Evaluacija tehnika otkrivanja kontrasta za potrebe selekcije atributa radi klasifikacije

Oreški, Dijana

Doctoral thesis / Disertacija

2014

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Organization and Informatics Varaždin / Sveučilište u Zagrebu, Fakultet organizacije i informatike Varaždin**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:211:337330

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2024-03-28**

SVEUČILIŠTE U ZAGREBU
FAKULTET ORGANIZACIJE I INFORMATIKE
VARAŽDIN

Repository / Repozitorij:

Faculty of Organization and Informatics - Digital Repository

DIGITALNI AKADEMSKI ARHIVI I REPOZITORIJI



Sveučilište u Zagrebu

Fakultet organizacije i informatike

Dijana Oreški

EVALUACIJA TEHNIKA OTKRIVANJA KONTRASTA ZA POTREBE SELEKCIJE ATRIBUTA RADI KLASIFIKACIJE

DOKTORSKI RAD

Varaždin, 2014.

PODACI O DOKTORSKOM RADU

I. AUTOR

Ime i prezime	Dijana Oreški
Datum i mjesto rođenja	16.05.1986., Varaždin
Naziv fakulteta i datum diplomiranja na VII/II stupnju	Fakultet organizacije i informatike, 04.12.2008.
Sadašnje zaposlenje	Fakultet organizacije i informatike, asistent

II. DOKTORSKI RAD

Naslov	Evaluacija tehnika otkrivanja kontrasta za potrebe selekcije atributa radi klasifikacije
Broj stranica, slika, tabela, priloga, bibliografskih podataka	156 stranica, 37 slika, 20 tablica, 32 stranice priloga i 103 bibliografska izvora
Znanstveno područje i polje iz kojeg je postignut doktorat znanosti	Društvene znanosti, Informacijske i komunikacijske znanosti
Mentori ili voditelji rada	Prof. dr.sc. Božidar Kliček Prof. dr.sc. Dunja Mladenić
Fakultet na kojem je obranjen doktorski rad	Fakultet organizacije i informatike
Oznaka i redni broj rada	

III. OCJENA I OBRANA

Datum sjednice Fakultetskog vijeća na kojoj je prihvaćena tema	16.10.2012.
Datum predaje rada	03.12.2013.
Datum sjednice Fakultetskog vijeća na kojoj je prihvaćena pozitivna ocjena rada	14.01.2014.
Sastav povjerenstva koje je rad ocijenilo	dr.sc. Dragan Gamberger Prof.dr.sc. Božidar Kliček Prof.dr.sc. Dunja Mladenić Prof.dr.sc. Vesna Dušak Prof.dr.sc. Mirko Maleković
Datum obrane doktorskog rada	06.02.2014.
Sastav povjerenstva pred kojim je rad obranjen	dr.sc. Dragan Gamberger Prof.dr.sc. Božidar Kliček Prof.dr.sc. Dunja Mladenić Prof.dr.sc. Vesna Dušak Prof.dr.sc. Mirko Maleković
Datum promocije	



Sveučilište u Zagrebu

Fakultet organizacije i informatike

DIJANA OREŠKI

EVALUACIJA TEHNIKA OTKRIVANJA KONTRASTA ZA POTREBE SELEKCIJE ATRIBUTA RADI KLASIFIKACIJE

DOKTORSKI RAD

Mentori: prof. dr. sc. Božidar Kliček
prof.dr. sc. Dunja Mladenić

Varaždin, 2014.



University of Zagreb

Faculty of Organization and Informatics

Dijana Oreški

EVALUATION OF CONTRAST MINING TECHNIQUES FOR FEATURE SELECTION IN CLASSIFICATION

DOCTORAL THESIS

Varaždin, 2014.

ZAHVALE

Na početku ovog rada želim zahvaliti svima koji su doprinijeli njegovom stvaranju. Iskrenu zahvalnost izražavam svojim mentorima, prof. dr.sc. Božidaru Kličeku i prof.dr.sc. Dunji Mladenić, na znanstvenom i stručnom usmjeravanju, poticanju, nesebičnom pomaganju i vođenju kod izrade ovog rada. Privilegirana sam što sam učila od Vas.

Najtoplije zahvaljujem svojem stricu mr.sc. Stjepanu Oreški na stručnim savjetima još od samih početaka pisanja disertacije te sugestijama i pomoći u cijelom procesu mog znanstvenog sazrijevanja.

Veliko hvala ide mojim dragim kolegama i prijateljima Jeleni Horvat, mag.oec., Ireni Kedmenec, mag.oec. i Nikoli Kadoiću, mag.inf. na pomoći u organizaciji posla, korisnim savjetima i podršci koju su mi svakodnevno pružali.

Naročito se zahvaljujem prijateljicama Dunji, Mariji, Renati i Vlatki na prijateljstvu i konstantnoj podršci.

Posebno zahvaljujem svojoj obitelji: roditeljima, sestri i baki jer su mi uvijek pružali podršku tijekom školovanja i svakodnevno me motivirali za rad. Zahvaljujem im na strpljenju, razumijevanju i velikoj količini ljubavi. Bez njih ne bi bilo ni ovog rada.

Naposljetku se zahvaljujem i svima ostalima koji nisu poimence spomenuti, a pomogli su mi kod pisanja ovog rada.

Sažetak: Zadnja dva desetljeća ogroman je porast količine podataka koji se pohranjuju u digitalnom formatu. Zahvaljujući današnjim tehnologijama prikupljanje podataka prestaje biti problem, a u fokus zanimanja dolazi njihova analiza i dobivanje vrijednih informacija iz podataka (znanja). Centralno za taj problem je proces otkrivanja znanja u podacima. Proces se sastoji od nekoliko koraka, a priprema podataka koja obuhvaća čišćenje podataka i selekciju atributa oduzima 60% - 95% ukupnog vremena cijelog procesa. Svrha istraživanja je definirati nove tehnike za korištenje u selekciji atributa s ciljem smanjenja vremena potrebnog za provođenje selekcije atributa, a time i otklanjanja uskog grla cjelokupnog procesa otkrivanja znanja u podacima. S tim ciljem definiraju se dvije nove tehnike selekcije atributa koja pripadaju grupi tehnika otkrivanja kontrasta. Predmet istraživanja predstavlja primjena i evaluacija tehnika otkrivanja kontrasta kao tehnika za selekciju atributa. U tu svrhu provodi se opsežno istraživanje s ciljem utvrđivanja na skupovima podataka kojih karakteristika tehnike otkrivanja kontrasta nadmašuju klasične tehnike selekcije atributa, te dobivanja općeg odgovora da li se ubuduće mogu i u kojim situacijama tehnike otkrivanja kontrasta koristiti kao superiorne tehnike selekcije aributa, i mogu li u znatnoj mjeri otkloniti usko grlo cijelokupnog procesa otkrivanja znanja u podacima.
Rezultati 1792 analize pokazuju da su u više od 80% analiziranih skupova podataka različitih karakteristika tehnike otkrivanja kontrasta rezultatirale točnijom klasifikacijom i brže provedenim procesom otkrivanja znanja nego dosad korištene tehnike.

Ključne riječi: rudarenje podataka, selekcija atributa, tehnike otkrivanja kontrasta, klasifikacija, karakteristike skupa podataka

Abstract: The last two decades there is a huge increase in the amount of data that is stored in digital format. Owing to today's technology data collection ceases to be a problem and in the focus of interest is their analysis and obtaining valuable information from the data (knowledge). Central for this issue is the process of knowledge discovery in data. The process consists of several steps and preparation of the data, which includes data cleaning and feature selection takes away from 60% till 95% total time of the whole process. The purpose of this research is to define new techniques for feature selection in order to reduce the time required to conduct feature selection, and thus removing the bottleneck of the entire process of knowledge discovery in data. For this purpose two new techniques of feature selection are defined, techniques that belong to the group of contrast mining field. The subject of this research is an application and evaluation of contrast mining techniques as a techniques for feature selection. The extensive empirical research is conducted in order to determine data sets characteristics for which contrast mining techniques outperform classical techniques of feature selection, and obtaining general answer can we, and in what kind of data sets, use contrast mining techniques as a superior feature selection techniques, and whether they can eliminate the bottleneck of the entire process of knowledge discovery in data.

Results of 1792 analysis showed that in the more than 80% of the analyzed data sets with different characteristics contrast mining techniques resulted with more accurate classification and quickly implemented process of knowledge discovery than previously used feature selection techniques.

Keywords: data mining, feature selection, contrast mining, classification, data set characteristis

SADRŽAJ

POPIS SLIKA III
POPIS TABLICA V
1. UVOD 1
1.1.Problem istraživanja 1
1.2.Svrha i ciljevi istraživanja 2
1.3.Hipoteze istraživanja 3
1.4.Metodologija istraživanja 4
1.5.Struktura rada 6
2. PROCES OTKRIVANJA ZNANJA U PODACIMA 7
2.1. Definicija procesa 7
2.2. Klasifikacija 9
2.2.1. Diskriminacijska analiza 10
2.2.2. Neuronske mreže 13
2.2.3. Komparacija klasifikatora 17
2.3. Evaluacija rezultata 17
2.3.1. Unakrsno vrednovanje 19
2.3.2. Testiranje statističke značajnosti 20
3. TEHNIKE SELEKCIJE ATRIBUTA 22
3.1. Definicija selekcije atributa 23
3.2.Pristupi selekciji atributa 24
3.2.1.Tehnike filtra 28
3.2.2.Tehnike omotača 28
3.3. Opis odabranih tehnika selekcije atributa 29
3.3.1. Informacijska dobit 30
3.3.2. Omjer dobiti 30
3.3.3. Tehnika *Relief* 31
3.3.4. Tehnika *linearni odabir unaprijed* 32
3.4. Pregled dosadašnjih istraživanja 34
3.4.1.Nedostaci prethodnih istraživanja 37
4. TEHNIKE OTKRIVANJA KONTRASTA 39
4.1. STUCCO algoritam 40

4.2. Magnum Opus 42
4.3. Prethodna istraživanja 45
4.4.Usporedba tehnika STUCCO i Magnum Opus 47
4.5. Diskusija o tehnikama otkrivanja kontrasta 47
5. KARAKTERISTIKE SKUPA PODATAKA 50
5.1. Standardne mjere 53
5.2. Mjere oskudnosti podataka 53
5.2.1. Mjera oskudnosti podataka 55
5.3. Statističke mjere 55
5.4. Mjere teorije informacija 57
5.5. Mjere šuma 57
6. DEFINICIJA TEHNIKA OTKRIVANJA KONTRASTA ZA SELEKCIJU ATRIBUTA 58
6.1. Mjere vrednovanja za selekciju podskupa atributa 58
6.2. Kriteriji rezanja za selekciju podskupa atributa 59
6.3.Selekcija atributa tehnikama otkrivanja kontrasta 62
7. EMPIRIJSKO ISTRAŽIVANJE 67
7.1. Karakterizacija skupova podataka 68
7.2. Selekcija atributa 85
7.2.1. Selekcija atributa dosad poznatim tehnikama 85
7.2.2. Selekcija atributa tehnikom Magnum Opus 88
7.2.3. Selekcija atributa tehnikom STUCCO 91
7.3. Klasifikacija 95
7.4. Vrijeme provođenja selekcije atributa 98
8. REZULTATI 100
8.1. Usporedba točnosti tehnika selekcije atributa - klasifikator neuronske mreže 101
8.2. Usporedba tehnika selekcije atributa – klasifikator diskriminacijska analiza 117
8.3. Usporedba rezultata dobivenih neuronskim mrežama i diskriminacijskom analizom 123
8.4. Usporedba tehnika selekcije atributa – vrijeme provedbe selekcija atributa 126
8.5. Ograničenja i preporuke za buduća istraživanja 143
9. ZAKLJUČAK 145
LITERATURA 148
PRILOZI 157

POPIS SLIKA

Slika 1. Model CRISP – DM 8
Slika 2. Struktura neuronske mreže 14
Slika 3. Koraci selekcije atributa 23
Slika 4. Princip selekcije atributa kod metoda filtra 27
Slika 5. Princip odabira atributa kod metoda omotača 27
Slika 6. Principi tehnike linearni odabir unaprijed 33
Slika 7. Pseudo kod STUCCO algoritma 42
Slika 8. Pseudo kod Magnum Opus algoritma 43
Slika 9. Fiksni broj atributa kao kriterij rezanja 60
Slika 10. Udio kao kriterij rezanja 60
Slika 11. Prag kao kriterij rezanja 60
Slika 12. Prag izražen kao udio kao kriterij rezanja 61
Slika 13. Razlika kao kriterij rezanja 61
Slika 14. Nagib kao kriterij rezanja 62
Slika 15. Dijagram tijeka selekcije atributa tehnikama otkrivanja kontrasta 64
Slika 16. *Pseudokod tehnike selekcije MOFS-a* 65
Slika 17. Pseudokod SfFS-a 66
Slika 18. Određivanje kategorija za karakteristiku *broj atributa* 70
Slika 19. Određivanje kategorija za karakteristiku *broj slučajeva* 71
Slika 20. Određivanje kategorija za karakteristiku *korelacija* 72
Slika 21. Određivanje kategorija za karakteristiku *omjer unutarnje dimenzionalnosti* 73
Slika 22. Postavke tehnike linearni korak unaprijed 85
Slika 23. Atributi selektirani tehnikom *linearni korak unaprijed* 86
Slika 24. Postavke tehnike informacijska dobit 86
Slika 25. Rezultati provedbe informacijske dobiti 86
Slika 26. Rezultati selekcije tehnikom omjer dobiti 87
Slika 27. Postavke Relief tehnike 87
Slika 28. Rezultati Relief tehnike 88
Slika 29. Postavke primjene *Magnum Opus FS* 88
Slika 30. Dopuštene vrijednosti u pravilima 89
Slika 31. Ovisnost točnosti klasifikacije neuronskim mrežama o karakteristikama skupa podataka 111
Slika 32. Doprinos karakteristika kod klasifikacije neuronskim mrežama 113

Slika 33. Ovisnost točnosti klasifikacije diskriminacijskom analizom o karakteristikama skupa podataka..... 121
Slika 34. Doprinos karakteristika podataka kod klasifikacije diskriminacijskom analizom. 122
Slika 35. Prosječne vrijednosti karakteristika..... 124
Slika 36. Ovisnost brzine izvođenja selekcije atributa o karakteristikama skupa podataka . 136
Slika 37. Doprinos karakteristika skupa podataka klasifikacija tehnika s obzirom na vrijeme provođenja selekcije atributa..... 137

POPIS TABLICA

Tablica 1. Matrica konfuzije 18
Tablica 2. Karakteristike podataka važne za klasifikaciju 51
Tablica 3. Standardne mjere 53
Tablica 4. Rezultati Box`s M testa 76
Tablica 5. Karakterizacija skupova podataka 78
Tablica 6. Srednja točnost neuronske mreže 96
Tablica 7. Friedman test 96
Tablica 8. Rangiranje tehnika selekcije za skup podataka *vote* 97
Tablica 9. Vrijeme provođenja selekcije atributa na skupu podataka 98
Tablica 10. Rangiranje tehnika na skupu *vote* prema brzini 98
Tablica 11. Statistika Friedman testa za brzinu 99
Tablica 12. Rangiranje tehnika selekcije atributa u klasifikaciji neuronskim mrežama 102
Tablica 13. Rangiranje tehnika selekcije u klasifikaciji diskriminacijskom analizom 118
Tablica 14. Rangiranje tehnika selekcije atributa s obzirom na brzinu izvođenja 126
Tablica 15. Povezanost karakteristika skupa podataka i vremena provođenja selekcije atributa 135
Tablica 16. Kvantitativne vrijednosti karakteristika skupa podataka 157
Tablica 17.Točnost neuronske mreže 166
Tablica 18. Točnost diskriminacijska analiza 174
Tablica 19. Vrijeme provođenja tehnika selekcije atributa 176
Tablica 20. Adrese skupova podataka 184

1. UVOD

Uvodno poglavlje rada detaljnije opisuje problem te ciljeve i hipoteze istraživanja, ukratko predstavlja metodologiju istraživanja te daju pregled strukture rada.

