Spektralne metode grupiranja podataka

Nemčić, Josipa

Sveučilište J.J. Strossmayera u Osijeku
Odjel za matematiku
Sveučilišni diplomski studij matematike i računarstva

Josipa Nemčić

Spektralne metode grupiranja podataka

Diplomski rad

Osijek, 2017.

Sveučilište J.J. Strossmayera u Osijeku
Odjel za matematiku
Sveučilišni diplomski studij matematike i računarstva

Josipa Nemčić

Spektralne metode grupiranja podataka

Diplomski rad

Mentor: doc. dr. sc. Snježana Majstorović

Osijek, 2017.

Sadržaj

1 Uvod

Grupiranje ili klasteriranje je postupak podjele nekog skupa podataka na najčešće unaprijed zadan broj grupa ili klastera tako da su podatci koji se nalaze unutar pojedine grupe što sličniji, a podatci koji pripadaju različitim grupama su najmanje slični. Možemo reći i da se podatci grupiraju tako da su unutar pojedine grupe podatci minimalno međusobno udaljeni, dok je udaljenost podataka koji se nalaze u različitim grupama što je moguće veća. Način na koji računamo sličnost, odnosno udaljenost među podatcima ovisi o vrsti podataka, odnosno o problemu kojeg rješavamo.

Problem grupiranja podataka prisutan je u gotovo svakoj znanstvenoj disciplini prilikom analize eksperimentalnih podataka čija nam struktura u početku nije poznata. Upravo zbog toga postoje brojni algoritmi koji problem grupiranja nastoje riješiti što učinkovitije.

U općenitom smislu, algoritme za grupiranje podataka možemo podijeliti u dvije skupine: hijerarhijske i particijske [9]. Hijerarhijski algoritmi rekurzivno pronalaze grupe i to ili u aglomerativnom, odnosno sakupljačkom smislu, gdje se u početku svaki podatak smjesti u jednu grupu, a onda se na temelju izmjerenih sličnosti među podatcima, manje grupe spajaju u veće, ili pak u razdvajajućem smislu, gdje se najprije svi podatci smjeste u jednu grupu, a onda se oni dalje rekurzivno dijele u manje grupe.

Za razliku od hijerarhijskih algoritama, particijski algoritmi istovremeno pronalaze odgovarajuće grupe zadanog skupa podataka.

Najpoznatiji hijerarhijski algoritmi su *single-link*, *complete-link*, *average-link* i *Wardov algoritam* [7].

Particijske metode možemo podijeliti u dvije skupine: *hard* grupiranje, gdje svaki podatak može pripadati jednoj i samo jednoj grupi, te *soft* grupiranje gdje svaki podatak pripada svakoj grupi 'do određene mjere'.

Najpoznatiji algoritmi za *soft* grupiranje podataka su *fuzzy k-means* [1], zatim *Expectation Maximization* algoritam [3], *smooth k-means* algoritam koji se bazira na Euklidskoj l_2-normi [12] ili na l_1-normi [19] itd.

Najpoznatiji i zbog jednostavnosti najpopularniji particijski algoritam je *k-means algoritam*.

U ovom radu ćemo predstaviti *spektralno grupiranje podataka*. Ono podrazumijeva široku klasu metoda za grupiranje koje koriste pažljivo izabrane objekte spektralne

teorije grafova. Spektralno grupiranje omogućuje identifikaciju grupa pomoću spektralnih svojstava grafova, a ključna ideja leži u promjeni reprezentacije originalnih podataka. Nova reprezentacija podataka je takva da i najjednostavniji algoritmi, poput npr. k-means algoritma, mogu bez poteškoća pravilno identificirati određene grupe podataka.

Ideja spektralnog grupiranja se prvi put pojavila sedamdesetih godina prošlog stoljeća, ali je vrlo brzo pala u zaborav. Tek 2000-ih godina spektralne metode se počinju razvijati i sve češće primjenjivati u praksi. Njihova popularnost leži u činjenici da uz pažljivu primjenju daju jako dobre rezultate. Premda je teorijski apsekt spektralnog grupiranja dosta kompliciran, sama implementacija metode je vrlo jednostavna.

Ovaj rad je većinskim dijelom baziran na literaturi o spektralnom grupiranju čija je autorica Ulrike von Luxburg [14]. Rad počinje terminologijom vezanom za grafove, a posebno se obrađuju i uspoređuju grafovi sličnosti. Zatim se definira pojam reza grafa te se precizno definiraju i opisuju tipovi Laplaceove matrice koji su ključni matematički objekti za spektralno grupiranje podataka. Detaljno su objašnjenje metode spektralnog grupiranja, a njihova učinkovitost je pokazana na raznim sintetičkim i realnim primjerima, uključujući primjenu na problem segmentacije slike.

2 Grafovi

Uz pretpostavku da imamo skup sastavljen od određenog broja podataka o kojima znamo malo ili gotovo ništa, prvi korak u analizi je proučavanje njihovih međusobnih odnosa, tj. sličnosti. Podatci ne moraju biti numerički, oni mogu biti objekti bilo kojeg tipa. Pretpostavimo da znamo način na koji ćemo utvrditi odnose između svaka dva podatka danog skupa. U tom slučaju podatcima možemo pridružiti specijalan graf koji se zove *graf sličnosti*, a uz pomoć kojeg ćemo grupirati podatke u izvjestan broj grupa. Konstrukcija grafa je takva da svaki podatak predstavlja jedan njegov vrh, a ako su dva podatka na neki način slična, pridružene vrhove ćemo spojiti bridom koji će nositi informaciju o mjeri njihove sličnosti. Čak i kad se radi o isključivo numeričkim podatcima, taj način razmišljanja nam bitno olakšava shvaćanje problema i pomaže u analizi.

Najprije ćemo definirati neke pojmove iz teorije grafova, a zatim ćemo definirati Laplaceovu matricu grafa te navesti njena svojstva koja su ključna za spektralno grupiranje. Više o grafovima čitatelj može pogledati u [2].

2.1 Osnovni pojmovi i notacija iz teorije grafova

Definicija 2.1.1 (Graf)**.** *Graf G je uređena trojka $G = (V(G), E(G), \psi_G)$ koja se sastoji od nepraznog skupa $V = V(G)$, čije elemente zovemo vrhovima od G, skupa $E = E(G)$ disjunktnog sa $V(G)$, čije elemente zovemo bridovima od G i funkcije incidencije ψ_G koja svakom bridu od G pridružuje neuređeni par (ne nužno različitih) vrhova od G.*

Vrhove grafa najčešće označavamo s u, v, w itd., a bridove grafa s e, f, g itd. Ako brid e spaja vrhove u i v, onda pišemo $e = uv$ ili $e = vu$. Za graf G kažemo da je konačan ako su $V(G)$ i $E(G)$ konačni skupovi. Ako su neka dva vrha u i v grafa G spojena s dva ili više bridova, onda kažemo da postoji *višestruki brid* između tih vrhova, a graf G zovemo *multigrafom*. Brid koji spaja vrh sa samim sobom zove se *petlja*. Graf u kojemu nema ni petlji ni višestrukih bridova zove se *jednostavan graf*. *Podgraf* grafa G je graf nastao uklanjanjem određenih bridova ili vrhova iz G. *Razapinjujući podgraf* H grafa G je onaj podgraf od G za koji vrijedi $V(G) = V(H)$.

Definicija 2.1.2 (Potpun graf)**.** *Potpun graf je jednostavan graf u kojemu je svaki par vrhova spojen bridom.*

Ciklus C duljine k u grafu G je netrivijalan konačan niz $C = v_0e_1v_1e_2 \dots e_{k-1}v_{k-1}e_kv_k$ u kojemu se naizmjenično pojavljuju međusobno različiti vrhovi i međusobno različiti bridovi i pritom vrijedi $e_i = v_{i-1}v_i$, $i = 1, \dots, k$, i $v_0 = v_k$.

Definicija 2.1.3 (Usmjereni graf)**.** *Usmjereni graf ili digraf D je uređena trojka $(V(D), A(D), \psi_D)$ koja se sastoji od nepraznog skupa $V(D)$ vrhova, skupa $A(D)$ lukova (ili usmjerenih bridova) i funkcije incidencije ψ_D koja svakom luku a pridružuje uređeni par (ne nužno različitih) vrhova u, v spojenih sa a. Vrh u je početni, a v krajnji vrh luka a.*

Definicija 2.1.4 (Težinski graf)**.** *Težinski graf G je graf čijim su bridovima pridruženi neki realni brojevi, tj. postoji težinska funkcija $w : E(G) \to \mathbb{R}$ pri čemu broj $w(e)$ zovemo težinom brida $e \in E(G)$.*

Uglavnom se u primjenama koriste nenegativne težinske funkcije, tj. one funkcije w za koje vrijedi $w : E \to \mathbb{R}_0^+$.
U nastavku ćemo vrhove nekog grafa G označavati s $v_1, v_2, \dots, v_n$, pri čemu je n broj vrhova u G, brid između vrhova v_i i v_j ćemo označavati s e_{ij}, a težinu brida v_iv_j ćemo označavati s w_{ij}, $i, j = 1, \dots, n$.

Definicija 2.1.5 (Težinska matrica susjedstva)**.** *Težinska matrica susjedstva $W = (w_{ij})_{i,j=1,\dots,n}$ grafa G je kvadratna matrica reda n koja na (i, j)-om mjestu sadrži težinu w_{ij} brida e_{ij} grafa G.*

Ako vrhovi v_k i v_l nisu spojeni bridom, onda stavljamo $w_{kl} = 0$. Za neusmjerene grafove matrica W je simetrična, tj. vrijedi $W^\tau = W$.
U ovome radu ćemo koristiti isključivo neusmjerene težinske grafove s nenegativnim težinama pa će se u daljnjem tekstu sve definicije i tvrdnje odnositi na takve tipove grafova.

Definicija 2.1.6 (Stupanj vrha)**.** *Neka je v_i, $i \in \{1, \dots, n\}$, vrh grafa G. Sumu težina svih bridova koji izlaze iz vrha v_i nazivamo stupnjem vrha v_i i označavamo s d_i. Pišemo*

$$d_i := \sum_{j=1}^{n} w_{ij}.$$

Definicija 2.1.7 (Matrica stupnjeva)**.** *Dijagonalnu matricu $D = diag(d_1, \dots, d_n)$ koja na dijagonali sadrži stupnjeve $d_1, \dots, d_n$ vrhova, redom $v_1, \dots, v_n$ grafa G, nazivamo matricom stupnjeva grafa G.*

Da pojednostavnimo notaciju, u nastavku ćemo umjesto $\{v_i | v_i \in A\}$, gdje je A neki podskup skupa vrhova V, pisati: $i \in A$. Komplement skupa A, $A^c = V \backslash A$ ćemo označavati s $\overline{A}$.
Neka je $A \subset V$ neki podskup skupa vrhova V grafa G. Sa $|A|$ ćemo označiti broj vrhova u skupu A, a sa $vol(A)$ sumu stupnjeva svih vrhova koji pripadaju skupu A, odnosno

$$vol(A) = \sum_{i \in A} d_i.$$

U grafu G mogu postojati neki disjunktni podskupovi A i B skupa vrhova V između kojih ne postoji niti jedan brid. Za takav graf kažemo da je *nepovezan*, a njegove dijelove sa skupovima vrhova A i B zovemo *komponentama povezanosti* grafa G. Nepovezani grafovi su posebno zanimljivi kod problema grupiranja jer intuitivno, baš ti nepovezani dijelovi grafa, odnosno komponente povezanosti predstavljaju grupe podataka.
Za graf kažemo da je *povezan* ako nije nepovezan.

Definicija 2.1.8 (Indikator vektor)**.** *Neka je $A \subset V$ podskup skupa vrhova grafa G i neka je $\overline{A}$ njegov komplement. Vektor $s_A = (s_1, \ldots, s_n)^T \in \mathbb{R}^n$ definiran s:*

$$s_i = \begin{cases} 1, & \text{ako je } v_i \in A \\ 0, & \text{ako je } v_i \in \overline{A} \end{cases}$$

nazivamo indikator vektor skupa A.

2.2 Grafovi sličnosti

Pojam susjedstva obično koristimo kada govorimo o vrhovima nekog grafa, dok se pojam sličnosti odnosi na podatke. Obzirom da ćemo podatke poistovjetiti sa vrhovima grafa, a njihovu sličnost sa težinama bridovima između vrhova, ta dva pojma smatrati ćemo ekvivalentnima. U nastavku ćemo definirati razne tipove grafa sličnosti, tj. grafa koji služi za reprezentaciju skupa podataka, a ima glavnu ulogu u metodi spektralnog grupiranja.

Definicija 2.2.1 (Graf sličnosti skupa podataka)**.** *Neka je $D = \{x_1, \ldots, x_n\}$ zadani skup podataka i neka su sa s_{ij}, $i, j = \{1, \ldots, n\}$ dane sličnosti među podatcima x_i i*

x_j.
Konstruiramo težinski graf G sa skupom vrhova $V = \{v_1, v_2, \ldots, v_n\}$ tako da svaki vrh predstavlja jedan podatak, tj. $v_i = x_i$, $i = 1, \ldots, n$, a sličnosti s_{ij} među podatcima x_i i x_j su predstavljene težinama bridova koji ih povezuju, odnosno $w_{ij} = s_{ij}$ $\forall i, j = 1, \ldots, n$.
Graf G nazivamo grafom sličnosti danog skupa podataka.

Postoji više različitih grafova sličnosti, a svaki od njih sadrži određene parametre koji odlučuju o tome koji će vrhovi biti spojeni bridom, a koji ne, odnosno hoće li rezultirajući graf biti povezan i koliko jako. Mi ćemo se usredotočiti samo na one tipove koji se najčešće koriste u primjenama.

U općenitom slučaju, konstruiranje grafa sličnosti nije nimalo trivijalan zadatak. Eksperimenti pokazuju da je spektralno grupiranje vrlo osjetljivo na izbor grafa sličnosti i njegovih parametara pa tu trebamo biti vrlo oprezni.
Graf sličnosti nije moguće konstruirati ako nemamo jasno definiranu funkciju koja svakom paru podataka računa sličnost. Zato ćemo najprije reći nešto o funkcijama sličnosti.

2.2.1 Funkcija sličnosti

Literatura o grupiranju podataka nudi mnogo različitih funkcija sličnosti, a njihov izbor ovisi o vrsti podataka koje želimo uspoređivati. Mi ćemo sličnost među podatcima uglavnom povezivati sa njihovom međusobnom udaljenošću obzirom da ćemo raditi sa numeričkim podatcima. tj. podatcima iz Euklidskog prostora $\mathbb{R}^d$. Funkcija sličnosti i funkcija udaljenosti su u određenom smislu u inverznom odnosu. Ako su podatci jako slični, to znači da im je vrijednost funkcije sličnosti velika, ali to onda znači da su ti podatci ujedno i bliži pa im je međusobna udaljenost mala. Ako je udaljenost među podatcima velika, onda je sličnost između njih svakako manja. Zato kao funkciju sličnosti možemo uzeti, primjerice, recipročnu vrijednost kvadrata udaljenosti. Međutim, kada su podatci jako blizu jedan drugome, takva funkcija postaje problematična. *Gaussova funkcija sličnosti* se pokazala kao vrlo dobar izbor u modeliranju lokalnog susjedstva podataka iz $\mathbb{R}^d$. Definirana je s

$$s_{ij} = e^{-\frac{d(x_i, x_j)^2}{2\sigma^2}},$$

pri čemu je $d(x_i, x_j)$ udaljenost među podatcima x_i i x_j. Obično za d uzimamo Euklidsku udaljenost. Parametar σ određuje veličinu lokalnog susjedstva i mora biti

pažljivo izabran.

2.2.2 Graf ϵ-susjedstva

Graf ϵ-susjedstva konstruiramo tako da spojimo bridom one vrhove v_i i v_j čija je međusobna udaljenost manja od ϵ, odnosno za koje vrijedi: $d(v_i, v_j) < \epsilon$, gdje $\epsilon > 0$. Udaljenosti spojenih vrhova u grafu ϵ-susjedstva su uglavnom numerički vrlo slične za većinu ϵ parametara pa dodjeljivanje težina postojećim bridovima neće pružiti dodatne informacije o odnosima između podataka. Iz tog razloga graf ϵ-susjedstva obično nema definirane težine bridova, tj. težina svakog brida jednaka je jedan. Optimalan parametar ϵ nije lako izabrati. Ako želimo da graf bude povezan, onda biramo najmanji ϵ tako da odgovara duljini najduljeg brida u minimalnom razapinjujućem stablu potpunog grafa, tj. u onom razapinjujućem podgrafu koji je povezan i ne sadrži cikluse, a ima svojstvo da mu je suma težina svih bridova najmanja moguća. Problem sa ovakvim pristupom nastaje kada postoje takozvani stršeći podatci (eng. *outliers*), tj. oni podatci koji su jako udaljeni od većine preostalih podataka i uglavnom se ne uzimaju u obzir. U tom slučaju će ϵ biti jako velik pa će se dogoditi da su i stršeći podatci povezani sa preostalim podatcima. Još jedan problem može nastati ako se podatci nalaze u nekoliko gustih područja čija je udaljenost vrlo velika. I tada će ϵ biti prevelik da bi dao uvid u najznačajnije dijelove skupa podataka. Ako bismo pak smanjili ϵ, onda dobiveni graf ne bi bio povezan. Možemo zaključiti da se upotreba ovog grafa ne preporučuje u situacijama kada se udaljenosti među nekim podatcima bitno razlikuju od udaljenosti među nekim drugim podatcima danog skupa.

