Klasterizacija gena
-Projektni zadatak-

Daniela Kotur, EE95/2014, danielakotur@gmail.com

I. UVOD

Genomika tumora teži da otkrije molekularnu osnovu tumora koristeći različite nivoe genomskih informacija, od kojih su najrasprostranjeniji podaci o ekspresiji gena. Zahvaljujući mikročip tehnologiji trenutno je moguće posmatrati gene čitavog organizma u različitim uslovima. Međutim, ogroman broj gena kao i kompleksnost bioloških mreža otežavaju tumačenje i razumijevanje velike količine podataka koji često sadrže milione mjerenja.

Klasterovanje se pokazalo kao idealno rješenje za otkrivanje strukture u visokodimenzionalnim podacima. Omogućuje razumijevanje funkcije gena, regulacije ćelijskih procesa kao i otkrivanje novih podtipova ćelija. Shodno tome, pomaže pri otkrivanju novih vrsta tumora kao i poboljšanju tretmana liječenja pacijenta, što je od velikog značaja za istraživanja u medicini.

II. BAZA PODATAKA

Bazu čini 21 set podataka o ekspresiji gena različitih tipova raka. Svaki set se sastoji od velikog broja uzoraka i atributa koji predstavljaju broj gena, odnosno dimenzionalnost. Podaci su dobijeni sa čipom Affymetrix i imaju dosta neujednačene vrijednosti ekspresija. Takođe, date su i labele koje ukazuju na broj podtipova raka.

TABELA 1: PODACI O EKSPESIJI GENA RAZLIČITIH TKIVA; K JE BROJ PRIRODNIH KLASTERA, N BROJ UZORAKA, A D REDUKOVANA DIMENZIONALNOST

Set podataka	Tkivo	k	N	D
Armstrong-V1	Blood	2	72	1081
Armstrong-V2	Blood	3	72	2194
Bhattacharjee	Lung	5	203	1543
Chowdary	Breast	2	104	182
Dyrskjot	Bladder	3	40	1203
Golub-V1	Bone marrow	2	72	1877
Golub-V2	Bone marrow	3	72	1877
Gordon	Lung	2	181	1626
Laiho	Colon	2	37	2202
Nutt-V1	Brain	4	50	1377
Nutt-V2	Brain	2	28	1070
Nutt-V3	Brain	2	22	1152
Pomeroy-V1	Brain	2	34	857
Pomeroy-V2	Brain	5	42	1379
Ramaswamy	Multi-tissue	14	190	1363
Shipp	Blood	2	77	798
Singh	Prostate	2	102	339
Su	Multi-tissue	10	174	1571
West	Breast	2	49	1198
Yeoh-V1	Bone marrow	2	248	2526
Yeoh-V2	Bone marrow	6	248	2526

III. ANALIZA PODATAKA

U ovom radu biće prikazani rezultati dobijeni spektralnim klasterovanjem nad nenormalizovanim, z-normalizovanim i range normalizovanim podacima. Cilj klasterovanja je pronalazak grupe uzoraka sa sličnim obrascima ekspresije gena, odnosno otkrivanje podtipova ćelija raka.

Kako bi se dobila realna predstava o performansama algoritma, algoritam je ponavljan 15 puta. Rezultati algoritma biće prikazani pomoću Adjusted Rand Index mjere validacije klasterovanja, čije varijacije u indeksima ukazuju na robusnost konsenzus rješenja.

.

A. Spektralno klasterovanje

Metoda spektralnog klasterovanja se bazira na definisanju sličnosti između svih tačaka koje želimo klasterovati, pridruženoj matrici sličnosti te njenoj spektralnoj analizi. Glavni cilj je kao i kod svakog algoritma – postići veliku sličnost među podacima unutar iste grupe, a malu sličnost između podataka u različitim grupama.

Ova metoda zahtjeva da se podaci koje je potrebno grupisati prikažu u formi neusmjerenog grafa u kojem čvorovi odgovaraju podacima a ivice koje povezuju tačke definišemo kao mjere sličnosti, odnosno rastojanja između tačaka. Kako bismo particionisali graf, potrebne su funkcije poput reza, razmjernog ili normalizovanog reza. Najbolji je normalizovani rez jer posjeduje svojstvo dualnosti, tj.svojstvo istovremene minimizacije sličnosti između klastera te maksimizacije sličnosti unutar klastera, što je i bio glavni cilj. Optimizacija normalizovanog reza predstavlja težak NP problem, koji se rješava korišćenjem Laplasijan grafa, njegovih svojstvenih vrijednosti i vektora.

