Generički algoritam klasterovanja zasnovan na optimizaciji rojem čestica

XLVIII Međunarodni simpozijum
o operacionim istraživanjima

Denis Aličić
denis_alicic@matf.bg.ac.rs

31. maj 2021.

Sažetak

U ovom radu je opisan razvijeni algoritam za rešavanje problema klasterovanja zasnovan na optimaciji rojem čestica [4]. Algoritam je zasnovan na centroidima, slično algoritmu K sredina. Najvažniji parametar algoritma je funkcija kvaliteta klasterovanja koja se optimizuje i time određuje tip klastera. Jedna čestica predstavlja niz centroida dužine k, zadate kao maksimalan broj klastera. U svakoj iteraciji algoritma se svakoj instanci iz skupa podataka dodeli klaster na osnovu najbližeg centroida. Funkcija udaljenosti se takođe prosleđuje algoritmu i od nje zavisi oblik klastera. Zatim se primenjuje funkcija kvaliteta i na osnovu pozicije najbolje čestice i najbolje pozicije trenutne čestice pomeri centroid svakoj čestici u skladu sa originalnim PSO (eng. *Particle swarm optimization*) algoritmom.
Prikazani su eksperimenti na nekim poznatim skupovima podataka, kao i na sintetičkim podacima u dvodimenzionalnom prostoru radi vizuelizacije.

1 Uvod

Klasterovanje je jedan od najpopularnijih problema nenadgledanog učenja. Predstavlja identifikaciju i grupisanje sličnih instanci u datom skupu podataka. Primene ovog metoda su vrlo široke. Od zamene grupa njihovim predstavnicima zarad smanjenja broja instanci u skupu podataka, do detekcije raznorodnih tkiva na medicinskim snimcima i identifikaciji sličnih grupa korisnika društvenih mreža u svrhu oglašavanja.

Obzirom da za mnoge primene nije moguće jednoznačno odrediti šta je dobro klasterovanje, razvijene su različite metode klasterovanja. Postojeći algoritmi se mogu svrstati u nekoliko kategorija u odnosu na vrstu i broj klastera koji su njihov rezultat. Nekim metodama se zadaje ciljni broj klastera, dok drugi kao izlaz mogu dati različit broj klastera. S druge strane, što se tiče vrste klastera koje daju kao izlaz razlikujemo: Globularne, dobro razdvojene, gustinske, hirejarhijske itd.[5]

Predloženi metod može u zavisnosti od njegovih parametara: funkcije sličnosti i funkcije kvaliteta, koje će biti opisane kasnije, da generiše različite vrste klastera, dok mu se kao poseban parametar zadaje maksimalan broj klastera. Algoritam je primarno zasnovan na centroidama.

2 Klasterovanje kao optimizacioni problem

Kod problema klasifikacije, koji je vid nadgledanog učenja, postoje dobro definisane funkcije koje nam mogu reći kakav je kvalitet dobijenog modela. To su pre svega tačnost i preciznost, mada postoje još neke.

Klasterovanje je problem nenadgledanog učenja, tako da za konkretno grupisanje ne postoje jednoznačne funkcije koje nam sa sigurnošću mogu reći koliko je ono dobro. Ipak, postoje neke funkcije, definisane tokom vremena od raznih istraživača, koje nam mogu dati ocenu kvaliteta klasterovanja [2][3].

Zanimljivo je primetiti, da iako na prvi pogled

ne deluje tako, algoritam K sredina se takođe može posmatrati kao optimizacioni algoritam. Funkcija koju taj algoritam optimizuje je:

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

gde je d rastojanje, najčešće euklidsko, ali može biti i neko drugo, a c_i centroida i-tog klastera.