2.2.3 Graf k-najbližih susjeda

Kod konstrukcije grafa k-najbližih susjeda spajamo bridom vrhove v_i i v_j ako je vrh v_j jedan od k najbližih susjeda vrha v_i, pri čemu je k unaprijed zadan prirodan broj. Ovakva konstrukcija daje usmjereni graf obzirom da u tom slučaju v_i općenito ne mora biti jedan od k-najbližih susjeda vrha v_j.
To nam ne predstavlja problem jer graf možemo vrlo lako pretvoriti u neusmjereni i to na dva načina:

1. spojimo bridom vrhove v_i i v_j ako je v_j jedan od k najbližih susjeda vrha v_i ili je v_i jedan od k najbližih susjeda vrha v_j;

2. spojimo vrhove v_i i v_j ako je v_j jedan od k najbližih susjeda vrha v_i i ako je v_i jedan od k najbližih susjeda vrha v_j.

U prvom slučaju govorimo o *grafu k-najbližih susjeda*, a u drugom slučaju govorimo o *grafu uzajamnih k-najbližih susjeda*. U oba slučaja grafovi su težinski, a težine bridova su jednake sličnostima podataka-vrhova spojenih tim bridovima. I ovdje je prilično teško odrediti optimalan k. On se obično bira tako da rezultirajući graf bude povezan ili da sadrži znatno manje komponenata povezanosti od broja grupa u koje želimo podijeliti zadani skup podataka.
Graf k-najbližih susjeda bez poteškoća povezuje podatke koji imaju različit raspon udaljenosti, što može biti vrlo korisno. Može biti nepovezan i to u slučaju kada postoje izrazito gusta područja podataka čija je međusobna udaljenost znatno velika.
Graf uzajamnih k-najbližih susjeda ne povezuje podatke koji se nalaze u područjima različitih gustoća. To je dobro jedino ako takva područja upućuju na tražene grupe podataka. Za fiksan k, ovaj graf ima puno manje bridova nego graf k-najbližih susjeda pa bi optimalan k morao biti znatno veći od onog k kojeg bi imao graf k-najbližih susjeda. Graf uzajamnih k-najbližih susjeda je najpogodniji za identifikaciju grupa podataka koji imaju različitu gustoću.

2.2.4 Potpun graf sličnosti

Potpun graf sličnosti konstruiramo tako da spojimo sve vrhove v_i i v_j koji imaju pozitivnu sličnost, odnosno za koje vrijedi: $s_{ij} > 0$. I ovdje se radi o težinskom grafu. Težine bridova su jednake vrijednostima sličnosti među podatcima, tj. $w_{ij} = s_{ij}$. Ovakav graf je dobar izbor jedino ako sama funkcija sličnosti dobro modelira lokalno susjedstvo među podatcima. Ako je tako, onda su podatci iz lokalnog susjedstva povezani bridovima relativno velikih težina, dok bridovi između udaljenijih podataka imaju pozitivne, ali zanemarive težine.
Obično se u slučaju ovakvog grafa koristi već spomenuta Gaussova funkcija sličnosti, a njen parametar σ kontrolira širinu susjedstva, odnosno, ima sličnu ulogu kao parametar ϵ kod grafa ϵ-susjedstva.
Glavni nedostatak ovog tipa grafa je u tome što odgovarajuća matrica susjedstva nije rijetka. Ne preporučuje se njegova upotreba ako je skup podataka jako velik.

2.3 Rez grafa

U prethodnom odjeljku smo objasnili kako pomoću grafa sličnosti možemo reprezentirati podatke i, što je još važnije, reprezentirati njihove međusobne odnose. Vidjeli smo da odabir grafa sličnosti i pripadnog parametra nije trivijalan posao, prvenstveno jer je moguće da za konkretan graf sličnosti jedna vrijednost parametra rezultira povezanim grafom, a neka druga vrijednost parametra može rezultirati nepovezanim grafom.
Kada bi graf sličnosti imao više komponenata povezanosti od zadanog broja grupa, onda bi metode spektralnog grupiranja na trivijalan način identificirale grupe: one bi odgovarale komponentama povezanosti grafa. To ima smisla jedino kada znamo da su to one grupe koje želimo identificirati. Obzirom da spektralne metode rade sa općenitim skupovima podataka o kojima u početku ne znamo puno, ne možemo biti sigurni da su komponente povezanosti koje ovise o tipu grafa sličnosti i njegovim parametrima točno one grupe koje tražimo. Zato obično težimo konstrukciji grafa sličnosti koji će biti povezan ili koji će imati vrlo malen broj komponenata povezanosti.
Problem particioniranja grafa star je više od 60 godina, a glavna ideja je podijeliti vrhove grafa u neprazne i međusobno disjunktne grupe tako da su težine bridova koji spajaju vrhove različitih grupa najmanje moguće, odnosno težine bridova između vrhova koji se nalaze u istoj grupi su što je moguće veće. Dakle, osnovna ideja particioniranja grafa leži u tome da želimo izdvojiti grupe u kojima se nalaze vrlo slični vrhovi.

Definicija 2.3.1 (Rez)**.** *Rez $R = (S, T)$ grafa G je particija skupa vrhova $V(G)$ u dva podskupa S i T.*

Ova se definicija može poopćiti pa imamo:

Definicija 2.3.2 (Rez)**.** *K-rez $(A_1, \ldots, A_k)$, $k \geq 2$, grafa G je particija skupa vrhova $V(G)$ u podskupove $A_1, \ldots, A_k$.*

Svaki k-rez definira skup svih bridova koji imaju jedan kraj u jednom podskupu, a drugi kraj u drugom podskupu k-particije. U literaturi se osnovni problem particioniranja grafa zove MIN-CUT problem, a ideja je pronaći onaj rez u danom grafu čija je suma težina bridova najmanja moguća. Analogno se definira MIN k-CUT problem gdje je cilj particionirati skup vrhova grafa u k-podskupova tako da za svaki par takvih podskupova odgovarajući rez ima najmanju moguću sumu težina

bridova. Ovakav pristup koriste i metode spektralnog grupiranja. Sada ćemo uvesti neke osnovne definicije vezane za problem minimalnog k-reza.

Definicija 2.3.3. *Neka je zadan graf G sa skupom vrhova V. Za dva ne nužno disjunktna skupa $A, B \subset V$ definiramo broj:*

$$W(A,B) := \sum_{i \in A, j \in B} w_{ij}.$$

Definicija 2.3.4 (Minimalan rez)**.** *Neka je W matrica susjedstva pridružena grafu sličnosti G i neka je $\{A_1, \ldots, A_k\}$, $k \in \mathbb{N}$, particija skupa vrhova $V(G)$. Optimizacijski problem minimizacije broja*

$$cut(A_1, \ldots, A_k) = \frac{1}{2} \sum_{i=1}^{k} W(A_i, \overline{A_i}),$$

po $A_1, \ldots, A_k \subset V$, zovemo problem minimalnog k-reza (eng. min k-cut).

Koeficijent $1/2$ je nužan jer bismo inače težinu svakog brida računali dva puta. Za slučaj $k = 2$ problem se može jednostavno i učinkovito riješiti primjerice *Stoer-Wagner* algoritmom [22]. Ipak, u praksi se najčešće dogodi da je rješenje *min-cut* problema particija u kojoj jedna grupa sadrži samo jedan vrh, a druga grupa sadrži sve preostale vrhove. To nije zadovoljavajući rezultat pa ćemo zato eksplicitno zahtijevati da skupovi $A_1, \ldots, A_k$ budu 'razumno' veliki.
Uveli smo dva načina na koja možemo mjeriti veličinu skupova A_i, $i = 1, \ldots, n$, pa ćemo *min k-cut* problem reformulirati tako da uključuje te mjere.

Definicija 2.3.5 (RatioCut)**.** *Neka je W matrica susjedstva grafa sličnosti G i neka $k \in \mathbb{N}$, $k \geq 2$. Problem minimizacije veličine*

$$RatioCut(A_1, \ldots, A_k) := \frac{1}{2} \sum_{i=1}^{k} \frac{W(A_i, \overline{A_i})}{|A_i|} \tag{1}$$

po svim k-particijama $(A_1, \ldots, A_k)$ skupa $V(G)$ zovemo problem minimalnog razmjernog reza (eng. RatioCut).

Definicija 2.3.6 (Ncut)**.** *Neka je W matrica susjedstva grafa sličnosti G i $k \in \mathbb{N}$, $k \geq 2$. Problem minimizacije veličine*

$$Ncut(A_1, \ldots, A_k) := \frac{1}{2} \sum_{i=1}^{k} \frac{W(A_i, \overline{A_i})}{vol(A_i)} \tag{2}$$

po svim k-particijama $(A_1, \ldots, A_k)$ *skupa* $V(G)$ *zovemo problem minimalnog normaliziranog reza (Ncut skraćeno od eng. normalized cut).*

Oba problema balansiraju grupe tako da prilikom traženja minimalnog reza uzimaju u obzir i veličinu skupova $A_1, \ldots, A_k$. *RatioCut* mjeri veličinu skupa pomoću broja vrhova u njemu, dok Ncut mjeri veličinu preko sume težina svih bridova koji imaju barem jedan kraj u tom skupu.

Rješenja ovih problema svakako daju 'bolje' grupe od onih koje daje obični *min k-cut* problem. Ipak, ovakvi problemi balansiranog particioniranja grafa su NP-teški (vidi: [11]). Radi se o problemima diskretne optimizacije po svim k-particijama konačnog skupa.
Sjetimo se da je broj svih k-particija nekog n-članog skupa jednak Stirlingovom broju druge vrste $\left\{ {n \atop k} \right\}$ kojeg računamo po formuli

$$\left\{ {n \atop k} \right\} = \frac{1}{k!} \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} j^n.$$

Metode spektralnog grupiranja služe za rješavanje relaksiranih *RatioCut* i *Ncut* problema, što ćemo objasniti nešto kasnije. U sljedećem pododjeljku ćemo definirati *Laplaceovu matricu* i njene varijante. Te matrice imaju glavnu ulogu u metodi spektralnog grupiranja.

2.4 Laplaceova matrica i njena svojstva

Svakom grafu možemo pridružiti neku matricu[1]. Primjerice, matrica susjedstva sadrži informacije o tome koliki je broj vrhova u grafu, koji su vrhovi spojeni bridom, a koji ne, kolike su težine bridova između pojedinih vrhova, a lako možemo izračunati i stupanj svakog vrha. Potencije matrice susjedstva daju informacije o broju šetnji između svih parova vrhova. Matrica udaljenosti daje informaciju o tome kolika je težinska udaljenost između svaka dva vrha u grafu itd.

Laplaceova matrica se prirodno pojavljuje u formulaciji *min-cut* problema. Ona također sadrži osnovne informacije o grafu, tj. daje uvid u njegovu strukturu. Vidjeti

[1]Postoji posebno područje teorije grafova koje proučava matrice pridružene grafovima, a zove se spektralna teorija grafova.

ćemo da je ova matrica ključan matematički objekt kod spektralnog grupiranja. Mi ćemo proučavati tri tipa Laplaceovih matrica: nenormaliziranu Laplaceovu matricu, zatim Laplaceovu matricu slučajnih šetnji te normaliziranu Laplaceovu matricu. Svaka od navedenih matrica biti će direktno povezana sa metodama spektralnog grupiranja.

Neka je G neusmjereni, težinski graf sa nenegativnim težinama bridova, neka mu je $W = (w_{ij})_{i,j=\{1,\dots,n\},n\in\mathbb{N}}$ matrica susjedstva, a D dijagonalna matrica sa stupnjevima d_i vrhova v_i grafa G, $i = 1, \dots, n$.

Definicija 2.4.1 (Nenormalizirana Laplaceova matrica)**.** *Nenormalizirana Laplaceova matrica L definirana je s*

$$L = D - W. \tag{3}$$

Navedimo najznačajnija svojstva matrice L:

Propozicija 2.4.1 (Svojstva nenormalizirane Laplaceove matrice)**.** *Matrica L zadovoljava sljedeća svojstva:*

1. *Za svaki vektor $y \in \mathbb{R}^n$ vrijedi:*

$$y^\tau L y = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij}(y_i - y_j)^2. \tag{4}$$

2. *Matrica L je simetrična i pozitivno semidefinitna.*

3. *Najmanja svojstvena vrijednost od L je 0, a odgovarajući svojstveni vektor je jedinični vektor $\mathbb{1}$.*

4. *Matrica L ima n nenegativnih, realnih svojstvenih vrijednosti:*

$$0 = \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n.$$

Dokaz.

1. Raspisati ćemo $y^\tau Ly$ pomoću W i D:

$$\begin{aligned}
y^\tau Ly &= y^\tau Dy - y^\tau Wy \\
&= \sum_{i=1}^{n} y_i^2 d_i - \sum_{i=1}^{n}\sum_{i=1}^{n} y_i y_j w_{ij} \\
&= \frac{1}{2}\left(\sum_{i=1}^{n} y_i^2 d_i - 2\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j w_{ij} + \sum_{j=1}^{n} y_j^2 d_j\right) \\
&\text{Sada koristimo: } d_i = \sum_{j=1}^{n} w_{ij}. \\
&= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} y_i^2 - 2\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j w_{ij} + \sum_{j=1}^{n}\sum_{i=1}^{n} w_{ij} y_j^2 \\
&= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(w_{ij} y_i^2 - 2y_i y_j w_{ij} + w_{ij} y_j^2) \\
&= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - y_j)^2.
\end{aligned}$$

2. Simetričnost matrice L slijedi iz simetričnosti matrica D i W jer je razlika dvije simetrične matrice opet simetrična matrica.
   Pozitivna semidefinitnost slijedi iz prvog svojstva. Zbog nenegativnih težina w_{ij} bridova, vrijedi

$$y^\tau Ly = \frac{1}{2}\sum_{i=j}^{n}\sum_{j=1}^{n} w_{ij}(y_i - y_j)^2 \geq 0$$

   za sve $y \in \mathbb{R}^n$, uz $w_{ij} \geq 0$.

3. Ako pomnožimo L sa jediničnim vektorom $\mathbb{1}$, imamo:

$$L\mathbb{1} = D\mathbb{1} - W\mathbb{1} = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_n \end{bmatrix} - \begin{bmatrix} \sum_{j=1}^{n} w_{1j} \\ \sum_{j=1}^{n} w_{2j} \\ \vdots \\ \sum_{j=1}^{n} w_{nj} \end{bmatrix} = 0.$$

   Dobili smo matričnu jednadžbu: $L\mathbb{1} = 0 \cdot \mathbb{1} = 0$, iz čega proizlazi da je 0 svojstvena vrijednost od L sa svojstvenim vektorom $\mathbb{1}$.

4. Dokazali smo da je L pozitivno semidefinitna i da ima svojstvenu vrijednost 0. Zbog pozitivne semidefinitnosti, L ne može imati negativne svojstvene vrijednosti, pa je 0 njena najmanja svojstvena vrijednost (vidi: [23]).

□

Važno svojstvo matrice L koje će nam trebati za spektralno grupiranje je dano sljedećom propozicijom:

Propozicija 2.4.2 (Broj komponenata povezanosti i spektar od L)**.** *Neka je G neusmjeren težinski graf s nenegativnim težinama bridova. Geometrijska kratnost k svojstvene vrijednosti 0 matrice L jednaka je broju komponenata povezanosti $A_1, \ldots, A_k$ grafa G.*
Svojstveni prostor svojstvene vrijednosti 0 je razapet indikator vektorima $\mathbb{1}_{A_1}, \ldots, \mathbb{1}_{A_k}$ tih komponenti.

Dokaz. Prvo ćemo dokazati slučaj kada je $k = 1$, odnosno kada je G povezan graf. Pretpostavimo da je y svojstveni vektor pridružen svojstvenoj vrijednosti 0.
Tada imamo

$$Ly = 0y \Rightarrow y^\tau Ly = y^\tau 0 = 0$$

i dalje iz Propozicije 2.4.1 slijedi:

$$0 = y^\tau Ly = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij}(y_i - y_j)^2.$$

Ako su v_i i v_j spojeni bridom, onda $w_{ij} > 0 \Rightarrow (y_i - y_j)^2 = 0 \Rightarrow y_i = y_j$.
Ako v_i i v_j nisu spojeni bridom, onda $w_{ij} = 0$.
Budući da je G povezan, iz svakog vrha možemo doći do svakog od preostalih vrhova prolazeći kroz neke bridove i/ili vrhove. Zaključujemo da svaki par vrhova mora imati istu vrijednost odgovarajuće komponente indikator vektora, tj. mora vrijediti $y_i = y_j$ za sve v_i, $v_j \in V$.
Dakle, jedinični vektor $\mathbb{1}$ pridružen svojstvenoj vrijednosti 0 od L, sa algebarskom kratnošću[2] 1, je indikator vektor komponente povezanosti grafa G. Time smo dokazali tvrdnju za $k = 1$.

