Minimal sum k clustering

Ilic Mihajlo

Klasterovanje

Klasterovanje predstavlja problem razbijanja skupa elemenata na podskupove tako da elementi jednog podskupa medjusobno zadovoljavaju dati kriterijum bolje nego sa elementima iz drugih podskupova.

Formalno klasterovanje je kategorizacija elemenata u podskupove sa ciljom maksimizacije homogenosti elemenata u istom podskupu I heterogenosti elemenata iz razlicitih podskupova

Problem k minimalne sume

Problem klasterovanja k minimalne sume se odnosi na klasterovanje sa zadatkom minimizacije ukupne sume udaljenosti izmedju elemenata istig podskupova sa ciljem pronalazenja tacno k podskupova.

Tj. Minimizuje se suma koja se racuna za svaki klaster kao zbir udaljenosti izmedju elemenata koji cine taj klaster.

Opis resenja

- Algoritmi koriste pomocne strukture koje im pomazu pri radu.
- Svi koriste matricu unapred izracunatih rastojanja izmedju svake dve instance kako ne bi vise puta morali to da racunaju.
- Takodje rade u mestu sa strukturama koje su napravili tj. Jedanput samo alociraju potreban prostor I uvek rade nad njim kako ne bi gubili vreme na sistemske pozive alokacije memorije.
- Na kraju se prikazuju prva dva atributa iz skupa u obliku rasprsenog dijagrama gde je bojom oznacen klaster koji mu je algoritam dodelio.
- Generalna forma u svakom algoritmu u kojoj se predstavljaju klasteri je kao niz celih brojeva gde broj odredjuje kom klasteru je dodeljena instanca na tom indeksu

111221113312 – moguce resenje sa 3 klastera

Algoritam grube sile

- Algoritam grube sile je implementiran kao algoritam generisanja varijacija I proveravanja da li je izgenerisana varijacija bolja od najbolje
- Izgenerisu se sve varijacije duzine n klase k , gde je n broj instanci u skupu a k broj trazenih klastera.
- Za svaku izgenerisanu varijaciju se racuna njena suma distanci unutar klastera
- Proverava se da li je bolja od trenutno optimalne I menja se sa njom ako jeste

Pohlepni algoritam

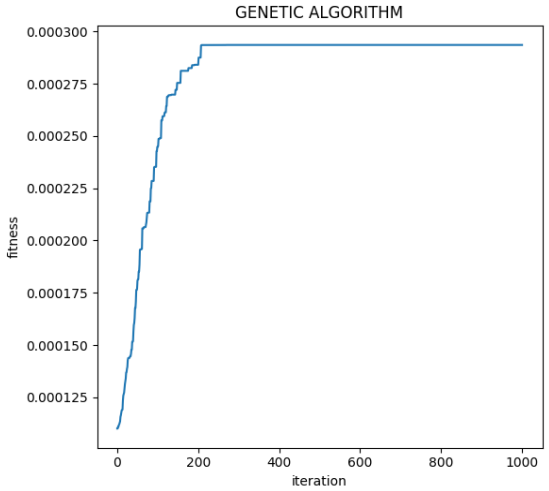
- Ne vrsi generaciju celog prostora resenja vec krece od jednog resenja I pokusava da ga popravi cime zapravo generise podskup prostora resenja.
- Popravku resenja radi tako sto za svaku instancu pokusava da je premesti u drugi skup I ako je ukupna suma tako manja uzima to resenje kao tekuce.
- Koristi pomocnu strukturu u obliku niza gde pamti trenutne sume u klasterima. Zbog toga kad vrsi popravku posto se menjaju dva klastera moze da izracuna celu sumu relativno brzo nego da mora sve ispocetka.

