Sadržaj

Uvod

U ovome radu promatrat ćemo metode grupiranja s posebnim fokusom na spektralne metode grupiranja.

Rad je podijeljen na šest poglavlja kroz koja se odrađuju temeljni koncepti i primjene grupiranja.

U prvom poglavlju objašnjava se važnost odabira mjere sličnosti pri grupiranju podataka i konstruiranje grafa sa odgovarajućom matricom.

U drugom poglavlju obrađen je algoritam grupiranja k-sredina u usporedbi s naprednim algoritmom k-sredina. Također su obrađene i dvije empirijske metode koje se koriste na određivanje optimalnog broja skupina kod grupiranja podataka.

U trećem poglavlju predstavljaju se temelji spektralnog grupiranja poput Laplaceove matrice i njenih svojstava popraćenih dokazima, svojstvenih vrijednosti i svojstvenih vektora Laplaceove matrice. Dodatno, predstavljena je i Lanczosova metoda za pronalaženje aproksimacija svojstvenih vrijednosti budući da se koristi u obrađenim metodama spektralnog grupiranja.

U četvrtom poglavlju obrađene su metode spektralnog grupiranja podijeljene na metode biparticije grafa i metode grupiranja u više skupina. Predstavljena je temeljna metoda grupiranja prema Fiedlerovom vektoru Laplaceove matrice uz tri alternativne metode spektralnog grupiranja, metode normaliziranog reza, NJW metode i KVV metode.

U petom poglavlju uspoređuju se navedeni algoritmi na IRIS skupu podataka uz analizu koristeći ARI i NMI mjere sličnosti.

U zadnjem poglavlju proučava se primjena spektralnog grupiranja u vidu segmentacije slika i grupiranja višedimenzionalnih podataka na MINST skupu podataka.

1. Grupiranje podataka

Grupiranje podataka je metoda nenadziranog strojnog učenja u kojoj se podatci grupiraju u skupine ovisno o njihovoj međusobnoj sličnosti, bez unaprijed poznatih rješenja. Struktura grupiranja proizlazi iz samih podataka.

Za mnoge algoritme grupiranja potrebno je podatke predstaviti u obliku grafa gdje su točke čvorovi u grafu, a bridovi poprimaju vrijednosti sličnosti te nam predstavljaju odnose između čvorova.

1.1. Mjere sličnosti

Mjera sličnosti predstavlja vrijednost kojom opisujemo koliko su podatci povezani, to jest koliko su slični jedno drugome. To uvelike utječe na to kako ćemo konstruirati graf, a time i na sami rezultat grupiranja. Stoga je ključno odabrati primjerenu mjeru sličnosti kako bismo dobili što kvalitetnije grupiranje. Mjerom sličnosti koja nije prigodna podatcima dobivamo loše odvojene skupine.

Mjera sličnosti na temelju geometrijske udaljenosti točaka naziva se **euklidska udaljenost**, izraz koristi koordinate točaka $\boldsymbol{x}$ *i* $\boldsymbol{y}$ kojom dobivamo najmanju udaljenost između dvije točke:

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2} \tag{1}$$

Jedan primjer kako možemo još računati mjeru sličnosti između vektora je **kosinusna mjera sličnosti**. Uzima se u obzir kosinus kuta između vektora $\boldsymbol{x}$ i $\mathbf{y}$ kojeg dobivamo kao skalarni umnožak podijeljen umnoškom njihovih duljina:

$$\cos \alpha = \frac{\boldsymbol{x}^T \boldsymbol{y}}{\|\boldsymbol{x}\| \|\boldsymbol{y}\|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}} \tag{2}$$

Mjera sličnost temeljena na **Gaussovoj jezgrenoj funkciji** između dvije točke $\boldsymbol{x}$ *i* $\boldsymbol{y}$ izražava se kao:

$$\kappa(\boldsymbol{x}, \boldsymbol{y}) = \exp(\ -\frac{\|\ \boldsymbol{x} - \ \boldsymbol{y}\|^2}{2\sigma^2}) \tag{3}$$

gdje imamo dodatni parametar σ kojim označujemo širinu pojasa [1].

Ova formula ovisi o parametru σ, pri čemu veća vrijednost σ označuje da i daljnje točke imaju veću razinu sličnosti, dok manja vrijednost da su samo bliske točke slične. Rezultat ove funkcije sličnosti su početne udaljenosti koje su preslikane u raspon $[0,1]$. Na taj način jasno vidimo kada su točke vrlo udaljene i vrlo blizu.

Svaka mjera sličnosti prigodna je za određeni tip podataka. Euklidska udaljenost (1) koristi se za podatke koji su centrirani, zbijeni, kosinusna sličnost (2) se koristi za višedimenzionalne podatke, a Gaussova jezgrena funkcija (3) za podatke s nelinearnom strukturom.

Nužno je da odabir metrike za mjeru sličnosti prilagodimo podatcima kako bismo dobili najbolju podjelu podataka u skupine.

1.2. Konstrukcija matrice grafa

Kako bismo mogli provesti grupiranje podataka, prvi korak je konstruirati graf na temelju dostupnog skupa podataka nad kojim kasnije provodimo particiju grafa. U tu svrhu formiramo neusmjereni, težinski graf definiran kao

$$G\ =\ (V, E)$$

gdje je V skup vrhova koji predstavljaju pojedinačne podatke, a E skup bridova koji označavaju povezanost između parova podataka [2].

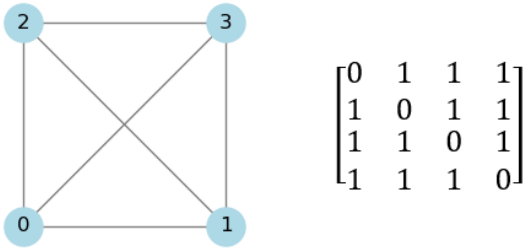
Takav graf matrično prikazujemo u obliku matrice susjedstva. Matrica susjedstva u tom slučaju je kvadratna matrica s elementima A_{ij} kojim nam označuju povezanost čvorova i i j gdje su vrijednosti matrice 1 ili 0 ovisno o tome jesu li točke povezane ili ne, pri čemu su elementi na dijagonali jednaki 0 [2].

U obzir moramo uzeti i tip grafa koji želimo konstruirati. Potpuno povezani graf pokazuje nam povezanost svih čvorova. U potpuno povezanom grafu vidimo sve odnose u grafu, no ne i njihovu jačinu, tako da se koristi za male skupove podataka u kojima nema šuma.

Popratnu matricu susjedstva konstruiramo prema izrazu [3]:

$$A_{ij} = \begin{cases} 1, & (i,j) \in E \\ 0, & (i,j) \notin E \end{cases}. \tag{4}$$

Primjer jednostavnog povezanog grafa i pripadne matrice susjedstva prikazan je na slici (Slika 1.1).



Slika 1.1 Potpuno povezani graf i pripadna matrica susjedstva

Budući da matrice susjedstva mogu biti velikih dimenzija kod velikih skupova podataka, često se koristi graf k-najbližih susjeda. U toj varijanti matrica susjedstva sadrži samo poveznice čvorova sa svojih k najbližih, najsličnijih čvorova. Kao rezultat dobivamo rjeđu matricu u kojoj je i dalje očuvana lokalna struktura podatka. Prednost je brža obrada budući da je manje osjetljiva na udaljenije točke te daje prednost sličnijim podatcima. Velik utjecaj na rezultat daje i parametar k. Ako je k premalen, postoji rizik da se izgube informacije o potencijalno bitnim vezama, dok s prevelikim k možemo dobiti previše povezan graf s kojim nismo smanjili računalnu složenost.

Varijanta matrice susjedstva je i ε-graf u kojem povezanost između čvorova uzimamo u obzir prema definiranom pragu udaljenosti ε. Za svaki par čvorova, dodaje se poveznica, brid ukoliko je njihova udaljenost manja od zadanog praga ε. Za razliku od grafa k-najbližih susjedstva, gustoća ovog grafa ovisi o prostornoj raspodjeli podataka jer se selekcijski parametar temelji na geometrijskoj udaljenosti, no i dalje dobivamo rjeđi graf koji je lakši za obradu.

Ukoliko povezanost točaka tretiramo kao sličnost istih, dobivamo matricu sličnosti, to jest težinsku matricu susjedstva gdje je težina brida jedna mjeri sličnosti između dva vrha. Ovisno o mjeri sličnosti koju koristimo dobivamo različite matrice, no sve dijele iste strukturalne karakteristike kao i matrice susjedstva.

Na sljedećem primjeru prikazanom na slici (Slika 1.2) za izračunavanje matrice sličnosti koristimo se metrikom euklidske udaljenosti i potpuno povezanim grafom. Budući da veće

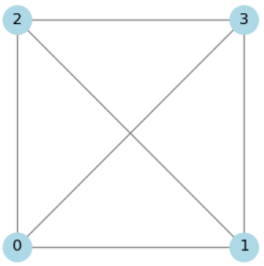
udaljenosti označavaju manju povezanost potrebno je izračunati inverz euklidske udaljenosti S_{ij} definirane formulom:

$$S_{ij} = \frac{1}{d_{ij} + \varepsilon} \tag{5}$$

gdje je

$$d_{ij} = \sqrt{\sum_{i=1}^{d}(x_i - y_i)^2}$$

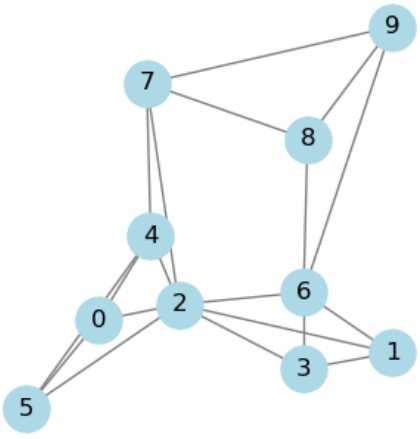
izraz za euklidsku udaljenost, a malu vrijednost ε dodajemo kako bismo izbjegli dijeljenje s nulom. [4]



$$A = \begin{bmatrix} 0 & 1 & 1 & 0.71 \\ 1 & 0 & 0.71 & 1 \\ 1 & 0.71 & 0 & 1 \\ 0.71 & 1 & 1 & 0 \end{bmatrix}$$

Slika 1.2 Potpuno povezani graf i pripadna matrica sličnosti

Ukoliko se radi o većim matricama, korisno je napraviti graf k- najbližih susjedstva kojim smanjujemo broj bridova, a s time i računalnu složenost potrebnu za obradu takve matrice. Na sljedećem primjeru od 10 točkama prikazan je graf dobiven iz matrice sličnosti gdje smo uzeli $k = 3$. Svaki redak matrice predstavlja tri najbliža susjeda pripadnog čvora, dok u stupcima vidimo koji čvorovi su odabrali pojedini čvor za susjeda. Tako na primjer na slici (Slika 1.3) čvor 2. su kao susjeda odabrala 7 ostalih čvorova, što vidimo i iz grafa.



$$A = \begin{bmatrix} 0 & 0 & 0.49 & 0 & 0.40 & 0.35 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.18 & 0.45 & 0 & 0 & 0.38 & 0 & 0 & 0 \\ 0.49 & 0.18 & 0 & 0.29 & 0.54 & 0.22 & 0.32 & 0.18 & 0 & 0 \\ 0 & 0.45 & 0.29 & 0 & 0 & 0 & 0.53 & 0 & 0 & 0 \\ 0.40 & 0 & 0.54 & 0 & 0 & 0.19 & 0 & 0.26 & 0 & 0 \\ 0.35 & 0 & 0.22 & 0 & 0.19 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.38 & 0.32 & 0.53 & 0 & 0 & 0 & 0 & 0.26 & 0.14 \\ 0 & 0 & 0.18 & 0 & 0.26 & 0 & 0 & 0 & 0.24 & 0.16 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.26 & 0.24 & 0 & 0.29 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.14 & 0.16 & 0.29 & 0 \end{bmatrix}$$

Slika 1.3 Povezani graf i pripadna matrica sličnosti

Grupiranje podataka je zapravo problem particioniranja težinskog grafa, to jest podjele podataka u skupine u kojima imamo snažnije povezane čvorove, a slabije veze između samih skupina. Nad takvim grafovima primjenom spektralnih metoda grupiranja otkrivamo skupine i kompleksne strukture unutar podataka.

2. Prikaz nekih osnovnih metoda grupiranja

Metode grupiranja dijelimo prema vrsti grupiranih podataka, broju skupina i načinu definiranja sličnosti podataka, a mogu biti hijerarhijske ili particijske.

U nastavku je opisan particijski algoritam k-sredina, temeljna metoda grupiranja koja se često koristi i u kombinaciji s drugim metoda radi dobivanja realnijih rezultata, a metoda spektralnog grupiranja koje se temelje na spektralnog teoriji grafova dotaknut ćemo se u sljedećem poglavlju.

Dodatno, u ovom poglavlju obrađuju se i metode za određivanje optimalnog broja skupina.

2.1. Algoritam k-sredina

Kod algoritma k-sredina postupak grupiranja je drugačiji nego kod spektralnog grupiranja. Određivanje skupina temelji se isključivo na prostornoj udaljenosti između podataka koja se izračunava preko euklidske formule za udaljenost.

Radi se o iterativnoj metodi koja iz skupa od N podataka $X = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$ započinje nasumičnim odabirom k središta $\boldsymbol{\mu} = \{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_k\}$ koja nam predstavljaju početna središta skupina. Broj $k = 1, \dots K$ je zadani broj konačnih skupina, a pripadnost pojedine točke skupini opisuje se prema izrazu:

$$r_{nk} = \begin{cases} 1 & ako\ k = \arg min_j \left\|\boldsymbol{x}_n - \boldsymbol{\mu}_j\right\|^2 \\ 0 & ostalo \end{cases} \tag{6}$$

gdje je $\boldsymbol{x}_n$ točka, r_{nk} binarna oznaka pripadnosti točke $\boldsymbol{x}_n$ k-toj skupini,a $\boldsymbol{\mu}_j$ nam označava središte *j-te* skupine [4].

Koristeći prethodni izraz možemo definirati ciljnu funkciju J [4] koja nam opisuje kvadriranu sumu udaljenosti točaka do svojih dodijeljenih središta kao:

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||\boldsymbol{x}_n - \boldsymbol{\mu}_k||^2. \tag{7}$$

Nakon što su sve točke pridružene najbližem središtu, ažuriraju se pozicije središta. Kako bismo dobili minimalnu vrijednost izraza (7) potrebno je ciljnu funkciju J derivirati po varijabli $\boldsymbol{\mu}_k$ te tako dobivamo uvjet za dobivanje stacionarne točke:

$$\frac{\partial J}{\partial \boldsymbol{\mu}_k} = 0$$

kojeg kada raspišemo dobivamo izraz kojeg koristimo za izračunavanje novih središta [4]:

$$\frac{\partial J}{\partial \boldsymbol{\mu}_k}\left(\sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\left||\boldsymbol{x}_n - \boldsymbol{\mu}_k\right||^2\right) = 0$$

$$2\sum_{n=1}^{N} r_{nk}(\boldsymbol{x}_n - \boldsymbol{\mu}_k) = 0$$

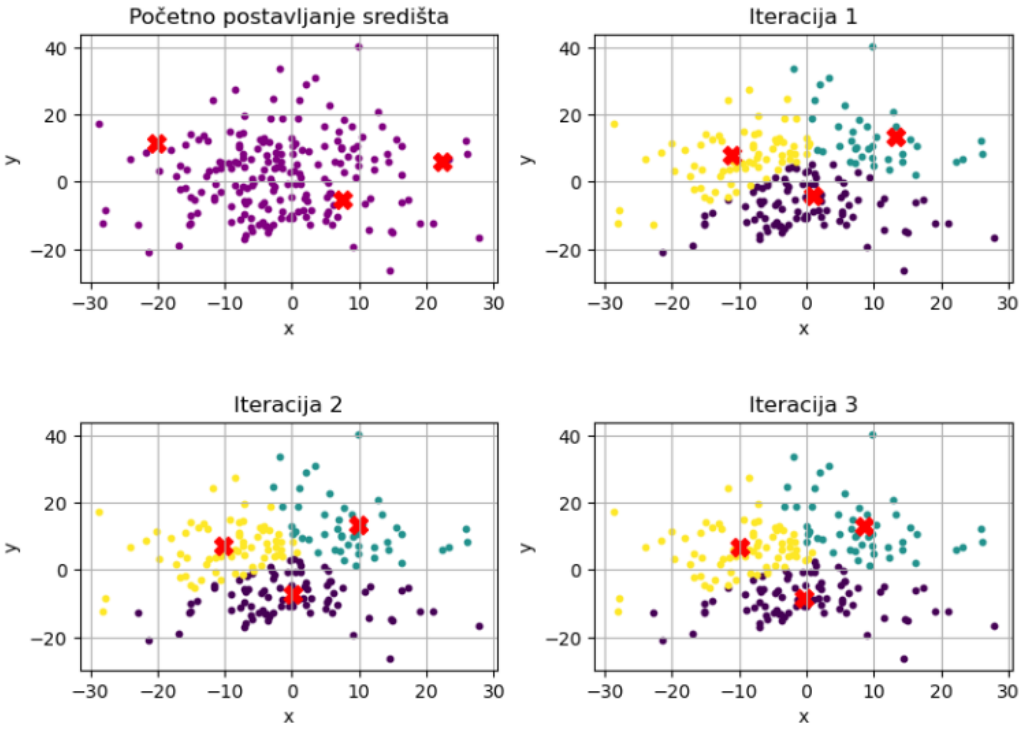
$$\sum_{n=1}^{N} r_{nk}\boldsymbol{x}_n - \boldsymbol{\mu}_k \sum_{n=1}^{N} r_{nk} = 0$$

To jest, nova vrijednost središta izračuna se kao srednja vrijednost svih točaka koje trenutno pripadaju skupini što se može i prikazati kao [4]:

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} r_{nk}\boldsymbol{x}_n}{r_{nk}}. \tag{8}$$

Postupak ažuriranja središta ponavlja se sve dok središta ne prestanu mijenjati svoju poziciju, točnije dok svaki podatak nije grupiran u skupinu čije mu je središte najbliže. Na slici (

Slika 2.1) prikazana su prva tri koraka iteracija uz početno postavljanje središta te kako se središta prilagođavaju nakon svake iteracije dok ne dođu do konačne pozicije.



Slika 2.1 Prikaz ažuriranja središta kod algoritma k-sredina

U situacijama kada su skupovi podataka veći i složeniji, može doći do slučajeva kada nekom inicijalno određenom središtu nije dodijeljena niti jedna točka, to jest svim točkama su bliža neka druga središta. Kako bismo to izbjegli, koristimo se naprednijom verzijom algoritma k-sredina koji se fokusira na efikasnije pronalaženje početnih središta.

Proces započinje tako da nasumično odaberemo točku za prvo početno središte, te za svako novo središte $\boldsymbol{\mu}_k$ izračunavamo udaljenost $D(\boldsymbol{x}_n)$ do najbližih trenutno odabranih središta $\boldsymbol{\mu}$ za svaku točku $\boldsymbol{x}_n$ koja nije odabrana kao središte:

$$D(\boldsymbol{x}_n) = \min_{\boldsymbol{\mu}_k \in \boldsymbol{\mu}} \|\boldsymbol{x}_n - \boldsymbol{\mu}_k\|^2 \tag{9}$$

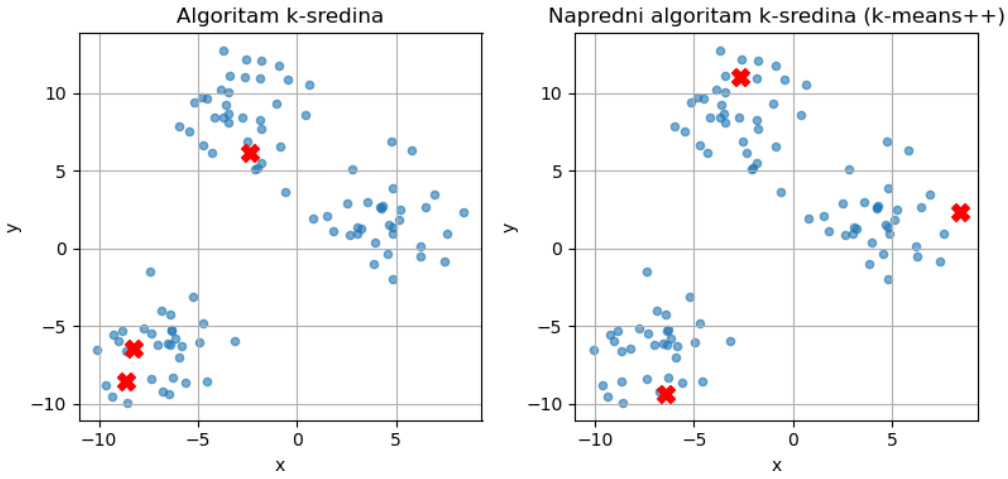
Tada izabiremo novo središte na temelju vjerojatnosti koju smo dobili preko sljedeće formule:

$$p(\boldsymbol{x}) = \frac{D(\boldsymbol{x})^2}{\sum_{x \epsilon X} D(\boldsymbol{x})^2} \tag{10}$$

Novo središte se i dalje nasumično odabire, ali središta koja su udaljena od već odabranih točaka imaju veću vjerojatnost da budu odabrana.

Proces se ponavlja dok ne dobijemo k središta [5].

Na slici (Slika 2.2) prikazana je usporedba inicijalizacije središta koristeći običan algoritam k-sredina i koristeći napredni algoritam k-sredina. Na lijevom grafu vidimo kako su pozicije dva od tri nasumična središta vrlo blizu te će biti potrebno velik broj iteracija kako bismo došli do optimalnog grupiranja podataka, dok na desnom grafu vidimo da je inicijalno postavljena središta u pogodnijem rasporedu za daljnje iteracije.



Slika 2.2 Usporedba početnog postavljanja središta kod običnog algoritma k-sredina i naprednog algoritma k-sredina

2.2. Optimalni broj skupina

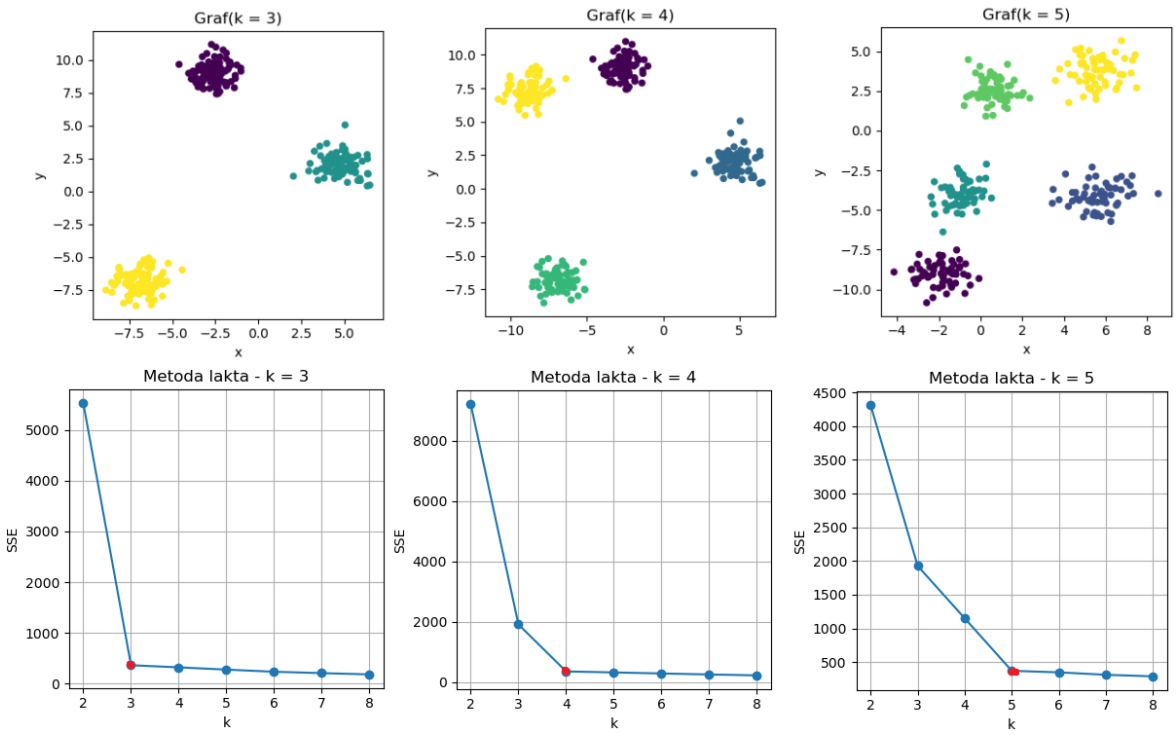
Pri grupiranju mnogo vanjskih parametara utječe na konačni rezultat grupiranja poput odabira mjere sličnosti, načina konstrukcija grafa te strukture samih podataka koji ne moraju biti jednostavno razdvojivi u skupine. U pravilu, nemamo predefinirani točan broj skupina prema kojem se možemo ravnati. U odabiru optimalnog broja skupina k koristimo se provjerenim metodama.