[2]U simetričnim matricama se geometrijska i algebarska kratnost svake svojstvene vrijednosti podudaraju.

Pretpostavimo da imamo $k \geq 2$ komponenti povezanosti $A_1, \ldots, A_k$.
Uz prigodno označavanje vrhova, Laplaceova matrica se može zapisati u blok dijagonalnoj formi na sljedeći način:

$$L = \begin{bmatrix} L_1 & 0 & \ldots & 0 \\ 0 & L_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & L_k \end{bmatrix}.$$

Svaki dijagonalni blok L_i odgovara Laplaceovoj matrici pridruženoj jednoj od k komponenti povezanosti grafa G.
Zbog blok dijagonalnosti, spektar od L je unija spektara od L_i, a svojstveni vektori od L odgovaraju svojstvenim vektorima od L_i sa nulama na pozicijama j za koje $v_j \notin A_i$, gdje je A_i komponenta koja odgovara bloku L_i.
Jer je L_i Laplaceova matrica pridružena komponenti povezanosti A_i, znamo da mora imati jednu svojstvenu vrijednost 0 i odgovarajući svojstveni vektor y gdje $y_i = 1$ kada je $v_i \in A_i$. Slijedi da matrica L ima svojstvenu vrijednost 0 sa algebarskom i geometrijskom kratnošću k, a pridruženi svojstveni vektori su indikator vektori komponenti povezanosti A_i, $i = 1, \ldots, k$.

Pokazali smo da je svojstveni prostor svojstvene vrijednosti 0 matrice L razapet sa k indikator vektora od kojih svaki predstavlja jednu komponentu povezanosti $A_1, \ldots, A_k$. □

U nastavku ćemo definirati dva nova tipa Laplaceove matrice.

Definicija 2.4.2. *Matricu L_{sym} definiranu s*

$$L_{sym} = D^{-1/2} L D^{-1/2} = I - D^{-1/2} W D^{-1/2} \tag{5}$$

zovemo normaliziranom Laplaceovom matricom, dok matricu L_{rw} definiranu s

$$L_{rw} = D^{-1} L = I - D^{-1} W \tag{6}$$

zovemo Laplaceovom matricom slučajnih šetnji.

Prvu matricu označavamo sa L_{sym} jer naglašavamo njeno svojstvo simetričnosti, a naziv i oznaka druge varijante Laplaceove matrice naglašava usku povezanost sa slučajnom šetnjom u grafu.

Valja uočiti da za slučaj grafa G koji ima izolirane vrhove ne možemo pravilno definirati ove matrice jer D nije regularna.
Ipak, ovakav problem možemo riješiti tako da svakom izoliranom vrhu pridružimo petlju čija je težina zanemarivo mala pozitivna vrijednost. To neće utjecati na strukturu grafa pa onda niti na grupiranje skupa vrhova, ali je time osigurana regularnost matrice D, tj. vrijedi $d_{ii} \geq w_{ii} > 0$, iz čega slijedi: $det(D) = \prod_{i=1}^{n} d_{ii} \neq 0$.

Sada ćemo navesti neka svojstva matrica L_{sym} i L_{rw}.

Propozicija 2.4.3. *Matrice L_{sym} i L_{rw} zadovoljavaju sljedeća svojstva:*

1. *Za svaki $y = (y_1, \ldots, y_n)^\tau \in \mathbb{R}^n$ vrijedi:*

$$y^\tau L_{sym} y = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij} \left(\frac{y_i}{\sqrt{d_i}} - \frac{y_j}{\sqrt{d_j}} \right)^2.$$

2. *λ je svojstvena vrijednost od L_{rw} sa pridruženim svojstvenim vektorom u ako i samo ako je λ svojstvena vrijednost od L_{sym} sa pridruženim svojstvenim vektorom $w = D^{1/2}u$.*

3. *λ je svojstvena vrijednost od L_{rw} sa svojstvenim vektorom u ako i samo ako λ i u rješavaju generalizirani svojstveni problem: $Lu = \lambda Du$.*

4. *0 je svojstvena vrijednost od L_{rw} sa pridruženim svojstvenim vektorom $\mathbb{1}$.*
   0 je svojstvena vrijednost od L_{sym} sa pridruženim svojstvenim vektorom $D^{1/2}\mathbb{1}$.

5. *L_{sym} i L_{rw} su pozitivno semidefinitne matrice i imaju n nenegativnih realnih svojstvenih vrijednosti $0 = \lambda_1 \leq \ldots \leq \lambda_n$.*

Dokaz.

1. Raspisujemo $y^\tau L_{sym} y$ pomoću Definicije 2.4.2:

$$
\begin{aligned}
& y^\tau L_{sym} y \\
&\quad = y^\tau I y - y^\tau D^{-1/2} W D^{-1/2} y \\
&\quad = \sum_{i=1}^{n} y_i^2 - (y_1, \ldots, y_n)^\tau \begin{bmatrix} \frac{w_{11}}{d_1} & \frac{w_{12}}{\sqrt{d_1 d_2}} & \cdots & \frac{w_{1n}}{\sqrt{d_1 d_n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{w_{n1}}{\sqrt{d_n d_1}} & \frac{w_{n2}}{\sqrt{d_n d_2}} & \cdots & \frac{w_{nn}}{d_n} \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \\
&\quad = \sum_{i=1}^{n} y_i^2 - (y_1, \ldots, y_n)^\tau \begin{bmatrix} \sum_{j=1}^{n} \frac{y_j w_{1j}}{\sqrt{d_1 d_j}} \\ \vdots \\ \sum_{j=1}^{n} \frac{y_j w_{nj}}{\sqrt{d_n d_j}} \end{bmatrix} \\
&\quad = \sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{y_i y_j w_{ij}}{\sqrt{d_i d_j}} \\
&\quad = \frac{1}{2} \left(\sum_{i=1}^{n} \frac{y_i^2 d_i}{d_i} - 2 \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{y_i y_j w_{ij}}{\sqrt{d_i d_j}} + \sum_{j=1}^{n} \frac{y_j^2 d_j}{d_j} \right) \\
&\quad = \frac{1}{2} \left(\sum_{i=1}^{n} \frac{y_i^2}{d_i} \sum_{j=1}^{n} w_{ij} - 2 \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{y_i y_j w_{ij}}{\sqrt{d_i d_j}} + \sum_{j=1}^{n} \frac{y_j^2}{d_j} \sum_{i=1}^{n} w_{ji} \right) \\
&\quad = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\frac{y_i^2 w_{ij}}{d_i} - 2 \frac{w_{ij} y_i y_j}{\sqrt{d_i d_j}} + \frac{y_j^2 w_{ij}}{d_j} \right) \\
&\quad = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \left(\frac{y_i}{\sqrt{d_i}} - \frac{y_j}{\sqrt{d_j}} \right)^2 .
\end{aligned}
$$

2. Neka je λ svojstvena vrijednost od L_{rw} sa svojstvenom vrijednošću u.

   $L_{rw} u = \lambda u \Leftrightarrow D^{-1} L u = \lambda u$
   $D^{-1/2} D^{-1/2} L u = \lambda u$
   Množimo slijeva s $D^{1/2}$ pa dobivamo:
   $D^{1/2} D^{-1/2} D^{-1/2} L u = \lambda D^{1/2} u$
   Kako je $D^{1/2} D^{-1/2} = I$, imamo: $(D^{-1/2} L D^{-1/2}) D^{1/2} u = \lambda D^{1/2} u$
   $L_{sym} D^{1/2} u = \lambda D^{1/2} u$
   Sada zamijenimo: $w = D^{1/2} u$ i konačno imamo: $L_{sym} w = \lambda w$

   Time smo dokazali da je λ svojstvena vrijednost od L_{sym} sa svojstvenim vektorom $w = D^{1/2} u$.

3. Imamo:

$$\begin{aligned} L_{rw}u &= \lambda u \\ D^{-1}Lu &= \lambda u. \end{aligned}$$

Množenjem zdesna s D dobivamo:

$$Lu = \lambda Du.$$

4. Imamo: $L_{rw} = I - D^{-1}W$. Sada tu matričnu jednadžbu množimo sa jediničnim vektorom zdesna:

$$\begin{aligned} L_{rw}\mathbb{1} &= I\mathbb{1} - D^{-1}W\mathbb{1} \\ &= \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} - \begin{bmatrix} \frac{1}{d_1}\sum_{j=1}^{n} w_{1j} \\ \frac{1}{d_2}\sum_{j=1}^{n} w_{2j} \\ \vdots \\ \frac{1}{d_n}\sum_{j=1}^{n} w_{nj} \end{bmatrix} \end{aligned}$$

Kako je $d_i = \sum_{j=1}^{n} w_{ij}$, imamo:

$$= \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} - \begin{bmatrix} \frac{1}{d_1}d_1 \\ \frac{1}{d_2}d_2 \\ \vdots \\ \frac{1}{d_n}d_n \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = 0$$

Dokazali smo da je 0 svojstvena vrijednost od L_{rw} sa pridruženim svojstvenim vektorom $\mathbb{1}$, jer $L_{rw}\mathbb{1} = \mathbb{1} \cdot 0$.
Iz druge tvrdnje propozicije slijedi da je 0 svojstvena vrijednost i od L_{sym} sa svojstvenim vektorom $D^{1/2}\mathbb{1}$.

5. Pokazali smo da je 0 svojstvena vrijednost od L_{sym}.
Iz svojstva 1 propozicije imamo:

$$y^\tau L_{sym} y = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left(\frac{y_i}{\sqrt{d_i}} - \frac{y_j}{\sqrt{d_j}}\right)^2,$$

što znači da je $y^\tau L_{sym} y \geq 0$ za sve $y \in \mathbb{R}^n$ jer: $\left(\frac{y_i}{\sqrt{d_i}} - \frac{y_j}{\sqrt{d_j}}\right)^2 \geq 0$ i sve težine w_{ij} su nenegativne.

Time smo pokazali semidefinitnost.
Kako je L_{sym} realna i simetrična, ima n realnih svojstvenih vrijednosti koje sve moraju biti veće ili jednake od nule zbog semidefinitnosti.
Iz svojstva 2 propozicije znamo da L_{rw} mora imati iste svojstvene vrijednosti $0 = \lambda_1 \leq \ldots \leq \lambda_n$, iz čega slijedi da je L_{rw} također pozitivno semidefinitna po definiciji.

□

Napomena 2.4.1.

1. *Matrice L i L_{sym} su kongruentne. Kogruentne matrice su one matrice A i B za koje postoji regularna matrica P takva da vrijedi $P^\tau AP = B$. U ovom slučaju ulogu matrice P igra $D^{-1/2}$. Prema Sylvesterovom zakonu inercije, kongruentne matrice imaju jednak broj pozitivnih, negativnih i nula svojstvenih vrijednosti [13].*

2. *Matrice L_{rw} i L_{sym} su slične. Slične matrice su one matrice A i B za koje postoji regularna matrica P takva da vrijedi $B = P^{-1}AP$. U ovom slučaju, ulogu matrice P ima $D^{1/2}$. Slične matrice imaju isti spektar [13].*

Kao i kod nenormalizirane Laplaceove matrice, kratnost svojstvene vrijednosti nula matrica L_{sym} i L_{rw} direktno je povezan sa brojem komponenata povezanosti odgovarajućeg grafa:

Propozicija 2.4.4 (Broj komponenata povezanosti i spektar od L_{sym} i L_{rw})**.** *Neka je G neusmjeren, težinski graf s nenegativnim težinama bridova.*
Tada je geometrijska kratnost k svojstvene vrijednosti 0 matrica L_{rw} i L_{sym} jednaka broju komponenata povezanosti $A_1, \ldots, A_k$ grafa G.
Svojstveni prostor od L_{rw} pridružen svojstvenoj vrijednosti 0 razapet je indikator vektorima $\mathbb{1}_{A_i}$ tih komponenti, dok je svojstveni prostor od L_{sym} pridružen svojstvenoj vrijednosti 0 razapet vektorima $D^{1/2}\mathbb{1}_{A_i}$, $i = 1, \ldots, k$.

Dokaz. Dokaz je analogan dokazu Propozicije 2.4.2 uz korištenje svojstava danih Propozicijom 2.4.3. □

3 Spektralno grupiranje i rez grafa

U ovom poglavlju ćemo objasniti relaksaciju *RatioCut* i *Ncut* problema, uspostaviti vezu sa Laplaceovim matricama L, L_{sym} i L_{rw} te objasniti kako funkcionira metoda spektralnog grupiranja.

3.1 Izvod nenormaliziranog algoritma

Pojam balansiranog reza grafa smo objasnili u pododjeljku 2.3.
U nastavku ćemo objasniti metodu spektralnog grupiranja koja nastaje relaksiranjem problema minimizacije razmjernog reza danog Definicijom 2.3.5. Takvu metodu još zovemo nenormalizirano spektralno grupiranje.
Zbog jednostavnosti, prvo ćemo gledati slučaj particioniranja grafa u dvije grupe, odnosno slučaj $k = 2$.

3.1.1 Aproksimacija razmjernog reza za $k = 2$

Za zadni graf G sa skupom vrhova $V = \{v_1, v_2, \ldots, v_n\}$ treba riješiti sljedeći optimizacijski problem:

$$\min_{A \subset V} RatioCut(A, \overline{A}). \tag{7}$$

Zadanom podskupu $A \subset V$ pridružiti ćemo vektor $y = (y_1, \ldots, y_n)^\tau \in \mathbb{R}^n$ i to na sljedeći način:

$$y_i := \begin{cases} \sqrt{\frac{|\overline{A}|}{|A|}}, & \text{ako } v_i \in A \\ -\sqrt{\frac{|A|}{|\overline{A}|}}, & \text{ako } v_i \in \overline{A} \end{cases} \tag{8}$$

Sljedeća propozicija daje vezu između razmjernog reza i nenormalizirane Laplaceove matrice:

Propozicija 3.1.1. *Neka je L nenormalizirana Laplaceova matrica grafa G. Za proizvoljan neprazan podskup A skupa vrhova V i vektor y definiran s (8) vrijedi:*

$$y^\tau L y = |V|\, RatioCut(A, \overline{A}). \tag{9}$$

Dokaz. Koristeći Propoziciju 2.4.1 imamo:

$$
\begin{aligned}
y^\tau Ly &= \frac{1}{2}\sum_{i=1}^{n} w_{ij}(y_i - y_j)^2 \\
&= \frac{1}{2}\sum_{i\in A, j\in\overline{A}} w_{ij}\left(\sqrt{\frac{|\overline{A}|}{|A|}} + \sqrt{\frac{|A|}{|\overline{A}|}}\right)^2 + \frac{1}{2}\sum_{i\in\overline{A}, j\in A} w_{ij}\left(-\sqrt{\frac{|\overline{A}|}{|A|}} - \sqrt{\frac{|A|}{|\overline{A}|}}\right)^2 \\
&= \frac{1}{2}\sum_{i\in A, j\in\overline{A}} w_{ij}\left(\frac{|\overline{A}|}{|A|} + 2\sqrt{\frac{|\overline{A}|}{|A|}\frac{|A|}{|\overline{A}|}} + \frac{|A|}{|\overline{A}|}\right) + \frac{1}{2}\sum_{i\in\overline{A}, j\in A} w_{ij}\left(\frac{|\overline{A}|}{|A|} + 2\sqrt{\frac{|\overline{A}|}{|A|}\frac{|A|}{|\overline{A}|}} + \frac{|A|}{|\overline{A}|}\right) \\
&= \frac{1}{2}\left(\sum_{i\in A, j\in\overline{A}} w_{ij} + \sum_{i\in\overline{A}, j\in A} w_{ij}\right)\left(\frac{|\overline{A}|}{|A|} + \frac{|A|}{|\overline{A}|} + 2\right) \\
&= cut(A,\overline{A})\left(\frac{|\overline{A}|}{|A|} + \frac{|A|}{|\overline{A}|} + 2\right) \\
&= cut(A,\overline{A})\left(\frac{|A| + |\overline{A}|}{|A|} + \frac{|A| + |\overline{A}|}{|\overline{A}|}\right) \\
&= (|A| + |\overline{A}|)\, cut(A,\overline{A})\left(\frac{1}{|A|} + \frac{1}{|\overline{A}|}\right).
\end{aligned}
$$

Koristeći definiciju razmjernog reza dobivamo:

$$y^\tau Ly = |V|\, RatioCut(A,\overline{A}).$$

□

Vektor y ima bitna svojstva koja će nam trebati za rješavanje optimizacijskog problema.

Propozicija 3.1.2. *Neka je y dan s jednadžbom (8). Tada vrijedi:*

1. $y \perp \mathbb{1}$
2. $\|y\|^2 = n$.

Dokaz.

