Klasterizacija

Motivacija

- Klasterizacija je proces razmatranja skupa tačaka i grupisanja tih tačaka u klastere na osnovu neke mjere rastojanja
  - Minimizovati rastojanje unutar klastera
- Veliki skup podataka, veliki broj atributa, prostor koji nije Euklidski

Osnovni pojmovi

- Skup podataka za klasterizaciju je kolekcija tačaka iz nekog prostora
  - Ekulidski prostor sadrži vektore realnih brojeva, komponente vektora su koordinate, dužina vektora je dimenzionalnost
  - Euklidsko rastojanje, Manhattan rastojanje, L_∞ rastojanje
- Prostori koji nijesu Euklidski
  - Mjera rastojanja: Jaccard, kosinus, Hamming, edit
- Rastojanje je funkcija dva argumenta koja je nenegativna, simetrična i zadovoljava nejednakost trougla

Strategije za klasterizaciju

- Hijerarhijsko, svaka tačka je inicijalno klaster, najbliži klasteri se iterativno spajaju sve dok se ne dostigne kriterijum zaustavljanja
- Point assignment, polazi se od nekog skupa klastera, a onda se svaka tačka dodjeljuje najbližem, u nekim varijantama je dozvoljeno spajanje ili/i dijeljenje klastera, pa i da neke tačke ne budu pridružene nijednom klasteru (izuzeci)
- Dodatna podjela na Euklidske ili ne (postojanje centroida), i na algoritme koji koriste eksternu memoriju (sumarna reprezentacija klastera) ili rade samo sa glavnom memorijom

The course of dimensionality

- Anomalije u Euklidskim prostorima sa velikim brojem dimenzija (atributa)
  - Skoro svi vektori su uzajamn normalni
  - Skoro svi vektori su jednako udaljeni jedni od drugih

Hijerarhijsko klasterisanje u Ekulidskom prostoru

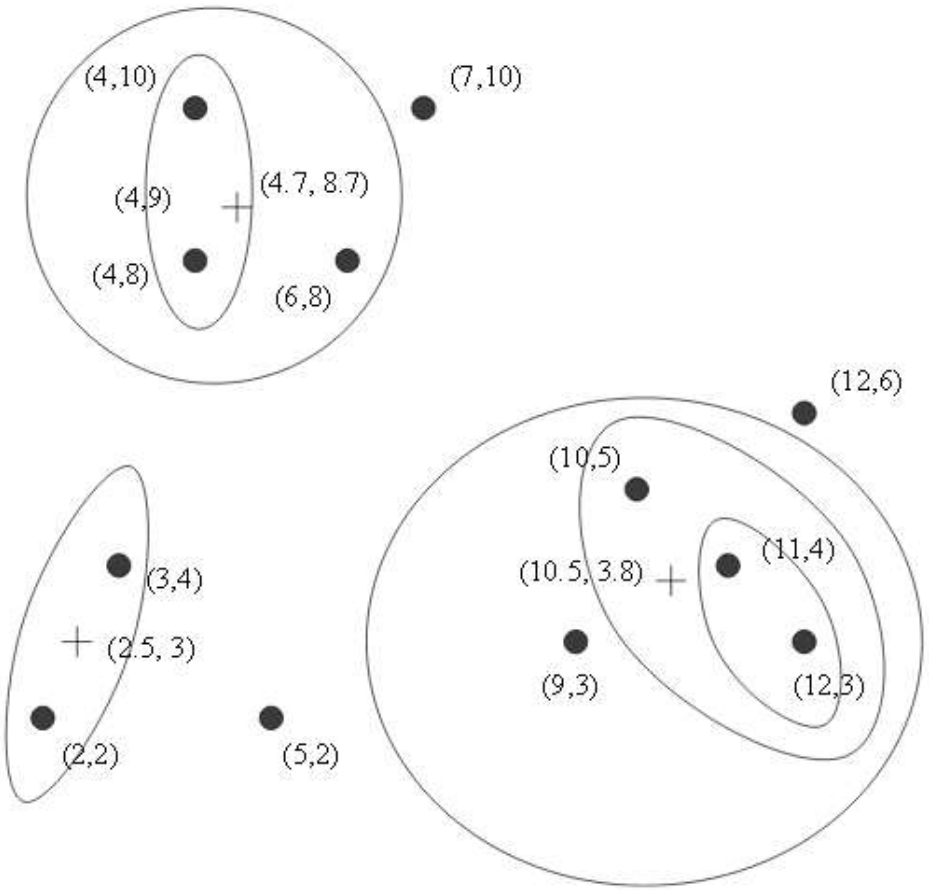
- Predstavljanje klastera centroidom, mali skup podataka
- Zaustavljanje kada se dostigne unaprijed definisani broj klastera, kada se generiše „loš“ klaster
- Drvo klastera, dendogram