U okviru ovog klasterovanja, koristi se jedan od najjednostavnijih i najpopularnijih algoritama, k-means,

koji na iterativan način minimizuje varijacije unutar klastera. Ovaj algoritam zavisi od početne inicijalizacije i neophodno je definisati broj klastera na koji će se uzorci dijeliti. U ovom slučaju, to je broj klasa koji je različit za svaki set podataka.

K-means metoda daje jako dobre rezultate kod problema particionisanja konveksnih skupova. U slučaju kada skupovi nisu konveksni, ova metoda će u većini slučaja potpuno zakazati. Upravo zbog toga se koriste spektralne metode jer će se u kombinaciji sa k-means metodom pokazati kao veoma dobar alat za particionisanje skupova.

B. *Mjere sličnosti*

Sličnost se najčešće tumači kao udaljenost između objekata. Što je udaljenost manja, objekti su sličniji, i po pravilu grupisani u istom klasteru. U ovom radu korišćeno je Helingerovo i Euklidsko rastojanje, kao i Pirsonova korelacija, koji su opisani u nastavku.

❖ Euklidsko rastojanje

Udaljenost između dva uzorka Xi i Xj u n-dimenzionalnom prostoru definisana je kao:

$$E(X_i, X_j) = \sqrt{\sum_{d=1}^{n} (X_{id} - X_{jd})^2}$$

Što je manja vrijednost dobijenih brojeva, veća je sličnost između uzoraka.

❖ Helingerovo rastojanje

$$H(X_i, X_j) = \frac{1}{\sqrt{2}} \sqrt{\sum_{k=1}^{n} (\sqrt{\frac{X_{ik}}{\sum X_{ik}}} - \sqrt{\frac{X_{jk}}{\sum X_{jk}}})^2}$$

Helingerovo rastojanje uzima vrijednosti iz opsega [0,1], gdje su vrijednosti sličnijih uzoraka bliži jedinici.

❖ Pirsonova korelacija

Pirsonova korelacija se koristi da bi se utvrdila korelisanost između uzoraka. Uzima vrijednosti između -1 i 1. Ako je vrijednost koeficijenta u negativnom opsegu to znači da je odnos između uzoraka negativno korelisan, odnosno kada se jedna vrijednost povećava, druga se smanjuje. U suprotnom, obe vrijednosti se zajedno smanjuju ili povećavaju.

Postoji ugrađena funkcija corr koja vraća Pirsonov koeficijent korelacije.

C. *Normalizacija*

S obzirom da su mjere rastojanja između uzoraka osjetljive na velike razlike u vrijednostima obilježja, potrebno je normalizovati podatke. Normalizacijom se vrši ekvalizacija vrijednosti obilježja i smanjuje se njihova varijabilnost odnosno usklađuju se relativne težine obilježja.

Z- normalizacijom se transformišu uzorci tako da im je srednja vrijednost jednaka nula, a varijansa jedan. Mana ovog algoritma je što je osjetljiv na outliere, i što može da se desi da izračunata srednja vrijednost i standardna devijacija ne aproksimiraju dobro njihove stvarne vrijednosti. U ovom radu, izvršena je z-normalizacija po vrsti i koloni pomoću ugrađene funkcije zscore.

Range normalizacijom dobijaju se vrijednosti u opsegu [0,1].. Takođe je rađena normalizacija po vrsti i koloni.

D. *Adjusted Rand Index*

Adjusted Rand Index (skraćeno ARI) pripada eksternim mjerama validacije koji za razliku od ostalih ne sagleda samo separaciju uzoraka koji pripadaju različitim klasama nego i odnose između uzoraka u toj klasi. Uzima vrijednosti iz intervala [-1,1], gdje 1 znači potpuno slaganje particija, a vrijednosti blizu 0 kao i negativne vrijednosti upućuju na slučajnost u njihovom slaganju.

ARI ne zavisi od vrste algoritma kao ni od broja klastera u particiji, a osnovna svrha mu je da evaluira sposobnost algoritma da razdvoji elemente koji ne pripadaju istoj klasi. U nastavku biće ukazano na značaj opsega vrijednosti ARI-ja primjenjenog na različite setove podataka.

