Istraživanje podataka

Vežbe 11

7. maj 2021

Outline

Outline

Klasterovanje

Primenom klasterovanja nad skupom podataka vrši se grupisanje instanci sa ciljem da instance jedne grupe budu što sličnije, a što udaljenije od instanci iz drugih grupa. Jedna grupa instanci dobijena klasterovanjem naziva se klaster.

Algoritam: K-sredina

Pronalazak klastera u algoritmu K-sredina je iterativni proces računanja centroida za svaki klaster i dodeljivanja instance klasteru.

Algoritam

1. Određivanje inicijalnih centroida za k klastera.
2. Svaku instancu dodeliti najbližem klasteru korišćenjem mere bliskosti.
3. Za svaki klaster ažurirati centroid na osnovu dodeljenih instanci tom klasteru.
4. Ponavljati korake 2 i 3 dok se ne ispuni uslov: nijedan centroid se nije promenio u odnosu na prethodnu iteraciju.

Algoritam: K-sredina

Za algoritam K-sredina potrebno je definisati

- parametar **K** - broj željenih klastera
- meru bliskosti (mera sličnosti ili različitosti) koja se koristi za računanje bliskosti između instance skupa i centroida klastera.
- inicijalne centroide.

Zadatak 1

Algoritmom K-sredina identifikovati 3 klastera u sledećim podacima. Pri tom, koristiti euklidsko rastojanje. Za polazne centroide uzeti prve tri instance.

X	Y	Z
1	0	2
2	0	0
-3	-1	1
-4	-2	2
0	4	9
1	5	9

Zadatak 1

Značenje oznaka koje se koriste u rešenju:
c_i - centroid klastera i
C_i - instance u klasteru i

Zadatak 1

Iteracija I

U tabeli je prikazana matrica rastojanja između instanci i inicijalnih centroida. Za svaku instancu je podebljano rastojanje do najbližeg centroida i njegovom klasteru se instanca dodeljuje.

centroid	i_1	i_2	i_3	i_4	i_5	i_6
c_1	0			5,4	**8,1**	**8,6**
c_2		0		6,6	10	10,3
c_3			0	**1,7**	9,9	10,8

Tabela: Matrica rastojanja između instanci i centroida za iteraciju 1

Zadatak 1

Nakon I iteracije podela instanci po klasterima je:

- $C_1 : i_1, i_5, i_6$
- $C_2 : i_2$
- $C_3 : i_3, i_4$

Novi centroidi su:

- $c_1 = \frac{i_1+i_5+i_6}{3} = (0,67; 3; 6,67)$
- $c_2 = i_2 = (2; 0; 0)$
- $c_3 = \frac{i_3+i_4}{2} = (-3,5; -1,5; 1,5)$

Zadatak 1

Iteracija II

Tabela: Matrica rastojanja između instanci i centroida za iteraciju II

centroid	i_1	i_2	i_3	i_4	i_5	i_6
c_1	5,6	7,4	7,8	8,3	**2,6**	**3,1**
c_2	**2,2**	**0**	5,2	6,6	10	10,3
c_3	4,8	5,9	**0,9**	**0,9**	9,9	10,9

Zadatak 1

Nakon II iteracije podela instanci po klasterima je:

- $C_1 : i_5, i_6$
- $C_2 : i_1, i_2$
- $C_3 : i_3, i_4$

Novi centroidi su:

- $c_1 = \frac{i_5 + i_6}{2} = (0,5; 4,5; 9)$
- $c_2 = \frac{i_1 + i_2}{2} = (1,5; 0; 1)$
- $c_3 = \frac{i_3 + i_4}{2} = (-3,5; -1,5; 1,5)$

Zadatak 1

Iteracija III

Tabela: Matrica rastojanja između instanci i centroida za iteraciju III

centroid	i_1	i_2	i_3	i_4	i_5	i_6
c_1	8,3	10,2	10,3	10,6	**0,7**	**0,7**
c_2	**1,1**	**1,1**	4,6	5,9	9,1	9,4
c_3	4,8	5,9	**0,9**	**0,9**	9,9	10,9

Zadatak 1

Nakon III iteracije podela instanci po klasterima je:

- $C_1 : i_5, i_6$
- $C_2 : i_1, i_2$
- $C_3 : i_3, i_4$

Primetiti da je podala instanci po klasterima ista u II i III iteraciji, zbog čega neće doći do promene u vrednostima centroidima, i time je klasterovanje završeno.