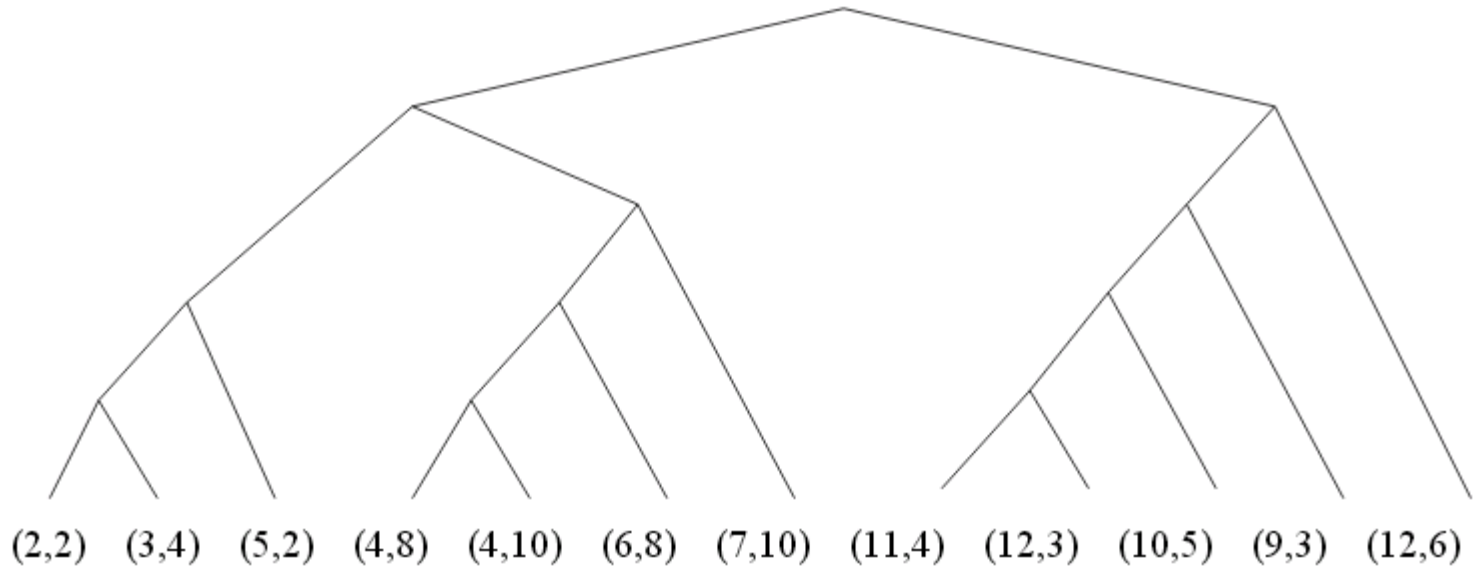
```
WHILE it is not time to stop DO
    pick the best two clusters to merge;
    combine those two clusters into one cluster;
END;
```

Složenost

- Osnovni algoritam je složenosti $O(n^3)$
- Upotreba prioritetnog reda $O(n^2 \log n)$
- Napomena
  - Koja dva klastera se spajaju: radius, dijametar
  - Zaustavljanje kada dijametar pređe unprijed određeni prag, kada je gustina ispod nekog praga

Primjer


(4,10)
(7,10)
(4,9)
(4.7, 8.7)
(4,8)
(6,8)
(12,6)
(10,5)
(11,4)
(10.5, 3.8)
(9,3)
(12,3)
(3,4)
(2.5, 3)
(2,2)
(5,2)


(2,2)
(3,4)
(5,2)
(4,8)
(4,10)
(6,8)
(7,10)
(11,4)
(12,3)
(10,5)
(9,3)
(12,6)

Prostori koji nijesu Euklidski

- Ne postoje centroidi
- Clustroid (klasteroid) je tačka iz klastera koja minimizuje
  - Sumu rastojanja ili kvadrata rastojanja do ostalih tačaka u klasteru
  - Maksimalno rastojanje do neke tačke u klasteru

	ecdab	abecb	aecdb
abcd	5	3	3
aecdb	2	2	
abecb	4		

Point	Sum	Max	Sum-Sq
abcd	11	5	43
aecdb	7	3	17
abecb	9	4	29
ecdab	11	5	45

K-Means algoritam

- Point assignment strategija, za Euklidske prostore, unaprijed je poznat broj klastera K

```
Initially choose k points that are likely to be in
    different clusters;
Make these points the centroids of their clusters;
FOR each remaining point p DO
    find the centroid to which p is closest;
    Add p to the cluster of that centroid;
    Adjust the centroid of that cluster to account for p;
END;
```