U nastavku predstavit ćemo dvije metode za određivanje broja skupina koje se koriste u praksi, metodu lakta i metodu siluete.

Metoda lakta je empirijska metoda koja se primjenjuje u kombinaciji s nekom metodom grupiranja, najčešće s algoritmom k-sredina. Za svako središte skupine izračunava se ukupna suma kvadrata udaljenosti do točaka unutar skupine (engl. SSE – Sum of Squared Errors):

$$SSE = \sum_{i=1}^{k} \sum_{\boldsymbol{x}_j \in c_i} \left\| \boldsymbol{x}_j - \boldsymbol{\mu}_i \right\|^2 \tag{11}$$

gdje je k odabrani broj skupina, c_i,oznaka za $i-tu$ skupinu, a $\boldsymbol{\mu}_i$ središte pripadne skupine i [6].



Slika 2.3 Prikaz rezultata metode lakta za više skupova podataka

Za svaki broj skupina u proizvoljnom rasponu provodimo odabranu metodu grupiranja te izračunavamo vrijednost SSE preko izraza (11) koje potom vizualiziramo na grafu. Na sljedećoj slici (Slika 2.3) prikazani su rezultati grupiranja metode k-sredina na tri skupa podataka koji su prirodno podijeljeni u 3,4, i 5 skupina i izračunate SSE vrijednosti za svaku skupinu u rasponu $k = [2,3,4,5,6,7,8]$. Uočljiv je oblik lakta, točka nakon koje se SSE vrijednosti počinju drukčije ponašati, na grafovima označena crveno. Upravo ta točka nam označuje optimalan broj skupina, a njihovi pripadni grupirani podatci prikazani su na grafu iznad.

Još jedna metoda je **metoda siluete** [7]. Kod nje izračunavamo srednju vrijednost udaljenosti točke od svih ostalih točaka u skupini a_i :

$$a_i = \frac{1}{|c_i| - 1} \sum_{\boldsymbol{x}_j \in c_i, i \neq j} d(\boldsymbol{x}_i, \boldsymbol{x}_j) \tag{12}$$

i srednju vrijednost udaljenosti točke do svih točaka u najbližoj susjednoj skupini b_i:

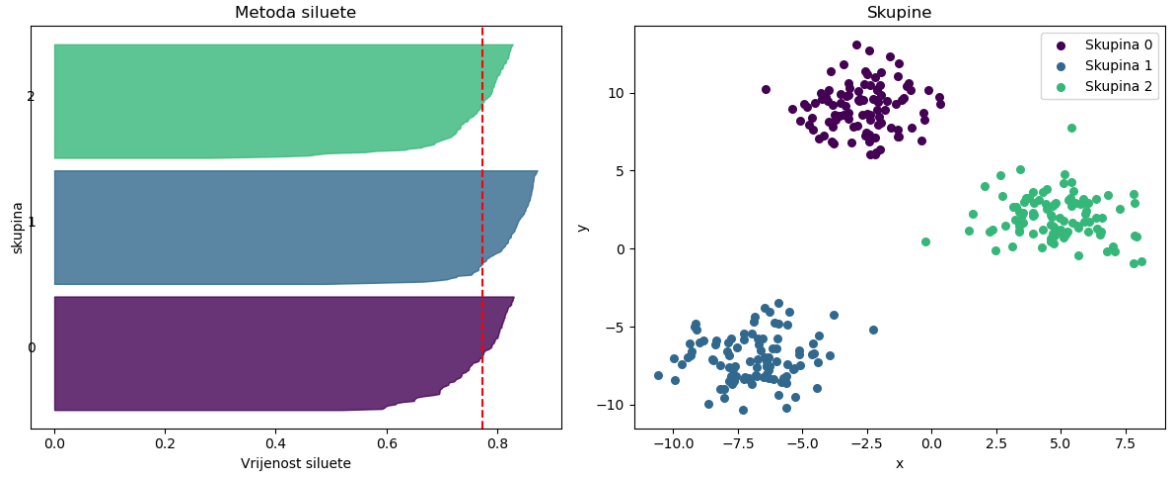
$$b_i = \min_{c_j \neq c_i} \frac{1}{|c_j|} \sum_{\boldsymbol{x}_j \in c_j} d(\boldsymbol{x}_i, \boldsymbol{x}_j) \tag{13}$$

gdje je c_i skupina pripadne točke $\boldsymbol{x}_i$, $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ euklidska udaljenost između točaka $\boldsymbol{x}_i$ *i* $\boldsymbol{x}_j$, a c_j za koju vrijedi $c_j \neq c_i$.

Koeficijent siluete s_i za točki i koristeći (12) i (13) izračunava kao:

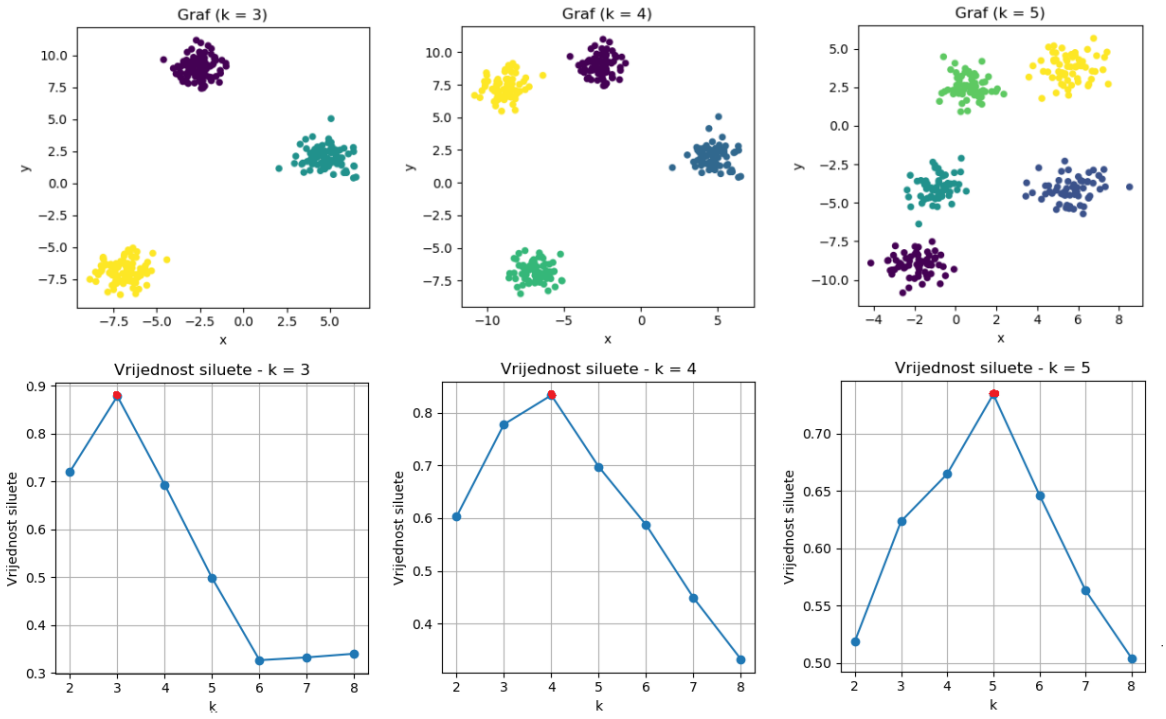
$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} \tag{14}$$

Koeficijent siluete opisuje odnos povezanosti točke unutar grupe i odvojenost od ostalih grupa. Koeficijenti se dobivaju u rasponu od $[-1,1]$, manje vrijednosti označavaju da točka je loše dodijeljena skupini, to jest nalazi se daleko od središta dodijeljene skupine, a više vrijednosti da je točka dobro dodijeljena skupini, to jest nalazi se daleko od ostalih skupina [7]. Kao što vidimo na sljedećem primjeru na slici (Slika 2.4), za svaku točku izračunata je vrijednost siluete prema izrazu (14), a budući da su podatci jasno vizualno grupirani, sve vrijednosti su vrlo visoke. Prosječna vrijednost siluete je 0.77, što označava dobro grupirane podatke.



Slika 2.4 Vrijednosti siluete za skup podataka

Na kraju za svaki broj skupina k u proizvoljnom rasponu primjenjujemo odabranu metodu grupiranja, te za svaku točku u različitim grupiranjima u k skupina izračunavamo prosječni broj siluete. Za optimalni broj skupina k odabiremo k s najvećim prosječnim koeficijentom siluete. Na sljedećem primjeru (Slika 2.5) vidimo vizualnu reprezentaciju prosječnog koeficijenta siluete za različite brojeve skupina k na tri različita skupa podataka, kao i popratne grafove grupiranih podataka.



Slika 2.5 Metoda siluete primijenjena na više skupova podataka

Budući da su obje navedene metode empirijske, njihovi rezultati ovise o podatcima nad kojima se provodi grupiranje, te ih se često koristi u kombinaciji s drugim postupcima za bolje rezultate. U prethodnim primjerima koristili smo prirodno grupiranje podatke tako da su rezultati bili jasni, što nije slučaj kod velikih skupova podataka.

Kod spektralnog grupiranja koristimo se drukčijim tehnikama za određivanje optimalnog broja skupina, koje ćemo prikazati u sljedećem poglavlju gdje se i govori o spektralnom grupiranju.

3. Uvod u spektralno grupiranje

Spektralno grupiranje temelji se na svojstvenim vektorima prema kojima određujemo grupe podataka sa sličnim svojstvima.

Glavni temelj ove metode je Laplaceova matrica koja nam opisuje strukturu grafa u obliku matrice. Ona predstavlja težinski, neusmjereni graf početnih podataka u kojem su prikazani odnosi između čvorova.

Za početak pretpostavimo da naš skup podataka sadrži n dvodimenzionalnih točaka $\boldsymbol{x_1, \dots, x_n}$. Spektralno grupiranje temelji se na analizi međusobnih odnosa točka pohranjenih u matrici sličnosti čije težine dobivamo prema udaljenosti ili razini njihove sličnosti. Ovisno o početnim podatcima, matrica $\boldsymbol{A}$ može predstavljati i matricu susjedstva gdje su vrijednosti matrice 1 ili 0 ovisno o tome jesu li točke povezane ili ne. U kontekstu spektralnog grupiranja, kreira se graf gdje su čvorovi točke, a bridovi su vrijednosti kojima je određena sličnost povezanih točaka, te je prema tome kreirana matrica sličnosti $\boldsymbol{A}$, to jest težinska matrica susjedstva.

Iz matrice sličnosti $\boldsymbol{A}$ izvodi se dijagonalna matrica tako da su vrijednosti na dijagonali $\boldsymbol{D_{ii}}$ sume redaka iz matrice sličnosti:

$$\boldsymbol{D_{ii}} = \sum_j \boldsymbol{A_{ij}}. \tag{15}$$

Ove vrijednosti predstavljaju ukupnu povezanost svake točke s ostalim točkama u skupu podataka. Navedene matrice koristimo za izračunavanje Laplaceove matice, ključne za spektralno grupiranje:

$$\boldsymbol{L = D - A} \tag{16}$$

Za rješavanje problema spektralnog grupiranja bit će nam potrebne svojstvene vrijednosti Laplaceove matrice i njeni pripadni svojstveni vektori.

Svojstveni vektor $\boldsymbol{v}$ matrice $\boldsymbol{L}$ za koji vrijedi $\boldsymbol{v \neq 0}$ definiramo uz pomoć skalara λ kao:

$$\boldsymbol{Lv} = \lambda \boldsymbol{v}.$$

Taj skalar λ naziva se svojstvenom vrijednosti matrice $\boldsymbol{L}$ koja odgovara vektoru $\boldsymbol{v}$ [8].

Kako bismo dobili svojstvene vrijednosti Laplaceove matrice, potrebno je riješiti karakterističnu jednadžbu:

$$\det(\boldsymbol{L} - \lambda \boldsymbol{I}) = 0$$

gdje je $\boldsymbol{L}$ Laplaceova matrica, λ svojstvena vrijednost, a $\boldsymbol{I}$ je jedinična matrica. Kao rješenje ove jednadžbe dobivamo polinom čiji su korijeni svojstvene vrijednosti. Nakon dobivenih svojstvenih vrijednosti, svojstvene vektore dobijemo rješavajući sustav:

$$\boldsymbol{Lv} = \lambda \boldsymbol{v}$$

gdje nam je $\boldsymbol{v}$ svojstveni vektor sa pripadnom λ svojstvenom vrijednosti.

Činjenicu da simetrične matrice imaju samo **realne svojstvene vrijednosti** dokazat ćemo koristeći skalarni umnožak kompleksnih vektora $\boldsymbol{v}$ i $\boldsymbol{w}$:

$$(\boldsymbol{v}, \boldsymbol{w}) = \boldsymbol{v} \cdot \boldsymbol{w} = \sum_i \overline{\boldsymbol{v}}_i \boldsymbol{w}_i$$

Kako bismo dokazali da su svojstvene vrijednosti realne, potrebno je dokazati da je u izrazu $\boldsymbol{Lv} = \lambda \boldsymbol{v}$, $\boldsymbol{L}$ kvadratna matrica za koju vrijedi $\boldsymbol{L} = \boldsymbol{L}^*, \boldsymbol{v} \neq 0$.

Ako prikažemo $\boldsymbol{Lv} = \lambda \boldsymbol{v}$ preko skalarnog umnoška dobivamo:

$$(\boldsymbol{Lv}, \boldsymbol{v}) = (\lambda \boldsymbol{v}, \boldsymbol{v}) = \lambda(\boldsymbol{v}, \boldsymbol{v})$$

Budući da za matricu $\boldsymbol{L}$ vrijedi $\boldsymbol{L} = \boldsymbol{L}^*$, to jest $(\boldsymbol{Lv}, \boldsymbol{v}) = (\boldsymbol{v}, \boldsymbol{Lv})$, $Lv = \lambda v$ možemo prikazati kao:

$$(\boldsymbol{v}, \boldsymbol{Lv}) = (\boldsymbol{v}, \lambda \boldsymbol{v}) = \bar{\lambda}(\boldsymbol{v}, \boldsymbol{v})$$

Usporedivši dva navedena izraza dobivamo:

$$\lambda(\boldsymbol{v}, \boldsymbol{v}) = \bar{\lambda}(\boldsymbol{v}, \boldsymbol{v})$$

te uz uvjet $(\boldsymbol{v}, \boldsymbol{v}) \neq 0$, svodimo na izraz kojim dokazujemo da su sve svojstvene vrijednosti matrice $\boldsymbol{L}$ realne vrijednosti, to jest λ je jednako svojem kompleksno konjugiranom obliku [9]:

$$\lambda = \bar{\lambda}.$$

Da su svojstveni vektori $\boldsymbol{v}$ simetrične matrice $\boldsymbol{L}$ **ortogonalni** dokazujemo spektralnim teoremom koji kaže da svaka realna simetrična matrica ima ortonormiranu bazu svojstvenih vektora i može se dijagonalizirati ortogonalnom matricom $\boldsymbol{S}$ takom da vrijedi:

$$\boldsymbol{S}^T\boldsymbol{L}\boldsymbol{S} = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{bmatrix}.$$

Navedeni teorem dokazujemo tako da svojstveni vektor $\boldsymbol{v}_1$ koji pripada svojstvenoj vrijednosti λ_1 normiramo i proširujemo do ortonormirane baze te dobivamo ortogonalnu matricu $\boldsymbol{S}_1$:

$$\boldsymbol{S}_1 = [\boldsymbol{v}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w}_n].$$

Pomoću te baze, transformiramo matricu $\boldsymbol{L}$ tako da poprimi oblik:

$$\boldsymbol{L}_1 = {\boldsymbol{S}_1}^T\boldsymbol{L}\boldsymbol{S}_1 = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & b_{22} & \dots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & b_{n2} & \dots & b_{nn} \end{bmatrix}$$

to jest:

$$\boldsymbol{L}_1 = {\boldsymbol{S}_1}^T\boldsymbol{L}\boldsymbol{S}_1 = \begin{bmatrix} \lambda_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{B} \end{bmatrix}.$$

Budući da je matrica $\boldsymbol{L}_1$ realna i simetrična i da je za $n = 2$ dijagonalna, krećemo od pretpostavke da je za sve matrice reda $n - 1$ također dijagonalna. Tada vrijedi da je i matrica $\boldsymbol{B}$ simetrična i dijagonalna, što znači da postoji ortogonalna matrica $\boldsymbol{S}_2$ koja je dijagonalizira:

$${\boldsymbol{S}_2}^T\boldsymbol{B}\boldsymbol{S}_2 = \begin{bmatrix} \lambda_2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_n \end{bmatrix}$$

Ako dodatno definiramo matricu:

$$\begin{bmatrix} 1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{S}_2 \end{bmatrix}$$

za koju znamo da je ortogonalna zbog podmatrice $\boldsymbol{S}_2$ koja je ortogonalna matrica, to jest uz ortogonalni vektor **1** imamo ortonormiranu bazu, možemo definirati konačnu matricu $\boldsymbol{S}$ kao:

$$\boldsymbol{S} = \boldsymbol{S}_1 \begin{bmatrix} 1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{S}_2 \end{bmatrix}$$

za koju znamo da je ortogonalna jer je rezultat umnoška dviju ortogonalnih matrica.

Dakle, da bismo dokazali polaznu tvrdnju raspisujemo:

$$\boldsymbol{S}^T\boldsymbol{L}\boldsymbol{S} = \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{S}_2{}^T \end{bmatrix}\begin{bmatrix} \lambda_1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{B} \end{bmatrix}\begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{S}_2 \end{bmatrix} = \begin{bmatrix} \lambda_1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{S}_2{}^T\boldsymbol{B}\boldsymbol{S}_2 \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{bmatrix}$$

što nas dovodi do zaključka da za svaki red $n \in \mathbb{N}$ postoji ortogonalna matrica $\boldsymbol{S}$ takva da je $\boldsymbol{S}^T\boldsymbol{L}\boldsymbol{S}$ dijagonalna matrica [11].

3.1. Laplaceova matrica

Laplaceovu matricu možemo promatrati u dva oblika, u nenormaliziranom i normaliziranom obliku.

Nenormaliziranu Laplaceovu matricu predstavili smo kao razliku matrice sličnosti i dijagonalne matrice prema izrazu (16).

Karakteristike koje zadovoljava nenormalizirana Laplaceova matrica [12] su sljedeće:

1. $\boldsymbol{L}$ je simetrična i pozitivno semidefinitna.
2. Najmanja svojstvena vrijednost λ_1 jednaka je 0, a pripadni svojstveni vektor je vektor **1.**
3. $\boldsymbol{L}$ sadrži *n* nenegativnih, realnih svojstvenih vrijednosti $0 = \lambda_1 \leq \cdots \leq \lambda_n$.
4. $\boldsymbol{L}$ ima ortonormiranu bazu svojstvenih vektora (dokazano u prethodnom poglavlju).

Budući da nenormaliziranu Laplaceovu matricu definiramo kao razliku dviju simetričnih matrica dobivenih iz neusmjerenog grafa, i ona sama posjeduje to svojstvo **simetričnosti**.

S obzirom na činjenicu da je nenormalizirana Laplaceova matrica realna i simetrična, sve njene svojstvene vrijednosti su realne, a svojstveni vektori su međusobno ortogonalni [10].

Matrica $\boldsymbol{L}$ je **pozitivno semidefinitna** ako su i sve njene svojstvene vrijednosti λ nenegativne, dakle:

$$\boldsymbol{v}^T\boldsymbol{L}\boldsymbol{v} \geq 0,$$

što se može i sažetije napisati kao:

$$\boldsymbol{L} \succcurlyeq 0. \quad [4]$$

Ovo svojstvo može se dokazati dekompozicijom simetrične matrice $\boldsymbol{L}$ uz $\boldsymbol{S}^T = \boldsymbol{S}^{-1}$ i dijagonalnu matricu svojstvenih vrijednosti $\boldsymbol{\Lambda}$:

$$\boldsymbol{L} = \boldsymbol{S\Lambda S^{-1}},$$

koristeći ne-nul vektor $\boldsymbol{v} \in \mathbb{R}^n$ *i* $\boldsymbol{y} = \boldsymbol{S^T v}$, kvadratnu formu prikazujemo kao sumu svojstvenih vrijednosti:

$$\boldsymbol{v^T L v} = \boldsymbol{v^T S \Lambda S^T v} = \boldsymbol{y^T \Lambda y} = \sum_{i=1}^{n} \lambda_i y_i^2 \geq 0\,,$$

Budući da matrica $\boldsymbol{\Lambda}$ sadrži svojstvene vrijednosti matrice $\boldsymbol{L}$, vrijednosti na dijagonali su pozitivne, to jest vrijedi da je $\boldsymbol{v^T L v} \geq 0$ za sve $\boldsymbol{v} \neq 0$ [3].

S obzirom na to da je Laplaceova matrica $\boldsymbol{L}$ definirana nad potpuno povezanim grafom, to jest n-regularnim grafom, možemo pokazati da je svojstvenom vektoru **1** pripadna svojstvena vrijednost 0. Definirajmo vektor **1** kao:

$$\mathbf{1} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \in \mathbb{R}^n.$$

Kod n-regularnog grafa zbroj elemenata u redcima matrice susjedstva **A** daje nam stupanj vrha:

$$\boldsymbol{A} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = \begin{bmatrix} d \\ d \\ \vdots \\ d \end{bmatrix},$$

Budući da je Laplaceova nenormalizirana matrica definirana kao $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$, slijedi:

$$(\boldsymbol{D} - \boldsymbol{A}) \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = \begin{bmatrix} d - d \\ d - d \\ \vdots \\ d - d \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = 0 \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix},$$

što možemo pojednostaviti kao:

$$\boldsymbol{L} \cdot \mathbf{1} = \mathbf{0} = 0 \cdot \mathbf{1}$$

i zaključiti da kod Laplaceove matrice svojstvena vrijednost 0 odgovara pripadnom svojstvenom vektorom **1** [11].