Outline

Suma kvadrata greške (*SSE* - sum of the squared error)

Kada se kao mera bliskosti koristi rastojanje u Euklidskom prostoru, za evaluaciju klasterovanja algortimom K-sredina često se koristi mera suma kvadrata greške (*SSE*)

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist(c_i, x)^2$$

gde je

- x instanca skupa
- C_i klaster
- c_i centroid klastera C_i

Cilj je da SSE bude što manja.

Silueta koeficijent, eng. Silhouette coefficient

Silueta koeficijent je mera koliko su instance grupisane sa instancama koje su slične njima samima. Prvo se silueta koeficijent računa za svaku instancu po formuli

$$s = \frac{b - a}{\max(a, b)}$$

gde je

- a - prosečno rastojanje između instance i ostalih instanci u istom klasteru
- b - prosečno rastojanje između instance i svih instanci iz najbližeg susednog klastera

Silueta koeficijent, eng. Silhouette coefficient

Silueta koeficijent za ceo skup je prosečna vrednost koeficijenata za pojedinačne instance. Vrednost silueta koeficijenta je između [-1,1] pri čemu je

- -1 za neispravno grupisanje
- +1 za gusto grupisanje

Vrednost koeficijenta je veća kada su klasteri gusti i dobro razdvojeni.

Outline

K-sredina u biblioteci scikit-learn

- *sklearn.cluster.KMeans*
- parametri
  - *n_clusters* - broj klastera, default=8
  - *init* - metod za inicijalizaciju centroida, (*k-means++*, *random*)
    - *k-means++* - inicijalni centroidi se biraju tako da budu generalno udaljeni jedan od drugog
    - *random* - inicijalni centroidi se nasumično biraju

K-sredina u biblioteci scikit-learn

- *sklearn.cluster.KMeans*
- parametri
  - *n_init* - koliko puta će algoritam K-sredina biti izvršen sa različitim inicijalnim centoridima
  - *max_iter* - maksimalan broj iteracija pri klasterovanju
  - *tol* - tolerancija za sumu kvadrata greške

K-sredina u biblioteci scikit-learn

- atributi
  - *cluster_centers_* - kooridinate centroida
  - *labels_* - oznake klastera kojima su instance dodeljene
  - *inertia_* - suma kvadrata rastojanja instanci do najbližeg centroida
  - *n_iter_* - broj izvršenih iteracija
- metode
  - *fit* - izvršavanje k-sredina klasterovanja
  - *fit_predict* - izvršavanje k-sredina klasterovanja i dodela oznake klastera svakoj instanci
  - *predict* - dodela oznake klastera svakoj instanci

Zadatak

Dat je skup *dogs* koji ima atribute:

- *breed* - rasa psa
- *height* - visina psa
- *weight* - težina psa

Primenom algoritma K-sredina izvršiti klasterovanje za 2, 3 i 4 klastera na osnovu visine i težine pasa.

Outline

K-sredina u alatu IBM SPSS Modeler

Čvor za algoritam K-sredina je **K-mean**. U okviru čvora **K-mean** prvo se vrši priprema atributa skupa za algoritam kako bi svaki atribut imao isti uticaj pri računanju euklidskog rastojanja.

Transformacija atributa

Numerički atributi

Sklariranje vrednosti u opseg $[0, 1]$ formulom

$$x_i' = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

Kategorički atributi

Za svaku kategoriju u kategoričkom atributu se pravi binarni atribut. U binarnom atributu za svaku kategoriju se instanciama te kategorije dodeljuje vrednost $\sqrt{\frac{1}{2}}$, a ostalim instancama 0.

Transformacija atributa

Npr. neka poredimo dve osobe prema dužini kose i veličini garderobe sa mogućim vrednostima *S, M, L*. Nakon transformacije originalnog skupa atribut

- dužina kose će imati vrednosti u opsegu [0,1]
- veličina odeće biće tri binarna atributa *S*, *M* i *L*.

