Analiza mononuklearnih krvnih ćelija

Seminarski rad u okviru kursa
Istraživanje podataka 2
Matematički fakultet

Aleksandar Jakovljević
mi15156@alas.matf.bg.ac.rs

7. juli 2019.

Sadržaj

1 Uvod

Sa porastom broja podataka raste i potreba za njihovom analizom. Podatke dobijene iz oblasti medicine je potrebno detaljno analizirati kako bi se otkrile anomalije ili šabloni koji bi ukazali na potencijalne probleme. Ulazne datoteke sadrže podatke dobijene iz perifernih mononuklearnih krvnih ćelija (engl. *Peripheral blood mononuclear cells, PBMCs*). PBMC ćelije uključuju ćelije različitih tipova: limfocite (B ćelije, T ćelije, NK ćelije (engl. *ntural killer cells*), monocite i dendritske ćelije. PBMC ćelije se koriste u istraživanju u različitim oblastima biomedicine, uključujući infektivne bolesti, imunologiju (uključujući i automune poremećaje), malignitet, transplantacionu imunlogiju, razvoj vakcina, i skrining. Mada mogu da imaju različite funkcije, glavna funkcija PBMC ćelija je imuna odbrana organizma. Svaki tip ćelije ima karakteristične obrasce ('mustre') proteina i gena koje ih međusobno razlikuju i mogu da se koriste za podelu prema njihovom tipu. Nazivi ulaznih datoteka su

- 018_Human_tumor_ascites_dendritic_cells_and_macrophages_csv.csv,
- 019_Human_tumor_ascites_dendritic_cells_and_macrophages_csv.csv,

u daljem tekstu kao 018 i 019, respektivno. Potrebno je instalirati programski jezik Pajton (engl. *Python*), kao i biblioteke *cluster*, *minisom*, *numpy*, *pandas*, *sklearn* i *joblib*. Instalacija biblioteka može se izvršiti pomoću *pip* alata izvršavanjem sledeće naredbe u komandnoj liniji

```
pip install libname
```

U radu su primenjeni različiti algoritmi klasterovanja iz biblioteke *cluster* koji implementiraju tehnike K-sredina, hijerarhijsko klasterovanje i BIRCH, kao i *MiniSom* koji implementira klasterovanje korišćenjem samo-organizujuće mape. Sve datoteke (modeli, rezultati i programski kodovi) mogu se pronaći na sledećoj veb lokaciji: https://drive.google.com/open?id=16kLM5aruERhR541YKKx4Nn-OHOFtDL_7.

2 Sadržaj datoteka i pretprocesiranje

Potrebno je imati uvid u sadržaj datoteka i pretprocesirati podatke pre nego što se uđe u detaljniju analizu jer može se poboljšati kvalitet dobijenih rezultata i smanjiti vremensko izvršavanje algoritama.

2.1 Sadržaj dateoteka

Dimenzije ulaznih podataka iz datoteka 018 i 019 su 31221×3196 i 31221×1613, respektivno. Redovi predstavljaju nazive gena, dok kolone predstavljaju redne brojeve ćelija. Vrednost svakog polja unutar matrice je nenegativna. Naziv svakog gena sadrži prefiks *hg38*, što označava da su podaci vezani za verziju 38 humanog genoma. Dakle, u prvoj koloni se nalaze nazivi gena, dok ostatak kolona predstavlja ekspresiju gena nad 3196 PBMC ćelija.

018.csv

	1	2	3	4	5	6	7	...	3190	3191	3192	3193	3194	3195	3196
Name								...							
hg38_A1BG	0	0	0	0	0	0	1	...	0	0	0	0	0	0	0
hg38_A1BG-AS1	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
hg38_A2M	0	0	0	0	1	0	0	...	1	0	1	0	0	0	0
hg38_A2M-AS1	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
hg38_AAAS	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
hg38_AACS	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
hg38_AAED1	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
hg38_AAGAB	0	0	0	1	0	0	0	...	0	0	0	0	0	0	0
hg38_AAK1	0	1	0	0	1	0	0	...	0	0	0	0	0	0	0
hg38_AAMDC	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
hg38_AAMP	0	0	0	0	0	0	1	...	0	0	0	0	0	0	0
hg38_AANAT	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
hg38_AAR2	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
...	..	..	..	..	..	..	..	...	...	...	...	...	...	...	...
hg38_ZSWIM7	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
hg38_ZSWIM8	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
hg38_ZSWIM9	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
hg38_ZUP1	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
hg38_ZW10	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
hg38_ZWILCH	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
hg38_ZWINT	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
hg38_ZXDA	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
hg38_ZXDB	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
hg38_ZXDC	0	0	0	0	0	0	0	...	0	0	1	0	0	0	0
hg38_ZYG11A	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
hg38_ZYG11B	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
hg38_ZYX	0	0	0	0	0	0	1	...	0	0	0	3	0	0	1
hg38_ZZEF1	0	0	1	0	0	0	0	...	0	0	0	0	0	0	0

2.2 Pretprocesiranje

Potrebno je ukloniti nula redove iz datoteka i smanjiti dimenzionalnost datoteka i povećati brzinu izvršavanja algoritama. Nula redovi ne doprinose analizi podataka, ali mogu uticati na kvalitet rezultata, jer bismo pri klasterovanju podataka dobili bar jedan klaster koji sadrži sve nula redove, što bi dovelo do pogrešnog zaključivanja prilikom analiziranja. Pretprocesiranje pokrećemo izvršavanjem komande 1 u komandnoj liniji.

```
python preprocessing.py naziv_datoteke
```

Listing 1: Pokretanje pretprocesiranja

Nakon poziva, pokreće se program 2 koji uklanja nula redove iz datoteka. Osim uklanjanja nula redova, program sadrži funkcije za učitavanje i čuvanje podataka koje će se koristiti u daljem radu sa podacima.

```
    ...
def preprocessing(data):
    print("Preprocessing data...")
    return data.loc[(data != 0).any(axis=1)]
    ...
```

Listing 2: Pretprocesiranje podataka

Nakon pretprocesiranja, broj redova i kolona datoteka 018 i 019 iznosi 14775× 3196 i 15965 × 1613. U narednoj tabeli nalaze informacije o datotekama nakon pretprocesiranja, gde prvi red ogovara datoteci 018, a drugi red datoteci 019.

preprocessing.txt

Br. ćelija	Br. nula r.	%nula r.	Broj ne-nula r.	%ne-nula r.
3196	16446	52.68%	14775	47.32%
1613	15256	48.86%	15965	51.14%

3 Analiza podataka

Klaster analiza predstavlja razdvajanje podataka u grupe, pri čemu su podaci unutar grupe homogeni, a grupe su međusobno heterogene. Klaster analiza predstavlja nenadgledano učenje, odnosno klasteri se prave pomoću informacija dobijenih iz skupa podataka i odnosa unutar skupa podataka.