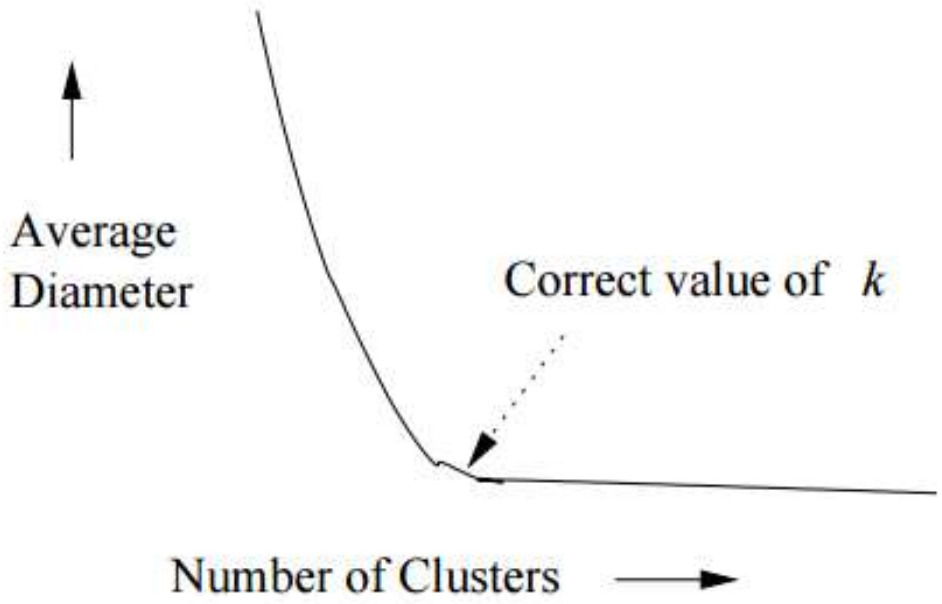
Inicijalizacija klastera

- Dvije strategije
  - Izabrati tačke na najvećem mogućem rastojanju, polazi se od slučajno izabrane tačke

```
WHILE there are fewer than k points DO
    Add the point whose minimum distance from the selected
      points is as large as possible;
END;
```

  - Klasterisati uzorak u K klastera

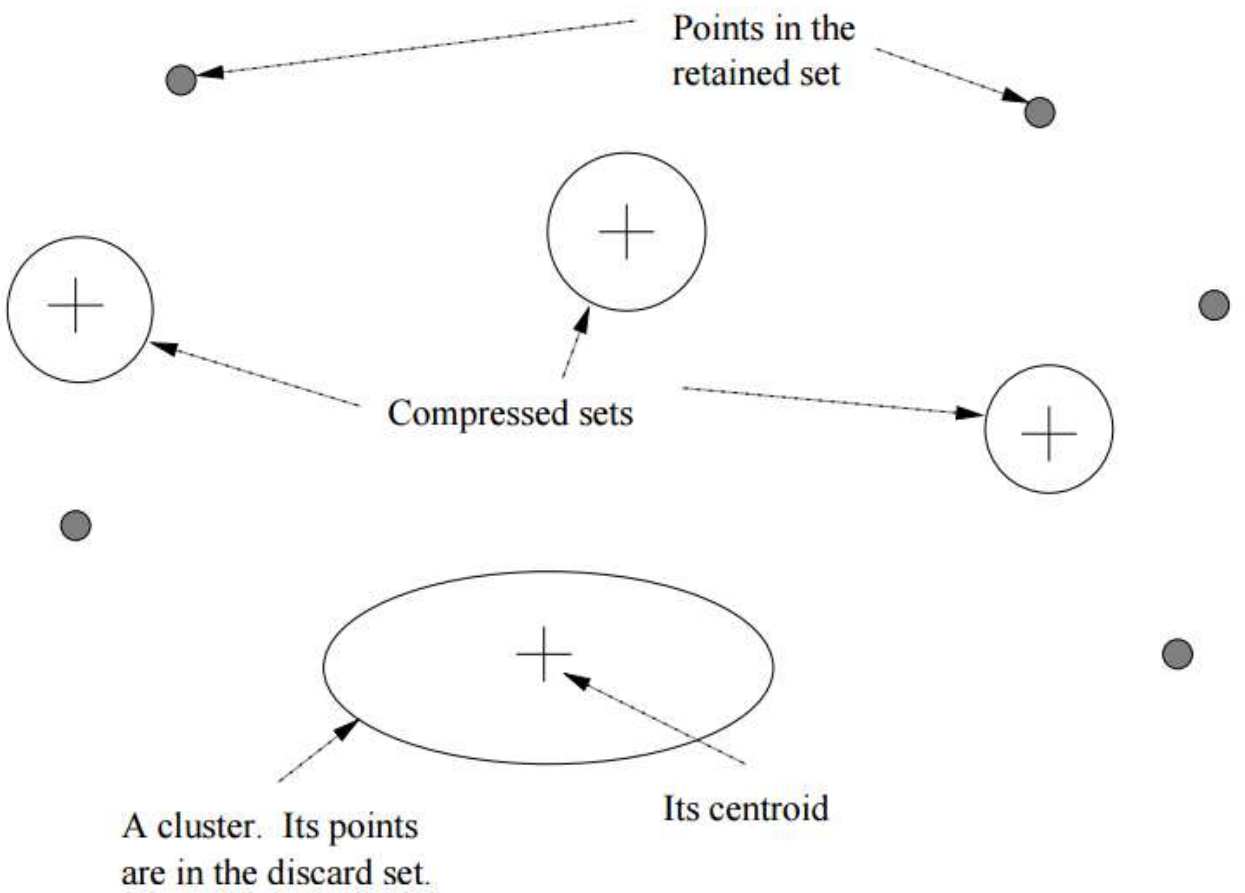
Kako odrediti broj klastera K?