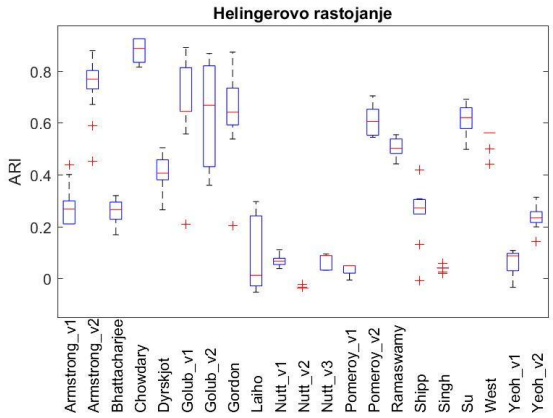
IV. REZULTATI I DISKUSIJA

Prvo će biti prikazani rezultati pomoću boxplota gdje je spektralno klasterovanje primjenjeno nad različitim napravljenim funkcijama u Matlabu koje sadrže kombinaciju rastojanja i normalizovanih ili nenormalizovanih podataka. Boxplotovi otkrivaju koliko se rezultati klasterovanja i stvarne labele koje odgovaraju tipovima raka razlikuju, kada se rezultati svih setova analiziraju zajedno.

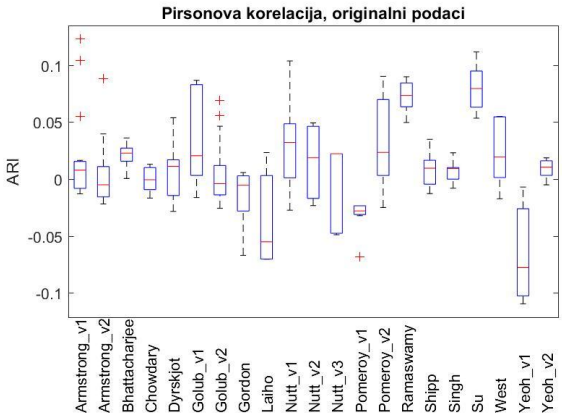
Na slici 1 izvršeno je spektralno klasterovanje korišćenjem Helingerovog rastojanja kao mjere sličnosti. U ovoj bazi izdvaja se set *Chowdary* gdje je vrijednost ARI indeksa najveći, odnosno u većini slučajeva se poklapaju stvarne labele sa labelama dobijene klasterovanjem. Set podataka *Golub v2* i pored toga što ima veliku vrijednost indeksa, opseg vrijednosti koji uzima značajno je veliki i podaci osciluju. Dobra strana ovog seta je što ne sadrži outliere, dok kod većine drugih to nije slučaj.

Normalizacija podataka po vrsti pokazala se kao bolji izbor od normalizacije podataka po koloni (Slika 2). Primjeri su set *Armstrong v2 i Chowdary* gdje je i kod z i kod range normalizacije primjećena konvergencija ka stvarnoj particiji, dok poklapanje stvarnih labela i labela dobijenih klasterovanjem kod normalizovanih podataka po koloni skoro i da ne postoji.

