XLVIII Simpozijum o operacionim istraživanjima
SYM-OP-IS 2021
Banja Koviljača, 20-23.9.2021.

GENERIČKI ALGORITAM KLASTEROVANJA ZASNOVAN NA OPTIMIZACIJI ROJEM ČESTICA
GENERIC CLUSTER ALGORITHM BASED ON PARTICLE SWARM OPTIMIZATION

DENIS ALIČIĆ[1], FILIP VIDOJEVIĆ[2], DUŠAN DŽAMIĆ[3], MIROSLAV MARIĆ[4]

[1] Matematički fakultet, Studentski trg 16, Beograd, Srbija, denis_alicic@matf.bg.ac.rs

[2] Matematički fakultet, Studentski trg 16, Beograd, Srbija, filip_vidojevic@matf.bg.ac.rs

[3] Fakultet organizacionih nauka, Jova Ilića 154, Beograd, Srbija, dusan.dzamic@fon.bg.ac.rs

[4] Matematički fakultet, Studentski trg 16, Beograd, Srbija, maricm@matf.bg.ac.rs

---

Rezime: *U ovom radu je predložen algoritam za rešavanje problema klasterovanja zasnovan na optimaciji rojem čestica [2]. Osnovna ideja algortima, kao i kod algoritma K sredina, je centroid [1]. 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 od najpoznatijih skupova podataka. Predloženi algoritam je nadmašio algoritam K sredina.

Ključne reči: *klasterovanje, optimizacija, pso, k-sredina*

Abstract: *In this paper, an algorithm for solving clustering problems based on particle swarm optimization [2] is proposed. The basic idea of algorithms, as with the K mean algorithm, is centroid [1]. The most important parameter of the algorithm is the cluster quality function which is optimized and thus determines the type of cluster. One particle represents a series of centroids of length k, given as the maximum number of clusters. In each iteration of the algorithm, each instance in the data set is assigned a cluster based on the nearest centroid. The distance function is also passed to the algorithm and the shape of the cluster depends on it. Then the quality function is applied and based on the position of the best particle and the best position of the current particle, the centroid moves to each particle in accordance with the original PSO (Particle swarm optimization) algorithm. Experiments on some of the most well-known data sets are presented. The proposed algorithm surpassed the K mean algorithm.*

Keywords: *clustering, optimization, pso, k-means*

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.[6]

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. Centroid jednoznačno određuje jedan klaster. Pripadnost klasteru jedne instance se određuje na osnovu najbližeg centroida. Centroida ima koliko i klastera.

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 poput f1 mere i odziva (eng. *recall*) [9].

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 [4][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:

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

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 [4] je zadata formulom:

$$\frac{1}{c} \sum_{i=1}^{c} \max_{i \neq j} \left(\frac{\sigma_i + \sigma_j}{d(C_i, C_j)} \right), \tag{2}$$

gde je c broj klastera, σ prosečno rastojanje svih instanci unutar 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 [4].

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 je 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, samim tim i beskonačno klastera jer je svaka instanca poseban klaster, 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 [3] 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 formulom:

$$s = \frac{tr(B_k)}{tr(W_k)} \times \frac{n_E - k}{k-1}, \tag{3}$$

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

Matrice se izračunavaju po formulama:

$$B_k = \sum_{q=1}^{k} n_q (c_q - c_E)(c_q - c_E)^T, \tag{4}$$

$$W_k = \sum_{q=1}^{k} \sum_{x \in C_q} (x - c_q)(x - c_q)^T, \tag{5}$$

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.

Što je ovaj indeks veći, to je klasterovanje bolje, centroidi su udaljeniji, a rastojanja unutar klastera manja. Zbog načina računanja disperzije klastera, koje se računa 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 [2]. Pseudokod algoritma je prikazan na slici 1.

Osnovnu PSO algoritma čini jedna čestica roja. Čestica predstavlja 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

Predloženi algoritam je testiran na poznatim skupovima podataka: IRIS i WINE. U eksperimentima su upoređeni rezultati algoritma K sredina sa PSO algoritmom. Vrednosti prikazane u tabelama 1: i 2: predstavljaju vrednosti funkcija evaluacije klasterovanja, koje su detaljno opisane u 2.1.1. i 2.1.2.
Za optimizaciju obe funkcije evaluacije korišćenje su vrednosti kognitivne i sociološke komponente, 1 i 2 redom. Taj izbor govori da na jednu česticu više utiče ceo roj, tj. najbolja čestica roja, nego najbolja pozicija trenutne čestice. Veličina roja je 20 čestica i broj iteracija je 500 što je bio i kriterijum zaustavljanja.

3.1. Iris

Iris [5] je jedan od najpoznatijih skupova podataka. Sastoji se od 4 numerička atributa:
- dužina krunice,
- širina krunice,
- dužina čašice,
- širina čašice.

