UNIVERZA V MARIBORU

FAKULTETA ZA ELEKTROTEHNIKO,
RAČUNALNIŠTVO IN INFORMATIKO

Matej Brumen

ORODJE ZA VIZUALNO ANALITIKO VEČDIMENZIONALNIH PODATKOV

Magistrsko delo

Maribor, oktober 2017

ORODJE ZA VIZUALNO ANALITIKO VEČDIMENZIONALNIH PODATKOV

Magistrsko delo

Študent:	Matej Brumen
Študijski program:	Računalništvo in informacijske tehnologije (MAG)
Mentor:	doc. dr. Domen Mongus, univ. dipl. inž. rač. in inf.
Somentor:	asist. dr. Niko Lukač, univ. dipl. inž. rač. in inf.

Univerza v Mariboru

Fakulteta za elektrotehniko,
računalništvo in informatiko
Smetanova ulica 17
2000 Maribor, Slovenija

Številka: E5017726
Datum in kraj: 16. 05. 2017, Maribor

Na osnovi 330. člena Statuta Univerze v Mariboru (Statut UM – UPB 11, Ur. l. RS, št. 44/2015) izdajam

SKLEP O ZAKLJUČNEM DELU

1. **Mateju Brumnu**, študentu študijskega programa 2. stopnje MAG RAČUNALNIŠTVO IN INFORMACIJSKE TEHNOLOGIJE, se dovoljuje izdelati zaključno delo.

2. Tema zaključnega dela je pretežno s področja Inštituta za računalništvo.

MENTOR: doc. dr. Domen Mongus
SOMENTOR: asist. dr. Niko Lukač

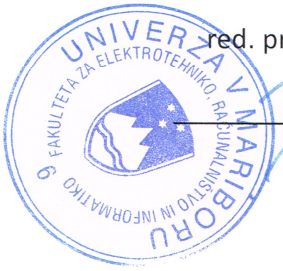
3. Naslov zaključnega dela:
ORODJE ZA VIZUALNO ANALITIKO VEČDIMENZIONALNIH PODATKOV

4. Naslov zaključnega dela v angleškem jeziku:
A TOOL FOR VISUAL ANALYTICS OF MULTIDIMENSIONAL DATA

5. Rok za izdelavo in oddajo zaključnega dela je 16. 05. 2018. Zaključno delo je potrebno izdelati skladno z "Navodili za izdelavo zaključnega dela" in ga v treh izvodih (dva trdo vezana izvoda in en v spiralo vezan izvod) oddati v pristojnem referatu članice. Hkrati se odda tudi izjava mentor-ja/-ice (in morebitnega somentor-ja/-ice) o ustreznosti zaključnega dela.

Pravni pouk: Zoper ta sklep je možna pritožba na Senat članice v roku 10 delovnih dni od dneva prejema sklepa.

Dekan:
red. prof. dr. Borut Žalik



Obvestiti:
- kandidata,
- mentor-ja/-ico,
- somentor-ja/-ico,
- odložiti v arhiv.

ZAHVALA

Zahvaljujem se mentorju doc. dr. Domnu Mongusu za pomoč in vodenje pri opravljanju magistrskega dela. Prav tako se zahvaljujem somentorju dr. Niku Lukaču za strokovne nasvete.

Posebna zahvala gre družini za podporo tekom študija.

Orodje za vizualno analitiko večdimenzionalnih podatkov

Ključne besede: vizualna analitika, masovni podatki, analiza odvisnosti, korelacija, odkrivanje znanja

UDK: 004.514.66:004.62(043.2)

Povzetek:

V magistrskem delu predstavimo orodje za vizualno analitiko večdimenzionalnih podatkov. V analizi sorodnega dela predstavimo tehnike vizualizacije masovnih podatkov in tradicionalne ter napredne tehnike odkrivanja znanja. Na tej osnovi podrobneje opišemo razvito orodje in s primeri uporabe demonstriramo njegovo učinkovitost. Z rezultati potrdimo pravilnost delovanja implementiranih funkcionalnosti.

A Tool for Visual Analytics of Multidimesional Data

Keywords: visual analytics, big data, correlation analysis, knowledge discovery

UDC: 004.514.66:004.62(043.2)

Abstract:

This Master's thesis introduces a tool for visual analytics of multidimensional data. During the analysis of the related work, techniques for big data visualization and traditional as well as advanced knowledge discovery methods are presented. On this basis, the developed tool is described in details, while its efficiency is demonstrated with several use-cases. The correctness of the implemented methods is proved with the results.

Kazalo

Poglavje 1

Uvod

Danes vse več procesov odločanja temelji na znanju, pridobljenem iz študij velikih množic multimodalnih podatkov. Takšne študije sestojijo iz zbiranja podatkov, statistične obdelave zbranih meritev ter testiranja postavljenih hipotez. Izmerjeni vzorci so pri tem definirani z množico spremenljivk, na njihovi osnovi pa običajno želimo predvidevati vrednosti ciljne spremenljivke. Slednje pripomore k izboljšanju procesov upravljanja in vodi v tako imenovano podatkovno podprto odločanje. Med tem ko so strukturirani vhodni podatki običajno podani v obliki tabel, so ciljne spremenljivke lahko podane kot diskretne (definirane z razredi) ali zvezne funkcije, ki opisujejo ključne parametre odločanja (na primer rast prihodkov podjetja, prioritete strank ali stroške proizvodnje). Natančne napovedi ciljne spremenljivke pa pri tem predstavljajo zgolj del informacijskih potreb. Pogosto je namreč enako pomembno tudi razumevanje vzročnoposledičnih odnosov znotraj vhodnih spremenljivk, ki omogoča razvoj ustreznih mehanizmov upravljanja. V ta namen pa je pomembna grafična predstava masovnih podatkov, ki uporabniku omogoča hiter vpogled v vzorce in njihovo obliko ter tako pripomore k odkrivanju novih znanj o obravnavanem problemu.

V tem magistrskem delu predstavljamo analitično orodje, ki omogoča analizo in vizualizacijo večdimenzionalnih masovnih podatkov. Aplikacija omogoča izračun več statističnih mer nad množico vhodnih podatkov ter uporabniku prijazen pregled rezultatov. Vsako spremenljivko v seznamu je mogoče tudi podrobneje vizualizirati v obliki grafa funkcije gostote verjetnosti in nad njo izvesti analizo soodvisnosti spremenljivk z grafom korelacij. Aplikacija prav

tako omogoča enostavno filtriranje vzorcev in izvedbo napovedovalne analize preko uporabniškega vmesnika.

V poglavju 2 tega magistrskega dela predstavimo nekatere sorodne rešitve. V poglavju 3 opišemo ključne komponente aplikacije, v poglavju 4 pa podamo podrobnejši opis implementiranih komponent. Validacija aplikacije je opisana v poglavju 5, poglavje 6 pa podaja zaključke tega magistrskega dela.

Poglavje 2

Sorodna orodja

Na trgu obstaja več komercialnih samostojnih aplikacij, odprtno-kodnih knjižnicah ter dodatkov za splošnonamenska orodja, ki omogočajo statistično obdelavo masovnih podatkov. Med najbolj popularna orodja za ta namen spadajo:

- **IBM SPSS Statistics**, ki je analitično orodje za obdelavo in vizualizacijo strukturiranih podatkov, ki nudi podporo številnim programskim jezikom, med drugim tudi programskim in skriptnim jezikom R, Java, Python, in C# [1].
- **Analyse-it**, ki je dodatek za orodje Microsoft Excel in ponuja množico statističnih metod, skupaj s podporo za različne analize porazdelitve podatkov ter poenostavitev dela [2].
- **Matlab z modulom Statistical toolbox**, ki vsebuje množico vgrajenih metod za statistično analizo podatkov in vizualizacijo [3].
- **Programski jezik R**, ki je programski jezik namenjen statistični analizi podatkov [4].
- **Weka**, ki je odprtokodno orodje izdelano v programskem jeziku Java in vsebuje številne algoritme podatkovnega rudarjenjenja, strojnega učenja in gručenja ter filtre namenjene predprocesiranju podatkov in izbiri smiselnih atributov. Služi lahko kot odlično izhodišče za raziskave in razvoj, saj podpira tudi platforme za vzdrževanje masivnih podatkov, kot je to na primer Hadoop [5].

Čeprav se našteta orodja v svojih specifikah bistveno razlikujejo, pa vsa našteta orodja vključujejo nekatere osnovne komponente za podatkovno analizo. Te podrobneje predstavimo v nadaljevanju.

2.1 Predstavitev podatkov

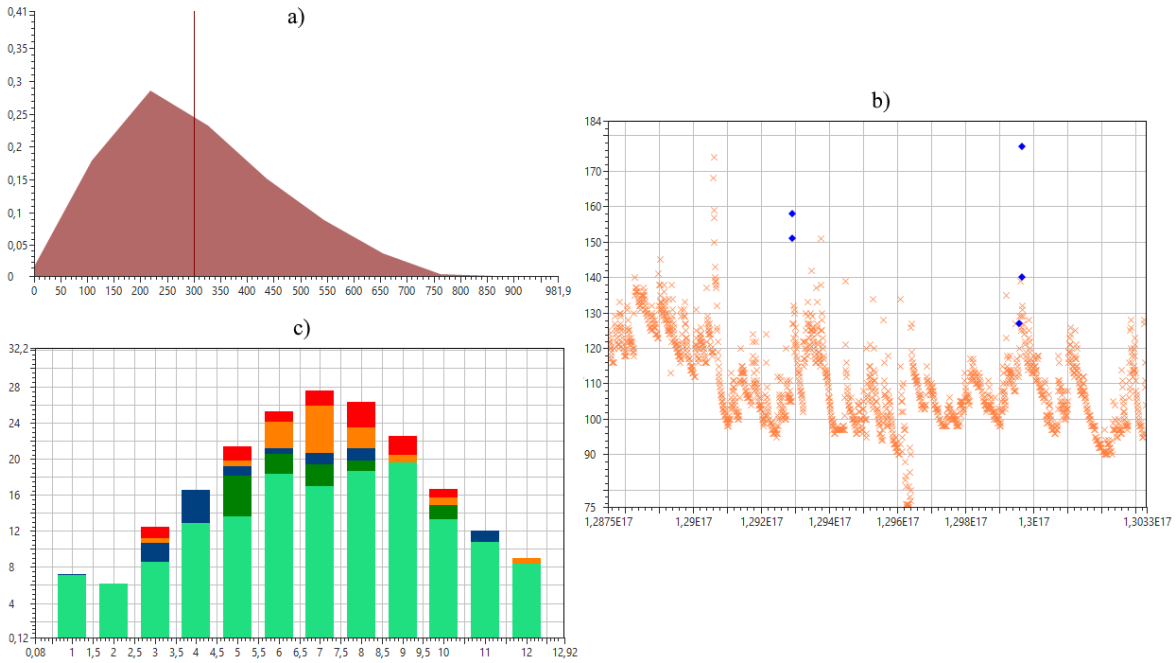
Zbirko podatkov si najlažje predstavljamo kot dvodimenzionalno matriko vrednosti (Tabela 2.1), pri čemer vrstice predstavljajo vzorce, stolpci pa njihove atribute oziroma spremenljivke. Število spremenljivk predstavlja dimenzionalnost prostora podatkov.

sepal length	sepal width	petal length	petal width	class
4.6	3.6	1.0	0.2	Iris-setosa
5.7	4.4	1.5	0.4	Iris-setosa
4.9	2.4	3.3	1.0	Iris-versicolor
5.0	2.3	3.3	1.0	Iris-versicolor
5.1	2.5	3.0	1.1	Iris-versicolor
5.0	2.0	3.5	1.0	Iris-versicolor
5.2	2.7	3.9	1.4	Iris-versicolor
5.6	2.9	3.6	1.3	Iris-versicolor
4.9	2.5	4.5	1.7	Iris-virginica
5.8	2.7	4.1	1.0	Iris-versicolor
6.2	2.2	4.5	1.5	Iris-versicolor

Tabela 2.1: Podatki podatkovne množice *Iris* s štirimi zveznimi spremenljivkami in tremi klasifikacijskimi razredi.

Takšne tabele pa so pogosto velike in, posledično, nepregledne. Danes poznamo več vizualizacijskih tehnik, ki so zmožne prikazati vsebino na kompaktnejši in preglednejši način. Nekatere izmed pogosteje uporabljenih so:

- **Funkcija gostote verjetnosti**, ki podaja relativno verjetnost, da bo slučajna spremenljivka zavzela določeno vrednost (Slika 2.1a),
- **graf raztrosa**, ki opisuje soodvisnost dveh spremenljivk z vizualizacijo podatkov v obliki dvodimenzionalnih točk (Slika 2.1b),
- **grafikoni**, ki opisujejo vrednosti diskretnih spremenljivk različnih kategorij (Slika 2.1c).



Slika 2.1: Prikaz a) funkcija gostote verjetnosti, b) graf raztrosa dveh spremenljivk in c) naložen stolpični grafikon.

Več izmed zgoraj naštetih osnovnih predstavitev obravnava podatke kot množice ali oblaka točk. Običajno pa so slednji definirani z mnogo več kot zgolj tremi spremenljivkami (dimenzijami), ki jih lahko predstavimo na človeku razumljiv način. Čeprav lahko kot dodatno dimenzijo uporabimo tudi barvo, tako problem zgolj omilimo, dejansko pa ga ne rešimo. Podatke je zato pogosto primerneje predhodno statistično obdelati in prikazati rezultate takšnih analiz. Primer osnovne analize predstavlja izračun korelacije, ki ovrednoti soodvisnost spremenljivk, med tem ko naprednejše analize izvajamo z metodami strojnega učenja [6]. Slednje delimo v tri skupine:

- **Nadzorovano strojno učenje**, kjer podatke razdelimo na učno in testno množico z znanimi vrednostmi ciljne spremenljivke za vsak vzorec. Metodo učimo nad učno množico podatkov, s testno množico pa preverjamo natančnost učenja. Med metode nadzorovanega strojnega učenja spadajo nevronske mreže, podporni vektorji, oziroma SVM (ang. Support Vector Machines) in odločitvena drevesa. V tem kontekstu so zanimiva predvsem slednja, saj jih lahko enostavno predstavimo in tako uporabniku odkrijemo potencialno novo znanje o vsebovanih vzorcih.

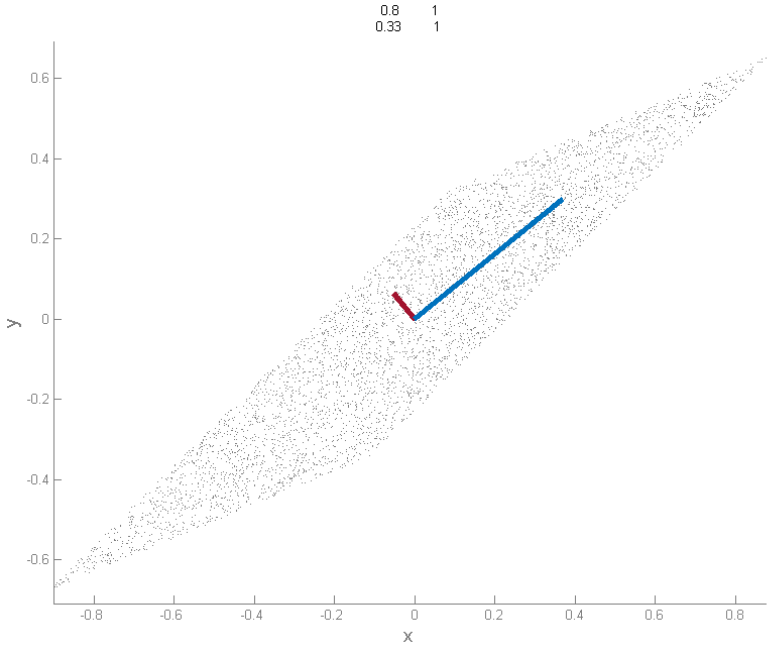
- **Nenadzorovano strojno učenje** uporabimo pri podatkih, kjer ne poznamo ciljne vre-

dnosti. V podatkih iščemo med seboj podobne vzorce in jih združujemo v večje množice. Primer nenadzorovanega strojnega učenja so algoritmi gručenja, ki nam omogočajo predstavitev skupin vzorcev s podobnimi karakteristikami.

- **Vzpodbujevalno strojno učenje**, kjer v metodo učenja vpeljemo koncept kaznovanja ali nagrajevanja (obteževanje določenih spremenljivk) na podlagi končnega izida. Izhodne uteži pa lažje predstavimo uporabnikom.

Opisani primeri uporabnikom podajajo karakteristike podatkov, namesto podatkov samih. Če to ni dovolj, in želimo uporabnikom prikazati dejanske podatke, moramo nad njimi izvesti redukcijo dimenzionalnosti. Vzorce v podatkih namreč pogosto sestavlja množica spremenljivk, ki tvorijo večdimenzionalen prostor. Hkratna obravnava vseh privede do težav povezanih s tako imenovano koncentracijo [7]. Slednja je določena kot razmerje med standardnim odklonom razdalj do referenčne točke in povprečjem teh razdalj. Višja koncentracija pomeni, da so vsi vzorci (oziroma točke v večdimenzionalnem prostoru) oddaljeni približno enako do vseh ostalih vzorcev. Izračun enostavnih statistik, kot so to na primer razdalje med točkami, njihova povprečna vrednost in standardni odkloni, zato niso več primerne metrike. Z analizo poglavitnih komponent, oziroma PCA (ang. Principle Component Analysis), pa lahko vzorce transformiramo v kompaktnejši prostor [8]. Pravimo, da so dimenzije tega prostora razvrščene glede na količino vsebovanih informacij, saj je razpršenost vzorcev v prvi dimenziji največja, v zadnji pa najmanjša (pogosto celo zanemarljiva). Vhod v PCA predstavljajo centrirani vzorci, kjer je od vsakega vzorca odšteta povprečna vrednost spremenljivke. Postopek PCA temelji na kovariančni analizi med podatkovnimi dimenzijami in izračunom lastnih vektorjev ter lastnih vrednosti nad kovariančno matriko.