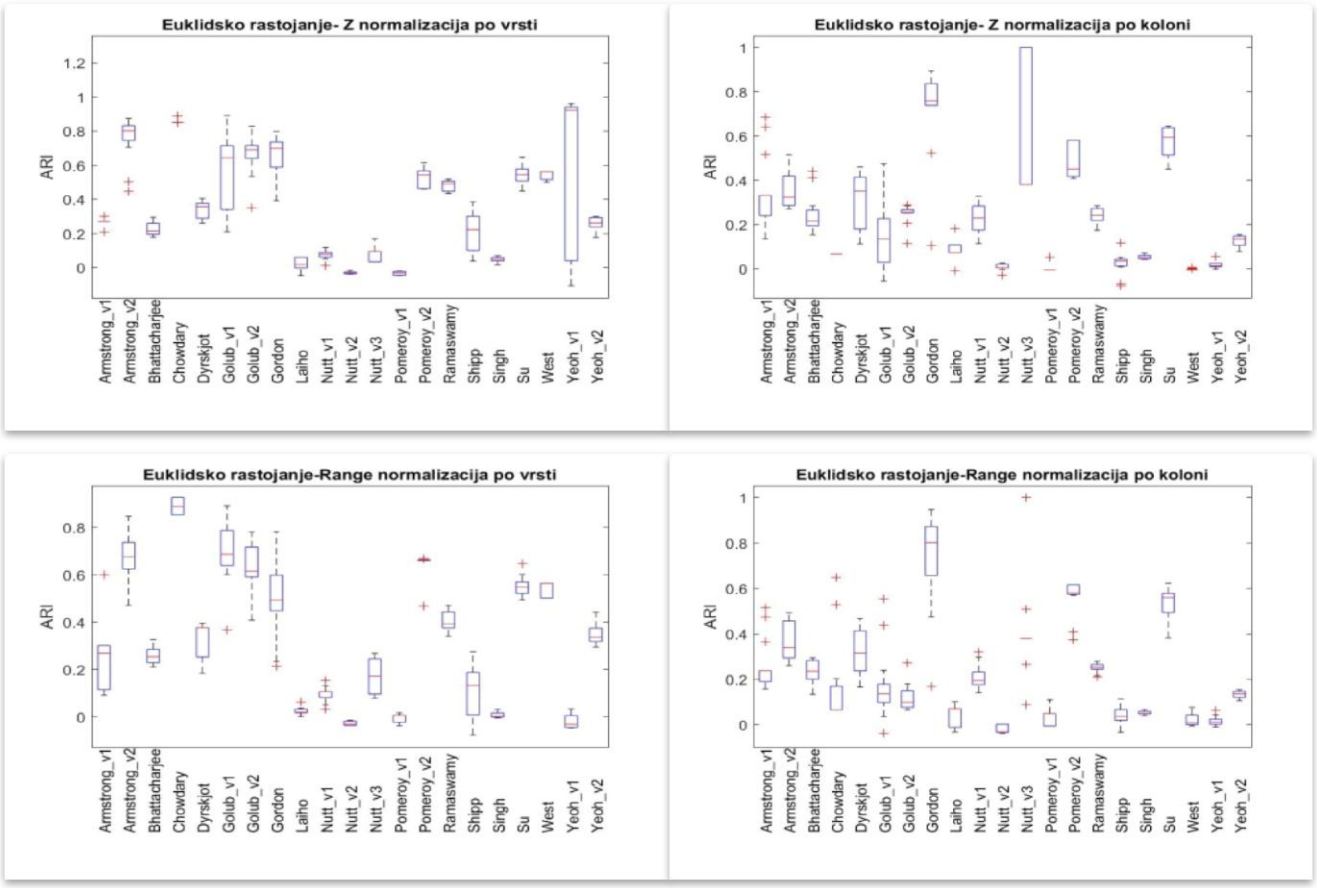
Pirsonov koeficijent korelacije se pokazao kao veoma loš izbor za ovaj algoritam klasterovanja. Uzima jako male vrijednosti ARI-ja, maksimalno 0.1, a u velikoj mjeri su prisutne i negativne vrijednosti (Slika 3).



Sl. 1. Boxplotovi za ARI vrijednosti za čitav *Affymetrix* set podataka kada se koristi Helingerovo rastojanje kao mjera sličnosti u spektralnom klasterovanju



Sl. 3. Boxplotovi za ARI vrijednosti za čitav *Affymetrix* set podataka kada se koristi Pirsonova korelacija kao mjera sličnosti u spektralnom klasterovanju



Sl. 2. Boxplotovi za ARI vrijednosti za čitav *Affymetrix* set podataka kada se koristi Euklidsko rastojanje kao mjera sličnosti u spektralnom klasterovanju nad z-normalizovanim i range normalizovanim podacima

Radi bolje preglednosti, na slici 4. je izvršeno poređenje korišćenih funkcija za svaki set podataka upotrebom ugrađene Matlab funkcije *bar*. Slika se sastoji iz 4 manje slike u kojoj se nalazi po 5 setova podataka, respektivno.