1.1.Problem istraživanja

Zadnja dva desetljeća ogroman je porast količine podataka koje se pohranjuju u digitalnom formatu (McKinsey Global Institute, 2011.). McKinsey Global Institute procjenjuje da se količina podataka u svijetu udvostručuje svakih dvadeset mjeseci (McKinsey Global Institute, 2011.). Jasno je da zahvaljujući današnjim tehnologijama prikupljanje podataka prestaje biti problem, a u fokus zanimanja dolazi njihova analiza i dobivanje vrijednih informacija iz podataka (znanja). Centralno za taj problem je proces otkrivanja znanja u bazama podataka (eng. *knowledge discovery in databases*). Proces otkrivanja znanja provodi se s ciljem postizanja jedne od slijedećih zadaća: klasifikacije, klasteriranja, vizualizacije, sumarizacije, detekcije devijacija ili procjene (Fayyad, Piatetsky-Shapiro i Smyth, 1996.). Klasifikacija se smatra temljenom zadaćom procesa otkrivanja znanja u podacima (Fayyad, Piatetsky-Shapiro i Smyth, 1996.) te je u fokusu interesa ovog rada.

Proces otkrivanja znanja u podacima sastoji se od nekoliko koraka, a priprema podataka koja obuhvaća čišćenje podataka i selekciju atributa oduzima 60% - 95% ukupnog vremena cijelog procesa (De Veaux, 2005.). Selekcija atributa, kao najvažniji dio toga koraka, odnosi se na problem odabira onih atributa koji daju najveću prediktivnu informaciju s obzirom na izlaz. To je problem koji se susreće u mnogim područjima i pronašao je veliku primjenu. Tehnike selekcije atributa imaju vrlo važnu ulogu jer dobra selekcija atributa u koraku pripreme podataka donosi višestruke koristi: smanjuje dimenzionalnost, miče irelevantne i redundantne atribute, olakšava razumijevanje podataka, smanjuje količinu podataka za učenje, poboljšava točnost predikcije algoritama i povećava interpretabilnost modela (Guyon i Elisseeff, 2003., Mladenić, 2006., Arauzo – Azofra, Aznarte i Benitez, 2011., Oreški, Oreški, i Oreški, 2012., Cadenas, Garrido, i Martinez, 2013.).

Unutar područja rudarenja podataka razvijeno je potpodručje otkrivanje kontrasta, (eng. *contrast mining*) kojemu je glavni cilj identifikacija onih atributa koji čine značajnu razliku između grupa. Tehnike otkrivanja kontrasta dosad su uspješno primjenjivane u mnogim

područjima: npr. analizi potrošačke košarice (Webb, Butler i Newlands, 2003.) i medicini (Kralj Novak et. al., 2009.) Vođeno tim rezultatima ovaj rad prepoznaje potencijal primjene tehnika otkrivanja kontrasta za selekciju atributa te ih definira i implementira kao tehnike selekcije atributa, problem na kojem se dosad još nisu primjenjivale. Predmet ovog istraživanja je primjena i evaluacija tehnika otkrivanja kontrasta kao tehnika za selekciju atributa. U tu svrhu provodi se opsežno istraživanje s ciljem utvrđivanja na skupovima podataka kojih karakteristika tehnike otkrivanja kontrasta nadmašuju klasične tehnike selekcije atributa, te dobivanja općeg odgovora da li se ubuduće mogu i u kojim situacijama tehnike otkrivanja kontrasta koristiti kao superiorne tehnike selekcije aributa, i mogu li u znatnoj mjeri otkloniti usko grlo cijelokupnog procesa otkrivanja znanja u podacima.

1.2.Svrha i ciljevi istraživanja

Odabir optimalnog skupa atributa za određeni zadatak problem je koji je važan u širokom rasponu različitih područja analize podataka. Stvarne primjene klasifikacije uključuju rad s velikim brojem atributa koji znatno povećavaju složenost klasifikacije. Za proces klasifikacije je bitan samo manji broj atributa s velikim diskriminacijskim mogućnostima. Odabir dobrog skupa atributa stoga je od ključne važnosti za izradu klasifikacijskog modela jer raste točnost modela i smanjuje se složenost i trajanje postupka.

Pregled literature je pokazao da je selekcija atributa još uvijek usko grlo procesa otkrivanja znanja u podacima. Stoga se u ovom istraživanju uočava potencijal primjene tehnika otkrivanja kontrasta, STUCCO i Magnum Opus, za potrebe selekcije atributa.

Ideja primjene tehnika otkrivanja kontrasta u selekciji atributa motivirana je temeljnim postavkama tehnika otkrivanja kontrasta. Naime, analiza razlika temeljeni je dio u razumijevanju kako i zašto stvari funkcioniraju (Satsanagi, 2011.). Da li osobe s visokom stručnom spremom zarađuju više od onih s višom? Zašto su neki studenti više uspješni, a drugi manje uspješni? Da bi se odgovorilo na prethodno navedena pitanja potrebno je napraviti usporedbu između više grupa. Traženje razlika među grupama središnji je problem u mnogim domenama. Razlikovanje određenih grupa posebno je važno u istraživanjima u društvenim znanostima. Stoga se 1999. godine, unutar rudarenja podataka, počinje razvijati područje otkrivanja kontrasta (Bay i Pazzani, 1999.). Značajan interes za ovo, relativno novo

područje, može se dobro razumijeti komentarom iz američkog crtanog filma *Get Fuzzy* autora Darbya Conley koji kaže da sve što je vrijedno treba uspoređivati jer će se tako još više cijeniti: „*Sometimes it is good to contrast what you like with something else. It makes you appreciate it even more"* (Conley, 2001.). Ova tvrdnja može se argumentirati slijedećim primjerima: praćenje promjena u prodaji od npr. 2002. do 2012. u jednom odjelu neće biti toliko informativno kao usporedba s prodajom u drugim odjelima. Poanta primjera je da kad neki objekat uspoređujemo s drugim, dobivamo više informacija. Boettcher u preglednom radu područja otkrivanja kontrasta iz 2011. godine navodi kao prednost ovog područja smanjenje kompleksnosti podataka, a da se pritom sačuva većina informacija iz orginalnog skupa podataka (Boettcher, 2011.). Umjesto da dva skupa podataka uspoređuje direktno, pristup otkrivanja kontrasta prvo nauči obrasce iz skupova podataka i onda ih uspoređuje. Ova činjenica predstavlja veliku motivaciju za primjenu tehnika otkrivanja kontrasta u selekciji atributa.

Motivirano prethodno navedenim, svrha ovog rada je utvrditi koliko točno i brzo tehnike otkrivanja kontrasta provode selekciju atributa. Primjenom tih tehnika cilj je poboljšati proces selekcije atributa, a time i cijeli proces otkrivanja znanja u podacima. Prednosti i nedostaci tehnika otkrivanja kontrasta utvrđuju se primjenjujući ih na referentnim skupovima podataka te uspoređujući s dosad korištenim tehnikama u ovu svrhu.

Kroz istraživanje provodi se evaluacija tehnika koje su se upotrebljavale za selekciju atributa. Postojeće metode se testiraju i uspoređuju s novim tehnikama predloženim u ove svrhe.

Glavni cilj istraživanja je proučiti primjenjivost tehnika otkrivanja kontrasta u selekciji atributa i identificirati za koje karakteristike podataka primjena tih tehnika poboljšava točnost i skraćuje vrijeme klasifikacije.

1.3.Hipoteze istraživanja

U skladu s prethodno navedenim ciljevima istraživanja, definiraju se slijedeće hipoteze istraživanja:

H1: Tehnike otkrivanja kontrasta za određene karakteristike podataka brže provode selekciju atributa od dosad šire korištenih tehnika selekcije atributa.

H2: Primjenom otkrivanja kontrasta u selekciji atributa za određene karakteristike podataka postiže se točnija klasifikacija nego dosad šire korištenim tehnikama selekcije atributa.

Hipoteza H1 će se koristiti za identificiranje karakteristika podataka za koje primjena tehnika otkrivanja kontrasta u selekciji atributa brže provodi selekciju atributa nego tehnikama koje su dosad najčešće korištene, a hipoteza H2 za identificiranje karakteristika podataka za koje se primjenom tehnika otkrivanja kontrasta u selekciji atributa točnije provodi klasifikacija nego tehnikama koje su dosad najčešće korištene. Pregledom literature utvrđeno je da su dosad šire korištenih tehnike selekcije atributa za potrebe klasifikacija: *Relief* algoritam, omjer dobiti (eng. *Gain ratio*), informacijska dobit (eng. *information gain*), linearni odabir unaprijed *(eng. linear forward selection) i* tehnika glasovanja *(*eng. *voting)*. Te tehnike će se koristiti kao referentne kod testiranja postavljenih hipoteza H1 i H2, a njihov odabir argumentira se u trećem poglavlju, selekcija atributa.

1.4.Metodologija istraživanja

Istraživanje se provodi slijedeći korake procesa otkrivanja znanja u podacima. Tijek istraživanja uključuje tri temeljne komponente:

(1) selekcija atributa

(2) klasifikacija i evaluacija

(3) analiza i komparacija rezultata

Prvi korak je prikupljanje skupova podataka koji imaju različite karakteristike nad kojima će se provoditi najprije selekcija atributa, a potom i klasifikacija. Dokazano je da karakteristike podataka utječu na postupak klasifikacije. Stoga je važno istražiti zašto neki algoritmi dobro djeluju na skupovima podataka s određenim karakteristikama, dok na drugima slabije. Različiti autori uzimaju u obzir različite karakteristike podataka. Ovo istraživanje vodi se rezultatima istraživanja Van der Walta (Van der Walt, 2008.) koji je identificirao nekoliko grupa karakteristika podataka važnih za klasifikaciju. U ovom istraživanju se pronalaze skupovi podataka uzimajući kao kriterije najvažnije karakteristike prepoznate u Van der Waltovom istraživanju (Van der Walt, 2008.), a više o tim karakteristikama napisano je u petom poglavlju, karakteristike podataka.

Kao izvor podataka koriste se javno dostupni repozitoriji koje sadrže referente skupove podataka s pratećom dokumentacijom za svaki skup. Popis svih korištenih skupova podataka s njihovim web adresama nalazi se u prilogu rada.

Kako bi se izdvojili atributi koji daju najveću informaciju za klasifikaciju, nad svakim skupom se provodi selekcija atributa. U sklopu selekcije atributa kompariraju se tehnike otkrivanja kontrasta, koje se u ovom istraživanju prvi put primjenjuju u selekciji atributa, s dosad najčešće korištenim tehnikama selekcije atributa. Nad svakim skupom podataka se provodi selekcija atributa sa svakom od tehnika selekcije atributa.

Nad selektiranim atributima se provodi klasifikacija primjenom algoritama koji predstavljaju različite pristupe klasifikaciji: statistički pristup (diskriminacijska analiza) i pristup neuralnog računarstva (neuronske mreže)**.** Klasifikacija se provodi primjenom svakog klasifikatora na svakom skupu podataka koji zadovoljava pretpostavke pojedinog algoritma.

Evaluacija tehnika selekcije atributa se odnosi na: 1) mjerenje vremena procesora potrebnog da se izvede selekcija atributa i 2) na komparaciju točnosti algoritama klasifikacije. Točnost algoritama klasifikacije predstavlja sposobnost algoritma da točno razvrsta što veći broj uzoraka iz skupa podataka. Za evaluaciju klasifikatora koristit će se matrica konfuzije (eng. *confusion matrix*). Matrica konfuzije je koristan alat za analiziranje koliko se rezultati dobiveni razvrstavanjem uzoraka razlikuju od stvarnih vrijednosti (Japkowicz i Shah, 2011., Oreški, Oreški i Oreški, 2012). Ako imamo *m* klasa, matrica konfuzije je tablica veličine najmanje *m* puta *m*.

Usporedba učinkovitosti tehnika otkrivanja kontrasta s ostalim tehnikama korištenim za selekciju atributa na skupovima podataka različitih karakteristika provest će se provedbom testova za procjenu statističke značajnosti razlika između pojedinih tehnika u brzini i točnosti. Svrha primjene testa je utvrđivanje da li su razlike procijenjenih srednjih vrijednosti točnosti klasifikacije i vremena potrebnog za izvršavanje algoritma selekcije atributa značajne.

Analizom i komparacijom rezultata utvrđuje se za koje karakteristike skupova su tehnike otkrivanja kontrasta u selekciji atributa dale bolje rezultate, tj.

(1) brže provele selekcije ili

2) dale točniju klasifikaciju od prethodno korištenih tehnika za selekciju atributa.

Pri tome se tehnike otkrivanja kontrasta definiraju kao metode filtra, znači neovisne su o klasifikatoru. Ideja je vođena prijedlozima Abea i suradnika koji zaključuju da je tehnikama selekcije atributa bolje odabrati onaj podskup atributa koji je efektivan na više klasifikatora (Abe et. al., 2006.).

1.5.Struktura rada

Drugo poglavlje ovog rada daje pregled osnovnih koncepata u području otkrivanja znanja u podacima. Definira se proces otkrivanja znanja u podacima, klasifikacija kao osnovna zadaća tog procesa koja je i u fokusu ovog rada te se daje pregled algoritama učenja i načina usporedbe rezultata pojedinih algoritama. Poglavlje 3 definira selekciju atributa, opisuje tehnike selekcije atributa koje će se koristi te daje pregled dosadašnjih komparacija učinkovitosti tehnika selekcije atributa.

Poglavlje 4 opisuje područje otkrivanja kontrasta, daje temeljne značajke ovog područja te temeljito opisuje dvije tehnike, STUCCO i Magnum Opus. Kroz pregled dosadašnjih istraživanja, završni dio četvrtog poglavlja ističe prednosti i nedostatke tehnika otkrivanja kontrasta.

Poglavlje 5 definira koje su karakteristike podataka važne za klasifikaciju te ih opisuje i objašnjava način na koji se izračunavaju. Poglavlje 6 predstavlja tehnike otkrivanja kontrasta u selekciji atributa; definira na koji način se primjenjuju i koje mjere se pritom koriste. Poglavlje 7 opisuje empirijsko istraživanje, a osmo poglavlje prikazuje i diskutira dobivene rezultate te navodeći ograničenja ovog istraživanja daje prijedloge za daljnja istraživanja. Zadnje poglavlje donosi zaključke o hipotezama i realizaciji ciljeva istraživanja.