Rezultat PCA sta matriki lastnih vektorjev in lastnih vrednosti. Lastni vektorji določajo smeri posameznih dimenzij glede na vhodni prostor, pri čemer lastne vrednosti predstavljajo dolžine vektorjev. Lastni vektor z največjo lastno vrednostnjo imenujemo poglavitna komponenta množice podatkov. Vsi lastni vektorji pa so pravokotni drug na drugega. Tako je transformacija podatkov neodvisna od položaja lastnih vektorjev v matriki (postopek jih samodejno uredi glede na lastne vrednosti od največje proti najmanjši). Rezultat je viden na sliki 2.2.



Slika 2.2: Prikaz naključnih vzorcev v 2D prostoru, kjer modra črta predstavlja poglavitni lastni vektor, medtem ko rdeča črta predstavlja pravokotni lastni vektor glede na poglavitnega.

Postopek PCA pa tako ustvari linearno transformiran prostor glede na osnovnega. Vzorce lahko med obema prostoroma enostavno preslikamo, če poznamo lastne vektorje, lastne vrednosti in povprečne vrednosti vzorcev v vsaki posamezni dimenziji. Takšna analiza nam torej definira dimenzije (spremenljivke) z največjo informacijsko vrednostjo, ki jih lahko uporabimo kot vhod v algoritme strojnega učenja. Pri tem pa lahko seveda ignoriramo lastne vektorje z nizko lastno vrednostjo, saj ne vsebujejo dovolj informacij. Te dimenzije so pogosto tudi zelo šumne. Preslikava v prostor PCA tako poleg same zmožnosti vizualizacije podatkov, pogosto pripomore tudi k dvigu učinkovitosti algoritmov strojnega učenja.

Med množico algoritmov strojnega učenja se v okviru tega magistrskega dela še posebej osredotočimo na metode hierarhičnega gručenja. Zato v nadaljevanju podajamo nekoliko podrobnejši opis upodobitve rezultatov takšnega pristopa.

2.2 Hierarhično gručenje

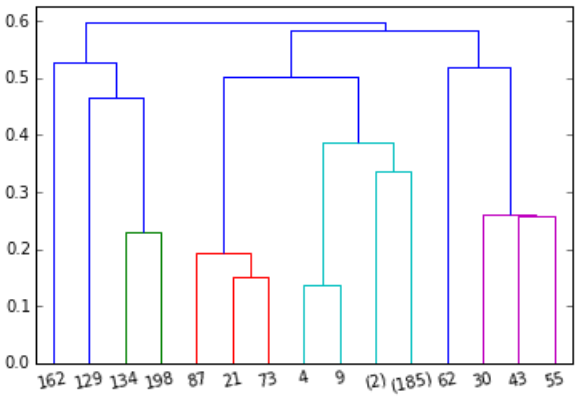
Gručenje spada med metode nenadzorovanega strojnega učenja, katerega cilj je združiti podobne vzorce skupine. Obstaja več različnih algoritmov gručenja. Med osnovne spadata K-Means in DBSCAN, danes pa poznamo tudi več tehnik naprednega hierarhičnega gručenja, ki izhajajo iz osnovnih tehnik [9].

Metode hierarhičnega gručenja omogočajo izgradnjo drevesne strukture gruč, pri čemer listi drevesa predstavljajo majhne množice vzorcev s podobnimi lastnostmi, združevanje slednjih v večje gruče pa je predstavljeno z notranjo strukturo vozlišč drevesa. Koren drevesa določa gručo, ki zajema vse podgruče in posledično vse vzorce. Obstajata dva načina gradnje hierarhije gruč:

- **Gradnja iz posameznih vzorcev (ang. bottom up)**, kjer na začetku vsak vzorec pripada svoji gruči. Algoritem gručenja na *i*-tem nivoju izvajamo nad gručami iz nivoja *i-1*. Nivo i_0 vsebuje toliko gruč, kot je vzorcev, ali številko gruč iz že obstoječega modela.
- **Gradnja iz ene gruče (ang. top down)**, kjer na začetnem nivoju vsi vzorci pripadajo eni gruči. Algoritem iterativno deli gruče na manjše, bolj podrobne gruče in se ustavi, ko struktura gruč zadosti vnaprej določenemu izstopnemu pogoju.

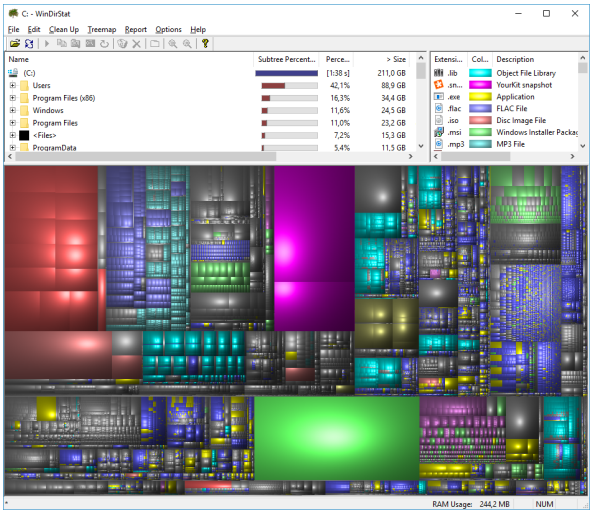
Ne glede na uporabljen pristop, danes poznamo več tehnik vizualizacije hierarhije gruč (oz. drevesnih struktur), ki jih lahko izvedemo v 2D ali 3D prostoru [10]. V naši rešitvi smo se omejili zgolj na 2D vizualizacijo predvsem zaradi težav, povezanih z navigacijo uporabnika v 3D prostoru. Tradicionalne tehnike, ki nam omogočajo tovrstne funkcionalnosti so:

- **Dendrogram**, ki omogoča tradicionalno vizualizacijo drevesnih struktur. Upodobitev se začne iz korena drevesa (zgoraj na sredi) in konča pri poravnanih listih (spodaj). Struktura drevesa med obema nivojema pa je dobro vidna, kot to prikazuje slika 2.3.



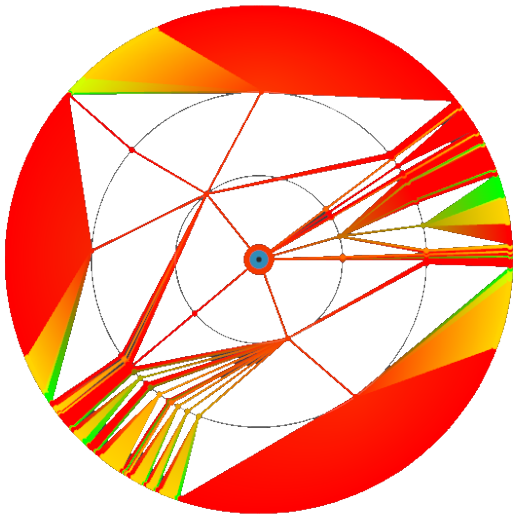
Slika 2.3: Primer klasičnega dvojiškega drevesa (dendogram).

- **Drevesno mapiranje**, ki omogoča predstavitev drevesnih struktur z delitvijo prostora na manjše podprostore znotraj večjih. Koren predstavlja celotno območje drevesne strukture. Gruče naslednjega nivoja so razporejene čez trenutnega, pri čemer je velikost njihovega območja proporcionalna glede na število vzorcev znotraj sinov. Primer vizualizacije prikazuje slika 2.4.



Slika 2.4: Primer drevesnega mapiranja s senčenjem. Vsak pravokotnik predstavlja ločeno gručo, njihova velikost pa je premosorazmerna s številom vsebovanih vzorcev.

- **Radialno drevo**, ki omogoča predstavitev gruč v polarnem koordinatnem sistemu. Koren se nahaja v središču kroga (koordinatnem izhodišču), nivoji drevesa pa so predstavljeni v koncentričnih krogih. Slednje omogoča razločno upodobitev strukture drevesa tudi kadar to vsebuje veliko količino vzorcev, kot je to prikazano na sliki 2.5.



Slika 2.5: Primer upodobitve hierarhije gruč z uporabo radialnega drevesa.

Poglavje 3

Metode in implementacija orodja

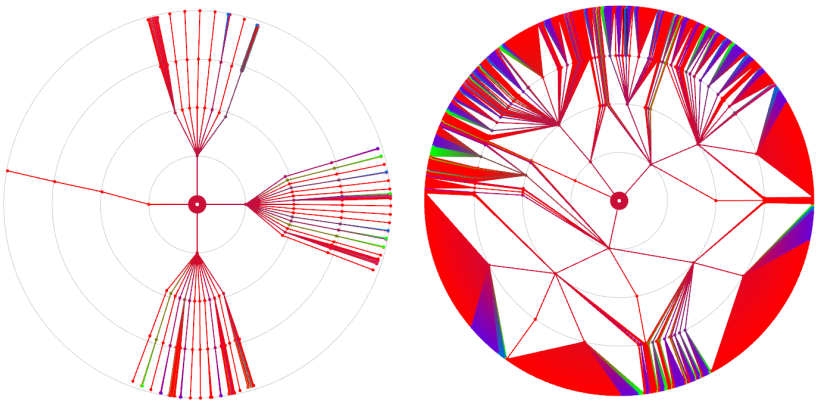
Na osnovi predstavljenih tehnik vizualizacije masivnih podatkov smo v okviru tega magistrskega dela razvili novo orodje za statistično analizo in odkrivanje novih znanj. V tem poglavju zato najprej predstavimo naš pristop k upodobitvi masivnih podatkov, čemur sledi opis analitičnih orodij, ki izhajajo iz tega. Orodje smo implementirali v programskem jeziku Java 1.8, s čimer smo omogočili njegovo delovanje na več platformah (Windows, Linux). Določene komponente za vizualizacijo uporabljajo programsko knjižnico OpenGL 3.0, ki je ovita v knjižnico JOGL (ang. Java OpenGL) [11]. Knjižnica JOGL omogoča popolni dostop do klicev programskega vmesnika OpenGL in je kompatibilna z različnimi ogrodji za gradnjo uporabniških vmesnikov. Med te spadajo AWT (ang. Abstract Widget Toolkit), Swing in SWT (ang. Standard Widget Toolkit). Slednji vključuje ogromno kvalitetnih visoko zmogljivih uporabniških komponent, ki ovijajo komponente operacijskega sistema (Win32, GTK+, Cocoa) [12]. Prav zaradi tega dejstva smo to ogrodje uporabili v okviru tega magistrskega dela.

3.1 Upodobitev masivnih podatkov

V predstavljeno orodje smo implementirali izgradnjo in vizualizacijo radialnega drevesa na osnovi hierarhičnega gručenja. V ta namen smo implementirali dva pristopa, prikazana na

sliki 3.1. To sta:

- **Gradnja iz korena gruč**, kjer drevo začnemo graditi iz središča gruč proti posameznim vzorcem. Koren umestimo v koordinatno izhodišče in nato razdelimo prostor na n enakih delov, kjer je vrednost n določena s številom podgruč. Težavi takšnega pristopa sta veliko nezapolnjenega prostora in nezmožnost intuitivne predstavitve velikosti posameznih gruč.
- **Gradnja iz posameznih vzorcev**, kjer drevo začnemo graditi iz posameznih vzorcev. Vzorce, ki pripadajo enakim gručam, skupaj umestimo na krožnico ločenimi s $\theta = \frac{n}{2\pi}$, pri čemer je n število vseh vzorcev. Vsaka gruča na višjem nivoju pa je predstavljena z vozliščem, ki se nahaja sredi površine njegovih sinov.



Slika 3.1: Vizualizacija hierarhije gruč, predstavljena z drevesom, zgrajenim iz korena (levo) in drevesom, ki je zgrajeno iz posameznih vzorcev (desno).

Ciljna spremenljivka v procesu podatkovne vizualizacije običajno določa klasifikacijske razrede oziroma ciljne ali napovedovalne vrednosti. Vrednosti te spremenljivke so lahko številske ali pa so podane v obliki opisnih lastnosti. Razločevanje vzorcev med razredi dosežemo tako, da jih ustrezno obarvamo (Slika 3.2). Podobno tehniko lahko uporabimo tudi pri vizualizaciji zveznih vrednosti, pri čemer pa vzorce najprej razdelimo v n razredov in tako diskretiziramo prostor. V naši rešitvi to izvedemo na dva načina:

1. **Delitev intervala na enake dele**, ki izvede delitev celotnega intervala vrednosti na n

Classes [class]	Count	Percentage%
Iris-setosa	50	33.33%
Iris-versicolor	50	33.33%
Iris-virginica	50	33.33%

#	sepal length	sepal width	petal length	petal width	TD: class
59	5.0	3.6	1.4	0.2	Iris-setosa
60	5.8	2.7	3.9	1.2	Iris-versicolor
61	5.8	2.6	4.0	1.2	Iris-versicolor
62	5.5	2.5	4.0	1.3	Iris-versicolor
63	5.2	2.7	3.9	1.4	Iris-versicolor
64	5.7	2.8	4.1	1.3	Iris-versicolor
65	5.6	2.7	4.2	1.3	Iris-versicolor
66	4.9	2.5	4.5	1.7	Iris-virginica

Slika 3.2: Prikaz treh klasifikacijskih razredov množice Iris in tabela vrednosti vzorcev.

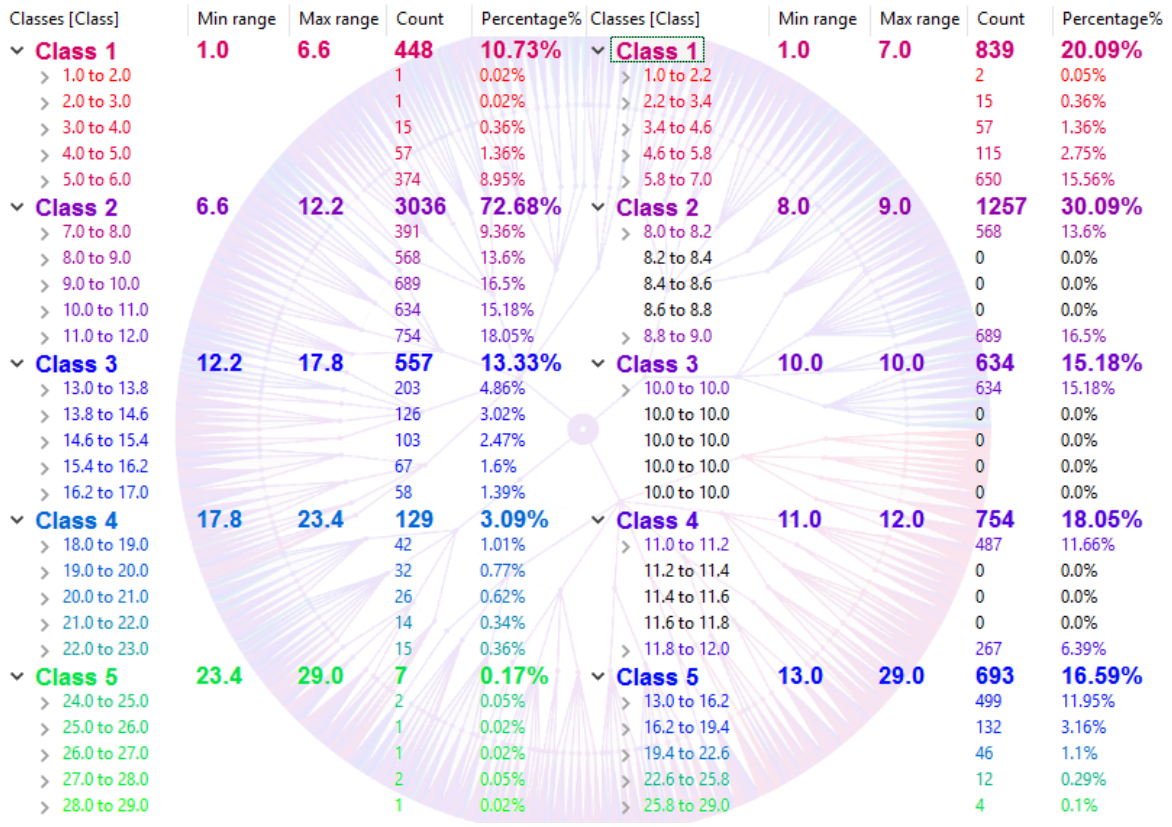
podintervalov enakih razponov. Na tak način zagotovimo enakomerno porazdelitev unikatnih napovedovalnih vrednosti čez vse razrede.

2. **Enakomerna razporeditev vzorcev čez razrede**, ki razdeli razpon vrednosti glede na vsebovano število vzorcev. Intervali so tako različnih dolžin in razporejeni tako, da noben vzorec z isto napovedovalno vrednostjo hkrati ne pripada več razredom.

Oba postopka sta prikazana na sliki 3.3, pri svojem delu pa lahko uporabnik poljubno spreminja način predstavitve v glavnem oknu.

Glavno okno v našem orodju omogoča še nalaganje podatkov, navigacijo po hierarhiji gruč in njihovo izbiro ter pregled ciljnih razredov podatkovne množice in njihovih vrednosti. Orodje omogoča tudi izbiro več gruč, s pomočjo katere lahko izvedemo medsebojno primerjavo statističnih mer. Z izbiro posamezne gruče se nad pripadajočimi vzorci samodejno izvede statistična analiza, katere rezultati pa so v obliki seznama spremenljivk in statistike prikazani v novem oknu. Glavno okno omogoča tudi hiter pregled informacij o gruči v obliki pojavnega okna, sestavljenega iz štirih komponent (Slika 3.4):

1. **Naslovna vrstica**, ki prikazuje nivo gruče, identifikacijsko številko ali poljubno poimenovanje in evklidsko razdaljo od izbrane gruče.



Slika 3.3: Delitev vzorcev v razrede, kjer leva stran prikazuje delitev na enake intervale, medtem ko desna stran prikazuje delitev intervalov glede na število vzorcev. Največja razlika delitve je opazna pri razredu, poimenovanem *Class 2*.