3.1 Klasterovanje metodom K-sredina

K-sredina algoritam razdvaja skup od N podataka X u K međusobno disjunktnih klastera $C_i, \forall i = \overline{1, K}$, pri čemu je svaki klaster opisan aritmetičkom sredinom podataka c_i. Sredine se nazivaju centroide i one nisu nužno iz skupa podataka, iako se nalaze u istom prostoru kao i podaci. K-sredina algoritam nasumično bira K centroida i nakon svake iteracije se vrednosti centroida ažuriraju na osnovu zadatog kriterijuma. Postupak ažuriranje se ponavlja sve dok podaci ne prestanu da menjaju klaster, odnosno sve dok centroida menja vrednost. Kriterijum se naziva inercija i predstavlja sumu kvadrata grešaka (engl. *sum of the squared errors, SSE*) i računa se korišćenjem formule

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

gde *dist* predstavlja euklidsko rastojanje između dva elementa, a $x \in X$ [1, 4]. Cilj je minimizovati sumu kvadrata grešaka. K-sredina algoritam zahteva da broj centroida bude unapred zadat. Broj centroida smo empirijski odredili minimizovanjem vrednosti inercije. Kvalitet klasterovanja je određen koeficijentom senke (engl. *silhouette score*), koji uzima vrednost od -1 do 1, gde negativna vrednost koeficijenta senke sugeriše loš kvalitet klasterovanja, odnosno da su podaci dodeljeni pogrešnim klasterima, a pozitivna da vrednost koeficijenta senke sugeriše da je klasterovanje odrađeno dobro. Vrednost koeficijenta senke u blizini 0 sugeriše da postoje preklapanja između klastera [4]. U datoteci *018_ kmeans_ ssd.txt* se nalaze vrednosti inercije u zavisnosti od broja klastera. Može se uočiti da nakon 10. iteracije, inercija se ne smanjuje previše, pa se za vrednost broja klastera uzima $k = 10$.

018_kmeans_ssd.txt

```
Sum of squared distances for k = 2: 11975521.6639
Sum of squared distances for k = 3: 8333361.22305
```

```
Sum of squared distances for k = 4: 6834980.05638
Sum of squared distances for k = 5: 6090281.16998
Sum of squared distances for k = 6: 5169364.25522
Sum of squared distances for k = 7: 4815364.36963
Sum of squared distances for k = 8: 4457133.02511
Sum of squared distances for k = 9: 4235142.40087
Sum of squared distances for k = 10: 4068753.06906
Sum of squared distances for k = 11: 3918015.32307
Sum of squared distances for k = 12: 3780990.32973
Sum of squared distances for k = 13: 3649025.76743
Sum of squared distances for k = 14: 3513269.60938
```

Nakon određivanja broja klastera, izvršava se klasterovanje gena i ćelija. Klasterovanje ćelija postižemo transponovanjem matrica ulaznih datoteka 018 i 019, a zatim pozivanjem algoritma nad transponovanim vrednostima. U datoteci *018_ kmeans_ result.txt* nalazi se rezultat klasterovanja gena. Može se uočiti da je najveći broj gena u 5. klasteru, dok su klasteri {1,2,4,9} veličine 1. Koeficijent senke je visok i iznosi 0.8055 što nam govori da je klasterovanje kvalitetno.

018_kmeans_reulst.txt

```
Size of 0 is:   251
Size of 1 is:   1
Size of 2 is:   1
Size of 3 is:   9
Size of 4 is:   1
Size of 5 is:   14412
Size of 6 is:   21
Size of 7 is:   76
Size of 8 is:   2
Size of 9 is:   1

Silhouette score:
0.8054873558696517
```

Klasterovanje ćelija je dalo ravnomernije klastere ali je koeficijent senke znatno lošiji i u blizini je nule što nam govori da postoje preklapanja između klastera. Rezultati se nalaze u datoteci *018_ kmeans_ transp_ result.txt*.

018_kmeans_transp_result.txt

```
Size of 0 is:   203
Size of 1 is:   1156
Size of 2 is:   42
Size of 3 is:   728
Size of 4 is:   217
Size of 5 is:   4
Size of 6 is:   294
Size of 7 is:   460
Size of 8 is:   65
Size of 9 is:   27

Silhouette score:
0.08782665848785386
```

Na isti način smo klasterovali podatke vezane za datoteku 019. Na početku je određen broj klastera k i analizom rezultata dobijenih prilikom računanja inercije za klastere veličine od $[2, 14]$, uzeli smo da je $k = 11$. Rezultati su dati u datoteci *019_ kmeans_ ssd.txt*.

019_kmeans_ssd.txt

```
Sum of squared distances for k = 2: 84278758.8322
Sum of squared distances for k = 3: 36283836.0829
Sum of squared distances for k = 4: 26630877.7065
Sum of squared distances for k = 5: 19975433.3273
Sum of squared distances for k = 6: 17046662.3173
Sum of squared distances for k = 7: 15588169.8226
Sum of squared distances for k = 8: 14554809.2774
Sum of squared distances for k = 9: 13602368.6154
Sum of squared distances for k = 10: 12582626.3815
Sum of squared distances for k = 11: 11927817.6373
Sum of squared distances for k = 12: 11453373.7961
Sum of squared distances for k = 13: 10934101.8256
Sum of squared distances for k = 14: 10513080.9127
```

Naredne dve datoteke *019_ kmeans_ result.txt* i *019_ kmeans_ transp_ result.txt* sadrže respektivno, rezultate klasterovanja gena i ćelija i može se uočiti kao i u prethodnom slučaju da su klasteri ravnomerniji prilikom klasterovanja ćelija, ali da je koeficijent senke lošiji. Koeficijent senke prilikom klasterovanje gena je visok, što ukazuje na kvalitetno klasterovanje. Osim toga, klaster 10 prilikom klasterovanja ćelija sadrži jedan podatak, pa se može razmisliti o smanjivanju broja klastera.

019_kmeans_result.txt

```
Size of 0 is:   15677
Size of 1 is:   1
Size of 2 is:   22
Size of 3 is:   1
Size of 4 is:   2
Size of 5 is:   51
Size of 6 is:   35
Size of 7 is:   2
Size of 8 is:   171
Size of 9 is:   1
Size of 10 is:  2

Silhouette score:
0.8793639218184727
```

019_kmeans_transp_result.txt

```
Size of 0 is:   264
Size of 1 is:   84
Size of 2 is:   93
Size of 3 is:   409
Size of 4 is:   81
Size of 5 is:   58
```

```
Size of 6 is:   47
Size of 7 is:   42
Size of 8 is:   154
Size of 9 is:   380
Size of 10 is:  1

Silhouette score:
0.14238007574341446
```

Algoritam K-sredine je pokazao dobre rezultate prilikom klaterovanja gena. U narednom delu sledi implementacija algoritma koji vrši analizu inercije klastera 4 i algoritma koji implementira tehniku klasterovanja K-sredina 5. Programi se pokreću iz komandne linije komandom

```
python kmeans_ssd.py filepath
python kmeans.py filepath
```

Listing 3: Komande za pokretanje analize inercije i algoritma K-sredina

Dakle, analiza inercije se vrši pokretanjem funkcije *KMeans* i analiza se vrši promenom broja klastera koja iterira od [2, 15], što se može uočiti u sledećem delu koda.

```
    ...
def kmeans(data):
    for i in range(2,15):
        model = cluster.KMeans(n_clusters=i, random_state=42).fit(data.values
    )
        print("Sum of squared distances for k = " + str(i) + ": "+ str(model.
    inertia_))
    ...
```

Listing 4: Analiza inercije (SSE)