2. PROCES OTKRIVANJA ZNANJA U PODACIMA

Tehnološke inovacije su revolucionarizirale proces znanstvenog istraživanja i otkrivanja znanja. Dostupnost podataka i izazovi na području obrade tih podataka preoblikovali su način statističkog razmišljanja, analize podataka i teorijskih studija. Izazovi rada s velikom količinom podataka javljaju se u različitim područjima znanosti, od biologije i medicinskih studija do financija i upravljanja rizicima. U svim tim područjima otkrivanje znanja u podacima je ključan proces istraživanja *(*Liu et. al., 2010.). Otkrivanje znanja u podacima multidisciplinarno je područje koje uključuje znanje baza podataka, strojnog učenja i statistike.

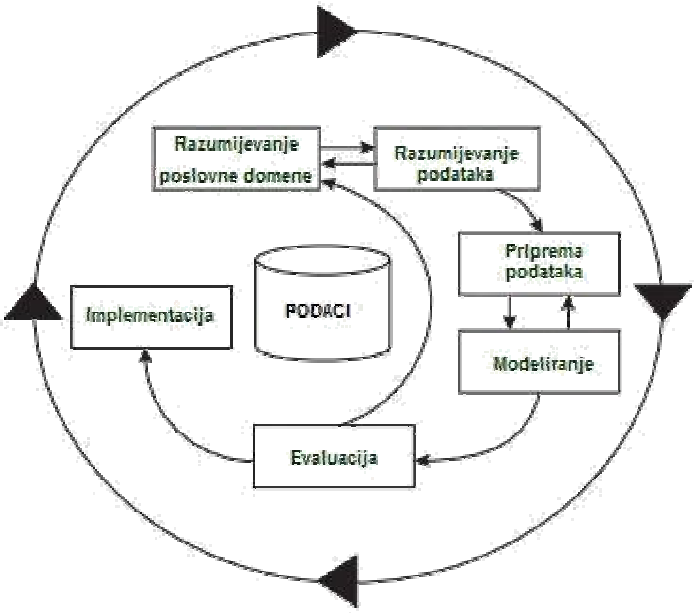
2.1. Definicija procesa

Process otkrivanja znanje je netrivijalni proces identifikacije:

- valjanih,
- novih,
- potencijalno korisnih
- i razumljivih

uzoraka u podacima. (Fayad et. al., 1996.)

Uslijed velikog porasta interesa i razvoja područja razvijaju se standardi i modeli koji definiraju korake u implementaciji procesa (Azevedo, Santos, 2008.). Azevedo i Santos rade usporedbu tri najpoznatija standardizirana modela, a to su: KDD (eng. *Knowledge discovery in databases*), model SEMMA (eng. *Sample, Explore, Modify, Model, Assess*), model CRISP - DM (eng. *CRoss – Industry Standard Process for Data Mining*) (Azevedo, Santos, 2008.).
U ovom radu proces otkrivanja znanja se opisuje kroz standardizirani model CRISP – DM, a koraci tog modela prikazani su na slici 1.



Slika 1. Model CRISP – DM

Izvor: Marbán, Segovia, Menasalvas, Fernández – Baizán, 2009.

CRISP – DM podrazumijeva slijedeće korake u procesu otkrivanja znanja (Marbán, Segovia, Menasalvas, Fernández – Baizán, 2009.):

1. **Razumijevanje poslovne domene** te zahtjeva i ciljeva krajnjeg korisnika
2. **Razumijevanje podataka** uključuje prikupljanje podataka, upoznavanje s podacima te procjenu valjanosti i kvalitete podataka
3. Korak **pripreme podataka** sadrži sve aktivnosti s ciljem da se konstruira konačan skup podataka iz početnih „sirovih“ podataka. Zadaće vezane uz pripremanje podataka: selekcija atributa, pretvorba i čišćenje podataka za alate modeliranja.
4. **Modeliranje** uključuje izbor i primijenu tehnika rudarenja podataka: traženje zakonitosti od interesa u određenom obliku: klasifikacijska pravila ili stablo, klasteri,..
5. **Evaluacija**: u ovoj je fazi izrađen model (ili više modela) za koje se čini da su kvalitetne forme iz perspektive analize podataka. Prije nego što se započne implementacija modela važno je temeljito evaluirati model i pregledati izvršene korake koji su učinjeni da se konstruira model zato da se utvrdi da li su doista postignuti poslovni ciljevi.
6. **Implementacija**: kreirani model nije kraj projekta. Iako je svrha modela povećanje znanja, dobiveno znanje treba biti organizirano i predstavljeno na način koji korisnik može upotrijebiti. To često uključuje primjenu „živih“ modela tijekom procesa donošenja organizacijskih odluka.

Proces otkrivanja znanja u podacima provodi se sa svrhom ispunjavanja jedne od zadaća: klasifikacija, regresija, klasteriranje, sumarizacija, modeliranje ovisnosti i otkrivanje promjena i devijacija (Fayyad et al. 1996, Witten i Frank, 2005.). Klasifikacija se smatra temeljnom i jednom od najvažnijih zadaća (Lavanya i Rani, 2011) te je i u fokusu ovoga rada.

2.2. Klasifikacija

Zadaća klasifikacije javlja se u širokom rasponu ljudskog djelovanja. U najširem poimanju, pojam klasifikacija uključuje svaki kontekst u kojem je odluka napravljena na temelju trenutno dostupne informacije, a klasifikacijska procedura je formalni postupak za opetovano donošenje takvih odluka u novim situacijama. U ovom radu razmatra se stroža definicija klasifikacije. Pretpostavit ćemo da se problem odnosi na izradu procedure koja se primjenjuje na niz *instanci* (slučajeva, entiteta), gdje svaki novi slučaj mora biti dodijeljen jednoj od unaprijed definiranih *klasa* na temelju promatranih *atributa*. Izrada klasifikacijske procedure iz skupa podataka za koji je poznata pripadnost slučajeva klasi naziva se *diskriminacija* ili *nadzirano učenje* (Michie, Spiegelhalter i Taylor, 1994.).

Klasifikacija ima dva različita značenja. U jednoj situaciji moguće je raspolagati sa skupom instanci i cilj je utvrđivanje postajanja klasa ili klastera u podacima. U drugoj situaciji možemo znati koliko je točno klasa i cilj je utvrditi pravila na temelju kojih možemo klasificirati nove instance u postojeće klase.

Prva situacija naziva se nenadzirano učenje, a druga nadzirano učenje.

U ovom radu termin klasifikacija koristi se u kontekstu nadziranog učenja, a istraživanje u okviru ove disertacije se fokusira samo na probleme s dvije klase, situacije u kojima je zavisni atribut binarni. Znači, zavisni atribut je diskretni. Ovo je i ključna karakteristika koja razlikuje klasifikaciju od regresije (zadaće kod koje je zavisni atribut kontinuirani).

Klasifikacija je zadaća procesa otkrivanja znanja koja, na temelju atributa, određuje objektu u koju od prije definiranih klasa pripada (Weiss and Kulikowski 1991., Hand 1981.). Znači, ulaz u proces klasifikacije je skup podataka koji se sastoji od određenog broja instanci. Cilj je upotrebom skupa podataka za treniranje izgraditi model koji se koristi u klasifikaciji novih podataka, koji nisu iz skupa za treniranje.

Postoje različite metode klasifikacije svaka s određenim prednostima i nedostacima (Lavanya i Usha Rani, 2011.). Ne postoji metoda koja bi davala najbolje rezultate za sve probleme klasifikacije (Lavanya i Usha Rani, 2011.). Michie i suradnici metode klasifikacije dijele u dvije grupe: statističke metode i metode strojnog učenja (Michie, Spiegelhalter i Taylor, 1994.). Lahiri u doktorskoj disertaciji kojoj je cilj usporediti tehnike iz ova dva pristupa radi opsežan pregled komparativnih istraživana te utvrđuje da su iz grupe metoda strojnog učenja najbolji izbor neuronske mreže (s obzirom na kriterij točnost klasifikacije), a iz grupe statističkih metoda diskriminacijska analiza (Lahiri, 2006.).

Stoga se u ovom radu koriste predstavnici dvije grupe indukcijskih algoritama, kao baza za komparaciju: neuronske mreže i diskriminacijska analiza. Dok je diskriminacijska analiza napredna statistička metoda s nizom pretpostavki, neuronske mreže su metoda strojnog učenja koja nema te pretpostavke. Ove metode koriste se i u najnovijim istraživanjima koji rade usporedbe više tehnike selekcije atributa na više klasifikatora (npr. Silva et al. 2013).

2.2.1. Diskriminacijska analiza

Diskriminacijska analiza je multivarijantna metoda koja se koristi za identificiranje atributa koji razlikuju pripadnost jednoj od dvije (ili više) klase. Analiza se vodi idejom da ustanovi koji atributi rade najveću razliku između uspoređivanih klasa instanci. Dakle, polazi se od nekoliko grupa instanci opisanih nizom atributa. Pritom se zahtijeva da se konstruiraju novi atributi (kojih treba biti manje nego polaznih) koji bi opisali razlike među klasama. Ti novi atributi zovu se *diskriminacijskim funkcijama,* dobivaju se kao linearne kombinacije izvornih atributa, a formiraju se prema zahtjevu da što bolje razlikuju klase.

Glavni ciljevi diskriminacijske analize su (Garson, 2008.):

1. identificiranje izvornih atributa koje najbolje diskriminiraju prethodno definirane klase;

2. korištenje ovih identificiranih atributa za izračun diskriminacijskih funkcija kao linearnih kombinacija izvornih atributa. Diskriminacijske funkcije daju najbolju separaciju ovih klasa;

3. definiranje zakonitosti po kojoj bi se buduća mjerenja (koja nisu bila uključena u definiranju ove zakonitosti) svrstala u jednu od definiranih klasa.

U ovom istraživanju diskriminacijska analiza će se koristiti u svrhu klasifikacije. Računalni program SAS JMP, koji implementira diskriminacijsku analizu i koji će se koristiti u ovom istraživanju, računa funkciju klasifikacije. Funkcija klasifikacije se koriste za određivanje kojoj klasi koji slučaj pripada, što je i potreba ovog rada. Svaka funkcija omogućuje izračun klasifikacijskih vrijednosti za svaki slučaj za svaku klasu, prema formuli:

$$S_i = c_i + w_{i1} * x_1 + w_{i2} * x_2 + \cdots + w_{im} * x_m$$

Pri čemu je:

S_i – vrijednost klasifikacije,

i –oznaka klase,

1,2,..m –atributi

C_i – konstantna za *i-tu* grupu

W_{ij} – vrijednost pondera za *j-ti* atribut u izračunu vrijednosti klasifikacije za *i-tu* klasu

x_j – vrijednost promatranog slučaja za *j-ti* atribut

Izračunom vrijednosti *S* za svaku od *I* klasa utvrđuje se kako klasificirati pojedini slučaj – u onu klasu za koju je vrijednost klasifikacije najveća.

Prije provedbe diskriminacijske analize potrebno je provjeriti da li skup podataka zadovoljava određene pretpostavke. Garson navodi slijedeće pretpostavke koje treba ispitati na skupu podataka prije nego se provede diskriminacijska analiza. Te pretpostavke su (Garson, 2008):

- *Stvarna kategorična ovisnost*. Zavisni atribut predstavlja pravu dihotomiju. Ako je atribut kontinuirani i njegov doseg se ograničava jedino u svrhu primjene diskriminacijske analize, korelacija slabi. Stoga nikad ne bi trebalo raditi dihotomiju kontinuiranog atributa jedino u svrhu provedbe diskriminacijske analize. Klase moraju biti međusobno isključive, tj. svaka instanca pripada samo jednoj klasi.
- *Nezavisnost*. Sve instance moraju biti nezavisne. Stoga ne može biti podataka koji koreliraju.

- *Nema značajnih razlika u veličini grupa*. Veličine grupa ne smiju se jako razlikovati. Ako se ova pretpostavka prekrši, bolje je provesti regresiju.

- *Prikladna veličina uzorka*. Moraju biti najmanje dvije jedinice za svaku klasu zavisnog atributa.

- *Varijanca*. Nijedan nezavisni atribut nema standardnu devijaciju 0 u jednoj ili više klasa formiranih pomoću zavisnog atributa.

- *Homogenost varijanci*. Između svake klase formirane zavisnim atributom, varijanca svakog intervala zavisnih atributa trebala bi biti slična među klasama. Nezavisni atributi mogu imati različite varijance, ali za jednu nezavisnu, grupe formirane zavisnim atributom trebale bi imati slične varijance i aritmetičke sredine na tom nezavisnom atributu. Diskriminacijska analiza je jako osjetljiva na atipične vrijednosti. Izostanak homogenosti varijanci može ukazivati na postojanje atipičnih vrijednosti u jednoj ili više grupa. Nedostatak homogenosti varijanci znači da je test značajnosti nepouzdan, osobito ako je veličina uzorka mala i podjela zavisne varijable neujednačena.

- *Homogenost kovarijanci/korelacija*. Između svake klase formirane zavisnim atributom, kovarijanca/korelacija između svaka dva prediktorska atributa treba biti slična odgovarajućoj kovarijanci/korelaciji u drugim klasama. Tj. svaka klasa ima sličnu matricu kovarijanci/korelacija.

- *Izostanak savršene multikolinearnosti*. Ako jedan nezavisni atribut visoko korelira s drugim, ili je funkcija (npr. suma) drugih nezavisnih atributa, tada se vrijednosti tolerancije za taj atribut približavaju 0 i matrica neće imati jedinstveno diskriminacijsko rješenje.

- *U svrhu testiranja značajnosti*, atributi slijede multivarijatnu normalnu distribuciju. Tj., svaki prediktorski atribut ima normalnu distribuciju oko fiksnih vrijednosti svih drugih nezavisnih atributa.

Ove pretpostavke provjeravaju se na svakom skupu podataka koji se koristi u istraživanju, te se diskriminacijska analiza provodi samo na onim skupovima koji zadovoljavaju navedene pretpostavke.

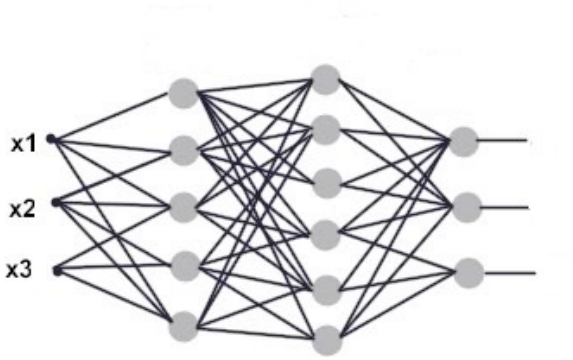
2.2.2. Neuronske mreže

Neuronske mreže su metoda umjetne inteligencije koja nastoji simulirati rad ljudskog mozga (Zekić, 1998.). Od početka prisutnosti u znanosti, neuronske mreže istraživane su s dva različita pristupa. Prvi, biološki pristup, istražuje neuronske mreže kao pojednostavljene simulacije ljudskog mozga i upotrebljava ih za testiranje hipoteza o funkcioniranju ljudskog mozga. Drugi pristup tretira neuronske mreže kao tehnološke sustave za složenu obradu informacija (Zahedi, 1993.). Ovaj rad je koncentriran na drugi pristup uz pomoć kojega se neuronske mreže ocjenjuju prema svojoj učinkovitosti u radu sa složenim problemima u području klasifikacije.