2.1 Funkcije evaluacije

Slično metrikama koje se koriste za evaluaciju klasifikacionih modela, postoje funkcije koje koriste informaciju o stvarnim klasama da bi ocenile kvalitet klasterovanja. Neke od njih su:

- Rand indeks
- Homogenost
- V-mera

U realnim primenama stvarne klase nisu dostupne, tako da ove funkcije nisu korišćene niti prilikom implementacije algoritma, niti prilikom evaluacije, te ni u radu neće biti dalje razmatrane.

Funkcije koje su korišćene prilikom implementacija i eksperimenata su:

- Davies-Bouldin indeks 2.1.1
- Calinski-Harabasz indeks 2.1.2

Za izračunavanje ovih funkcija potrebna je informacija o centroidima klastera i dodeli klastera svakoj instanci iz skupa podataka.

2.1.1 Davies-Bouldin indeks

Metrika je zadata formulom:

$$\frac{1}{c} \sum_{i=1}^{c} \max_{i \neq j} \left(\frac{\sigma_i + \sigma_j}{d(C_i, C_j)} \right)$$

gde je c broj klastera, σ prosečno rastojanje svih instanci jednog klastera od njegovog centroida, d rastojanje, najčešće euklidsko i C_i centroid i-tog klastera.

Minimizacijom ove funkcije po C_i dobijamo grupe koje imaju malo rastojanje unutar istog klastera, a veliko rastojanje između različitih klastera [1].

Generalno, to je ideja većine ovih funkcija zasnovanih na centroidama, s tim što na različite načine kvantifikuju rastojanja unutar klastera i između različitih klastera.

Minimalna vrednost ove funkcije se dostiže u 0. Treba biti obazriv sa minimizacijom ove funkcije, jer metode sa mogućnošću povećavanja broja klastera, prilikom optimizacije ove mere teže tome da svaka tačka bude pojedinačni klaster.

U situaciji da postoji beskonačno instanci, deo funkcije $\frac{1}{c}$ bi težio ka nuli, onda bi i cela funkcija težila minimalnoj vrednosti.

2.1.2 Calinski-Harabasz indeks

Calinski-Harabasz indeks je nešto složeniji za izračunavanje i interpretaciju.
Za dati skup podataka E, veličine n_E, koji se klasteruje u k klastera, vrednost indeksa je definisana kao odnos proseka disperzija klastera i disperzija unutar klastera.
Definisan je formulom:

$$s = \frac{tr(B_k)}{tr(W_k)} \times \frac{n_E - k}{k - 1}$$

gde je $tr(B_k)$ trag matrice[1] disperzije centroida klastera i $tr(W_k)$ trag matrice disperzije unutar pojedinačnog klastera.

Matrice se izračunavaju po formulama:

$$B_k = \sum_{q=1}^{k} n_q (c_q - c_E)(c_q - c_E)^T$$

$$W_k = \sum_{q=1}^{k} \sum_{x \in C_q} (x - c_q)(x - c_q)^T$$

gde je C_q skup instanci u klasteru q, c_q centroid klastera q, c_E centroida celog skupa instanci E i n_q broj instanci u klasteru q.

[1] Trag matrice predstavlja sumu elemenata na glavnoj dijagonali: $tr(A) = \sum_{i=1}^{n} a_{ii}$.

Što je ovaj indeks veći, to je klasterovanje bolje, centroide su udaljenije, a rastojanja unutar klastera manja.
Zbog načina računanja disperzije klastera, koje se izračunava kao kvadriran zbir udaljenosti instance od centroida, ovaj indeks je generalno veći za konveksne klastere, što je za algoritam koji je predstavljen u ovom radu prednost jer očekivani izlaz algoritma jesu konveksni klasteri.

2.2 PSO algoritam

Optimizacioni algoritam je zasnovan na optimizaciji rojem čestica [4]. Pseudokod algoritma je prikazan na slici 1.

