

# Fazi klasterovanje

Seminarski rad u okviru kursa  
Racunarska inteligencija  
Matematički fakultet

Luka Marković, Tamara Radovanović  
luka.markovic.d@gmail.com, radovanovic.tamara.t@gmail.com

5. april 2019.

## Sažetak

Ovaj seminarski rad je posvćen fazi klasterovanju podataka. Predstavljena su dva algoritma, FCM i Gustafson-Kessel, kao i rezultati njihovih izvršavanja na nekoliko skupova podataka. Pored ovih algoritama, pomenute su jos neke metode bez njihove implementacije.

## Sadržaj

<b>1</b>	<b>Uvod</b>	<b>2</b>
<b>2</b>	<b>Klasterovanje</b>	<b>2</b>
2.1	Fuzzy logika i klasterovanje . . . . .	2
<b>3</b>	<b>FCM</b>	<b>3</b>
3.1	Opis . . . . .	3
3.2	Primena FCM . . . . .	4
3.3	Prednosti i mane FCM algoritma . . . . .	4
<b>4</b>	<b>Druge vrste FCM algoritma</b>	<b>6</b>
4.1	Elipsoidni algoritam (Gustafson-Kessel) . . . . .	6
4.2	Algoritmi za prepoznavanje linija i ravni . . . . .	7
4.3	Algoritmi skoljkastih klastera . . . . .	9
<b>5</b>	<b>Zaključak</b>	<b>9</b>
	<b>Literatura</b>	<b>9</b>

# 1 Uvod

Fazi klasterovanje predstavlja jednu podvrstu algoritama klasterovanja. Ova vrsta algoritama je dobra za podatke za koje se ne može sa sigurnošću reći da pripadaju određenom klasteru. FCM algoritam je jedan od najpopularnijih u ovoj oblasti, pre svega zbog svoje jednostavne implementacije. Podred ovog algoritma rasprostranjen je i Gustafson-Kessel algoritam, koji je unapređenje FCM-a i prepoznaje klastere drugih oblika. Postoji nekoliko verzija ovog algoritma kao što su varijante koje prepoznaju linijske i šel klastere [1], [2].

## 2 Klasterovanje

Ideja klaster analize je da podeli dati skup podataka ili objekata u klastere (grupe, klase). Podela treba da ima sledeće osobine:

- Podaci koji pripadaju jednom klasteru treba da budu što je moguće sličniji,
- Podaci koji pripadaju različitim klasterima treba da budu što je moguće više različiti.

Najčešće se kao mera sličnosti/različitosti, koristi euklidsko rastojanje jer su podaci obično vektori realnih brojeva. Podaci treba da budu skalirani. Problemi mogu da nastanu ako nisu samo realni podaci.

Deterministička klaster analiza nije pogodna za svaki skup podataka. Podaci koji se nalaze na jednakoj udaljenosti od centroida se, zbog karakteristika algoritma, moraju dodeliti jednom klasteru, iako podjednako pripadaju u oba klastera. To predstavlja problem koji može da se prevaziđe korišćenjem fazi skupova i fazi klasterovanja, koji nam omogućavaju da podatak može da pripada većem broju klastera istovremeno sa određenim stepenom.

### 2.1 Fuzzy logika i klasterovanje

Fazi skupovi su osmišljeni da bi se omogućilo da podaci ne pripadaju u potpunosti jednom klasteru. U običnim skupovima element mora ili da pripada ili ne pripada skupu, kod fazi skupova to nije slučaj. Da bismo ovo predstavili matematički, koristimo stepen pripadnosti. Fazi skup  $A$  se može predstaviti:  $\mu_A : X \rightarrow [0, 1]$ . Za svako  $x$  pripada  $X$ ,  $\mu_A(x)$  pripada intervalu  $[0, 1]$  i predstavlja stepen pripadnosti elementa  $x$  skupu  $A$ . Može se smatrati da su obični skupovi specijalan slučaj fazi skupova, gde element pripada skupu sa stepenom pripadnosti 0 ili 1.