Neuronska mreža je paralelno distribuirana informacijska struktura, koja se sastoji od elemenata obrade (koji mogu imati lokalnu memoriju i mogu izvršavati procese operacije obrade informacija) međusobno povezanih komunikacijskim kanalima zvanima veze (Hand, Manila i Smyth, 2001.). Neuronska mreža je tehnika područja umjetne inteligencije, a nastala je prema uzoru na živčani sustav čovjeka. Čovjekov živčani sustav se sastoji od neurona – živčanih stanica koje obrađuju podatke. Neuroni su međusobno povezani i između njih putuju živčani impulsi. Svaki neuron se sastoji od dendrita – izdanaka koji primaju živčane impulse od drugih stanica ili osjetila, aksona – izdanaka koji prenose impulse do drugih stanica, jezgre za obradu podataka i sinapsi – spojeva između neurona. Drugi neuron tada prima živčani impuls i kreće u obradu podataka. Nakon što obradi podatke šalje ih drugim neuronima. Podaci se u mozgu obrađuju kroz nekoliko slojeva neurona različitih vrsti. 1943. godine neuropsiholog Warren McCulloch i logičar Walter Pits razvili su prema uzoru na biološki neuron, umjetni neuron, koji se danas koristi u umjetnoj inteligenciji.

Umjetni ili računalni neuron obrađuje podatke u računalu i on je element obrade. U neuronskim mrežama su neuroni složeni u slojeve: ulazni sloj (jedan), skrivene slojeve (jedan ili više) i izlazni sloj (jedan). Neuron ima funkciju prijenosa, koja služi za prijenos ulaznih signala prema izlaznom signalu a ona je najčešće sigmoida ili tangens hiperbolni. Ulaznih signala može biti više i oni dolaze od drugih neurona iz ulaznog sloja ili nekog prethodnog skrivenog sloja. Iz svakog neurona izlazi jedan izlazni signal, koji može biti ulazni signal drugim neuronima u drugim slojevima ili izlaz izlaznom sloju. Izlaz iz mreže ili neurona je u intervalu [0,1]. Podaci zvani signali (podaci su u biološkom neuronu živčani impulsi) ulaze u element obrade i tamo se zbrajaju pomoću neke zbrojne funkcije. Zbroj signala s težinskim vezama aktivira element obrade (interna aktivacija). Zbroj signala se obrađuje u skrivenim

slojevima neurona i funkcijom prijenosa prenosi prema izlaznom sloju. Zbrojna i prijenosna funkcija skupa čine aktivacijsku funkciju neurona. Na slici 2. prikazana je struktura neuronske mreže s tri sloja.



Slika 2. Struktura neuronske mreže

Temeljni princip učenja neuronske mreže su veze između uzoraka odnosno ulaznih podataka. Neuronske mreže poznaju autoasocijativno i heteroasocijativno učenje. U prvom se slučaju uzorci (podaci) pridružuju sebi samima (klasteriranje), a u drugom se dva različita tipa uzoraka pridružuju jedan drugome. Učenje može biti nadzirano i nenadzirano. Nadzirano učenje znači da se metodom povratne veze uklanjaju razlike između željenog i stvarno dobivenog izlaza. Kod nenadziranog učenje nema povratne veze.

Da bi neuronske mreže mogle predviđati buduće vrijednosti nekih atributa treba ih najprije istrenirati. U raznim programskim alatima neuronske mreže mogu imat različitit broj skrivenih slojeva i svaka mreža može imati različitu pouzdanost predviđanja. Neuronska mreža je karakteristična po tome što svaki put, pa čak i istim postavkama, može dati drugačije rezultate jer postoji nedeterminizam u svrhu boljeg pretraživanja prostora mogućih rješenja. U ovom istraživanju za implementaciju neuronske mreže korišten je alat SAS JMP koji implementira neuronsku mrežu s unazadnom propagacijom (eng. *backpropagation neural network)* te se u slijedećem podpoglavlju opisuje taj algoritam neuronske mreže.

2.2.2.1 Mreže s unazadnom propagacijom

Mreže s unazadnom propagacijom su najčešće korištene neuronske mreže (Zahedi, 1993., Zekić, 1998.). One uče širenjem greške unatrag, a prilikom učenja modificiraju se težine veza

između neurona. Postoje četiri osnovna koraka algoritma za širenje greška unatrag (eng. *backpropagation algorithm*) u mrežama s više slojeva (Hand, Manila i Smyth, 2001.):

1. računanje unaprijed,
2. širenje greške unatrag u izlazni sloj,
3. širenje greške unatrag u skriveni sloj,
4. ažuriranje težina veza.

Algoritam staje kad je vrijednost greške dovoljno mala da bi mogli koristiti mrežu za klasifikaciju. U prvom koraku se računa vektor **o**, koji se šalje sve do izlaznog sloja kroz sve slojeve neurona. Za vrijeme prenošenja tog vektora se računaju interne aktivacije i izlazi iz svakog neurona. Kada u bilo koji neuron ulaze ulazne vrijednosti one se sumiraju. Prilikom izlaza iz pojedinog neurona na njih djeluje funkcija prijenosa (tangens hiperbolni ili sigmoida). Izlaz iz nekog neurona *j* u *s*- tom sloju neuronske mreže se računa po formuli (Hand, Manila i Smyth, 2001.):

$$x_j^{[s]} = f\left(\sum w_{ji}^{[s]} \cdot x_i^{[s-1]}\right) = f\left(I_j^{[s]}\right).$$

Pri čemu je:

f funkcija prijenosa,

$w_{ji}^{[s]}$ težina veze koja povezuje *i* – ti i *j* – ti neuron u *s* – tom sloju,

$x_i^{[s]}$ trenutačno stanje izlaza *i* – tog neurona u *s-1* (prethodnom) sloju.

Suma umnožaka svih težina i izlaza iz neurona što znači interna aktivacija neurona *j* u *s* – tom sloju. Na dobiveni broj se primjenjuje neka od funkcija prijenosa: iz prethodnog sloja je interna aktivacija neurona i označava se s $I_j^{[s]}$,

$$Tanh(I) = \frac{e^{I} - e^{-I}}{e^{I} + e^{-I}} \quad Sigmoida(I) = \frac{1}{1 + e^{-I}}$$

Na taj se način mogu izračunati izlazi za sve neurone pa na kraju i izlazi za neurone u izlaznom sloju i tako saznati izlaz iz neuronske mreže. Dobiveni broj je očekivani izlaz iz neuronske mreže, a ukoliko se dogodi da on nije jednak stvarnom izlazu, treba ga ispraviti

širenjem nazad. To je drugi korak algoritma. Formula za računanje greške za neurone u izlaznom sloju je (Hand, Manila i Smyth, 2001.):

$$E = f'(I) \cdot (D - A).$$

Nakon vraćanja greške u izlazni sloj, dolazi treći korak u kojem se greška širi i prema skrivenim slojevima. Greška za neurone u skrivenim slojevima se računa po formuli (Hand, Manila i Smyth, 2001.):

$$E^j = f'(I^j) \cdot \sum (E^{j+1} \cdot w^{j+1}).$$

Kod obje formule se koristi derivacija funkcije prijenosa. U prvoj formuli se derivirani izlaz množi s razlikom traženog i stvarnog izlaza. U drugoj formuli se greška za neuron u *j* – tom sloju računa kao umnožak derivacije izlaza u *j* – tom sloju s zbroja umnožaka grešaka i težina veza u *j+1* sloju (sljedećem sloju gledamo li od naprijed prema unatrag). Da bi se mogle izračunati dane derivacije izlaza treba znati derivirati funkcije prijelaza. Derivacija funkcije tangens hiperbolni je:

$$f'(I) = (1 + f(I)) \cdot (1 - f(I))$$

a derivacija funkcije sigmoide:

$$f' = f(I) \cdot (1 - f(I)).$$

Kod učenja se greške iz viših slojeva neuronske mreže prosljeđuju sve do ulaznog sloja i korigiraju se težine veza. Ažuriranje težina veza je zadnji korak u algoritmu širenaj greške unatrag. Za korekciju težina neke veze se koristi greška neurona u *j* – tom sloju i izlaz iz neurona iz *j-1* sloja čija se veza s neuronom u j – tom sloju korigira te koeficijent učenja.

Formula za korekciju težine veze je (Hand, Manila i Smyth, 2001.):

$$\Delta(w) = Lcoef \cdot E^j \cdot x^{j-1}.$$

Mreže s unazadnom propagacijom mogu imati jako puno skrivenih slojeva, pri čemu se može stvoriti problem, jer je potrebno više vremena za računanje svih grešaka i izlaza.

Vraćanje grešaka natrag u mrežu je treniranje. Nakon treninga neuronska mreža je spremna za rad sa stvarnim podacima.

2.2.3. Komparacija klasifikatora

Nekoliko je važnih točaka koje treba spomenuti kada govorimo o klasifikatorima, a koje su važne kod odabira klasifikatora jer služe za usporedbu različitih klasifikatora. Neke od njih su (Michie et al., 1994.):

- *Točnost*: točnost klasifikacije prikazuje se postotkom točno klasificiranih instanci, iako može biti da su neke greške „ozbiljnije“ od drugih te može biti važno kontrolirati stopu pogreške samo za neke, najvažnije klase
- *Brzina:* u nekim slučajevima brzina klasifikatora može biti izuzetno važna. Klasifikator točnosti 90% može biti bolji izbor od klasifikatora koji postiže točnost od 95% ako je 100 puta brži. U okruženju koje se izrazito brzo mijenja, mogućnost da se klasifikacijska pravila brzo nauče jako je važno.

Prethodna istraživanja su pokazala je da priprema podataka (koja uključuje korake čišćenja podataka i redukcije broja atributa) usko grlo cjelokupnog procesa otkrivanja znanja u podacima te da oduzima čak 60% - 95% ukupnog vremena procesa (De Veaux, 2005.). Zbog toga se priprema podataka smatra najvažnijim korakom u procesu otkrivanja znanja u podacima te je selekcija atributa u fokusu interesa ovog rada jer dobra selekcija atributa može znantno ubrzati cjelokupni proces klasifikacije.

2.3. Evaluacija rezultata

Kao kriteriji evaluacije u ovom istraživanju koristit će se točnost klasifikacije i brzina provođenja selekcije atributa. Točnost klasifikacije odnosi se na sposobnost modela da ispravno odredi pripadnost klasi za nove podatke (Japkowicz i Shah, 2011.). Za mjerenje točnosti koristit će se matrica konfuzije. Matrica konfuzije je koristan alat za analiziranje koliko se rezultati dobiveni razvrstavanjem uzoraka razlikuju od stvarnih vrijednosti (Japkowich i Shah, 2011., Oreški, Oreški, i Oreški, 2012.). Matrica konfuzije za dvije klase

prikazana je u tablici u nastavku. Ako imamo m klasa, matrica konfuzije je tablica veličine najmanje *m* puta *m*. Klasifikator daje dobru točnost ako je većina uzoraka na dijagonali, a vrijednosti van dijagonale blizu 0.

Tablica 1. Matrica konfuzije

		Predviđanje	
		Nema promjene	Promjena
Razred uzorka	Nema promjene	TN	FP
	Promjena	FN	TP

Matrica se sastoji od slijedećih vrijednosti:

TN - broj *ispravno* predviđenih *negativnih* ishoda

FP - broj *pogrešno* predviđenih *pozitivnih* ishoda

FN - broj *pogrešno* predviđenih *negativnih* ishoda

TP - broj *ispravno* predviđenih *pozitivnih* ishoda

Mjere koje se koriste kod matrice konfuzije su slijedeće (Japkowicz i Shah, 2011.):

- **točnost** (eng. accuracy) je omjer uzoraka kojima je razred točno predviđen i ukupnog broja uzoraka. Računa se prema slijedećoj formuli:

$$\frac{TN + TP}{TN + FP + FN + TP}$$

- **opoziv** (eng. recall) ili mjera točno predviđenih pozitivnih uzoraka (TP) (eng true positive rate):

$$\frac{TP}{FN + TP}$$

- **mjera pogrešno predviđenih pozitivnih uzoraka** (PP) (eng. false positive rate) je omjer uzoraka koji su pogrešno svrstani u pozitivan razred i ukupnog broja negativnih uzoraka:

$$\frac{FP}{TN + FP}$$

- **mjera točno predviđenih negativnih uzoraka** (eng. true negative rate):

$$\frac{TN}{TN + FP}$$

- **mjera pogrešno predviđenih negativnih uzoraka** (eng. false negative rate) je omjer uzoraka koji su pogrešno svrstani u negativan razred i ukupnog broja pozitivnih uzoraka:

$$\frac{FN}{FN + TP}$$

- **preciznost** (eng. precision) je omjer točno predviđenih pozitivnih uzoraka i ukupnog broja uzoraka za koje je predviđen pozitivan razred:

$$\frac{TP}{FP + TP}$$

Ovaj rad u komparaciji rezultata koristi točnost klasifikatora kao mjeru. Kao metoda validacije rezultata koristi se unakrsno vrednovanje, koje se opisuje u podpoglavlju 2.3.1.

2.3.1. Unakrsno vrednovanje

Unakrsno vrednovanje (eng. *cross validation*) je metoda za procjenu generalizacije performansi tehnika s ciljem utvrđivanjem najboljeg načina korištenja raspoloživih podataka. Unakrsno vrednovanje je tehnika validacije koja se koristi u statistici, a posebno u strojnom učenju (Bonev, 2010). Tehnika se sastoji u podjeli skupa na nekoliko podskupova i provođenju statističke analize na različitim kombinacijama tih podskupova (Bonev, 2010). Najčešća metoda unakrsnog vrednovanja je *unakrsno vrednovanje s k preklapanja* (eng. k-*fold cross validation*) (Bonev, 2010). Ova metoda dijeli skup u *k* podskupova i provodi *k* analiza računajući točnost za svaki podskup. Jedan podskup se koristi za testiranje, a ostali za treniranje. Postupak se ponavlja sve dok je svaki podskup jednom skup za testiranje. Particioniranje skupa se provodi samo jednom (Bonev, 2010). U ovom istraživanje uzima se srednja dobivena točnost u evaluaciji rezultata.

2.3.2. Testiranje statističke značajnosti

Mjere performansi tehnika opisane na početku poglavlja 2.3. same nisu dovoljne za punu evaluaciju razlika u performansama tehnika. Preciznije rečeno, iako su performanse tehnika različite na određenom skupu podataka, potrebno je provjeriti da li su uočene razlike statistički značajne ili samo slučajne (Japkowicz i Shah, 2011). Svrha statističkog testiranja je prikupljanje dokaza o stupnju u kojem su razultati evaluacijskih mjera reprezentativni za generaliziranje o ponašanju tehnika (Japkowicz i Shah, 2011). Za tu svrhu razvijeni su statistički testovi. Japkowicz i Shah rade pregled statističkih testova i, s obzirom na situaciju u kojoj se koriste, dijele ih na (Japkowicz i Shah, 2011):

- testove za komparaciju dva algoritma na jednoj domeni
- testove za komparaciju više algoritama na jednoj domeni
- testove za komparaciju više algoritama na više domena

U ovom istraživanju radi se komparacija više algoritama na više domena te se u posljednoj skupini treba tražiti statistički test. Za ovakav slučaj Japkowicz i Shah predlažu parametrijski test (ANOVA) i neparametrijski test (Friedman test). Pošto paramterijski test zahtijeva normalnu distribuciju, a svi skupovi podataka u ovom istraživanju ne ispunjavaju tu pretpostavku, koristi se neparametrijski test, Friedman test.