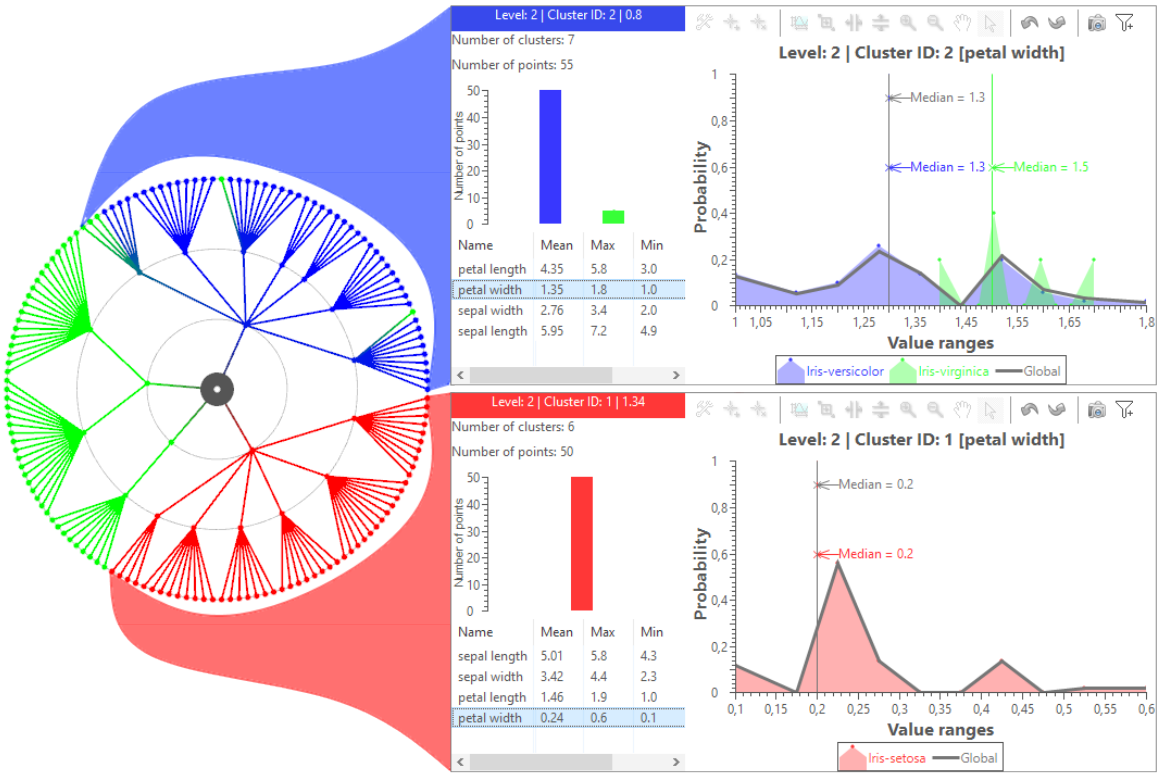
2. **Informacije o vzorcih**, ki prikazujejo število vzorcev in podgruč na naslednjem nivoju ter stolpične grafikone količine vzorcev po klasifikacijskih razredih.

3. **Seznam spremenljivk**, ki prikazuje prvih n spremenljivk z najmanjšim odstotkom medrazrednih prekrivanj na intervalu in štiri osnovne statistične mere (povprečna vrednost, najvišja in najnižja vrednost ter Pearsonova korelacija).

4. **Pregled funkcije gostote verjetnosti** nad izbrano spremenljivko.

3.2 Statistična analiza

Statistična analiza se nad izbrano gručo izvede samodejno, prikaz rezultatov pa je izveden v ločenem oknu. Okno je sestavljeno iz treh horizontalnih komponent. To so:



Slika 3.4: Pregled informacij o gruči. Levo je prikazana hierarhija gruč, medtem ko desna stran slike prikazuje pojavna okna z osnovnimi informacijami za modro (zgoraj) in rdečo (spodaj) gručo na nivoju 2.

1. **Seznam spremenljivk**, ki prikazuje rezultate izračunanih statističnih metrik, katerih podrobnejši opis je podan v nadaljevanju. Omogoča tudi vpogled v vrednosti posameznih vzorcev (tabela podatkov), iskanje po imenu spremenljivke in pregled grafa raztrosa med izbrano ter napovedovalno spremenljivko. Z izbiro spremenljivke izvedemo prikaz funkcije gostote verjetnosti, ki se za ta namen izvede v ločeni komponenti (glej točko 3).

2. **Pregled razredov** ciljne spremenljivke izbranih vzorcev, kjer so prikazane tudi informacije o intervalih in deležu vzorcev klasifikacijskih razredov.

3. **Komponenta prikaza funkcije gostote verjetnosti** prikazuje funkcijo gostote verjetnosti po klasifikacijskih razredih za izbrano spremenljivko. Komponenta omogoča tudi enostavno izbiro intervala za določen razred in izbiro poljubnega intervala ter enostavno dodajanje novih pravil v trenutno drevo pravil.

3.2.1 Statistične metrike

Za lažjo interpretacijo podatkov smo poleg vizualizacije implementirali še množico statističnih mer, ki podajajo dodatne informacije o posameznih spremenljivkah. V naši rešitvi smo jih organizirali v naslednje kategorije:

- **Meje vzorčne populacije** določajo najvišje in najnižje vrednosti v vzorčni populaciji.
- **Povprečne vrednosti** vzorcev opazovane spremenljivke, ki jih delimo na:
  - Aritmetično sredino.
  - Mediano, ki je definirana kot srednja vrednost urejenega seznama vrednosti opazovane spremenljivke.
  - Modus, ki je definiran kot vrednost z največjo frekvenco pojavitve opazovane spremenljivke.
- **Razpršenost vzorčnih podatkov**, ki jo določajo naslednje metrike:
  - Varianca, ki je definirana kot povprečje kvadratnih razlik od povprečne vrednosti opazovane spremenljivke po enačbi:

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{n}(x - \bar{x}) \tag{3.1}$$

  - Standardni odklon, ki opisuje odstopanje od povprečne vrednosti opazovane spremenljivke in je podan kot $s = \sqrt{s^2}$.
  - Standardna napaka povprečja, ki predstavlja mero odstopanja v podatkih in je odvisna od velikosti vzorčne populacije. Nizka standardna napaka pomeni majhno odstopanje vrednosti opazovane spremenljivke od povprečja. Standardna napaka je definirana kot:

$$s_e = \frac{s}{\sqrt{N}}. \tag{3.2}$$

- **Oblika podatkov [13]**, ki je ovrednotena z:

- Frekvenco ničelnih vrednosti
- Nagnjenost (ang. Skewness), ki predstavlja mero simetrije porazdelitve podatkov. Negativna vrednost predstavlja nagnjenost vzorčne populacije v levo, pozitivna vrednost v desno in vrednost 0 popolno simetrijo vzorčne populacije. Primer popolne simetrije je normalna porazdelitev. Simetrijo porazdelitve lahko približno ocenimo tudi kot razliko med aritmetično sredino in mediano.
- Sploščenost (ang. Kurtosis), ki predstavlja mero sploščenosti ali koničastosti porazdelitve vzorcev spremenljivke relativno na normalno porazdelitev. Negativna vrednost predstavlja sploščenost funkcije, medtem ko pozitivna koničavost.

Poleg predstavljenih tradicionalnih statističnih metrik naša rešitev vsebuje tudi nekatera naprednejša statistična orodja, kot so to interval zaupanja, statistični testi in korelacije.

Interval zaupanja podaja verjetnost, s katero bo povprečna vrednost naključne množice vzorcev pripadala danemu intervalu [14]. V orodje smo vključili interval zaupanja pri stopnji zaupanja 95 %. Izračunamo ga tako, da nad celotno populacijo zgradimo tabelo porazdelitve vrednosti. 95 % vsebovanih vrednosti v tabeli nato pomnožimo s standardnim odklonom naključne populacije ali populacije podgruče. Na dobljeni rezultat se lahko zanesemo le v primeru normalne porazdelitve podatkov. Zaradi strukture podatkovne baze (hierarhično gručenje) je očitno, da vzorci v podgručah niso naključni vzorci. Kadar se povprečja posameznih podgruč opazno razlikujejo med seboj in s povprečjem celotne populacije, je takšna metrika neuporabna.

Statistični testi predstavljajo alternativno metriko, ki jo lahko uporabimo tudi v primerih, ko interval zaupanja ni uporaben. V našo orodje smo vključili naslednja statistična testa:

1. **Test T (ang. Two-sample test)** je test enakih povprečij [15]. V ta namen, populacijo razdelimo v dva razreda in nato izračunamo razliko povprečij po naslednji enačbi:

$$\mathrm{T} = \frac{\bar{X}_a - \bar{X}_b}{\sqrt{\frac{s_a^2 + s_b^2}{N}}} \tag{3.3}$$

Ker je test omejen na 2 razreda, smo ga v naši rešitvi zamenjali z analizo variance ozi-

roma ANOVA (ang. **AN**alysis **O**f **VA**riance) in statistiko F, ki ju opišemo v naslednjem koraku.

2. **ANOVA in statistika F [16]**, kjer v prvem koraku za vsak razred izračunamo povprečno vrednost $\bar{x}_g$ in določimo skupno povprečno vrednost kot povprečje povprečij vseh razredov $\bar{X}$. V naslednjem koraku izračunamo skupni vsoti kvadratov znotraj razredov (SS_W) in med razredi (SS_B) po naslednji enačbi:

$$\mathrm{SS}_W = \sum_{g=1}^{m} \sum_{i=1}^{n} (x_{gi} - \bar{x}_g) \tag{3.4}$$

$$\mathrm{SS}_B = \sum_{g=1}^{m} m(\bar{x}_g - \bar{X}) \tag{3.5}$$

Vrednost testa F lahko nato določimo kot:

$$F = \frac{\frac{\mathrm{SS}_b}{m-1}}{\frac{\mathrm{SS}_w}{m(n-1)}}, \tag{3.6}$$

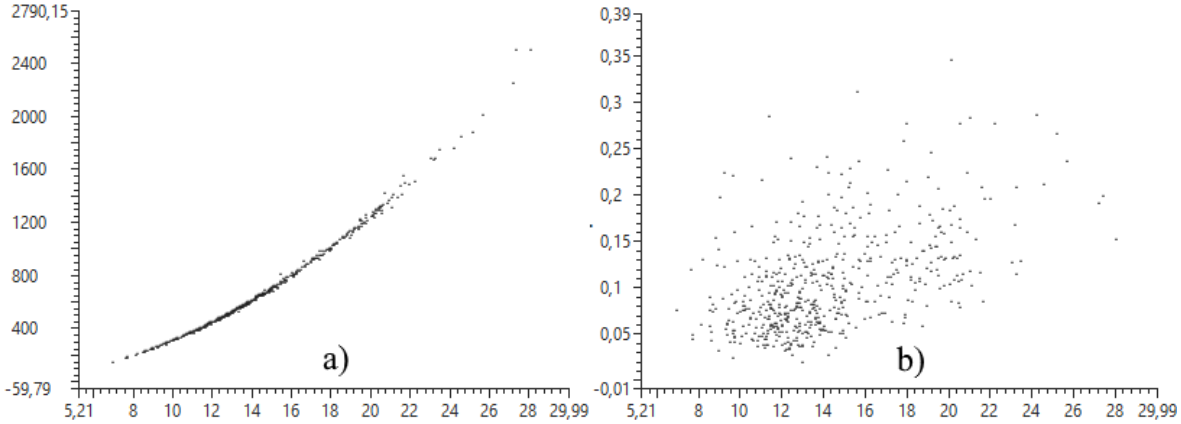
pri čemer $m-1$ in $m(n-1)$ predstavljata število prostostnih stopenj.

V statističnih analizah pogosto želimo tudi identificirati tiste spremenljivke, ki so karakteristične za napovedni razred. Način ocenitve medsebojnega vpliva dveh spremenljivk imenujemo korelacija.

Korelacijski koeficient predstavlja mero odvisnosti opazovanih spremenljivk. Interval koeficienta je med $[-1, 1]$, pri čemer vrednost proti -1 predstavlja negativno korelacijo, vrednosti proti 1 pozitivno korelacijo in vrednost blizu 0 medsebojno neodvisnost spremenljivk (Slika 3.5).

Oceno korelacije lahko izvedemo z več metrikami. V našem primeru smo se osredotočili na naslednje pogosto uporabljene pristope:

1. **Pearsonov koeficient** je mera linearne odvisnosti dveh spremenljivk, ki je podan je z



Slika 3.5: Primer porazdelitve vzorcev dveh odvisnih spremenljivk (a) in porazdelitve vzorcev dveh neodvisnih spremenljivk (b).

naslednjo enačbo [17]:

$$c = \frac{n\sum_{i=1}^{n}(x_i y_i) - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n}(x_i^2) - (\sum_{i=1}^{n} x_i)^2} * \sqrt{n\sum_{i=1}^{n}(y_i^2) - (\sum_{i=1}^{n} y_i)^2}} \tag{3.7}$$

2. **Spearmanov koeficient** je za razliko od Pearsonovega koeficienta, izveden nad rangi spremenljivk namesto nad njihovimi dejanskimi vrednostmi [18]. Range določimo določimo z indeksom vrednosti v urejenem seznamu števil, pri čemer odstranimo podvojene vrednosti.

3. **Korelacija razdalje (ang. Distance Correlation)** spada med mere nelinearne odvisnosti [19]. Za njeno ocenitev najprej izračunamo matriko razdalj vrednosti za vsako spremenljivko posebej. Vrednostim v matriki odštejemo tri povprečne vrednosti (povprečje vrstice, stolpca in celotne matrike):

$$\mathrm{dcd}_{i,j} = d_{i,j} - \sum_{i=1}^{n} d_{i_j} - \sum_{j=1}^{m} d_{i_j} - \sum_{i=1}^{n}\sum_{j=1}^{m} d_{i_j} \tag{3.8}$$

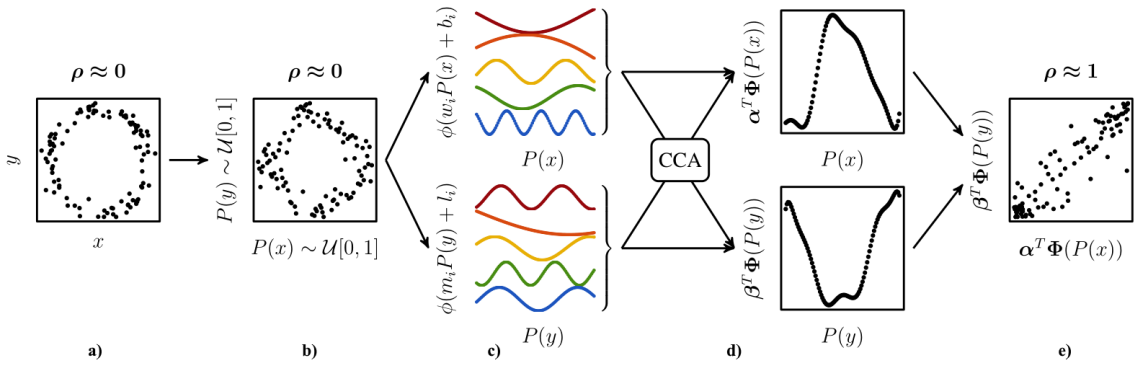
Korelacijo nato izračunamo po naslednji enačbi:

$$c = \frac{\mathrm{Cov(X,\ Y)}}{\sqrt{\mathrm{Cov(X,\ X)Cov(Y,\ Y)}}} \tag{3.9}$$

4. **Koeficient maksimalne informacije oziroma MIC (ang. Maximal Information Coefficient)**, ki temeljo na diskretizaciji prostora dveh spremenljivk [20]. Prostor v ta

namen razdelimo v mreže različnih dimenzij. To izvedemo tako, da dosežemo maksimizacijo skupne informacije. Algoritem nato poišče najboljše kandidate in jih shrani v matriko mrež različnih dimenzij. Vrednost z najvišjo oceno v tej matriki predstavlja korelacijo med spremenljivkama.

5. **Koeficient naključne odvisnosti** predstavlja mero nelinearne odvisnosti naključnih spremenljivk [21]. Njegovo oceno izvedemo tako, da najprej opazovani spremenljivki pretvorimo v funkcije skupne empirične verjetnosti oziroma ECDF (ang. Empirical Comulative Distribution Function). Vzorci slednje predstavljajo rang vhodnih vzorcev in so porazdeljeni enakomerno na intervalu [0, 1]. Nato za vsako spremenljivko ustvarimo matriko naključnih vrednosti normalne porazdelitve in jo zmnožimo z matriko skupnih empiričnih verjetnosti. To nato preslikamo v nelinearen prostor z uporabo funkcij *sin* in *cos*. V zadnjem koraku izvedemo še analizo kanoničnih korelacij oziroma CCA (ang. Canonical Correlation Analysis) med dvema transformiranima spremenljivkama. Opisan postopek prikazuje slika 3.6.