Transformacija atributa

Poređenje osobe sa najkraćom kosom (dužina je 0) koja nosi veličinu L (O_1) i osobe sa najdužom kosom (dužina je 1) koja nosi veličinu M (O_2) pri:

- kodiranju T/F sa $1/0$

$$dist(O1, O2) = \sqrt{(O1_d - O2_d)^2 + (O1_S - O2_S)^2 + (O1_M - O2_M)^2 + (O1_L - O2_L)^2} = \sqrt{(0-1)^2 + (0-0)^2 + (0-1)^2 + (1-0)^2} = \sqrt{1+0+1+1} = \sqrt{3}$$

Transformacija atributa

- kodiranju T/F sa $\sqrt{\frac{1}{2}}/0$

$$dist(O1, O2) = \sqrt{(O1_d - O2_d)^2 + (O1_S - O2_S)^2 + (O1_M - O2_M)^2 + (O1_L - O2_L)^2} =$$
$$\sqrt{(0-1)^2 + (0-0)^2 + (0-\sqrt{\tfrac{1}{2}})^2 + (\sqrt{\tfrac{1}{2}} - 0)^2} =$$
$$\sqrt{1 + 0 + \tfrac{2}{4} + \tfrac{2}{4}} = \sqrt{2}$$

Algoritam K-sredina u alatu IBM SPSS Modeler

Pronalazak klastera u algoritmu K-sredina je iterativni proces računanja centroida za svaki klaster i dodeljivanja instance klasteru.

Koraci

1. Računaju se inicijalni centroidi za k klastera.
2. Svaka instanca se dodeljuje najbližem klasteru korišćenjem euklidskog rastojanja.
3. Za svaki klaster se ažurira centroid na osnovu dodeljenih instanci tom klasteru.
4. Ponavljaju se koraci 2 i 3 dok se ne ispuni jedan od uslova:
   - Nijedan centroid se nije promenio u odnosu na prethodnu iteraciju.
   - Izvršen je maksimalan broj iteracija.

Inicijalni centroidi

Primena maxmin algoritma

1. Prva instanca u skupu se postavlja za centroid prvog klastera.
2. Za svaku instancu se računa rastojanje do definisanih centroida klastera.
3. Pronalazi se najudaljenija instanca od definisanih centroida i ona se dodaje kao novi centroid.
4. Ponavljaju se koraci 2 i 3 dok se ne definiše k inicijalnih centroida.

Ažuriranje centroida klastera

Za svaki klaster C_j se centroid ažurira na kraju svake iteracije po formuli:

$$c_{qj} = \frac{\sum_{i=1}^{n_j} x_{qi}(j)}{n_j}$$

gde je

- n_j broj instanci u klasteru C_j
- $x_{qi}(j)$ je vrednost q. transformisanog atributa instance i koja je dodeljena klasteru C_j

Parametri u SPSS modeleru

- Broj klastera
- Maksimalan broj iteracija
- Tolerancija greške
  Ukoliko je za svaki klaster C_j u iteraciji i euklidsko rastojanje centroida u iteraciji i i centroida u iteraciji $i-1$ manje od zadate vrednosti tolerancije greške, vraća se dobijeni model.
- Vrednost sa kojom se kodira T u transformisanim binarnim atributima za kategoričke atribute.

Zadatak

Izvršiti klasterovanje predmeta primenom algoritma K-sredina u alatu IBM SPSS Modeler. Skup *podaci_o_predmetima.csv* ima atribute:

- *predmet* - naziv predmeta
- *upisalo* - broj studenata koji su upisali predmet
- *polozilo* - broj studenata koji su položili ispit iz predmeta
- *prosek* - prosečna ocena na položenim ispitima iz predmeta. Za predmete koje nijedan student nije položio, prosek je 5.

Outline

Algoritam hijerarjijsko sakupljajuće klasterovanje

Algoritam

Svaka instanca je zaseban klaster. Računa se matrica bliskosti klastera (tj. matrica bliskosti instanci).