1. Obzirom da je $y \neq 0$, dovoljno je pokazati da je skalarni produkt vektora y i vektora $\mathbb{1}$ jednak nuli. Imamo

$$\sum_{i=1}^{n} y_i = \sum_{i \in A} \sqrt{\frac{|\overline{A}|}{|A|}} - \sum_{i \in \overline{A}} \sqrt{\frac{|A|}{|\overline{A}|}} = |A|\sqrt{\frac{|\overline{A}|}{|A|}} - |\overline{A}|\sqrt{\frac{|A|}{|\overline{A}|}} = 0$$

   pa je okomitost tih vektora dokazana.

2.
$$\|y\|^2 = \sum_{i=1}^{n} y_i^2 = |A|\frac{|\overline{A}|}{|A|} + |\overline{A}|\frac{|A|}{|\overline{A}|} = |\overline{A}| + |A| = |V| = n.$$

   Time smo dokazali da je norma vektora y jednaka $\sqrt{n}$.

□

Kombinirajući rezultate iz Propozicija 3.1.1 i 3.1.2, problem (7) možemo zapisati u matričnom obliku:

$$\min_{y \in \mathbb{R}^n} y^\tau L y, \quad y \text{ dan jednadžbom (8)}, \, y \perp \mathbb{1}, \, \|y\| = \sqrt{n}.$$

Jasno je da se ovdje radi o diskretnom optimizacijskom problemu jer treba raditi minimizaciju po svim vektorima y čije komponente dolaze iz diskretnog skupa podataka, tj. poprimaju samo dvije moguće vrijednosti. Ranije smo ustanovili da je ovakav problem NP-težak i da ćemo morati raditi relaksaciju koja će olakšati traženje optimalnog y. Relaksacija će biti takva da ćemo dozvoliti da komponente vektora y budu proizvoljni realni brojevi, ali ćemo zadržati uvjet o normi, tj. i dalje ćemo imati uvjet da je $\|y\|^2 = n$. To nas dovodi do relaksiranog optimizacijskog problema:

$$\min_{y \in \mathbb{R}^n} y^\tau L y \text{ uz uvjet } y \perp \mathbb{1}, \, \|y\| = \sqrt{n}, \tag{10}$$

pa u geometrijskom smislu treba minimizirati funkciju $y^\tau L y$, koja je zapravo simetrična pozitivno semidefinitna kvadratna forma, po svim radijvektorima točaka koje se nalaze na sferi smještenoj u $\mathbb{R}^n$ sa polumjerom $\sqrt{n}$ i središtem u ishodištu koordinatnog sustava. Dakle, problem minimizacije se više ne odnosi samo na radijvektore nekih točaka na danoj sferi, već se odnosi na radijvektore svih njenih točaka.
Rješenje ovakvog problema postoji, a dobiti ćemo ga na vrlo jednostavan način, i to korištenjem poznatog rezultata iz linearne algebre, tzv. Rayleigh-Ritz teorema [8].

Teorem 3.1.1 (Rayleigh-Ritz)**.** *Neka je $A \in \mathbb{R}^{n\times n}$ simetrična matrica sa svojstvenim vrijednostima $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$ i neka je $(u_1, u_2, \ldots, u_n)$ ortonormirana baza svojstvenih vektora od A, gdje je u_i jedinični svojstveni vektor pridružen svojstvenoj vrijednosti λ_i, $i = 1, \ldots, n$.*
Tada vrijedi:

1. $$\min_{x\neq 0} \frac{x^\tau Ax}{x^\tau x} = \lambda_1$$
   i minimum se postiže u $x = u_1$.

2. $$\min_{x\neq 0, x\in\{u_1,\ldots,u_{i-1}\}^\perp} \frac{x^\tau Ax}{x^\tau x} = \lambda_i$$
   i minimum se postiže u $x = u_i$ gdje je $2 \leq i \leq n$.

Dokaz. Vidi [8]. □

Propozicija 3.1.3. *Rješenje problema (10) je dano svojstvenim vektorom pridruženim drugoj po redu najmanjoj svojstvenoj vrijednosti matrice L.*

Dokaz. Iz Propozicije 2.4.1 znamo da je L simetrična matrica. To znači da možemo napraviti spektralnu dekompoziciju od L i time naći n svojstvenih vektora $u_1, \ldots, u_n$ koji su međusobno ortogonalni i imaju normu $\sqrt{n}$. Dakle, vrijedi $u_i \perp \mathbb{1}$ za sve $i = 2, \ldots, n$ i $\|u_i\| = \sqrt{n}$ za sve $i = 1, \ldots, n$. Drugim riječima, svi svojstveni vektori osim u_1 ispunjavaju uvjet problema (10).
Stoga je rješenje problema (10) dano svojstvenim vektorom pridruženim drugoj po redu najmanjoj svojstvenoj vrijednosti[3] od L, a najmanja vrijednost relaksirane verzije *RatioCut*-a aproksimirana je s drugom najmanjom svojstvenom vrijednosšću od L. □

Da bismo napravili biparticiju grafa za koju će *RatioCut* biti minimalan, moramo se nekako vratiti iz relaksiranog u diskretni slučaj.
Jedan od načina je da koristimo predznake komponenti optimalnog vektora y kao indikatore grupa:

$$\begin{cases} v_i \in A \text{ ako } y_i \geq 0 \\ v_i \in \overline{A} \text{ ako } y_i < 0. \end{cases}$$

[3]U literaturi se drugi svojstveni vektor Laplaceove matrice najčešće zove Fiedlerov vektor, a odgovarajuća svojstvena vrijednost poznata je pod nazivom algebarska povezanost grafa G.

Ovakav način je korektan jer svaki svojstveni vektor koji nije $\mathbb{1}$, a zbog okomitosti na $\mathbb{1}$, ima i pozitivnih i negativnih komponenti. Ipak, većina algoritama za spektralno grupiranje vraćanje u diskretan slučaj obavlja na drugi način: koordinate vektora y_i smatraju točkama u $\mathbb{R}$ pa koriste popularnu k-means metodu za određivanje optimalnih grupa C i $\overline{C}$. Dakle,

$$\begin{cases} v_i \in A \text{ ako } y_i \in C \\ v_i \in \overline{A} \text{ ako } y_i \in \overline{C}. \end{cases}$$

Možemo zaključiti da je problem minimizacije razmjernog reza za slučaj $k = 2$ vrlo jednostavno riješiti: treba naći odgovarajući svojstveni vektor Laplaceove matrice i upotrijebiti k-means.

3.1.2 Aproksimacija razmjernog reza za proizvoljni $k \in \mathbb{N}$

Slično kao i u prethodnom pododjeljku, rješavamo optimizacijski problem:

$$\min_{A_1,\ldots,A_k \subset V} RatioCut(A_1, \ldots, A_k). \tag{11}$$

Danoj particiji $A_1, \ldots, A_k \subset V$ pridružujemo indikator vektor $h_j = (h_{1j}, \ldots, h_{nj})^\tau$ na način:

$$h_{ij} = \begin{cases} \frac{1}{\sqrt{|A_j|}} & \text{ako } v_i \in A_j,\ i = 1, \ldots, n,\ j = 1, \ldots, k \\ 0 & \text{inače.} \end{cases} \tag{12}$$

Definiramo matricu $H \in \mathbb{R}^{n \times k}$ čiji su stupci vektori h_j, $j = 1, \ldots, k$.

Propozicija 3.1.4. *Neka je matrica H dana s (12). Tada vrijedi:*

1. $h_i^\tau L h_i = \dfrac{cut(A_i, \overline{A_i})}{|A_i|}$
2. $h_i^\tau L h_i = (H^\tau L H)_{ii}$.

Dokaz. Dokaz je sličan dokazu Propozicije 3.1.1. □

Propozicija 3.1.5. *Za Laplaceovu matricu L i matricu H definiranu s (12) vrijedi:*

$$RatioCut(A_1, \ldots, A_k) = tr(H^\tau L H),$$

pri čemu je $tr(H^\tau H)$ trag matrice $H^\tau L H$.

Dokaz. Koristeći Propoziciju 3.1.4, imamo:

$$RatioCut(A_1, \ldots, A_k) = \sum_{i=1}^{k} h_i^\tau L h_i = \sum_{i=1}^{k} (H^\tau L H)_{ii} = tr(H^\tau L H)$$

□

Kombinirajući Propoziciju 3.1.4 i Propoziciju 3.1.5, problem (11) možemo zapisati u matričnom obliku:

$$\min_{H \in \mathbb{R}^{n \times k}} tr(H^\tau L H) \text{ gdje je } H \text{ dana s (12) i } H^\tau H = I. \tag{13}$$

Uočimo da jednakost $H^\tau H = I$ vrijedi jer su stupci od H međusobno ortogonalni vektori.

Napomena 3.1.1. *Za matricu $H \in \mathbb{R}^{n \times k}$ kažemo da je semiortogonalna ako vrijedi $H^\tau H = I$.*

Kao i u slučaju $k = 2$, radi se o NP-teškom diskretnom optimizacijskom problemu pa relaksiramo problem tako da oslabimo uvjet na elemente matrice H, odnosno dozvolimo da njeni elementi budu proizvoljni realni brojevi, ali i dalje vrijedi $H^\tau H = I$.

Relaksacija problema (13) glasi:

$$\min_{H \in \mathbb{R}^{n \times k}} tr(H^T L H) \text{ uz uvjet } H^T H = I.$$

Ovo je standardni problem minimizacije traga pa će nam opet pomoći Rayleigh-Ritz teorem, ovaj put verzija za proizvoljan $k \in \mathbb{N}$.

Teorem 3.1.2 (Rayleigh-Ritz za proizvoljni k)**.** *Neka je $A \in \mathbb{R}^{m \times m}$ simetrična matrica sa svojstvenim vrijednostima $\lambda_1, \ldots, \lambda_m$ i pridruženim ortonormiranim svojstvenim vektorima $u_1, \ldots, u_m$.*
Tada je rješenje optimizacijskog problema:

$$\min_{X \in \mathbb{R}^{m \times n}} tr(X^\tau A X) \text{ uz uvjet } X^\tau X = I$$

za neki $n \in \{1, \ldots, m\}$ dano sa matricom X koja za stupce ima prvih n svojstvenih vektora. □

Dakle, prema Rayleigh-Ritz teoremu rješenje problema (13) dano je matricom H koja za stupce ima prvih k svojstvenih vektora, odnosno onih k svojstvenih vektora koji su pridruženi prvim k najmanjim svojstvenim vrijednostima matrice L.
Za vraćanje u diskretni slučaj koristimo k-means metodu nad retcima matrice H. Vrhove grafa reprezentiramo uređenim k-torkama realnih brojeva koji odgovaraju retcima od H i njih grupiramo k-means metodom.

3.2 Izvod normaliziranih algoritama

Balansirani rez $Ncut$ smo definirali u sekciji 2.3. Sada ćemo vidjeti kako relaksacija minimizacije normaliziranog reza (Definicija 2.3.6) vodi do normaliziranog spektralnog grupiranja. Najprije ćemo promatrati slučaj $k = 2$.

3.2.1 Aproksimacija normaliziranog reza za $k = 2$

Treba riješiti optimizacijski problem

$$\min_{A \subset V} Ncut(A, \overline{A}). \tag{14}$$

Skupu $A \subset V$ pridružimo indikator vektor $y = (y_1, \ldots, y_n)^\tau$:

$$y_i = \begin{cases} \sqrt{\frac{vol(\overline{A})}{vol(A)}} \text{ ako } v_i \in A \\ -\sqrt{\frac{vol(A)}{vol(\overline{A})}} \text{ ako } v_i \in \overline{A}. \end{cases} \tag{15}$$

Propozicija 3.2.1. *Za vektor y definiran s (15) vrijedi*

1. $(Dy)^\tau \mathbb{1} = 0$,
2. $y^\tau D y = vol(V)$.

Dokaz. Svojstva vektora y dokazujemo vrlo jednostavnim računom:

1.

$$\begin{aligned}(Dy)^\tau \mathbb{1} &= \sum_{i=1}^{n} d_i y_i \\ &= \sum_{i\in A} d_i \sqrt{\frac{vol(\overline{A})}{vol(A)}} - \sum_{i\in \overline{A}} d_i \sqrt{\frac{vol(A)}{vol(\overline{A})}} \\ &= \sqrt{\frac{vol(\overline{A})}{vol(A)}} \sum_{i\in A} d_i - \sqrt{\frac{vol(A)}{vol(\overline{A})}} \sum_{i\in \overline{A}} d_i \\ &= \sqrt{\frac{vol(\overline{A})}{vol(A)}} vol(A) - \sqrt{\frac{vol(A)}{vol(\overline{A})}} vol(\overline{A}) \\ &= 0.\end{aligned}$$

2.

$$\begin{aligned}y^\tau D y &= \sum_{i=1}^{n} y_i^2 d_i \\ &= \sum_{i\in A} \frac{vol(\overline{A})}{vol(A)} d_i + \sum_{i\in \overline{A}} \frac{vol(A)}{vol(\overline{A})} d_i \\ &= \frac{vol(\overline{A})}{vol(A)} vol(A) + \frac{vol(A)}{vol(\overline{A})} vol(\overline{A}) \\ &= vol(\overline{A}) + vol(A) \\ &= vol(V).\end{aligned}$$

□

Propozicija 3.2.2. *Neka je y definiran s (15). Tada vrijedi:*

$$y^\tau L y = vol(V) Ncut(A, \overline{A}).$$

Dokaz. Upotrijebiti ćemo Propoziciju 2.4.1:

$$\begin{aligned}
y^\tau Ly &= \frac{1}{2}\sum_{i,j=1}^{n} w_{ij}(y_i - y_j)^2 \\
&= \frac{1}{2}\sum_{i\in A, j\in\overline{A}} w_{ij}\left(\sqrt{\frac{vol(\overline{A})}{vol(A)}} + \sqrt{\frac{vol(A)}{vol(\overline{A})}}\ \right)^2 + \\
&\qquad + \frac{1}{2}\sum_{i\in\overline{A}, j\in A} w_{ij}\left(-\sqrt{\frac{vol(A)}{vol(\overline{A})}} - \sqrt{\frac{vol(\overline{A})}{vol(A)}}\ \right)^2 \\
&= \frac{1}{2}\left(\sum_{i\in A, j\in\overline{A}} w_{ij} + \sum_{i\in\overline{A}, j\in A} w_{ij}\right)\left(\frac{vol(\overline{A})}{vol(A)} + \frac{vol(A)}{vol(\overline{A})} + 2\right) \\
&= cut(A,\overline{A})\left(\frac{vol(\overline{A})}{vol(A)} + \frac{vol(A)}{vol(\overline{A})} + 2\right) \\
&= cut(A,\overline{A})\left(\frac{vol(\overline{A}) + vol(A)}{vol(A)} + \frac{vol(A) + vol(\overline{A})}{vol(\overline{A})}\right) \\
&= vol(V)Ncut(A,\overline{A}).
\end{aligned}$$

□

Kombinirajući Propozicije 3.2.1 i 3.2.2, problem (14) možemo zapisati ovako:

$$\min_{y\in\mathbb{R}^n} y^\tau Ly \text{ pri čemu je } y \text{ dan sa (15)}, Dy\perp\mathbb{1}, y^\tau Dy = vol(V).$$

Pripadni relaksirani problem je

$$\min_{y\in\mathbb{R}^n}\ y^\tau Ly \text{ uz uvjet } Dy\perp\mathbb{1}, y^\tau Dy = vol(V).$$

Uvedemo supstituciju $g := D^{1/2}y$ pa možemo pisati:

$$\min_{g\in\mathbb{R}^n} g^\tau D^{-1/2}LD^{-1/2}g \ \text{ uz uvjet } g\perp D^{1/2}\mathbb{1}, \|g\|^2 = vol(V).$$

Uvjet $Dy\perp\mathbb{1}$ je nakon supstitucije ekvivalentan uvjetu $g\perp D^{1/2}\mathbb{1}$ jer vrijedi:

$$Dy\perp\mathbb{1} \Rightarrow (D^{1/2}g)^\tau\mathbb{1} = g^\tau D^{1/2}\mathbb{1} = 0 \Rightarrow g\perp D^{1/2}\mathbb{1},$$

a uvjet $\|g\|^2 = vol(V)$ vrijedi jer

$$vol(V) = y^\tau Dy = g^\tau D^{-1/2}DD^{-1/2}g = g^\tau g.$$

Sada se u problemu minimizacije pojavljuje matrica $D^{-1/2}LD^{-1/2}$ koja je po definiciji normalizirana Laplaceova matrica L_{sym}. Iz Propozicije 2.4.3 znamo da je svojstveni vektor pridružen najmanjoj svojstvenoj vrijednosti od L_{sym}, tj. nuli, jednak $D^{1/2}\mathbb{1}$. Kako taj vektor ne ispunjava zadane uvjete optimizacije, rješenje g je dano svojstvenim vektorom pridruženom drugoj po redu najmanjoj svojstvenoj vrijednosti od L_{sym}, što opet daje Rayleigh-Ritz teorem.
Vraćanjem supstitucije $y = D^{-1/2}g$, a uz Propoziciju 2.4.3, zaključujemo da je y svojstveni vektor pridružen drugoj po redu svojstvenoj vrijednosti od L_{rw}.