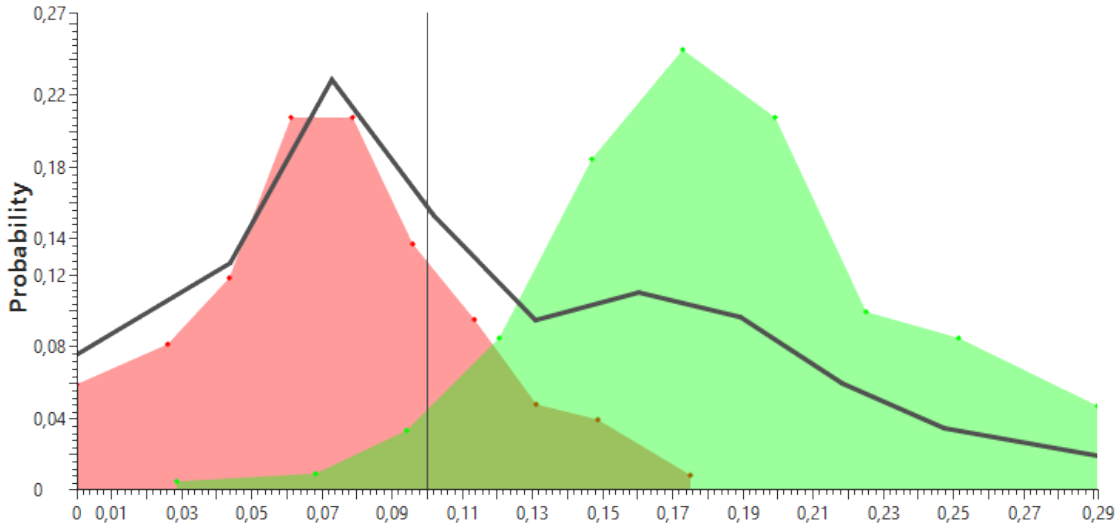


Slika 3.6: Postopek izračuna koeficienta naključne odvisnosti, kjer a) prikazuje vzorce dveh spremenljivk, b) zmnožek matrik ECDF in matrike naključne porazdelitve, c) preslikavo v nelinearen prostor, d) izračun kanoničnih korelacij, e) porazdelitev preslikanega prostora spremenljivk in p predstavlja rezultat Pearsonovega koeficienta pri dani distribuciji (a, b, e).

3.2.2 Funkcija gostote verjetnosti

Vrednosti opazovane spremenljivke lahko predstavimo s funkcijo gostote verjetnosti oziroma PDF (ang. Probability Density Function). Slednja nam poda verjetnosti pojavitve vredno-

sti vzorcev na določenih intervalih. V primeru številskih vrednosti interval razdelimo na n podintervalov, za katere vodimo frekvenco pojavitve vrednosti v normalizirani obliki. PDF gradimo nad razredi ločeno in jih vizualiziramo kot grafikon območja (ang. area chart), kar prikazuje slika 3.7. V primeru opisnih vrednosti delitev na podintervale ni potrebna in za njihovo predstavitev zadostuje že preprost histogram, ki ga prikažemo kot grafikon stolpičnih vrednosti.



Slika 3.7: Funkcija gostote verjetnosti za nad dvema, prikazanima z rdečo in zeleno barvo. Temna črta predstavlja PDF obeh razredov.

3.3 Graf odvisnosti spremenljivk

Predstavljene metode izračuna korelacij nam omogočajo odkrivanje povezav med različnimi spremenljivkami in izvedbo ocene njihovega vpliva na končno napoved. Spremenljivke lahko v ta namen predstavimo kot posamezna vozlišča v grafu in jih med seboj povežemo. Povezave tvorimo z analizo medsebojne odvisnosti spremenljivk, ki jo ocenimo z izračunom korelacijskega koeficienta po eni izmed implementiranih metod. Povezavo sprejmemo, kadar je odvisnost med spremenljivkama dovolj visoka. Graf odvisnosti zgradimo po naslednjem postopku:

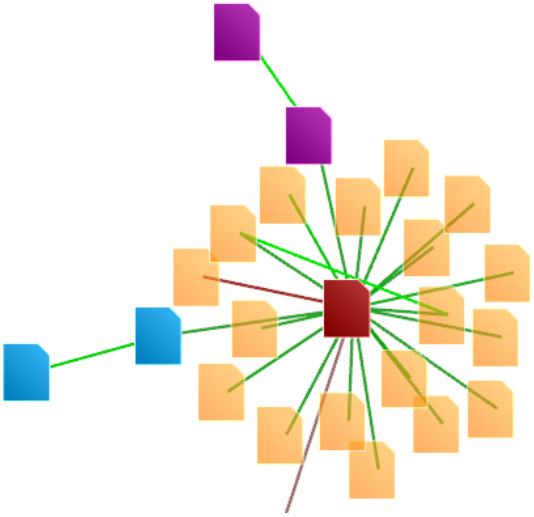
1. Izmed množice n spremenljivk izberemo ciljno spremenljivko.

2. Izračunamo korelacije do ostalih spremenljivk.
3. Iz padajoče urejenega seznama korelacij sprejmemo korelacije, ki zadoščajo določenemu pragu in največjemu dovoljenemu številu povezav med spremenljivkami (obe mejni vrednosti poda uporabnik).
4. Postopek rekurzivno ponavljamo za vsako sprejeto spremenljivko, dokler ne zadostimo izstopnim pogojem.

V računalništvu obstaja množica metod upodobitve grafov, med katerimi ima vsaka svoje prednosti in slabosti [22]. V našem primeru smo implementirali naslednje tehnike:

- **Planarna metoda**, pri kateri neplanaren graf (vsebuje sekajoče povezave) pretvorimo v planarnega z dodajanjem navideznih vozlišč v presečiščih povezav.
  - ***Prednosti*** so hitrost, delovanje nad različno geometrijo povezav (linije, polilinije) in dobro delovanje v praksi tudi nad večjimi grafi.
  - ***Slabosti*** sta zahtevna implementacija in nezmožnost zadostitve vsem omejitvam.
- **Sistem vzmeti**, ki za vsako vozlišče hrani vrednosti fizikalnih sil z namenom minimizacije sil celotnega sistema [23]. Časovna zahtevnost je kvadratna $O(|V|^2)$, pri čemer je $|V|$ red grafa, določen s številom vozlišč. Rezultat je sprejemljiva postavitev vozlišč, kar pomeni da so odvisnosti spremenljivk jasno razvidne.
  - ***Prednosti*** so preprosta implementacija in sprotna vizualizacija razvoja mreže ter enostavna razširitev v 3D prostor.
  - ***Slabosti*** so počasnost v smislu konvergence, rezultat ni optimalen, hkrati pa metoda tudi ni zmožna zadostiti vsem omejitvam.

Ciljno spremenljivko v orodju privzeto pobarvamo z rumeno barvo, neposredne sosede s svetlo modro in ostale spremenljivke s svetlo sivo barvo. Povezave so pobarvane z rdečo ali zeleno barvo. Rdeča barva predstavlja negativno korelacijo, zelena pa pozitivno. Svetlejša barva predstavlja močno korelacijo, medtem ko temnejša predstavlja šibko korelacijo med povezanima spremenljivkama (Slika 3.8).



Slika 3.8: Primer grafa medsebojne odvisnosti spremenljivk. Povezave so obarvane z zeleno ali rdečo barvo. Spremenljivke oranžne barve neposredno vplivajo na spremenljivko rdeče barve. Spremenljivke vijolične in modre barve so primer verižnega vpliva.

Gradnjo odvisnosti spremenljivk upravljamo s štirimi parametri. Prvi parameter je ime izhodiščne spremenljivke. Drugi parameter je maksimalno število sosednjih vozlišč in se uporablja skupaj s tretjim parametrom, ki določa pragove koeficienta odvisnosti po različnih nivojih in s tem preprečuje gradnjo polno povezanega grafa. Četrti parameter določa uporabljeno metodo za izračun korelacije med spremenljivkami. Te metode so Pearsonov koeficient, Spearmanovo rangiranje, korelacija razdalje, RDC in MIC.

Predstavljeni graf medsebojne odvisnosti omogoča izvedbo **analize medsebojne odvisnosti več spremenljivk**. V podatkih namreč pogosto pride do tako imenovanega pojava multivariabilnosti, ki opredeljuje skupen vpliv množice spremenljivk na izbrano opazovano spremenljivko. V praksi to pomeni, da imajo posamične opazovane spremenljivke šibko korelacijo s ciljno spremenljivko, medtem ko pa opazovane skupaj tvorijo močno korelacijo. Metoda za odkrivanje večvariabilnih korelacij se imenuje maksimalna analiza korelacije oziroma MAC (ang. Maximal Analysis of Correlation) in je generalizacija obstoječe metode MIC nad večdimenzionalnim prostorom [24]. V orodju lahko spremenljivke po želji združujemo v skupine, nad katerimi opravimo multivariabilno analizo nad ciljno spremenljivko grafa odvisnosti. Praviloma ne poznamo skupine spremenljivk, ki najbolje vplivajo na rezultate vrednosti ciljne spremenljivke. V postopku analize jih moramo zato najprej poiskati. Zaradi množice obravnavanih spremenljivk in visoke računske zahtevnostni pa naivnega (ang. brute force) pristopa v tem primeru ni mogoče izvesti. V predlaganem orodju smo zato implementirali

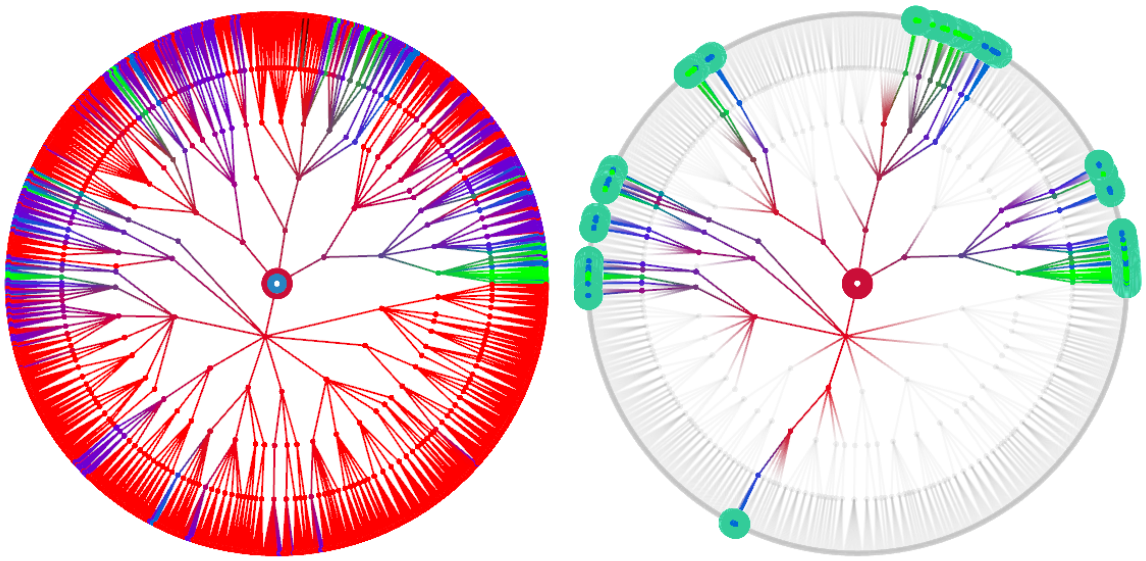
dva inteligentnejša pristopa k identifikaciji takšnih spremenljivk. Prva metoda izvede maksimizacijo korelacije iz kombinacije majhnih množic skupine spremenljivk, druga metoda pa maksimizacijo iz celotne množice skupine. Potek prve metode je naslednji:

1. Tvorba vseh možnih kombinacij parov spremenljivk iz celotne množice in izračuna njihovih odvisnosti z ciljno spremenljivko.
2. Par z najvišjo oceno (glede na prejšnji korak) nato dopolnimo z novo spremenljivko.
3. Če je nova ocena višja od prejšnje ocene, potem ponovimo korak 2.
4. V primeru da nova ocena ni višja, postopek zaključimo in vrnemo najboljšo kombinacijo, skupaj z njeno oceno.

Postopek druge metode je podoben, pri čemer pa spremenljivke odstranjujemo, namesto da bi jih dodajali.

3.4 Filtriranje vzorcev

Poleg predstavljenih statističnih analiz smo v opisanem orodju omogočili tudi postopek odkrivanja znanja s pomočjo podatkovnih filtrov. Filtriranje tako služi kot pomoč izbire podmnožice vzorcev, ki nas zanimajo. Pravila sestavljajo pogoji spremenljivk, ki skupaj tvorijo drevo pravil. Ob filtriranju vsak vzorec pošljemo skozi drevo in preverimo, ali ta zadosti vsem pogojem filtriranja. Vsak pogoj je pri tem predstavljen kot vozlišče v drevesu in vsebuje informacije o spremenljivki (ime, identifikacijska številka), vrsto pogoja (večje, manjše, enakost, ...) in vrednost ter barvo ozadja vzorca. V drevesu sosednji pogoji predstavljajo logični *ALI*, medtem ko neposredni sinovi predstavljajo logični *IN*. Vzorce, ki ne zadostijo pogojem, v vizualizaciji spremenimo v manj opazne in jih izključimo iz nadaljnih analiz (Slika 3.9). Aplikacija omogoča izvoz in uvoz drevesa pogojev v ali iz datoteke formata XML (ang. Extensible Markup Language).



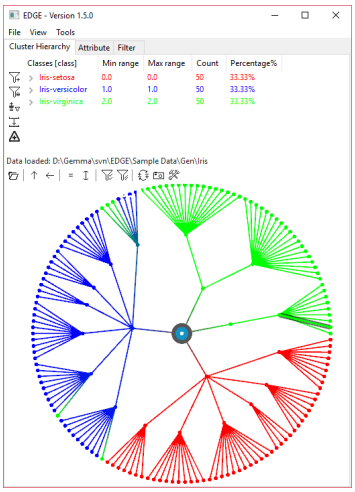
Slika 3.9: Prikaz hierarhije gruč a) brez in b) z vključenim filtriranjem. Zeleno ozadje predstavlja izbrano barvo pravila.

Poglavje 4

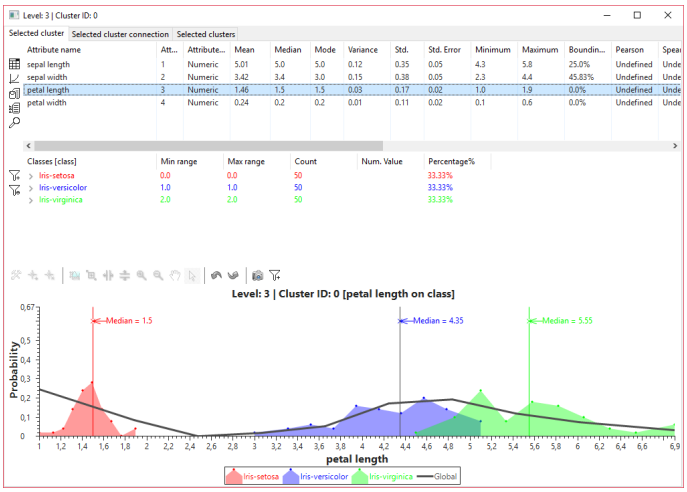
Primeri uporabe predstavljenega orodja

Prvi primer uporabe se osredotoča na analizo podatkov iz zbirke *Iris* [25], ki vsebuje meritve rastlin. To so njihove dolžine, širine venčnih in ćašnih listov ter njihovo klasifikacijo v tri podrazrede (Iris-setosa, Iris-versicolor, Iris-virginica). Skupno število vzorcev je 150, in sicer po 50 vzorcev vsakega razreda. Podrazred *Iris-setosa* je linearno ločljiv od ostalih dveh razredov, medtem ko se vzorci razredov *Iris-versicolor* in *Iris-virginica* med seboj prekrivajo. To postane očitno takoj, ko podatke naložimo in prikažemo strukturo gručenja v obliki radialnega drevesa (Slika 4.1). Takšen prikaz pa nam omogoča izbiro množice vzorcev z izbiro poljubnega vozlišča (podgruče) v drevesu. Zunanja vozlišča so pobarvana glede na vrednost ciljne spremenljivke. Najnižja vrednost predstavlja rdečo barvo, srednja rumeno (modra v primeru svetlega ozadja) in najvišja zeleno. Barva starševskega vozlišča je določena kot povprečje barv sinov. Na enak način se določi tudi barva pripadajočemu razredu. Za analizo vseh vzorcev izberemo srednje vozlišče. V tem primeru se ustvari novo okno, ki prikazuje rezultate statistične analize (Slika 4.2).

Z izbiro spremenljivke sprožimo gradnjo funkcije gostote verjetnosti, ki jo prikažemo v spodnji tretjini okna. Pri opazovanju spremenljivk lahko na tak način ugotovimo, na katerih intervalih so posamezni razredi ločljivi med seboj. Druga tretjina okna nam omogoča izbiro posameznega razreda in grafično označbo na grafu PDF. Meni konteksta nad razredom omogoča nastavitev intervala grafa PDF znotraj mej razreda, s čimer lahko določamo pravila za gradnjo filtrov.



Slika 4.1: Prikaz hierarhije gruč podatkov *Iris* v glavnem oknu, ki se deli na tri klasifikacijske razrede. Označeno je vozlišče v središču kroga, ki predstavlja koren ter vključuje vse podatke.



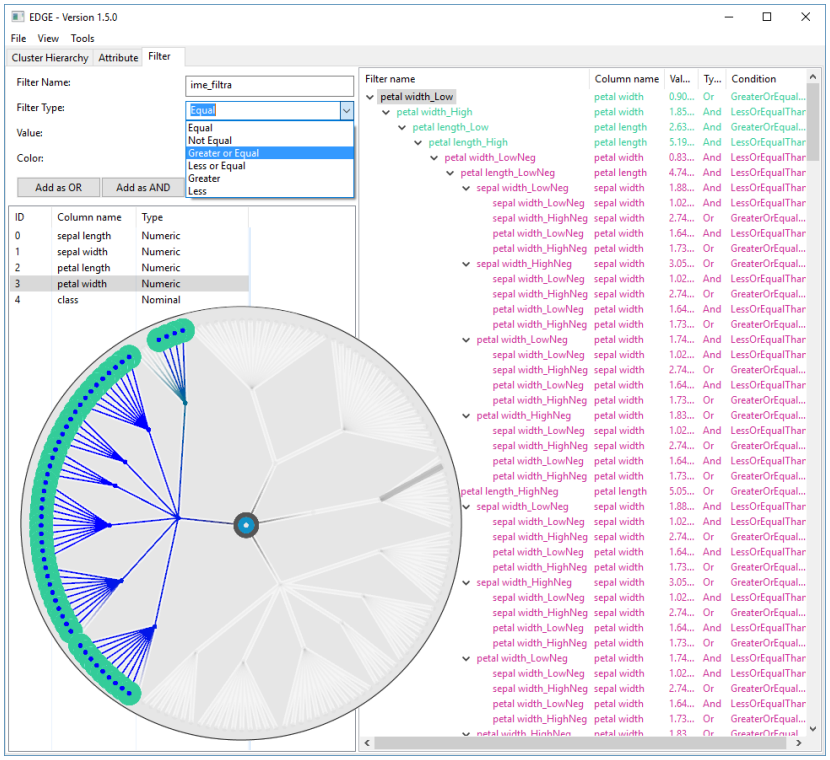
Slika 4.2: V oknu sta prikazana statistična analiza podatkov izbrane gruče s slike 4.1 in graf PDF izbrane spremenljivke v seznamu. Razvidno je tudi, da je razred rdeče barve *Iris-setosa* linearno ločljiv od ostalih razredov že po izbrani spremenljivki.