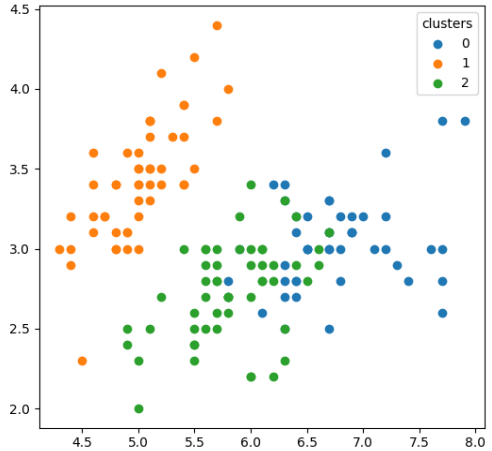
Genetski algoritam

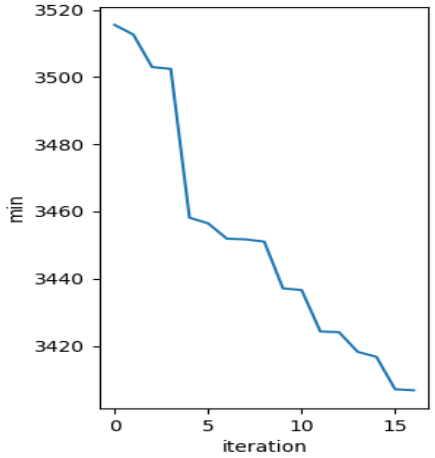
- Predstavlja implementaciju standardnog genetskog algoritma koji je prilagodjen za rad sa klasterima.
- Genotip jedinke je predstavljen kao niz celih brojeva (mana je memorijska konzumcija)
- Fitnes funkcija je predstavljena kao 1 / suma , zato sto genetski algoritam maksimizuje fitnes funkcije sto u nasem kontekstu znaci da maksimizacijom nje mi minimizujemo sumu
- Operator ukrstanja je uradjen kao standardni jednopozicioni operator ukrstanja
- Operator mutacije je uradjen tako sto se odabere nasumican element I samo dodeli drugom klasteru