Slično kao i kod *RatioCut* problema, na elemente vektora y treba primijeniti 2-means algoritam kako bismo grupirali vrhove grafa u dvije grupe.

3.2.2 Aproksimacija normaliziranog reza za proizvoljni $k \in \mathbb{N}$

Sada ćemo analizirati $Ncut$ problem za proizvoljnan broj grupa, a po uzoru na slučaj $k = 2$.
Skupu $A_j \subset V$ pridružimo indikator vektor $h_j = (h_{1j}, \ldots, h_{nj})^\tau$ na način:

$$h_{ij} = \begin{cases} \frac{1}{\sqrt{vol(A_j)}} & \text{ako } v_i \in A_j,\ i = 1, \ldots, n,\ j = 1, \ldots, k \\ 0 & \text{inače.} \end{cases} \tag{16}$$

Definiramo matricu H čiji su stupci indikator vektori skupova $A_1, \ldots, A_k$.

Propozicija 3.2.3. *Neka je matrica $H = (h_{ij})_{i=1,\ldots,n, j=1,\ldots,k}$ definirana s (16). Tada vrijedi:*

1. $H^\tau H = I$
2. $h_i^\tau D h_i = 1$
3. $h_i^\tau L h_i = \dfrac{cut(A_i, \overline{A_i})}{vol(A_i)}$.

Dokaz. Dokaz je sličan dokazu Propozicija 3.2.1 i 3.2.2. □

Problem $Ncut$ zapisujemo ovako:

$$\min_{H \in \mathbb{R}^{n \times k}} tr(H^\tau L H) \text{ uz uvjet } H^\tau D H = I, \text{ gdje je } H \text{ dana s (16).} \tag{17}$$

Uvodimo supstituciju $T = D^{1/2}H$.
Relaksiramo problem (17) tako da dozvolimo da elementi od H budu realni brojevi,

ali i dalje vrijedi $H^\tau DH = I$.
Imamo

$$\min_{T \in \mathbb{R}^{n \times k}} tr(T^\tau D^{-1/2} L D^{-1/2} T) \text{ uz uvjet } T^\tau T = I.$$

Opet imamo standardni problem minimizacije traga pa je prema Rayley-Ritzovom teoremu rješenje T dano matricom čiji su stupci svojstveni vektori pridruženi prvim k najmanjim svojstvenim vrijednostima od L_{sym}.
Vraćanjem supstitucije $H = D^{-1/2}T$ zaključujemo da su stupci matrice H ujedno i svojstveni vektori pridruženi prvim k najmanjim svojstvenim vrijednostima matrice L_{rw}.
Preostaje primjeniti k-means algoritam na retke matrice H da bismo dobili odgovarajuće grupe vrhova.

Dakle, relaksacija $Ncut$ problema nas vodi do algoritma koji koristi svojstvene vektore od L_{rw}. To je metoda normaliziranog spektralnog grupiranja autora Shi i Malik (2000).
Druga verzija ove metode koristi svojstvene vektore od L_{sym} koji čine matricu T, a predstavili su je Ng, Jordan i Weiss (2002).

4 Spektralno grupiranje i slučajna šetnja

Metode spektralnog grupiranja mogu se objasniti i pomoću slučajne šetnje u grafu sličnosti.

Slučajna šetnja u grafu je stohastički proces[4] u kojem vrhove grafa obilazimo na slučajan način.
Vjerojatnost da ćemo iz vrha v_i doći u vrh v_j proporcionalnam je težini brida w_{ij} i dana je s

$$p_{ij} = \frac{w_{ij}}{d_i}.$$

Možemo definirati tranzicijsku matricu P slučajne šetnje s elementima p_{ij}, $i, j = 1, \ldots, n$, odnosno

$$P = D^{-1}W.$$

Odmah možemo uočiti vezu između matrice P i matrice L_{rw} obzirom da vrijedi $L_{rw} = I - P$. Stoga je naziv "Laplaceova matrica slučajnih šetnji" svakako opravdan!
Znamo da je λ svojstvena vrijednost od L_{rw} sa svojstvenim vektorom u ako i samo ako je $1 - \lambda$ svojstvena vrijednost od P sa svojstvenim vektorom u.
Ova veza nam je iznimno korisna. Poznato je da se mnoga svojstva grafa mogu izraziti preko odgovarajuće tranzicijske matrice P slučajne šetnje. Jedno od tih svojstava će nam pomoći u pronalasku grupa sličnih vrhova u grafu.

Ako je graf povezan i nije bipartitan, tj. skupovi vrhova mu se ne mogu particionirati u dva podskupa tako da svi njegovi bridovi spajaju vrhove jednog podskupa sa vrhovima drugog, tada slučajna šetnja ima jedinstvenu distribuciju: $\pi = (\pi_1, \ldots, \pi_n)^\tau$, gdje je $\pi_i = \frac{d_i}{vol(V)}$. Matrica P je stohastička, odnosno sve njezine vrijednosti su nenegativne i suma elemenata u svakom retku jednaka je 1.

[4]Stohastički proces je opis tranzicije objekta kroz vrijeme. U svakom stanju imamo jednu ili više mogućnosti za prijelaz i svaka pozicija ima određenu vjerojatnost. Iako ne možemo znati točan put, na temelju vjerojatnosti ga možemo pretpostaviti.

4.1 Slučajna šetnja i normalizirani rez

Propozicija 4.1.1 (*Ncut* preko prijelaznih vjerojatnosti)**.** *Neka je G povezan graf koji nije bipartitan. Pretpostavimo da imamo slučajnu šetnju $(X_i)_{i\in\mathbb{N}}$ sa početkom u X_0 i stacionarnom distribucijom π. Za međusobno disjunktne podskupove $A, B \subset V$ definiramo: $P(B|A) := P(X_1 \in B|X_0 \in A)$ kao vjerojatnost da će slučajna šetnja doći u skup B ako se trenutno nalazi u skupu A.*
Tada vrijedi:

$$Ncut(A, \overline{A}) = P(\overline{A}|A) + P(A|\overline{A}).$$

Dokaz. Uočimo da vrijedi:

$$P(X_0 \in A, X_1 \in B) = \sum_{i\in A, j\in B} P(X_0 = i, X_1 = j) = \sum_{i\in A, j\in B} \pi_i p_{ij}$$
$$= \sum_{i\in A, j\in B} \frac{d_i}{vol(V)} \frac{w_{ij}}{d_i} = \frac{1}{vol(V)} \sum_{i\in A, j\in B} w_{ij}.$$

Koristeći tu jednakost, dobivamo:

$$P(X_1 \in B|X_0 \in A) = \frac{P(X_0 \in A, X_1 \in B)}{P(X_0 \in A)}$$
$$= \left(\frac{1}{vol(V)} \sum_{i\in A, j\in B} w_{ij}\right)\left(\frac{vol(A)}{vol(V)}\right)^{-1} = \frac{1}{vol(A)} \sum_{i\in A, j\in B} w_{ij},$$

čime smo pokazali da je vjerojatnost da ćemo iz vrha $X_0 \in A$ doći u vrh $X_1 \in B$ slučajnom šetnjom, jednaka upravo vrijednosti normaliziranog reza *Ncut* skupova A i B. □

Prethodna propozicija na vrlo lijep način interpretira minimizaciju *Ncut* problema: pokušavamo particionirati skup vrhova grafa u grupe takve da slučajna šetnja, jednom kad je u jednoj grupi, ima tendenciju u njoj i ostati.

Ovo poglavlje zaključujemo jednom bitnom napomenom:

Napomena 4.1.1. *Iako iskustvo pokazuje da metode spektralnog grupiranja, kao relaksirane metode odgovarajućih diskretnih optimizacijskih problema, daju jako dobre rezultate, jamstvo na kvalitetu rješenja i dalje ne postoji. Također, relaksacije koje smo predstavili nisu jedinstvene. Prednost korištenja metode spektralnog grupiranja*

nije u tome što daje savršene rezultate, već je prednost u samom načinu funkcioniranja te metode. Kao što smo vidjeli, ona se bazira na nekoliko jednostavnih rezultata iz područja linearne algebre.

5 k-means grupiranje

Algoritam k-means je najpoznatiji i najjednostavniji algoritam za grupiranje podataka. Kod izvoda spektralnih metoda smo vidjeli da se upravo taj algoritam koristi u posljednjim koracima kako bi se vratili na diskretni slučaj nakon relaksacije $RatioCut$ i $Ncut$ problema. Istina je zapravo da umjesto k-means algoritma možemo koristiti bilo koji drugi particijski algoritam grupiranja koji bi trebao pronaći jednako kvalitetne grupe podataka uz pretpostavku da smo u prethonom koraku spektralnih metoda promijenili reprezentaciju podataka na odgovarajući način. Algoritam k-means se najčešće koristi upravo zbog svoje jednostavnosti i brzine.

Cilj k-means metode je pronaći grupe podataka iterativno minimizirajući udaljenosti između podataka i neke točke koja predstavlja pojedinu grupu, a koju nazivamo *centrom*. Za centar grupe se obično uzima aritmetička sredina udaljenosti podataka.

Neka je $D = \{x_1, \ldots, x_n\}$ dani skup podataka i neka je k traženi broj grupa. U početku nasumično izaberemo k centara $c_1^{(0)}, \ldots, c_k^{(0)}$ grupa $C_1^{(0)}, \ldots, C_k^{(0)}$, gdje $c_j^{(0)}$ označava j-ti centar C_j-e grupe u nultoj iteraciji. Svaku točku $x_i \in D$, $i = 1, \ldots, n$ pridružimo najbližem centru tako da dobijemo grupe nulte iteracije formalno dane sa:

$$C_j^{(0)} = \{x_i \in D : d(c_j^{(0)}, x_i) \leq d(c_l^{(0)}, x_i)\}, \quad l, i = 1, \ldots, n,$$

pri čemu vodimo račun o tome da svaki podatak pripada samo jednoj grupi. Za d najčešće uzimamo Euklidsku metriku. Nove centre računamo prema formuli

$$c_j^{(1)} = \frac{1}{|c_j^{(0)}|} \sum_{x_i, x_l \in C_j^{(0)}, i \neq l} d(x_i, x_l)$$

i pridružujemo svaku točku skupa D novom najbližem centru, odnosno imamo:

$$C_j^{(1)} = \{x_i \in D : d(c_j^{(1)}, x_i) \leq d(c_l^{(1)}, x_i)\}, \quad l, i = 1, \ldots, n.$$

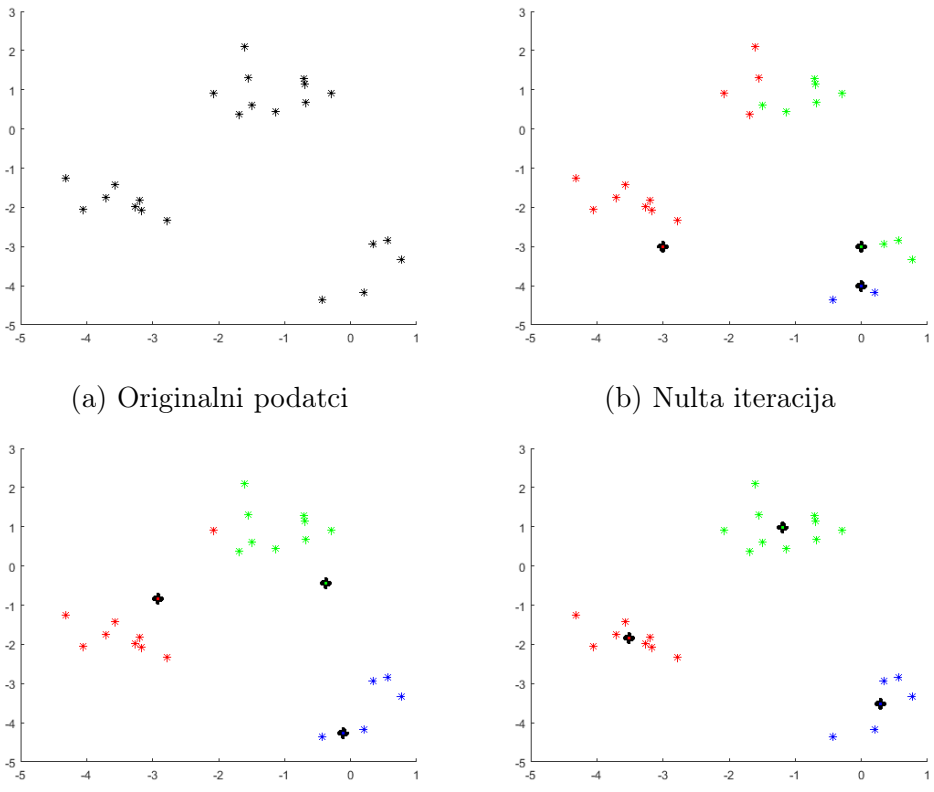
Na ovaj način minimiziramo takozvanu *funkciju cilja*:

$$\sum_{j=1}^{k} \sum_{x_i \in C_j} d(c_j, x_i).$$

Postupak ponavljamo sve dok novi centri grupa ne budu jednaki (do na eventualnu

dogovorenu grešku) centrima u prethodnoj iteraciji, ili dok ne prijeđemo unaprijed definiran maksimalan broj iteracija.

Postupak grupiranja k-means metodom možemo vidjeti u primjeru na Slici 1. Imamo zadan jednostavan skup točaka u ravnini i želimo ga particionirati u tri grupe k-means metodom. Centre označavamo znakom '+'. Nasumično smo odabrali početne centre, a zatim smo svaki podatak pridružili grupi čiji je centar najbliži i obojali ga bojom karakterističnom za tu grupu, što možemo vidjeti na Slici 1b. Za nove centre uzimamo aritmetičku sredinu podataka (Slika 1c) i ponavljamo postupak sve dok centri ne poprime jednake vrijednosti kao u prethodnoj iteraciji.



(a) Originalni podatci

(b) Nulta iteracija

(c) Prva iteracija

(d) Druga iteracija

Slika 1: k-means algoritam

Algoritam 1 k-means algoritam

Ulaz: $D = \{x_1, \ldots, x_n\}$ - skup podataka, k - broj grupa, $c_1^{(0)}, \ldots, c_k^{(0)}$ - početni centri, it - oznaka iteracije (početak $it = 0$), it_{max} - maksimalan broj iteracija

1: Odredimo grupe

$$C_j^{(it)} = \{x_i \in D : d(c_j^{(it)}, x_i) \leq d(c_l^{(it)}, x_i)\}, \quad i, l = 1, \ldots, n, \quad j = 1, \ldots, k,$$

tako da svaki podatak x_i pripada samo jednoj grupi $C_j^{(it)}$, i izračunamo njihove centre

$$c_j^{(it+1)} = \frac{1}{|c_j^{(it)}|} \sum_{x_i, x_l \in C_j^{(it)}, i \neq l} d(x_i, x_l), \quad j = 1, \ldots, k$$

2: Ako je $c_1^{(it)} \approx c_1^{(it+1)}, \ldots, c_k^{(it)} \approx c_k^{(it+1)}$ ili je $it = it_{max}$ stanemo, inače postavimo $it = it + 1$ i idemo na korak 1

Izlaz: Grupe $C_1^{(it)}, \ldots, C_k^{(it)}$, gdje je $it \leq it_{max}$ broj iteracija.

Možemo zaključiti da k-means metoda daje jako dobre rezultate kod problema particioniranja konveksnih skupova. U slučaju kada skupovi nisu konveksni, ova metoda će u većini slučajeva potpuno zakazati. Upravo zbog toga koristimo spektralne metode jer će se u kombinaciji sa k-means metodom pokazati kao jako dobar alat za particioniranje nekonveksnih skupova.

6 Algoritmi spektralnog grupiranja

U prethodnom poglavlju smo detaljno objasnili metode spektralnog grupiranja podataka koje se baziraju na tri različite Laplaceove matrice: L, L_{sym} i L_{rw}. Sada možemo predstaviti odgovarajuće algoritme, po jedan za svaki tip Laplaceove matrice.

Treba, dakle, krenuti sa određenim skupom podataka $D = \{x_1, \ldots, x_n\}$ kojeg želimo particionirati u k grupa tako da unutar pojedine grupe budu podatci koji su što sličniji, a u različitim grupama trebaju biti podatci koji su najmanje međusobno slični.

Obzirom na tip podataka, biramo funkciju sličnosti koja će svakom paru podataka pridružiti nenegativan realan broj. Što je taj broj veći, odgovarajući par podataka je sličniji. Zatim biramo odgovarajući tip grafa sličnosti i primijenimo neku od metoda spektralnog grupiranja. Slijede opisi algoritama za spektralno grupiranje.

6.1 Nenormalizirano spektralno grupiranje

U ovom pododjeljku predstavljamo algoritam spektralnog grupiranja podataka koji koristi nenormaliziranu Laplaceovu matricu L:

Algoritam 2 Metoda nenormaliziranog spektralnog grupiranja

Ulaz: Matrica sličnosti $S \in \mathbb{R}^{n \times n}$, broj grupa k.