Primarno je namenjen za testiranje algoritama koji rešavaju problem klasifikacije jer sadrži i peti, kategorički atribut, koji predstavlja vrstu cveta iris. Skup podataka sadrži 3 klase od po 50 instanci. Za potrebe ovog rada iskorišćena su gore navedena četiri atributa bez informacije o pripadajućoj klasi.
Zanimljivo je primetiti da je za Davies-Bouldin index 2.1.1. broj klastera 2. U sekciji 2.2. je naznačeno da se razvijenom algoritmu prosledjuje maksimalan broj klastera i da se taj broj može smanjiti tokom izvršavanja algoritma. Prilikom testiranja algoritma nad ovim skupom podataka za različite parametre kognitivne i sociološke komponente, broja čestica, broja iteracija itd. primećeno je da algoritam kao izlaz da 2 klastera. Analizom skupa podataka je utvrđeno da je jedna klasa linearno razdvojiva od druge dve, što je i navedeno u opisu skupa podataka [7]. Tako da se može opravdano pretpostaviti da je ovo jedno od validnih klasterovanja.
Očekivano je PSO algoritam nadmašio algoritam K sredina, jer je direktno optimizovao funkcije koje su prikazane u tabeli 1:.

Algoritam	IRIS			
	DB		CZ	
	c	index	c	index
K Sredina	2	0.40	3	561.62
PSO	2	**0.28**	3	**601.05**

Tabela 1: Vrednosti funkcija evaluacije nad skupom podataka IRIS.

Algoritam	WINE			
	DB		CZ	
	c	index	c	index
K Sredina	3	0.53	3	561.81
PSO	3	**0.43**	3	**585.38**

Tabela 2: Vrednosti funkcija evaluacije nad skupom podataka WINE.

3.2. Wine

Skup podataka Wine [8] je takođe jedan od poznatih skupova. Sastoji se od 13 numeričkih atributa i ciljne klase. Atributi predstavljaju vrednosti različitih hemijskih supstanci do kojih se došlo hemijskom analizom 3 vrste vina koja potiču iz Italije. Skup se sastoji od 178 instanci. Kao i u prethodnom skupu podataka, ni u ovom nije korišćena informacija o klasi.
Algoritam je kao parametar prosleđen broj 5 kao maksimalan broj klastera. Prilikom optimizacije obe funkcije, skoro svaki put je algoritam smanjio broj klastera na 3, što itekako ima smisla s obzirom da u skupu podataka zaista postoje 3 klase i za taj broj klastera su funkcije bila minimalne. To takođe znači da su izabrane funkcije pogodne za rešavanje problema klasterovanja nad ovim skupom.
Skup podataka Wine se smatra za jedan od lakših primera problema klasifikacije, ali treba imati u vidu da je ovde reč o klasterovanju i da algoritam prilikom izvršavanju ni na koji način nije imao informaciju od broju klasa.
U tabeli 2: su prikazani rezultati izvršavanja PSO algoritma i algoritma K sredina. Očekivano, PSO je i u ovom slučaju nadmašio K sredina.

4. Zaključak

U ovom radu predložen je PSO algoritam za rešavanje problema klasterovanja i predstavljeni su rezulati nad skupovima podataka Iris [5] i Wine [8]. Predloženi su parametri razvijenog algoritma zasnovanog na optimizaciji rojem čestica za koje su dobijeni najbolji rezultati nad oba skupa podataka. Razvijeni algoritam je generički u smislu izbora funkcije koja je optimizuje i funkcije blizine dve instance, što omogućava da algoritam kao izlaz ima različite oblike i vrste klastera. Takođe, algoritam se može koristiti za određivanje broja klastera. Buduća istraživanja bi bila usmerena ka hibridizaciji algoritma sa drugim poznatim heurističkim algoritmima globalne optimizacije.

LITERATURA

[**1**] Stuart P. Lloyd (1982). Least squares quantization in pcm. IEEE Transactions on Information Theory, 129-137 vol 28.

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

[**3**] Harabasz, Calinski T and Karoński, M. (1974). Communications in Statistics - A dendrite method for cluster analysis, 1-27 vol. 3.

[**4**] David L. Davies and D. Bouldin (1979). A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, 224-227 vol. PAMI-1.

[**5**] Edgar Anderson (1936). The Species Problem in Iris. Annals of the Missouri Botanical Garden, 457-509, vol. 23.

[**6**] Nikolić M. (2019). Mašinsko učenje. http://ml.matf.bg.ac.rs/readings/ml.pdf.

[**7**] UCI Machine Learning repository: Iris data. https://archive.ics.uci.edu/ml/datasets/iris, accessed: 2021-06-14.

[**8**] UCI Machine Learning repository: Wine data. https://archive.ics.uci.edu/ml/datasets/wine, accessed: 2021-06-14.
[**9**] Jones, K Sparck and Van Rijsbergen, Cornelis Joost (1976). Information retrieval test collections. Journal of documentation, MCB UP Ltd.