Fazi klasterovanje

Seminarski rad u okviru kursa
Računarska 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 izvšavanja na nekoliko skupova podataka. Pored ovih algoritama, pomenute su jos neke metode bez njihove implementacije.

Sadržaj

1 Uvod

Fazi klasterovanje predstavlja jednu podvrstu algoritama klasterovanja. Ova vrsta algoritama je dobra za podatke za koje se ne može sa sigurnošu 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 dručijih oblika. Postoji nekolik verzija ovog algoritma kao što su varijante koje prepoznaju linijske i šel klastere [1], [2].

2 Klasterovanje

Ideja klster 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 prestavlja problem koji može da se prevaziđe korišćenjem fazli skupova i fazi klasterovanja, koji nam omogučavaju 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 siskretnim skupovima element mora ili da pripada ili ne pripada skupu, dok kod fazi skupova to nije slučaj. Da bismo ovo predstavili matematički, koristimo stepen pripadnosti. Fazi skup A se može predstaviti funkcijom: $\mu A : X \to [0, 1]$. 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 diskretni skupovi specijalan slučaj fazi skupova, gde element prpada 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 defisane, 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 unapređenom 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:

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

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

- c je broj klastera za koje se pretposatalja da postoje u skupu X
- m > 1 je fazi eksponent koji određuje koliko će fazi biti skup. Ako je m=1 algoritm 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 klasičnom FCM algoritmu se koristi euklidsko rastojanje, a opšti način izračunavanja je dat formulom 2:

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

  gde je A pozitivno definitna matrica.

Ovde postoji ograničenje koje mora da se ispuni:

$$\sum_{i=1}^{c} u_{ik} = 1 \qquad \forall k \in \{1...N\} \tag{3}$$

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

$$\mathbf{p}_i = \frac{\sum_{i=1}^{N} u_{ik}^m \mathbf{x}_k}{\sum_{i=1}^{N} u_{ik}^m} \tag{4}$$

i

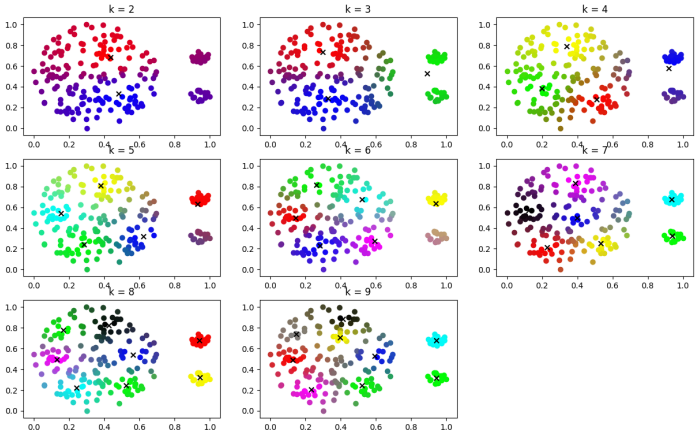
$$u_{ik} = \frac{1}{\sum_{j=1}^{c} (\frac{d_{ik}^2}{d_{jk}^2})^{\frac{1}{m-1}}} \tag{5}$$

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 pripadnsti U
2. Iznaračunati c cetroida fazi klasera P kao u jednačini 4
3. Ažurirati matricu pripadnosti po formuli 5
4. Alogoritam se završava nakon unapred zadatog broja iteracia 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 različita skupa za različit broj klastera. U prvom i trećem 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. Stepen pripadnosti se smanjuje što je element 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

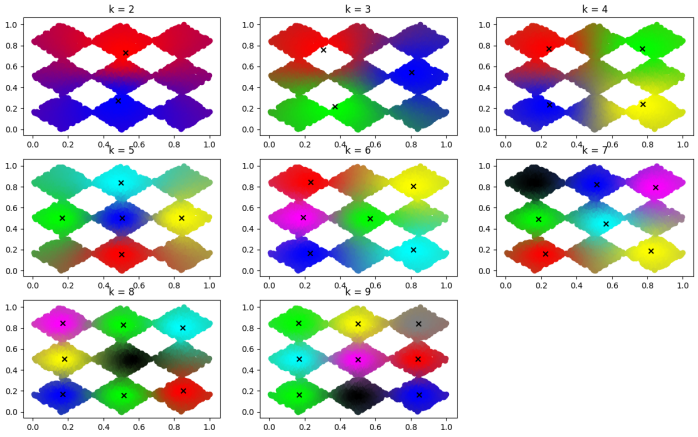
Pri klasifikaciji podataka sa slike 1 očekivno je da algoritam da najbolje rezultete za $c = 3$, ali FCM ne radi dobro za klastere različitih veličina, pa zbog toga ne daje očekicane rezultate.