Koristeći matricu sličnosti, iz Laplaceove matrice možemo analizirati jačinu veza, to jest povezanosti između točaka te lako primijetiti gdje su najslabije veze u grafu što nam omogućuje da efikasno grupiraju podatci.

Kod nenormalizirane Laplaceove matrice može doći do problema da čvorovi s većim stupnjem povezanosti ili većim težinama dominiraju, što dovodi do neravnoteže u dobivenim

grupama. Kako bismo osigurali veću stabilnost pri grupiranju, potrebno je normalizirati Laplaceovu matricu.

Postoje dva načina na koji se Laplaceova matrica može normalizirati. Ukoliko želimo dobiti simetričnu normaliziranu matricu $\boldsymbol{L}_{sym}$, koristimo sljedeći izraz:

$$\boldsymbol{L}_{sym} := \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{L}\boldsymbol{D}^{-\frac{1}{2}} = \boldsymbol{I} - \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}} \tag{17}$$

u kojem je $\boldsymbol{L}$ nenormalizirana Laplaceova matrica, a s izrazom $\boldsymbol{D}^{-\frac{1}{2}}$ dobivamo inverznu matricu od $\boldsymbol{D}^{\frac{1}{2}}$, to jest dijagonalnu matricu gdje su elementi na dijagonali recipročne vrijednosti kvadratnog korijena elemenata na dijagonali matrice $\boldsymbol{D}$ [12].

Ukoliko se odlučimo za drugi način dobivanja normalizirane matrice, koristi se izraz kojim dobivamo matricu koju nazivamo matricom nasumične šetnje:

$$\boldsymbol{L}_{rw} := \boldsymbol{D}^{-1}\boldsymbol{L} = \boldsymbol{I} - \boldsymbol{D}^{-1}\boldsymbol{A} \tag{18}$$

Izraz $\boldsymbol{D}^{-1}\boldsymbol{A}$ označava tranzicijsku matricu za nasumičnu šetnju po grafu, to jest vjerojatnosnu matricu koja predstavlja nasumična kretanja „šetača“ po grafu [12]. Ona sadrži vjerojatnosti kretanja šetača iz jednog čvora u drugi. Ta ideja je korisna jer se vodimo mišlju da će šetač ostati duže u dobro povezanim skupinama, to jest promatramo gdje se proces koji predstavlja šetač zadržava u grafu.

Prethodno navedene karakteristike nenormalizirane Laplaceove matrice vrijede i za normalizirani oblik, uz iznimku da kod normalizirane simetrične Laplaceove matrice $\boldsymbol{L}_{sym}$ za najmanju svojstvenu vrijednost λ_1 pripadni svojstveni vektor je oblika $\boldsymbol{v}_1 = \boldsymbol{D}^{1/2}\mathbf{1}$ [12].

3.2. Numerička metoda za nalaženje svojstvenih vrijednosti

Numeričke metode za nalaženje svojstvenih vrijednosti služe nam kako bismo aproksimirali svojstvene vrijednosti i pripadne svojstvene vektore matrica. Budući da nalaženje svojstvenih vrijednosti Laplaceove matrice često mogu biti računalno zahtjevno, posežemo za navedenim metodama kako bismo dobili brže i bolje rezultate. Postoje razne iterativne metode poput metode potencija [13] metode inverzne iteracije ili QR metode za pronalaženje svojstvenih vrijednosti, za više detalja pogledati u literaturi [14].

U ovom poglavlju bavit ćemo se **Lanczsovom metodom** za izračun svojstvenih vrijednosti i pripadnih svojstvenih vektora kako bismo riješili problem velikih i računalno zahtjevnih Laplaceovih matrica.

Lanczosovom metodom dobivamo aproksimacije svojstvenih vrijednosti i vektora Laplaceove matrice tako da Laplaceovu matricu svodimo na trodijagonalnu matricu. U sklopu metode koristi se metoda Krylovljevog potprostora kojom projiciramo rješenje problema u određeni potprostor, točnije Krylovljev potprostor. Množenjem zadanog vektora $\boldsymbol{v}$ s matricom $\boldsymbol{L}$ dobivamo Krylovljev niz, skup vektora $\boldsymbol{v}, \boldsymbol{Lv}, \boldsymbol{L^2v}, \boldsymbol{L^3v}$, a Krylovljev potprostor reda *k* je razapet između prvih *k* vektora [15].

Ovisno o odabiru početnog normaliziranog vektora $\boldsymbol{r}$ konstruiramo Krylovljev potprostor koristeći simetričnu, realnu matricu $\boldsymbol{L}$:

$$\kappa(\boldsymbol{L}, \boldsymbol{r}, k) = span\{\boldsymbol{r}, \boldsymbol{Lr}, \boldsymbol{L^2r}, \dots, \boldsymbol{L^{k-1}r}\}$$

Kako bismo osigurali ortonormiranost baze potprostora $\kappa(\boldsymbol{L}, \boldsymbol{v}, k)$, provodimo Gram-Schmidtovu ortonorminalizaciju nad vektorima Krylovljevog potprostora te dobivamo matricu $\boldsymbol{Q}_k$ čiji su stupci ortonormirani Lanczosovi vektori:

$$\boldsymbol{Q}_k = [\boldsymbol{q}_1, \boldsymbol{q}_2, \dots, \boldsymbol{q}_k] \quad [16]$$

Algoritam koji prati metoda uzet je iz literature [17]:

$$\boldsymbol{q}_1 = \frac{\boldsymbol{b}}{||\boldsymbol{b}||_2}, \boldsymbol{\beta}_0 = 0, \boldsymbol{q}_0 = 0$$

$for\ j = 1\ to\ k$

$\boldsymbol{z} = \boldsymbol{L}\boldsymbol{q}_j$

$\alpha_j = \boldsymbol{q}_j^T z$

$\boldsymbol{z} = \boldsymbol{z} - \alpha_j \boldsymbol{q}_j - \boldsymbol{\beta}_{j-1}\boldsymbol{q}_{j-1}$

$\boldsymbol{\beta}_j = ||\boldsymbol{z}||_2$

$if\ \boldsymbol{\beta}_j = 0, stop$

$\boldsymbol{q}_{j+1} = \boldsymbol{z}/\boldsymbol{\beta}_j$

uz dodatnu varijablu nasumično odabranog početnog vektora $\boldsymbol{b}$ različitog od 0.

Kao rezultat k iteracija dobivamo matricu $\boldsymbol{L}$ projiciranu u Krylovljev potprostor, to jest simetričnu trodijagonalnu matricu **T** čije su svojstveni vektori i svojstvene vrijednosti aproksimacije matrice **L**:

$$\boldsymbol{T}_k = \boldsymbol{Q}_k^T \boldsymbol{L} \boldsymbol{Q}_k$$

$$\boldsymbol{T}_k = \begin{bmatrix} \alpha_1 & \beta_1 & \dots & 0 \\ \beta_1 & \alpha_2 & \beta_2 & \vdots \\ \vdots & \beta_2 & \ddots & \beta_k \\ 0 & \dots & \beta_k & \alpha_k \end{bmatrix}. \qquad [18]$$

U postupku spektralnog grupiranja, trodijagonalnu matricu dobivenu Lanczosovom metodom koristimo u izračunima svojstvenih vrijednosti i pripadnih svojstvenih vektora kao aproksimaciju početne Laplaceove matrice $\boldsymbol{L}$.

Ova transformacija omogućuje nam da iz trodijagonalne matrice $\boldsymbol{T}$ lakše izračunamo svojstvene vrijednosti λ i dobijemo njene odgovarajuće svojstvene vektore $\boldsymbol{v}_T$. Dobivene vrijednosti predstavljaju aproksimacije svojstvenih vrijednosti i svojstvenih vektora Laplaceove matrice $\boldsymbol{L}$.

Ako je $\boldsymbol{v}_T$ svojstveni vektor trodijagonalne matrice $\boldsymbol{T}$, onda pripadni svojstveni vektor matrice $\boldsymbol{L}$ dobivamo koristeći izraz:

$$\boldsymbol{v} = \boldsymbol{Q}_k \boldsymbol{v}_T$$

gdje je $\boldsymbol{Q}_k$ matrica čiji stupci čine ortonormiranu bazu k-dimenzionalnog potprostora dobivenog Lanczosovom metodom [19].

4. Metode spektralnog grupiranja

U kontekstu spektralne metode grupiranja, svojstveni vektori predstavljaju optimalne smjerove za podjelu podataka, omogućujući mapiranje višedimenzionalnih podataka u niždimenzionalni prostor, što olakšava proces grupiranja.

U nastavku su prikazane spektralne metode koje koriste svojstvene vektore kako bi grupirale podatke u dvije skupine, koristeći samo svojstveni vektor koji pripada drugoj najmanjoj svojstvenoj vrijednosti i koristeći više svojstvenih vektora.

4.1. Prikaz metoda biparticije grafa

Kao što smo već spomenuli kod karakteristika Laplaceove matrice, najmanja svojstvena vrijednost λ_1 dobivena iz Laplaceove matrice je 0, a svojstveni vektor koji pripada toj vrijednosti je konstantni vektor **1**. Važno je istaknuti da svojstvena vrijednost λ_1 može biti vrijednost više višestrukosti što ukazuje na nepovezanost grafa. Višestrukost svojstvene vrijednosti λ_1 jednaka je broju povezanih komponenti grafa. Ako je graf povezan, tada je ta višestrukost jednaka 1 [11].

Druga najmanja svojstvena vrijednost λ_2 još se naziva i Fiedlerovom vrijednosti te ju možemo definirati preko minimiziranja Rayleighovog kvocijenta koji nam govori koliko se vrijednosti vektora $\boldsymbol{v}$ razlikuju između međusobno povezanih čvorova grafa:

$$\lambda_2 = \min_{v \perp \mathbf{1}} \frac{\boldsymbol{v}^T L \boldsymbol{v}}{\boldsymbol{v}^T \boldsymbol{v}},$$

gdje promjenu vektora između bridova grafa vidimo u dijelu izraza:

$$\boldsymbol{v}^T \boldsymbol{L} \boldsymbol{v} = \sum_{(i,j)\in E} (\boldsymbol{v}_i - \boldsymbol{v}_j)^2$$

uz uvjet $\boldsymbol{v} \perp \mathbf{1}$ da je traženi vektor ortogonalan na svojstveni vektor za $\lambda_1 = 0$. Za izvod detaljnije u literaturi [2].

λ_2 je minimalna svojstvena vrijednost čiji pripadni vektor prikazuje minimalne razlike između povezanih čvorova, a ortogonalan je na konstantni vektor uz pretpostavku da je višestrukost svojstvene vrijednosti jednaka 1. U slučaju da je višestrukost λ_1 veća od 1, uvjet ortogonalnosti proširuje se na cijeli potprostor svojstvenih vektora za koje vrijedni $\lambda = 0$.

Ukoliko smo Laplaceovu matricu gradili iz potpuno povezanog grafa, druga svojstvena vrijednost λ_2 je veća od 0 te nam označuje jačinu povezanosti grafa. To nam osigurava teorem koji govori da je višestrukost svojstvene vrijednosti jednaka broju povezanih komponenti grafa [11]. Vrijednost λ_2 je proporcionalna jačini povezanosti grafa, manje vrijednosti označuju manje povezane grafove iz kojih lako možemo grupirati podatke, a veće vrijednosti jače povezanije grafove.

Svojstveni vektor koji odgovara drugoj najmanjoj svojstvenoj vrijednosti Laplaceove matrice λ_2 naziva se Fiedlerovim vektorom. U slučaju da svojstvena vrijednost λ_2 ima veću višestrukost, za Fiedlerov vektor koristi se linearna kombinacija vektora iz pripadnog vektorskog potprostora za svojstvenu vrijednost λ_2 [12].

Budući da je jedan od svojstvenih vektora nenormalizirane Laplaceove matrice vektor jedinica **1**, a svojstveni vektori simetrične matrice su ortogonalni, Fiedlerov vektor $\boldsymbol{v}$ zadovoljava [2]:

$$\sum_{i=1}^{n} \boldsymbol{v}_i = 0. \tag{19}$$

4.1.1. Metoda grupiranja koristeći samo Fiedlerov vektor

Može se pokazati da uz pomoć Fiedlerovog vektora možemo razdijeli čvorove grafa u dvije skupine. Koristimo ga u određivanju skupina jer minimizira udaljenost između povezanih čvorova u početnom skupu podataka te nam daje najbolju podjelu grafa na dva dijela [2].

U slučaju kada je graf moguće podijeliti na dva dijela uklanjanjem bridova, suma težina uklonjenih bridova označava mjeru razdijeljenosti između dva dijela grafa i naziva se rezom:

$$cut(A, B) = \sum_{u \epsilon A, v \epsilon B} w(u, v) \tag{20}$$

gdje su A i B oznake disjunktnih skupova čvorova grafa, u i v pojedinačni čvorovi unutar tih skupova, a $w(u, v)$ označava težinu brida koji povezuje čvorove u i v.

Idealno za grupiranje $P = (A, B)$ htjeli bismo minimizirati vrijednost reza uzimajući $|A| = |B|$, odnosno dva jednako brojna skupa. Taj problem možemo formulirati preko vektora $\boldsymbol{v} \epsilon \{-1,1\}^n$ tako da je $cut(A, B)$ minimalan uz uvjet (19). Budući da je ovaj problem NP-

težak problem možemo ga relaksirati tako da uvedemo realne komponente vektoru $\boldsymbol{v}$ uz uvjete:

$$\sum_{i=1}^{n} \boldsymbol{v}_i = 0 \quad \text{i} \quad \sum_{i=1}^{n} \boldsymbol{v}_i{}^2 = n.$$

Može se pokazati da u tom slučaju vrijedi:

$$cut(A,B) = \frac{1}{4} \sum_{i,j \in E} (\boldsymbol{v}_i - \boldsymbol{v}_j)^2$$

te da upravo Fiedlerov vektor pridružene Laplaceove matrice minimizira ovaj izraz, detaljnije u literaturi [20].

Postoje razni načini podjele Fiedlerovog vektora kako bismo mogli grupirati podatke.

U osnovnom algoritmu na temelju pozitivnosti u Fiedlerovom vektoru $\boldsymbol{v}$ određujemo pripadnost podatka i iz skupa $N = \{1,2, \ldots., n\}$ skupini A i B:

$$A = \{\, i \in N : \boldsymbol{v}_i < 0\}$$

$$B = \{\, i \in N : \boldsymbol{v}_i > 0\}$$

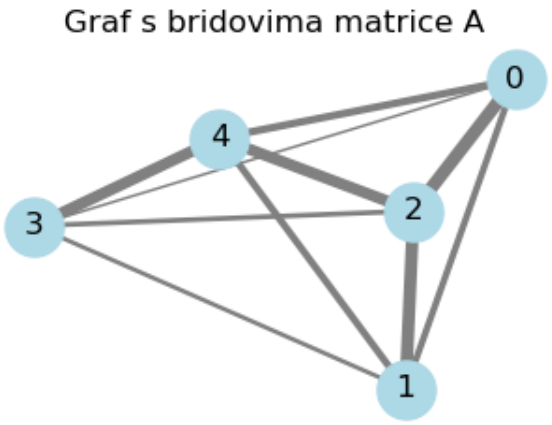
te time dobivamo prvu podjelu podataka u skupine A i B u obliku vektora $\boldsymbol{y}$ za svaki podatak i [21]:

$$\boldsymbol{y_i} = \begin{cases} 1\,, & ako\ vrh\ i\ pripada\ skupini\ A \\ -1\,, & ako\ vrh\ i\ pripada\ skupini\ B \end{cases}.$$

Ako je vrh $\boldsymbol{v}_i = 0$, možemo ga pridružiti bilo kojoj od skupina bez utjecaja na rezultat grupiranja.

Granica podjele čvorova ne mora nužno biti vrijednost 0, može se definirati kao medijan vrijednosti podataka, prag koji minimizira vrijednost reza (detaljnije u sljedećem poglavlju) ili neki drugi kriterij prilagođen konkretnom problemu [22].

U sljedećem primjeru prikazan je postupak dobivanja Fiedlerovog vektora na skupu od 5 nasumično odabranih podataka. Iz podataka kreira se matrica sličnosti preko Gaussove jezgrene funkcije na temelju koje dobivamo potpuno povezani graf. Na primjeru grafa prikazanog na slici (Slika 4.1) bridovi kojima je težina 0 nisu prikazani, a ostale težine prikazane su vizualno debljinom brida. Laplaceova matrica pripadnog grafa sadrži pozitivne vrijednosti na dijagonali, i negativne vrijednosti ili nule u ostatku matrice. Konačni rezultat grupiranja vidimo na slici (Slika 4.2).



Slika 4.1 Početni graf

$$\begin{bmatrix} 1.58 & -0.34 & -0.76 & -0.08 & -0.4 \\ -0.34 & 1.64 & -0.73 & -0.19 & -0.38 \\ -0.76 & -0.73 & 2.37 & -0.24 & -0.65 \\ -0.08 & -0.19 & -0.24 & 1.17 & -0.66 \\ -0.4 & -0.38 & -0.65 & -0.66 & 2.09 \end{bmatrix}$$

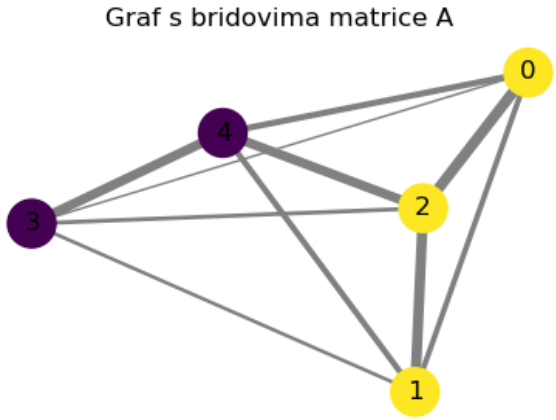
(Laplaceova matrica)

$$[\ 0 \quad 1.24 \quad 1.96 \quad 2.54 \quad 3.11\ ]$$

(svojstvene vrijednosti)

$$[\ 0.44 \quad 0.28 \quad 0.23 \quad -0.81 \quad -0.13\ ]$$

(Fiedlerov vektor)



Slika 4.2 Prikaz grupiranja podataka Fiedlerovim vektorom

4.1.2. Metoda normaliziranog reza

Metoda normaliziranog reza pisana je prema radu [22] , gdje su uz pristup s normaliziranim rezom predložili i korištenje iterativne Lanczosove metode za računanje svojstvenih vrijednosti.

Metoda normaliziranog reza razlikuje se od prethodno opisanog spektralnog algoritma u posljednjem koraku kada se koristi Fiedlerov vektor za podjelu skupa podataka. Glavni nedostatak metode za biparticiju prikazane u prethodnom poglavlju jest to što se ne uzima u obzir veličina novonastalih skupina, što dovodi do loše grupiranih podataka.

Budući da prethodno opisani pristup traženja reza ima svojih mana poput neprirodnog izdvajanja malenog skupa kao particije, učinkovitijim se pokazao pristup kada vrijednost reza gledamo u odnosu na ukupnu vrijednost svih težina bridova u grafu. Ovaj pristup naziva se normalizirani rez i definiramo ga kao:

$$Ncut(A,B) = \frac{cut(A,B)}{assoc(A,V)} + \frac{cut(B,A)}{assoc(B,V)} \tag{21}$$

u kojem $assoc(A,V) = \sum_{u\epsilon A, t\epsilon V} w(u,t)$ označava sumu svih težina bridova koji povezuju čvorove iz grupe vrhova A prema svim ostalim čvorovima u grafu. Isto vrijedi i za $assoc(B,V)$ gdje se sumiraju težine bridova koji povezuju čvorove iz skupa B sa svim čvorovima [22].

Budući da problem minimiziranja vrijednosti normaliziranog reza spada u NP-teške probleme, pokazat ćemo da se aproksimacija diskretnog optimizacijskog problema može pronaći ukoliko početni problem pretvorimo u optimizacijski problem u području s realnim vrijednostima [22].

Koristeći prethodno definirane matrice sličnosti $\boldsymbol{A}$ i dijagonalne matrice $\boldsymbol{D}$ prikazat ćemo izraz za normalizirani rez $Ncut(A,B)$ matričnim računom. Cilj izvoda je podijeliti graf V u dvije grupe A i B. Definirat ćemo i vektor $\boldsymbol{x}$:

$$x_i = \begin{cases} 1 & ako\ je\ čvor\ i\ \epsilon\ A \\ -1 & ako\ je\ čvor\ i\ \epsilon\ B \end{cases},$$

koji će pohranjivati informacije o tome koji čvor pripada kojoj grupi.

Neka je $\boldsymbol{d}_i = \sum_j w(i,j)$ ukupna povezanost čvora i prema ostalim čvorovima, a $\boldsymbol{A}(i,j) = a_{ij}$. Tada izraz (21) možemo zapisati kao:

$$Ncut(A,B) = \frac{cut(A,B)}{assoc(A,V)} + \frac{cut(B,A)}{assoc(B,V)}$$

$$= \frac{\sum_{(x_i>0, x_j<0)} -a_{ij}\boldsymbol{x}_i\boldsymbol{x}_j}{\sum_{(x_i>0)} d_i} + \frac{\sum_{(x_i<0, x_j>0)} -a_{ij}\boldsymbol{x}_i\boldsymbol{x}_j}{\sum_{(x_i<0)} \boldsymbol{d}_i}.$$

Dodatno, definiramo i novu varijablu k koja nam predstavlja koliki je udio svih vrijednosti stupnjeva čvorova skupa A u ukupnome grafu. Njezina uloga je normalizacija vektora čime omogućujemo bolju procjenu povezanosti [22]:

$$k = \frac{assoc(A,V)}{assoc(V,V)} = \frac{\sum_{x_i>0} \boldsymbol{d}_i}{\sum_i \boldsymbol{d}_i}.$$

Slijedi izraz na normalizirani rez izražen koristeći prethodno definirane vektore i matrice:

$$4Ncut(A,B) = \frac{(\mathbf{1}+\boldsymbol{x})^T(\boldsymbol{D}-\boldsymbol{A})(\mathbf{1}+\boldsymbol{x})}{k\mathbf{1}^T\boldsymbol{D}\mathbf{1}} + \frac{(\mathbf{1}-\boldsymbol{x})^T(\boldsymbol{D}-\boldsymbol{A})(\mathbf{1}-\boldsymbol{x})}{(1-k)\mathbf{1}^T\boldsymbol{D}\mathbf{1}}.$$

Ovaj izraz prikazuje nam normalizirani rez preko Laplaceove matrice. Usporedivši ga s početnim izrazom (21) primjećujemo:

$$cut(A,B) = (\mathbf{1}+\boldsymbol{x})^T(\boldsymbol{D}-\boldsymbol{A})(\mathbf{1}+\boldsymbol{x});$$

gdje vektorom $(\mathbf{1}+\boldsymbol{x})$ određujemo pripadnost grupi A, dok se izraz $(\boldsymbol{D}-\boldsymbol{A})$ odnosi na ne-normaliziranu Laplaceovu matricu. Rezultat ove matrične operacije je skalar koji predstavlja vrijednost reza između grupa. Slično, za $cut(B,A)$ koristimo izraz $(\mathbf{1}-\boldsymbol{x})$ za određivanje pripadnosti grupi B.

Zajednički faktor u nazivniku izraza jest $\mathbf{1}^T\boldsymbol{D}\mathbf{1}$ koji daje ukupni stupanj grafa. Uz taj faktor pojavljuje se i prethodno opisani skalar k, to jest $(1-k)$. Budući da je graf podijeljen na dva disjunktna skupa, vrijednost za grupu B jednostavno se dobiva kao $(1-k)$.