1: Konstruiramo graf sličnosti na jedan od načina objašnjenih u odjeljku 2.2 i pridružimo mu matricu susjedstva W.
2: Izračunamo nenormaliziranu Laplaceovu matricu L.
3: Izračunamo prvih k svojstvenih vektora $u_1, \ldots, u_k$ od L.
4: Formiramo matricu $U \in \mathbb{R}^{nxk}$ čiji su stupci redom vektori $u_1, \ldots, u_k$.
5: Za $i = 1, \ldots, n$ neka je $y_i \in \mathbb{R}^k$ vektor koji odgovara i-tom retku od U.
6: Grupiramo točke $(y_i)_{i=1,\ldots,n} \in \mathbb{R}^k$ pomoću k-means algoritma u grupe $C_1, \ldots, C_k$.

Izlaz: Grupe $A_1, \ldots, A_k$ podataka dane s $A_i = \{j : y_j \in C_i\}$.

6.2 Spektralno grupiranje uz pomoć Laplaceove matrice slučajnih šetnji L_{rw}

Kao što smo ranije napomenuli, spektralno grupiranje podataka koje koristi Laplaceovu matricu slučajnih šetnji L_{rw} predstavili su 2000. godine Shi i Malik [16]. Evo

algoritma:

Algoritam 3 Metoda spektralnog grupiranja koja koristi matricu L_{rw}

Ulaz: Matrica sličnosti $S \in \mathbb{R}^{n \times n}$, broj grupa k.

1: Konstruiramo graf sličnosti na jedan od načina objašnjenih u odjeljku 2.2 i pridružimo mu matricu susjedstva W.
2: Izračunamo Laplaceovu matricu slučajnih šetnji L_{rw}.
3: Izračunamo prvih k svojstvenih vektora matrice L_{rw}.
4: Formiramo matricu $U \in \mathbb{R}^{nxk}$ čiji su stupci redom vektori $u_1, \ldots, u_k$.
5: Za $i = 1, \ldots, n$ neka je $y_i \in \mathbb{R}^k$ vektor koji odgovara i-tom retku od U.
6: Grupiramo točke $(y_i)_{i=1,\ldots,n} \in \mathbb{R}^k$ pomoću k-means algoritma u grupe $C_1, \ldots, C_k$.

Izlaz: Grupe $A_1, \ldots, A_k$ podataka dane s $A_i = \{j : y_j \in C_i\}$.

6.3 Spektralno grupiranje uz pomoć normalizirane Laplaceove matrice L_{sym}

Sada još treba predstaviti algoritam za grupiranje koje koristi normaliziranu Laplaceovu matricu L_{sym}, a kojeg su 2002. godine predstavili Andrew Y. Ng, Michael I. Jordan i Yair Weiss.

Algoritam 4 Metoda spektralnog grupiranja koja koristi matricu L_{sym}

Ulaz: Matrica sličnosti $S \in \mathbb{R}^{n \times n}$, broj grupa k.

1: Konstruiramo graf sličnosti na jedan od načina objašnjenih u odjeljku 2.2 i pridružimo mu matricu susjedstva W.
2: Izračunamo normaliziranu Laplaceovu matricu L_{sym}.
3: Izračunamo prvih k svojstvenih vektora matrice L_{sym}.
4: Neka je $U \in \mathbb{R}^{n \times k}$ matrica koja se sastoji od vektora $u_1, \ldots, u_k$ po stupcima.
5: Formiramo matricu $T \in \mathbb{R}^{n \times k}$ iz matrice U tako da normaliziramo retke od U. Imamo:
$$t_{ij} = \frac{u_{ij}}{\sqrt{\sum_k u_{ik}^2}}.$$
6: Za $i = 1, \ldots, n$ neka je $y_i \in \mathbb{R}^k$ vektor koji odgovara i-tom retku od T.
7: Grupiramo točke $(y_i)_{i=1,\ldots,n} \in \mathbb{R}^k$ pomoću k-means algoritma u grupe $C_1, \ldots, C_k$.

Izlaz: Grupe $A_1, \ldots, A_k$ podataka dane s $A_i = \{j : y_j \in C_i\}$.

Primijetimo da ovaj algoritam ima dodatan korak u kojem normiramo retke od U, što nije potrebno kod ostalih algoritama. Objašnjenje je krajnje jednostavno. U idealnom slučaju, tj. kada je graf sličnosti takav da su mu komponente povezanosti upravo one grupe koje želimo dobiti spektralnim grupiranjem, indikator vektori grupa se podudaraju sa svojstvenim vektorima matrica L i L_{rw}. Kod matrice L_{sym} to nije slučaj jer su u idealnom slučaju traženi svojstveni vektori jednaki $D^{1/2}\mathbb{1}_{A_i}$. Tek kada normiramo retke matrice koja sadrži svojstvene vektore kao stupce, tj. tek kada normiramo retke od T, dobiti ćemo indikator vektore traženih grupa.

Možemo zaključiti da sva tri algoritma imaju istu ideju: promijeniti reprezentaciju danih podataka $x_i \in \mathbb{R}^n$ u apstraktne podatke $y_i \in \mathbb{R}^k$. Dakle, krenuli smo od matrice sličnosti čiji se retci mogu shvatiti kao točke iz $\mathbb{R}^n$. Svaki redak odgovara određenom podatku, a svaki broj u tom retku objašnjava odnos tog podatka i svih ostalih podataka, tj. podatak je matricom sličnosti reprezentiran pomoću n-dimenzionalnog vektora koji sadrži informacije o sličnosti tog podatka sa svim ostalim podatcima. Nakon što konstruiramo odgovarajuću Laplaceovu matricu, nađemo joj prvih k svojstvenih vektora i definiramo matricu U, podatci mijenjaju reprezentaciju. Zapravo smo na taj način reducirali dimenziju prostora jer podatci više nisu iz $\mathbb{R}^n$, već ih promatramo kao k-torke realnih brojeva koji su dovoljno blizu tome da ukazuju na jasnu strukturu grupa. Takva promjena je moguća upravo zbog svojstava Laplaceovih matrica te omogućuje trivijalno otkrivanje grupa nekom od popularnih metoda kao što je primjerice k-means.

Treba još napomenuti da se u praksi najčešće koriste spektralne metode koje koriste matrice L_{rw} ili L_{sym} obzirom da daju najbolje rezultate.

7 Primjeri

Vidjeli smo da metode spektralnog grupiranja završavaju uporabom k-means metode, odnosno da se ta metoda primjenjuje u koracima 6 kod algoritama koji koriste matricu L, odnosno matricu L_{rw} i u koraku 7 kod algoritma koji koristi matricu L_{sym}. Zapravo iskorištavamo jednostavnost i brzinu k-means algoritma na novo reprezentiranim podatcima čije je grupiranje nakon toga trivijalno.

U prethodnom odjeljku smo spomenuli da k-means ne može pravilno grupirati skupove koji nisu konveksni. U ovom odjeljku ćemo kroz nekoliko sintetičkih i 'real world' primjera pokazati uspješnost spektralnih metoda u rješavanju problema nekonveksnog particioniranja.
Sve implementacije su rađene u programskom paketu MATLAB.

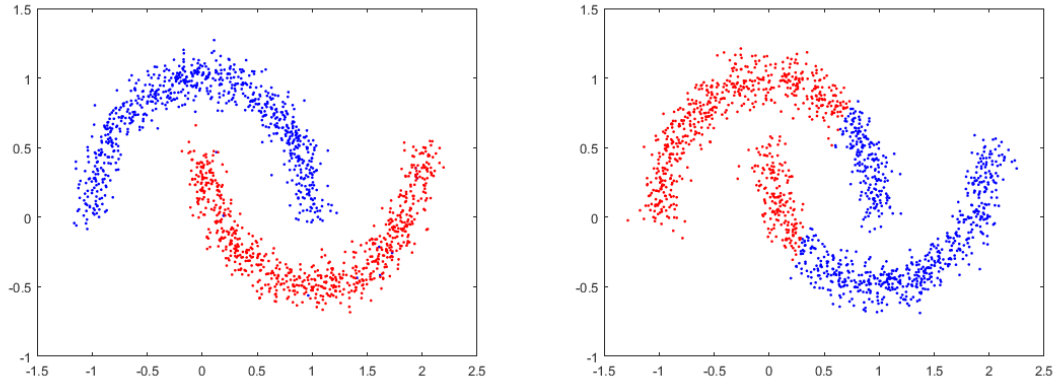
7.1 Sintetički primjeri

Najprije ćemo predstaviti primjer koji pokazuje slabost k-means metode grupiranja u odnosu na spektralne metode.

7.1.1 Primjer - polumjeseci

Generiran je skup podataka - točaka u ravnini koji su grupirani tako da tvore oblike koji nalikuju na dva polumjeseca okrenuta jedan prema drugom. Jasno je da se ovdje radi o nekonveksnim skupovima. Testiranje MATLAB-ove funkcije *kmeans* uz zadani broj grupa $k = 2$ daje rezultate prikazane na Slici 2b. Čak i na tako jednostavnom skupu podataka, k-means ne daje dobre rezultate. Na Slici 2a prikazani su rezultati dobiveni metodom spektralnog grupiranja koja koristi matricu L_{rw}. Za graf sličnosti smo odabrali graf k-najbližih susjeda, pri čemu je $k = 11$. Dobivene grupe obojane su različitim bojama.

Ovdje je na prvi pogled jasno o kakvim se grupama radi i sa sigurnošću možemo reći da podatci jednog polumjeseca tvore jednu grupu, dok podatci drugog polumjeseca drugu. Međutim, procjena kvalitete dobivenih grupa u općenitom slučaju nije trivijalan posao, posebno kada imamo podatke velikih dimenzija koje ne možemo prikazati grafički.



(a) Spektralno grupiranje (b) k-means

Slika 2: Usporedba spektralne metode i k-means algoritma

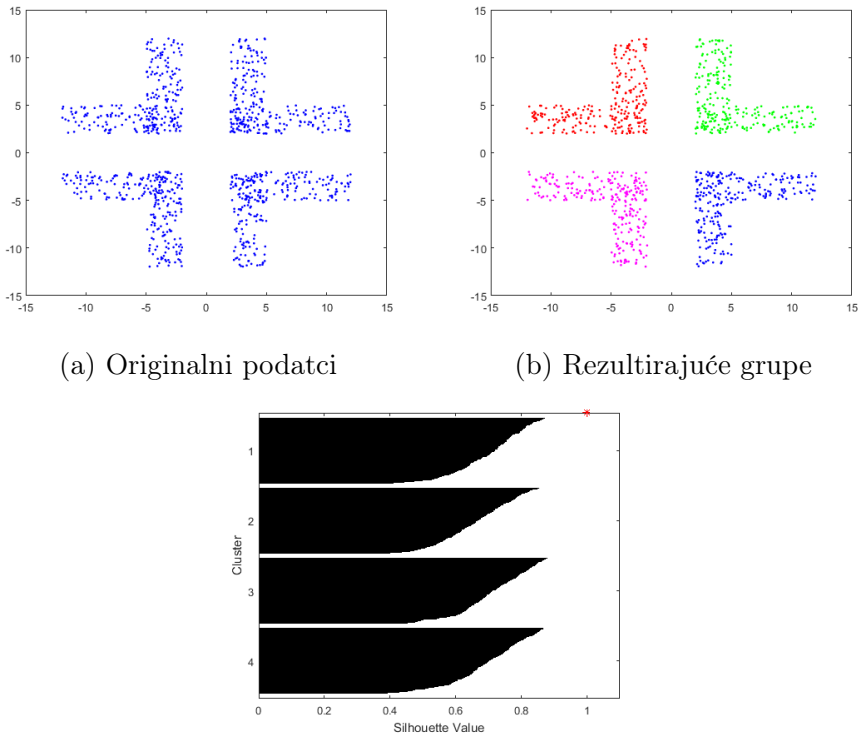
Postoji više metoda za automatsku validaciju dobivenih grupa podataka, a mi ćemo koristiti dvije: *silhouette graf* i *ARI indeks*. U nastavku ćemo ih ukratko opisati.

• Silhouette metoda interpretira i validira konzistenciju podataka unutar dobivene grupe, odnosno pokušava ocijeniti rezultat grupiranja na temelju dva fundamentalna cilja grupiranja: sličnosti podataka unutar pojedinih grupa i različitosti podataka između grupa. Metoda pruža grafički prikaz ocjene kvalitete, tzv. silhouette graf podataka. Silhouette vrijednost proizvoljnog podatka se nalazi unutar skupa $[-1, 1]$. Što je ova vrijednost veća, to bi se podatak trebao što bolje uklopiti u grupu kojoj je pridružen. Na primjerima ćemo vidjeti da to ipak nije uvijek tako. Validacija uvelike ovisi o grafičkim svojstvima podataka pa silhouette metoda ponekad daje potpuno pogrešne ocjene. Metoda je prvi put opisana 1986. godine u radu *Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis* (vidi [17]) autora Peter J. Rousseeuwa.

• Hubert-Arabie ARI indeks (eng. *adjusted Rand index*) je metoda koja značajno bolje opisuje intuitivne pretpostavke o grupama. ARI indeks uspoređuje rezultat grupiranja sa nekim vanjskim kriterijem i poprima vrijednosti iz skupa $[-1, 1]$. Što je vrijednost veća, to je dobivena particija bliža idealnoj. Kako mi imamo egzaktna rješenja grupiranja za sve skupove podataka koje promatramo, referentna particija će nam biti upravo ona koja je idealna. Više o Hubert-Arabie indeksu se može naći u literaturi [18].

7.1.2 Primjer - četiri kuta

Promatramo generirani skup dvodimenzionalnih podataka koji su grupirani u obliku četiri kuta kao što je prikazano na Slici 3a. Ovdje se idealna podjela odmah vidi sa slike, no kvalitetu particije ćemo ipak provjeriti prethodno opisanim metodama. ARI indeks daje realne ocjene isključivo iz razloga što uspoređuje rezultat s egzaktnim rješenjem, koje ne uzimamo u obzir kod silhouette metode zbog čega ne čudi što ona ponekad griješi.
Dakle, jasno je da tražimo četiri grupe podataka od kojih svaka formira jedan od četiri nekonveksna skupa. Koristimo potpun graf susjedstva i algoritam nenormaliziranog spektralnog grupiranja, odnosno onaj koji koristi matricu L. Slika 3b prikazuje rezultat grupiranja gdje je svaka grupa označena posebnom bojom. Silhouette kriterij iznosi 0.6988. ARI indeks iznosi 1. Bez obzira što prosječna silhouette vrijednost nije maksimalna, zaključujemo da je algoritam podatke grupirao bez greške.



(a) Originalni podatci (b) Rezultirajuće grupe

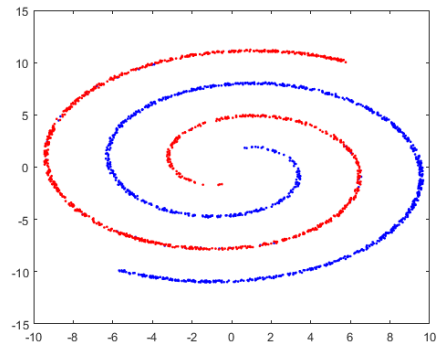
(c) Silhouette graf

Slika 3: Grupiranje spektralnom metodom koja koristi L

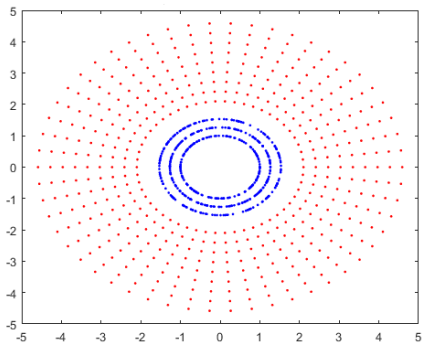
7.1.3 Primjer - spirale i kružnice

Sada ćemo predstaviti dva vrlo slična primjera. Rezultat grupiranja skupa podataka u ravnini koji formira dvije spirale prikazan je na Slici 4a, a rezultat grupiranja skupa podataka koji se sastoji od kružnica različitih gustoća prikazan je na Slici 4b. U oba slučaja imamo prirodnu podjelu podataka na dvije grupe od kojih smo jednu označili plavom, a drugu crvenom bojom. Koristili smo algoritam spektralnog grupiranja koji koristi matricu L_{rw}. Dvije spirale smo grupirali pomoću grafa uzajamnih 10-najbližih susjeda, a kružnice pomoću ϵ-grafa susjedstva.

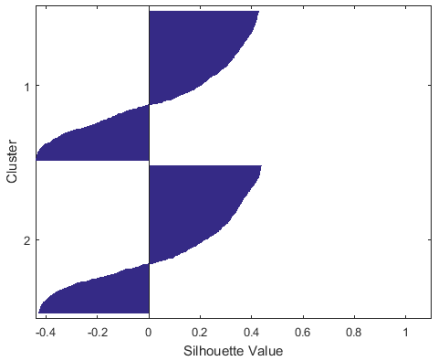
Silhouette grafovi prikazuju negativne vrijednosti zbog geometrijskih svojstava podataka. ARI indeks prvog i drugog skupa iznosi 1 kao i u prethodnom primjeru.



