Istraživanje podataka

Vežbe 12

13. maj 2021

Outline

Outline

Algoritam hijerarjijsko sakupljajuće klasterovanje

Algoritam

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

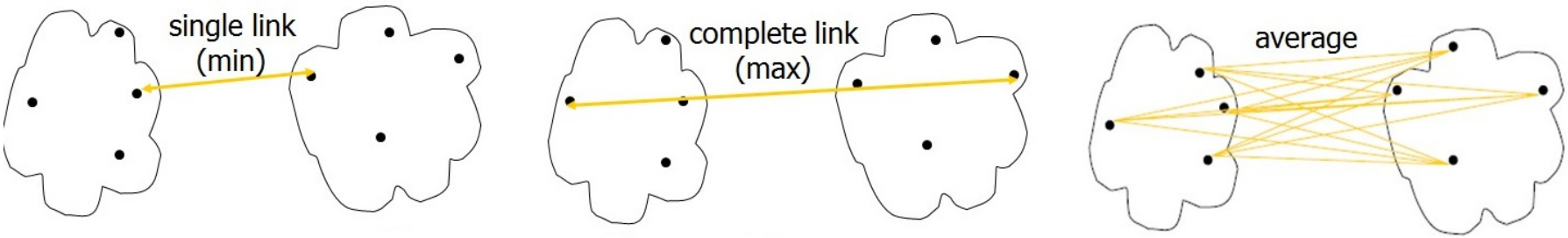
1. Spojiti dva najbliža klastera.
2. Ažurirati matricu bliskosti klastera.
3. Ponavljati korake 1 i 2 dok ne ostane jedan klaster.

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

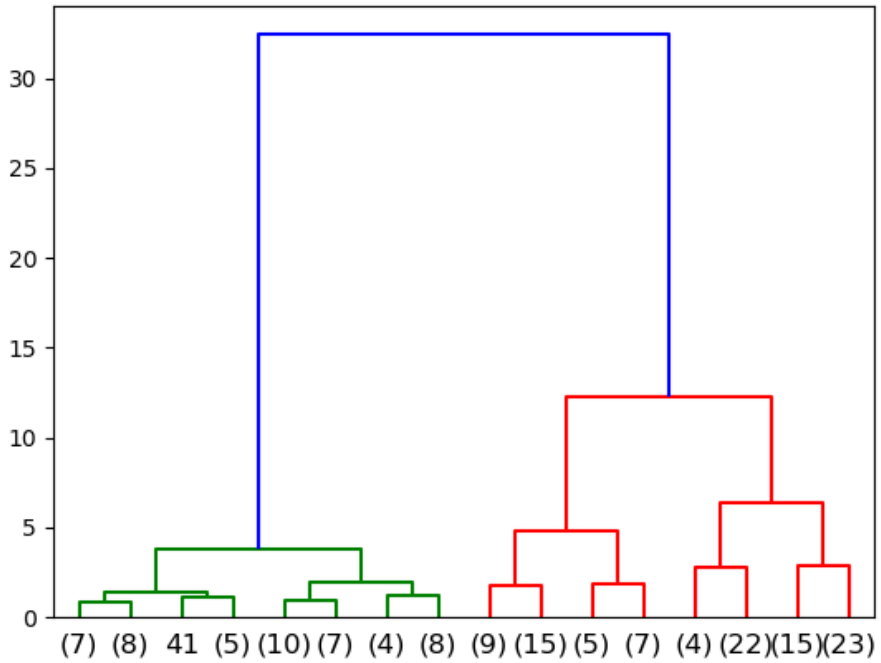
Algoritam hijerarjijsko sakupljajuće klasterovanje



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

Hijerarjijsko sakupljajuće klasterovanje u biblioteci scikit-learn

- *sklearn.cluster.AgglomerativeClustering*
- parametri
  - *n_clusters* - broj klastera, default=8
  - *affinity* - mera bliskosti (*euclidean*, *l1*, *l2*, *manhattan*, *cosine*), default: *euclidean*
  - *linkage* - kriterijum za određivanje blizine klastera (*ward*, *complete*, *average*, *single*), default:*ward*

Hijerarjijsko sakupljajuće klasterovanje u biblioteci scikit-learn

- atributi
  - *labels_* - oznake klastera kojima su instance dodeljene
  - *children_* - matrica koja predstavlja decu unutrašnjih čvorova. U *i*-toj iteraciji, *deca[i][0]* i *deca[i][1]* se spajaju da bi formirali čvor *broj_instanci + i*.

