Istraživanje podataka 1 - vežbe 8, 2020.

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.

Jedan od najpoznatijih algoritama klasterovanja je algoritam K-sredina.

1 Algoritam: K-sredina

Algoritam K-sredina pripada grupi algoritama klasterovanja koji su zasnovani na reprezentativnim predstavnicima. Svaki klaster se opisuje preko svog predstavnika - centroida.
Pronalazak klastera u algoritmu K-sredina je iterativni proces računanja centroida za svaki klaster i dodele svake instance klasteru čijem centroidu je najsličnija (najbliža). U jednoj iteraciji se prvo svaka instanca dodeljuju klasteru sa čijim centroidom je najsličnija, a zatim se za svaki klaster centroid ažurira na osnovu instanci koje su mu dodeljene u toj iteraciji.
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. Ako se koristi **mera sličnosti**, instanca se dodeljuje klasteru sa čijim centroidom ima najveću sličnost, odnosno vrednost mere bliskosti je **najveća**, a ako se koristi **mera različitosti**, instanca se dodeljuje klasteru sa čijim centroidom ima najmanju različitost, odnosno vrednost mere bliskosti je **najmanja**.
- inicijalne centroide. Mogu biti npr. nasumično izabranih k instanci iz skupa. Različitim izborom inicijalnih centroida mogu se dobiti različiti rezultati klasterovanja.

Algoritam 1 K-sredina

1: Odrediti inicijalne centroide+ 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 dalje najbolje rezultate za globularne podatke, a loše za klastere proizvoljnog oblika (neglobularnog), različitih gustina ili veličina.

1.1 Zadatak 1

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

instanca	**X**	**Y**	**Z**
i_1	1	0	2
i_2	2	0	0
i_3	-3	-1	1
i_4	-4	-2	2
i_5	0	4	9
i_6	1	5	9

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

Iteracija I

Kako su prve tri instance izabrane za centoride klastera, za njih nema potrebe računati rastojanje do centoida jer je $dist(c_i, c_i) = 0$, tako da će biti izračunata samo rastojanja za instance i_4, i_5 i i_6 do izabranih centroida.

U tabeli 1 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 1: Matrica rastojanja između instanci i centroida za iteraciju 1

Nakon I iteracije podela instanci po klasterima je:

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

Za svaki klaster se ažuriraju centroidi na osnovu dodeljenih instanci, tako što se za svaki atribut računa srednja vrednost koju imaju instance klastera.
Novi centroidi su:

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

Iteracija II

U tabeli 2 je prikazana matrica rastojanja između instanci i ažuriranih centroida. Za svaku instancu je podebljano rastojanje do najbližeg centroida i njegovom klasteru se instanca dodeljuje u ovoj iteraciji.

Tabela 2: 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

Nakon II iteracije podela instanci po klasterima je:

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

Za svaki klaster se ažuriraju centroidi na osnovu dodeljenih instanci, tako što se za svaki atribut računa srednja vrednost koju imaju instance klastera.
Novi centroidi su:

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

Iteracija III

U tabeli 3 je prikazana matrica rastojanja između instanci i ažuriranih centroida. Za svaku instancu je podebljano rastojanje do najbližeg centroida i njegovom klasteru se instanca dodeljuje u ovoj iteraciji.

Tabela 3: 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

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.

2 Kvalitet klasterovanja

2.1 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.

2.2 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 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.

3 K-sredina u biblioteci scikit-learn

Algoritam K-sredina je implementiran klasom *sklearn.cluster.KMeans*. Mera bliskosti je Euklidsko rastojanje. Klasa ima

- parametre
  - *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
  - *n_init* - koliko puta će algoritam K-sredina biti izvršen sa različitim inicijalnim centoridima. Bira se klasterovanje sa najmanjom sumom kvadrata rastojanja instanci do najbližeg centroida (vrednost atributa *inertia_*).
  - *max_iter* - maksimalan broj iteracija pri klasterovanju. Jedan od uslova za ranije zaustavljanje klasterovanja.
  - *tol* - tolerancija za promenu sume kvadrata greške. Ako je promena u dve uzastopne iteracije manja od tolerencije, klasterovanje se zaustavlja. Jedan od uslova za ranije zaustavljanje klasterovanja.

- 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

Primer 1: 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.
Rešenje: **k_means.py**

4 K-sredina u alatu IBM SPSS Modeler

4.1 Transformacija atributa

Čvor za algoritam K-sredina u alatu IBM SPSS Modeler se zove **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:

- Numerički atributi
  Sklariraju se vrednosti svakog numeričkog atributa na 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.

  Ako se vrednost T kodira sa 1, a F sa 0 (što je najčešći slučaj) tada kategorički atributi iz originalnog skupa imaju veći uticaj pri računanju euklidskog rastojanje između dve instance od numeričkih atributa koji nakon transformacije imaju vrednosti u opsegu [0,1]. 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], a umesto atributa veličina odeće biće tri binarna atributa *S*, *M* i *L*. Ako se u binarnim atributima za kodiranje T/F koriste vrednosti $1/0$ pri poređenju osobe sa najkraćom kosom (dužina je 0) koja nosi veličinu L (dalje u tekstu je označena sa $O1$) i osobe sa najdužom kosom (dužina je 1) koja nosi veličinu M (dalje u tekstu je označena sa $O2$), rastojanje je
  $dist(O1, O2) = \sqrt{(O1_{duzina} - O2_{duzina})^2 + (O1_S - O2_S)^2 + (O1_M - O2_M)^2 + (O1_L - O2_L)^2} = \sqrt{(0-1)^2 + (0-0)^2 + (0-1)^2 + (1-0)^2} = \sqrt{1+0+1+1} = \sqrt{3}$
  te originalan atribut veličina garderobe ima uticaj kao dva numerička atributa.