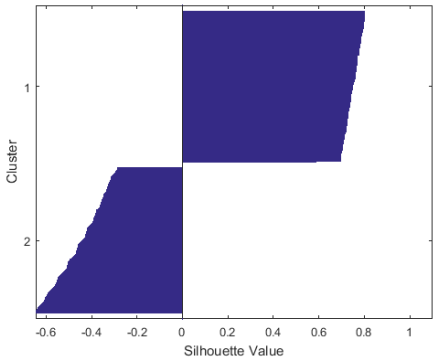
(a) Dvije spirale



(b) Koncentrične kružnice



(c) Silhouette - dvije spirale



(d) Silhouette - koncentrične kružnice

Slika 4: Grupe dobivene spektralnom metodom koja koristi matricu L_{rw}

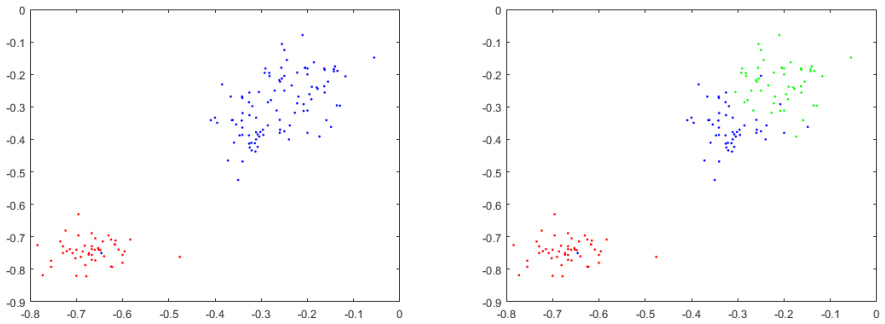
7.2 'Real world' primjeri

U nastavku ćemo predstaviti neke realne primjere u kojima će nas zanimati možemo li spektralnim metodama dobro odrediti grupe i u slučaju velikog broja višedimenzionalnih podataka gdje grupe nisu očigledne. Rezultate ćemo usporediti sa egzaktnim rješenjima.

7.2.1 Primjer - Iris

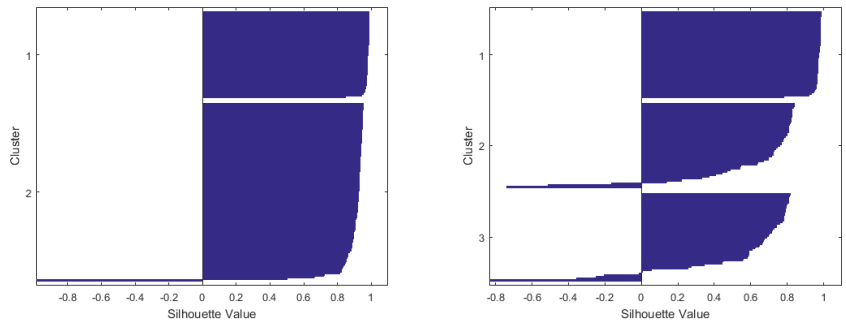
Prvi primjer koji ćemo promatrati je vrlo poznata baza podataka pod nazivom Iris iz 1936. godine, koja se prvi put pojavljuje u radu [4]. Podatci su preuzeti sa [13]. Baza sadrži različite mjere latica i čašice cvijeta irisa pomoću kojih možemo saznati podvrstu kojoj svaka biljka pripada. Podvrste koje imamo su Iris Setosa, Iris Versicolour i Iris Virginica. Imamo tri grupe od kojih jednu možemo linearno odvojiti od ostale dvije (Iris Setosa) čije zajedničke karakteristike nisu linearno odvojive (Iris Versicolor i Iris Virginica). Prvo tražimo dvije, a zatim tri grupe. Za prikaz višedimenzionalnih podataka koristimo tehniku *star coordinates* (vidi: [10]).

Na Slici 5a su prikazani rezultati grupiranja uz zadani broj grupa $k = 2$, a na Slici 5b rezultati za slučaj $k = 3$. Koristili smo algoritam normaliziranog spektralnog grupiranja koji koristi L_{sym} uz odabir grafa uzajamnih 30-najbližih susjeda.
Na Slici 6a je prikazan silhouette graf podataka podijeljenih u dvije grupe. Silhouette vrijednost je ovaj put bolja nego u prethodnim slučajevima i iznosi 0.9196. Vrijednost ARI indeksa je 0.9730. Na Slici 6b je prikazan silhouette graf podataka podijeljenih u tri grupe. Silhouette vrijednost iznosi 0.7139, a ARI indeks 0.9799.



(a) Dvije grupe (b) Tri grupe

Slika 5: Rezultati grupiranja baze podataka Iris

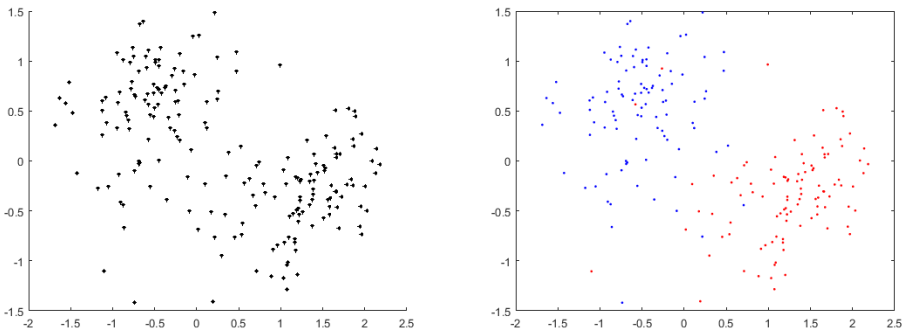


(a) Dvije grupe (b) Tri grupe

Slika 6: Silhouette graf baze podataka Iris

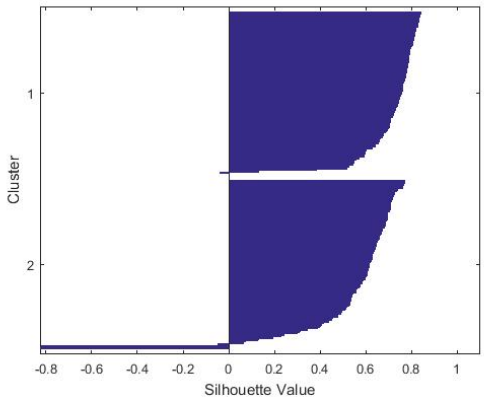
7.2.2 Primjer - Banknotes

Sljedeći skup podataka *Banknotes* sadrži različite mjere lažnih i pravih novčanica od 1000 švicarskih franaka (vidi [5]). Imamo 200 7-dimenzionalnih podataka, od kojih je 100 dobiveno mjerenjem lažnih, a 100 mjerenjem pravih novčanica. Pomoću algoritma spektralnog grupiranja pokušavamo identificirati te dvije grupe podataka. Na Slici 7 su prikazani rezultati. Podatke prikazujemo isto kao u prethodnom slučaju, koristeći tehniku *star coordinates*. Koristimo algoritam normaliziranog spektralnog grupiranja koji koristi L_{rw} pri čemu smo odabrali graf 10-najbližih susjeda. Podatke smo prije grupiranja normirali kako bi bili u istom rasponu. Silhouette vrijednost iznosi 0.6299. ARI indeks iznosi 0.9212. Usporedbom sa egzaktnim rješenjem vidimo da je algoritam krivo grupirao samo 4 podatka.



(a) Originalni podatci (b) Dvije grupe

Slika 7: Rezultirajuće grupe podataka skupa *Banknotes*



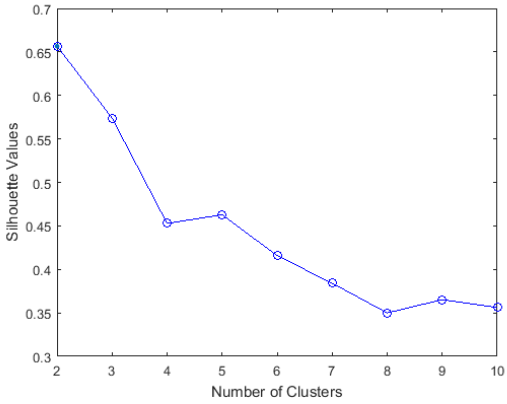
Slika 8: Silhouette graf podataka *Banknotes* uz $k = 2$

8 Optimalan broj grupa

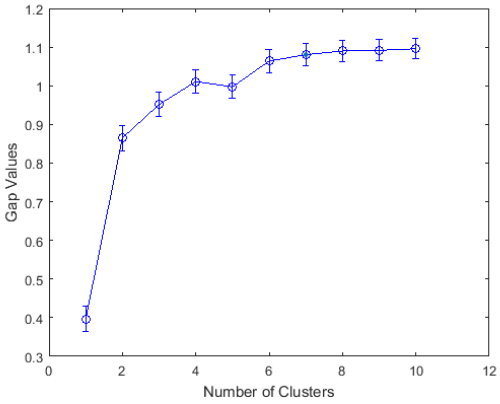
Primijetimo da smo u svim prethodnim primjerima mogli lako pretpostaviti optimalan broj grupa zbog prirode problema. U općenitom slučaju to nećemo moći. Odabir optimalnog broja grupa k oduvijek je bio veliki problem kod svih metoda grupiranja, te je do danas ostao neriješen. Situaciju dodatno komplicira uvođenje raznih parametara od kojih smo neke vidjeli u našim metodama. Osim što ne postoji automatski način za pronalazak točnog broja grupa, ne postoji niti formalna definicija "točnosti" grupa. Ipak, postoje heuristike i smjernice za koje se u praksi pokazalo da nam ponekad mogu pomoći. U svakom slučaju, moramo imati na umu da su to samo heuristike i biti pažljivi kod prihvaćanja rezultata. Jedna od heuristika koju smo već vidjeli u ovom poglavlju je upravo *silhouette kriterij* koji, osim što služi za validaciju rezultata grupiranja, može pomoći i kod odabira odgovarajućeg broja grupa. Sada ćemo promotriti još dvije heuristike: *kriterij svojstvenog skoka* i *Calinski-Harabasz kriterij*.

Opet ćemo promotriti primjer baze podataka *Banknotes* te računati optimalan broj grupa na temelju svojstava podataka prema sva tri navedena kriterija. Na Slici 9 je prikazano kako se vrijednosti kriterija mijenjanju obzirom na broj grupa. U slučaju silhouette i Calinski-Harabasz kriterija predloženi optimalni broj grupa je 2, dok je kod kriterija svojstvenog skoka taj broj jednak 7. Kako znamo da baza sadrži informacije o lažnim i pravim novčanicama koje se razlikuju po danim izmjerenim vrijednostima, zaključujemo da su Calinski-Harambash i silhouette kriterij u ovom slučaju predložili bolji broj grupa.

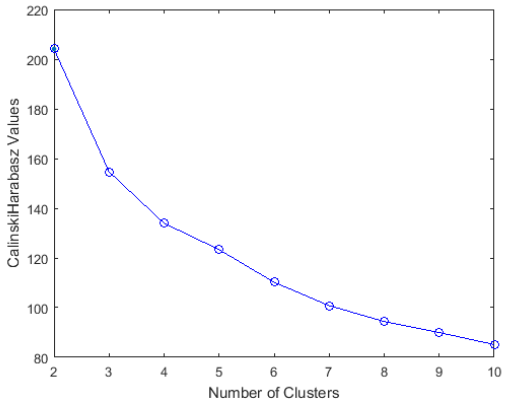
U slučaju baze podataka Iris, kriteriji svojstvenog skoka i Calinski-Harambasz kriterij predlažu tri optimalne grupe, dok silhouette kriterij predlaže dvije. Vrijednosti kriterija u odnosu na broj grupa možemo vidjeti na Slici 10. Kako znamo da baza sadrži podatke o tri podvrste biljke Iris od kojih dvije imaju male razlike u mjerama latica i čašice cvijeta, možemo zaključiti da su svi kriteriji dobro odredili broj grupa.



(a) Silhouette kriterij

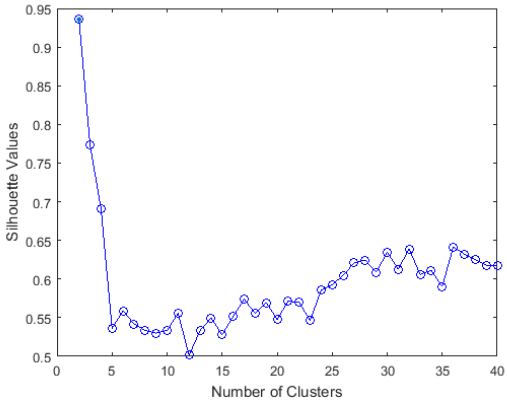


(b) Kriterij svojstvenog skoka

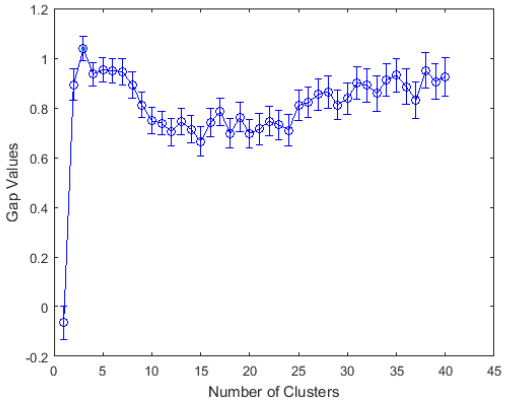


(c) Calinski-Harambasz kriterij

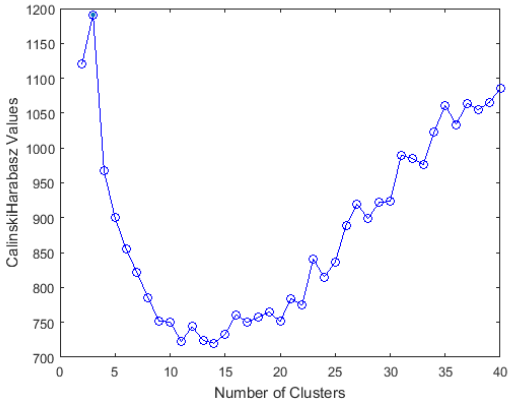
Slika 9: Analiza podataka Banknotes u svrhu traženja optimalnog broja grupa



(a) Silhouette kriterij



(b) Kriterij svojstvenog skoka



(c) Calinski-Harambasz kriterij

Slika 10: Analiza podataka baze Iris u svrhu traženja optimalnog broja grupa

9 Primjena spektralnih metoda u problemu segmentacije slike

Segmentacija slike je proces grupiranja piksela slike, tj. grupiranja osnovnih sastavnih dijelova slike koji imaju nešto zajedničko.
Najčešće se koristi za lociranje, izdvajanje ili prepoznavanje pojedinih objekata, detekciju rubova objekata, kompresiju, analizu sa gledišta baza podataka preko upita na podatcima izvedenih iz slike.

Prije same segmentacije moramo odgovoriti na nekoliko pitanja:

- Što određuje problem? Kako problem postaviti matematički?
- Kako riješiti problem?
- Kako prikazati rješenje i kako ga interpretirati?
- Kako možemo dobro segmentirati sliku bez prepoznavanja objekata?
- Što čini dobar segment?

Problem ćemo definirati na nekoliko načina koje ćemo vidjeti u nastavku. Rješavati ćemo ga uporabom metoda spektralnog grupiranja.
Rezultirajuću sliku ćemo u slučaju dvije grupe prikazati u crno-bijeloj boji, a u slučaju k grupa ćemo pomoću originalne slike pronaći boju koja će biti reprezentant pojedine grupe, a koju ćemo iskoristiti za bojanje svakog piksela koji pripada toj grupi.
Dobar segment čine regije koje su uniformne i homogene s obzirom na neku karakteristiku (svjetlina, boja, tekstura...). Interior bi trebao biti jednostavan bez malih rupa. Susjedne regije bi trebale imati značajno različite vrijednosti.
Vidjeti ćemo da nije lako pronaći segment koji ispunjava sve uvjete. U realnom slučaju obrade slika, striktno uniformne i homogene regije su pune malih rupa i imaju neravne rubove. Ljudi prirodno koriste prepoznavanje objekata kod segmentacije što program naravno ne može. Regije koje čovjek vidi kao homogene, ne moraju biti homogene kod digitalnog prikaza. Zbog toga rezultati nisu savršeni prema ljudskim standardima, ali cilj je automatizirati rješenje problema.