BFR algoritam

- Za Euklidske prostore velike dimenzionalnosti, large-scale clustering algoritam, rastojanje tačaka od centroida zadovoljava normalnu raspodjelu, očekivanje i disperzija mogu da se razlikuju za razne dimenzije ali su dimenzije nezavisne
- Na početku se bira K tačaka, onda se tačke čitaju u chunk-ovima odgovarajuće veličine kako bi bili obrađeni u operativnoj memoriji
- Dodatno u operativnoj memoriji smještaju se:
  - Discard set: reprezentacija klastera
  - Compressed set: reprezentacija grupe tačaka koje nijesu pridružene nijednom klasteru, miniklasteri
  - Retained set: tačke koje nijesu u prethodne dvije reprezentacije

BFR algoritam (1)



BFR algoritam (2)

- Discard i compressed skup predstavljeni su sa 2d+1 komponenti ako je prostor dimenzije d
  - Broj tačaka u klasteru, N
  - Vektor SUM sa d komponenti, gdje je SUM_i suma po i-toj komponenti svih tačaka u klasteru
  - Vektor SUMSQ, sa sumama kvadrata za svaku dimenziju
- Na ovaj način skup tačaka reprezentujemo sa njihovim brojem, centroidom i standardnom disperzijom za svaku dimenziju

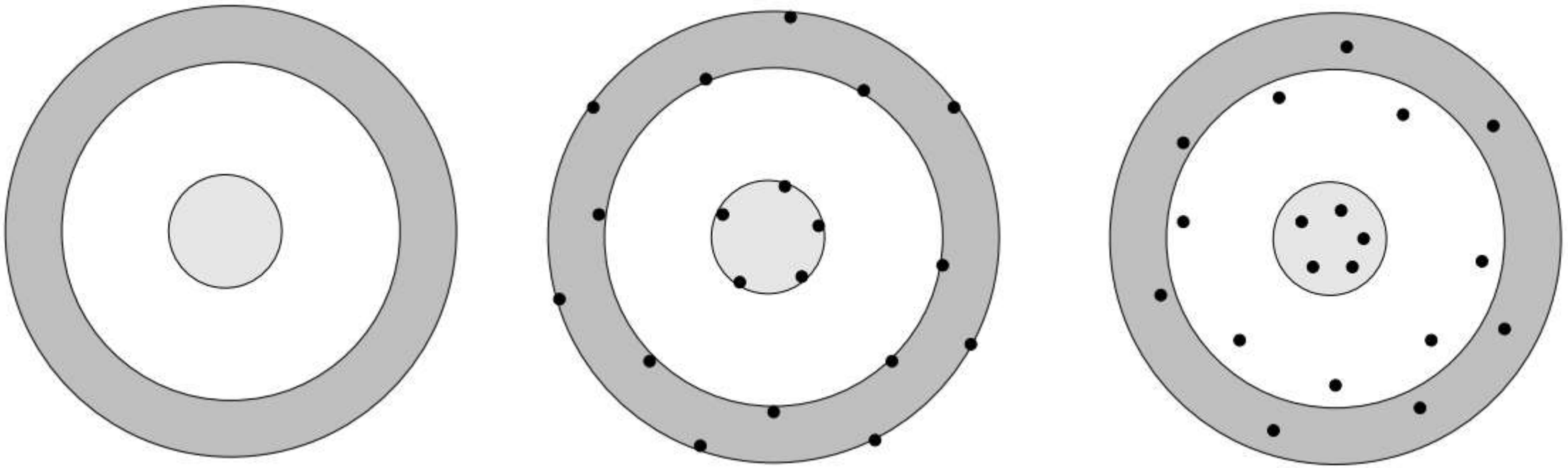
BFR algoritam (3)