Provedbom odgovarajućih supstitucija i algebarskih transformacija dolazimo do konačnog izraza koji predstavlja Rayleighov kvocijent za matricu $\boldsymbol{L}$:

$$min_{\boldsymbol{x}} Ncut(\boldsymbol{x}) = \min_{y} \frac{\boldsymbol{y}^T(\boldsymbol{D}-\boldsymbol{A})\boldsymbol{y}}{\boldsymbol{y}^T\boldsymbol{D}\boldsymbol{y}} \quad (22)$$

gdje je $\boldsymbol{y}$ definiran kao $\boldsymbol{y} = (\mathbf{1}+\boldsymbol{x}) - b(\mathbf{1}-\boldsymbol{x})$, s pripadnim ograničenjima: $y_i \epsilon \mathbb{R}$; $\boldsymbol{y}^T\boldsymbol{D}\mathbf{1} = 0$, gdje je $y_i \in \{-1, b\}$, a b je definiran kao $b = \frac{k}{1-k}$. Tako definiran b osigurava da je vektor $\boldsymbol{y}$ ortogonalan na $\boldsymbol{D}\mathbf{1}$, točnije da je uvjet $\boldsymbol{y}^T\boldsymbol{D}\mathbf{1} = 0$ zadovoljen.

Glavna karakteristika Rayleighovog kvocijenta jest ta da ako je $\boldsymbol{y}$ svojstveni vektor simetrične matrice $\boldsymbol{A}$, onda $Ncut(x)$ izračunava pripadnu svojstvenu vrijednost.

Ukoliko gornji problem (22) relaksiramo tako da uzmemo da su komponente vektora $\boldsymbol{y}$ realne vrijednosti, može se pokazati da rješenje problema dobivamo kao rješenje generaliziranog problema svojstvenih vrijednosti [23]:

$$(\boldsymbol{D}-\boldsymbol{A})\boldsymbol{y}=\lambda \boldsymbol{D}\boldsymbol{y}.$$

Ovaj problem definiramo kao generalizirani problem zato što nije u klasičnoj formi svojstvenog problema: $\boldsymbol{A}y=\lambda y$, a dodatno imamo i $\boldsymbol{D}$ matricu na desnoj strani jednadžbe. Kako bismo olakšali rješavanje ove jednadžbe pretvorit ćemo ju u standardni svojstveni problem koristeći $\boldsymbol{z}=\boldsymbol{D}^{1/2}\boldsymbol{y}$ te tako dobivamo simetričnu matricu koja nam osigurava realne svojstvene vrijednosti:

$$\boldsymbol{D}^{-\frac{1}{2}}(\boldsymbol{D}-\boldsymbol{A})\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{z}=\lambda \boldsymbol{z}.$$

Najmanji dobiveni vektor u tom slučaju je $\boldsymbol{z}_0=\boldsymbol{D}^{1/2}\boldsymbol{1}$, sa pripadnom svojstvenom vrijednosti 0. Drugi najmanji svojstveni vektor z_1, ortogonalan na z_0 , dobiva se rješavanjem:

$$\boldsymbol{z}_1=arg.min_{\boldsymbol{z}^T\boldsymbol{z}_0=0}\frac{\boldsymbol{z}^T\boldsymbol{D}^{-\frac{1}{2}}(\boldsymbol{D}-\boldsymbol{A})\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{z}}{\boldsymbol{z}^T\boldsymbol{z}},$$

a uvrstimo li $\boldsymbol{z}=\boldsymbol{D}^{1/2}\boldsymbol{y}$, dobivamo rješenje početnog minimizacijskog problema $\boldsymbol{y}_1$:

$$\boldsymbol{y}_1=arg.min_{\boldsymbol{y}^T\boldsymbol{D}\boldsymbol{1}=0}\frac{\boldsymbol{y}^T(\boldsymbol{D}-\boldsymbol{A})\boldsymbol{y}}{\boldsymbol{y}^T\boldsymbol{D}\boldsymbol{y}}$$

U odnosu na komponente vektora $\boldsymbol{y}_1$ dijelimo naš graf na dvije skupine, to jest on nam daje najbolju realnu aproksimaciju segmentacije.

U implementaciji metode normaliziranog reza, za prema Fiedlerovom vektoru određujemo element nad kojim je normalizirani rez minimalan, te na temelju njegove pozicije dijelimo podatke na dvije skupine.

Dodatno, u implementaciji koristi se tehnika „sweep cut“ za pronalaženje optimalnog reza, to jest čvorovi grafa uzlazno se sortiraju prema vrijednostima u Fiedlerovom vektoru i nad tako sortiranim podatcima pronalazi se optimalni rez, detaljnije u literaturi [24].

Opisani algoritam normaliziranog reza možemo i prikazati kao:

1. ulazni podatci: podatci, gornja granica vrijednosti normaliziranog reza

2. konstrukcija matrice sličnosti $\boldsymbol{A}$ i dijagonalne matrice $\boldsymbol{D}$
3. konstrukcija normalizirane simetrične Laplaceove matrice $\boldsymbol{L}_{sym}$
4. izračunaj svojstvene vrijednosti i svojstvene vektore matrice $\boldsymbol{L}_{sym}$
5. uzlazno sortiraj čvorove prema vrijednostima Fiedlerovog vektora i podijeli ih u dvije skupine na temelju pozicije točke koja ima najmanju vrijednost normaliziranog reza
6. (opcionalno: rekurzivni koraci za daljnje podjele objašnjeni u sljedećem poglavlju)

Uzmimo za primjer skup od deset nasumično odabranih podataka:

$$D = \{(0.90, -0.37),\ (0.92,0.65),\ (0.37, -0.02),\ (-0.57,0.64),\ (0.66,0.29),$$

$$(0.26, -0.07),\ (-0.34, -0.11),\ (0.58, -0.41),\ (1.54,0.26),\ (0.21, -0.20)\}$$

Za računanje Laplaceove matrice koristit ćemo normalizirani izraz uz parametre : $\sigma_X = 0.5$, $n_{cut} = 0.78$ i matricu sličnosti prema Gaussovoj jezgrenoj funkciji (3), no bez korištenja iterativne Lanczosove metode budući da se radi o manjem skupu podataka. Za uzlazno sortirane čvorove prema Fiedlerovom vektoru normalizirane Laplaceove matrice računamo listu vrijednosti normaliziranog reza za svaki čvor, te odabiremo poziciju s najmanjom vrijednosti za rez.

$$\begin{bmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ -0.15 & -0.51 & 0.09 & 0.33 & -0.31 & 0.19 & 0.42 & 0.03 & -0.48 & 0.24 \end{bmatrix}$$

(Fiedlerov vektor)

$$[\ 1\ \ 8\ \ 4\ \ 0\ \ 7\ \ 2\ \ 5\ \ 9\ \ 3\ \ 6\ ]$$

(uzlazno sortirani čvorovi prema Fiedlerovom vektoru)

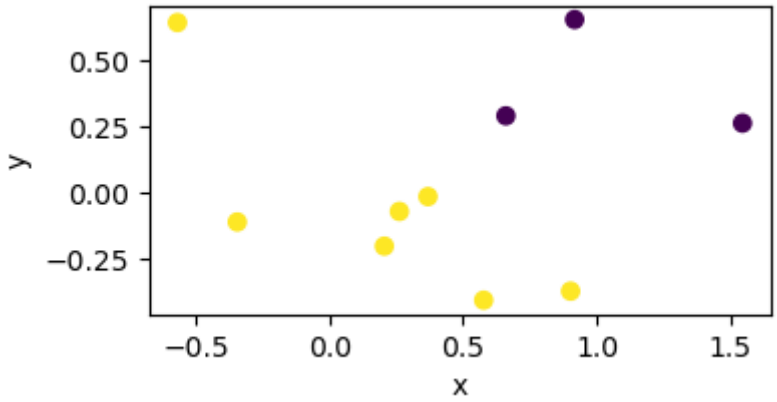
Na primjer, za element na indeksu 2 vrijednost normaliziranog reza dobivamo prema formuli (21) gdje nam je skup $A = \{1,8,4\}$ i $B = \{0,7,2,5,9,3,6\}$:

$$Ncut(A,B) = \frac{cut(A,B)}{assoc(A,V)} + \frac{cut(B,A)}{assoc(B,V)} = \frac{3.6}{6.06} + \frac{3.6}{21.44} = 0.76$$

$$\begin{bmatrix} 1 & 8 & \mathbf{4} & 0 & 7 & 2 & 5 & 9 & 3 & 6 \\ 1.07 & 0.82 & \mathbf{0.76} & 0.83 & 0.85 & 0.91 & 0.95 & 0.85 & 1.08 & x \end{bmatrix}$$

(vrijednosti normaliziranog reza sa označenim minimalnim elementom)

Kod elementa na poziciji 2. pronađena je najmanja vrijednost normaliziranog reza 0.76, dakle podjelom podataka na dvije skupine prema 2. elementu dolazimo do sljedećeg grupiranja na slici (Slika 4.3).



Slika 4.3 Prikaz grupiranih podataka korištenjem metode normaliziranog reza

Kako bismo poboljšali performanse implementiranog algoritma, vrijednost normaliziranog reza izračunava se za svaki l-ti element.

Postupak ćemo objasniti na prethodnom primjeru skupa od 10 podataka uz parametre : $\sigma_X = 0.5\,, n\_cut = 0.86,$ za $l = 2$ što znači da za svaki parni element (označen sivo) sortiranih čvorova prema Fiedlerovom vektoru izračunavamo vrijednost normaliziranog reza i iz tog smanjenog skupa vrijednosti biramo najmanju kao konačni rez. Postupak određivanja vrijednosti normaliziranog reza za određeni l-ti element istovjetan je postupku u prethodnom primjerom kada se vrijednost izračunavala za svaki element sortiranih čvorova prema Fiedlerovom vektoru.

$$\begin{bmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ -0.15 & -0.51 & 0.09 & 0.33 & -0.31 & 0.19 & 0.42 & 0.03 & -0.48 & 0.24 \end{bmatrix}$$

(Fiedlerov vektor)

$$[\ 1\ \ 8\ \ 4\ \ 0\ \ 7\ \ 2\ \ 5\ \ 9\ \ 3\ \ 6\ ]$$

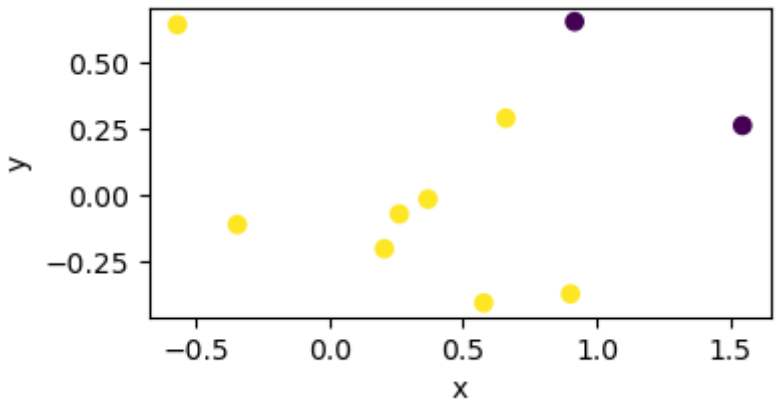
(sortirani čvorovi prema Fiedlerovom vektoru)

$$\begin{bmatrix} \mathbf{8} & 0 & 2 & 9 & 6 \\ \mathbf{0.820} & 0.826 & 0.915 & 0.865 & 0.994 \end{bmatrix}$$

(vrijednosti normaliziranog reza sa označenim minimalnim elementom)

U ovom slučaju više se ne provjerava vrijednost reza za element na poziciji 2., dakle najmanja vrijednost u novom smanjenom skupu je na poziciji 0. i čvorovi se dijele na

temelju te pozicije. Dakle, u konačnom grupiranju na slici (Slika 4.4) primjećujemo da se razlikuje od grupiranja (Slika 4.3), to jest da je skupina označeno ljubičasto manja budući da točka reza koja se provjeravala za grupiranje bez parametra *l* nije bila provjeravana za grupiranje s parametrom *l=2*.



Slika 4.4 Prikaz grupiranih podataka korištenjem metode normaliziranog reza s parametrom *l*

Kvaliteta metode ovisi o veličini podataka te opada proporcionalno s povećanjem vrijednosti l.

4.1.3. Metoda KVV

Metoda KVV (Kannan-Vempala-Vetta) slijedi algoritam predstavljen u literaturi [25], a detaljnije opisan u literaturi [26]. Ona je zapravo varijanta metode normaliziranog reza, no koristi se drukčiji pristup u pronalaženju optimalnog reza. Umjesto vrijednosti normaliziranog reza, u ovom se algoritmu koristi Cheegerova vodljivost.

Optimalni rez pronalazimo na temelju Cheegerove vodljivosti $\phi(A,B)$*:*

$$\phi(A,B) = \frac{cut(A,B)}{\min(assoc(A,V)\,, assoc(B,V))}. \tag{23}$$

Dakle, ovisno o Cheegerovoj vodljivosti pronalazimo rez tako da uzimamo minimalnu vrijednost, te na temelju nje dijelimo čvorove u dvije skupine.

Uzmimo za primjer skup od deset nasumično odabranih podataka:

$$D = \{(0.48,-0.36),(1.39,0.37),(0.99,-0.33),(1.21,-0.25),(-0.25,-0.45),$$
$$(0.35,0.20),(0.90,0.11),(-0.60,-0.24),(-0.27,0.12),(-0.38,0.31)\}.$$

Za računanje Laplaceove matrice korištena je Gaussova jezgrena funkcija (3) sa $\sigma_X = 0.5$. Za uzlazno sortirane čvorove prema Fiedlerovom vektoru računamo listu vrijednosti Cheegerove vodljivosti prema izrazu (23) za svaki čvor, te odabiremo poziciju s najmanjom vrijednosti Cheegerove vodljivosti za rez:

$$\begin{bmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ -0.11 & -0.3 & -0.34 & -0.37 & 0.32 & 0.01 & -0.33 & 0.39 & 0.39 & 0.37 \end{bmatrix}$$

(Fiedlerov vektor)

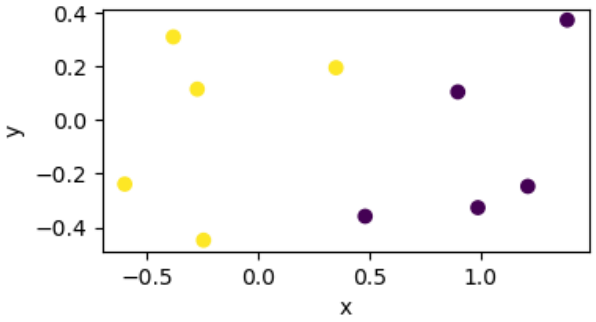
$$[\ 3\ \ 2\ \ 6\ \ 1\ \ 0\ \ 5\ \ 4\ \ 9\ \ 8\ \ 7\ ]$$

(uzlazno sortirani čvorovi prema Fiedlerovom vektoru)

$$\begin{bmatrix} 3 & 2 & 6 & 1 & \mathbf{0} & 5 & 4 & 9 & 8 & 7 \\ 1.0 & 0.661 & 0.465 & 0.288 & \mathbf{0.216} & 0.229 & 0.426 & 0.743 & 1.0 & x \end{bmatrix}$$

(vrijednosti Cheegerove vodljivosti sa označenim minimalnim elementom)

Element s najmanjom vrijednosti Cheegerove vodljivosti nalazi se na poziciji 4., dakle sortirani čvorovi dijele se u dvije skupine po pet čvorova, konačno grupiranje prikazano na slici (Slika 4.5).



Slika 4.5 Prikaz grupiranih podataka korištenjem KVV metode

Opisani algoritam možemo i prikazati kao:

1. ulazni podatci: podatci, gornja granica vrijednosti Cheegerove vodljivosti
2. konstrukcija matrice sličnosti $\boldsymbol{A}$ i dijagonalne matrice $\boldsymbol{D}$
3. konstrukcija normalizirane simetrične Laplaceove matrice $\boldsymbol{L}_{sym}$
4. izračunaj svojstvene vrijednosti i svojstvene vektore matrice $\boldsymbol{L}_{sym}$
5. uzlazno sortiraj čvorove prema vrijednostima Fiedlerovog vektora i podijeli ih u dvije skupine na temelju vrijednosti komponente za koju dobivamo najmanju vrijednost Cheegerove vodljivosti

6. (opcionalno: rekurzivni koraci za daljnje podjele objašnjeni u sljedećem poglavlju)

4.2. Prikaz metoda grupiranja u više skupina

Za dobivanje većeg broja grupa, prethodno opisani postupci se rekurzivno primjenjuju na svakoj novoj skupini dok ne dobijemo željeni broj skupina. Dodatno, opisana je i spektralna metoda koja za podjelu podataka koristi veći broj svojstvenih vektora.

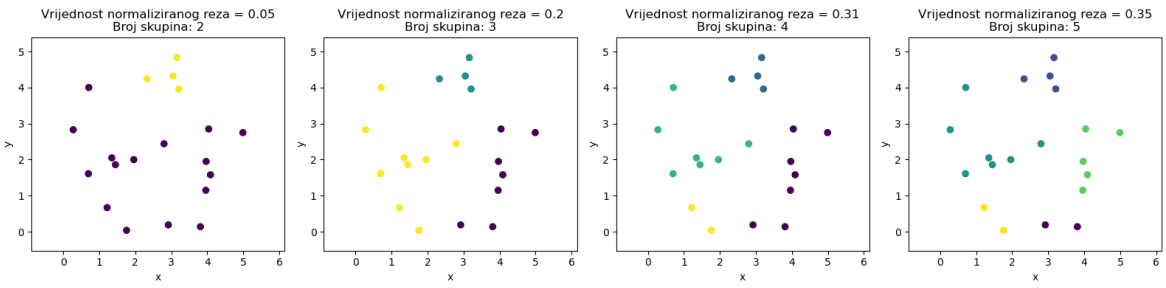
4.2.1. Podjela u više skupina koristeći Fiedlerov vektor

Postupak grupiranja prema Fiedlerovom vektoru nenormalizirane Laplaceove matrice rekurzivno se ponavlja nad svakom novonastalom skupinom dok ne dosegnemo predodređeni kriterij. Matricu sličnosti koja ulazi u sljedeći rekurzivni korak dobivamo tako da iz matrice sličnosti iz prethodnog koraka uzimamo retke i stupce koji odgovaraju čvorovima u novonastaloj skupini.

Kao kriterij zaustavljanja kod osnovnog algoritma odabiremo maksimalni broj skupina.

Kao kriterij zaustavljanja rekurzivnog algoritma kod metode normaliziranog reza moguće je odabrati maksimalnu dopuštenu vrijednost normaliziranog reza kojom određujemo prihvatljivu razinu grupiranja podataka, a time i konačni broj skupina. Ako je vrijednost normaliziranog reza veća od zadanog praga, to znači da podjela grupa nije zadovoljavajuća, to jest grupe nisu dovoljno dobro odvojene i daljnja podjela gubi smisao. Alternativna opcija je odrediti najveći željeni broj skupina, kao i kod osnovnog algoritma.

Na sljedećim primjerima na slici (Slika 4.6) prikazani su rezultati grupiranja skupa od 20 podataka nakon svake podjele dok se nije ispunio zaustavni uvjet grupiranja podataka u 5 skupina korištenjem metode normaliziranog reza.



Slika 4.6 Grupiranje podataka u više skupina

Slično, kriterij zaustavljanja kod metode KVV bira se na isti način uz promjene u načinu na koji pripremamo ulaznu matricu za sljedeći rekurzivni korak. U svakom rekurzivnom koraku koristimo izvornu matricu sličnosti na temelju koje gradimo Laplaceovu matricu iz koje izdvajamo odgovarajuće retke i stupce na temelju indeksa točaka koje čine trenutnu skupinu.

Sljedeći primjer je kreiranje ulazne normalizirane Laplaceove matrice koristeći Gaussovu jezgrenu funkciju (3) na skupu podataka:

$$D = \{(1,1), (2,1), (1,2), (2,2), (1.5,1.5), (3,1), (4,1), (3,2), (4,2), (3.5,1.5)\}$$

u drugom rekurzivnom koraku gdje ulazimo sa skupinom podataka rednih brojeva [5,6,7,8,9], dakle dobivamo podmatricu od pet redaka i pet stupaca, prikazanu na slici (Slika 4.7). U ovom slučaju odabrani su pozicije [5,6,7,8,9] jer je to dio implementiranog algoritma u Pythonu, za prvu grupu koja ide u daljnje grupiranje uzima se podskup desno od pozicije reza.

$$L=\begin{bmatrix} 1 & -0.18 & -0.21 & -0.02 & -0.37 & 0 & 0 & 0 & 0 & 0 \\ -0.18 & 1 & -0.2 & -0.17 & -0.33 & -0.17 & 0 & -0.02 & 0 & -0.01 \\ -0.21 & -0.02 & 1 & -0.18 & -0.37 & 0 & 0 & 0 & 0 & 0 \\ -0.02 & -0.17 & -0.18 & 1 & -0.33 & -0.02 & 0 & -0.17 & 0 & -0.01 \\ -0.37 & -0.33 & -0.37 & -0.33 & 1 & -0.01 & 0 & -0.01 & 0 & 0 \\ 0 & -0.17 & 0 & -0.02 & -0.01 & 1 & -0.18 & -0.17 & -0.02 & -0.33 \\ 0 & 0 & 0 & 0 & 0 & -0.18 & 1 & -0.02 & -0.21 & -0.37 \\ 0 & -0.02 & 0 & -0.17 & -0.01 & -0.17 & -0.02 & 1 & -0.18 & -0.33 \\ 0 & 0 & 0 & 0 & 0 & -0.02 & -0.21 & -0.18 & 1 & -0.37 \\ 0 & -0.01 & 0 & -0.1 & 0 & -0.33 & -0.37 & -0.33 & -0.37 & 1 \end{bmatrix}$$

Slika 4.7 Kreiranje Laplaceove podmatrice kod metode KVV

Budući da u tom slučaju novonastala matrica nije normalizirana, potrebno ju je normalizirati jednom od dviju sljedećih metoda tako da je suma svakog retka jednaka 1. Prva metoda koristi se množenjem kako bismo skalirali sve elemente u retku tako da je njihova suma jednaka 1. Druga metoda koristi se zbrajanjem, elementima na dijagonali pridodajemo vrijednost potrebnu da zbroj elemenata retka bude jednak 1. Detaljni za obje metode nalaze se u literaturi [26].

4.2.2. Metoda NJW

Metoda NJW (Ng-Jordan-Weiss) slijedi algoritam opisan u radu [27], uz iznimku da se za izračunavanje svojstvenih vrijednosti i vektora koristi Laplaceova matrica kao što je u literaturi [12].

Od osnovnog spektralnog algoritma grupiranja metoda NJW razlikuje se u tome kako koristimo dobivene svojstvene vektore. Za ovu metodu specifično je da se uzima k svojstvenih vektora koji pripadaju k najvećim svojstvenim vrijednostima normalizirane matrice sličnosti $\boldsymbol{A}_{norm}$ [27], to jest ti vektori odgovaraju k najmanjim svojstvenim vrijednostima normalizirane Laplaceove matrice $\boldsymbol{L}_{sym}$ [12]. Budući da se ne uzima isključivo svojstveni vektor koji odgovara drugoj najmanjoj svojstvenoj vrijednosti, izbjegava se potreba za rekurzivnim pozivima algoritama koji dijele graf u dvije skupine kako bismo ostvarili željeni broj skupina.