Pomoću osnovnih tehnika klasterovanja, kao što je k-mean, posmatrane elemente možemo sa sigurnošću da stavimo da pripadaju jednom i samo jednom klasteru. To nije poželjno u nekim situacijama, kao u slučaju kada je element na sličnim udaljenostima između dva klastera i kada ne možemo sa sigurnošću odrediti kom klasteru pripada. Zato fazi klasterovanje dozvoljava klastere sa granicama koje nisu strogo definisane, to omogućava da neki objekat pripada u više klastera sa nekim stepenom sigurnosti. Ako nam vektor  $X$  predstavlja  $n$  objekata i vektor  $K$  predstavlja  $m$  klastera, tada svaki element  $x_i$  mora da pripada sa nekim stepenom svakom klasteru. Takođe, mora da važi da je suma pripadnosti nekog elementa svim klasterima jednaka 1. Iz ovoga se može zaključiti da je obično klasterovanje specijalni slučaj fazi klasterovanja.

### 3 FCM

Fcm predstavlja osnovni algoritam fazi klasterovanja. Ovaj algoritam se može smatrati unapredjenom verzijom k-means algoritma, otuda i njegov prvobitan naziv Fuzzy k-means ili c-means. Ovaj algoritam prepoznaje globularne oblike u p-dimenzionom prostoru. Klasteri su približno iste veličine i imaju svoje centre. Kao mera rastojanja koristi se euklidsko rastojanje između podataka i centroida.

#### 3.1 Opis

Cilj algoritma je minimizacija funkcije:

$$J_{FCM}(P, U; X, c, m) = \sum_{i=1}^c \sum_{k=1}^N (u_{ik})^m d_{ik}^2(\mathbf{x}_k, \mathbf{p}_i) \quad (1)$$

Gde su  $X, c, m$  ulzani parametri, a  $P$  i  $U$  vrednosti koje se traže.

- $c$  je broj klastera za koje se pretpostavlja da postoje u skupu  $X$
- $m > 1$  je fazi eksponent koji određuje koliko će fazi biti skup. Ako je  $m=1$  algoritam se svodi na k-means algoritam. Ukoliko je  $m$  veliki broj za svaki element sve funkcije pripadnosti će biti iste.
- $U_{ik}$  je stepen pripadnosti elementa  $X_k$  klasteru  $p_i$
- $N$  je broj elemenata
- $d_{ik}^2$  je rastojanje između elementa  $X_k$  i centroida  $p_i$ . U klasicnom FCM algoritmu se koristi euklidsko rastojanje, a opsti način izracunavanja je dat formulom 2:

$$d_{ik}^2(\mathbf{x}_k, \mathbf{p}_i) = \|\mathbf{x}_k - \mathbf{p}_i\|_A^2 = ((\mathbf{x}_k - \mathbf{p}_i)^T A (\mathbf{x}_k - \mathbf{p}_i)) \quad (2)$$

gde je  $A$  pozitivno definitna matrica.

Ovde postoji ograničenje koje mora da se ispuni:

$$\sum_{i=1}^c u_{ik} = 1 \quad \forall k \in \{1 \dots N\} \quad (3)$$

Ovo je iterativni algoritam u kome se u svakom koraku ažuriraju vrednosti  $p_i$  i vrednosti matrice  $U$  po formulama 4 i 5:

$$\mathbf{p}_i = \frac{\sum_{k=1}^N u_{ik}^m \mathbf{x}_k}{\sum_{k=1}^N u_{ik}^m} \quad (4)$$

i

$$u_{ik} = \frac{1}{\sum_{j=1}^c \left( \frac{d_{ik}^2}{d_{jk}^2} \right)^{\frac{1}{m-1}}} \quad (5)$$