Friedman test koristi se za statističko testiranje razlika u točnosti klasifikacije i vremenu provođenja selekcije atributa. Za ovu vrstu evaluacije predlažu ga Japkowicz i Shah (Japkowicz i Shah, 2011.) a koristili su ga i Čehovin i Bosnić (Čehovin i Bosnić, 2010.) te Demšar (Demšar, 2007) u istraživanjima koja su se također odnosila na situaciju komparacije više tehnika selekcije atributa na više skupa podataka.

Friedman test je neparametrijski test koji pod nul hipotezom pretpostavlja da su rezultati svih algoritama jednaki, dok odbijanje nul hipoteze sugerira postojanje razlika između performansi proučavanih algoritama.

Friedman test radi na slijedeći način: rangira algoritme za svaki skup podataka odvojeno, pri čemu algoritam s najboljim rezultatom dobiva rang 1, onaj s drugim najboljim rezultatom rang 2, itd. U slučaju jednake pozicije dvaju algoritama, računa se prosjek koji se pridaje oba algoritma.

Neka je r_i^j rang *j*-tog od k algoritama n *i*-tom od *N* skupova podataka. Friedman test uspoređuje prosječne rangove algoritama, $R_j = \frac{1}{N}[\sum_j R_j^2 - \frac{k(k+1)^2}{4}]$. Pod nul hipotezom koja kaže da su rezultati svih algoritama jednaki, pa samim time i njihovi rangovi R_j trebaju biti jednaki, Friedman statistika

$$X_F^2 = \frac{12N}{k(k+1)}[\sum_j R_j^2 - \frac{k(k+1)^2}{4}]$$

je distribuirana prema X_F^2 s *k-1* stupnjem slobode, kada su *N* i *k* dovoljno veliki (*N*>10 i *k*>5) (Demšar, 2006).

Pretpostavke Friedman testa su slijedeće (Demšar, 2006):

- Jedna grupa se mjeri na tri ili više slučaja
- Postoji jedna zavisna varijabla koja je ordinalna, intervalna ili odnosna
- Uzorak ne mora imati normalnu distribuciju

Svi skupovi podataka koji se koriste u empirijskom istraživanju u sklopu ovog rada zadovoljavaju navedene pretpostavke.

3. TEHNIKE SELEKCIJE ATRIBUTA

Dimenzionalnost skupa podataka može se smanjiti upotrebom prikladnih tehnika. Te se tehnike svrstavaju u dvije grupe: one koje konstruiraju nove atribute iz početnog skupa atributa (eng. *feature transformation*) i one koje odabiru podskup od početnog skupa atributa (**tehnike selekcije atributa**, eng. *feature selection*, npr. *Relief*, informacijska dobit,..). Kod tehnika selekcije atributa odabire se manji skup atributa na temelju evaluacijske funkcije.

Selekcija atributa je vrlo aktivno i plodonosno područje istraživanja u strojnom učenju, statistici i rudarenju podataka (Ramaswami i Bhaskaran, 2009.). Glavni cilj provedbe selekcije atributa je izabrati podskup ulaznih atributa kako bi se eliminirali atributi koji nisu relevantni i koji ne daju prediktivnu informaciju te konačno, postizanje visoke točnosti klasifikacije (Ramaswami i Bhaskaran, 2009.). Selekcija atributa se u teoriji i praksi pokazala učinkovita u povećanju djelotvornosti učenja, povećanju točnosti predviđanja i smanjenju složenosti rezultata (Koller, Sahami, 1996.)

Problem selekcije atributa odnosi se na pronalaženje onog podskupa unutar orginalnog skupa atributa na kojem će algoritam učenja generirati klasifikacijski model s najvećom točnošću. Da bi se to postiglo nužno je odabrati atribute koji su **relevantni** za klasifikacijski problem i koji **nisu redundatn**i (Liu et al., 2010; Blum & Langley, 1997).

Blum definira **relevantnost** atributa na slijedeći način:

„atribut *f* je relevantan ako promjena vrijednosti atributa rezultira promjenom vrijednosti atributa klase“ (Blum i Langley, 1997).

„Atribut *i* je **redundantan** s obzirom na klasu varijable *C* i drugi atribut *j* ako ako *i* ima veću klasifikacijsku moć za *j* nego za klasu varijable *C“* (Blum i Langley, 1997).

Još jedan pojam koji je potrebno definirati je *prokletstvo dimenzionalnosti* (eng. *curse of dimensionality).* Hand i suradnici ovaj problem definiraju kao eksponencijalnu stopu rasta broja instanci u prostoru kako raste broj atributa. Redukcija broja atributa smanjuje taj prostor i smanjuje kompleksnost klasifikacijskog problema (Hand, Manila i Smyth, 2001).

3.1. Definicija selekcije atributa

Selekcija atributa se definira na slijedeći način.

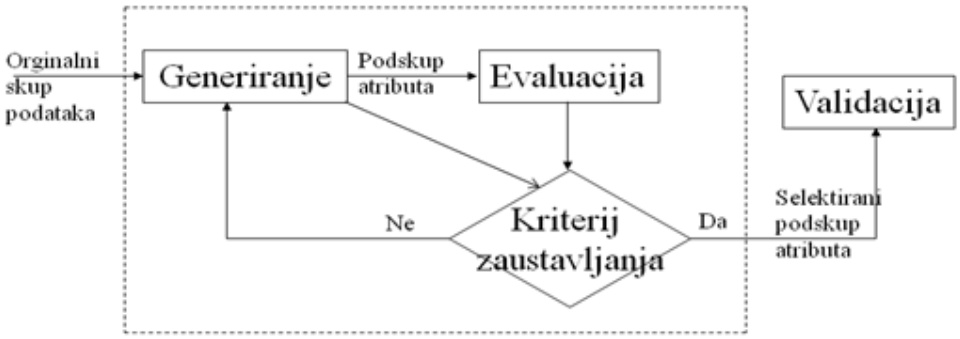
„Pretpostavimo da je *F* orginalni skup atributa s kardinalnošću *n* (*n* označava broj atributa u skupu *F*), a $\overline{F}$ selektirani podskup atributa s kardinalnošću $\overline{n}$ ($\overline{n}$ označava broj atributa u skupu *F*), te je $\overline{F} \subseteq F$. Nadalje, neka $J(\overline{F})$ bude kriterij za selekciju atributa u skup $\overline{F}$. Pretpostavljamo da veća vrijednost od *J* ukazuje na bolji podskup atributa. Stoga je cilj maksimizirati *J*(). Problem selekcije atributa je pronalaženje podskupa atributa $\overline{F} \subseteq F$ takvih da je,

$$J(\overline{F}) = \max_{Z \subseteq F, |Z|=n} J(Z)$$“

(Chrysostomou, 2009.)

Dash i Liu proces selekcije atributa provode u četiri koraka, a to su:

- generiranje podskupa,
- evaluacija podskupa,
- kriterij zaustavljanja i
- validacija (Dash i Liu, 1997).



Slika 3. Koraci selekcije atributa

Izvor: Dash i Liu, 1997.

Četiri glavna koraka su prikazana na slici 3. Na orginalnom skupu podataka se generira podskup atributa. Kod traženja podskupa u svakom se koraku atributi dodaju u podskup (ako

je početni skup atributa prazan), uklanjaju iz podskupa (ako je početni skup cijeli skup) ili se generira slučajni podskup (kod kojeg se atributi mogu dodavati, uklanjati ili se u svakoj novoj iteraciji generira novi podskup). Dobiveni podskup se evaluira s obzirom na definiranu evaluacijsku funkciju kojom se određuje se da li je definirani podskup optimalan. Kriterij zaustavljanja se može definirati s obzirom na način generiranja podskupa ili na način evaluacije podskupa. U prvom slučaju se provjerava da li je odabran zadani broj atributa ili je proveden određeni broj iteracija. U drugom slučaju se provjerava da li dodavanje ili uklanjanje atributa iz podskupa daje bolje rezultate ili da li je dobiven optimalan podskup.
Zadnji korak je validacija. Taj korak nije dio procesa odabira, već se koristi za provjeru da li je definirani podskup valjan s obzirom na definirane potrebe (npr. usporedba dobivenih rezultata s rezultatima drugih metoda) (Dash i Liu, 1997.).

3.2.Pristupi selekciji atributa

Pregledom literature utvrđeno je da postoji mnogo pristupa problemu selekcije atributa. Ipak, u svojoj biti svi pristupi uključuju dvije temeljne komponente:

- **Strategiju traženja** koja istražuje skup svih podskupova atributa na svrhovit način
- **Kriterij** na temelju kojeg se **vrednuju** podskupovi

Strategija traženja je neovisna o kriteriju koji se koristi (Devijver, Kittler, 1982). Najbolji podskup atributa je pronađen optimiziranjem (najčešće maksimiziranjem) evaluacijske funkcije. Najbolji učinak se postiže kada se selekcija atributa i kasnije klasifikacija optimiziraju koristeći isti kriterij (Bishop, 1996.), npr. točnost klasifikacije.
Heuristika se primjenjuje u procesu traženja kako bi se smanjila kompleksnost. Stoga se selekcija atributa i konačna klasifikacija često rade odvojeno.
Strategije traženja uključuju tehnike rangiranje atributa (Guyon i Elisseeff, 2003., Kirra i Rendell, 1992.) ili traženja podskupa (Devijver, Kittler, 1982, Pudil et. al., 1994.). Oba pristupa mogu biti temeljena na deterministčkim ili slučajnim principima koji usmjeravaju traženje kroz skup atributa.

S obzirom na način pretraživanja atributa, tehnike se dijele u tri grupe (Dash i Liu, 1997., Yang i Honavar, 1997.):

- Cjelovito pretraživanje – pretražuje se kompletan skup atributa, a na temelju evaluacijske funkcije se određuje najbolji podskup. S ciljem smanjivanja prostora traženja, koristi se princip povratnog pretraživanja koji garantira da dodavanje atributa ne pogoršava učinkovitost. Metode koje koriste ovaj princip su metode grananja i ograničavanja (Foroutan i Sklansky, 1987). Iako te metode selekcije atributa daju zadovoljavajuće rezultate kad se kao klasifikatori koriste tradicionalne statističke metode, rezultati su izrazito loši kada se koriste nelinearni klasifikatori (npr. neuronske mreže).
- Heurističko pretraživanje – karakterizira ga upotreba heuristike u pretraživanju. Primjeri heurističkog pretraživanja su: odabir najboljeg atributa za slijedeći korak (eng. stepwise forward selection) (Kohavi i John, 1997.), eliminacija najgoreg atributa u slijedećem koraku (eng. eng. stepwise backward elimination) (Kohavi i John, 1997.),, Relief (Kirra i Rendell, 1992., John et al., 1994.).
- Slučajno pretraživanje – podskup se generira slučajno, a pretražuje se onaj broj podskupova koji je prethodno zadan. U ovakvo pretraživanje svrstavaju se genetski algoritimi (Jang i Honavar, 1997.) i LVF (Dash i Liu, 1997.).

Evaluacijska funkcija se koristi za određivanje relevantnosti atributa. Cilj primjene funkcije je mjerenje sposobnosti atributa (ili skupa atributa) da uzorke podataka svrsta u neku od klasa. Relevantnost atributa ovisi o evaluacijskoj funkciji te se odabirom različitih evaluacijskih funkcija dobivaju različiti podskupovi kao optimalni.
S obzirom na ono što mjere, evaluacijske funkcije se dijele u pet grupa (Dash i Liu, 1997.):

1. Mjera udaljenosti
2. Mjera informacija
3. Mjera ovisnosti
4. Mjera dosljednosti
5. Mjera pogreške klasifikatora

Mjera udaljenosti određuje relevantnost atributa na temelju vjerojatnosne udaljenosti između gustoća uvjetne vjerojatnosti pripadnost uzorka jednoj od klasa. Kao takva, ova mjera

zahtjeva informaciju o pripadnosti pojedine instance klasi (Liu i Yu, 2005.). Primjeri mjera udaljenosti su: Euklidska mjera udaljenosti i Chernoffova mjera udaljenosti.

Mjera informacije određuje informacijsku dobit atributa. Jedan atribut je relevantniji od drugog ako je informacijska dobit prvog atributa veća od informacijske dobiti drugog atributa (Dash i Liu, 1997.). Primjer mjere informacije je Shannonova mjera entropije.

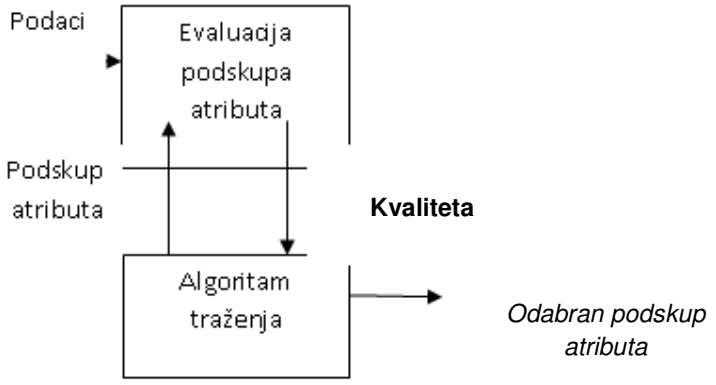
Mjera ovisnosti određuje relevantnost atributa na temelju korelacije nezavisnog atributa i zavisnog (klase). Atribut *A* je relevantniji od atributa *B*, ako je atribut *A* u većoj korelaciji sa zavisnim od atributa *B* (Dash i Liu, 1997.).

Mjera dosljednosti traži minimalan skup atributa koji daje dovoljno veliku dosljednost na podskupu skupa uzoraka. Ova mjera se jako oslanja na informaciju o pripadnosti klasama i nastoji pronaći minimalni broj atributa koji razdvajaju klase toliko dosljedno kao cijeli skup atributa (Liu i Yu, 2005.).

Od navedenih mjera dvije koriste oznaku klase u računanju (mjera udaljenosti i mjera dosljednosti), a dvije ju ne koriste (mjera informacije i mjera ovisnosti).

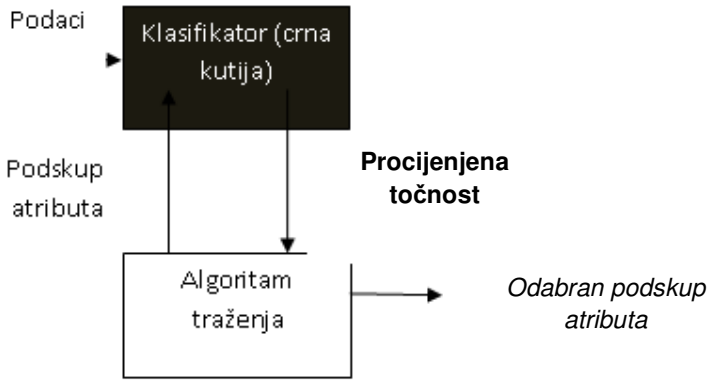
Mjera pogreške klasifikatora je točnost koju odabrani podskup atributa daje za zadani klasifikator. Kod ove mjere klasifikator služi kao evaluacijska funkcija za odabir relevantnih atributa.