Na temelju dobivenih svojstvenih vektora konstruiramo matricu $\boldsymbol{X} \in \mathbb{R}^{nxk}$, u kojoj su vektori poredani u stupce:

$$\boldsymbol{X} = [\boldsymbol{x}_1,\ \boldsymbol{x}_2, \boldsymbol{x}_3, \dots\ \boldsymbol{x}_k]\ ,$$

gdje su $\boldsymbol{x}_1,\ \boldsymbol{x}_2,\ \boldsymbol{x}_3, \dots\ \boldsymbol{x}_k$ svojstveni vektori Laplaceove matrice $\boldsymbol{L}_{sym}$, k odabrani broj skupina, a n duljina svojstvenih vektora, to jest veličina početne matrice.

Konstrukcijom matrice $\boldsymbol{X}$ dobivamo novi prikaz strukture podatka, svaki redak te matrice predstavlja jednu točku, to jest njenu projekciju, u k-dimenzionalnom prostoru definiranom strukturom povezanosti u grafu.

Sljedeći važan korak je normalizacija matrice $\boldsymbol{X}$ čime dobivamo matricu $\boldsymbol{Y}$ u kojoj su svi redci matrice $\boldsymbol{X}$ skalirani tako da imaju jediničnu duljinu:

$$\boldsymbol{Y}_{ij} = \frac{\boldsymbol{X}_{ij}}{\sqrt{\sum_j \boldsymbol{X}_{ij}^2}}. \tag{24}$$

Nužno je da su svi vektori istih duljina kako različite norme ne bi utjecale na izračun udaljenosti među podatcima u kasnijim koracima. Tim postupkom zadržavamo samo smjerove vektora, tako da sličnosti ovise isključivo o kutu između vektora, a ne o njihovoj normi. Rezultat ove normalizacije jest mapiranje podataka na površinu jedinične kugle u k-dimenzionalnom prostoru $\mathbb{R}^k$ [27].

Normalizacijom dobivenih vektora sadržanih u matrici $\boldsymbol{Y}$ imamo novu reprezentaciju podataka nad kojom moramo iskoristiti algoritam k-sredina kako bi se provelo grupiranje.

Algoritam se može podijeli u 7 koraka, detaljnije u literaturi [27]:

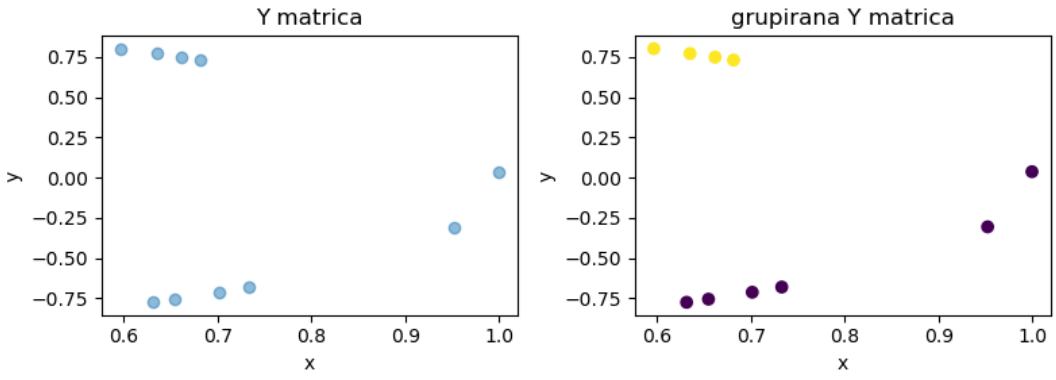
1. ulazni podatci: podatci, broj skupina k

2. konstrukcija matrice sličnosti $\boldsymbol{A}$ i dijagonalne matrice $\boldsymbol{D}$
3. konstrukcija normalizirane simetrične Laplaceove matrice $\boldsymbol{L}_{sym}$
4. izračunaj svojstvene vrijednosti i svojstvene vektore matrice $\boldsymbol{L}_{sym}$
5. kreiraj matricu $\boldsymbol{X}$ koristeći svojstvene vektore matrice $\boldsymbol{L}_{sym}$
6. koristeći svojstvene vektore pripadne k najmanjim svojstvenim vrijednostima konstruiraj matricu $\boldsymbol{Y}$ prema (24)
7. nad matricom $\boldsymbol{Y}$ provedi algoritam k-sredina

Uzmimo isti primjer skupa od 10 nasumičnih točaka kao kod KVV metode kako bismo prikazali opisano. Odredit ćemo da će se podatci grupirati u dvije skupine, dakle iz svojstvenih vektora koji pripadaju dvjema najmanjim svojstvenim vrijednostima stvaramo matricu $\boldsymbol{X}$ i pripadnu matricu $\boldsymbol{Y}$:

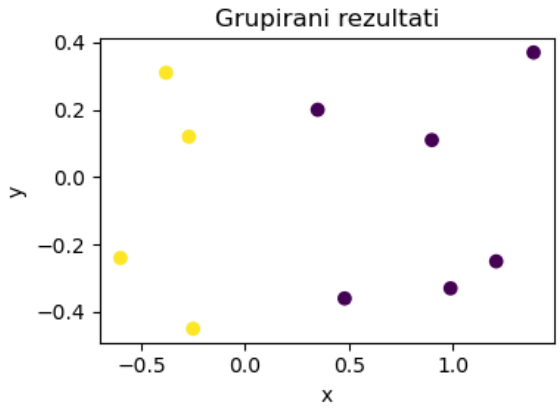
$$\boldsymbol{X} = \begin{bmatrix} 0.33 & -0.11 \\ 0.24 & -0.3 \\ 0.34 & -0.34 \\ 0.32 & -0.37 \\ 0.3 & 0.32 \\ 0.33 & 0.01 \\ 0.35 & -0.33 \\ 0.29 & 0.39 \\ 0.34 & 0.39 \\ 0.3 & 0.37 \end{bmatrix} \quad \boldsymbol{Y} = \begin{bmatrix} 0.95 & -0.31 \\ 0.63 & -0.78 \\ 0.7 & -0.71 \\ 0.65 & -0.76 \\ 0.68 & 0.73 \\ 1.0 & 0.03 \\ 0.73 & -0.68 \\ 0.6 & 0.8 \\ 0.66 & 0.75 \\ 0.63 & 0.77 \end{bmatrix}$$

Na slici (Slika 4.8) lijevo su prikazane vrijednosti dobivene $\boldsymbol{Y}$ matrice, a desno je prikazan rezultat primijenjenog algoritma k-sredina nad točkama matrice $\boldsymbol{Y}$.



Slika 4.8 Prikaz matrice $\boldsymbol{Y}$ prije i nakon grupiranja algoritmom k-sredina

Posljedično, na slici (Slika 4.9) možemo primijetiti da kada primijenimo istu dodjelu točaka skupinama na naše početne podatke dobivamo grupirane podatke u dvije skupine.



Slika 4.9 Prikaz grupiranih podataka metodom NJW

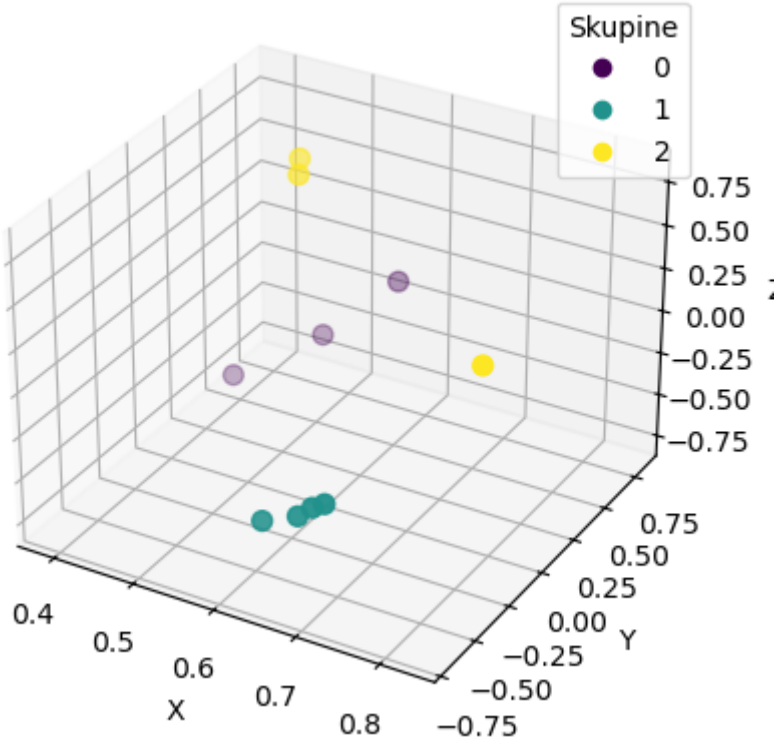
Na sljedećem primjeru prikazat ćemo kako izgleda grupiranje skupa podataka od 10 nasumično odabranih točaka u tri skupine. Za grupiranje u tri skupine potrebna su nam tri svojstvena vektora koji odgovaraju trima najmanjim svojstvenim vrijednostima:

$$\boldsymbol{v_1} = \begin{bmatrix} 0.33 \\ 0.29 \\ 0.33 \\ 0.32 \\ 0.32 \\ 0.34 \\ 0.28 \\ 0.33 \\ 0.35 \\ 0.26 \end{bmatrix} \quad \boldsymbol{v_2} = \begin{bmatrix} -0.32 \\ 0.43 \\ -0.32 \\ 0.25 \\ -0.33 \\ 0.15 \\ 0.42 \\ -0.31 \\ -0.12 \\ 0.35 \end{bmatrix} \quad \boldsymbol{v_3} = \begin{bmatrix} -0.1 \\ -0.32 \\ -0.14 \\ 0.49 \\ -0.2 \\ 0.53 \\ -0.08 \\ -0.08 \\ 0.2 \\ -0.5 \end{bmatrix}$$

U tom slučaju matrice **Y** i **X** imaju tri stupca:

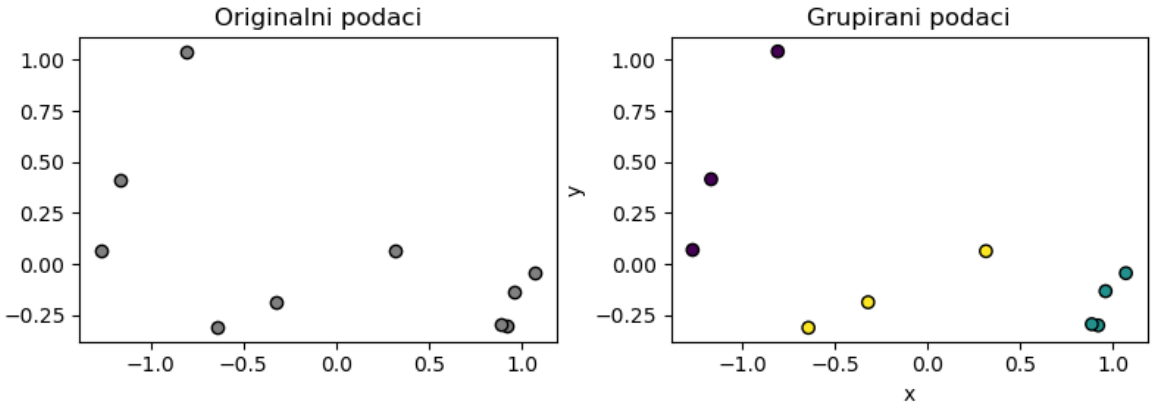
$$\boldsymbol{X} = \begin{bmatrix} 0.33 & -0.32 & -0.1 \\ 0.29 & 0.43 & -0.32 \\ 0.33 & -0.32 & -0.14 \\ 0.32 & 0.25 & 0.49 \\ 0.32 & -0.33 & -0.2 \\ 0.34 & 0.15 & 0.53 \\ 0.28 & 0.42 & -0.08 \\ 0.33 & -0.31 & -0.08 \\ 0.35 & -0.12 & 0.2 \\ 0.26 & 0.35 & -0.5 \end{bmatrix} \quad \boldsymbol{Y} = \begin{bmatrix} 0.7 & -0.68 & -0.21 \\ 0.47 & 0.71 & -0.56 \\ 0.68 & -0.67 & -0.29 \\ 0.5 & 0.4 & 0.77 \\ 0.64 & -0.66 & -0.39 \\ 0.53 & 0.23 & 0.82 \\ 0.55 & 0.82 & -0.16 \\ 0.72 & -0.68 & -0.17 \\ 0.83 & -0.28 & 0.48 \\ 0.39 & 0.53 & -0.76 \end{bmatrix},$$

a njihova reprezentacija u 3-dimenzionalnom prostoru prikazana je na slici (Slika 4.10).



Slika 4.10 Vizualizacija matrice **Y**

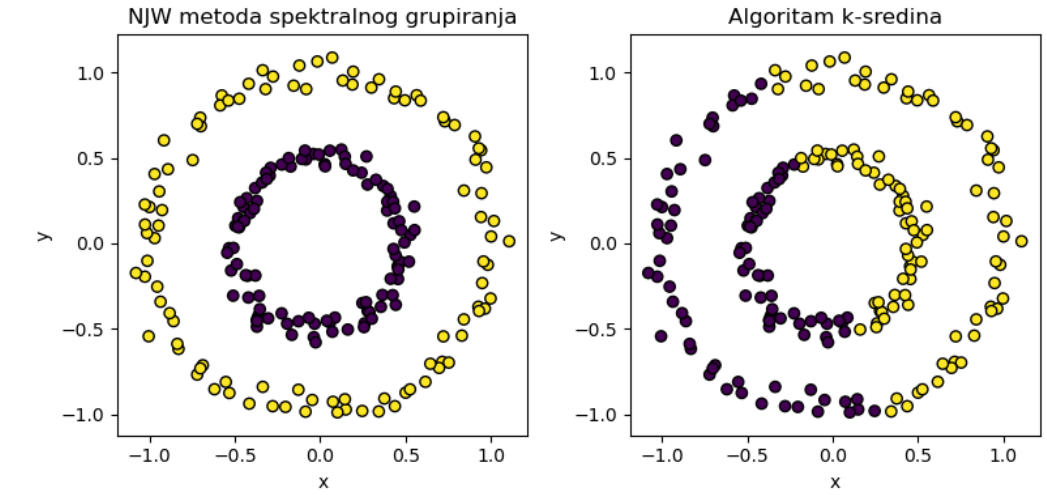
Na slici (Slika 4.11 Grupiranje podataka u 3 skupine metodom NJW) prikazan je početni skup podataka i njihovo konačno grupiranje.



Slika 4.11 Grupiranje podataka u 3 skupine metodom NJW

Prednost ove metode je u tome što, za razliku od običnog algoritma *k*-sredina koji ovisi samo o prostornoj udaljenosti, ovdje grupiranje ovisi i o strukturnoj povezanosti podataka. Kao i kod prethodnih metoda, njezini rezultati ovise o parametrima povezanim s matricom sličnosti, a uz to i odabranom broju skupina.

Na sljedećem primjeru na slici (Slika 4.12) desno vidimo rezultate koje algoritam *k*-sredina daje na skupovima podataka s kompleksnijim, nelinearnim strukturama, u ovom slučaju to su koncentrične kružnice generirane u Pythonu korištenjem biblioteke *sklearn.datasets* [28]. Formirane su dvije skupine, ali su one određene isključivo na temelju euklidske udaljenosti (1), pri čemu algoritam ne uspijeva prikazati stvarnu povezanost među podacima. U usporedbi s rezultatima NJW algoritma u kombinaciji s matricom sličnosti konstruirane Gaussovom jezgrenom funkcijom (3) gdje su skupine jasno razdvojene u skladu s kružnicama, može se jasno uočit da korištenjem Laplaceove matrice podaci se mapiraju u novi prostor u kojem je moguće razdvojiti i nelinearne strukture. Na taj način spektralno grupiranje omogućuje preciznije prepoznavanje grupa unutar složenih oblika, što je jasno vidljivo na ovih grafovima.



Slika 4.12 Usporedba grupiranja koncentričnih krugova NJW metodom i algoritmom k-sredina

Sličan pristup uzimanja prvih *k* svojstvenih vektora za grupiranje može se i primijeniti nad metodama iz poglavlja s biparticijom, metodi normaliziranog reza i KVV metodi [29]. Uzmimo za primjer metodu normaliziranog reza, ako uzmemo k' svojstvenih vektora, gdje je $k' > k,$ i primjerimo algoritam *k*-sredina nad njima, dobit ćemo veći broj skupina od željenog. Sljedeći korak bi bio spajati skupine $A_1, \dots, A_k$ koje minimiziraju vrijednost normaliziranog reza prema izrazu:

$$Ncut_k = \frac{cut(A_1, V - A_1)}{assoc(A_1, V)} + \cdots + \frac{cut(A_k, V - k)}{assoc(A_k, V)} \tag{25}$$

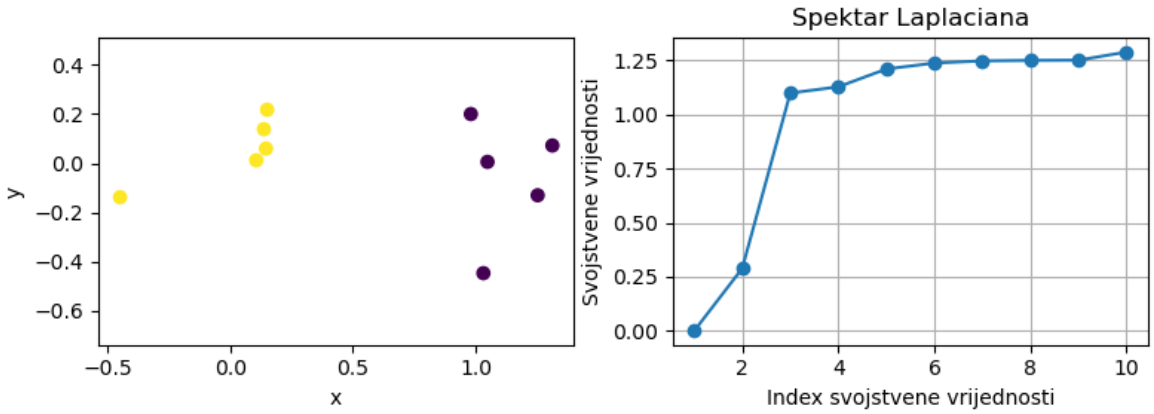
gdje A_i predstavlja podskup *i* skupa V [22].

Alternativa opisanom koraku je kreiranje manjeg grafa u kojem su čvorovi skupine iz prethodnog koraka, a bridovi veze između skupina. Nad njime rekurzivno se primjenjuje metoda normaliziranog reza za dobivanje željenog broja skupina [22].

4.2.3. Određivanje optimalnog broja skupina kod spektralnog grupiranja

Svojstvene vrijednosti pomažu nam odrediti optimalan broj grupa, to jest prikazuju nam jačinu poveznica između potencijalnih grupa. Na sljedećem primjeru koristeći prethodne podatke, prikazane su njihove svojstvene vrijednosti. Niske svojstvene vrijednosti označavaju da graf ima dobro definirane grupe koje su međusobno slabo povezane. Kod njih tražimo prvi nagli skok u vrijednosti koji nam govori da daljnje podjele podataka neće biti korisne jer su podatci previše povezani [12].

Gledajući lijevi graf na slici (Slika 4.13) vidimo da grupiranje podataka ima najviše smisla za dvije grupe, te na desnom grafu nakon druge svojstvene vrijednosti vidimo veliki skok kojim nam to potvrđuje.



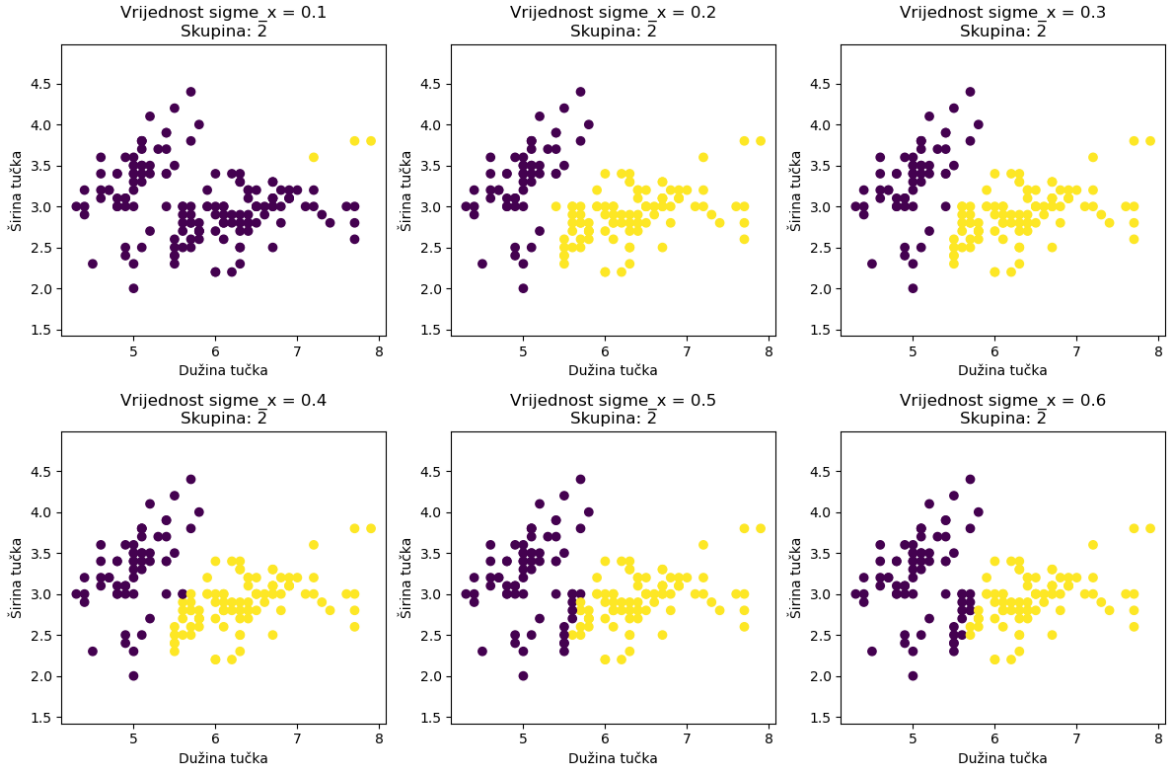
Slika 4.13 Prikaz svojstvenih vrijednosti Laplaceove matrice

5. Rezultati i usporedba metoda

5.1. Primjeri s grafovima

Sljedeći primjeri rađeni su na skupu podataka IRIS [30] koji sadrži morfološke značajke za 150 uzoraka cvijeća. Radi lakše vizualizacije rezultata algoritama grupiranja iz skupa su uzete samo prve dvije značajke širine i dužine tučka, to jest podatke ćemo prikazivati samo u dvodimenzionalnom prostoru.