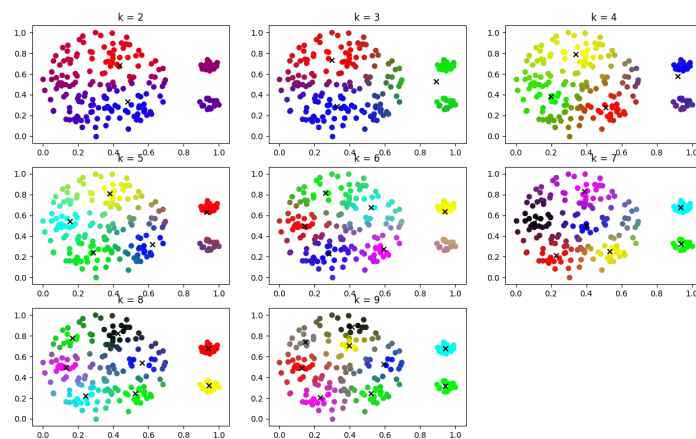
Vremenom ova iteracija konvergira ka minimumu i dobijaju se neki šabloni. Algoritam je predstavljen sledećim koracima:

1. Odabrati broj centroida  $c$ ,  $2 \leq c < N$   
Odabrati  $m$ ,  $1 \leq m < \infty$   
Inicijalizovati matricu pripadnosti  $U$
2. Izračunati c centroida fazi klasera  $P$  kao u jednačini 4
3. Ažurirati matricu pripadnosti po formuli 5
4. Algoritam se završava nakon unapred zadatog broja iteracija ili kada u funkciji  $J_{FCM}$  nema značajnih promena. Inače nazad na koraka 2.

### 3.2 Primena FCM

U ovom delu je dat primer upotrebe FCM algoritma. Podaci su preuzeti sa [adrese](#). Prikazani su rezultati algoritma na tri razlicita skupa za razlicit broj klastera. U prvom i trecem primeru uzet je uzorak od 200 elemenata dok je u drugom uzorak na 3000 elemnata. Broj iteracija je ogranicen na 100, ali ako promena funkcije koja se minimizuje postane manja od 0.001 prekida se izvršavanje.

Na slikama 1, 2 i 3 su rezultati za  $m=2$  i broj klastera između 2 i 9 na različitim skupovima. Može se videti da elementi koji su bliži centroidu klastera sa većim stepenom pripadaju tom klasteru. Stepem pripadnosti se smanju što su elementi dalji od centroida klastera. Slabost ovog algoritma je što se očekuje da će svi klasteri biti jednake veličine.



Slika 1: *FCM* algoritam na prvom skupu

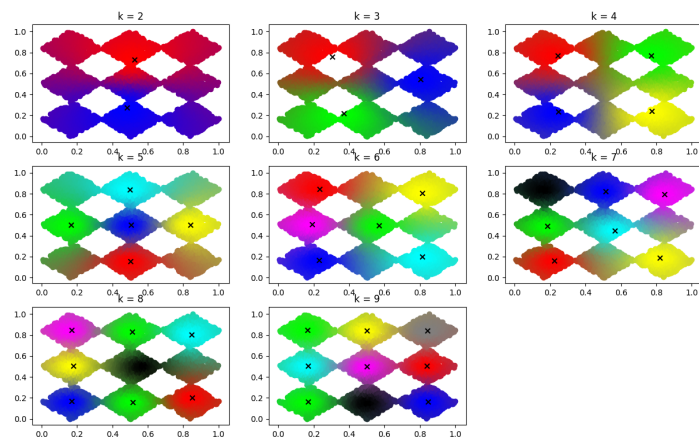
Pri klasifikaciji podataka sa slike 1 očekivano je da algoritam da najbolje rezultate za  $c = 3$ , ali FCM ne radi dobro za klasterne različite veličine, pa zbog toga ne daje očekivane rezultate.

Na slici 4 se vide rezultati rada algoritma kada je parametar  $m$  postavljen na 3. Povećavanjem ovog parametra povećava se rasplinitost skupa, odnosno svaki element sa sličnijim stepenom pripada svakom skupu i zbog ovoga granice klastera postaju nejasnije.