4.1 Drevo pravil

Aplikacija omogoča gradnjo drevesa pravil na več različnih načinov. Prvi način je ročni in se izkaže za precej zamudnega in napornega (Slika 4.3). Postopek dodajanja novega pravila poteka v naslednjih korakih:

1. Najprej izberemo očeta iz drevesa (če le-to že vsebuje vnose).
2. Nato izberemo spremenljivko iz seznama. Prikazane so samo tri vrednosti (id, ime in tip spremenljivke).
3. Pogoj filtra lahko določimo z izborom ene izmed naslednjih operacij: $=, \neq, >, <, \geq, \leq$. Slabost tega pristopa je dvojno dodajanje pravil za vrednosti na intervalu (najprej vnos za $>$, nato pa še $<$).
4. Glede na dodan pogoj lahko nato določimo vrednosti pogoja. Te so lahko numerične ali opisne, njihov tip pa je odvisen od tipa spremenljivke.
5. Na koncu lahko določimo še barvo, s katero bodo obarvani vzorci, ki zadostijo drevesu pravil.

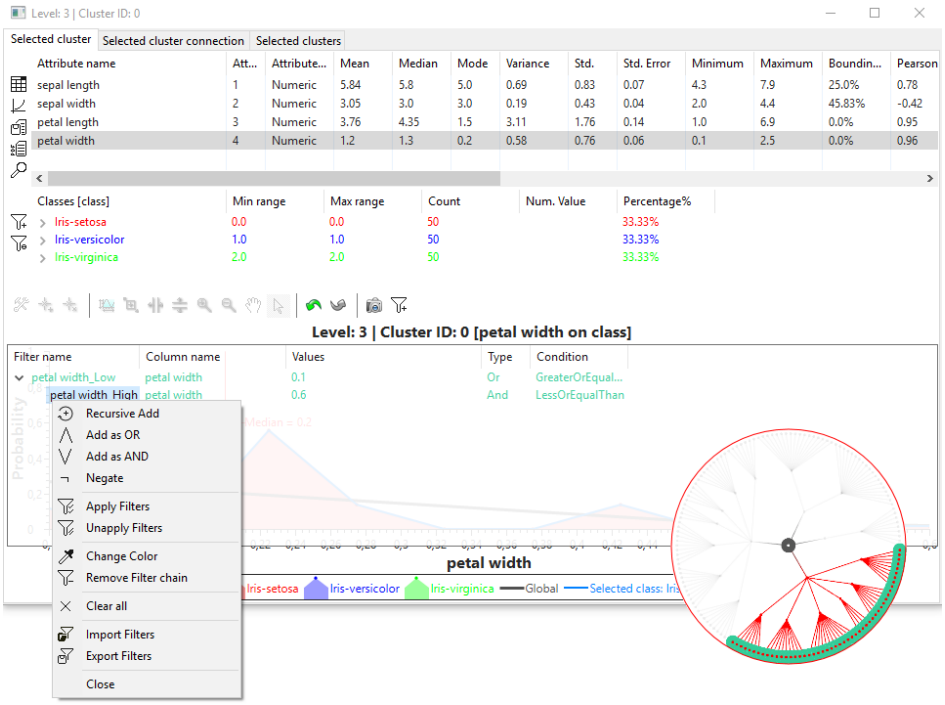
Vzorcem, ki ne zadoščajo pogojem, nastavimo manj opazno barvo. Vse nadaljnje statistične analize pa tako izvajamo zgolj nad vzorci, ki zadoščajo pogojem.



Slika 4.3: Pregled drevesa pravil za razred *Iris-versicolor*.

Gradnjo drevesa pravil lahko izvajamo tudi enostavneje z uporabo grafičnega vmesnika. V komponentah, ki prikazujejo grafe PDF ali porazdelitve vzorcev med spremenljivkami, imamo možnost gradnje pravil z izbiro mej na samem grafu. Z možnostmi v meniju konteksta lahko vstavimo novo pravilo v obstoječe drevo. V primeru velikega števila spremenljivk lahko te uredimo po glede na želene statistične mere, drevo pravil pa gradimo sprotno. Pri dodajanju novih pravil nam orodje ponuja več možnosti, kar prikazuje slika 4.4. Pri tem pa je vsako novo pravilo pa je sestavljeno iz verige $l \leq x \leq r$, pri čemer sta l spodnja meja in r zgornja meja uporabniško določenega intervala. Tako lahko izvedemo:

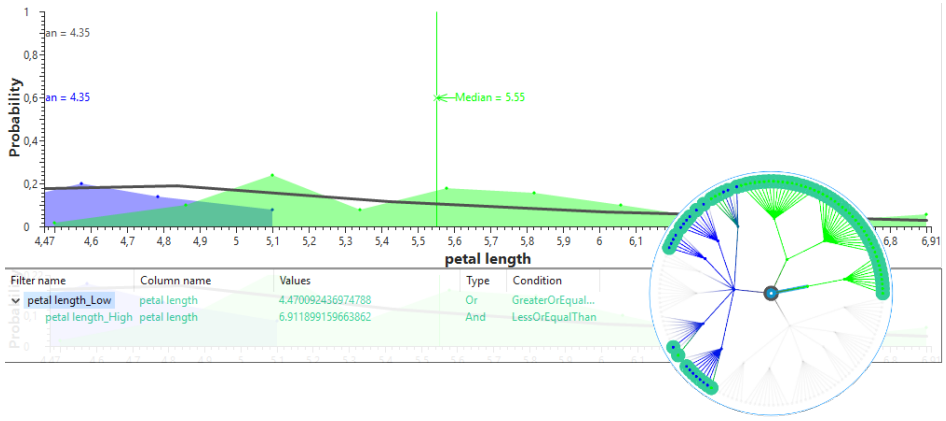
- **Rekurzivno dodajanje**, kjer dodamo novo pravilo vsem posrednim listom izbranega pravila v drevesu kot pravilo *IN*.
- **Dodaj pravilo ALI (∧)**, kjer staršu izbranega pravila dodamo novo pravilo *ALI*.
- **Dodaj pravilo IN (∨)**, kjer izbranemu pravilu dodamo novo pravilo kot pravilo *IN*.



Slika 4.4: Filtriranje vzorcev, ki pripadajo intervalu [0,1 0,6] glede na spremenljivko *petal width*. S tem ohranimo samo vzorce razreda *Iris-setosa* (vozlišča rdeče barve).

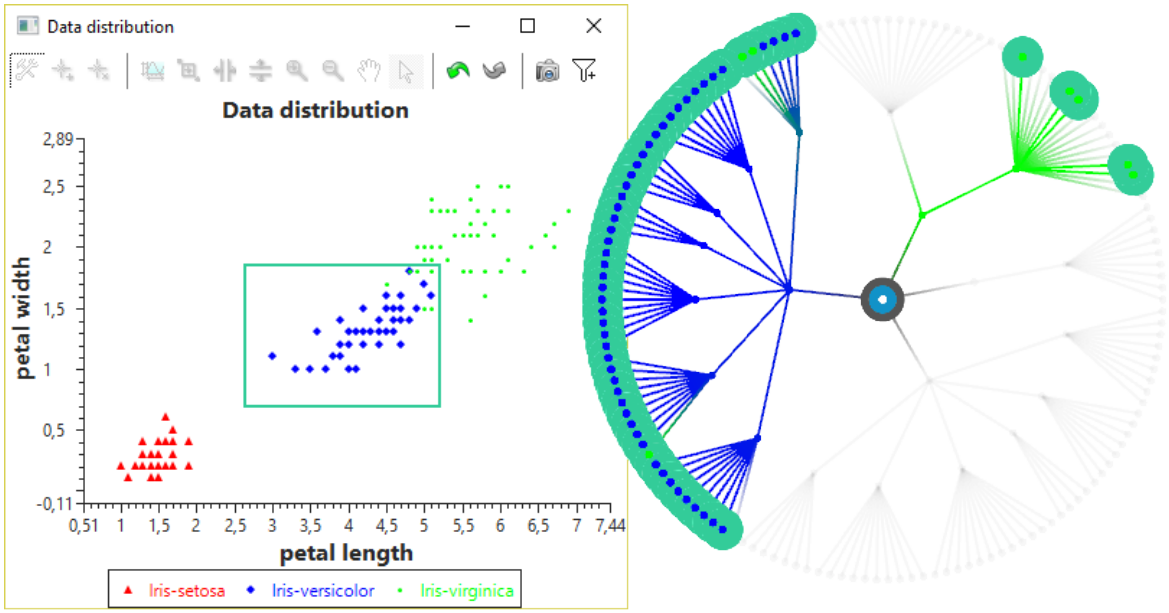
- **Negacija pravila**, kjer izbrano pravilo zamenjamo z nasprotnim pravilom. V primeru pravila večje/manjše, se pri tem spremeni struktura drevesa. Tako na primer pravilo $l \leq x \leq r$ postane $x < l \wedge x > r$.
- **Uporabi drevo pravil**, kjer filtriramo vse vzorce s trenutnim drevesom pravil. Pri tem se ponovno izvede statistična analiza in ovrednoti PDF izbrane spremenljivke.
- **Odstrani verigo pravil**, kjer odstranimo vsa pravila, ki so posredni ali neposredni sinovi izbranega pravila.
- **Uvoz/Izvoz**, kjer pravila shranimo ali naložimo.

V našem primeru smo v drevo pravil dodali spremenljivko *petal width* na intervalu, ki ga pokriva razred *Iris-setosa* (Slika 4.4). V grafu lahko tudi povečamo ali zmanjšamo opazovani interval in ga dodamo kot novo pravilo. V našem primeru smo to storili nad spremenljivko *petal length* in se pri tem osredotočili na vrednosti razreda *Iris-virginica*, prikazanega na sliki 4.5.



Slika 4.5: Filtriranje vzorcev, ki pripadajo intervalu [4,47 6,91] glede na spremenljivko *petal length*. S tem ohranimo vse vzorce razreda *Iris-virginica* (vzorci zelene barve) in nekaj vzorcev iz razreda *Iris-versicolor* (vzorci modre barve).

Nad izbranimi vzorci lahko nato izdelamo graf verjetnosti porazdelitve. Graf porazdelitve vzorcev med dvema spremenljivkama aktiviramo z izbiro dveh poljubnih spremenljivk iz grafa odvisnosti. Čez prikazane točke v ta namen označimo območje, ki ga želimo dodati ali izločiti iz drevesa. V primeru dodajanja k izbranemu pravilu dodamo novo verigo štirih pravil $l_X \leq x \leq r_X \vee l_Y \leq y \leq r_Y$, pri čemer sta l_X in r_X meji prve izbrane spremenljivke ter l_Y in r_Y meji druge izbrane spremenljivke, kot to prikazuje slika 4.6. V primeru izbire možnosti izločanja vzorcev na izbranem območju se prejšnja veriga doda v drevo kot negacija.



Slika 4.6: Prikaz porazdelitve vzorcev med spremenljivkama *petal length* in *petal width* iz nabora podatkov *Iris*. Zeleno območje predstavlja izbrane meje za novo pravilo, katerega vzorce vidimo v pregledu hierarhije gruč.

4.2 Analiza medsebojne odvisnosti spremenljivk

Če analizirani nabor podatkov vsebuje veliko število spremenljivk, lahko podroben pregled vsake izmed njih postane dolgotrajen. V pomoč smo zato implementirali analizo odvisnosti spremenljivk, ki omogoča vpogled v njihove medsebojne odvisnosti. To omogoča tudi statistično analizo in pregled porazdelitve vzorcev ter funkcij gostote verjetnosti med poljubno izbranima spremenljivkama. Ta del orodja je namenjen napredni obdelavi in je neodvisen od izbrane ciljne spremenljivke. Vsebina okna analize odvisnosti spremenljivk je sestavljena iz devetih komponent in je prikazana na sliki 4.7. Te komponente so:

1. **Informacija o izbrani gruči in nivoju**, ki nam poda informacijo o izbrani množici vzorcev (gruče) in stanje, ki opredeljuje, ali je graf odvisnosti zgrajen iz množice izbrane gruče ali ne.

2. **Parametri gradnje grafa odvisnosti**, ki služijo upravljanju pragov korelacij po nivojih, omejevanju števila povezav vozlišča ter izbiri izhodiščne spremenljivke in metode analize korelacij.

3. **Orodna vrstica**, ki vsebuje gumbe za začetek gradnje grafa odvisnosti, prikaz surovih vrednosti vzorcev izbrane spremenljivke (tabela), prikaz porazdelitve vrednosti vzorcev med izbranima spremenljivkama, vklop/izklop izrisa imen slednjih in vrednosti njihovih korelacij v komponenti vizualizacije grafa odvisnosti, iskanje po imenu spremenljivke in spremembo njene barve ter uvoz/izvoz grafa odvisnosti v datoteko *XML*.

4. **Seznam sosedov**, ki prikazuje seznam povezanih spremenljivk s trenutno izbrano. Seznam služi izbiri druge spremenljivke za operacije, kot sta to prikaz porazdelitve vzorcev in statistična analiza med izbranima spremenljivkama. Izbira spremenljivke premakne pogled komponente vizualizacije grafa v njeno središče.

5. **Prikaz statistike izbrane spremenljivke**, ki prikazuje rezultate statistične analize izbrane spremenljivke. Analiza se opravi ob njeni izbiri in zajema metode, kot so povprečja, standardni odklon, najmanjša, največja vrednost, nagnjenost in sploščenost ter interval zaupanja.

6. **Prikaz statistike med izbranima spremenljivkama**, ki prikazuje rezultate statistične analize odvisnosti med izbranima spremenljivkama. Analiza se opravi ob izbiri druge spremenljivke iz komponente vizualizacije grafa ali seznama sosedov in zajema vse implementirane izračune korelacije ter statistične teste.

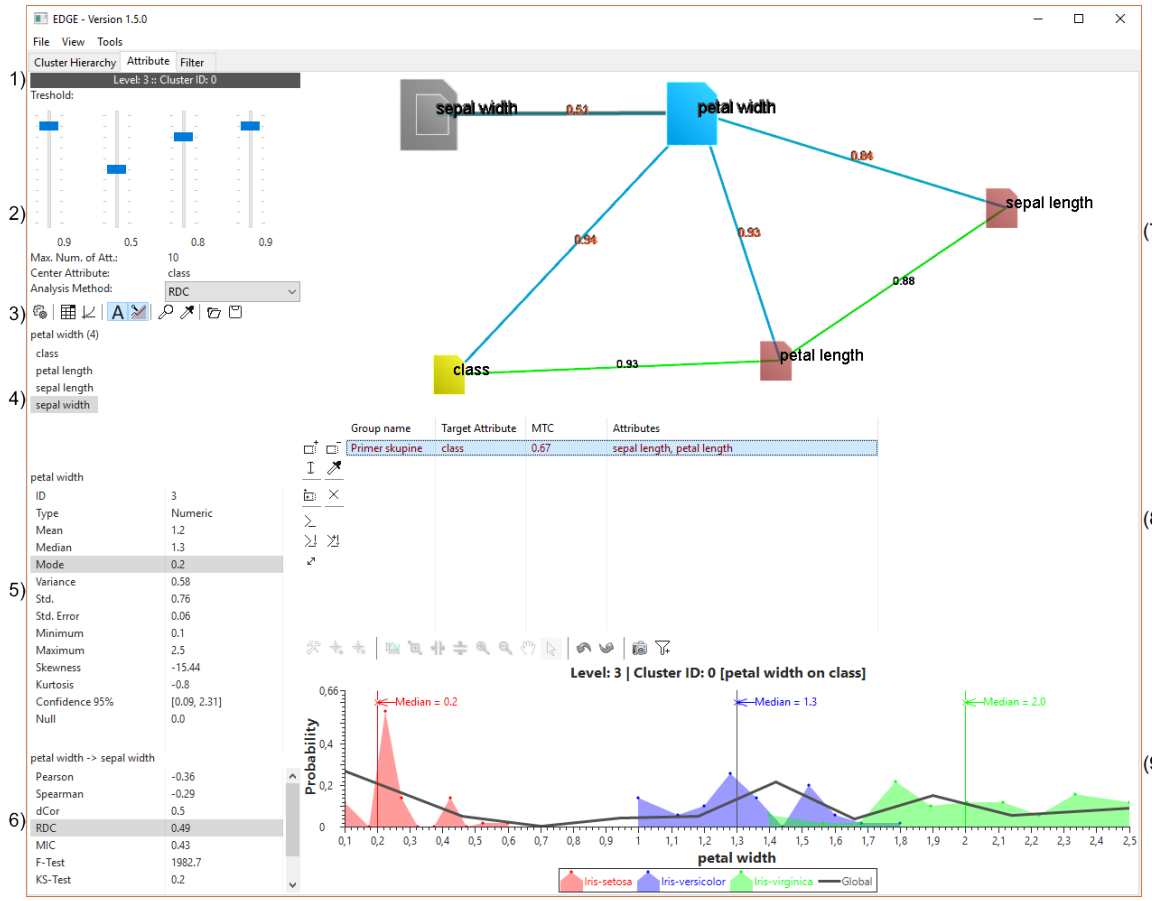
7. **Komponenta vizualizacije** grafa odvisnosti spremenljivk, ki prikazuje spremenljivke v obliki vozlišč in povezav z ostalimi spremenljivkami. Zraven vozlišč in povezav slednjih so prikazana tudi njihova imena in medsebojne ocene korelacij. Z izbiro spremenljivke se upodobi še funkcija gostote verjetnosti v komponenti PDF. Komponenta vizualizacije je opremljena tudi s kontekstnim menijem, ki omogoča hiter dostop do ostalih funkcionalnosti grafičnega vmesnika. Te možnosti so:

    - prikaz tabele podatkov označene spremenljivke,
    - prikaz komponente porazdelitve vzorcev ali komponente PDF med označeno spremenljivko in njej povezanimi,
    - prikaz porazdelitve vzorcev in PDF med označeno in izbrano spremenljivko in
    - združevanje označenih spremenljivk v skupine.