Dalje sledi kod vezan za tehniku klasterovanja K-sredina. Poziv je sličan, samo što ovaj put imamo fiksan broj klastera $\_cluster\_size$ i dohvatamo kojim klasterima su pridruženi geni pomoću promenljive $model.labels\_$.

```

def kmeans_fun(data, filename):
    ...
    model = cluster.KMeans(n_clusters= _cluster_size, random_state=42).fit(
    data.values)
    cluster_ids = model.labels_
    ...
```

Listing 5: Algoritam K-sredina

Na kraju prikazujemo delove datoteka koje sadrže ili gene ili ćelije grupisane po klasterima. Grupisanje gena po klasterima datoteke 018:

018_kmeans_model_grouped.txt

```
...
------------------6------------------
['hg38_ATP5F1E', 'hg38_CST3', 'hg38_MT-CO1', 'hg38_MT-CO2',
'hg38_MT-CO3', 'hg38_RPL10', 'hg38_RPL21', 'hg38_RPL27A',
'hg38_RPL28', 'hg38_RPL32', 'hg38_RPL34', 'hg38_RPL36',
'hg38_RPL37A', 'hg38_RPLP2', 'hg38_RPS12', 'hg38_RPS14',
'hg38_RPS15A', 'hg38_RPS18', 'hg38_RPS19', 'hg38_RPS27',
'hg38_SERF2']
```

```
-------------------7-------------------
['hg38_AIF1', 'hg38_APOC1', 'hg38_APOE', 'hg38_C1QA',
'hg38_C1QB', 'hg38_C1QC', 'hg38_CD14', 'hg38_CD74',
'hg38_CTSB', 'hg38_CTSD', 'hg38_CTSL', 'hg38_EEF1A1',
'hg38_FAU', 'hg38_FCER1G', 'hg38_FN1', 'hg38_GPX1',
'hg38_HLA-B', 'hg38_HLA-DRA', 'hg38_LAPTM5', 'hg38_LYZ',
'hg38_MT-ATP6', 'hg38_MT-ND1', 'hg38_MT-ND2', 'hg38_MT-ND3',
'hg38_MT-ND4', 'hg38_NEAT1', 'hg38_NPC2', 'hg38_OAZ1',
'hg38_PABPC1', 'hg38_PFN1', 'hg38_PTMA', 'hg38_RNASE1',
'hg38_RPL11', 'hg38_RPL12', 'hg38_RPL13', 'hg38_RPL13A',
'hg38_RPL15', 'hg38_RPL18A', 'hg38_RPL19', 'hg38_RPL22',
'hg38_RPL23A', 'hg38_RPL26', 'hg38_RPL27', 'hg38_RPL30',
'hg38_RPL31', 'hg38_RPL35', 'hg38_RPL35A', 'hg38_RPL36A',
'hg38_RPL37', 'hg38_RPL38', 'hg38_RPL6', 'hg38_RPL9',
'hg38_RPS13', 'hg38_RPS15', 'hg38_RPS16', 'hg38_RPS17',
'hg38_RPS2', 'hg38_RPS21', 'hg38_RPS23', 'hg38_RPS24',
'hg38_RPS25', 'hg38_RPS26', 'hg38_RPS27A', 'hg38_RPS3A',
'hg38_RPS6', 'hg38_RPS8', 'hg38_RPS9', 'hg38_S100A11',
'hg38_S100A4', 'hg38_S100A6', 'hg38_S100A9', 'hg38_SH3BGRL3',
'hg38_SRGN', 'hg38_TPT1', 'hg38_TYROBP', 'hg38_UBA52']
-------------------8-------------------
['hg38_FTH1', 'hg38_TMSB10']
-------------------9-------------------
['hg38_SPP1']
```

Grupisanje ćelija po klasterima datoteke 018:

018_kmeans_transp_model_grouped.txt

```
...
-------------------7-------------------
['5', '12', '21', '23', '24', '33', '41', '43',
'49', '64', '74', '85', '86', '94', '96', '98',
'103', '112', '115', '117', '123', '127', '137',
'151', '158', '172', '173', '176', '213', '222',
...
'2922', '2924', '2943', '2966', '2982', '2988',
'2991', '3004', '3008', '3023', '3047', '3052',
'3054', '3069', '3084', '3090', '3091', '3106',
'3111', '3114', '3115', '3119', '3129', '3131',
'3140', '3156', '3159',
-------------------8-------------------
['57', '62', '160', '175', '210', '247', '278',
'402', '508', '688', '696', '752', '773', '840',
'855', '915', '919', '935', '937', '960', '963',
'997', '1015', '1021', '1050', '1062', '1187',
...
'1357', '1383', '1392', '1397', '1403', '1486',
'1631', '1704', '1748', '1781', '1809', '1821',
'1860', '1922', '2033', '2155', '2159', '2254',
'2279', '2327', '2337', '2408', '2410', '2568',
'2578', '2748', '2793',
-------------------9-------------------
['362', '490', '515', '652', '713', '971', '1089',
'1179', '1193', '1341', '1480', '1670', '1718',
'1828', '1847', '1866', '1894', '1934', '1966',
'1989', '2065', '2096', '2467', '2619', '2990',
'3061', '3121']
```

Grupisanje gena po klasterima datoteke 019:

019_kmeans_model_grouped.txt

```
...
-------------------1-------------------
['hg38_MALAT1']
-------------------2-------------------
['hg38_ACTB', 'hg38_RPL10', 'hg38_RPL13', 'hg38_RPL13A',
'hg38_RPL21', 'hg38_RPL26', 'hg38_RPL27A', 'hg38_RPL28',
'hg38_RPL32', 'hg38_RPL34', 'hg38_RPL37A', 'hg38_RPL39',
'hg38_RPLP1', 'hg38_RPLP2', 'hg38_RPS12', 'hg38_RPS14',
'hg38_RPS15A', 'hg38_RPS18', 'hg38_RPS19', 'hg38_RPS27',
'hg38_RPS28', 'hg38_RPS29']
-------------------3-------------------
['hg38_TMSB4X']
-------------------4-------------------
['hg38_RPL41', 'hg38_TMSB10']
-------------------5-------------------
['hg38_ATP5F1E', 'hg38_CST3', 'hg38_EIF1', 'hg38_FAU',
'hg38_GPX1', 'hg38_H3F3A', 'hg38_HLA-DQB1', 'hg38_LGALS1',
'hg38_MT-ATP6', 'hg38_MT-CO3', 'hg38_MT-ND1', 'hg38_MT-ND2',
'hg38_MT-ND3', 'hg38_MT-ND4', 'hg38_MYL6', 'hg38_PFN1',
'hg38_RPL10A', 'hg38_RPL14', 'hg38_RPL18', 'hg38_RPL22',
'hg38_RPL23', 'hg38_RPL24', 'hg38_RPL27', 'hg38_RPL29',
'hg38_RPL3', 'hg38_RPL30', 'hg38_RPL31', 'hg38_RPL36A',
'hg38_RPL5', 'hg38_RPL6', 'hg38_RPL7', 'hg38_RPL7A',
'hg38_RPL8', 'hg38_RPS10', 'hg38_RPS11', 'hg38_RPS20',
'hg38_RPS21', 'hg38_RPS26', 'hg38_RPS3', 'hg38_RPS5',
'hg38_RPS7', 'hg38_S100A11', 'hg38_S100A4', 'hg38_S100A6',
'hg38_S100A8', 'hg38_S100A9', 'hg38_SERF2', 'hg38_SH3BGRL3',
'hg38_SRGN', 'hg38_TYROBP', 'hg38_UBA52']
...
```