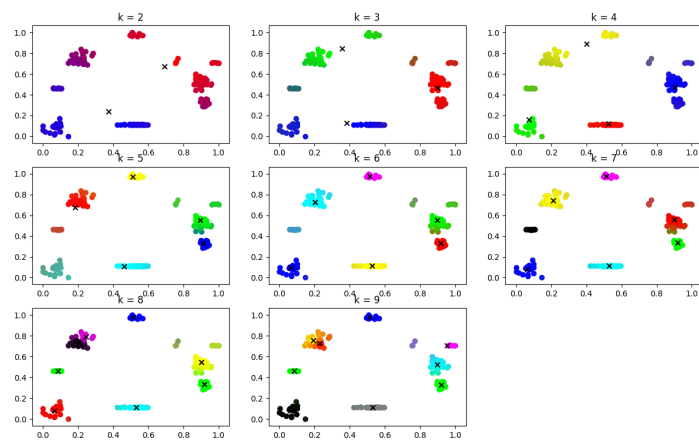
### 3.3 Prednosti i mane FCM algoritma

*FCM* je postao jedna od popularnijih metoda fazi klasterovanja. Implementacija je jednostavna i pravolinijska. Za sve sferne tipove klastera, *FCM* radi prilično tačno.

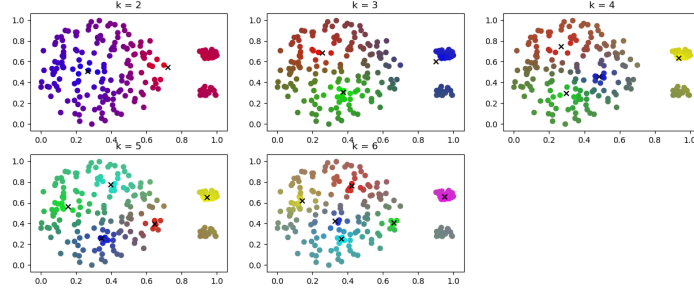
Jedna od najvećih mana je to što zahteva broj klastera unapred, što nije uvek moguće znati. Takođe, na početku rada algoritma treba inicijalizovati početne centre, to dosta utiče na pretragu. Iz tog razloga, dva pokretanja istog koda mogu dovesti do potpuno različitih rezultata. Algoritam može da prepozna samo klasterne istog oblika, a takođe je i osetljiv na šumove i elemente van granica.



Slika 2: *FCM* algoritam na drugom skupu



Slika 3: *FCM* algoritam na trećem skupu



Slika 4: *FCM* algoritam na prvom skupu za parametar  $m = 3$

## 4 Druge vrste FCM algoritma

Fcm algoritam, iako ima svoje slabosti, nastavio je da se razvija i poboljšava. Koristeći drugačije funkcije rastojanja moguće je podesiti algoritam da prepoznaje elipsoidne klastere. Takođe postoje varijante algoritma koje prepoznaju linije ili ravni i školjkaste oblike (shell clustering). Ipak sve ove nadogradnje idalje ne mogu istovremeno da prepoznaju klastere različitih oblika. Isto kao i FCM, osetljivi su na početnu inicijalizaciju, ali i dodatno su skuplji u odnosu na njega.

### 4.1 Elipsoidni algoritam (Gustafson-Kessel)

Gustafson i Kessel su predstavili verziju algoritma koja za svaki klaster koristi različitu  $A$  normu za računanje distance između centara i elemenata. Zahvaljujući ovome, svaki klaster ima različiti elipsoidni oblik. Modifikovana funkcija minimizacije sad izgleda:

$$J(P, U; X, c, m) = \sum_{i=1}^c \sum_{k=1}^N u_{ik}^m \|\mathbf{x}_k - \mathbf{p}_i\|_{A_i}^2 = \sum_{i=1}^c \sum_{k=1}^N u_{ik}^m ((\mathbf{x}_k - \mathbf{p}_i)^T A_i (\mathbf{x}_k - \mathbf{p}_i)) \quad (6)$$

gde je  $A_i$  pozitivno definitna simetrična matrica. Pored ograničenja 3 postoji i dodatno ograničenje:

$$\|A_i\| = \rho_i = \text{constant}$$

Koraci algoritma su:

- Odabrati broj centroida  $c$ , fazi eksponent  $m$ , inicijalizovati početne vrednosti centroida i svaku matricu  $A_i$
- Ažurirati matricu  $U$  po formuli 5
- Ažurirati centroide  $P$  po formuli 4