- Za svaki chunk
- 1. tačke se pridružuju najbližem centroidu (klasteru)
- 2. tačke koji nijesu dovoljno blizu nijednom centroidu klasterišu se u odnosu na retained set. Klasteri sa bar dvije tačke dodaju se u compressed set. Singlton klasteri ostaju u retained set
- 3. spajanje klastera iz tačke 2 sa klasterima iz „starog“ compressed set-a
- Tačke koje nijesu u retained set-u brišu se iz operativne memorije
- Ako je posljednji chunk obrađen, dodatno obraditi retained i compressed set-ove

CURE algoritam

- Za Euklidske prostore, nema ograničenja u smislu raspodjele tačaka, umjesto centroida upotrebljava skup reprezenata
- Inicijalizacija CURE algoritma
  - Klasterizacija uzorka hijerarhijski
  - Biranje reprezenata, najudaljenije tačke unutar istog klastera
  - Pomjeranje svakog reprezenta za 20% rastojanja do centroida, u pravcu centroida

CURE algoritam (1)



CURE algoritam (2)

- Spajaju se dva klastera ukoliko postoji par reprezenata koji su dovoljno blizu
- Čitanje tačaka i dodjela klasteru sa najbližim reprezentima

GRGPF algoritam

- Ne-Euklidski prostor, non-main-memory data, kombinacija hijerarhijske i point assignment strategija
- Stablo klastera, klasteri u listovima su blizu, klasteri dostižni iz jednog unutrašnjeg čvora su takođe relativno blizu
- Reprezentacija klastera u operativnoj memoriji – features
- Ako je p tačka iz klastera, ROWSUM(p) je suma kvadrata rastojanja p do svih ostalih tačaka iz klastera
- Mora da postoji neka mjera rastojanja d iako prostor nije Euklidski

Reprezentacija klastera

- Sljedeća svojstva čine reprezentaciju klastera
  - N, broj tačaka u klasteru
  - Clustroid c, tačka iz klastera za koju je ROWSUM minimalno
  - ROWSUM(c)
  - Za neko k, najbližih k tačaka u odnosu na clustroid c, njihovi ROWSUM
  - najudaljenijih k tačaka u odnosu na clustroid c, njihovi ROWSUM
- Veličina reprezentacije klastera ne zavisi od broja tačaka u njemu

Inicijalizacija stabla klastera

- Genriše se uzorak koji se hijerarhijskom strategijom klasteriše, rezultat je stablo T
- Dalje se biraju čvorovi iz T, koji reprezentuju klastere unaprijed definisane veličine n
- Ovi se klasteri smještaju u listove stabla GRGPF algoritma
  - Klasteri sa zajedničkim pretkom u T spajaju se u unutrašnje čvorove stabla GRGPF algoritma
  - Balansiranje stabla, slično kao za B-stablo

Procesiranje tačaka

- Tačke se čitaju sa diska i pridružuju najbližem klasteru, počinje se od korijena i traže se najbliži clustroidi, proces se zaustavlja u listu
- U listu se mijenja reprezentacija klastera sa najbližim clustroidom
  - Inkrementira se N
  - Dodaje se rastojanje između tekuće tačke p i svih tačaka iz reprezentacije q (clutroid, k najbližih i k najudaljenijih tačaka)
  - Ukoliko je potrebno procjenjuje se ROWSUM(p) = ROWSUM(c) + $Nd^2(p, c)$
- Dijeljenje klastera sa radijusom većim od SQRT(ROWSUM(c)/N)
  - Dijeljenje čvorova kao u B stablu
- Spajanje klastera ukoliko stablo prelazi veličinu operativne memorije