Postoji velik broj različitih tehnika selekcije atributa. Dash i Liu kategorizirali su ih na temelju načina generiranja podskupa i evaluacijske funkcije koju koriste (Dash i Liu, 1997). Ipak, najčešća podjela tehnika selekcije atributa je na filter tehnike (eng. *filter techniques*) i tehnike omotača (eng. *wraper techniques*). Temeljna razlika između ova dva pristupa očituje se u načinu na koji vrednuju podskup (Kohavi i John, 1997). Tehnike omotača uključuju selekciju atributa u izgradnju modela, dok je kod filter tehnika evaluacija podskupa neovisna o algoritmu učenja (Blum i Langley, 1997; Kohavi i John, 1997). Pojmove "tehnike filtra" i "tehnike omotača" uveo je John 1994.godine (John et. al., 1994.). Blok dijagrami koji prikazuju osnovni princip obiju skupina tehnikama prikazani su na slikama 4. i 5.



Slika 4. Princip selekcije atributa kod metoda filtra

Izvor: John et. al., 1994



Slika 5. Princip odabira atributa kod metoda omotača

Izvor: John et. al., 1994

Principi rada obiju skupina tehnika objašnjavaju se u podpoglavljima 3.2.1 i 3.2.2.

3.2.1.Tehnike filtra

Tehnike filtere se mogu koristiti kao predkorak za kasnije učenje. Vrlo često koriste heuristički pristup kod kojeg evaluacijska funkcija nije direktno vezana na učinkovitost određenog klasifikatora. Umjesto toga, rezultat ovisi o unutrašnjim svojstvima podataka. Atributi se evaluiraju s obzirom na kriterije kao što su mjera udaljenosti, Pearsonov koeficijent korelacije, entropija ili neke druge mjere informacije (Devijver i Kittler, 1982; Guyon i Elisseeff, 2003). Možemo reći da filter tehnike predstavljaju opći pristup selekciji atributa dajući rješenje prikladno za velik skup klasifikatora. Filter tehnike su uglavnom vrlo brze i kao takve izrazito korisne za visoko dimenzionalne probleme gdje druge tehnike nisu konkurentne s obzirom na računalnu kompleksnost. Filtri koriste statističke mjere za „filtiranje“ atributa koji nisu potrebni, prije izrade modela. No, selektirani optimalni atributi nužno nisu garancija najbolje učinkovitosti klasifikatora.

3.2.2.Tehnike omotača

Tehnike omotača vrednuju podskup atributa procjenjujući točnost algoritma učenja. Strategija traženja koristi točnost predikcije kao funkciju koja vodi potragu za najboljim podskupom te nastoji pronaći one atribute koji maksimiziraju točnost. Algoritam učenja djeluje kao crna kutija što tehnike omotača čini jednostavnim i univerzalnim. Naravno, atributi su optimizirani za prethodno odabranu tehniku te vrlo vjerojatno nisu optimalni za drugi algoritam učenja. Tehnike omotača zahtijevaju puno računanja jer velik broj modela klasifikacije mora biti napravljen tijekom procesa traženja najboljeg podskupa atributa. Brzina bi se mogla postići upotrebom efikasne strategije traženja. Ali to traženje postaje gotovo nemoguće s porastom dimenzionalnosti, osobito kad se radi s računalno intenzivnim tehnikama za učenje. Kod tehnika omotača često dolazi i do pretreniranosti. Unatoče tome, neki autori (John et. al., 1994., Kohavi i John, 1997.) su pokazali da daju veću točnost od tehnika filtra.

U poglavlju 3.3. opisuju se tehnike selekcije atributa koje će se koristiti u istraživanju, a to su: informacijska dobit, omjer dobiti, *Relief* i linearni odabir unaprijed.

3.3. Opis odabranih tehnika selekcije atributa

Ovo podpoglavlje daje kratak teorijski pregled tehnika selekcije atributa koje će se koristiti u ovom istraživanju. Tehnike su izdvojene s obzirom na nekoliko kriterija:

- predstavljaju različite pristupe selekciji atributa
- često su korištene u prethodnim istraživanjima
- predstavljaju referentne tehnike selekcije atributa za zadaću klasifikacije

Referetne tehnike selekcije atributa prema istraživanju Hall i Holmesa su informacijska dobit i *Relief* (Hall i Holmes, 2003.), a prema istraživanju Gaucheva i suradnika referentne tehnike selekcije atributa su: informacijska dobit i omjer dobiti (Ganchev et. al., 2006.).

U nastavku se navode neka od istraživanja koja su primjenjivala tehnike selekcije atributa koje će se komparirati u ovom istraživanju :

- *informacijska dobit:* Novakovic, Strbac i Bulatovic, 2011., Oreski, Oreski i Oreski, 2012., Sun et. al., 2012.. Silva et. al., 2013.,
- *omjer dobiti*: Novakovic, Strbac i Bulatovic, 2011., Oreski, Oreski i Oreski, 2012., Sun et. al., 2012. Silva et. al., 2013.,
- *Relief* :Novakovic et al, 2011. Novakovic, Strbac i Bulatovic, 2011., Kirra i Rendell, 1992., Dietterich, 1997., Silva et. al., 2013.,
- *linearni odabir unaprijed:* Novakovic et al, 2011. Novakovic, Strbac i Bulatovic, 2011., Pudil, P., Novovičova, J., Kittler, J., 1994. Oreski, Oreski i Oreski, 2012.

3.3.1. Informacijska dobit

Informacijska dobit (eng. *information gain*) je tehnika selekcije atributa koja radi rangiranje atributa. Temelji se na radu Claudea Shannona u području teorije informacija gdje je proučavao „informacijski sadržaj“ poruke. Atribut s najvećom vrijednosti informacijske dobiti minimizira potrebnu informaciju za klasifikaciju skupa. Takav pristup minimizira očekivani broj testova u procesu klasifikacije. Očekivana informacija potrebna za klasifikaciju u nekom skupu *D* je:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

Pri čemu je p_i vjerojatnost da je proizvoljni element u *D* pripada klasi $C_{i.}$

3.3.2. Omjer dobiti

Prethodno opisana tehnika, informacijska dobit, je pristrana kod testova s mnogo ishoda. Kako bi se prevladao ovaj nedostatak razvijena je tehnika *omjer dobiti* (eng. *information gain*). Ova tehnika primjenjuje određenu vrstu normalizacije u odnosu na informacijsku dobit koristeći nazvanu „dijeljenje informacije“ (eng. *split information*) i definiranu kao:

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2(\frac{|D_j|}{|D|}).$$

Omjer dobiti se definira kao:

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)}$$

Atribut s većom vrijednošću omjera dobiti bolje je rangirani.

3.3.3. Tehnika *Relief*

Neke tehnike selekcije atributa temelje se na procjeni atributa. Procjena znači dodjeljivanje vrijednosti važnosti svakom atributu te odabir atributa koji imaju najveću vrijednost važnosti. Najpoznatija takva tehnika selekcije atributa je *Relief*. Relief algoritam bolje radi procjenu od drugih statističkih procjena (korelacije ili kovarijance) jer uzima u obzir međuodnose atributa (Kira i Rendell, 1997., Arauzo-Azofra, Benitez i Castro, 2004.)

U ovom kontekstu, kada se govori o važnosti, misli se na važnost atributa za izlazni atribut. Vrijednost važnosti svakog atributa odražava njegovu sposobnost da razlikuje klase. Atributi su rangirani po važnosti i oni koji imaju veću važnost od prethodno definirane korisničke vrijednosti ulaze u odabrani podskup.

Relief je tzv. radnomizirani algoritam (eng. *randomized algorithm*) jer radi slučajno uzorkovanje skupa za treniranje i ažurira važnost temeljem razlike između selektirane instance i dvije najbliže instance iste i suprotne klase. Relief nastoji pronaći sve relevantne atribute (Kohavi i John, 1997.). Kira i Rendell eksperimentalno dokazuju da *Relief* algoritam efikasno identificira relevantne atribute kada su interakciji, međutim, ne radi dobro s redundantnim atributima (Kirra i Rendell, 1997.). „Ako je većina atributa u skupu relevantna za izlaznu varijablu, Relief algoritam selektira većinu danih atributa iako je mali broj atributa doista nužan za opisivanje izlaza.“ (Kirra i Rendell,

Relief algoritm radi samo na problemima s dvije klase zavisne varijable (Čehovin i Bosnić, 2010.)

U realnim skupovima podataka mnogi atributi imaju visoku korelaciju sa zavisnom varijablom te su mnogi atributi slabo relevantni i neće biti maknuti od strane Relief algoritma.

Pretpostavimo da su primjeri $I_1, I_2, \dots, I_N$ u prostoru instanci i opisani su vektorima atributa $A_I, i = 1, \dots a$ gdje je a broj atributa, a izlazna varijabla je τ_j. Znači, instance su točke u *a-dimenzionalnom* prostoru (Robnik-Šikonja, Kononenko, 2003.)

Pseudokod temeljenog Relief algoritma dan je u nastavku:

Ulaz: za svaku instancu vrijednost vektora atributa i vrijednost klase

Izlaz: vektor W procjena kvalitete atributa

1. Postaviti sve težine $W[A] := 0.0$;

2. **Z**a **i:=1 do** m **radi slijedeće**

3. Nasumično odabrati instancu R_i;

4. Naći najbliži pogodak *H* i najbliži promašaj *M*;

5. **Za** A:=1 **do** *a* raditi slijedeće:

6. $W[A] := W[A] - \frac{diff(A,R_i,H)}{m} + \frac{diff(A,R_i,M)}{m}$;

7. **Kraj**;

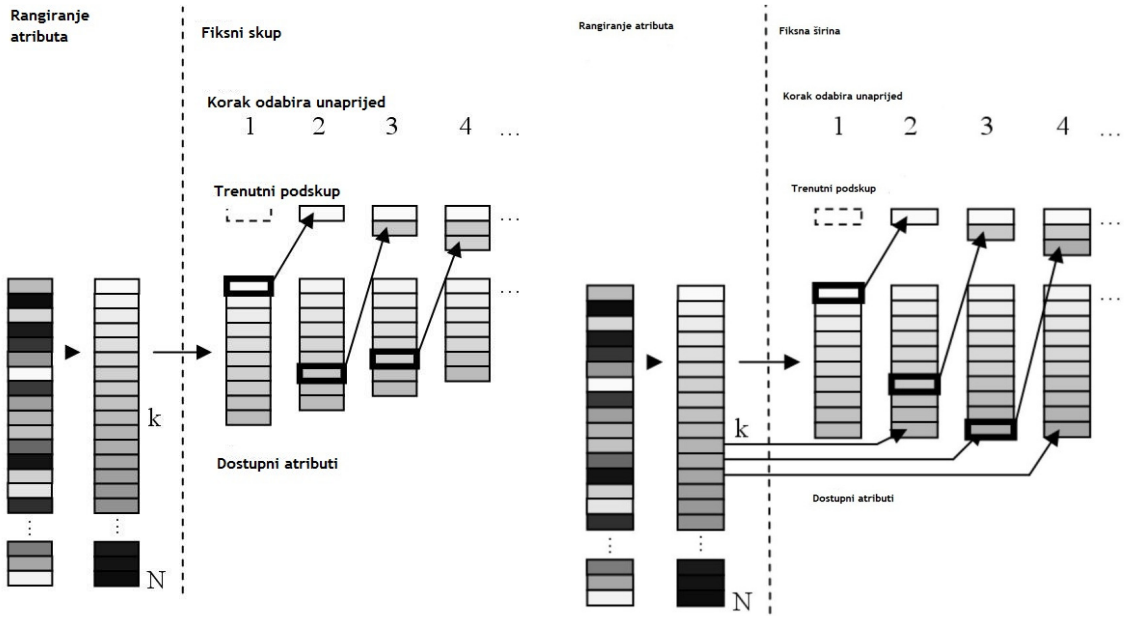
(Robnik-Šikonja, Kononenko, 2003.)

Kao što je već rečeno, osnovna ideja algoritma je procijeniti kvalitetu atributa s obzirom na to koliko dobro vrijednosti atributa razlikuju instance koje su jedna blizu druge. „U tu svrhu, za nasumično izabranu instancu R_i (linija 3), Relief traži dva najbliža susjeda: jednog iz iste klase, koji se naziva *najbliži pogodak H*, i jednog iz druge klase, koji se naziva *najbliži promašaj M* (linija 4). Dalje se ažurira procjena kvalitete *W[A]* za sve atribute *A* ovisno o njihovoj vrijednosti za R_i, *M* i *H* (linije 5 i 6). Ako instance R_i i *H* imaju različite vrijednosti atributa *A* tada atribut *A* razdvaja dvije instance unutar iste klase što nije poželjno te se smanjuje vrijednost vrijednost procjene kvalitete *W[A]*. S druge strane, ako instance R_i i *M* imaju različite vrijednosti atributa A tada atribut A razdvaja dvije instance iz različitih klasa, što i hoćemo postići, pa se povećava vrijednost procjene kvalitete *W[A]*. Cijeli proces se ponavlja *m* puta, pri čemu je *m* parametar definiran od strane korisnika.“ (Robnik-Šikonja, Kononenko, 2003.)

3.3.4. Tehnika *linearni odabir unaprijed*

Linearni odabir unaprijed traženje počinje s praznim podskupom i dodajući atrubut po atribut vrednuje sva moguća proširenja postojećeg podskupa. Atribut koji daje najbolji rezultat dodaje se u podskup. Traženje prestaje kada više nema atributa koji bi poboljšao rezultat. U slučaju klasičnog sekvencijalnog odabira unaprijed broj vrednovanja kvadratno raste s brojem atributa: broj vrednovanja u svakom koraku je jednak broju preostalih atributa koji trenutno nisu u podskupu. Taj podskup raste u svakom koraku, sve dok algoritam ne završi. Brz rast

broja vrednovanja je problematičan kod skupova podataka s velikim brojem atributa. Stoga se koristi pristup linearni odabir unaprijed kod kojeg se limitira broj atributa koji se razmatraju u svakom koraku da se ne prelazi od korisnika definirana konstanta. Time se uvelike reducira broj evalucija i poboljšava vrijeme izvođenja algoritma. Linearni odabir unaprijed koristi jedan od dva načina za ograničavanje broja atributa: fiksni broj atributa (eng. *fix set*) ili fiksna širina (eng. *fix width*). Kod prvog pristupa inicijalno se rangiraju svi atributi i odabire se prvih *k* atributa koji služe kao ulaz u daljnju selekciju. Inicijalno rangiranje odvija se vrednovanjem svakog atributa posebno i rangiranjem s obzirom na dobivene vrijednosti. Samo je *k* najboljih atributa uključeno u kasniji odabir unaprijed, dok se ostatak atributa odbacuje. Ova jednostavna metoda skraćuje gornju granicu broja evaluacija koje se razmatraju tijekom traženja na broj od $\frac{1}{2} \times k(k+1)$, neovisno o početnom broju atributa. Motivacija za korištenje ovog pristupa može se naći u činjenici da miče većinu atributa koji nisu relevantni, pa se algoritam dalje fokusira samo na relevantne. Nedostatak ovog pristupa je da se odbacuju atributi koji imaju manju relevantnost kad djeluju sami, ali bi mogli povećati točnost klasifikacije kada bi djelovali s drugim atributima. Također, *k* možda nije dovoljan broj atributa. Ovaj pristup kod tehnike *linearni odabir unaprijed* naziva se *fiksni skup* jer su dostupni atributi reducirani na fiksni skup od *k* atributa. Ovaj pristup prikazan je s lijeve strane na slici 6.