- Izračunati nove vrednosti matrice A:

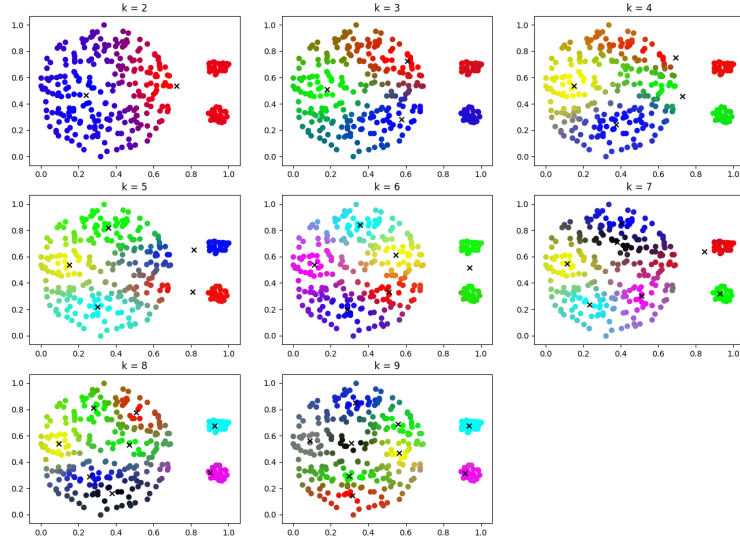
$$A_i^{-1} = \left( \frac{1}{\rho_i |C_i|} \right)^{\frac{1}{m}} C_i \quad (7)$$

gde je  $C_i$  fazi matrica kovarijanse:

$$C_i = \frac{\sum_{k=1}^N u_{ik}^m (\mathbf{x}_k - \mathbf{p}_i)(\mathbf{x}_k - \mathbf{p}_i)^T}{\sum_{k=1}^N u_{ik}^m} \quad (8)$$

- Ako uslovi zaustavljanja nisu ispunjeni, vratiti se na korak 2.

Korišćenjem različite matrice A za računanje dobijaju se elipsoidni klasteri različite orijentacije. Ograničenje je što A određuje veličinu klastera i tera ih da svi budu iste veličine. Rezultati primene ovog algoritma su prikazani na slikama 5, 6 i 7. Parametar m je 2, a broj klastera varira od 2 do 9.



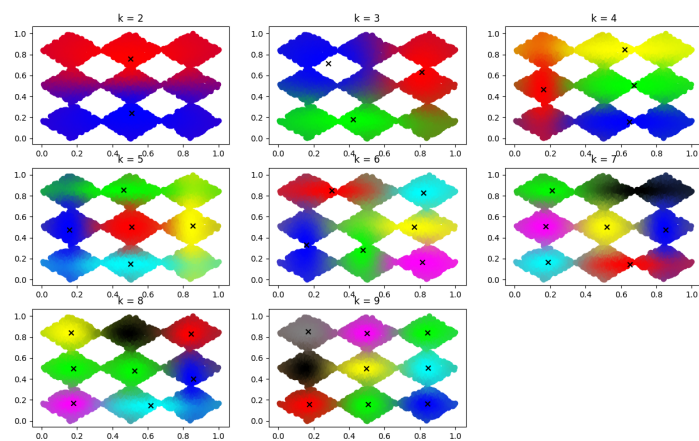
Slika 5: Gustafson i Kessel algoritam na prvom skupu

Na slici 6 se najlakše može primetiti da ovaj algoritam daje klasterne elipsoidnog oblika.