Grupisanje ćelija po klasterima datoteke 019:

019_kmeans_transp_model_grouped.txt

```
...
-------------------4-------------------
['18', '32', '43', '74', '95', '108', '118',
'125', '134', '140', '155', '161', '228', '239',
'244', '253', '270', '311', '319', '320', '337',
'390', '410', '461', '492', '547', '549',
...
'1151', '1168', '1218', '1270', '1298', '1311',
'1317', '1321', '1331', '1391', '1407', '1463',
'1469', '1473', '1496', '1523', '1525', '1533',
'1543', '1573', '1613']
-------------------5-------------------
['2', '11', '56', '66', '79', '132', '150',
'184', '210', '259', '271', '283', '295',
'322', '341', '372', '421', '431', '479',
'556', '603', '684', '836', '901', '911',
...
'1161', '1181', '1222', '1237', '1240',
'1290', '1302', '1342', '1343', '1360',
'1373', '1406', '1409', '1419', '1440',
'1443', '1480', '1532', '1539', '1581',
'1598', '1612']
```

```
------------------6------------------
['45', '78', '146', '185', '206', '230',
'293', '294', '327', '392', '405', '407',
'428', '494', '509', '562', '563', '570',
'584', '606', '632', '636', '641', '654',
...
'841', '848', '852', '858', '998', '1039',
'1040', '1075', '1106', '1171', '1189', '1225',
'1282', '1296', '1354', '1451', '1474', '1484',
'1548', '1604']
------------------7------------------
['14', '139', '154', '172', '193', '237',
'282', '291', '310', '379', '477', '512',
'568', '580', '583', '616', '644', '645',
'646', '735', '755', '799', '814', '815',
'844', '847', '933', '957', '967', '985',
'1016', '1038', '1091', '1135', '1160', '1162',
'1180', '1185', '1187', '1303', '1376', '1415']
...
```

Grupisanje smo dobili pokretanjem sledećeg programa u komandnoj liniji:

```
python group_genes.py filename
```

Koeficijent senke računamo pokretanjem sledećeg programa u komandnoj liniji:

```
python python clustering_quality.py filename labelsfilename
```

3.2 Hijerarhijsko klasterovanje

Hijerarhijsko klasterovanje je tehnika klasterovanja koja gradi ugnježdene klastere spajanjem (engl. *agglomerative*) ili razdvanajem (engl. *divisive*). Ovaj pristup, poput K-sredina, spada u najranije tehnike klasterovanja, ali je još uvek rasprostranjen. U ovom radu prezentovani su rezultati dobijeni tehnikom klasterovanja koja gradi ugnježdene klastere spajanjem. Ideja je da se krene od jednog objekta i da se spoji sa njemu najbližim klasterom. Osim izbora tehnike izgradnje klastera, bitno je odrediti meru sličnosti i metodu povezivanja. U ovom istraživanju kao mera sličnosti je iskorišćeno euklidsko rastojanje, a za metod povezivanja je izabran metod Vard (engl. *Ward*). Metod povezivanja Vard minimizuje sumu kvadratnih razlika unutar klastera i koristi je kao kriterijum spajanja klastera, odnosno maksimizuje homogenost klastera. Zbog neuklanjanja šumova u procesu pretprocesiranja koristi se ova metoda, jer daje dobre rezultate prilikom klasterovanja skupova podataka koji sadrže veliki broj šumova [1].

U narednom delu ovog poglavlja slede rezultati dobijeni hijerarhijskim klasterovanjem kao i komanda poziva i deo programskog koda. Kao i kod K-sredine, klasterovali smo gene i ćelije. Rezultati su slični kao i prilikom klasterovanja tehnikom K-sredine. Naime, prilikom klasterovanja gena datoteke 018, dobijen je klaster koji sadrži veliki broj elemenata, ali je koeficijent senke blizu jedinice, tako da je kvalitet klasterovanja visok.

018_agglomerative_results.txt

```
Size of 0 is:   16
Size of 1 is:   2
Size of 2 is:   266
Size of 3 is:   38
Size of 4 is:   47
Size of 5 is:   1
Size of 6 is:   1
Size of 7 is:   1
Size of 8 is:   14402
Size of 9 is:   1

Silhouette score:
0.8058630871936969
```

Klasterovanje ćelija je dalo ravnomernije klastere, ali je koeficijent senke dosta mali, što ukazuje na preklapanja između klastera. Sledeća datoteka sadrži veličine klastera datoteke 018.

018_agglomerative_transp_results.txt

```
Size of 0 is:   417
Size of 1 is:   590
Size of 2 is:   17
Size of 3 is:   42
Size of 4 is:   1375
Size of 5 is:   5
Size of 6 is:   498
Size of 7 is:   135
Size of 8 is:   72
Size of 9 is:   45

Silhouette score:
0.08568095787713036
```

Naredne dve datoteke sadrže veličine klastera dobijene prilikom klasterovanja gena i klasterovanja ćelija, respektivno, datoteke 019. Rezultati su slični onima koje smo dobili prilikom klasterovanja datoteke 018.

019_agglomerative_results.txt

```
Size of 0 is:   8
Size of 1 is:   33
Size of 2 is:   67
Size of 3 is:   2
Size of 4 is:   2
Size of 5 is:   15597
Size of 6 is:   37
Size of 7 is:   1
Size of 8 is:   1
Size of 9 is:   1
Size of 10 is:  216

Silhouette score:
0.8465005173194241
```

019_agglomerative_transp_results.txt

```
Size of 0 is:   612
Size of 1 is:   133
Size of 2 is:   223
Size of 3 is:   126
Size of 4 is:   36
Size of 5 is:   342
Size of 6 is:   39
Size of 7 is:   1
Size of 8 is:   11
Size of 9 is:   5
Size of 10 is:  85

Silhouette score:
0.184014418620219
```

Hijerarhijsko klasterovanje se poziva izvršavanje sledeće komande

```
python agglomerative_clustering.py filepath
```

Listing 6: Komanda za pokretanje algoritma hijerarhijskog klasterovanja

Program poziva iz Pajton biblioteke *cluster* funkciju *AgglomerativeClustering* koja prihvata kao argument broj klastera. Podrazumevana mera sličnosti je euklidsko rastojanje, a metoda povezivanja Vard.

```
def agglomerative_clustering(data, filename):
    ...
    hierarchy_model = cluster.AgglomerativeClustering(n_clusters=
     _cluster_size).fit(data.values)
    cluster_ids = hierarchy_model.labels_
    ...
```

Listing 7: Algoritam hijerarhijskog klasterovanja

Na kraju prikazujemo deo datoteka koje sadrže ili gene ili ćelije grupisane po klasterima. Grupisanje gena po klasterima datoteke 018:

018_agglomerative_model_grouped.txt

```
...
-------------------3-------------------
['hg38_AIF1', 'hg38_APOC1', 'hg38_APOE', 'hg38_C1QA',
'hg38_C1QB', 'hg38_C1QC', 'hg38_CD14', 'hg38_CD74',
'hg38_CST3', 'hg38_CTSB', 'hg38_CTSD', 'hg38_CTSL',
...
'hg38_PABPC1', 'hg38_RNASE1', 'hg38_S100A11', 'hg38_S100A4',
'hg38_S100A6', 'hg38_S100A9', 'hg38_SERF2', 'hg38_SH3BGRL3',
'hg38_SRGN', 'hg38_TYROBP']
-------------------4-------------------
['hg38_EEF1A1', 'hg38_FAU', 'hg38_LYZ', 'hg38_PTMA',
'hg38_RPL10', 'hg38_RPL11', 'hg38_RPL12', 'hg38_RPL13',
'hg38_RPL13A', 'hg38_RPL18A', 'hg38_RPL19', 'hg38_RPL21',
...
'hg38_RPS21', 'hg38_RPS23', 'hg38_RPS24', 'hg38_RPS25',
'hg38_RPS26', 'hg38_RPS27A', 'hg38_RPS6', 'hg38_RPS8',
'hg38_RPS9', 'hg38_TPT1', 'hg38_UBA52']
```

```
-------------------5-------------------
['hg38_FTL']
-------------------6-------------------
['hg38_SPP1']
-------------------7-------------------
['hg38_MALAT1']
...
```

Grupisanje ćelija po klasterima datoteke 018:

018_agglomerative_transp_model_grouped.txt

```
...
-------------------7-------------------
['10', '18', '119', '120', '121', '304',
'358', '375', '388', '408', '438', '444',
'454', '479', '492', '493', '505', '519',
...
'2824', '2875', '2896', '2923', '2931', '2998',
'3002', '3014', '3021', '3063', '3067', '3071',
'3076', '3105', '3126', '3169', '3178']
-------------------8-------------------
['57', '62', '175', '210', '247', '278',
'329', '391', '402', '508', '685', '688',
'696', '752', '773', '840', '855', '915',
...
'2408', '2568', '2578', '2748', '2789', '2793',
'2912', '2956', '2960', '3007', '3053', '3107',
'3124', '3147', '3149', '3171', '3193']
-------------------9-------------------
['45', '148', '182', '249', '353', '607',
'652', '689', '731', '733', '809', '880',
'971', '1089', '1139', '1179', '1193', '1321',
...
'2096', '2372', '2481', '2619', '2634', '2698',
'2703', '2808', '3108', '3109', '3113', '3116',
'3121', '3130', '3179']
```

Grupisanje gena po klasterima datoteke 019:

019_agglomerative_model_grouped.txt

```
-------------------0-------------------
['hg38_ACTB', 'hg38_CD74', 'hg38_HLA-DRA', 'hg38_RPL10',
'hg38_RPL39', 'hg38_RPLP1', 'hg38_RPS27', 'hg38_RPS29']
-------------------1-------------------
['hg38_EEF1A1', 'hg38_HLA-DPA1', 'hg38_HLA-DPB1',
'hg38_HLA-DRB1', 'hg38_MT-CO1', 'hg38_MT-CO2', 'hg38_PTMA',
'hg38_RPL11', 'hg38_RPL13', 'hg38_RPL13A', 'hg38_RPL18A',
...
'hg38_RPS14', 'hg38_RPS15', 'hg38_RPS15A', 'hg38_RPS18',
'hg38_RPS19', 'hg38_RPS2', 'hg38_RPS23', 'hg38_RPS24',
'hg38_RPS27A', 'hg38_RPS28']
-------------------2-------------------
['hg38_ACTG1', 'hg38_AIF1', 'hg38_ATP5MC2', 'hg38_ATP5MG',
'hg38_CFL1', 'hg38_COX4I1', 'hg38_COX7C', 'hg38_CYBA',
'hg38_DDX5', 'hg38_EEF1B2', 'hg38_EIF1', 'hg38_FCER1G',
```

```
...
'hg38_S100A10', 'hg38_S100A11', 'hg38_S100A8', 'hg38_S100A9',
'hg38_SAT1', 'hg38_SH3BGRL3', 'hg38_SRGN', 'hg38_TMA7',
'hg38_TOMM7', 'hg38_UQCR11', 'hg38_UQCRB', 'hg38_VIM']
------------------3------------------
['hg38_B2M', 'hg38_FTH1']
------------------4------------------
['hg38_RPL41', 'hg38_TMSB10']
...
```

Grupisanje ćelija po klasterima datoteke 019:

019_agglomerative_transp_model_grouped.txt

```
...
------------------6------------------
['150', '166', '170', '184', '210', '271',
'283', '295', '341', '372', '421', '479',
'556', '603', '743', '836', '873', '1010',
'1013', '1037', '1067', '1116', '1155', '1161',
'1219', '1222', '1237', '1290', '1302', '1325',
'1343', '1360', '1406', '1409', '1419', '1440',
'1458', '1480', '1598']
------------------7------------------
['587']
------------------8------------------
['18', '140', '641', '786', '1008', '1090',
'1296', '1317', '1331', '1473', '1543']
------------------9------------------
['293', '310', '327', '563', '1225']
-----------------10------------------
['12', '20', '33', '43', '71', '88', '92',
'113', '125', '127', '147', '175', '192',
'221', '228', '231', '248', '254', '323',
...
'1265', '1270', '1283', '1340', '1393', '1395',
'1403', '1407', '1413', '1424', '1464', '1492',
'1520', '1533', '1549', '1564', '1569', '1573']
```

3.3 Klasterovanje metodom *BIRCH*

BIRCH (engl. *balanced iterative reducing and clustering using hierarchies*) predstavlja uopštenje algoritma k-sredina na hijerarhijsku metodu odozgo-naniže koje gradi drvo karakterističnih osobina (engl. *Characteristic Feature Tree, CFT*). Podaci su kompresovani sa zanemarljivim gubicima na skup čvorova sa karakterističnim osobinama. Čvorovi imaju potklastere (engl. *Characteristic Feature subclusters*) koji se ne nalaze u listovima mogu da sadrže druge čvorove. potklasteri čuvaju neophodne podatke razdvajanja i time omogućavaju da se u memoriji čuvaju delovi ulaznih podataka [1]. Algoritam je otporan na šum i koristi se u radu sa velikim skupom podataka [**?**]. Ulazni parametri algoritma su prag (engl. *treshold*) i faktor grananja (engl. *branching factor*). Prag ograničava rastojanje između ulaznog uzorka i postojećih potklastera, dok faktor grananja ograničava broj potklastera unutar čvora [4].

U narednom delu slede rezultati dobijeni korišćenjem algoritma BIRCH nad datotekama 018 i 019. Rezultat klasterovanja gena datoteke 018 sadrži klastere različitih veličina i veliki broj gena se našao u klasteru 8. Kvalitet klasterovanja je visok jer je koeficijent senke veliki.

018_birch_result.txt

```
Size of 0 is:        16
Size of 1 is:        2
Size of 2 is:        310
Size of 3 is:        38
Size of 4 is:        47
Size of 5 is:        1
Size of 6 is:        1
Size of 7 is:        1
Size of 8 is:        14358
Size of 9 is:        1

Silhouette score:
0.795237165198908
```

Rezultat klasterovanja ćelija datoteke 018 je dao ravnomernije klastere, ali nizak kvalitet zbog koeficijenta senke.