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

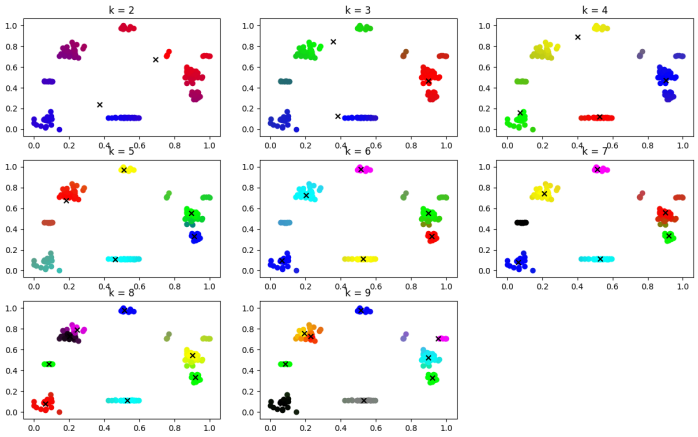
3.3 Prednosti i mane FCM algoritma

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

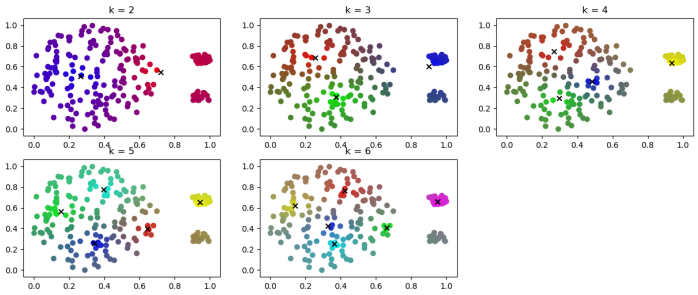
Jedna od najvećih mana je to što zahteva broj klastera unapred, što nije uvek moguće znati. Takođe, na početku rada algoritma treba inicijalizovati početne centre, to dosta utiče na pretragu. Iz tog razloga, ako je početna inicijalizacija klastera slučajna, dva pokretanja istog koda mogu dovesti do potpuno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



Slika 2: FCM algoritam na drugom skupu



Slika 3: FCM algoritam na trećem skupu



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

granica.

4 Druge vrste FCM algoritma

Fcm algoritam, iako ima svoje slabosti, nastavio je da se razvija i poboljšava. Koristeći drugačije funkcije rastojanja mogće je podesiti algoritam da prepoznaje elipsoidne klastere. Takođe postoje varijante algoritma koje prepoznaju linije ili ravni i šel klastere. 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:

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

gde je A_i pozitivno definitna simetrična matrica. Pored ograničenja 3 postiji i dodatno ogranicenje:

$$||A_i|| = \rho_i = constant$$

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:

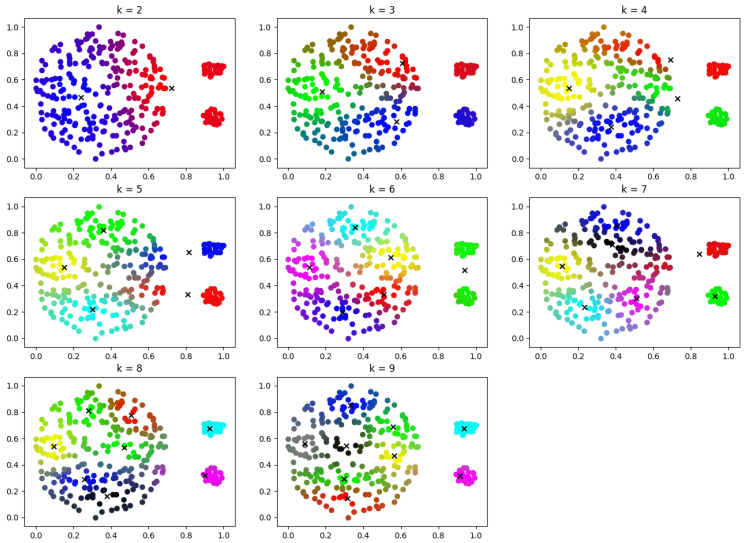
$$A_i^{-1} = (\frac{1}{\rho_i |C_i|})^{\frac{1}{m}} C_i \tag{7}$$

gde je C_i fazi matrica kovarijanse:

