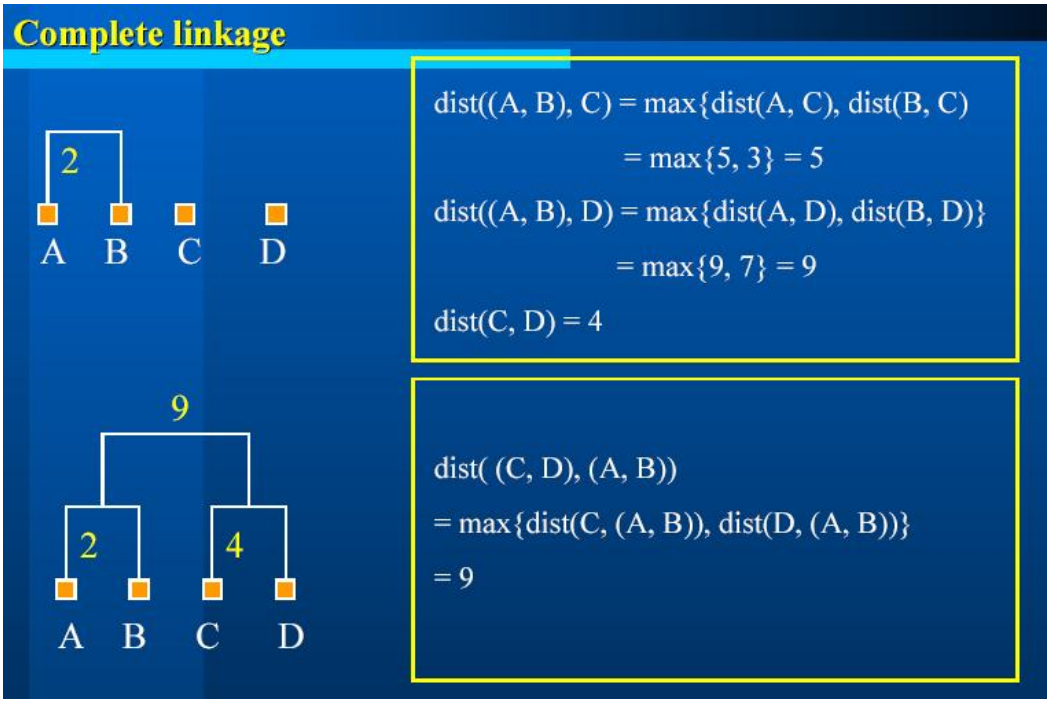
(Slika preuzeta od Jeong-Ho Chang, Seoul National University Seoul, Korea)

Average Linkage – Povezivanje na temelju srednje udaljenosti između grupa. Udaljenost je srednja vrijednost udaljenosti svih parova u grupama.
(Slika preuzeta od Jeong-Ho Chang, Seoul National University Seoul, Korea)
(Slika preuzeta od Jeong-Ho Chang, Seoul National University Seoul, Korea)
(Slika preuzeta od Jeong-Ho Chang, Seoul National University Seoul, Korea)



(Slika preuzeta od Jeong-Ho Chang, Seoul National University Seoul, Korea)

Povezivanje na temelju maksimalne udaljenosti – udaljenost između dvije grupe (elementa) je određena najvećom udaljenošću.
Osigurava da su svi objekti u grupi unutar neke maksimalne udaljenosti.



(Slika preuzeta od Jeong-Ho Chang, Seoul National University Seoul, Korea)

Uočava se sličnost dendrograma complete linkage i average linkage, ali se povezivanje dešava na različitim razinama udaljenosti.

Ulaz u postupak povezivanja može biti i korelacijska matrica. Sličnost između dviju varijabli mjeri se produkt-moment korelacijskim koeficijentom. Varijable s velikim negativnim korel. koef. smatraju se jako udaljenima, a one s većim pozitivnim smatraju se bliskima.