018_birch_transp_result.txt

```
Size of 0 is:        1738
Size of 1 is:        603
Size of 2 is:        396
Size of 3 is:        42
Size of 4 is:        7
Size of 5 is:        161
Size of 6 is:        45
Size of 7 is:        72
Size of 8 is:        127
Size of 9 is:        5

Silhouette score:
0.13356316795928133
```

Što se tiče datoteke 019, rezultati su slični kao kod datoteke 018. Klasterovanjem gena, dobijamo klastere visokog kvaliteta.

019_birch_result.txt

```
Size of 0 is:        4
Size of 1 is:        8
Size of 2 is:        54
Size of 3 is:        33
Size of 4 is:        203
Size of 5 is:        15623
Size of 6 is:        37
Size of 7 is:        1
Size of 8 is:        1
Size of 9 is:        1
```

```
Silhouette score:
0.8572286846946008
```

Rezultati dobijeni klasterovanjem ćelija datoteke 019 su lošiji u odnosu na klasterovanje gena datoteke 019. Najveći broj podataka je u klasteru 2.

019_birch_transp_result.txt

```
Size of 0 is:        308
Size of 1 is:        133
Size of 2 is:        612
Size of 3 is:        126
Size of 4 is:        36
Size of 5 is:        342
Size of 6 is:        39
Size of 7 is:        1
Size of 8 is:        11
Size of 9 is:        5

Silhouette score:
0.18575782355595447
```

Program koji pokreće algoritam BIRCH se poziva izvršavanjem sledeće komande u komandnoj liniji

```
python birch.py filepath
```

Programski kod algoritma 8 poziva funkciju *Birch* iz biblioteke *cluster* i kao argument joj se prosleđuje broj klastera. Vrednost prag je 0.5, a faktora grananja 50 i te vrednosti predstavljaju podrazumevane vrednosti.

```
def birch_clustering(data, filename):
    ...
    model = cluster.Birch(n_clusters=_cluster_size).fit(data.values)
    ...
```

Listing 8: Algoritam BIRCH

U nastavku sledi prikaz delova datoteka koje sadrže ili gene ili ćelije grupisane po klasterima ili datoteke 018 ili datoteke 019. Grupisanje gena po klasterima datoteke 018:

018_birch_model_grouped.txt

```
...
------------------3------------------
['hg38_AIF1', 'hg38_APOC1', 'hg38_APOE', 'hg38_C1QA',
'hg38_C1QB', 'hg38_C1QC', 'hg38_CD14', 'hg38_CD74',
...
'hg38_PABPC1', 'hg38_RNASE1', 'hg38_S100A11', 'hg38_S100A4',
'hg38_S100A6', 'hg38_S100A9', 'hg38_SERF2', 'hg38_SH3BGRL3',
'hg38_SRGN', 'hg38_TYROBP']
------------------4------------------
['hg38_EEF1A1', 'hg38_FAU', 'hg38_LYZ', 'hg38_PTMA',
'hg38_RPL10', 'hg38_RPL11', 'hg38_RPL12', 'hg38_RPL13',
```

```
...
'hg38_RPS26', 'hg38_RPS27A', 'hg38_RPS6', 'hg38_RPS8',
'hg38_RPS9', 'hg38_TPT1', 'hg38_UBA52']
------------------5------------------
['hg38_FTL']
------------------6------------------
['hg38_SPP1']
------------------7------------------
['hg38_MALAT1']
...
```

Grupisanje ćelija po klasterima datoteke 018:

018_birch_transp_model_grouped.txt

```
...
------------------3------------------
['238', '252', '274', '336', '604', '629', '702', '730',
'920', '926', '959', '999', '1032', '1076', '1104', '1140',
'2152', '2233', '2257', '2290', '2299', '2317', '2425',
'2971', '3039', '3045', '3146', '3155']
------------------4------------------
['713', '1480', '1828', '1894', '2467', '2990', '3061']
------------------5------------------
['26', '37', '47', '54', '63', '116', '146', '159', '163',
'174', '187', '189', '195', '206', '214', '220', '234',
...
'2950', '2972', '3010', '3040', '3080', '3088', '3117',
'3142', '3166', '3168', '3170', '3173', '3177']
------------------6------------------
['45', '148', '182', '249', '353', '607', '652', '689',
'731', '733', '809', '880', '971', '1089', '1139', '1179',
...
'2096', '2372', '2481', '2619', '2634', '2698', '2703',
'2808', '3108', '3109', '3113', '3116', '3121', '3130',
'3179']
------------------7------------------
['57', '62', '175', '210', '247', '278', '329', '391', '402',
'508', '685', '688', '696', '752', '773', '840', '855',
... '2279', '2327', '2408', '2568', '2578', '2748', '2789',
'2793', '2912', '2956', '2960', '3007', '3053', '3107',
'3124', '3147', '3149', '3171', '3193']
...
```

Grupisanje gena po klasterima datoteke 019:

019_birch_model_grouped.txt

```
------------------0------------------
['hg38_B2M', 'hg38_FTH1', 'hg38_RPL41', 'hg38_TMSB10']
------------------1------------------
['hg38_ACTB', 'hg38_CD74', 'hg38_HLA-DRA', 'hg38_RPL10',
'hg38_RPL39', 'hg38_RPLP1', 'hg38_RPS27', 'hg38_RPS29']
------------------2------------------
['hg38_ACTG1', 'hg38_AIF1', 'hg38_CD14', 'hg38_CFL1',
'hg38_COX7C', 'hg38_CYBA', 'hg38_EIF1', 'hg38_FCER1G',
...
```

```
'hg38_RPS7', 'hg38_RPSA', 'hg38_S100A10', 'hg38_S100A11',
'hg38_S100A8', 'hg38_S100A9', 'hg38_SAT1', 'hg38_SH3BGRL3',
'hg38_SRGN', 'hg38_VIM']
------------------3------------------
['hg38_EEF1A1', 'hg38_HLA-DPA1', 'hg38_HLA-DPB1',
'hg38_HLA-DRB1', 'hg38_MT-CO1', 'hg38_MT-CO2', 'hg38_PTMA',
'hg38_RPL11', 'hg38_RPL13', 'hg38_RPL13A', 'hg38_RPL18A',
...
'hg38_RPS14', 'hg38_RPS15', 'hg38_RPS15A', 'hg38_RPS18',
'hg38_RPS19', 'hg38_RPS2', 'hg38_RPS23', 'hg38_RPS24',
'hg38_RPS27A', 'hg38_RPS28']
...
```

Grupisanje ćelija po klasterima datoteke 019:

019_birch_transp_model_grouped.txt

```
...
------------------6------------------
['150', '166', '170', '184', '210', '271', '283', '295', '341',
'372', '421', '479', '556', '603', '743', '836', '873', '1010',
'1013', '1037', '1067', '1116', '1155', '1161', '1219', '1222',
'1237', '1290', '1302', '1325', '1343', '1360', '1406', '1409',
'1419', '1440', '1458', '1480', '1598']
------------------7------------------
['587']
------------------8------------------
['18', '140', '641', '786', '1008', '1090', '1296', '1317',
'1331', '1473', '1543']
------------------9------------------
['293', '310', '327', '563', '1225']
```