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

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

Korišćenjem različite matrice A za računanje dobijaju se elipsoidni klasteri različite orjentacije. 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.



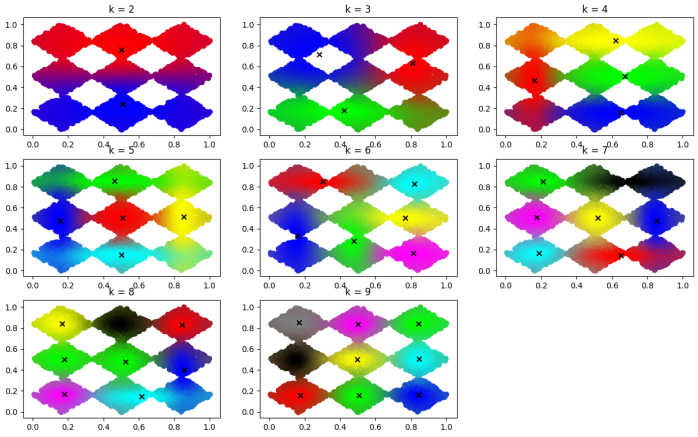
Slika 5: Gustafson i Kessel algoritam na prvom skupu

Na slici 6 se najlakše može primetit da ovaj algotiam daje klastere elipsoidnog oblika.

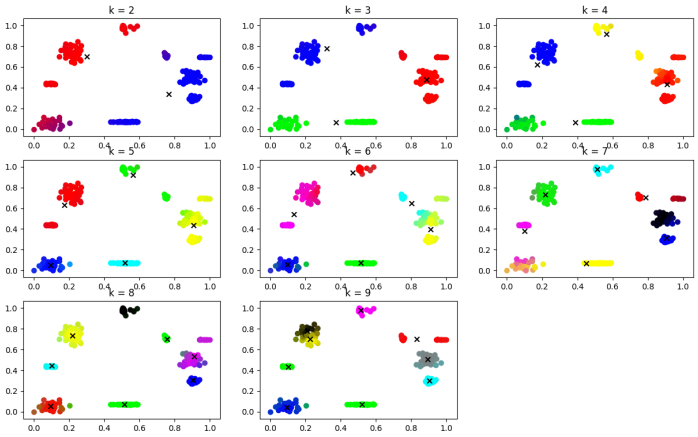
4.2 Algoritmi za prepoznavanje linija i ravni

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:

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



Slika 6: Gustafson i Kessel algoritam na drugom skupu



Slika 7: Gustafson i Kessel algoritam na trećem skupu

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

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

gde je $r \in [1, p]$ i e_{ij} je j-ti sopstveni vektor matrice kovarijanse C_i klastera i. ($\cdot$ ovde predstavlja skalarni proizvod dva vektora) Kada je $r = 1$, $d^2_{V_{ik}}$ se može koristiti za detekciju linija, dok u slucaju da je $r = 2$ se može koristiti za detekciju ravni. Parametar α u jednačini 9 može da ima vrednosti između 0 i 1 i mora da bude zadat 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 šel klastera

Ova vrsta algoritama se najčesce koristi u obradi slika. Slike se pretprocesiraju kako bi se pronašle ivice, a zatim se pikseli koji se nalaze na tim ivicama prosleđuju ovom algoritmu kako bi se odredila granica. Postoji više algoritama za šel klastere, ali za sve njih je novina mera rastojanja koju koriste. Centroidi su opisani centralnom tackom p_i i poluprečnikom r_i. Šel klaster algoritmi su skupi zbog rešavanja nelinearnih jednacina koje zahteva iterativne metode

5 Zaključak

U radu je predstavljeno fazi klasterovanje kao i vodeći algoritmi iz ove oblasti. Može se reći 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.