Hijerarjijsko sakupljajuće klasterovanje u biblioteci scikit-learn

- metode
  - *fit* - izvršavanje klasterovanja
  - *fit_predict* - izvršavanje klasterovanja i dodela oznake klastera svakoj instanci

Hijerarjijsko sakupljajuće klasterovanje u biblioteci scipy

- *scipy.cluster.hierarchy*
- funkcije
  - *scipy.cluster.hierarchy.linkage* - izvršava klasterovanje
    - *y* - skup podataka ili matrica rastojanja
    - *method* - kriterijum za određivanje blizine klastera (*complete*, *average*, *single*, *ward*, *centroid*), default:*single*
    - *metric* - mera različitosti (*euclidean*, *cityblock*, *cosine*), ...) default: *euclidean*
    - vraća matricu spajanja Z (u *i.* iteraciji dobija se *n+i.* klaster spajanjem klastera Z[i,0] i Z[i,1] čije je rastojanje Z[i,2], a nakom spajanja sadrži Z[i,3] instanci).

Hijerarjijsko sakupljajuće klasterovanje u biblioteci scipy

- funkcije
  - *scipy.cluster.hierarchy.dendrogram* - predstavlja rezultat hijerarhijskog klasterovanja pomoću dendograma
    - *Z* - matrica spajanja
    - *color_threshold* - sva spajanja koja imaju rastojanje iznad zadatog praga se boje plavom bojom. (default: 0.7*max(Z[:,2]))
    - *labels* - oznake instanci
    - *leaf_font_size* - veličina slova za ispis oznaka

Hijerarjijsko sakupljajuće klasterovanje u biblioteci scipy

- funkcije
  - *scipy.cluster.hierarchy.fcluster* - dodeljuje id klastera svakoj instanci
    - Z - matrica spajanja
    - t - prag za određivanje klastera
    - *criterion* - kriterijum za određivanje klastera. Koristimo samo *distance* - klasteri čije je rastojanje iznad zadatog praga t neće biti spojeni.

Outline

Algoritam DBSCAN

- Density-based spatial clustering of applications with noise
- Algoritam zasnovan na gustini
- Algoritam DBSCAN može pronaći klastere proizvoljnog oblika

Algoritam DBSCAN

Parametri

- *Eps* - prag za rastojanje suseda. Dve instance su susedne ako im je rastojanje manje ili jednako *Eps*
- *MinPts* - prag za broj suseda instanci

Algoritam

Podela instanci

- Instance u jezgru klastera - instanca je u jezgru klastera ako je broj suseda na rastojanju *Eps* bar *MinPts* .
- Instance na granici klastera - instanca nije u jezgru, ali je na rastojanju do *Eps* nekoj instanci koja je u jezgru klastera.
- Šum - instanca koja nije ni u jezgru ni na granici klastera.

Algoritam

Koraci

1. Za svaku instancu odrediti tip: u jezgru, na granici ili šum.
2. Eliminisati instance koje su šum.
3. Povezati sve instance u jezgru koje su na međusobnom rastojanju do *Eps*.
4. Napraviti poseban klaster za svaku grupu instanci u jezgru koje su povezane.
5. Svaku instancu na granici dodeliti klasteru kojem pripada instanca u jezgru u čijem je susedstvu ta instanca na granici.

DBSCAN u biblioteci scikit-learn

- *sklearn.cluster.DBSCAN*
- parametri
  - *eps* - maksimalna udaljenost između dve instance da bi se smatralo da su u susedstvu, default:0,5
  - *min_samples* - neophodan broj instanci u susedstvu da bi instanca imala status *u jezgru*. Ovaj broj ukuljučuje i samu instancu.
  - *metric* - mera rastojanja (*euclidean*, *l1*, *l2*, *manhattan*, *cosine*)

DBSCAN u biblioteci scikit-learn

- atributi
  - *core_sample_indices_* - indeksi instanci u jezgru
  - *labels_* - oznake klastera kojima su instance dodeljene. Šum ima oznaku -1.
- metode
  - *fit* - izvršavanje klasterovanja
  - *fit_predict* - izvršavanje klasterovanja i dodela oznake klastera svakoj instanci