  Ako se u binarnim atributima za kodiranje T/F koriste vrednosti $\sqrt{\frac{1}{2}}/0$ rastojanje je
  $dist(O1, O2) = \sqrt{(O1_{duzina} - O2_{duzina})^2 + (O1_S - O2_S)^2 + (O1_M - O2_M)^2 + (O1_L - O2_L)^2} = \sqrt{(0-1)^2 + (0-0)^2 + (0-\sqrt{\frac{1}{2}})^2 + (\sqrt{\frac{1}{2}} - 0)^2} = \sqrt{1+0+\sqrt{2}4+\sqrt{2}4} = \sqrt{2}$
  te originalan atribut veličina garderobe ima isti uticaj kao numerički atribut.

4.2 Algoritam K-sredina

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:
   - Svi centroidi su se pomerili za manje od zadate tolerancije greške.
   - Izvršen je maksimalan broj iteracija.

Za računanje udaljenosti koristi se kvadrat euklidskog rastojanja.

4.2.1 Inicijalni centroidi

Za određivanje inicijalnih centroida primenjuje se maxmin algoritam:

1. Prva instanca u skupu se postavlja za centroid prvog klastera.
2. Za svaku instancu skupa 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.

4.2.2 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

4.3 Parametri algoritma K-sredina

- Parametri čvora K-means u odeljku *Model*
  - *Use partitioned data* - ukoliko je izvršena podela podataka na trening i test skup, za klasterovanje se koristi samo skup za treniranje
  - *Specified number of clusters* - broj željenih klastera
  - *Generate distance field* - čvor sa modelom klasterovanja, koji se dobija kao rezultat klasterovanja, će sadržati i atribut sa euklidskim rastojanjem od instance do najbližeg centroida
  - *Cluster label* - da li oznaka klastera da bude niska (Cluster 1, Cluster 2, ...) ili samo broj
- Parametri čvora K-means u odeljku *Expert*
  - kriterijum za zaustavljanje klasterovanja
    - *Maximum Iterations* - maksimalan broj iteracija
    - *Change tolerance* - 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.

4.4 Rukovanje nedostajućim (blanko) vrednostima

Nedostajuće (blanko) vrednosti se zamenjuju sa neutralnim vrednostima. U numeričkim i binarnim atributima nedostajuća vrednost se zamenjuje sa 0,5. U kategoričkim atributima, pri pojavi nedostajuće vrednosti u tom atributu, vrednosti u svim izvedenim atributi se postavljaju na 0.

4.5 Podaci koje sadrži model klasterovanja

Skupu za klasterovanje se dodaju još dva atributa

- *$KM-imemodela* sa podatkom o klasteru kome instanca pripada
- *$KMD-imemodela* - sa rastojanjem između instance i najbližeg centroida

Pri interpretaciji rezultata klasterovanja

- utvrđuju se karakteristike koje su jedinstvene za klaster
- analizaraju se vrednosti atributa po klasterima

Kada se napravi model, postoji nekoliko načina za grafički prikaz statistika i distribucije vrednosti atributa po klasterima:

- *Model Summary* - za brzu odluku koliko je dobro klasterovanje. Prikazuje broj atributa korišćenih za pravljenje modela, broj dobijenih klastera i silueta koeficijent.
- *Clusters* - prikaz tabele sa podacima o svakom klasteru. Za svaki klaster se izdvaja:
  - *Cluster* - ime klastera
  - *Size* - veličina klastera u broju instanci, kao i procenat instanci koji klaster sadrži u odnosu na ukupan broj instanci u skupu
  - *Inputs* - za svaki atribut koji je korišćen pri pravljenju modela prikazuje se njegova značajnost i srednja vrednost po klasteru
- *Predictor Importance* - prikazuje značajnost atributa pri pravljenju modela u opsegu [0-1]
- *Cluster Sizes* - grafički prikaz veličine klastera
- *Cell Distribution* - za detaljniji prikaz distribucije vrednosti izabranog atributu u izabranom klasteru koji se biraju preko pogleda *Clusters*. Svetlijom bojom su prikazani podaci za ceo skup, a tamnijom za izabrani klaster.
- *Cluster Comparison* - za poređenje izabranih klastera (biraju se klikom) preko pogleda *Clusters*. Za svaki atribut distribucija vrednosti po klasteru se prikazuje pomoću kućica. Svakom klasteru je dodeljena jedinstvena boja za lakše porđenje klastera. Ovim pogledom se najlakše uočava šta je specifično za svaki klaster.

Primer 2: 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.

Rešenje:

- *klasterovanje_podaci_o_predmetima_k_sredina.str* - radni tok
- *ipVezbe8Primer2.pdf* - komentari