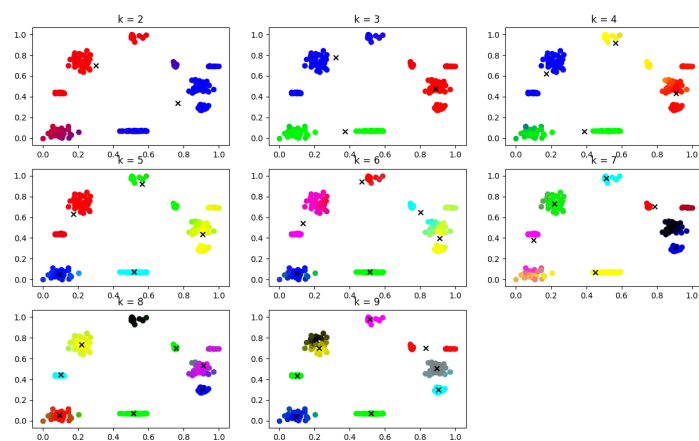
## 4.2 Algoritmi za prepoznavanje linija i ravni

Ova varijanta FCM algoritma je predstavljena od strane Bezdeka i služi za detekciju klastera koji imaju oblik linija ili ravni. Glavna ideja ovog algoritma je da se euklidsko rastojanje olabavi za računanje vrednosti koje pripadaju istom klasteru (leže na istoj liniji ili ravni), dok se primenjuje normalno euklidsko rastojanje za ostale tačke. Ovo se postiže tako što se rastojanje računa kao kombinacija dve mere rastojanja:

$$d^2(\mathbf{x}_k, \mathbf{p}_i) = \alpha d_{V_{ik}}^2 + (1 - \alpha) d_{E_{ik}}^2 \quad (9)$$



Slika 6: Gustafson i Kessel algoritam na drugom skupu



Slika 7: Gustafson i Kessel algoritam na trećem skupu



Ovde je  $d_{E_{ik}}^2$  euklidsko rastojanje, dok je  $d_{V_{ik}}^2$ :

$$d_{V_{ik}}^2 = \|\mathbf{x}_i - \mathbf{p}_i\|^2 - \sum_{j=1}^r ((\mathbf{x}_i - \mathbf{p}_i) \cdot \mathbf{e}_{ij})$$

gde je  $r \in [1, p]$  i  $\mathbf{e}_{ij}$  je  $j$ -ti sopstveni vektor matrice kovarijanse  $C_i$  klastera  $i$ . ( $\cdot$  ovde predstavlja skalarni proizvod dva vektora) Kada je  $r = 1$ ,  $d_{V_{ik}}^2$  se može koristiti za detekciju linija, dok u slučaju da je  $r = 2$  se može koristiti za detekciju ravni. Parametar  $\alpha$  u jednačini 9 može da ima vrednosti između 0 i 1 i mora da bude zadata apriori. Postoji varijanta gde se dinamički određuje parametar  $r$ , čime se postiže da klasteri ne moraju da budu iste veličine. Međutim, kako pronalazi samo linearne strukture, može se desiti da uhvati podatke koji mu ne pripadaju.

### 4.3 Algoritmi skoljkastih klastera

Ova vrsta algoritama se najčešće koristi u obradi slika. Slike se pretprocesiraju kako bi se pronašle ivice, a zatim se pikseli koji se nalaze na tim ivicama prosledjuju ovom algoritmu kako bi se odredila granica. Postoji više algoritama za skoljkaste klasterne, ali za sve njih je novina mera rastojanja koju koriste. Centroidi su opisani centralnom tačkom  $p_i$  i poluprečnikom  $r_i$ . Skoljkasti klaster algoritmi su skupi zbog rešavanja nelinearnih jednačina koje zahteva iterativne metode

## 5 Zaključak

U radu je predstavljeno fazi klasterovanje kao i vodeći algoritmi iz ove oblasti. Može se reci da ove metode mogu dati dobre rezultate za probleme gde se klasteri preklapaju, ali su takođe spore i izbegava se rad sa njima u realnom vremenu. Imaju primenu u raznim oblastima, kao što su medicina, segmentacija slika i mnoge druge.

## Literatura

- [1] Frederick Evers. *The bases of competence : skills for lifelong learning and employability*. Jossey-Bass, San Francisco, Calif, 1998.
- [2] Ahmed Ismail Shihab. *FUZZY CLUSTERING ALGORITHMS AND THEIR APPLICATION TO MEDICAL IMAGE ANALYSIS*. PhD thesis, University of London, 2000.