8. **Komponenta za analizo multivariabilnosti skupine spremenljivk**, ki služi združevanju posameznih spremenljivk v skupine in analizi skupne odvisnosti z izhodiščno spremenljivko.

9. **Komponenta PDF**, ki prikaže funkcijo gostote verjetnosti za izbrano spremenljivko in omogoča enostavno dodajanje novih pravil.

Vzorce povezanih spremenljivk lahko vzporedno opazujemo v obliki grafa raztrosa. Porazdelitev vzorcev spremenljivke *petal width* z njej povezanimi (sepal width, class, petal length in sepal length) spremenljivkami prikazuje slika 4.8. Barva vzorca je določena kot vrednost pripadajočega razreda. V komponenti lahko tudi odstranimo grafe, ki nas ne zanimajo, ali dodajamo nova pravila v drevo.

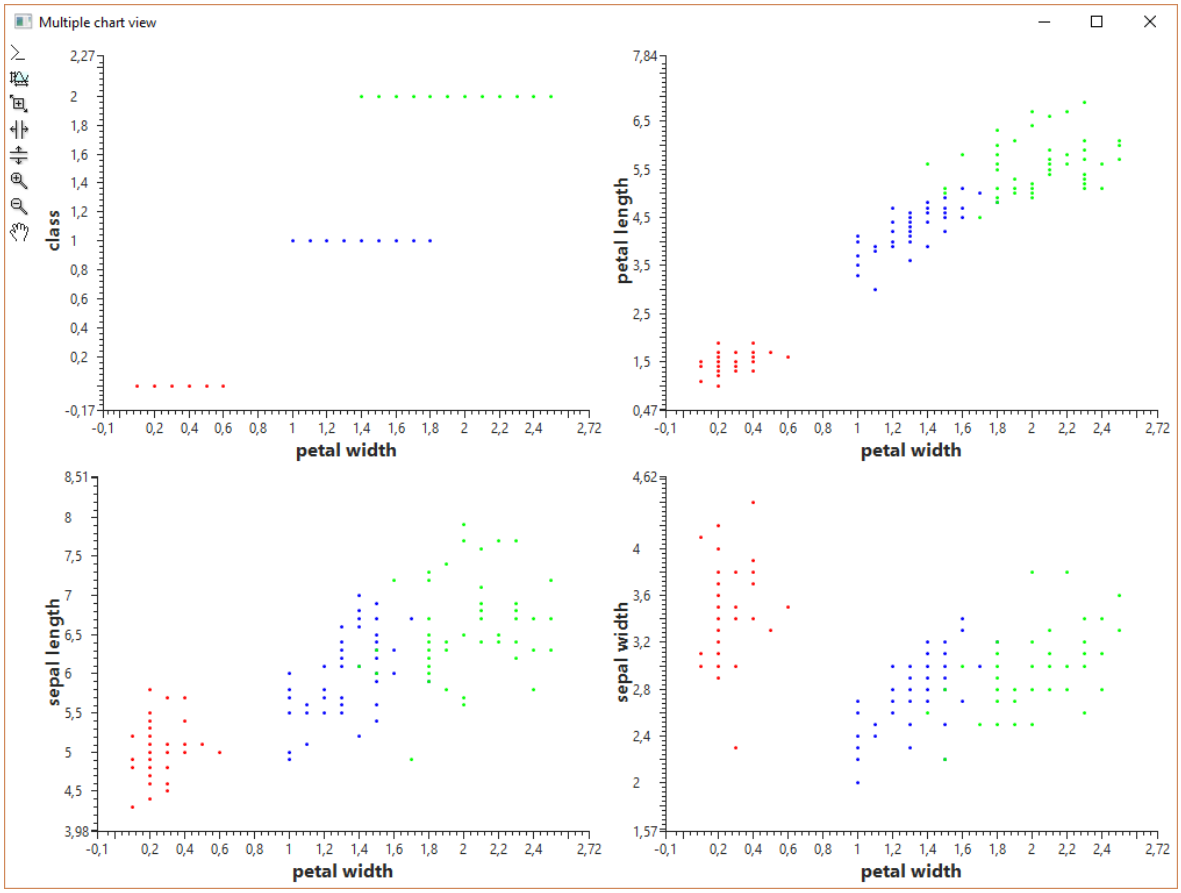


Slika 4.7: Pogled komponent analize medsebojne odvisnosti spremenljivk za podatke *Iris* za točke 1 - 9.

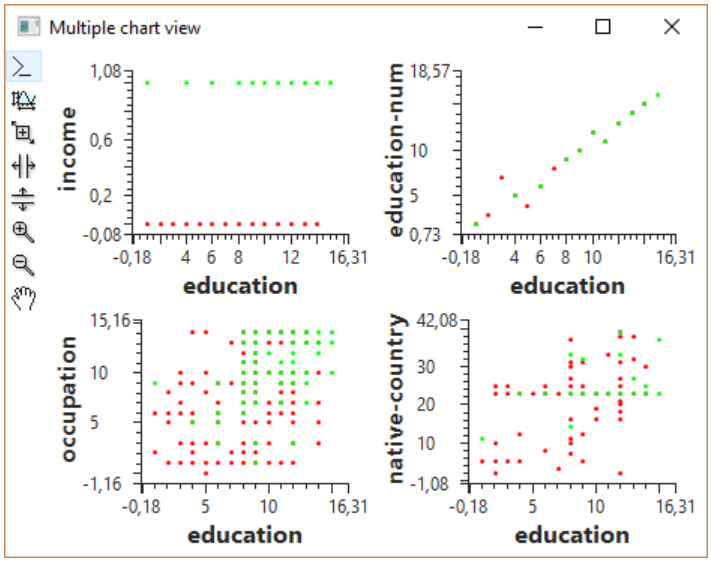
4.3 Prikaz gostote verjetnosti povezanih spremenljivk

Za demostracijo prikaza gostote verjetnosti povezanih spremenljivk smo izbrali podatke *Adult* [26], ki predstavljajo cenzus. Podatke sestavlja 14 spremenljivk in 32561 vzorcev. Večina spremenljivk v teh podatkih je opisnih. V primeru takšnih spremenljivk nam prikaz razpršitve vzorcev ne predstavlja dovolj informacije zaradi prekrivanja posameznih razredov na enakih območjih (Slika 4.9).

Cilj napovedi pri podatkih *Adult* je oceniti letni prihodek posameznika glede na 13 danih spremenljivk. V tem primeru smo zgradili graf odvisnosti spremenljivk in se odločili analizirati vpliv izobrazbe (education) na prihodek (income), poklic (occupation) in matično državo (native-country). Najprej smo analizirali porazdelitev vzorcev med izobrazbo in naštetimi spremenljivkami in opazili prekrivanje vzorcev na enakih območjih. V ta namen smo imple-



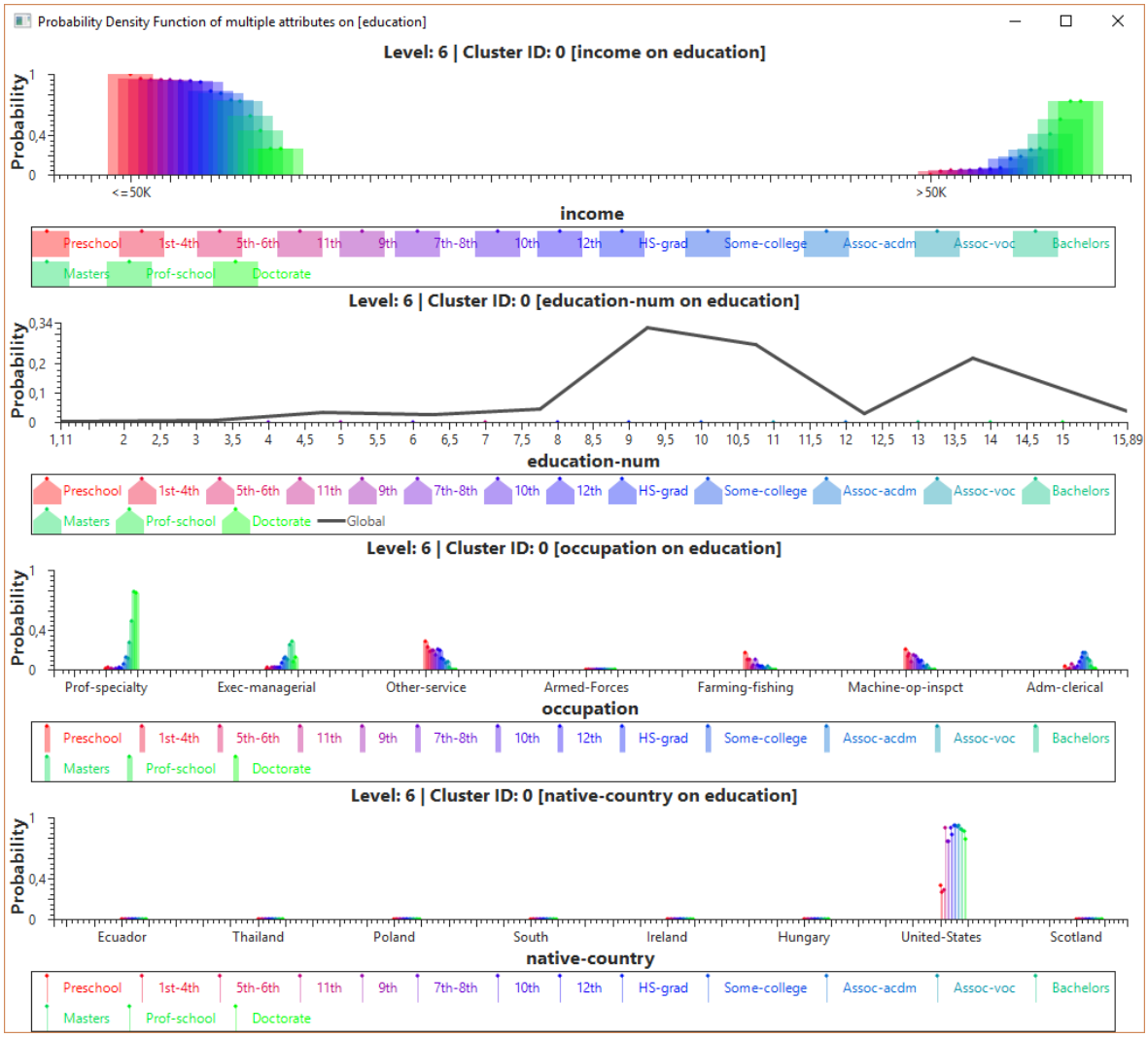
Slika 4.8: Prikaz porazdelitve vzorcev spremenljivke *petal width* glede na ostale povezane spremenljivke.



Slika 4.9: Prikaz porazdelitve vzorcev opisnih spremenljivk podatkov *Adult*.

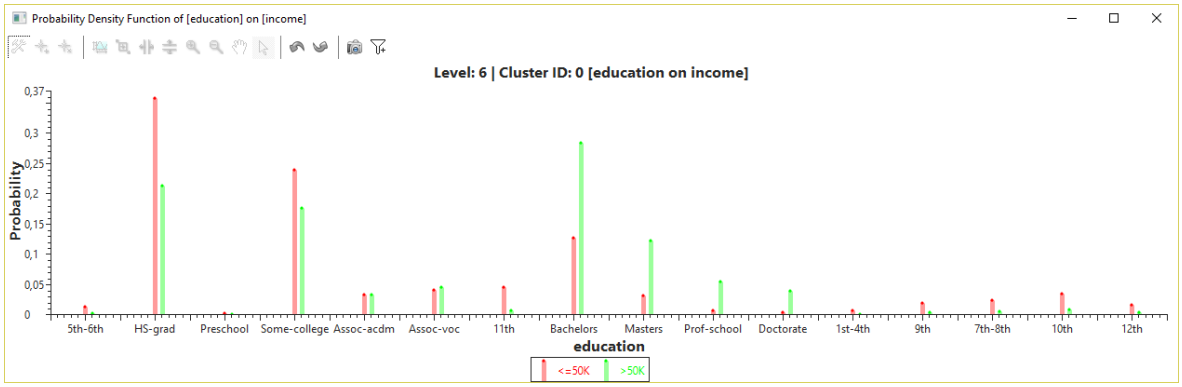
mentirali možnost sočasnega pregleda funkcij gostot verjetnosti nad takšnimi spremenljivkami, ki nam omogoča hiter pregled gostote vrednosti pojavitve vzorcev. V tej komponenti

barva ne predstavlja napovedovalnega razreda ampak intenziteto. Na sliki 4.10 vidimo funkcije gostote verjetnosti za vzorce iz slike 4.9. Iz tega lahko sklepamo, da stopnja izobrazbe in

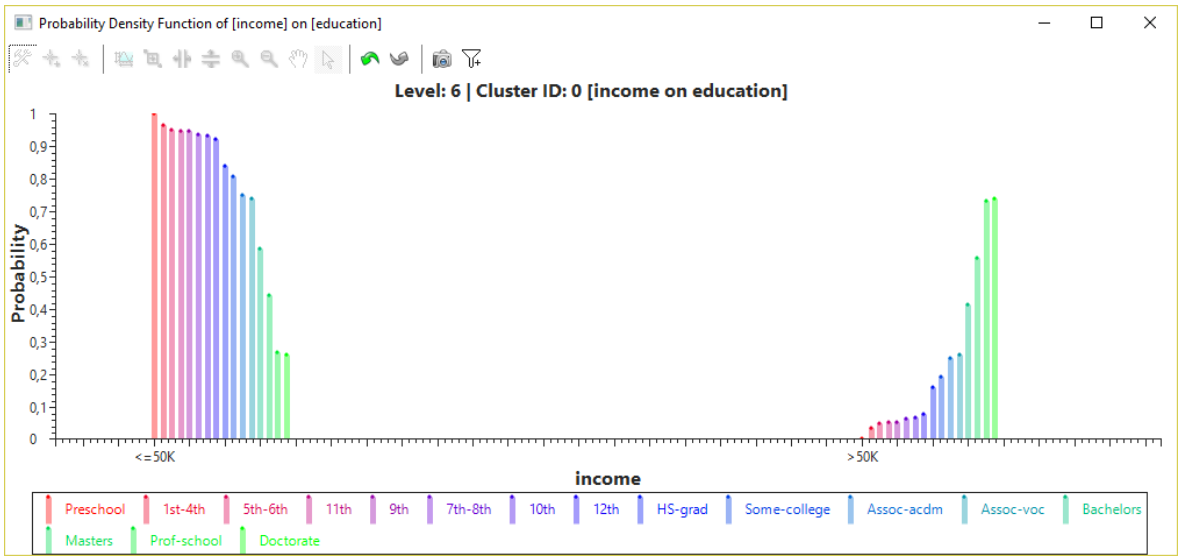


Slika 4.10: Prikaz funkcij gostot verjetnosti spremenljivke *education* in ostalih (*income*, *education-num*, *occupation* in *native-country*) spremenljivk v naboru podatkov *Adult*.

poklic vplivata na prihodek. Matična država nima velikega vpliva, vendar podatki nakazujejo na višjo verjetnost nižjega prihodka za vzorce iz Južne Amerike in podobne verjetnosti obeh razredov za ZDA. V primeru opisnih spremenljivk imamo na voljo dve predstavitvi vrednosti histogramov. To pomeni verjetnost vrednosti iz prve spremenljivke na drugo ali obratno, kot to prikazujeta sliki 4.11 in 4.12.



Slika 4.11: Prikaz histograma vpliva izobrazbe na letni prihodek.



Slika 4.12: Prikaz histograma vpliva letnega prihodka glede na izobrazbo.

Poglavje 5

Rezultati

Delovanje orodja smo preizkusili tudi nad različnimi zbirkami meteoroloških [27] in hidroloških [28] meritev območja Celja in okolice. Namen te študije je bil preučiti zmožnosti vpeljave podatkovno podprtega odločanja v primeru kriznih situacij. V ta namen smo se osredotočili na podatke, zbrane med letoma 2009 in vključno 2013. V tem intervalu so bile namreč v Sloveniji dvakrat dokumentirane večje poplave, ki so konkretno nastopile septembra 2010 [29] in novembra 2012 [30]. V ta namen pa smo kot ciljno (napovedovalno) spremenljivko izbrali pretok hidrološke postaje *Celje II brv* z identifikatorjem 6140 (Slika 5.1).