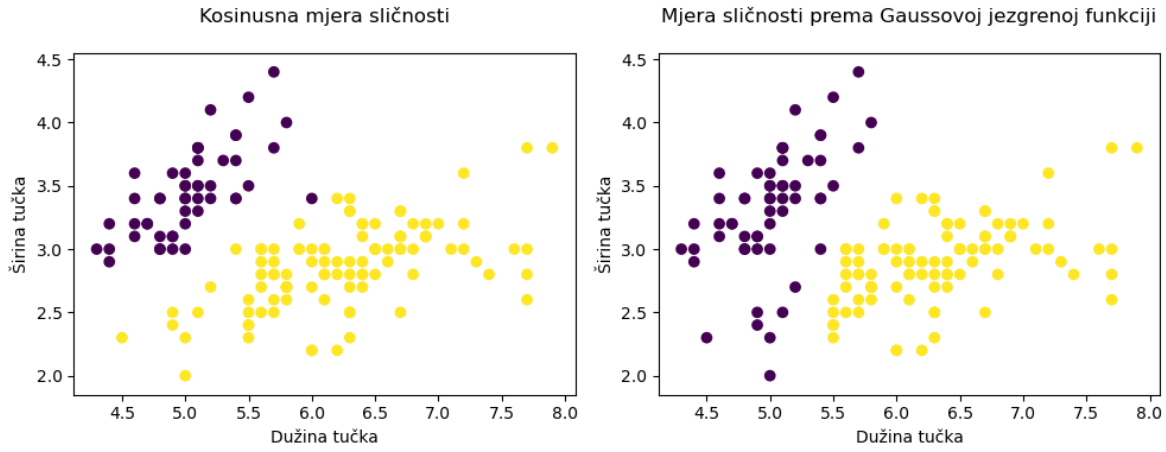
Budući da u sljedećim primjerima matrice sličnosti napravljene koristeći Gaussovu jezgrenu funkciju (3) promotrit ćemo na slici (Slika 5.1) utjecaj odabira parametra σ_x na rezultate osnovnog algoritma spektralnog grupiranja, to jest metode grupiranja koristeći samo Fiedlerov vektor opisanog u poglavlju 4.1.1..



Slika 5.1 Utjecaj parametra σ_x na rezultate grupiranja podataka

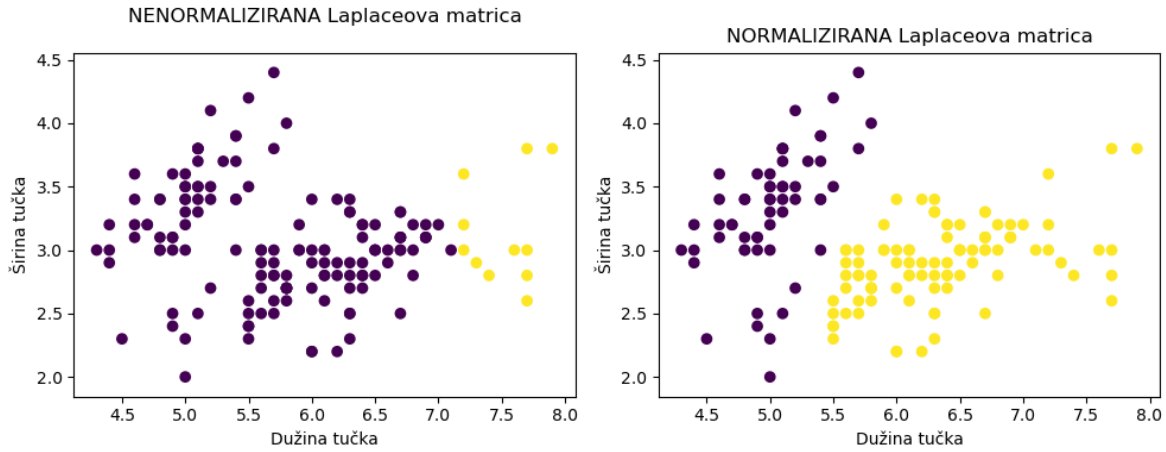
Primjećujemo da parametar σ_x igra ključnu ulogu u rezultatima grupiranja određujući prostorno mjerilo na temelju kojeg određujemo sličnost podataka. Odabirom premalenog parametra možemo dobiti izolirane podatke.

Na konačno grupiranje utječe i odabir mjere sličnosti na kojoj se temelj matrica sličnosti $\boldsymbol{A}$. U sljedećem primjeru (Slika 5.2 Utjecaj mjere sličnosti na rezultate grupiranja podataka) prikazani su rezultati grupiranja sa dvije različite mjere sličnosti, desno – prema Gaussovoj jezgrenoj funkciji (3) $\sigma_X = 0.3$ (3), i lijevo – prema kosinusnoj mjeri sličnosti (2). Korištena je osnovna metoda spektralnog grupiranja prema Fiedlerovom vektoru.



Slika 5.2 Utjecaj mjere sličnosti na rezultate grupiranja podataka

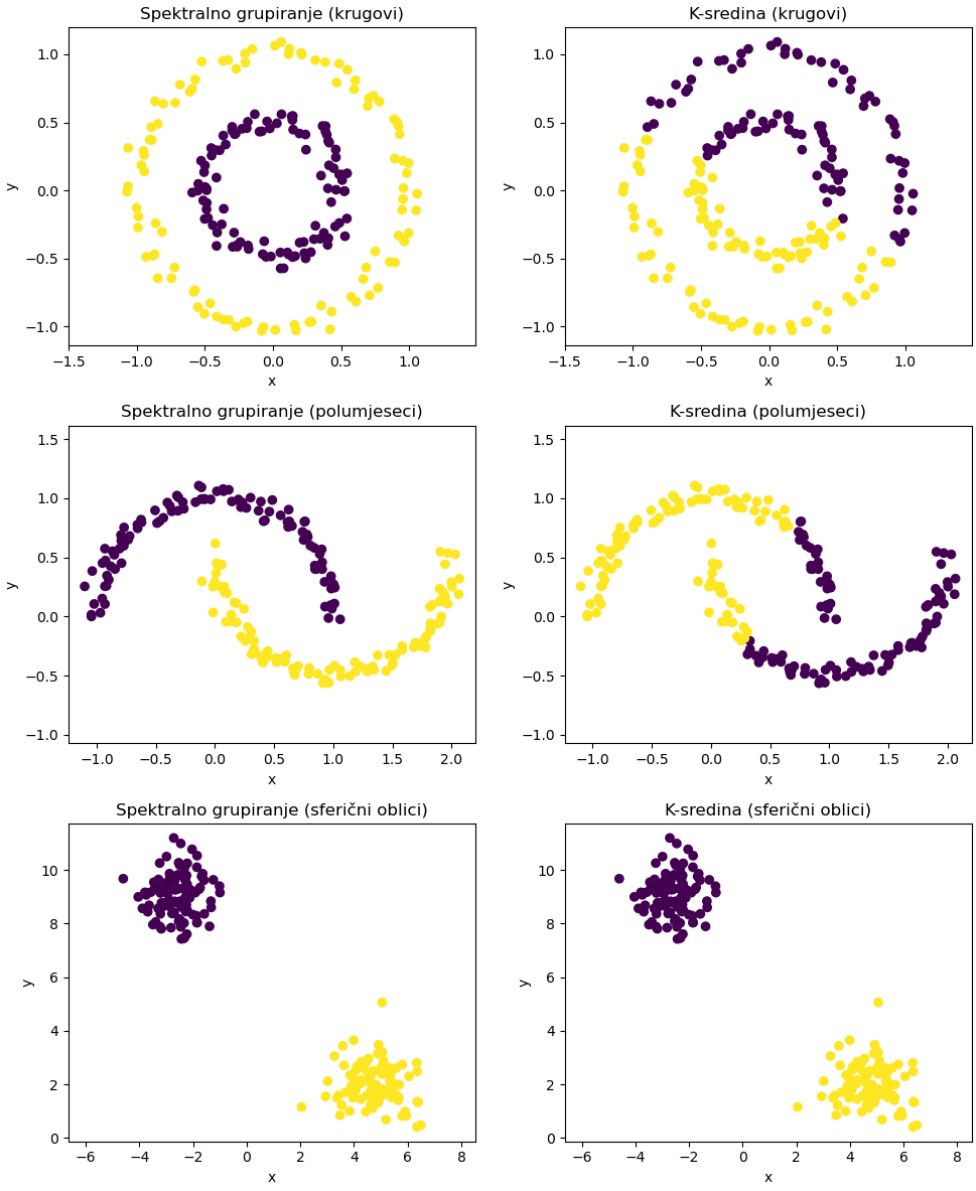
Još jedna stvar koju je potrebno uzeti u obzir je i odabir normalizirane ili nenormalizirane Laplaceove matrice jer to može dovesti do velikih razlika u rezultatima. Na sljedećim grafovima na slici (Slika 5.3) korištena je Gaussova jezgrena funkcija (3) za matricu sličnosti s parametrom $\sigma_X = 0.3$ s temeljnim spektralnim algoritmom iz potpoglavlja 4.1.1. Normalizacijom Laplaceove matrice uzimamo u obzir i ukupnu povezanost podataka te dobivamo uravnoteženije skupine.



Slika 5.3 Utjecaj normalizacije Laplaceove matrice na rezultate grupiranja podataka

Jedna od glavnih prednosti metode spektralnog grupiranja u usporedbi s algoritmom k-sredina jest da oblik podataka ne utječe na uspješnost grupiranja. Nije potrebno da oblici

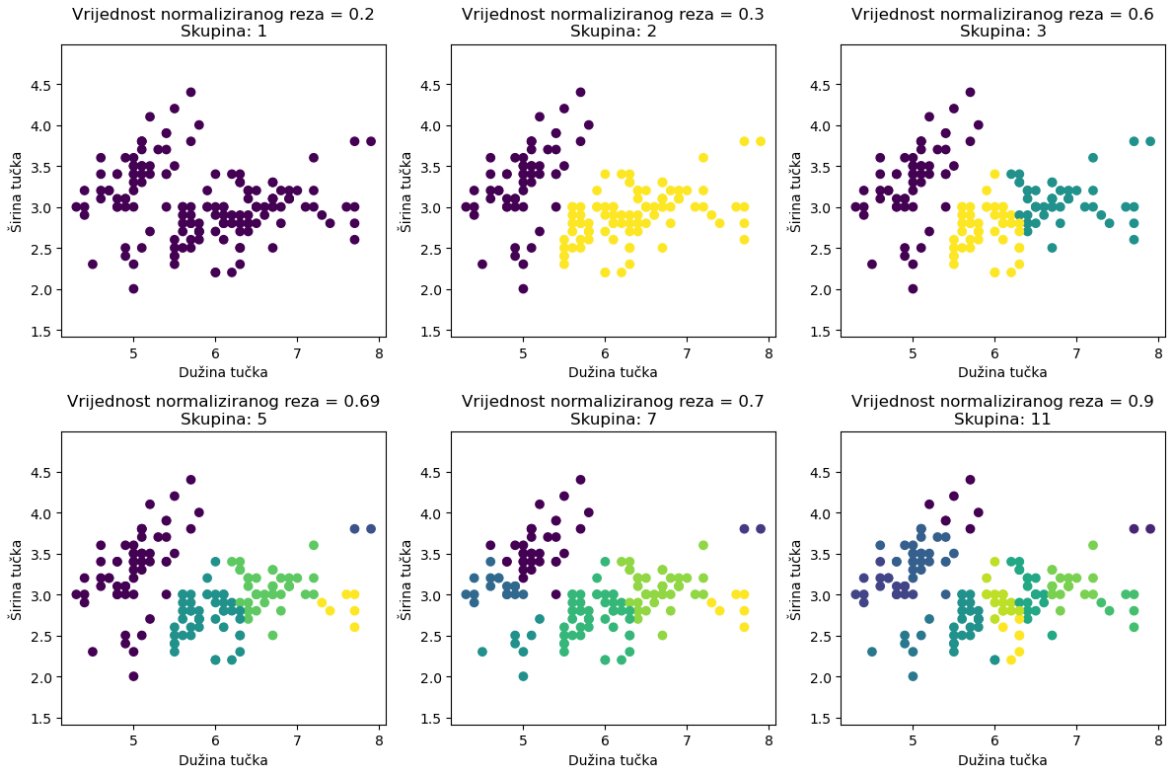
budu konveksni ili linearni. Kao što vidimo na grafovima na slici (Slika 5.4) s raznim oblicima (lijevo – osnovna metoda spektralnog grupiranja, desno – algoritam *k*-sredina), jedino uspješno grupiranje algoritma *k*-sredina je u zadnjem retku u kojem su podatci sfernog oblika. Skupovi su generirani u Pythonu korištenjem biblioteke *sklearn.datasets* [28].



Slika 5.4 Usporedba temeljnog spektralnog grupiranja i algoritma *k*-sredina na raznim oblicima podataka

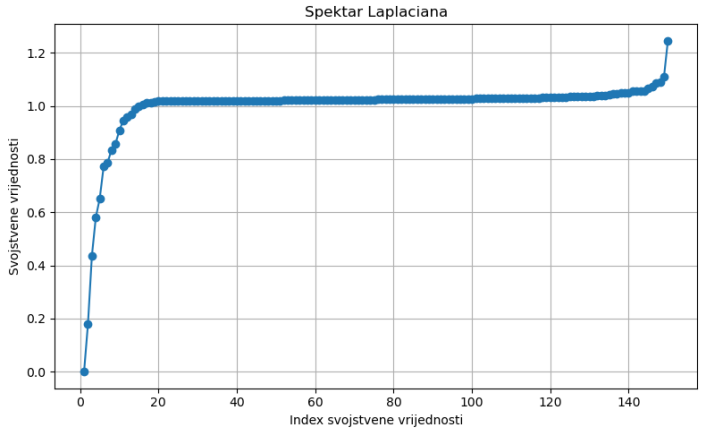
Kod metode normaliziranog reza postoji ovisnost o određenim parametrima. Uz parametre potrebne za matricu sličnosti, dodatno imamo i prag normaliziranog reza, točnije najveća vrijednost normaliziranog reza nakon koje se podatci više neće dijeliti na dodatne podskupine. Na sljedećem primjeru na slici (Slika 5.5) prikazano je 6 različitih vrijednosti

normaliziranog reza uz parametar $\sigma_X = 0.5$ za Gaussovu jezgrenu funkciju sličnosti (3). Možemo primijetiti da s višim vrijednostima normaliziranog reza imamo i više skupina.



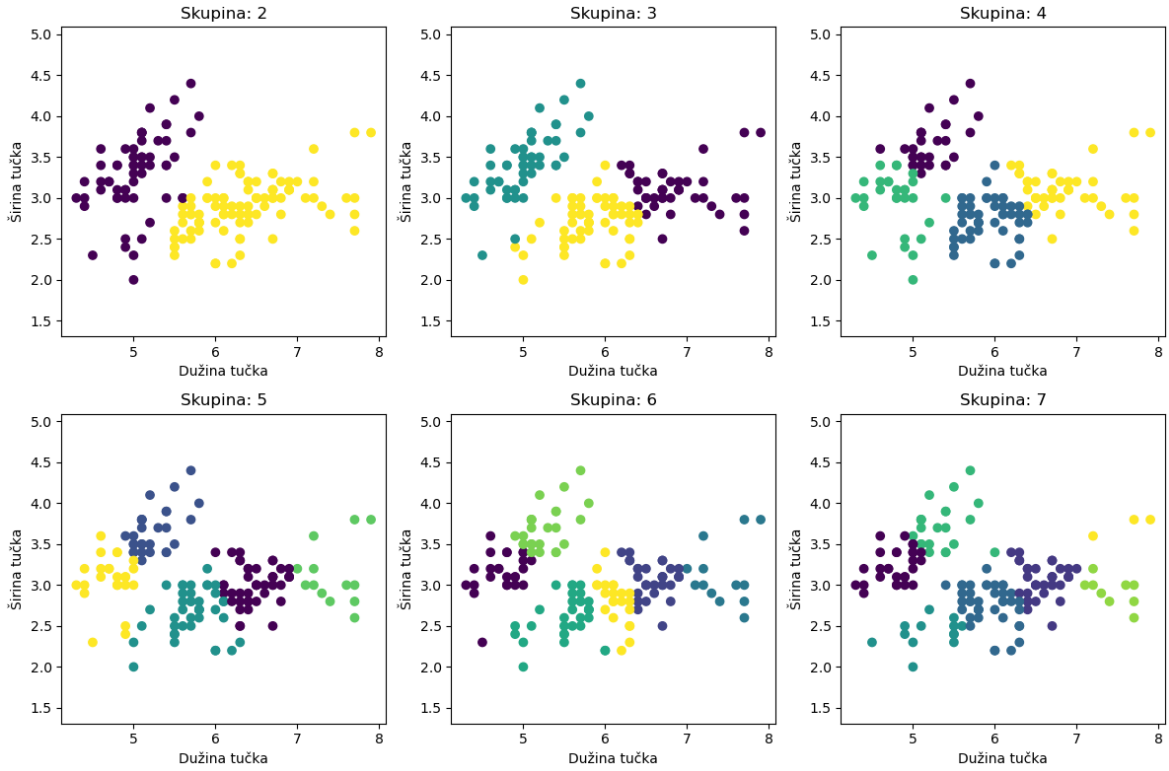
Slika 5.5 Utjecaj vrijednosti normaliziranog reza na grupiranje podataka

Na sljedećem linijskom grafu na slici (Slika 5.6) prikazane su svojstvene vrijednosti početne normalizirane Laplaceove matrice. Najveći skok je nakon svojstvene vrijednosti na drugoj poziciji, što nam govori da je najbolja podjela podataka na dva skupa, a da druge ne daju značajnije rezultate, što i vidimo na prethodnih grafovima. Iako je za IRIS skup podataka očekivano razdvajanje u tri skupine, podatci se prirodno razdvajaju u dvije jasno odvojene skupine [1].



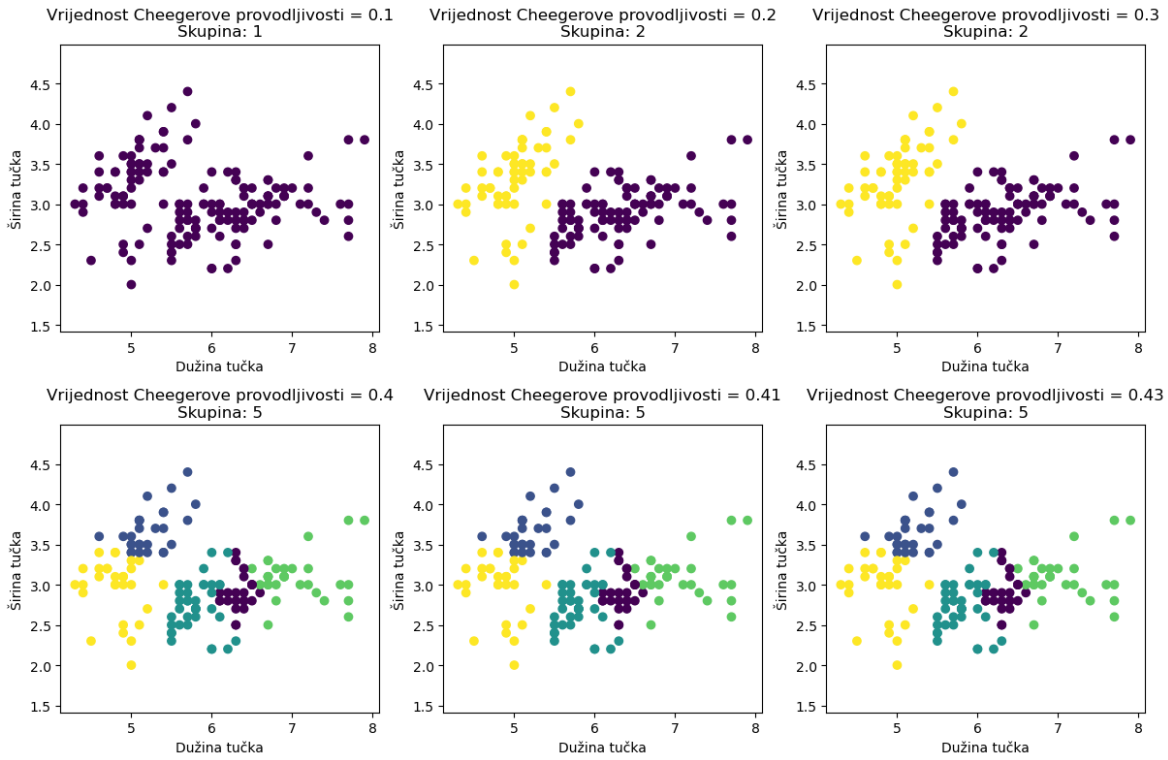
Slika 5.6 Prikaz svojstvenih vrijednosti

Kod metode NJW ulazni parametar je broj skupina koje konačno želimo, dakle imamo veću kontrolu nad ishodom nego kod metode normaliziranog reza, a ovisnost o tom parametru prikazana je na sljedećim grafovima na slici (Slika 5.7), uz dodatni ulazni parametar $\sigma_X = 0.5$.

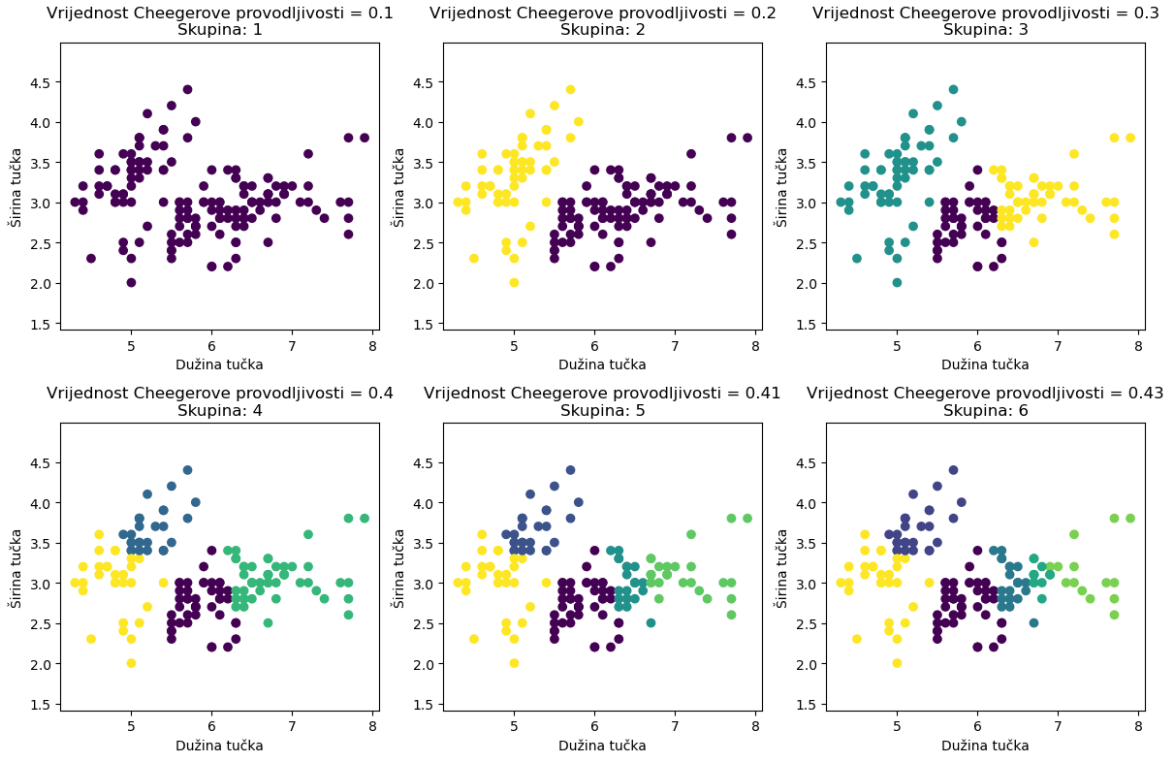


Slika 5.7 Utjecaj broja skupina na grupiranje podataka kod NJW metode

Kod KVV metode slična je situacija kao i kod metode normaliziranog reza, parametar koji se mijenja je prag za Cheegerovu vodljivost. Na sljedećim primjerima na slici (Slika 5.8) prikazana je ovisnost broja skupina o pragu za Cheegerovu vodljivost za dvije različite metode normaliziranja Laplaceove podmatrice u rekurzivnim pozivima, metodu normaliziranja množenjem i metodu normaliziranja zbrajanjem.