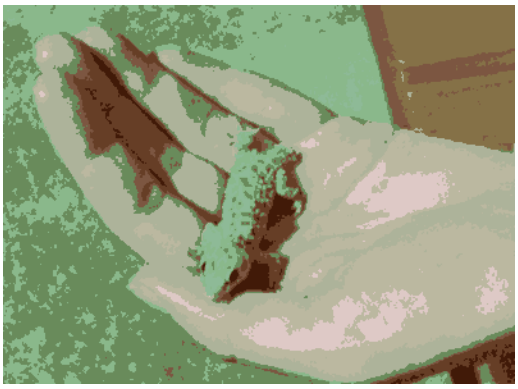
Postupak je sljedeći: sliku spremimo kao matricu podataka dimenzije $m \times n \times d$, gdje je m širina slike u pikselima, n visina, a d su RGB vrijednosti, odnosno intenzitet crvene, zelene i plave boje. Dakle podatci su peterodimenzionalni. Za računanje sličnosti koristimo Gaussovu funkciju pa podatke moramo normirati obzirom da se raspon vrijednosti prve dvije koordinate (pozicije dane prirodnim brojevima) uvelike razlikuje od raspona zadnje tri (RGB vrijednosti iz intervala $[0, 1]$), inače bi rezultat više ovisio o poziciji piksela nego o njegovoj boji.
Uočavamo da je naš skup podataka u ovom slučaju vrlo velik čak i za male dimenzije. Na primjer, originalne slike koje smo obrađivali su dimenzija otprilike 300×400 piksela, što rezultira s $1.728 \cdot 10^{12}$ 5-dimenzionalnih podataka za računanje. Zato koristimo graf k-najbližih susjeda (stavimo $k = 10$) i metode za rijetke matrice kod spektralnog grupiranja. Jedna takva metoda je Lanczos metoda ili metoda Krylovljevih potprostora za računanje svojstvenih vektora Laplaceove matrice, a koristi je MATLAB-ova funkcija *eigs*. U zadnjem koraku pokrenemo k-means algoritam nad retcima matrice čiji su stupci prvih k svojstvenih vektora odgovarajuće Laplaceove matrice. Rezultat prikazujemo pomoću MATLAB-ove funkcije *imshow*.

Na Slici 11 je primjer segmentacije pomoću spektralnog grupiranja koje koristi matricu L_{rw} pri čemu je zadani broj grupa jednak 2. Sa lijeve strane su originalne slike, a sa desne segmentirane. Rezultirajuće slike su pojednostavljenjene i svedene na samo dvije boje, crnu i bijelu. Detalji su gotovo potpuno izostavljeni i zanemarena je pozadina.

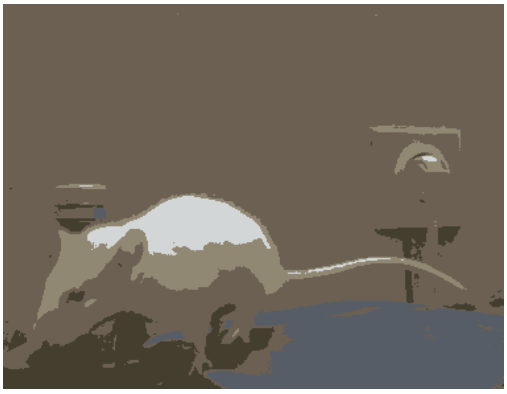
Slika 11: Segmentacija pomoću spektralnog grupiranja koje koristi L_{rw} za $k = 2$.

Sljedeći problem koji rješavamo je prepoznavanje objekata. Zanima nas možemo li segmentirati slike preko spektralnog grupiranja uz prilagođeni broj grupa k tako da

dobijemo smislene regije koje će određivati različite objekte.
Na Slici 12 vidimo rezultate.
Slika 12a prikazuje ruku i guštera, što možemo zaključiti tek uz broj grupa $k = 9$. Na slici 12b vidimo miša kojeg bez problema raspoznajemo uz $k = 4$ broj grupa. Slika 12c vrlo jasno prikazuje pauka čak i uz mali broj grupa $k = 3$, dok na Slici 12d vidimo mačku.
Zaključujemo da kvaliteta segmentacije ovisi najviše o bojama originala i o kontrastima, ali na nju utječe i količina detalja i svjetlina. Blaga zamućenost originala ne smeta.

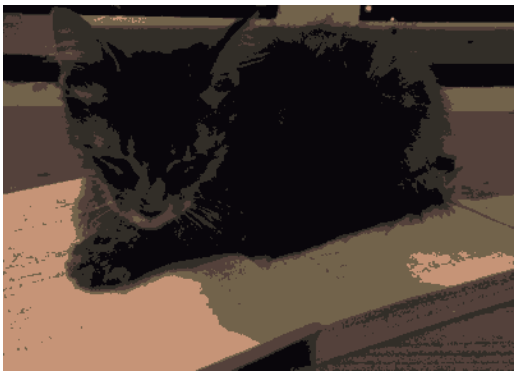


(a) $k = 9$



(b) $k = 4$

(c) $k = 3$

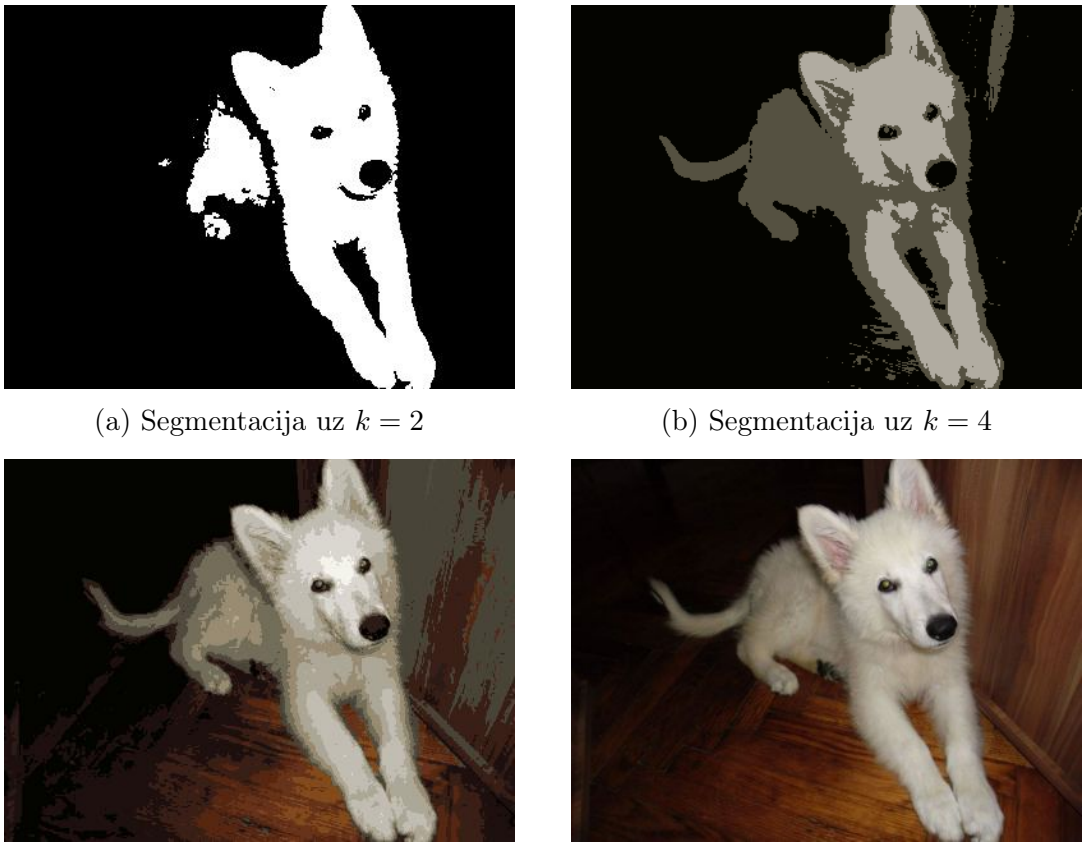


(d) $k = 5$

Slika 12: Segmentacija u svrhu određivanja objekata

Za raspoznavanje objekata nam odgovara što jednostavniji prikaz pa smo tražili minimalan broj grupa uz koji bez problema možemo raspoznati objekte.

Uz povećanje broja grupa, razlike između segmentirane slike i originala su sve manje. To pokazujemo na Slici 13.
Ako zadamo $k = 2$, jedini objekt koji možemo prepoznati je pas, te ga možemo izdvojiti od pozadine, svi detalji su ovdje zanemareni. Prikaz je puno bolji već uz $k = 4$ gdje slika počinje dobivati dubinu pomoću sjena. Za $k = 15$ grupa, i dalje vidimo dosta razlika u odnosu na original prikazan na četvrtoj slici, ali u ovom slučaju imamo detaljniji prikaz, na primjer vidimo teksturu dlake i parketa. Broj grupa uvijek zadajemo u ovisnosti o problemu kojeg rješavamo.



(a) Segmentacija uz $k = 2$

(b) Segmentacija uz $k = 4$

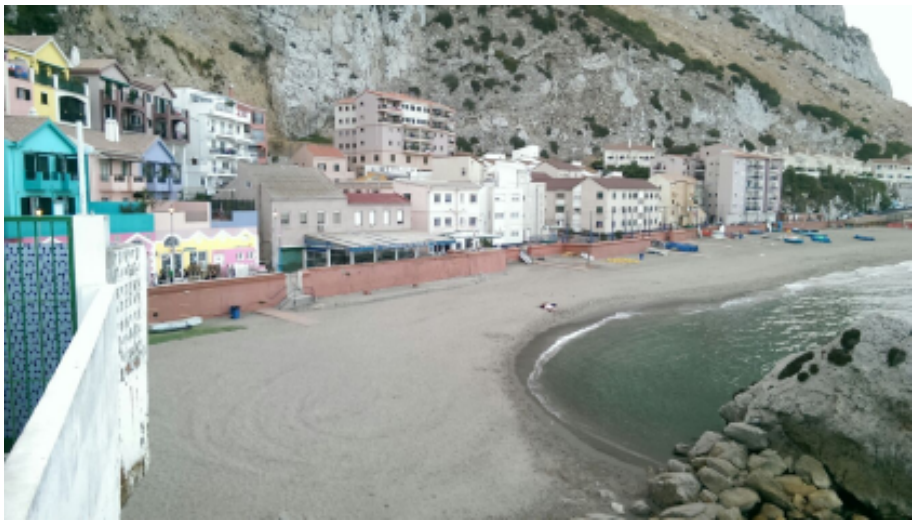
(c) Segmentacija uz $k = 15$

(d) Originalna slika

Slika 13: Segmentacija ovisno o broju grupa

Na kraju pokazujemo kako se segmentacija može koristiti za kompresiju. Slika 14 prikazuje originalnu i kompresiranu sliku. Razlika u kvaliteti je primjetna, ali svi detalji obrađene slike su i dalje dobro vidljivi. Uglavnom homogeni dijelovi slike sa dobro definiranim rubovima su jako dobro grupirani, dok to nije slučaj na dijelovima

gdje imamo puno nijasni iste boje (pijesak, more). Uspjeli smo smanjiti veličinu sa $240KB$ na $123KB$.



(a) Originalna slika

(b) Kompresirana slika

Slika 14: Segmentacija u svrhu kompresije

Literatura

[1] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Kluwer Academic, Publishers Norwell, 1981.

[2] R. Diestel, *Graph Theory (Graduate Texts in Mathematics)*, 5th ed. 2010. Corr. 3rd 2012 Edition

[3] R. O. Duda, P. E. Hart, D. G. Stork, *Pattern Classification*, 2nd Edition, Wiley, 2001.

[4] R. A. Fisher, *The use of multiple measurements in taxonomic problems*, Annual Eugenics, 7, Part II(1936) 179-188.

[5] B. Flury, H. Riedwyl, *Angewandte multivariate Statistik*, Stuttgart, Fischer, 1983.

[6] J. Friedman, T. Hastie, R. Tibshirani,*The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*, Springer Series in Statistics, 2009.

[7] G. Gan, C. Ma, J. Wu, *Data Clustering: Theory, Algorithms, and Applications*, SIAM, Philadelphia, 2007.

[8] R. A. Horn, C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, 1985.

[9] A. K. Jain, *Data clustering: 50 years beyond k-means*, Patern Recognition Letters, 31 (2010), 651-666.

[10] E. Kandogan, *Star Coordinates: A Multi-dimensional Visualization Technique with Uniform Treatment of Dimensions*, In Proceedings of the IEEE Information Visualization Symposium, Late Breaking Hot Topics (2000)

[11] R. M. Karp, *Reducibility Among Combinatorial Problems*, Complexity of Computer Computations (R.E. Miller and J.W. Thatcher, eds.), Plenum Press (1972), 85–103.

[12] J. Kogan, *Introduction to Clustering Large and High-Dimensional Data*, Cambridge University Press, 2007.

[13] M. Lichman, *UCI repository of machine learning databases*, University of California, Department of Information and Computer Science, Irvine, CA, 1998.

[14] U. von Luxburg, *A Tutorial on Spectral Clustering*, Statistics and Computing 17 (4), 2007.

[15] M. Meila, J. Shi, *A random walks view of spectral segmentation*, IEEE International Conference on Artificial Intelligence and Statistics, 2001.

[16] J. Malik, J. Shi, *Normalized Cuts and Image Segmentation*, IEEE Transaction on Pattern Analysis and Machine Intelligence Vol. 22 No. 8, 2000.

[17] P. J. Rousseeuw, *Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis, Computational and Applied Mathematics*, Journal of Computational and Applied Mathematics 20, North-Holland (1987), 53-65.

[18] W. L. Ruzzo, K. Y. Yeung, *Details of the Adjusted Rand index and Clustering algorithms*, Bioinformatics, 2001.

[19] K. Sabo, *Center-based l1-clustering method*, International Journal of Applied Mathematics and Computer Science, 24 (2014), 151–163.

[20] K. Sabo, R. Scitovski, *Interpretation and optimization of the k-means algorithm*, Applications of mathematics, No. 4, Osijek (2012), 391–406.

[21] K. Sabo, R. Scitovski, I. Vazler, *Grupiranje podataka: klasteri*, Osječki matematički list 10 (2010), 149–178.

[22] M. Stoer, F. Wagner, *A Simple Min Cut Algorithm*, Algorithms - ESA '94, LNCS 855 (1994), 141-147.

[23] L. Trevisan, *Spectral Graph Theory and Graph Partitioning*, Computer Science Department, Stanford University, NSF CCF-1161812, 2011.

Sažetak

Grupiranje ili klasteriranje je postupak organiziranja podataka u disjunktne grupe (klastere) takve da su podatci unutar jedne grupe međusobno slični te različiti od podataka u drugim grupama. Sličnost mjerimo preko neke funkcije sličnosti pa podatke i odnose među njima prikazujemo pomoću grafa sličnosti. Imamo više vrsta takvih grafova, a razlikuju se po načinu na koji oblikuju lokalno susjedstvo. Svakom od tih grafova možemo pridružiti Laplaceovu matricu koja predstavlja ključni matematički objekt u spektralnim metodama. Spektralno grupiranje koristi svojstvene vektore Laplaceovih matrica za promjenu reprezentacije originalnih podataka i po tome je dobilo ime. Na promijenjenim podatcima se zatim primijeni k-means algoritam koji nam daje konačne grupe. Na prvi pogled, spektralne metode se mogu činiti kao proširenje k-means algoritma, ali to zapravo nije točno. Ključni korak kod spektralnih metoda je promjena reprezentacije podataka na vrlo pametan način, tako da nakon toga čak i jednostavan algoritam kao k-means može trivijalno otkriti grupe. Spektralno grupiranje se u praksi pokazalo vrlo dobro, čak i na skupovima koji nisu dani normalnom distribucijom i nisu konveksni, međutim rezultati ovise o dosta parametra i zapravo nemamo nikakvu teorijsku garanciju na kvalitetu rješenja. S obzirom da nam ove metode omogućuju da vrlo komplicirane probleme riješimo na vrlo jednostavan način i to preko standardnih operacija linearne algebre, to je mana preko koje možemo prijeći.

Ključne riječi: spektralno grupiranje, graf sličnosti, Laplaceova matrica, k-means, segmentacija slike

Summary

Clustering is a process of organizing data into distinct groups (clusters) such that data points in one group are similar to each other and dissimilar to the data points in the other groups. We measure similarity through some similarity function and represent data by similarity graphs. There are several types of such graphs and they differ by the way they model local adjacency. We use those graphs to obtain Laplace matrix, which is the crucial mathematical object for spectral clustering methods. Spectral clustering use the eigenvectors of Laplace matrix in order to change the representation of the original data, and that's the reason for naming. On the changed data, we just run k-means algorithm and we end up with final clusters. On the first look, we could take spectral methods as k-means improvement, but it's not really true. The main step in spectral methods is very smart change of representation, so that afterwards, even a simple algorithm like k-means can trivially detect groups. Spectral clustering shows very well results in application, even on a data sets which are not normally distributed and those that are nonconvex. However, the results depend on many parameters and we don't have any kind of theoretical support to guarantee it's quality. Considering this methods allow us to solve really complicated problems in a really simple way, by standard operations of linear algebra, it's the disadvantage that we can live with.

Keywords: spectral clustering, similarity graph, Laplace matrix, k-means, image segmentation

Životopis

Rođena sam u Zagrebu. Osnovnu školu i opću gimnaziju sam završila u Svetom Ivanu Zelini. Preddiplomski studij matematike sam završila na prirodoslovno-matematičkom fakultetu u Zagreb nakon čega sam upisala smjer matematika i računarstvo na Odjelu za matematiku u Osijeku.