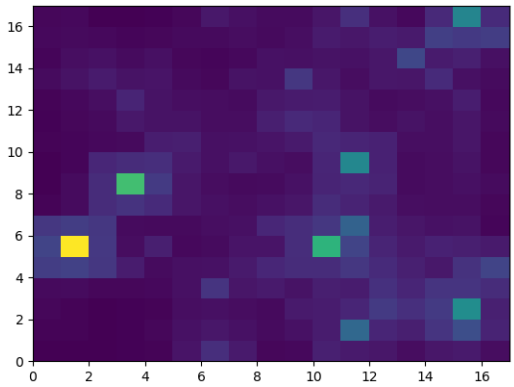
3.4 Samo-organizujuće mape

Kohonen je razvio samo-organizujuće mape inspirisan samo-organizovanošću ljudskog cerebralnog korteksa. Samo-organizujuća mapa (SOM) je tehnika višedimenzionog skaliranje ulaznih signala iz prostora dimenzije I u diskretan prostor manjih dimenzija J. Zbog vizuelizacije, diskretan prostor je najčešće dimenzije $J = 2$. Ideja se zasniva na zadržavanju topološke strukture ulaznih podataka, gde važi da ukoliko su dva podatka blizu u ulaznom prostoru, biće blizu i u izlaznom prostoru [2, 4]. Klasterovanje samo-organizujućim mapama se sastoji iz dva dela: učenja i preslikavanja. Učenje je zasnovano na kompetitivnoj strategiji i ulazni podaci su povezani sa odgovarajućim neuronima u mapi. Mapa je uglavnom kvadratnog oblika i dimenzija mape se određuje pomoću formule

$$M^2 \approx 5\sqrt{N}, \tag{2}$$

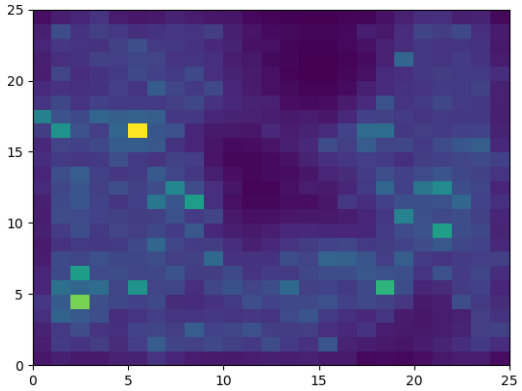
gde je M^2 veličina izlaznih podataka, odnosno broj neurona, a N veličina podataka [5]. SOM koristi toplotnu mapu kako bi predstavio rezultat dvodimenzionalnog diskretnog izlaza i na osnovu predefinsane boje toplota, sa mape se može odrediti broj klastera. U nastavku sledi osam toplotnih mapa dobijenih izršavanjem tehnike klasterovanja SOM. Prilikom izvršavanja algoritma, korišćen je ili nasumični ili sekvencijalni izbor ili gena ili ćelija.

Naredna toplotna mapa 1 je dobijena izvršavanjem programa nad genima datoteke 018. Izbor gena je nasumičan. Sa nje se može uočiti između 7 i 8 klastera od kojih je jedan dominantan klaster, odnosno najveći broj gena je raspodeljen u dobijeni klaster. Dakle, mape se čitaju tako što svetliji delovi mape predstavljaju klastere sa velikim brojem elemenata, odnosno tamniji predstavljaju elemente van granica ili praznine. Dimenzija ove mape, kao i svih narednih, određena je pomoću prethodno navedene formule (2).



Slika 1: Toplotna mapa dobijena sekvencijalnim izborom gena datoteke 018

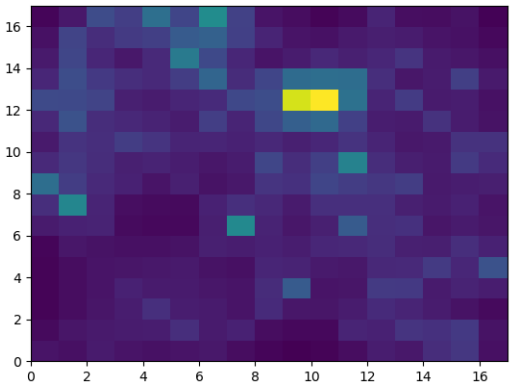
Naredna toplotna mapa 2 je dobijena klasterovanjem ćelija datoteke 018. Broj klastera je u opsegu od 8 do 14. Može se uočiti dva dominantna klastera i veliki broj manjih klastera.



Slika 2: Toplotna mapa dobijena sekvencijalnim izborom ćelija datoteke 018

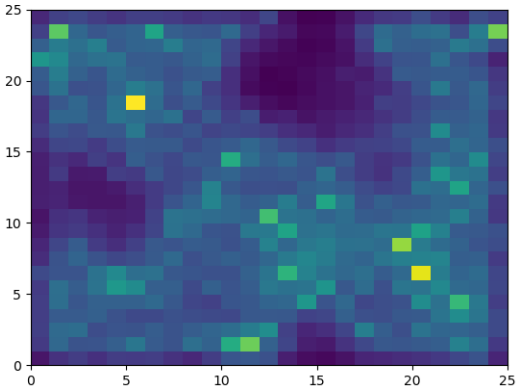
Toplotna mapa 3 dobijena nasumničnim izborom gena sadrži jedan domi-

nantan klaster i nekoliko manjih klastera. Broj klastera varira između 5 i 7.



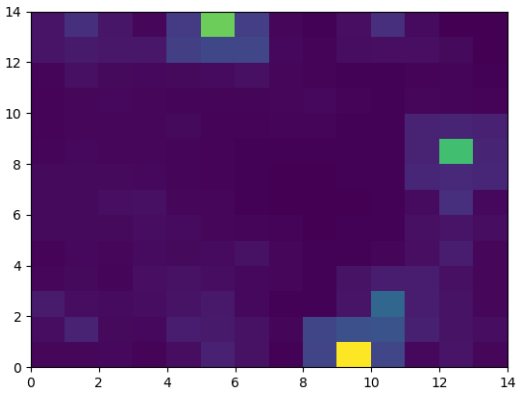
Slika 3: Toplotna mapa dobijena nasumičnim izborom gena datoteke 018

Toplotna mapa 4 dobijena nasumičnim izborom ćelija sadrži bar četiri klastera, dok je gornju granicu teško odrediti na osnovu dobijenog rezultata.



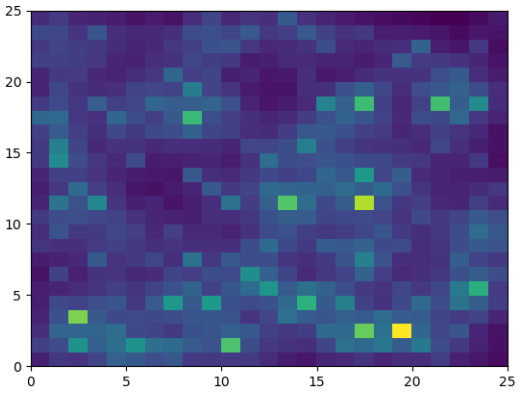
Slika 4: Toplotna mapa dobijena nasumičnim izborom ćelija datoteke 018

Pređimo na datoteku 019. Toplotna mapa 5 dobijena je sekvencijalnim izborom gena. Može se uočiti da nedvosmisleno postoje tri klastera.