(uz korištenje metode normaliziranja množenjem)



(uz korištenje metode normaliziranja zbrajanjem)

Slika 5.8 Utjecaj praga za Cheegerovu vodljivost na grupiranje podataka za različite metode normalizacija podmatrica

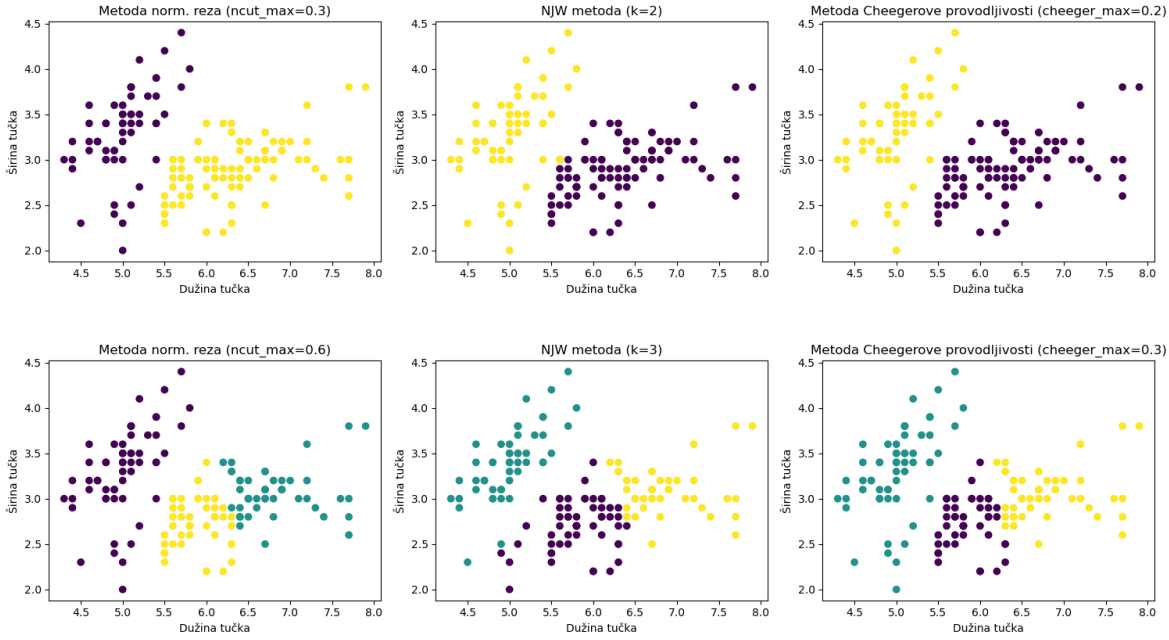
5.2. Mjere sličnosti

Kako bismo mogli usporediti tri implementirane metode spektralnog grupiranja potrebno je koristiti se mjerama sličnosti između različitih metoda grupiranja.

Prva mjera sličnosti koju ćemo iskoristiti je ARI koja mjeri sličnost između dobivenih grupiranja. Ona uspoređuje nalazi li se isti par podataka u istoj ili različitim skupinama u oba grupiranja kod dvije korištene metode. Niže vrijednosti označuju malu sličnost kod grupiranja podataka, dok veći podatci označuju vrlo sličnu podjelu. Očekivane vrijednosti su u rasponu $[-1,1]$ [31]. U implementaciji korištena je gotova funkcija *.adjusted_rand_score()* iz Python biblioteke *sklearn.metrics* [28].

Kratice „SM“, „NJW“ i „KVV“ koje se koriste u sljedećim primjerima odnose se na tri metode spektralnog grupiranja, metodu normaliziranog reza, NJW metodu i KVV metodu.

Na sljedećem primjeru na slici (Slika 5.9) usporedili smo tri metode grupiranja podataka u 2 i 3 skupine.



Slika 5.9 Usporedba rezultata tri spektralna algoritma grupiranja podataka u dvije i tri skupine

Uspoređivali smo za dvije skupine svaku metodu sa svakom i dobili smo ove rezultate:

Tablica 1 Rezultati metrike ARI za 2 skupine

NJW vs KVV	0.9469
NJW vs KVV	0.9469
KVV vs SM	1

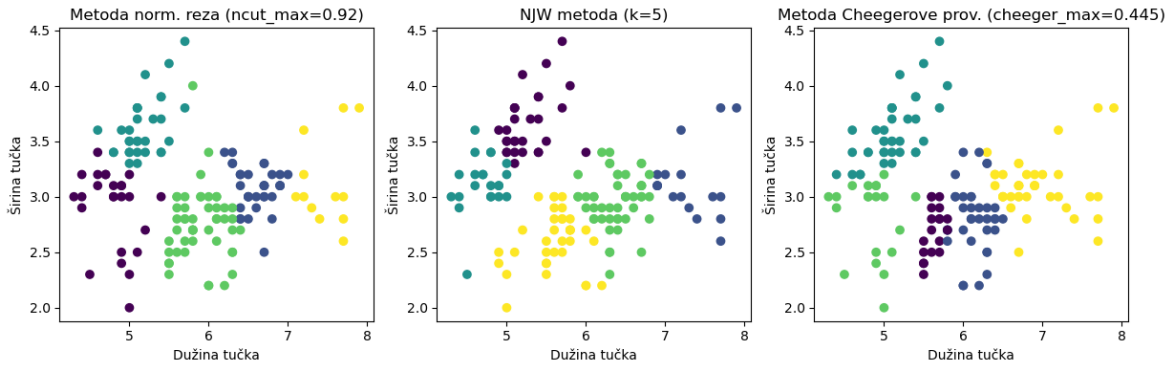
Kao i što vidimo iz grafova, metode KVV i SM imaju identične podjele te zato imaju i najviši mogući ARI rezultat, ostale kombinacije metoda također imaju visoki rezultat zbog velike sličnosti grupiranja. Ni na primjeru (Tablica 2) s tri skupine rezultati ARI metrike se značajno ne razlikuju zbog malih odudaranja u skupinama različitih metoda:

Tablica 2 tablica rezultata metrike ARI za 3 skupine

NJW vs KVV	0.8083
NJW vs SM	0.8433
KVV vs SM	0.9633

Druga mjera sličnosti koju ćemo iskoristiti je NMI metrika. Temelji na teoriji informacije i mjeri koliku količinu informacije dijele uspoređena grupiranja. Vrijednosti koje može poprimiti su u rasponu [0,1], te veće vrijednosti označavaju bolje grupiranje podataka [32]. U implementaciji korištena je gotova funkcija *.normalized_mutual_info_score()* iz Python biblioteke *sklearn.metrics* [28].

Budući da su rezultati za dvije i tri skupine metrike NMI vrlo slični metrici ARI, sljedeći primjer napravit ćemo na 5 skupina i usporediti s rezultatima ARI metrike, prikazano na slici (Slika 5.10).



Slika 5.10 Rezultati grupiranja tri spektralne metode za pet skupina

Na prvi pogled možemo pretpostaviti da će rezultati metrika za parove metoda biti drugačiji u usporedbi s prethodnim primjerom (Tablica 1) (Tablica 2). Dobiveni rezultati metrika su:

Tablica 3 tablica rezultata metrika za 5 skupine

NJW vs KVV	0.4134	NJW vs KVV	0.6004
NJW vs SM	0.5491	NJW vs SM	0.6788
KVV vs SM	0.5850	KVV vs SM	0.7257
	(ARI metrika)		(NMI metrika)

Rezultati metrika (Tablica 3) se znatno razlikuju od rezultata za 2 i 3 skupine. Niže vrijednosti pokazuju da se skupine ne poklapaju u velikoj mjeri kao prije, no i dalje skupine su kreirane na relativno istim područjima.

Konačno, odabir metode ima velik utjecaj na krajnji rezultat, a za procjenu rezultata i usporedbu potrebno je koristiti se raznim metrika kako bismo donijeli dobru odluku.

6. Primjena

Spektralno grupiranje ima široku primjenu zbog svoje sposobnosti otkrivanja skrivenih veza među podatcima. U poglavlju proučit ćemo spektralno grupiranje kod segmentacije slika gdje nam je cilj izdvojiti područja na temelju sličnosti između piksela te segmentaciju MINST skupa podataka gdje se pokazuje kako spektralne metode grupiraju podatke u višedimenzionalnom prostoru.

6.1. Segmentacija slika

Budući da podatci dobiveni iz slika nisu dvodimenzionalni, morat ćemo prilagoditi postupak kreiranja matrice sličnosti kako bismo obuhvatili sve bitne informacije potrebne za segmentaciju slika. Sličnost ćemo definirati kombiniranjem prostorne udaljenosti između piksela i njihove razlike u intenzitetu boje kao što je to i prikazano u literaturi [22].

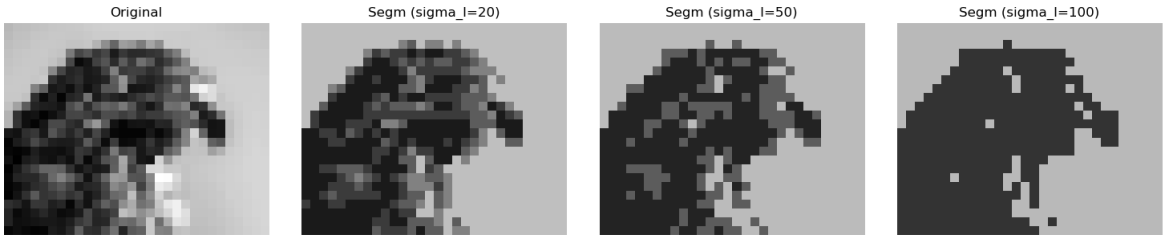
Za izračunavanje sličnosti između podataka koristit ćemo formulu koja se temelji na kombinaciji dvije Gaussove jezgrene funkcije. prvom se obuhvaća intenzitet svjetline piksela dobivenih iz slike označenih s F_i i F_j, a drugom prostorna lokacija piksela označenih s $\boldsymbol{x}_i$ i $\boldsymbol{x}_j$:

$$A_{ij} = e^{\frac{-\left\|F_i - F_j\right\|_2^2}{\sigma_I^2}} * \begin{cases} e^{\frac{-\left\|x_i - x_j\right\|_2^2}{\sigma_X^2}}, & ako \left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|_2 < r \\ 0, & ostalo \end{cases} \tag{26}$$

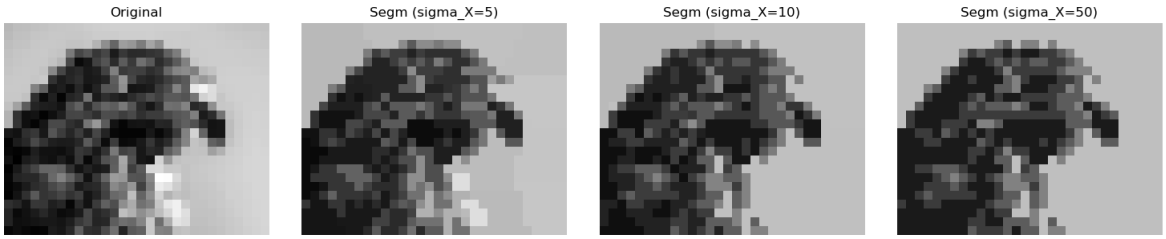
Dodatni parametri potrebni za ovu formulu su σ_I koji označava osjetljivost sličnosti na razlike u intenzitetu svjetline, σ_X koji određuje utjecaj prostorne udaljenosti, te r koji nam predstavlja prag za prostornu udaljenost unutar koje prihvaćamo dobivenu sličnost.

Za sljedeće primjere implementirani su prethodno opisani algoritmi grupiranja te primijenjeni na slike dimenzija 30x24 piksela (motiv orla) i 40x27 piksela (motiv oka). Dodatno, vrijednosti intenziteta svjetline skalirani su u raspon [0,255], a s time korištene su i veće vrijednosti parametara σ_I i σ_X.

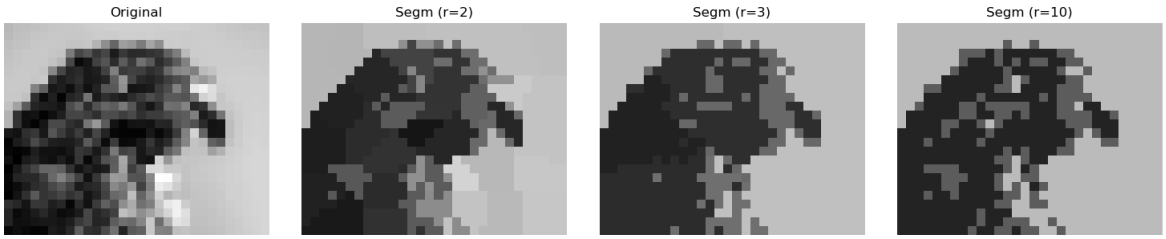
Na primjerima prikazanim na slici (Slika 6.1) vidimo utjecaj navedenih parametara na rezultate segmentiranja korištenjem metode normaliziranog reza. Temeljne vrijednosti su: $\sigma_I = 50, \sigma_X = 100, r = 10, lancoz\_k = 10, l = 10, n\_cut = 0.7$, a za svaki primjer mijenja se jedan od prva tri parametra.



(ovisnost o parametru σ_I)



(ovisnost o parametru σ_X)

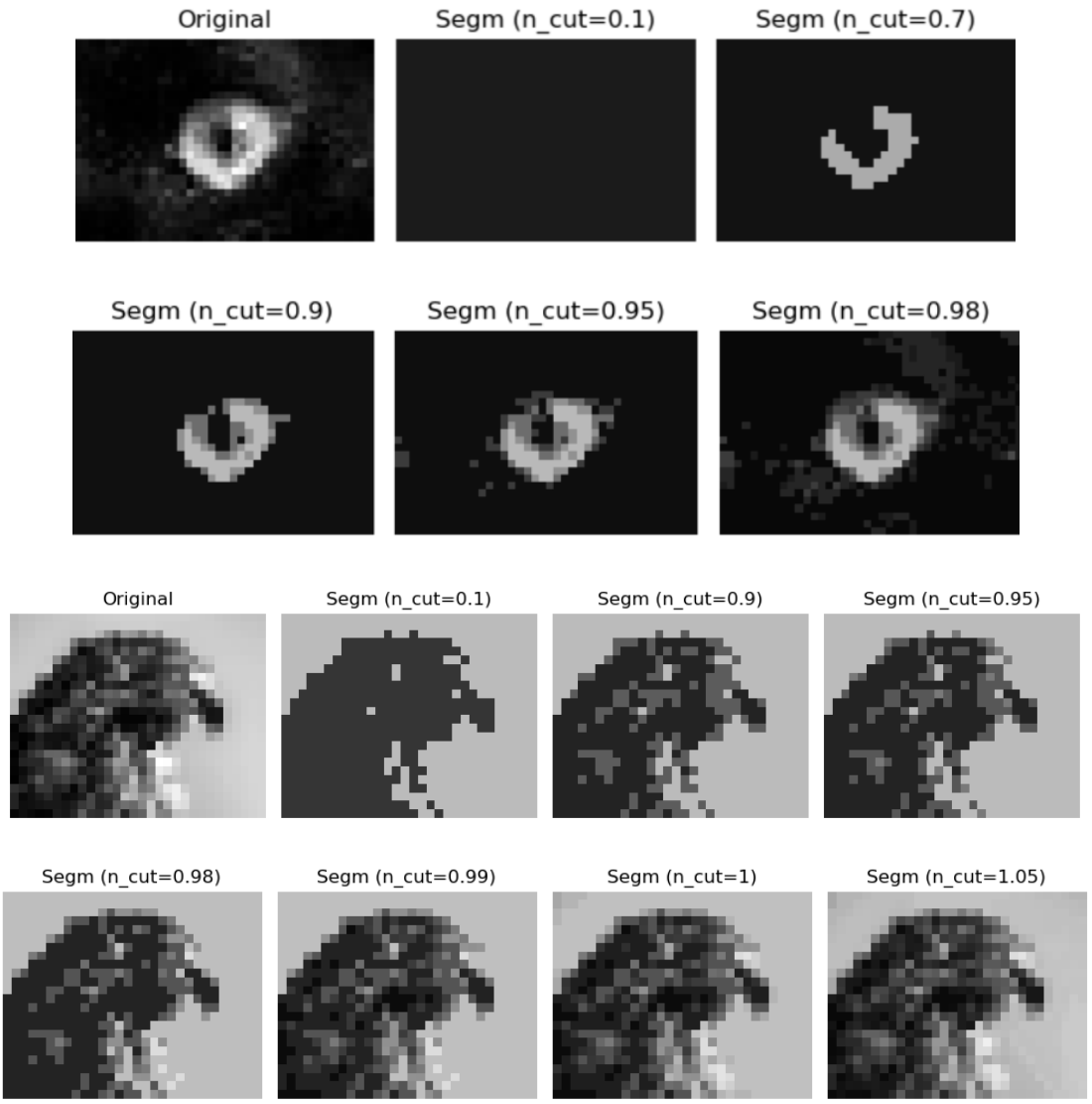


(ovisnost o parametru r)

Slika 6.1 Utjecaj ulaznih parametara na rezultate segmentacija slika

Na sljedećem primjeru na slici (Slika 6.2) prikazano je kako povećanjem vrijednosti maksimalno dopuštene vrijednosti normaliziranog reza dolazi do povećanja broja dobivenih skupina. Segmentacija je finija, ima više manjih grupa i više detalja je uočljivije na slikama.

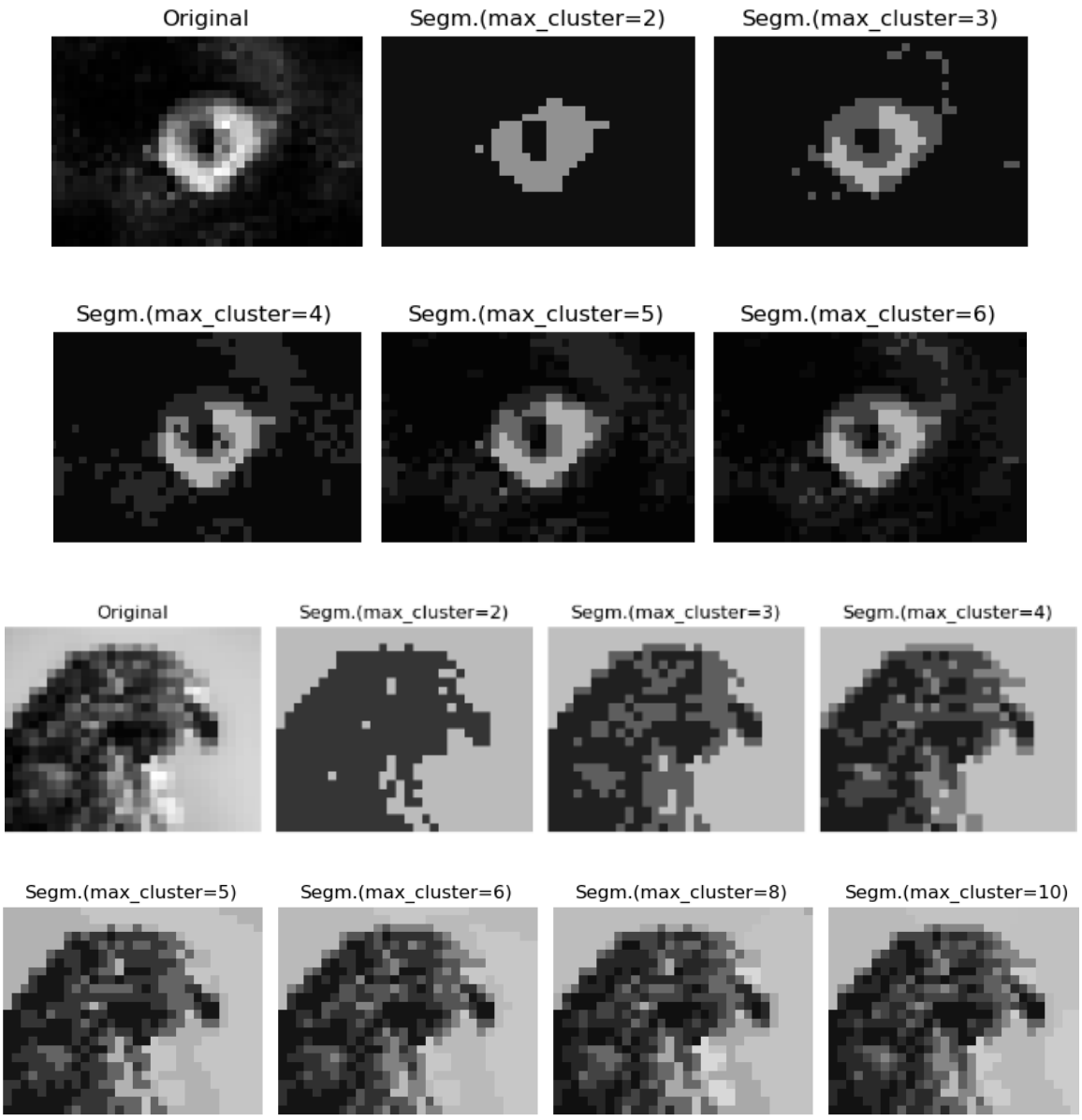
Ostale vrijednosti su: σ_I = 50, σ_X =100, r=10, lancoz_k=10, l=10.



Slika 6.2 Segmentacija slika pomoću metode normaliziranog reza

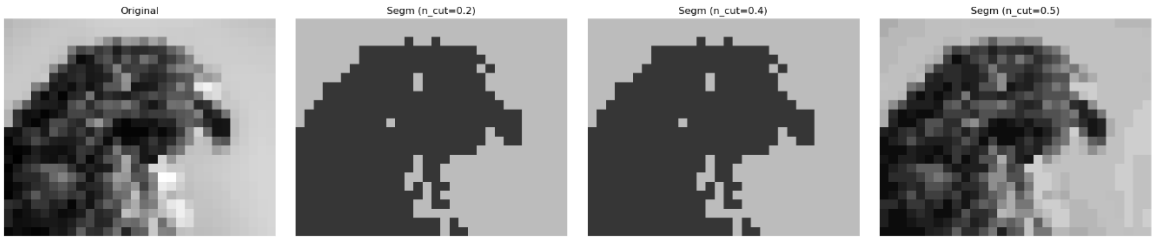
I za sljedeće primjere NJW metode korištene su prethodne slike te je i matrica sličnosti izrađena na isti način i temelji se na prethodnim podatcima.

Na sljedećim primjerima na slici (Slika 6.3) prikazani su rezultati u ovisnosti o zadanom brojem skupina. Sa manjim zadanim brojem skupina ističe se obris originalne slike, dok sa većim brojem skupina dobivamo više fragmentirane slike s više detalja.