Slika 6. Principi tehnike linearni odabir unaprijed

Izvor: Gütlein et. al, 2009

Drugi pristup prikazan je s desne strane slike 6. i naziva se *fiksna širina*. Ovaj pristup zadržava broj proširenja u svakom koraku odabira unaprijed konstantnim, na fiksnoj širini *k* (slika). Inicijalno rangiranje se radi kao i kod prethodnog pristupa i potraga kreće s *k* top atributa. U svakom koraku broj atributa se povećava za 1 tako da se dodaje slijedeći najbolje rangirani atribut. Time se osigurava da se skup kandidata za proširenje sastoji od atributa koji su individualno najbolji, a dosad nisu bili odabrani. Kod ovog pristupa teoretska gornja granica broja evalucija je: $\boldsymbol{N \times k - \frac{1}{2} \times k(k-1)}$. Ovaj pristup uzima u obzir jako relevantne atribute u prvom koraku, ali i atribute s manjom relevantnošću u slijedećim koracima (Gütlein et. al, 2009.).

3.4. Pregled dosadašnjih istraživanja

Selekcija atributa je vrlo aktivno područje u računalnim znanostima (Novaković, Strbac, i Bulatović, 2011). i dosad je razvijeno mnogo tehnika u tu svrhu, a i napravljeno komparativnih studija koje ih analiziraju i međusobno uspoređuju. U ovom pregledu dosadašnjih istraživanja fokus neće biti na predstavljanju razvijenih tehnika, nego na istraživanjima koja su ih uspoređivala, s naglaskom na opseg istraživanja (u smislu broja analiziranih skupova podataka, broja korištenih tehnika i algoritama učenja te kriterija i aspekata usporedbe). Pregled je dan kronološki.

John i suradnici predlažu tehniku selekcije atributa temeljenu na unakrsnom vrednovanju, a primjenjiva je na bilo koji algoritam učenja. Preloženu tehniku testiraju koristeći C4.5 kao algoritam učenja (John, Kohavi i Pfleger, 1994.)

Na 18 skupova podataka dvije tehnike (odabir unaprijed i odabir unatrag) uspoređuju Kohavi i Sommerfield (Kohavi i Sommerfield, 1995.).

Koller i Sahami uvode tehniku selekcije atributa koja se bazira na teoriji informacija. Predloženu tehniku testiraju na 5 skupova podataka i dokazuju da je učinkovita na skupovima podataka s velikim brojem atributa (Koller i Sahami, 1996.).

Kohavi i John uvode tehnike omotača i uspoređuju ih s Relief algoritmom. Kao algoritme učenja koriste stablo odlučivanja i naivni Bayesov klasifikator (Kohavi i John, 1997).

Dash i Liu u članku iz 1997. daju pregled dotad poznatih tehnika selekcije atributa. Kategorizirali su tehnike s obzirom na način generiranja podskupa i funkcije vrednovanja. U komparativnoj analizi sedam tehnika selekcije atributa koristili su 3 skupa podataka. Kao rezultat istraživanja daju smjernice koje tehnike selekcije atributa koristiti u kojoj domeni (Dash i Liu, 1997.).

Jain i Zongker u istraživanju iz 1997. dokazuju da je algoritam korak unaprijed superioran u odnosu na druge korištene algoritme (Jain i Zongker, 1997.)
Weston uvodi tehniku selekcije atributa za SVM i testira ju na 4 skupa podataka. Tehnika se pokazuje superiornom u odnosu na ostale na testiranim skupovima podataka (Weston et. al, 2001.).
Liu i suradnici testiraju Relief algoritam na 16 podataka (Liu, Motoda i Yu, 2002.). Autori uzimaju u obzir standardne mjere karakteristika skupa podataka o kojima će kasnije u ovom radu biti riječi.

Geng i suradnici uvode novu tehniku selekcije atributa koja se temelji na pronalaženju sličnosti između dva atributa. Predloženu tehniku testiraju na 2 skupa podataka (Geng et. al, 2007.)

Alibeigi i suradnici predlažu novu tehniku selekcije atributa koju definiraju kao filter tehniku i uspoređuju ju s tri poznate tehnike selekcije atributa samo na 3 skupa podataka (Alibeigi, Hashemi i Hamzeh, 2009.). Nadalje, u njihovom istraživanju u obzir uzimaju samo standardne mjere karakteristika podataka (broj atributa i broj instanci).

Janecek u doktorskoj disertaciji iz 2009. komparira tehnike selekcije atributa na 3 skupa podataka iz 2 domene (klasifikacija emaila i otkrivanje droga).

Drugan i Wiering na 15 skupova podataka testiraju 1 tehniku selekcije atributa za klasifikaciju Bayesovim mrežama (Drugan i Wiering, 2010.).
Čehovin i Bosnić (Čehovin i Bosnić, 2010.) uspoređuju 5 tehnika selekcije atributa: ReliefF, *random forest feature selector, sequential forward selection*, *sequential backward selection* i

Gini index kroz točnost klasifikacije na čak 6 klasifikatora, između ostalih i stabla odlučivanja, neuronske mreže i naivni Bayesov klasifikator.

Lavanya i Usha Rani uspoređuju performance tehnika selekcije atributa na 3 skupa podataka koji se odnose na rak dojke. Dobiveni rezultati vode do zaključka da nijedna tehnika selekcije atributa nije superiorna na sva tri skupa podataka. Autori utvrđuju da izbor tehnike selekcije atributa ovisi o broju atributa u skupu podataka te broju instanci (Lavanya i Usha Rani, 2011.). Nadalje, isti autori navode točnost klasifikacije i brzinu kao kriterije za usporedbu, no u prvi plan, kao najvažniji kriterij, ističu točnost klasifikacije.

Novakovic i suradnici uspoređuju 6 tehnika selekcije atributa na 2 skupa podataka i pri tome kao kriterij za komparaciju uzimaju točnost klasifikacije. (Novakovic, Strbac i Bulatovic, 2011.)
Haury i suradnici uspoređuju 8 tehnika selekcije atributa na 4 skupa podataka. (Haury, Gestraud i Vert, 2011)

Silva i suradnici uspoređuju 4 postojeće tehnike selekcije atributa (Informacijska dobit, omjer dobiti, hi kvadrat, korelaciju) na 1 skupu podataka u domeni poljoprivrede (Silva et. al., 2013).

Pregled istraživanja je pokazao da se tehnike selekcije atributa ocjenjuju na temelju vremena koje im je potrebno za izvedbu i kvalitete odabranih podskupova atributa (Jain, Zongker, 1997.). Metodologija za vrednovanje rezultata nije standardizirana i razlikuje se od članka do članka. Stoga je vrlo teško izvući zaključke ili napraviti usporedbu između tehnika selekcije atributa.

Kao kriteriji kod uspredbe tehnika uglavnom su se koristili učinkovitost klasifikacije, i vrijeme izvršenja algoritma. Jain i Zongker tvrde da je vrijeme izvršenja algoritma manje važan kriterij od konačne učinkovitosti klasifikacije (Jain, Zongker, 1997.). Iako je ovo točno, novije primjene procesa otkrivanja znanja u podacima uključujuju podatke s tisućama atributa. U takvim slučajevima računalni zahtjevi odabrane tehnike postaju izrazito važni.

Tehnike omotača često daju točnije rezultate od metoda filtra, ali je vrijeme izvršenja puno veće. Stoga u problemima s nekoliko tisuća atributa tehnike omotača nisu ni primjenjive. Dok neki autori navode da je najveći nedostatak tehnika filtra zanemarivanje učinka odabranog podskupa na točnost algoritma učenja (Guyon i Elisseeff, 2003.), drugi autori (Abe et al. 2006.) nezavisnost tehnike selekcije atributa od algoritma učenja navode kao prednost iz razloga jer je najbolje odabrati podskup atributa koji daje dobre rezultate na više klasifikatoras (Abe et. al., 2006.).

3.4.1.Nedostaci prethodnih istraživanja

Na temelju prethodnih empirijskih komparacija vrlo je teško izvući zaključke o superiornosti jedne tehnike selekcije atributa nad drugom. Rezultati jednog istraživanja su često u izravnoj kontradikciji s rezultatima nekog drugog istraživanja. Stoga je nemoguće tvrditi da je neka tehnika brža ili točnije od druge. U ovom radu prepoznaju se ti nedostaci i dokazuje se da se rješenje nalazi u nekoj vrsti kategorizacije karakteristika skupova podataka. Tako bi se mogli donijeti zaključci da je za *tu-i-tu* vrstu skupa podataka *taj-i-taj* algoritam bolji.

Pregledavajući komparativne analize tehnika selekcije atributa iz prethodnih istraživanja uočava se njihov glavni nedostatak, a to je nedovoljna opsežnost. Detaljnije, u prethodnim istraživanjima uočeni su slijedeći nedostaci:

- izbor tehnika selekcije atributa je vrlo uzak
- koristi se jedan klasifikator i ne može se utvrditi kako koja tehnika selekcije atributa djeluje na kojem klasifikatoru
- skupovi podataka su obično mali i/ili simulirani i kao takvi ne predstavljaju stvarne probleme
- ne uzimaju se u obzir karakteristike skupa podataka
- broj skupova podataka na kojima se radi selekcija atributa je jako mali
- samo jedan kriterij je korišten kod usporedbe

Navedeni nedostaci se u ovom istraživanju nastoje izbjeći kroz sveobuhvatno istraživanje koje predstavlja korak naprijed u odnosu na prethodne komparativne analize u slijedećim aspektima:

- komparacija tehnika se provodi na 128 skupova podataka

- uzima se u obzir 5 grupa karakteristika podataka
- komparira se 7 tehnika selekcije atributa
- u procesu učenja koriste se 2 klasifikatora različitih pristupa učenju

4. TEHNIKE OTKRIVANJA KONTRASTA

„*There is no quality in this world that is not what it is merely by contrast. Nothing exists in itself.*“

Herman Melville

Područje rudarenja podataka jedna je od najzanimljivijih tehnologija informacijskih znanosti u 21. stoljeću (Dong i Bailey, 2012.). Ovo područje danas je neizostavan mehanizam u interpretaciji informacija skrivenih u velikim skupovima podataka. Visoka zastupljenost u širokom spektru domena vodi do značajnog razvoja područja i razvoja specijaliziranih podpodručja unutar rudarenja podataka. Jedan od najnovijih i najinteresantnijih podskupa jest područje *otkrivanja kontrasta* (eng. *contrast mining*).

Boettcher i suradnici ističu da su tehnike rudarenja podataka orjentirane na analizu statičkog svijeta u kojem se podaci prikupljaju, spremaju i analiziraju kako bi se dobili modeli koji opisuju sadašnjost (Boettcher et. al. 2011.). Mnogi autori ističu da je otkrivanje kako se nešto u domeni mijenja jednako važno kao i stvaranje preciznih modela (npr. Boetcher et al., 2011). Štoviše, temeljnom zadaćom analize podataka smatra se razumijevanje razlika između kontrastnih grupa. Stoga je danas razvijanje metoda za analizu i razumijevanje razlika jedno od glavnih pitanja kada se radi s podacima (Gaber, Zaslavsky i Krishnaswamy, 2005.).

Vođeno ovim potrebama, unutar rudarenja podataka razvija se područje otkrivanja kontrasta (eng. *contrast mining*). Počeci razvoja područja traženja kontrasta između grupa bili su 1999.g., a danas to je jedno od najizazovnijih i najvitalnijih područja u domeni rudarenja podataka (Boettcher, 2011.). Odgovore na pitanja zašto se danas ovo područje vrlo intenzivno istražuje moglo bi se naći u slijedećim izjavama:

“*There is no quality in this world that is not what it is merely by contrast. Nothing exists in itself.”*

Herman Melville

„*Sometimes it is good to contrast what you like with something else. It makes you appreciate it even more*.“

Darby Conley

Generalni zaključak koji se može izvući iz ovih izjava je slijedeći: kada neki objekt uspoređujemo s drugim, možemo otkriti više informacija. Slijedom toga dolazi se do ideje upotrebe tehnika otkrivanja kontrasta u svrhu selekcije atributa.

U području otkrivanja kontrasta, kontrast između grupa se izražava kroz *kontrastni skup* (eng. *contrast set*). Kontrastni skup se definira kao kombinacija parova atribut-vrijednost. Cilj otkrivanja kontrasta je pronaći kontrastni skup koji jednu grupu značajno razlikuje od druge. Formalni opis algoritama STUCCO i Magnum Opus daje se u slijedeća dva poglavlja, a u poglavlju 4.4. se uspoređuju.

4.1. STUCCO algoritam

Problem otkrivanja kontrasta prvi put je predložen od strane Baya i Pazzanija kao pronalaženje kontrastnih skupova koji predstavljaju „vezu atribut-vrijednost koja značajno razlikuje distribucije među grupama“ (Bay i Pazzani, 1999.). STUCCO (*Search and Testing for Understandable Consistent Contrast*) algoritam prvi je razvijen u te svrhe (Bay i Pazzani, 1999.).

STUCCO algoritam pretražuje stablo sa svim mogućim kontrastnim skupovima. U korijen se postavlja prazan skup, a čvorovi djeca su atributi koji nas zanimaju. Dalje se stablo grana na kombinacije atributa. Pretraživanje se vrši na način da se kreće od najopćenitijih čvorova prema specifičnijim, odnosno od pojedinačnih atributa prema kombinacijama. Svaki čvor će tijekom pretraživanja biti posjećen samo jednom, ili nijednom ukoliko dođe do potkresivanja (eng. *pruning*) stabla.

Za svaki kontrastni skup ispituje se podrška (eng. *support*) pojedinih grupa. Podrška kontrastnog skupa od strane grupe jednaka je broju primjeraka iz te grupe čije se vrijednosti podudaraju s kontrastnim skupom podijeljenog kardinalnim brojem te grupe. tj. podrška je

postotak podržanih entiteta neke grupe za kontrastni skup. Npr. ako imamo podatke o izborima gdje su grupe različiti gradovi, podrška za kontrastni skup

{socijalna kategorija='siromašan' AND izbor = '1'}

može biti sljedeća:

*Podrška*GRAD=A({socijalna kategorija='siromašan' AND izbor = '1'})=15%

*Podrška*GRAD=B{(socijalna kategorija='siromašan' AND izbor = '1'})= 60%

Dakle, 15% siromašnih birača koji su odabrali opciju 1 su iz grada A, dok ih je za istu opciju iz grada B 60%.

Pomoću dobivene podrške provjeravamo razlike među grupama te izdvajamo one koje su dovoljno značajne i zanimljive.

Formalno definirano, to izgleda ovako:

Skup podataka je skup grupa $G_1, G_2 \dots G_l$. Svaka grupa se sastoji od objekata $O_1 \dots O_u$. Svaki objekt O_i je skup k parova atribut-vrijednost, jedan za svaki od atributa $A_1 \dots A_k$. Atribut A_j ima vrijednosti $V_{j1} \dots V_{jm}$.