Podatki v tej študiji so bili zbrani s tremi različnimi tipi postaj, od katerih je vsaka zmožna izvajati različen nabor meritev. Uporabljeni tipi postaj so:

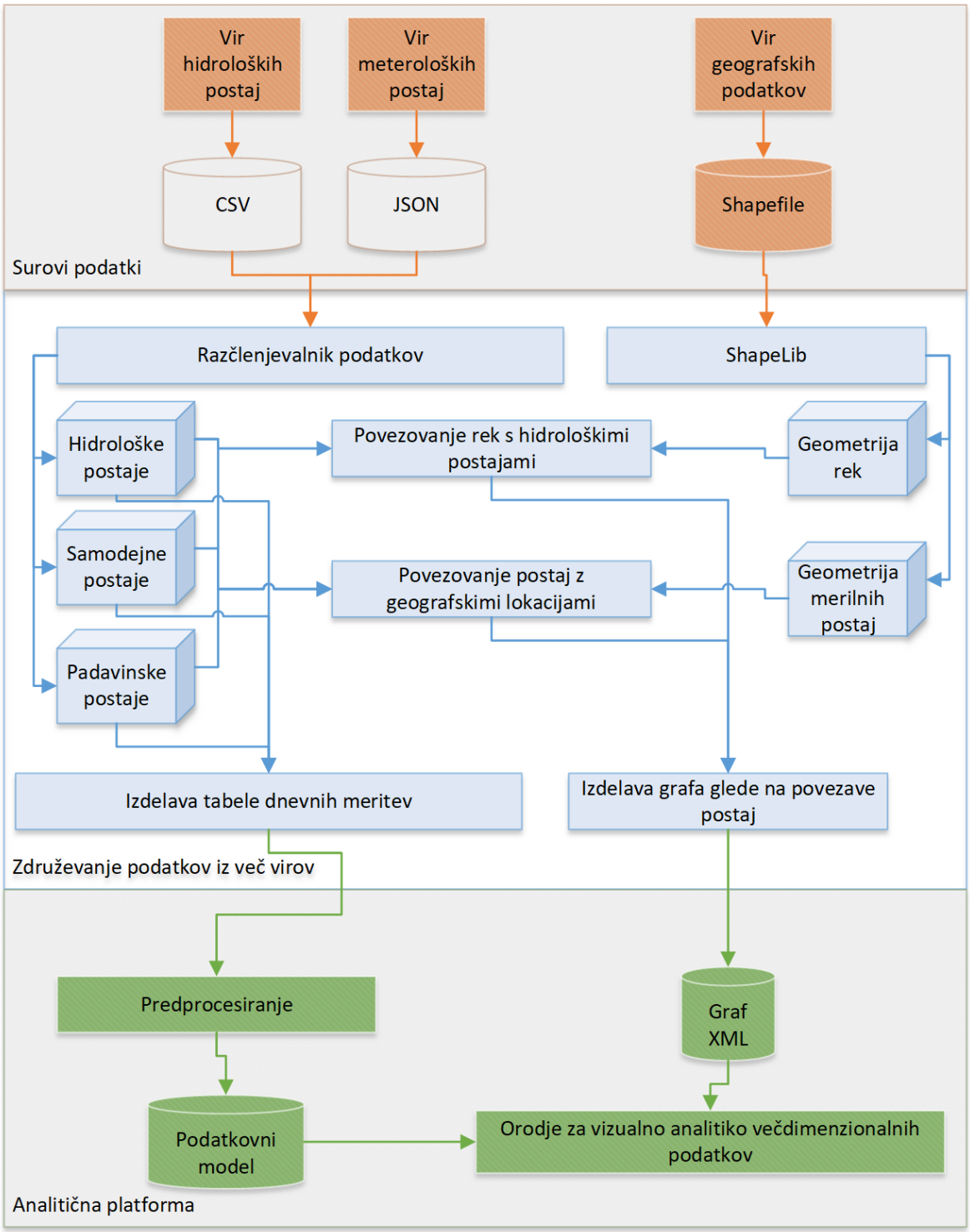
- **Samodejne postaje**, ki zajemajo visokoločljivostne meritve meteoroloških spremenljivk, kot so zračni tlak, temperatura, relativna vlaka, količina padavin, hitrost in smer vetra. Časovna ločljivost teh meritev je polurna. Ključni podatek pa je v našem primeru bila količina padavin.
- **Padavinske postaje**, ki zajemajo dnevne meritve meteorološkega stanja in vsoto količine padavin. Razen količine padavin je večina zbranih meritev opisnih. Količino padavin smo uporabili v nadaljnji analizi.

Slika 5.1: Prikaz območja pogostejših poplav v Celju. Puščica označuje geografsko lokacijo opazovane hidrološke postaje (6140) na reki Savinji.

- **Postaje hidroloških meritev**, ki zajemajo meritve dnevnih vrednosti pretokov, vodostajev in temperatur rek. Število različnih vrst meritev teh postaj je odvisno od tipa merilnika.

Postavitev samodejnih postaj med letoma 2009 in 2013 je bila redka v primerjavi z današnjim stanjem, zato smo si morali pomagati tudi s podatki padavinskih postaj. Iz hidrološkega arhiva smo izbrali meritve do vključno leta 2015 (po nekod le do 2013), kar je zadostovalo za namene naše študije.

Surove podatke smo najprej razčlenili in iz njih izluščili uporabne informacije. Pridobljene postaje smo obogatili z geografskimi in opisnimi informacijami, kot so položaj, ime in pripadajoči vodotok. Z geometrijo in topologijo rek smo tvorili usmerjen graf povezav med hidrološkimi postajami in nato iz podatkov pridobljenih postaj ustvarili tabelo časovnih vrst dnevne ločljivosti. Vsaka vrstica v vhodni tabeli tako opisuje meritve iz različnih postaj v istem časovnem intervalu, stolpci tabele pa podajajo meritev in metapodatke o meritvi sami (ti so pretok, vodostaj in količina padavin). Iz povezav rek in hidroloških postaj smo nato izvozili še mrežo povezav med postajami v obliki datoteke *XML*. Celoten postopek od pridobitve podatkov do obdelave in nalaganja v naše orodje prikazuje slika 5.2. Po končani predobde-



Slika 5.2: Postopek predobdelave meritev, kjer na vhodu prejmemo neobdelane podatke, jih strukturiramo v podatkovne tabele in v tej obliki naložimo v predstavljeno aplikacijo.

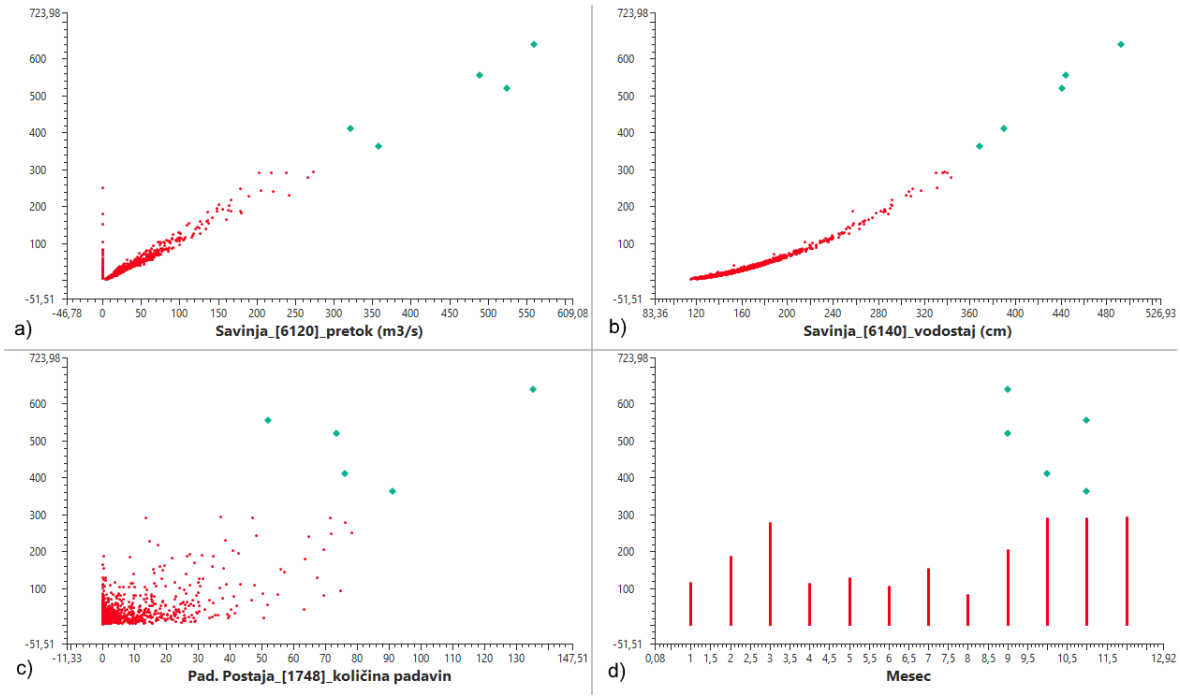
lavi smo v predstavljeno orodje naložili nabor podatkov, ki skupaj zajema 1826 vzorcev s 132 spremenljivkami. Interval pretoka smo razdelili na dva razreda $[4 - 321)$ in $[321 - 639]$, pri čemer v slednjega pade 5 vzorcev.

5.1 Iskanje odvisnosti

S predstavljenim orodjem smo najprej poiskali spremenljivke posameznih postaj, ki so korelirane s pretoki postaje 6140. Najvišje korelacije s slednjimi so imeli pretoki in vodostaji ostalih hidroloških postaj. Ta je znašala tudi več kot 95 %, glede na Pearsonov koeficient. Količina padavin pri tem ni imela neposrednega vpliva na opazovan pretok. To primerjavo prikazuje tabela 5.1, medtem ko je sama razpršenost vzorcev prikazana na sliki 5.3.

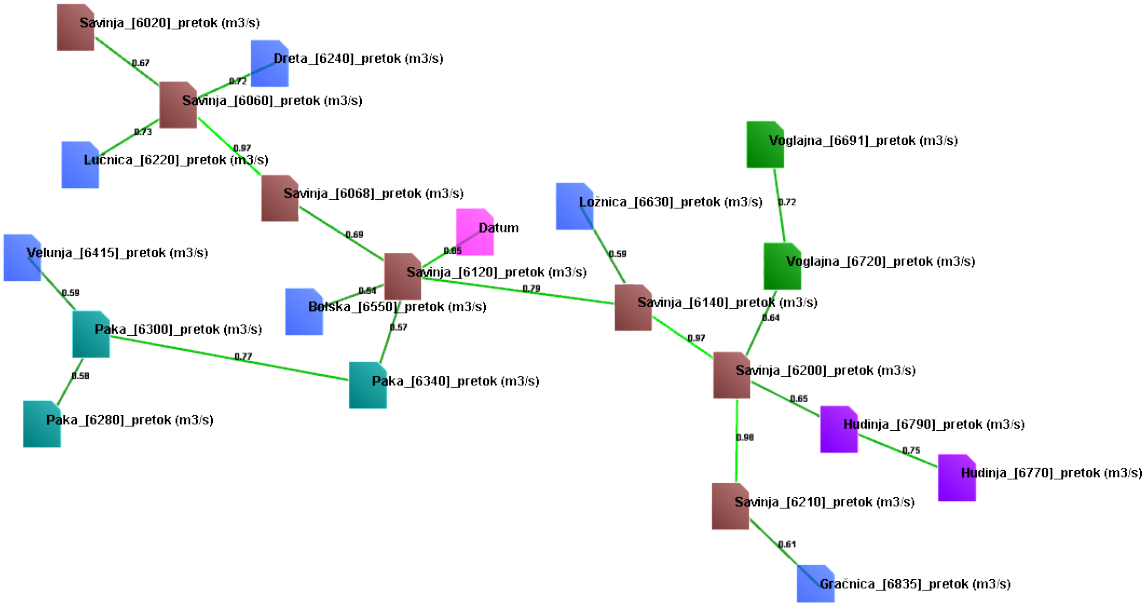
Tabela 5.1: Prikaz ocen različnih metod korelacij izbranih spremenljivk s pretokom hidrološke postaje 6140.

Postaja	Spremenljivka	Pearson	Spearman	Distance	RDC	MIC
6200	Pretok	0.99	0.99	0.99	0.99	0.97
6210	Pretok	0.98	0.98	0.98	0.99	0.94
6140	Vodostaj	0.95	0.88	0.97	0.99	0.92
6120	Pretok	0.95	0.86	0.91	0.94	0.79
6020	Pretok	0.77	0.86	0.78	0.87	0.61
1748	Količina padavin	0.67	0.42	0.54	0.47	0.21
2471	Količina padavin	0.59	0.38	0.49	0.43	0.21



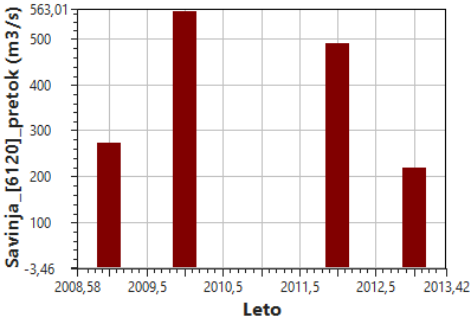
Slika 5.3: Prikaz odvisnosti vrednosti vzorcev med pretokom postaje 6140 in ostalimi postajami, kjer zelena barva prikazuje visoke vrednosti pretoka, rdeča pa nizke. Graf raztrosa a) prikazuje razpršenost vzorcev s pretokom sosednje postaje 6120, b) razpršenost vzorcev z vodostajem na postaji 6140, c) razpršenost vzorcev s količino padavin postaje 1748 in d) najvišje pretoke po mesecih.

V orodje smo nato uvozili še graf topoloških povezav pretokov in naredili analizo odvisnosti med povezanimi postajami (Slika 5.4). Analiza nakazuje na linearno odvisnost med vsemi



Slika 5.4: Graf korelacije med posameznimi pretoki za reko Savinjo od prve do zadnje merilne postaje (od leve proti desni). Povezave so bile tvorjene izven našega orodja in predstavljajo dejanske povezave med merilnimi postajami. Svetlejše povezave predstavljajo višjo oceno korelacije, temnejše pa nižjo.

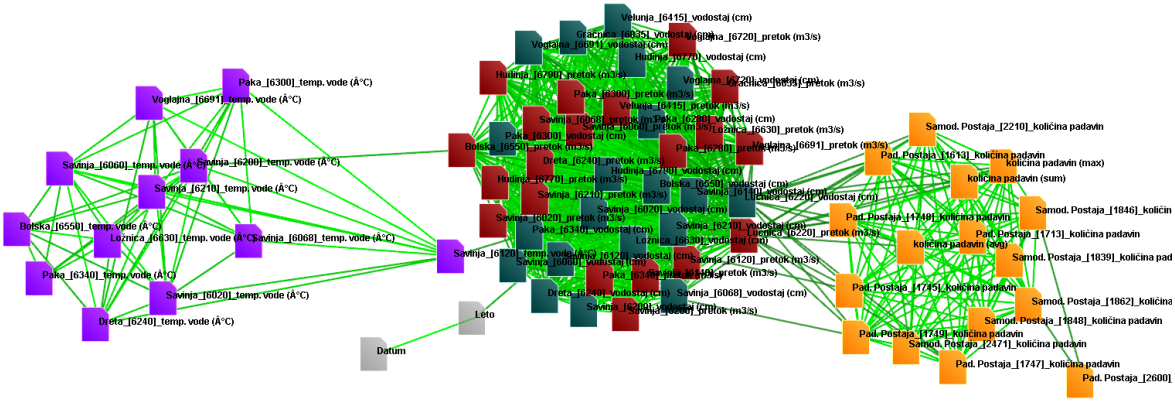
izmerjenimi pretoki razen s postajo 6120. Razlog za to je šum v podatkih, saj merilnik na postaji v letu 2011 ni deloval (Slika 5.5). V naslednjem koraku smo želeli ugotoviti še ostale



Slika 5.5: Prikaz grafikona najvišjih vrednosti pretokov po letih za postajo 6120, kjer je razvidno, da za leto 2011 ni podatkov.

spremenljivke, ki lahko vplivajo na pretoke. V ta namen smo izvedli analizo odvisnosti vhodnih spremenljivk s ciljno spremenljivko s pomočjo grafa povezav. Slednje je zanimivo predvsem s stališča samodejne tvorbe povezav med spremenljivkami enakega tipa (Slika 5.6). V orodju smo spremenljivke nato združili v štiri skupine (pretoki, vodostaji, količina padavin

in temperature) in nad njimi izvedli multivariabilno analizo s pretoki različnih postaj na reki Savinji. Rezultate prikazuje tabela 5.2. Pri tem pa lahko opazimo, da identifikacija ključnih spremenljiv ni bila očitna, saj so bile ocene multivariabilne odvisnosti nad različnimi pretoki precej podobne. To nakazuje na redundanco v podatkih.



Slika 5.6: Prikaz rezultata samodejne analize medsebojne odvisnosti. Barve predstavljajo štiri tipe spremenljivk. Ti so Pretoki (rdeča), Vodostaji (modro zelena), Temperature rek (vijolična) in Količina padavin (oranžna barva).

Tabela 5.2: Ocene multivariabilne analize s pretoki postaj na reki Savinji.

Postaja	Pretoki	Vodostaji	Količina padavin	Temperature
6020	0.79	0.77	0.68	0.81
6060	0.75	0.78	0.74	0.8
6068	0.75	0.78	0.69	0.81
6120	0.75	0.77	0.75	0.81
6140	0.74	0.78	0.77	0.78
6200	0.77	0.78	0.74	0.79
6210	0.76	0.77	0.74	0.78

5.2 Validacija

Poleg uporabniške validacije, predstavljene v prejšnjem poglavju, smo v okviru te magistrske naloge izvedli tudi validacijo implementiranih statističnih metrik. To smo izvedli z referenčnimi implementacijami v programskem jeziku R, pri tem pa smo uporabili podatke o pretokih rek iz prejšnjega poglavja. Najprej smo opravili analizo osnovnih statističnih metod nad tremi različnimi spremenljivkami, ki se med seboj statistično razlikujejo. Dobljene rezultate smo

nato med seboj primerjali z enotskimi testi (ang. unit tests). Pri tem smo ugotovili, da so naši rezultati bili identični rezultatom referenčnih implementacij v programskem jeziku R (Tabela 5.3). V naslednjem koraku smo se odločili opraviti analizo odvisnosti pretoka po-

Tabela 5.3: Primerjava rezultatov analize osnovne statistike nad izbranimi spremenljivkami referenčnih implementacij v R in v predstavljenem orodju.