Osnovnu PSO algoritma čini jedna čestica roja. Čestice predstavljaju jedno rešenje optimizacionog problema. Jednu česticu u algoritmu razvijenom za potrebe ovog rada predstavlja niz centroida klastera. Centroidi klastera zajedno sa funkcijom blizine predstavljaju jedinstveno određeno rešenje problema klasterovanja. Za svaku instancu skupa nad kojim se vrši klasterovanje se određuje pripadajući klaster kao najbliži centroid koristeći funkciju daljine koja je prosleđena algoritmu kao parametar. Od prosleđene funkcije daljine (euklidska distanca, kosinusna ili neka druga), zavisi oblik klastera. Algoritam je generički u smislu izbora funkcije daljine i funkcije evaluacije koja se optimizuje.

Važan deo svakog algoritma optimizacije je njegova sposobnost pretrage širokog prostora rešenja u odnosu na postizanje lokalnog optimuma u nekoj manjoj oblasti. Kod PSO algoritma ovaj problem je rešen korišćenjem dva parametra koji se odnose na kognitivnu i sociološku komponentu. Kognitivna komponenta čestice daje značajnost njenoj najboljoj poziciji, dok se sociološka komponenta odnosi na najbolju poziciju celog roja. Menjanjem vrednosti ove dve komponente, koje su realni brojevi, balansira se izmedju nalaženja globalnog i lokalnog optimuma.

Algoritam 1: PSO algoritam klasterovanja

Rezultat: Broj klastera k, k centroida i dodeljivanje klastera svakoj instanci
Ulaz: Skup podataka, maksimalan broj klastera k, funkcija evaluacije, funkcija udaljenosti
Izlaz: Skup podataka sa pripadajucim klasterima, maksimum k centroida

```
inicijalizuj početni roj;
odredi najbolju česticu;
dok nije postignut kriterijum zaustavlja čini
    za svaki česticu u roju čini
        Ažuriraj brzinu čestice na osnovu pozicije najbolje čestice i najbolje pozicije trenutne čestice;
        Promeni poziciju čestice na osnovu izračunate brzine;
        Na osnovu funkcije udaljenosti odredi pripadnost svakoj instanci skupa odgovarajućem klasteru;
        Izračunaj funkciju evaluacije na osnovu dodeljenog klasterovanja;
        ako nova pozicija bolja od prethodne najbolje pozicije čestice;
        onda
            ažuriraj najbolju vrednost trenutne čestice;
        kraj
        ako nova pozicija bolja od pozicije najbolje čestice;
        onda
            ažuriraj poziciju najbolje čestice;
        kraj
    kraj
kraj
vrati klasterovanje najbolje čestice, broj klastera i njene centroide
```

3 Eksperimentalni rezultati

Algoritam za rešavanje problema klasterovanja, razvijen za potrebe ovog rada, je testiran na

Algoritam	IRIS				WINE			
	DB		CZ		DB		CZ	
	c	index	c	index	c	index	c	index
K Sredina	2	0.40	3	561.62	3	0.53	3	561
PSO	2	**0.28**	3	**575.69**	3	**0.47**	3	**576**

poznatim skupovima podataka: IRIS i WINE. U eksperimentima su upoređeni rezultati algoritma K sredina sa PSO algoritmom. Vrednosti prikazane u tabeli **??** predstavljaju vrednosti funkcija evaluacije klasterovanja, koje su detaljno opisane u 2.1.1 i 2.1.2.

4 Zaključak

Literatura

[1] David L. Davies and D. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1:224–227, 1979.

[2] J. C. Dunn†. Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, 4(1):95–104, 1974.

[3] Calinski T Harabasz and M Karoński. A dendrite method for cluster analysis. In *Communications in Statistics*, volume 3, pages 1–27. 1974.

[4] J. Kennedy and R. Eberhart. Particle swarm optimization. In *Proceedings of ICNN'95 - International Conference on Neural Networks*, volume 4, pages 1942–1948 vol.4, 1995.

[5] M. Nikolić. *Mašinsko učenje*. 2019.