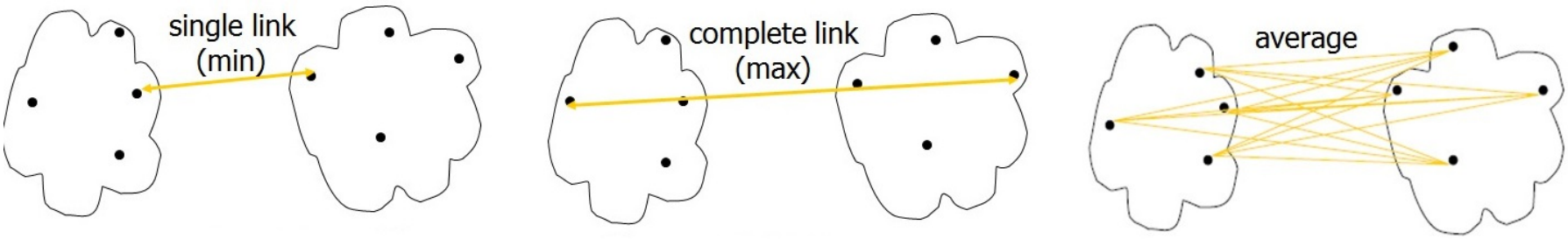
1. Spojiti dva najbliža klastera.
2. Ažurirati matricu bliskosti klastera.
3. Ponavljati korake 1 i 2 dok ne ostane jedan klaster.

Algoritam hijerarjijsko sakupljajuće klasterovanje

Kriterijumi pri određivanju blizine klastera:

- Najbolja (min, single) veza - bliskost dva klastera je jednaka bliskosti najbližeg para instanci iz različitih klastera
- Najgora (max, complete) veza - bliskost dva klastera je jednaka bliskosti najudaljenijeg para instanci iz različitih klastera
- Prosečna (avg) veza - bliskost dva klastera je jednaka prosečnoj bliskosti parova instanci iz različitih klastera

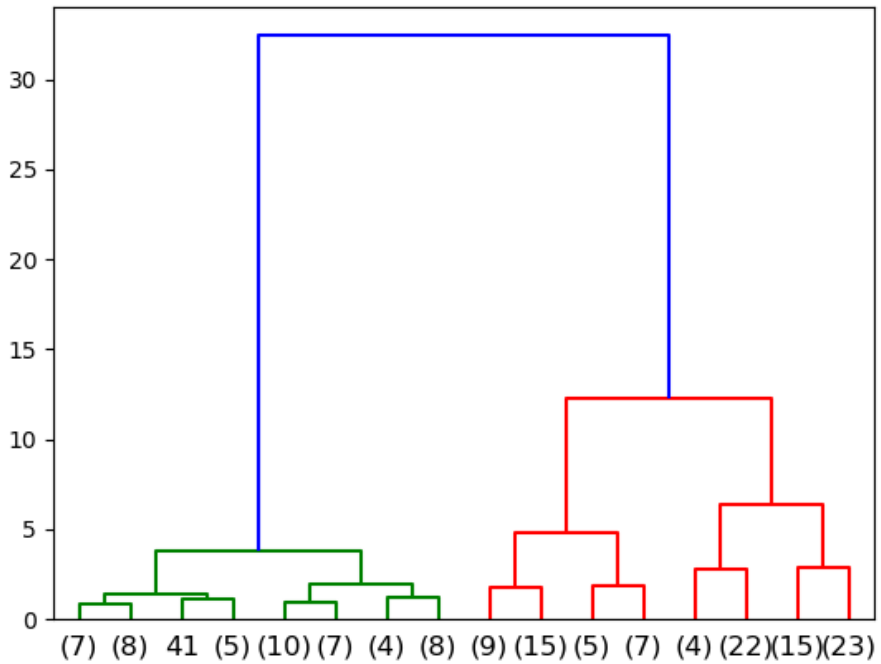
Algoritam hijerarjijsko sakupljajuće klasterovanje



Slika: Prikaz veza koje se mogu birati kao kriterijum za određivanje bliskosti dva klastera

Algoritam hijerarjijsko sakupljajuće klasterovanje

Rezultat hijerarhijskog klasterovanja se obično prikazuje pomoću dendograma ili dijagrama sa ugnježdenim klasterima.



Slika: Na x-osi su prikazane oznake instanci, a na y-osi udaljensot klastera koji se spajaju.

Zadatak

Data je matrica sličnosti skupa podataka. Izvršiti hijerarhijsko klasterovanje korišćenjem min i max veze. Rezulatat prikazati dendogramom.

	p1	p2	p3	p4	p5
p1	1.00	0.10	0.41	0.55	0.35
p2	0.10	1.00	0.64	0.47	0.98
p3	0.41	0.64	1.00	0.44	0.85
p4	0.55	0.47	0.44	1.00	0.76
p5	0.35	0.98	0.85	0.76	1.00