Statistika	Količina padavin 2471		Temp. vode 6240		Sunki vetra 1848	
	R	Orodje	R	Orodje	R	Orodje
Povprečje	2.9	2.9	9.58	9.58	346.12	346.12
Mediana	0	0	9.4	9.4	307.95	307.95
Modus	0	0	0	0	0	0
Varianca	61.02	61.02	26.85	26.85	25533.32	25533.32
Standardni odklon	7.81	7.81	5.18	5.18	159.79	159.79
Standardna napaka	0.18	0.18	0.12	0.12	3.74	3.74
Min.	0	0	0	0	0	0
Maks.	116.7	37.7	21.8	21.8	1024	1024
Nagnjenost	5.18	5.18	0.05	0.05	0.99	0.99
Sploščenost	44.65288	44.65288	2.09	2.09	4.01	4.01
Frekv. ničel	1131	1131	52	52	52	52

staje 6140 s tremi izbranimi spremenljivkami, pri katerih je vsaka definirana z različnimi karakteristikami. Te spremenljivke so:

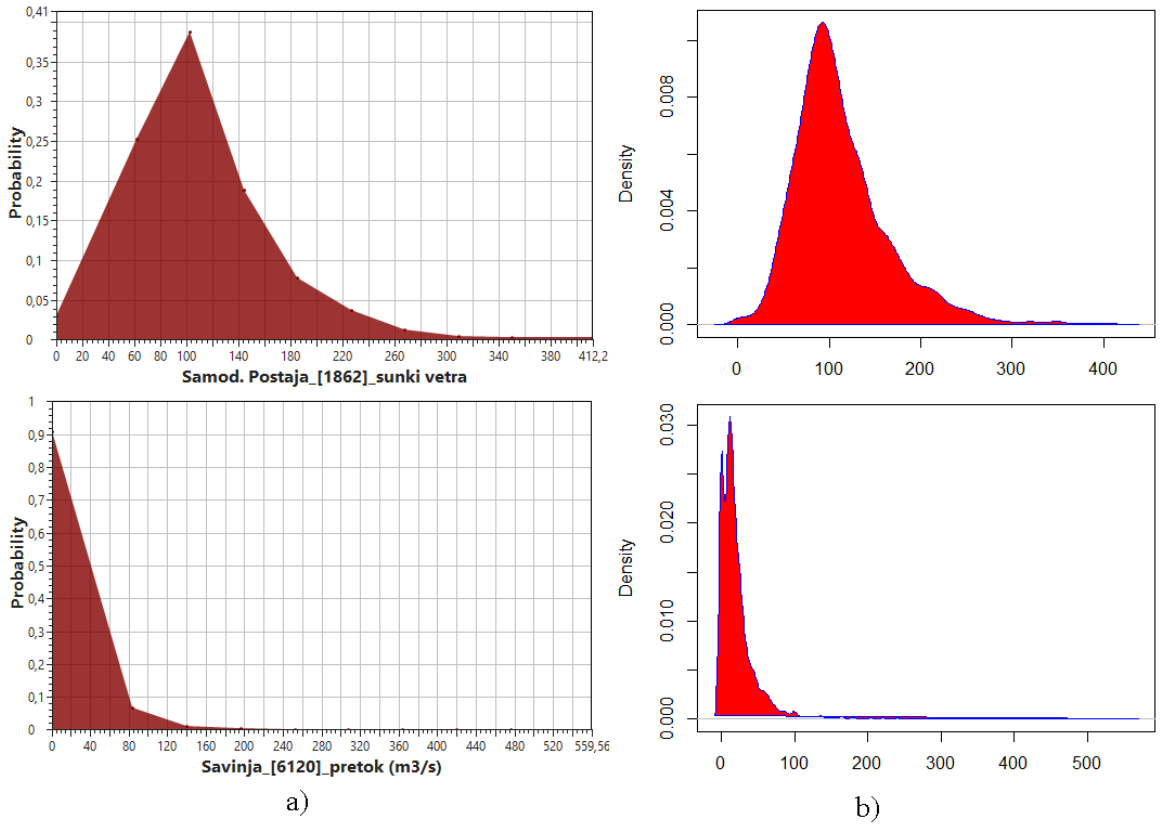
- **Pretok na postaji 6200**, ki je v našem orodju predstavljen kot najbolj koreliran pretok s ciljnim pretokom.
- **Pretok na postaji 6120**, ki v podatkih vsebuje šum.
- **Hitrost vetra na postaji 1862**, ki ima šibke negativne korelacije s ciljnim pretokom.

Analizo odvisnosti smo opravili zgolj z metodami za izračun korelacij, za katere obstajajo referenčne implementacije v programskem jeziku R. Te metode so Pearsonov koeficient, Spearmanov koeficient, korelacija razdalje in RDC. Z uporabo naštetih metod smo izvedli analizo odvisnosti za vsako izmed omenjenih spremenljivk. Primerjavo dobljenih rezultatov prikazuje tabela 5.4. Razlike v rezultatih so se pojavile pri metodah *Spearman* in *RDC*. Razlog slabih rezultatov slednje je napačno delovanje referenčne implementacije v zadnji različici programskega jezika R. Nazadnje smo preverili še vizualizacijo funkcij gostot verjetnosti.

Tabela 5.4: Primerjava rezultatov korelacij

Korelacije	Pretok 6200		Pretok 6120		Hitrost vetra 1862	
	R	**Orodje**	**R**	**Orodje**	**R**	**Orodje**
Pearson	0.99	0.99	0.99	0.95	-0.03	-0.03
Spearman	0.99	0.99	0.81	0.86	-0.09	-0.08
Distance	0.99	0.99	0.91	0.91	0.07	0.07
RDC	0.71	0.99	0.98	0.94	0.86	0.13

Programski jezik R namreč omogoča tako njihovo gradnjo, kakor tudi vizualizacijo z uporabo orodja RStudio [31]. Primerjavo smo opravili nad celotno množico vzorcev izbrane spremenljivke brez delitve na razrede. Do razlik v vizualizaciji pride zaradi načina gradnje funkcije gostote verjetnosti, saj jo v našem orodju privzeto omejimo na 10 podintervalov, medtem ko pa je v R določena dinamično.



Slika 5.7: Prikaz primerjave funkcij gostot verjetnosti v a) našem orodju in b) orodju RStudio.

Poglavje 6

Sklep

V magistrskem delu smo predstavili orodje za vizualno analitiko večdimenzionalnih podatkov. V orodju smo implementirali več statističnih metod in algoritmov, ki omogočajo analize odvisnosti izbranih spremenljivk. Implementirali smo tudi različne metode upodobitve rezultatov analiz v obliki grafikonov, funkcij gostot verjetnosti, drevesnih struktur in grafov, ki so namenjeni odkrivanju novega znanja v podatkih. Z demostracijo funkcionalnosti orodja smo pokazali, da predlagane metode omogočajo odkrivanje novih znanj v različnih domenah, ki vključujejo tako statično razpoznavo vzorcev, kakor tudi analizo časovnih vrst. Skozi validacijo smo pokazali pravilnost delovanja implementiranih funkcionalnosti, medtem ko smo samo učinkovitost predlaganih rešitev demonstrirali s primeri.

Literatura

[1] IBM SPSS Software. Dostopno na: https://www.ibm.com/analytics/us/en/technology/spss/. [08.10.2017].

[2] Analyse-it. Dostopno na: https://analyse-it.com/. [08.10.2017].

[3] Statistics and Machine Learning Toolbox. Dostopno na: https://www.mathworks.com/products/statistics.html. [08.10.2017].

[4] The R Project for Statistical Computing. Dostopno na: https://www.r-project.org/. [08.10.2017].

[5] Weka. Dostopno na: https://www.cs.waikato.ac.nz/ml/weka/. [08.10.2017].

[6] R., Stuart in P., Norvig. *Artificial Intelligence: A Modern Approach*. Pearson Education, Inc., Upper Saddle River, New Jersey 07458., 2009. str. 693 - 852.

[7] François D., Wertz V. in M., Verleysen. The Concentration of Fractional Distances. *IEEE Transactions on Knowledge and Data Engineering*, 19(7):873 - 886, 2007.

[8] I., Smith Lindsay. A tutorial on Principal Components Analysis. *Cornell University, USA,*, 202.

[9] O., Maimon in L., Rokach. *Data Mining and Knowledge Discovery Handbook*. Springer, 2010. str. 269 - 295.

[10] Nguyen, Q. V. in Huang, M. L. A Space-Optimized Tree Visualization. *IEEE Symposium on Information Visualization*, 2002.

[11] JOGL. Dostopno na: http://jogamp.org/jogl/www/. [08.10.2017].

[12] Standard Widget Toolkit. Dostopno na: https://www.eclipse.org/swt/. [08.10.2017].

[13] Asimetrija in sploščenost. Dostopno na: http://www.benstat.si/blog/koeficient-asimetrije-sploscenosti. [08.10.2017].

[14] Interval zaupanja. Dostopno na: http://stattrek.com/estimation/confidence-interval.aspx. [08.10.2017].

[15] T-Test. Dostopno na: www.itl.nist.gov/div898/handbook/eda/section3/eda353.htm. [08.10.2017].

[16] E., Heron. Analysis of Variance. Dostopno na: www.itl.nist.gov/div898/handbook/eda/section3/eda353.htm, 2009. [08.10.2017].

[17] G., Hall. Pearson's correlation coefficient. Dostopno na: www.hep.ph.ic.ac.uk/~hallg/UG_2015/Pearsons.pdf, 2015. [08.10.2017].

[18] Spearman's correlation. Dostopno na: www.statstutor.ac.uk/resources/uploaded/spearmans.pdf. [08.10.2017].

[19] Székely, Rizzo M. L., G. J. in Bakirov, N. K. Measuring and testing dependence by correlation of distances. *The annals of statistics*, 35(6):2769-2794., 2007.

[20] Reshef D. N., Y. A. Finuncane H. K. Grossman S. R. McVean G. Turnbaugh P. J. Lander E. S. Mitzenmacher M., Reshef in Sabeti, P. C. Detecting novel associations in large data sets. *Science*, 334(6062):1518-1524, 2011.

[21] Lopez-Paz, Hennig P., D. in Schölkopf, B. The randomized dependence coefficient. *In Advances in neural information processing systems*, strani 1-9, 2013.

[22] Cruz, I. F. in Tamassia, R. Graph drawing tutorial, 1998. Dostopno na: www.cs.brown.edu/rt/papers/gd-tutorial/gd-constraints.pdf.

[23] Kobourov, S. G. Force-directed drawing algorithms, 2004.

[24] Nguyen, Müller E. Vreeken J. Efros P., H. V. in Böhm, K. Multivariate maximal correlation analysis. *In Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, strani 775 - 783, 2014.

[25] Fisher, R. A. Iris Data Set. Dostopno na: `https://archive.ics.uci.edu/ml/datasets/Iris`. [08.10.2017].

[26] Kohavi, R. in Becker, B. Adult Data Set. Dostopno na: `https://archive.ics.uci.edu/ml/datasets/adult`. [08.10.2017].

[27] Opazovani in merjeni meteorološki podatki po Sloveniji. Dostopno na: `http://meteo.arso.gov.si/met/sl/archive/`. [08.10.2017].

[28] Arhiv hidroloških podatkov. Dostopno na: `http://vode.arso.gov.si/hidarhiv/pov_arhiv_tab.php`. [08.10.2017].

[29] Hidrološko poročilo o povodnji v dneh od 17. do 21. septembra 2010. Dostopno na: `http://www.arso.gov.si/vode/poročila%20in%20publikacije/Poplave%2017.%20-%2021.%20september%202010.pdf`, September 2010. [08.10.2017].

[30] Hidrološko poročilo o poplavah v dneh med 4. in 6. novembrom 2012. Dostopno na: `www.arso.gov.si/vode/poročila%20in%20publikacije/Poplave%205.%20-%206.%20november%202012.pdf`, November 2012. [08.10.2017].

[31] RStudio. Dostopno na: `https://www.rstudio.com/`. [08.10.2017].

ZA ŠTUDENTE MEH

Univerza v Mariboru

Fakulteta za elektrotehniko,
računalništvo in informatiko
Koroška cesta 46
2000 Maribor, Slovenija

IZJAVA O USTREZNOSTI ZAKLJUČNEGA DELA

Podpisani mentor/-ica : Domen Mongus
(ime in priimek mentor-ja/-ice)

in somentor/-ica (eden ali več, če obstajajo): Niko Lukač
(ime in priimek somentor-ja/-ice)

Izjavlja-m/-va/-mo, da je študent/-ka

Ime in priimek: Matej Brumen, *ID številka:* 1001983343,

vpisna številka: E5017726, *na študijskem programu:*

Računalništvo in informacijske tehnologije (MAG)

izdelal/-a zaključno delo z naslovom:

Orodje za vizualno analitiko večdimenzionalnih podatkov

(naslov zaključnega dela v slovenskem jeziku)

v skladu z odobreno temo zaključnega dela, navodili o pripravi zaključnih del in mojimi (najinimi/našimi) navodili.

Preveril/-a/-i in pregledal/-a/-i sem/sva/smo poročilo o preverjanju podobnosti vsebin z drugimi deli (priloga) in potrjujem/potrjujeva/potrjujemo, da je zaključno delo ustrezno.

Datum in kraj:
9.10.2017, Maribor

Podpis mentor-ja/-ice:

Datum in kraj:
9.10.2017, Maribor

Podpis somentor-ja/-ice (če obstaja):

Priloga:
- Poročilo o preverjanju podobnosti vsebin z drugimi deli.

Univerza v Mariboru

Fakulteta za elektrotehniko,
računalništvo in informatiko

IZJAVA O AVTORSTVU IN ISTOVETNOSTI TISKANE IN ELEKTRONSKE OBLIKE ZAKLJUČNEGA DELA

Ime in priimek študent-a/-ke: Matej Brumen
Študijski program: Računalništvo in informacijske tehnologije (MAG)
Naslov zaključnega dela: Orodje za vizualno analitiko večdimenzionalnih podatkov

Mentor: doc. dr. Domen Mongus, univ. dipl. inž. rač. in inf.
Somentor: asist. dr. Niko Lukač, univ. dipl. inž. rač. in inf.

Podpisan-i/-a študent/-ka Matej Brumen

- izjavljam, da je zaključno delo rezultat mojega samostojnega dela, ki sem ga izdelal/-a ob pomoči mentor-ja/-ice oz. somentor-ja/-ice;
- izjavljam, da sem pridobil/-a vsa potrebna soglasja za uporabo podatkov in avtorskih del v zaključnem delu in jih v zaključnem delu jasno in ustrezno označil/-a;
- na Univerzo v Mariboru neodplačno, neizključno, prostorsko in časovno neomejeno prenašam pravico shranitve avtorskega dela v elektronski obliki, pravico reproduciranja ter pravico ponuditi zaključno delo javnosti na svetovnem spletu preko DKUM; sem seznanjen/-a, da bodo dela deponirana/objavljena v DKUM dostopna široki javnosti pod pogoji licence Creative Commons BY-NC-ND, kar vključuje tudi avtomatizirano indeksiranje preko spleta in obdelavo besedil za potrebe tekstovnega in podatkovnega rudarjenja in ekstrakcije znanja iz vsebin; uporabnikom se dovoli reproduciranje brez predelave avtorskega dela, distribuiranje, dajanje v najem in priobčitev javnosti samega izvirnega avtorskega dela, in sicer pod pogojem, da navedejo avtorja in da ne gre za komercialno uporabo;
- dovoljujem objavo svojih osebnih podatkov, ki so navedeni v zaključnem delu in tej izjavi, skupaj z objavo zaključnega dela;
- izjavljam, da je tiskana oblika zaključnega dela istovetna elektronski obliki zaključnega dela, ki sem jo oddal/-a za objavo v DKUM.

Uveljavljam permisivnejšo obliko licence Creative Commons:_____(navedite obliko)

Začasna nedostopnost:

Zaključno delo zaradi zagotavljanja konkurenčne prednosti, zaščite poslovnih skrivnosti, varnosti ljudi in narave, varstva industrijske lastnine ali tajnosti podatkov naročnika:
_____(naziv in naslov naročnika/institucije) ne sme biti javno dostopno do _____(datum odloga javne objave ne sme biti daljši kot 3 leta od zagovora dela). To se nanaša na tiskano in elektronsko obliko zaključnega dela.

Temporary unavailability:

To ensure competition priority, protection of trade secrets, safety of people and nature, protection of industrial property or secrecy of customer's information, the thesis _____(institution/company name and address) must not be accessible to the public till_____(delay date of thesis availability to the public must not exceed the period of 3 years after thesis defense). This applies to printed and electronic thesis forms.

Datum in kraj: 9.10.2017, Maribor

Podpis študent-a/-ke:

Podpis mentor-ja/-ice: _____
(samo v primeru, če delo ne me biti javno dostopno)

Ime in priimek ter podpis odgovorne osebe naročnika in žig:

_____
(samo v primeru, če delo ne sme biti javno dostopno)

ZA ŠTUDENTE GING

Univerza v Mariboru

Fakulteta za elektrotehniko, računalništvo in informatiko

Ekonomsko-poslovna fakulteta

IZJAVA O OBJAVI OSEBNIH PODATKOV

Ime in priimek diplomant-a/ magistrant-/-ke: Matej Brumen

ID številka: 1001983343

Študijski program: Računalništvo in informacijske tehnologije (MAG)

Naslov zaključnega dela: Orodje za vizualno analitiko večdimenzionalnih podatkov

Mentor/-ica FERI: Domen Mongus

Somentor/-ica: Niko Lukač

Podpisan-i/-a izjavljam, da dovoljujem objavo osebnih podatkov, vezanih na zaključek študija (ime, priimek, leto zaključka študija, naslov zaključnega dela) na spletnih straneh Univerze v Mariboru in v publikacijah Univerze v Mariboru.

Datum in kraj:
9.10.2017, Maribor

Podpis diploman-ta/magistran-ta/-ke:

Brumen