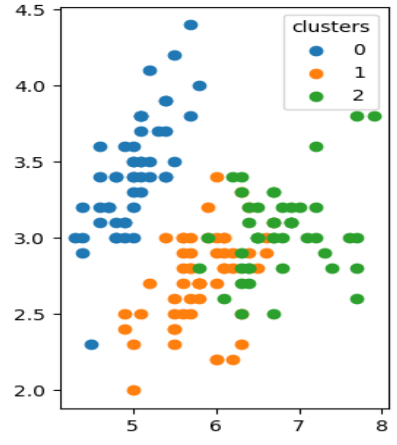
Rezultati eksperimenta nad iris skupom

algoritam	suma	vreme
Gruba sila	/	/
pohlepni	3406	3.8
genetski	3406	93.8



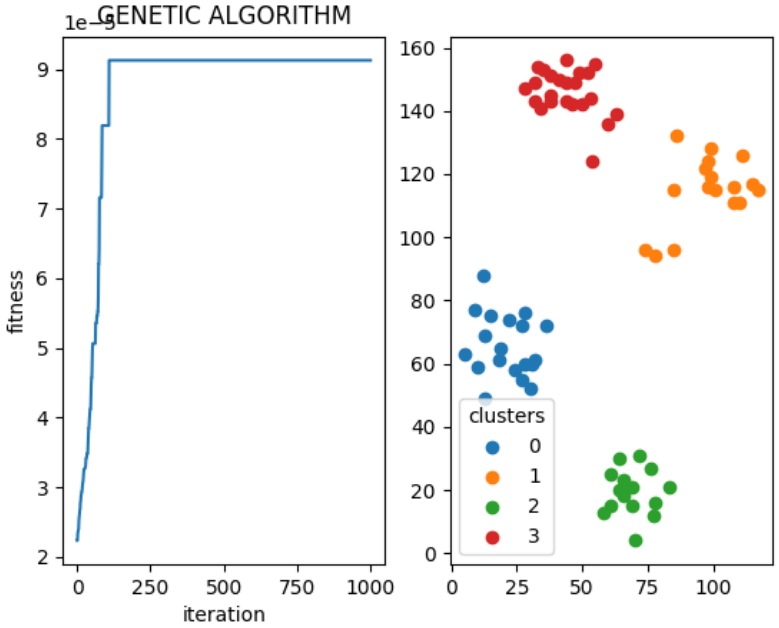


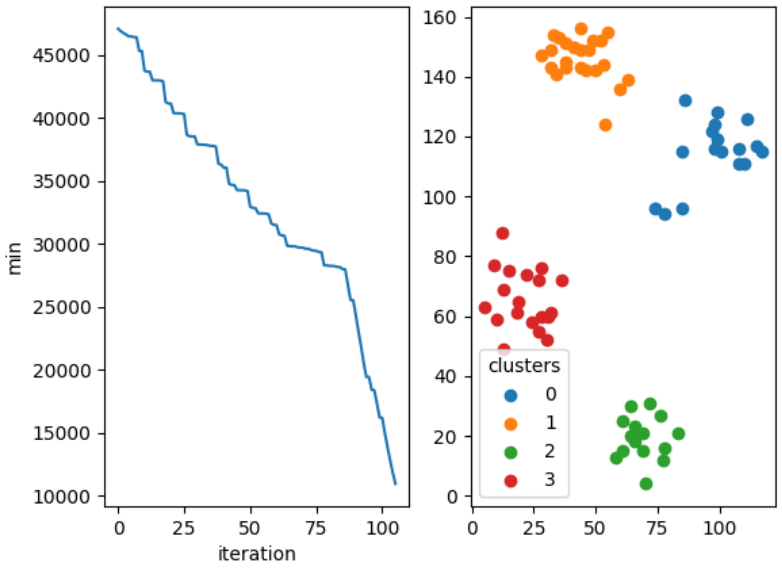




Rezultati eksperimenta nad skupom ruspini

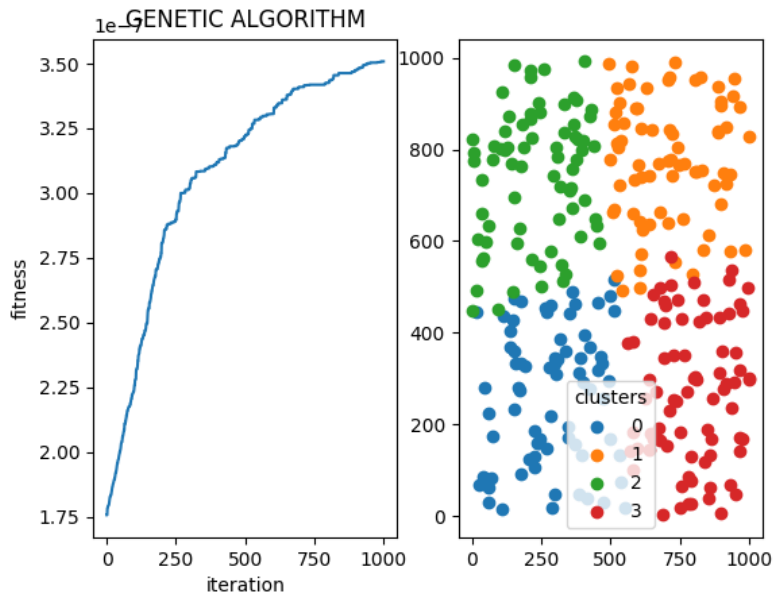
algoritam	suma	vreme
Gruba sila	/	/
pohlepni	10956	1.14
genetski	10956	20.59

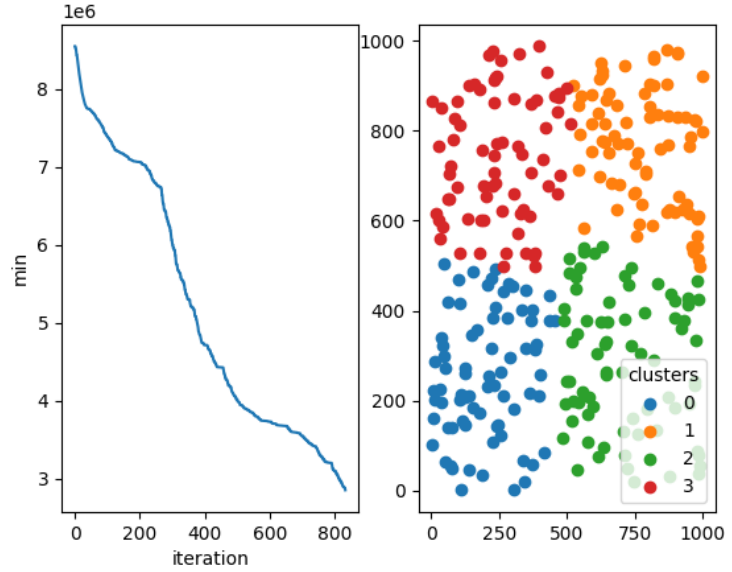




Rezultati eksperimenta nad skupom random

algoritam	suma	vreme
Gruba sila	/	/
pohlepni	2851536	562
genetski	2850149	290





Zakljucak

Prednost algoritma grube sile je sto garantuje optimalno resenje medjutim vremenska slozenost mu je lose.

Pohlepni algoritam daje veoma dobre rezultate za male ulaze I radi brze od genetskog.

Za velike I kompleksne ulaze genetski algoritam brze radi (mada I on je generalno spor) I daje pristojne rezultate.

Jedna od prednosti genetskog algoritma je takodje sto teorijski sa operatorima ukrstanja I mutacije pretrazuje veci prostor resenja.

Kraj