Zaključci:

- hijerarhijske aglomerativne metode su osjetljive na tzv. *outliere* - vrijednosti koje odudaraju
- nema mogućnosti preispitivanja već pridjeljenih (krivo) objekata grupama
- dobro je probati više metoda i više mjera udaljenosti te provjeriti konzistentnost rješenja
- stabilnost grupiranja može se provjeriti dodavanjem perturbacija. Ako su grupe jasno odjeljive grupiranje prije i poslije perturbacija se trebaju slagati

ALGORITAM k – SREDNJIH VRIJEDNOSTI

(k-means algorithm)

Odnosi se na particiju objekata, a ne varijabli.
Ne koristi matricu sličnosti pa je zahvalnija metoda za veći skup podataka.

Ukratko:

1. odabere se k početnih centara grupa (centroida)
2. sve se objekti rasporede u *k* grupa po pravilu minimalne udaljenosti
3. računa se novih k centroida
4. ponavljaj korake 2 i 3 dok više nema promjena

Algoritam k - srednjih vrijednosti je postupak grupiranja na temelju **minimizacije kriterijske funkcije**:

$$J = \Sigma_{j=1}^{N_c} J_j, \quad \text{pri čemu je} \quad J_j = \Sigma_{x \in S_j} || x - Z_j ||^2.$$

N_c predstavlja broj elemenata od k grupa, dok S_j predstavlja skup uzoraka čiji je centar Z_j.

Cilj algoritma je naći *k* središta grupa Z_1, Z_2, ..., Z_k za *N* početnih neraspodijeljenih uzoraka. Broj *k* se zadaje na početku, zajedno sa uzorcima, i za njega vrijedi:

$$0 < k < N.$$

Specifičnost algoritma je ta da ovisi o redoslijedu uzimanja uzoraka.

Algoritam k-srednjih vrijednosti:

1. izabiremo k središta grupa Z_1 (1), Z_2 (1), ..., Z_k (1). Metoda izbora početnih središta grupa je proizvoljna. Postoji nekoliko tipova uobičajenih izbora pa prema tome i nekoliko tipova algoritma k – srednjih vrijednosti.

2. u m – tom koraku (iteraciji) razdjeljujemo uzorke $x_1, x_2, ..., x_N$ u k grupa pomoću relacije:

   $x \in S_j$ (m) ako je $|| x - Z_j (m) || < || x - Z_i (m) ||$, i = 1, 2, .., N; $i \neq j$.

   S_j (m) predstavlja skup uzoraka u m – tom koraku čiji je centar Z_j.

3. izračunavamo nova središta grupa Z_j (m+1), j = 1, 2, ..., k tako da je kriterijska funkcija

   $J = \Sigma^k_{j=1} \Sigma_{x \in Sj(m)} || x - Z_j (m+1) ||^2$ minimalna.

   Središta grupa koja minimiziraju kriterijsku funkciju u m – toj iteraciji su aritmetičke srednje vrijednosti uzoraka pojedinih grupa

   $Z_j (m+1) = 1/N_j (\Sigma_{x=Sj(m)} x)$ za j = 1, 2, ..., k; N_j je broj uzoraka u grupi.

4. ako je Z_j (m+1) = Z_j (m) za sve j =1, 2, ..., k, postupak završava. Ukoliko taj uvjet nije ispunjen, ponavljamo postupak od koraka 2.

Na rezultat grupiranja pomoću algoritma k – srednjih vrijednosti utječe:

- broj grupa
- izbor početnih središta grupa

Algoritam zahtjeva eksperimentiranje s različitim brojem grupa k i različitim početnim rasporedom k centara grupa.

Nema općenitog dokaza o konvergenciji algoritma.

Metoda glavnih komponenata i grupiranje

Može se raditi PCA prije grupiranja kako bi se reducirao veliki broj varijabli i time smanjilo ukupno računanje. Rezultati se sa i bez pretprocesiranja s PCA mogu razlikovati!