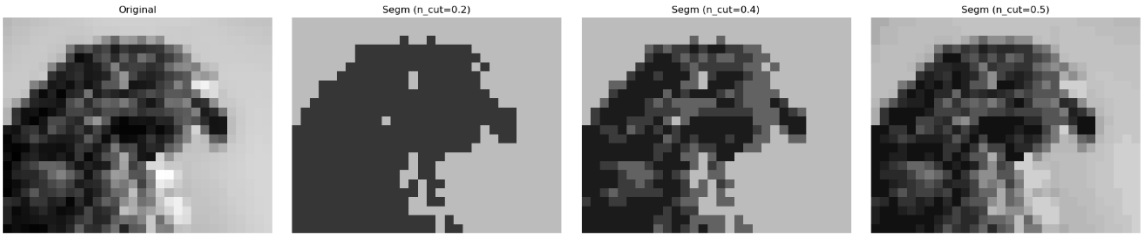


Slika 6.3 Segmentacija slika pomoću NJW metode

Na istim slikama prikazat ćemo rezultate segmentiranja slika koristeći KVV metodu (Slika 6.4). Ulazni parametri isti su kao i kod primjera s metodom normaliziranog reza osim pragova vrijednosti Cheegerove vodljivosti koje mijenjamo. Prvi redak slika dobiven je korištenjem metode normalizacija Laplaceove podmatrice množenjem, a drugi redak metodom normalizacija Laplaceove podmatrice zbrajanjem.



(metoda normalizacija Laplaceove podmatrice množenjem)



(metoda normalizacija Laplaceove podmatrice zbrajanjem)

Slika 6.4 Segmentacija slika pomoću metode KVV

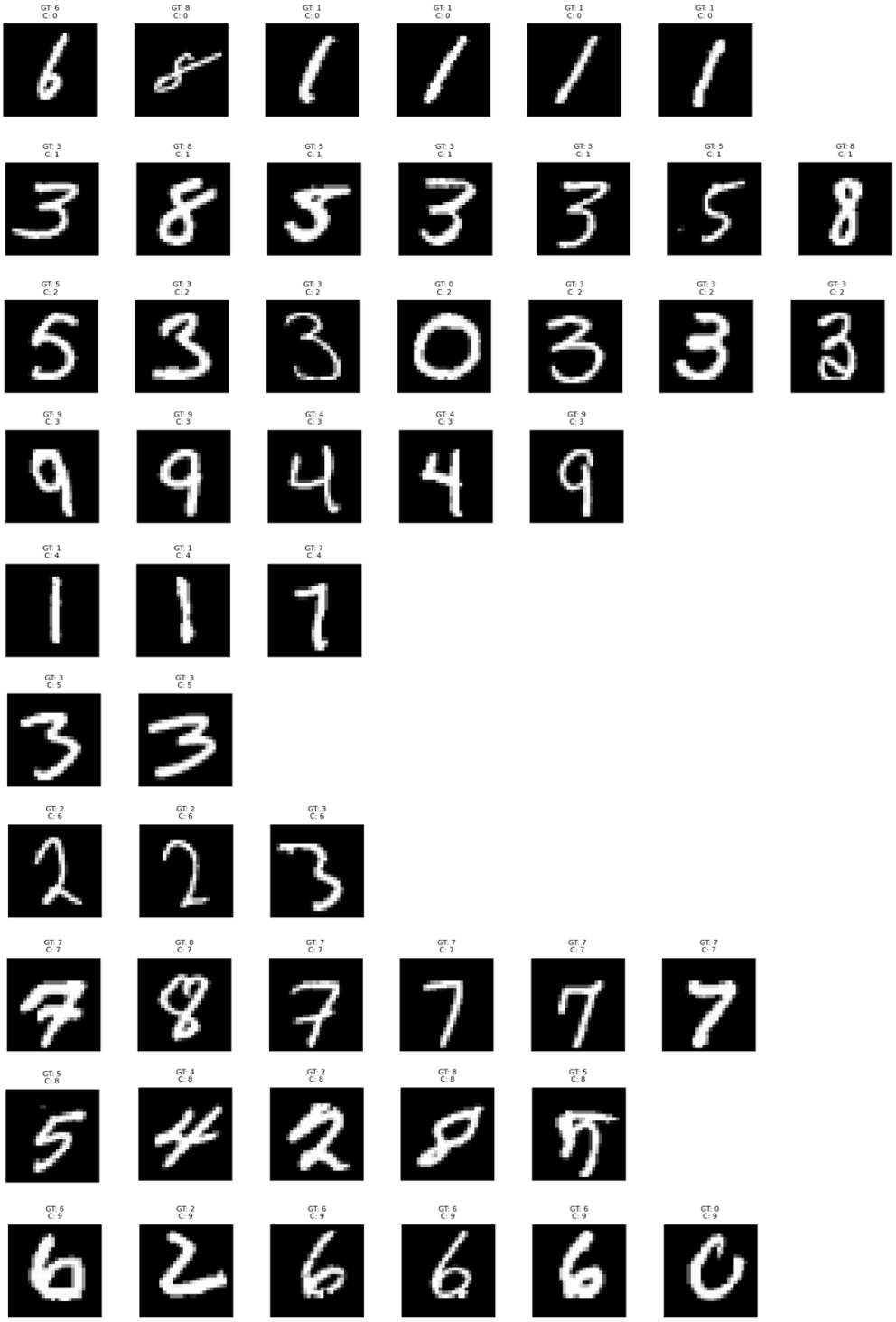
6.2. Grupiranje višedimenzionalnih podataka

Kako bismo prikazali spektralnu metodu grupiranja na višedimenzionalnim podatcima koristit ćemo MINST skup podataka koji sadrži slike ručno pisanih znamenki dimenzija 28x28 piksela [33].

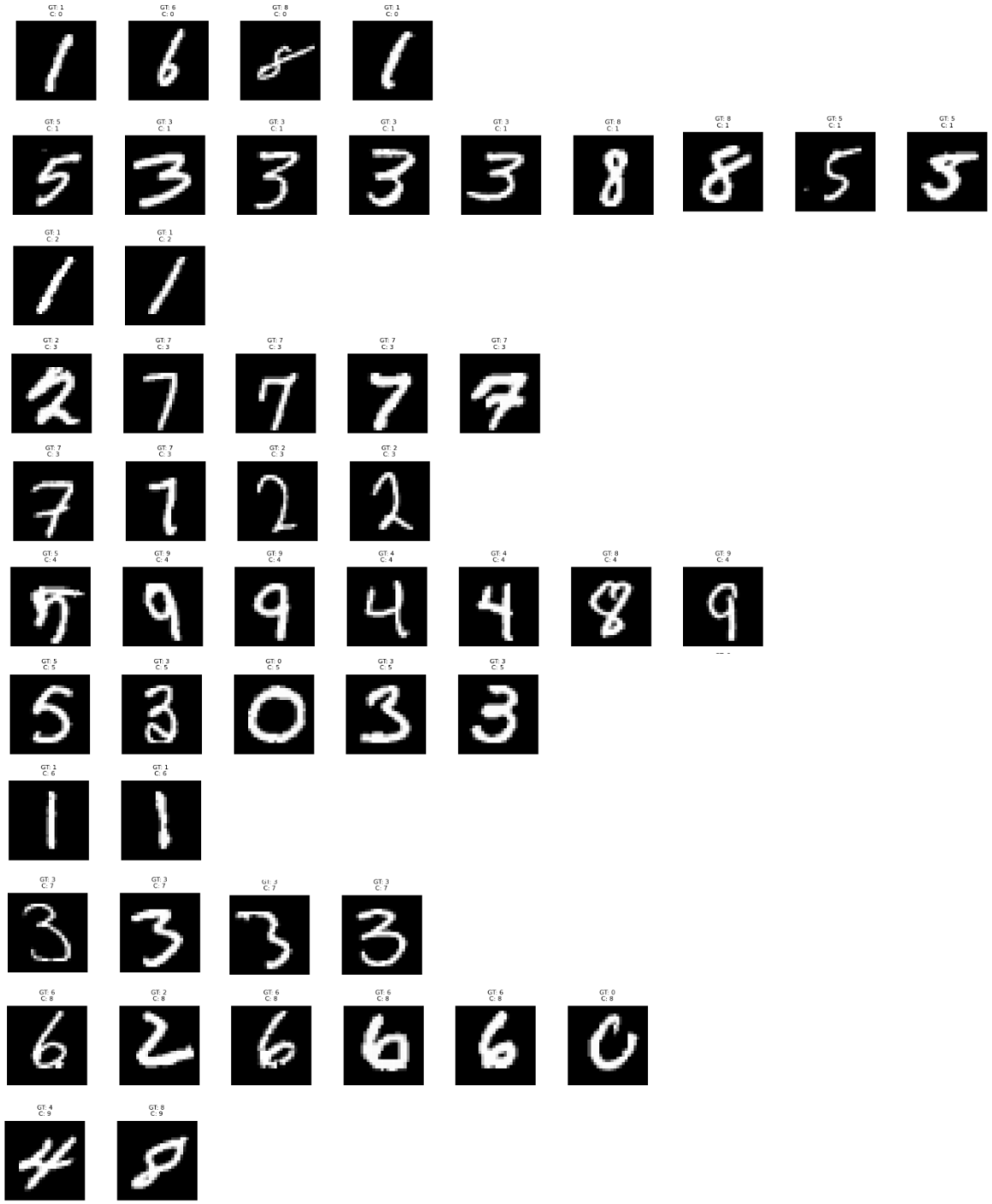
Za grupiranje znamenki koristit ćemo spektralnu metodu NJW na podskupu 50 nasumično odabranih slika znamenki. Svaka slika pretvara se u vektor od 784 elemenata koje tretiramo kao višedimenzionalne točke.

S obzirom na poteškoće pri određivanju parametara za postizanje optimalnog grupiranja, na sljedećim slikama (Slika 6.5), (Slika 6.6 Grupiranje na MINST skupu podataka u 10 skupina koristeći NJW metodu), prikazani su rezultati dva grupiranja podataka u deset skupina koristeći spektralne metode NJW i normaliziranog reza ($ncut = 1.11$) uz kosinusnu mjeru sličnosti (2). Svaki redak slika predstavlja jednu grupiranu skupinu. Budući da su slike grupirane na temelju njihovih vektorskih reprezentacija, metode grupiraju slike sa sličnim vizualnim karakteristikama, a ne isključivo prema semantičkom sadržaju. Takav pristup dovodi nam do rezultata u kojem u istoj grupi imamo više znamenki koje dijele sličan oblik. Na oba primjera možemo primijeniti da su strukture grupa slične s određenim varijacijama, to jest određene podgrupe znamenki se ponavljaju.

Nijedno od prikazanih grupiranja ne možemo jednoznačno odrediti kao ispravno ili pogrešno jer nam različite metode nude drugačije, ali jednako ispravne raspodjele podataka.

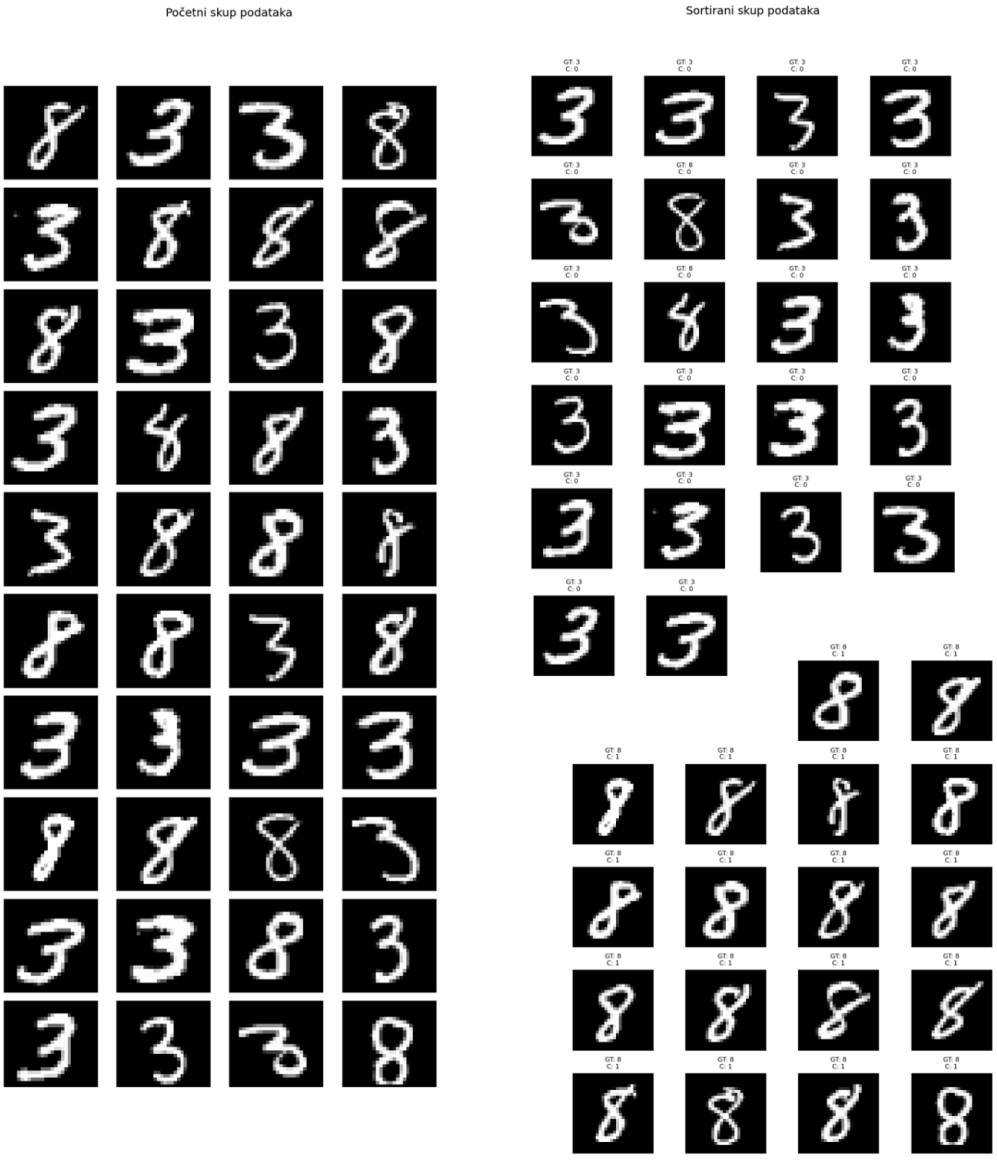


Slika 6.5 Grupiranje na MINST skupu podataka u 10 skupina koristeći metodu normaliziranog reza



Slika 6.6 Grupiranje na MINST skupu podataka u 10 skupina koristeći NJW metodu

Dodatno, prikazan je i primjer grupiranja znamenki 3 i 8 u dvije skupine (Slika 6.7). Za grupiranje korištena je NJW metoda uz matricu sličnosti konstruiranu pomoću Gaussove jezgrene funkcije (3) uz σ_X =1. Grupiranje je u ovom slučaju bilo uspješnije, budući da su samo dva uzorka pogrešno grupirana s obzirom na njihovo semantičko značenje. Ti pogrešno grupirani primjeri sadrže obilježja obje klase te zaključujemo da kvalitetu grupiranja značajno određuje i kvaliteta samih podataka. Budući da se ovdje radi primjerima ručno pisanim znamenki, postoje odstupanja u primjercima iste klase što dodatno otežava odvajanje skupina.



Slika 6.7 Grupiranje MINST znamenki 3 i 8 NJW metoda u dvije skupine

Zaključak

U radu su predstavljene i implementirane različite metode grupiranja s posebnim fokusom na metode spektralnog grupiranja. U raznim primjerima pokazalo se da su spektralne metode grupiranja uspješnije u prepoznavanju kompleksnih struktura u podatcima u odnosu na metodu *k*-sredina.

Također, pokazano je da velik utjecaj na konačni rezultat grupiranja čini odabir prigodnih parametara potrebnih za izračun matrica sličnosti i kriterija za podjelu skupina. Ove metode primijenjene su na primjerima segmentacije slika i grupiranja višediomenzionalnih podataka.

S obzirom na veliku osjetljivost na ulazne parametre i računalnu složenost, za provođenje optimalnog spektralnog grupiranja potrebno je znanje o specifičnostima pojedinog problema i značajkama ulaznih podataka.

Literatura

[1] P. Favati, G. Lotti, O. Menchi i F. Romani, »Construction of the similarity matrix for the,« *Journal of Computational and Applied Mathematics; Volume 375,* 2020.

[2] D. A. Spielman i S.-H. Teng, »Spectral partitioning works: Planar graphs and finite element meshes,« *Linear Algebra and its Applications;Volume 421, Issues 2–3,* pp. 284-305, Ožujak 2007.

[3] B. Slininger, »Fiedler's Theory of Spectral Graph Partitioning,« Department of Computer Science, University of California, Davis.

[4] C. M. Bishop, Pattern Recognition and Machine Learning, New York: Springer Science+Business Media, LLC, 2006.

[5] D. Arthur i S. Vassilvitskii, »k-means++: the advantages of careful seeding,« u *SODA '07: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithm*, 2006.

[6] E. Umargono, J. E. Suseno i V. G. S.K, »K-Means Clustering Optimization Using the ElbowMethod and Early Centroid Determination Based onMean and Median Formula,« *Advances in Social Science, Education and Humanities Research, volume 474,* 2019.

[7] J. Pavlopoulos, G. Vardakas i A. Likas, »Revisiting Silhouette Aggregation«.

[8] N. Elezović i A. A. Aglić, Linearna Algebra, Zagreb: Element, 2016..

[9] O. Knill, »Linear Algebra And Vector Analysis, Math 22b, Harvard College,« 2019.

[10] H. Li, »Properties and Applications Of Graph Laplacians,« The University of Chicago Department of Mathematics, 2022.

[11] A. Mardsen, »Eigenvalues Of The Laplacian And Their Relationship To The Connedtedness Of A Graph,« The University of Chicago Department of Mathematics, 2013.

[12] U. v. Luxburg, »A Tutorial on Spectral Clustering,« *n Statistics and Computing, 17 (4),,* 2007.

[13] S. C. Chapra i R. P. Canale, »Numerical Methods for Engineers, seventh edition,« Published by McGraw-Hill Education, New York, 2015.

[14] J. Francis, »The QR Transformation - Part 2,« *The Computer Journal, Volume 4, Issue 4,,* pp. 332-345, 1962.

[15] L. N. I. Trefethen i D. Bau, Numerical Linear Algebra, Philadelphia: Society for Industrial and Applied Mathematics, 1997.

[16] D. Jandrić, »Diplomski rad: Lanczosova i Arnoldijeva metoda,« Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet, Matematički odsjek, 2022.

[17] »7 Iterative Algorithms for Eigenvalue Problems,« Statistics, Univesity of Chicago Department of Statistics.

[18] Q. Yuan, »A Tour of the Lanczos Algorithm and its ConvergenceGuarantees throught the Decades,« Department of Mathematics, 2018. [Mrežno]. Available: https://math.berkeley.edu/. [Pokušaj pristupa 1. lipanj 2025.].

[19] G. H. Golub i C. F. V. Loan, Matrix Computations, Baltimore: The Johns Hopkins University Press, 2013.

[20] J. Wyss-Gallifent, »Graph Theory,« u *Applications of Linear Algebra Textbook*.

[21] »Linear Algebra for Data Science with examples in R: Chapter 22 Algorithms for Graph Partitioning,« [Mrežno]. Available: https://shainarace.github.io/LinearAlgebra/chap1-5.html. [Pokušaj pristupa 10. lipanj 2025.].

[22] J. Shi i J. Malik, »Normalized Cuts and Image Segmentation,« *IEEE Transactions On Patterm Analysis and Machine Learning, vol. 22, no. 8,* kolovoz 2000.

[23] G. Chen, »Math 253: Mathematical Methods for Data Visualization: Lecture 4: Rayleigh Quotients,« San José State University, [Mrežno]. Available: https://www.sjsu.edu/faculty/guangliang.chen/Math253S20/lec4RayleighQuotient.pdf. [Pokušaj pristupa 25. svibanj 2025.].

[24] T. Wu, A. R. Benson i D. F. Gleich, »General Tensor Spectral Co-clustering for Higher-Order Data«.

[25] D. Verma i M. Meila, »A Comparison of Spectral Clustering Algorithms«.

[26] R. Kannan, S. Vempala i A. Vetta, »On Clusterings: Good, Bad and Spectral,« M.I.T., Cambridge, Massachusetts, 2004.

[27] A. Y. Ng, M. I. Jordan i Y. Weiss, »On Spectral Clustering: Analysis and an Algorithm,« 2001.

[28] »Scikit-learn,« [Mrežno]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_circles.html#sklearn.datasets.make_circles. [Pokušaj pristupa 30. svibanj 2025.].

[29] R. Mondal, E. Ignatova, D. Walke, D. Broneske, G. Saake i R. Heyer, »Clustering graph data: the roadmap to spectral techniques, vol.4, članak br.4,« *Discover Artificial Intelligence,* 2022.

[30] »The Iris Dataset,« [Mrežno]. Available: https://scikit-learn.org/1.4/auto_examples/datasets/plot_iris_dataset.html. [Pokušaj pristupa 5. lipanj 2025.].

[31] J. M. Santos i M. Embrechts, »On the Use of the Adjusted Rand Index as a Metric for Evaluating Supervised Classification,« u *Conference: Artificial Neural Networks - ICANN 2009*, Limassol, 2009.

[32] A. Amelio i C. Pizzuti, »Correction for Closeness: Adjusting Normalized Mutual Information Measure for Clustering Comparison: Correction For Closeness: Adjusting NMI,« *Computational Intelligence,* 2016.

[33] »Minst dataset,« Tensorflow. [Mrežno]. [Pokušaj pristupa 20. lipanj 2025].

Sažetak

Naslov: Spektralno grupiranje podataka

Sažetak: U ovom radu objašnjenje su odabrane metode grupiranja s naglaskom na spektralno grupiranje.

Uz algoritam k-sredina kao particijske metode grupiranja, pojašnjene su i spektralna metoda grupiranja koristeći Fiedlerov vektor Laplaceove matrice i tri alternativne metode, metoda normaliziranog reza, NJW metoda i KVV metoda.

U drugom dijelu rada, implementirani su i analizirani navedeni algoritmi u ovisnosti o ulaznim parametrima i prema mjerama sličnosti. Primjene spektralnog grupiranja prikazane su na primjerima segmentacije slika i višedimenzionalnih podataka iz MINST skupa podataka.

Dobiveni rezultati prikazuju prednosti spektralnog grupiranja u prepoznavanju kompleksnih struktura unutar podataka koje klasične metode grupiranja ne uspijevaju grupirati.

Ključne riječi: grupiranje podataka, algoritam k-sredina, spektralno grupiranje podataka, Laplaceova matrica, Fiedlerov vektor, segmentacija slika, grupiranje višedimenzionalnih podataka

Summary

Title: Spectral Data Clustering

Summary: This paper explains selected clustering methods with an emphasis on spectral clustering. Alongside the k-means algorithm as a partitioning clustering method, spectral clustering methods using the Fiedler vector of the Laplacian matrix and three alternative methods are presented: the normalized cut method, the NJW method, and the KVV method.

In the second part of the paper, the mentioned algorithms are implemented and analyzed depending on input parameters and similarity measures. Applications of spectral clustering are demonstrated on examples of image segmentation and high-dimensional data from the MNIST dataset.

The obtained results show the advantages of spectral clustering in recognizing complex structures within data that classical clustering methods fail to group effectively.

Keywords: data clustering, k-means algorithm, spectral clustering, Laplacian matrix, Fiedler vector, image segmentation, high-dimensional data clustering

Privitak

U sklopu diplomskog rada izrađena je Jupyter bilježnica koja sadrži sve implementirane metode opisane i radu, kao i sve pripadne primjere. Primjeri su organizirani u šest poglavlja koja prate strukturu poglavlja ovog rada.

Jupyter bilježnica nalazi se na sljedećoj poveznici: https://github.com/antoniamstr1/diplomski_rad . Za njeno pokretanje potrebno je imati instaliran programski jezik Python i odgovarajuću podršku za rad s Jupyter bilježnicama.