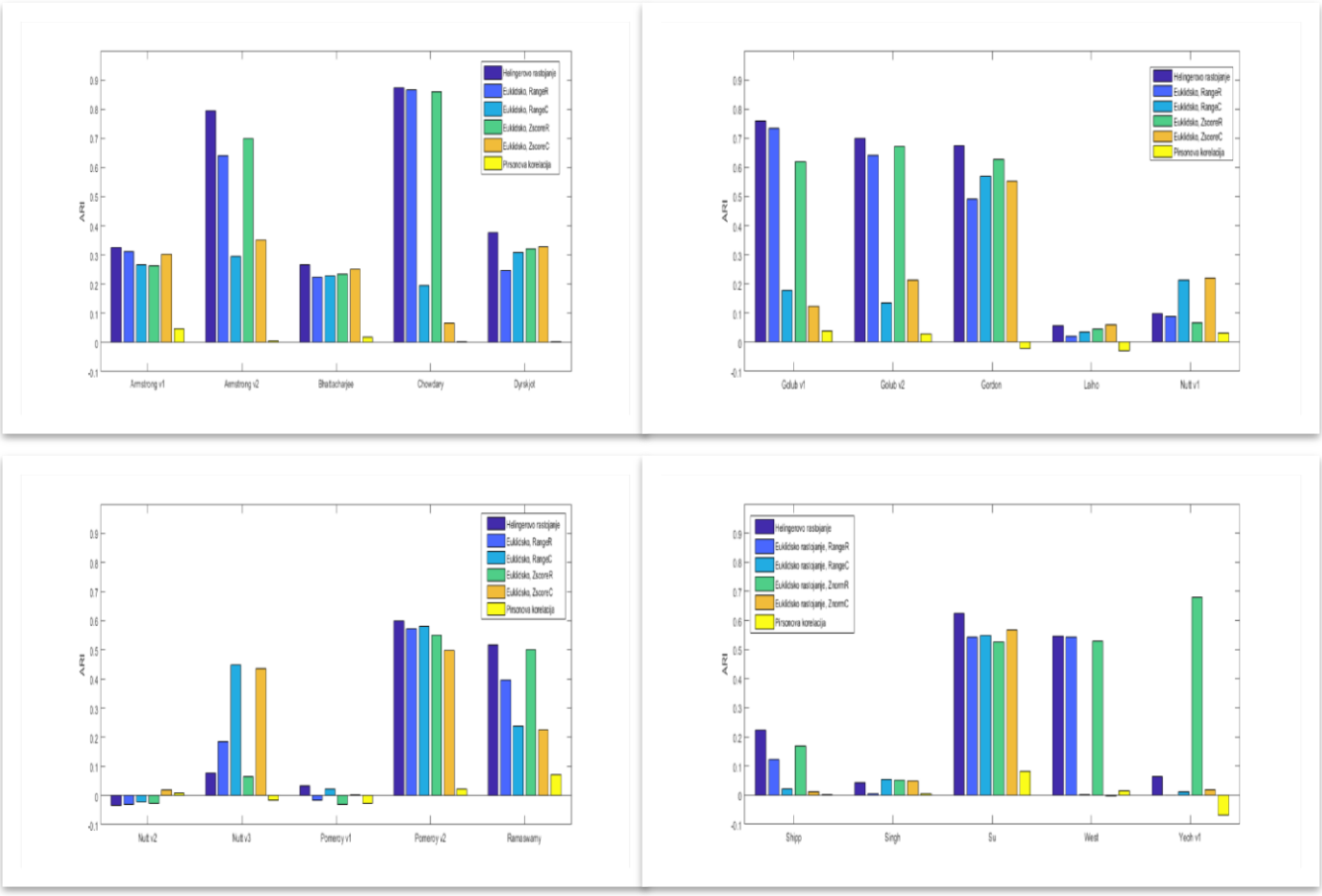
Stubići obojeni najtamnijom nijansom plave boje predstavljaju rezultate spektralnog klasterovanja vršeno Helingerovim rastojanjem, svijetlija nijansa plave odgovara Euklidskom rastojanju i range normalizovanim podacima, a najsvjetlija nijansa prikazuje rezultate klasterovanja vršeno Euklidskim rastojanjem i range normalizacijom podataka po koloni. Stubići zelene boje odgovaraju funkciji koja kombinuje Euklidsko rastojanje i z-normalizaciju po vrsti, dok stubići narandžaste boje prikazuju rezultate dobijene kombinacijom Euklidskog rastojanja i z-normalizacije podataka po koloni. Žutom bojom prikazana je Pirsonova korelacija.

Ne analizirajući *Laiho, Nutt v1, Nutt v2, Pomeroy v1 i Singh* setove podataka, koji uzimaju veoma male vrijednosti ARI indeksa, svaki drugi set daje različite rezultate za zadane funkcije.

Ukoliko se u Matlabu definišu funkcije Helingerovog rastojanja, Euklidskog rastojanja nad z i range normalizovanim podacima po vrsti, i pozove funkcija spektralnog klasterovanja za svaku od njih, zaključuje se da će se dobiti najbolji rezultati. U poređenju sa ostalim funkcijama, ovi rezultati su sa najvećom vrijednošću ARI-ja, što rezultuje skoro pa potpunim slaganjem particija.

Takođe, kod većeg broja setova primjećuje se blaga prednost Helingerovog rastojanja kao najbolje mjere sličnosti, dok kod *Yeoh v1* seta ubjedljivo se izdvaja Euklidsko rastojanje izvršeno nad range normalizovanim podacima po vrsti.

Kao što je već utvrđeno, Pirsonov koeficijent korelacije ne igra nikakvu ulogu u metodi spektralnog klasterovanja.



Slika 4. Različita vrijednost ARI indeksa za svaku korišćenu funkciju primjenjenu nad *Affymetrix* setom podataka