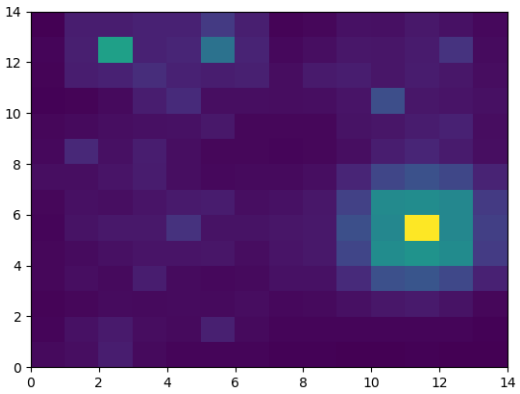
Slika 5: Toplotna mapa dobijena sekvencijalnim izborom gena datoteke 019

Kod klasterovanja ćelija i primenom sekvencijalnog izbora, dobijena je toplotna mapa 6 i veliki broj klastera, pa je teško uočiti tačan broj. Ono što se može uočiti jeste da postoje bar četiri klastera.



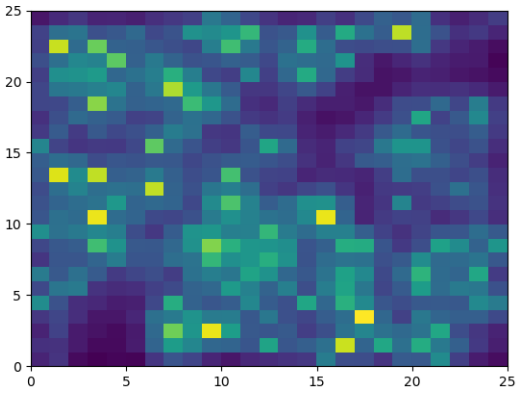
Slika 6: Toplotna mapa dobijena sekvencijalnim izborom ćelija datoteke 019

Korišćenjem nasumičnog izbora gena, dobijena je toplotna mapa 7 koja sadrži jedan klaster u kojem se nalazi najveći broj gena i dva manja klastera.



Slika 7: Toplotna mapa dobijena nasumičnim izborom gena datoteke 019

Na toplotnoj mapi 8 se može uočiti bar 5 klastera. Klasteri nisu srazmerni, odnosno klaster koji se nalazi u donjem delu mape je veći od ostalih klastera.



Slika 8: Toplotna mapa dobijena nasumičnim izborom ćelija datoteke 019

U nastavku sledi programski kod koji se pokreće izvršavanjem naredne komande u komandnoj liniji

```
python som.py filepath
```

Listing 9: Algoritam klasterovanj pomoću SOM

Kod 10 sadrži funkciju *MiniSom* kojoj se prosleđuju dimenzije toplotne mape, kao i veličina ulaznih podataka. U zavisnosti da li koristimo ili nasumičan ili sekvencijalan izbor, koristimo funkcije *train_ random* ili *train_ batch*, respektivno. Na kraju kod se vrši iscrtavanje toplotne mape.

```
def minisom_fun(data, filename):
    ...
    model = minisom.MiniSom(17, 17, len(data.values[1]))
    model.random_weights_init(data.values)
    model.train_random(data.values, 100)
    # model.train_batch(data.values, 100)
    plt.pcolor(model.distance_map().T)
    plt.show()
    ...
```

Listing 10: Algoritam SOM

3.5 Poređenje koeficijenata senke

Već je napomenuto da koeficijent senke određuje kvalitet klasterovanja. U ovom poglavlju uporedićemo koeficijente senke dobijene izvršavanjem različitih algoritama nad genima i ćelijama datoteka 018 i 019. U tabeli 1 se nalaze dobijeni rezultati i na osnovu njih možemo zaključiti da je klasterovanje gena odrađeno sa visokim kvalitetom, dok klasterovanje ćelija je je uglavnom dalo koeficijente senki koji su bliži 0, nego 1, što znači da postoje preklapanja među klasterima. Najbolje klasterovanje gena datoteke 018 dobili smo korišćenjem hijerarhijskog klasterovanja, ali na četvrtoj decimali. Zbog vremenske složenosti hijerarhijskog algoritma koji je implementiran i iznosi $O(n^3)$ [6], bolje je koristiti ovaj algoritam jer je vremenska složenost algoritma K-sredina $O(n^{dk+1})$[3], gde su d i k dimenzija i broj klastera, koji su unapred poznati. BIRCH nam je dao lošiji rezultat od ostala dva algoritma, ali ono što ide u korist BIRCH-a jeste što može da postigne vremensku složenost $O(n)$ [7]. Dakle, ukoliko nam je vreme izračunavanja bitno, BIRCH je najbolji izbor. Današnji računari uglavnom prevazilaze problem memorijske složenosti, pa je fokus samo na vremenskoj složenosti izvršavanja algoritma. Klasterovanje gena datoteka 019 je najbolje dobijeno pomoću K-sredina i razlikuje se u odnosu na ostale na drugoj decimali. Što se tiče klasterovanja ćelija, u svakoj situaciji je BIRCH dao najbolje rezultate, ali je koeficijent senke u blizini nule, pa smatramo da klasterovanje nije odrađeno kvalitetno.

Tabela 1: Koeficijenti senke dobijeni izvršavanjem različitih algoritama

datoteka	K-sredina	hijerarhijsko	BIRCH
018 geni	0.8055	0.8059	0.7952
018 ćelije	0.0878	0.0857	0.1336
019 geni	0.8795	0.8465	0.8572
019 ćelije	0.1424	0.1840	0.1858

4 Zaključak

Kvalitetna analiza podataka koja sadrži gene i njihovu ekspresiju može doprineti ka njihovom boljem razumevanju. Ideja ovog rada je prikaz i primena različitih tehnika klasterovanja nad podacima velikih dimenzija. Ispostavilo se da visoka dimenzionalnost podataka predstavlja pravi izazov prilikom klasterovanja gena, kao i ćelija. Klasterovanjem gena dobijali smo neravnomerne

klastere u odnosu na klasterovanje ćelija, ali je koeficijent senke obećavajući, pa se može smatrati da su geni uspešno razvrstani u klastere. Koeficijent senke kod klasterovanja ćelija je u blizini nule, pa odbacujemo ovaj pristup uz primenjene algoritme. Daljim istraživanjem i detaljnijim radom sa vrednostima parametra korišćenih algoritama, mogu se unaprediti rezultati klasterovanja.

Literatura

[1] Charu C Aggarwal. *Data mining: the textbook*. Springer, 2015.

[2] Andries P Engelbrecht. *Computational intelligence: an introduction*. John Wiley & Sons, 2007.

[3] Mary Inaba, Naoki Katoh, and Hiroshi Imai. Applications of weighted voronoi diagrams and randomization to variance-based k-clustering: (extended abstract). In *Proceedings of the Tenth Annual Symposium on Computational Geometry*, SCG '94, pages 332–339, New York, NY, USA, 1994. ACM.

[4] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[5] Ignacio Rojas, Gonzalo Joya, and Andreu Catala. *Advances in Computational Intelligence: 13th International Work-Conference on Artificial Neural Networks, IWANN 2015, Palma de Mallorca, Spain, June 10-12, 2015. Proceedings*, volume 9094. Springer, 2015.

[6] R. Sibson. SLINK: An optimally efficient algorithm for the single-link cluster method. *The Computer Journal*, 16(1):30–34, 01 1973.

[7] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. Birch: An efficient data clustering method for very large databases. *SIGMOD Rec.*, 25(2):103–114, June 1996.