Pri tome se mjeri *podrška* kontrastnog skupa. Podrška se definira u odnosu na svaku grupu. Podrška kontrastnog skupa u odnosu na grupu G_i je proporcija objekata $o \in G_i$ i označava se kao $podrška(cset, G_i)$. STUCCO algoritam je definiran tako da traži kontrastne skupove čija podrška se značajno razlikuje među grupama. Točnije, traže se kontrastni skupovi koji zadovoljavaju slijedeće:

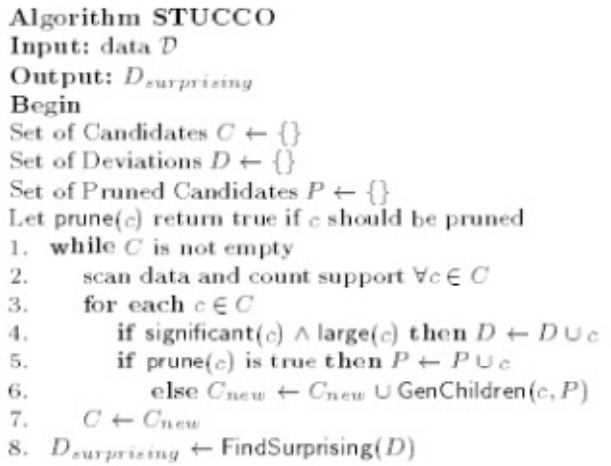
$$\exists ij P(cset|G_i) \neq P(cset|G_j)$$

i

$$\max(i,j)\,|podrška(cset, G_i) - podrška\big(cset, G_j\big)| \geq$$

Gdje je δ korisnički definirana granica nazvana *minimum podrške razlici* (eng. *minimum support –difference*). Kontrastni skupovi za koje je prva jednadžba statistički značajna se zovu *značajni* (eng. *significant*), a oni za koje je zadovoljena druga jednadžba zovu se *veliki* (eng. l*arge*). Kada su zadovoljene obje, taj kontrastni skup se naziva *devijacija*.

Statistička značajnost prve jednadžbe je procijenjena korištenjem *hi-kvadrat* testa kako bi se procijenila nul hipoteza o čijim je članovima grupe neovisna podrška kontrastnim skupovima. Pseudo kod STUCCO algoritma dan je na slici 7.



Slika 7. Pseudo kod STUCCO algoritma

Izvor: Bay i Pazzani, 1999.

STUCCO ima dodirnih točaka s tehnikom *Odds ratio*, koji se pokazao uspješan za selekciju atributa u klasifikaciji teksta (Mladenić,2006.). Sličnost ova dva pristupa je u slijedećem: ono što je kod STUCCO algoritma atribut-vrijednost, kod *Odds ratio* je ekvivalentno pojavi riječi u dokumentu (riječ = atribut, pojava riječi =vrijednost binarnog atributa je 1).

4.2. Magnum Opus

Magnum Opus predstavlja implementaciju OPUS algoritma. OPUS (engl. ***O**ptimized **P**runing for **U**nordered **S**earch)* je optimalna pretraga za pretraživanje nesortiranog prostora. Ovaj pristup je poznat je po tome što ima sposobnost uspješnog pronalaženja zadanog broja pravila koji maksimiziraju proizvoljnu funkciju mjereći kvalitetu pravila. OPUS je algoritam koji omogućuje učinkovitu pretragu kroz zadani skup podataka u kojem redosljed pretraživanja nije bitan. Algoritam pretraživanja je vrlo učinkovit s obzirom na postojeće algoritme. Pseudo kod OPUS algoritma dan je na slici 8.

```
Algorithm:  OPUS_AR(CurrentLHS, AvailableLHS,
AvailableRHS)

 com CurrentLHS is the set of conditions in the LHS
     of the rule currently being considered.

 com AvailableLHS is the set of conditions that may
     be added to the LHS of rules to be explored
     below this point

 com AvailableRHS is the set of conditions that
     may appear on the RHS of a rule in the search
     space at this point and below

  1. SoFar := {}

  2. FOR EACH P in AvailableLHS

     (a) NewLHS := CurrentLHS ∪ {P}

     (b) AvailableLHS := AvailableLHS - P

     (c) IF pruning rules cannot determine that
         ∀x ⊆ AvailableLHS: ∀y ∈ AvailableRHS:
         ¬credible(x ∪ NewLHS → y) THEN

          i. NewAvailableRHS = AvailableRHS

         ii. FOR EACH Q in AvailableRHS

             A. IF credible(NewLHS → Q) THEN

                  record NewLHS → Q

             B. IF pruning rules determine that
                ∀x ⊆ AvailableLHS: x = {} ∨
                ¬credible(x ∪ NewLHS → Q) THEN

                  NewAvailableRHS :=
                  NewAvailableRHS - Q

        iii. IF NewAvailableRHS ≠ {} THEN
                OPUS_AR(NewLHS, SoFar,
                NewAvailableRHS)

         iv. SoFar := SoFar ∪ {P}
```

Slika 8. Pseudo kod Magnum Opus algoritma

Izvor: Webb, 2000.

Magnum Opus, implementacija Opus algoritma, objedinjuje nekoliko jedinstvenih tehnologija za otkrivanje povezanosti i time čine korak naprijed u odnosu na sam Opus algoritam. U samoj biti Magnum Opusa je upotreba k-*optimalnih* (poznata i kao *top-k*) tehnika za otkrivanje povezanosti. Većina tehnika orijentirana je na pronalaženje čestih uzoraka. Mnogi od pronađenih uzoraka nisu zanimljivi za primjenu. Suprotno tome, *k-optimalne* tehnike omogućuju korisniku da sam definira što mu čini zanimljivu povezanost i koliko (*k*) veza želi naći. Algoritam tada pronalazi *k* najkvalitetnijih asocijacija sukladno definiranim kriterijima od strane korisnika.

Korisnik odabire jedan od kriterija, a dostupni kriteriji za mjerenje kvalitete su: *snaga, podrška, utjecaj, interes* i *pokrivenost.* Mjere se definiraju u nastavku.

Neka je:

D = dani skup podataka,

X = LHS (*Left Hand Side*, lijeva strana pravila),

Y = RHS (*Right Hand Side*, desna strana pravila)

Tada se mjere za pravila definiraju na slijedeći način:

- ***pokrivenost*** ***(****eng.coverage)* – prikazuje broj slučajeva u kojima se stavka nalazi s lijeve strane pravila

$$pokrivenost(X \rightarrow Y, D) = pokrivenost(X, D)$$

- ***podrška*** *(eng. support)*- prikazuje broj slučajeva kod kojih se odnos lijeve i desne strane ponavlja

$$podrška(X \rightarrow Y, D) = pokrivenost(X \cup Y, D)$$

- ***snaga*** (eng.*Strength,Confidence*) – broj dobiven dijeljenjem iznosa *podrške* s iznosom *pokrivenosti*. Prikazuje procjenu vjerojatnosti u kojoj će se stavka u desnoj strani pravila prikazati u slučaju kad je prikazana i lijeva strana pravila

$$snaga(X \rightarrow Y, D) = \frac{podrška(X \rightarrow Y, D)}{pokrivenost(X \rightarrow Y, D)}$$

- ***interes*** *(eng. Lift)*- broj dobiven dijeljenjem iznosa podrške s iznosom *podrške* koja bi bila kad ne bi postojala veza izmedu lijeve i desne strane pravila. Viša vrijednost upućuje na jaču povezanost, dok niža vrijednost upućuje na slabiju povezanost

$$interes(X \rightarrow Y, D) = \frac{podrška(X \rightarrow Y)}{pokrivenost(X) * pokrivenost(Y)}$$

- ***utjecaj*** *(eng. leverage)* – broj dobiven razlikom iznosa podrške s iznosom podrške kada ne bi postojala veza izmedu stavki desne i lijeve strane.

$$utjecaj(X \rightarrow Y, D) = pokrivenost(X \rightarrow Y, D) * (snaga(X \rightarrow Y, D) - pokrivenost(Y, D))$$

– ***P*** – rezultat statističke procjene značajnosti pravila. Što je niža vrijednost, manja je vjerojatnost da je pravilo neispravno, ili zbog toga što su lijeva i desna strana nepovezane, ili zbog toga što jedna ili više stavki u lijevoj strani ne pridonosi povezanosti sa stavkom u desnoj strani

Ako je $p<0.05$ je statistički značajno. Ako je $p>0.05$ pravilo nije statističi značajno te postoji vjerojatnost da je rezultat slučajnosti, a ne stvarnog stanja.

4.3. Prethodna istraživanja

Iako relativno nove, tehnike otkrivanja kontrasta dosad su primjenjivane u nekoliko područja u svrhu identifikacije atributa koji daju najveći kontrast između klasa i pokazale su dobre rezultate. Tehnike otkrivanja kontrasta uspješno su primjenjene u svrhu predviđanja pacijenata s moždanim udarom i razlika u odnosu na pacijente koji imaju druge neurološke poremećaje, a imaju iste simptome kao i pacijenti s moždanim udarom (Kralj et. al., 2007.). Otkrivanje kontrasta primjenjivalo se u istraživanju razlika između uspješnih i neuspješnih studenata koristeći podatke obrazovnih sustava (Perera, 2009.)

Lin i Keogh upotrebljavaju tehnike otkrivanja kontrasta u istraživanju vremenskih serija i multimedijskih podataka (Lin i Keogh, 2006.), a Alqadah i Bhatnagar za pridjeljivanje labela klasterima nakon procesa klasteriranja (Alqadah i Bhatnagar, 2009.). An i suradnici primjenjuje otkrivanje kontrasta za identificiranje zakonitosti u velikim bazama podataka (An et. al., 2009.). Motivaciju im predstavlja brzina izvođenja algoritama za otkrivanja kontrasta. Loekito i Bailey koriste otkrivanje kontrasta u istraživanju dinamičkih promjena u podacima i dobivaju odlične rezultate (Loekito i Bailey, 2008.). Wong i Tseng rješavaju probleme povezane s traženjem negativnih kontrastnih skupova (Wong i Tseng, 2005.)

Autori STUCCO algoritma, Bay i Pazzani, rade evaluaciju algoritma na dva skupa podataka. Kao izvor podataka služi im repozitorij sveučilišta California. Prvi skup koriste za ispitivanje

razlika između doktora i prvostupnika u 14 varijabli. Na drugom skupom provode STUCCO algoritam s ciljem ispitavanja promjena u broju aplikanata na sveučilište California u razdoblju od 1993. do 1998. (Bay i Pazzani, 1999.) U ovom istraživanju koristili su 17 ulaznih varijabli.

Nazari et al. primjenjuju STUCCO algoritam u predviđanju avionskih nesreća i incidenata. Cilj primjene ove tehnike rudarenja kontrastnih skupova bio je analiza nesreća u suprotnosti s incidentima, točnije, nastojalo se identificirati obrasce koji ukazuju na nesreće. (Nazeri et. al, 2008.). STUCCO algoritam koriste kako bi proveli četiri grupe analiza. U svakoj analizi, vektori nesreća su upareni s vektorima incidenata iz jedne od četiri baze podataka incidenata. Svaka analiza identificira uzorak faktora koji su značajno asocirani s nesrećama (ili incidentima). Podaci korišteni u analizi sastoje se od nesreća i incidenata koji su se događali komercijalnim letovima u periodu od 1995. do 2004. godine. Nesreće su prikupljene iz jedne baze podataka koja sadrži zapise o svim nesrećama. Incidenti su prikupljeni iz četiri baze koje sadrži zapise incidenata. U analizu su bila uključena 184 incidenta.

Simeon i Hilderman kompariraju STUCCO algoritam s COSINE algoritmom (jedan od najnovojih algoritama razvijen u svrhu traženja kontrasta) na četiri skupa podataka čije karakteristike su dane u tablici u nastavku (Simeon i Hilderman, 2011.).

Hilderman i Peckham u svom radu iz 2007. u kojem analiziraju statističku pozadinu STUCCO algoritma također koriste skupove podataka iz repozitorija sveučilišta California (3 skupa podataka).

4.4.Usporedba tehnika STUCCO i Magnum Opus

U ovom poglavlju uspoređuju se dvije ovdje korištene tehnike otkrivanja kontrasta, STUCCO i Magnum Opus. Glavni aspekt u kojem se ove tehnike razlikuju je u primjeni filtera koji identificiraju kontrastne skupove. Magnum Opus koristi binomni test (eng. *binomial test*), dok STUCCO koristi hi-kvadrat test. Prednost hi-kvadrat test je u činjenici da je osjetljiv na mali raspon ekstremnih oblika kontrasta te je kao takav bolji za zadaću otkrivanja kontrasta. Ovo ne znači da će u svim praktničnim primjenama otkrivanja kontrasta dati bolje rezultate od Magnum Opusa. Najveća razlika u filtrima između ova dva pristupa je STUCCOva upotreba korekcije kod višestrukih usporedbi i provedba traženja minimalne razlike. Magnum Opus ne provodi takvu korekciju kako bi izbjegao povećanje greške tipa 2. Nadalje, STUCCO primjenuje ograničenje na minimalnu veličinu razlike između grupa, dok Magnum Opus to ne radi. Može se zaključiti kako je filter koji primjenjuje Magnum Opus blaži od onog koji primjenjuje STUCCO.

Magnum Opus radi usporedbu unutar grupa, a ne između grupa (kao STUCCO) te je za očekivati kako će generirati manje pravila od STUCCO algoritma.

4.5. Diskusija o tehnikama otkrivanja kontrasta

U ovom poglavlju tehnike otkrivanja kontrasta se uspoređuju s postojećim tehnikama: stablom odlučivanja i asocijativnim pravilima. Na taj se način identificiraju njihove prednosti na temelju kojih se dolazi do zaključka da tehnike otkrivanja kontrasta imaju velik potencijal za primjenu u svrhu selekcije atributa.
U nastavku ovog poglavlja uspoređuju se razlike tehnika otkrivanja kontrasta u odnosu na redom: stabla odlučivanja i asocijativna pravila.

Jedan od pristupa koji se u literaturi koristio za razlikovanje dvije ili više grupa je stablo odlučivanja. Stablo odlučivanja ima prednost brzine generiranja razumljivih modela, ali i slijedeće nedostatake (Bay i Pazzani, 1999.):

(1) Metoda stabla odlučivanja nije potpuna jer postiže brzinu primjenom heuristike kako bi se kresao velik dio prostora za pretraživanje. Time se mogu propustiti alternativni načini razlikovanja jedne grupe od druge.
(2) Stabla odlučivanja se fokusiraju na diskriminacijsku sposobnost te propuštaju razlike među grupama koje nisu dobri diskriminatori, ali su ipak važne.
(3) Pravila dobivena stablom se obično interpretiraju u fiksnom poretku gdje se pravilo primjenjuje samo ako sva prethodna pravila nisu zadovoljena. Ova činjenica interpretaciju svakog pojedinog pravila čini teškom s obzirom na to da ga treba interpretirati u kontekstu.
(4) Teško je definirati dobar kriterij kao što je minimalna podrška.

Područje koje je usko povezano s otkrivanjem kontrasta su asocijativna pravila. Asocijativna pravila prikazuju vezu između varijabli u obliku X->Y. U analizi potrošačke košarice X ili Y su stavke poput kruha ili mlijeka. Kad imamo kategorijske podatke X i Y su parovi atribut-vrijednost kao npr. zanimanje = inženjer ili prihod > 5000 kn. Kako traženje asocijativnih pravila tako otkrivanje kontrasta zahtijeva pretraživanje prostora skupa stavaka ili parova atribut-vrijednost. U rudarenju asocijativnih pravila, tražimo skupove koji imaju podršku veću od određene definirane vrijednosti. Kod otkrivanja kontrasta tražimo skupove koji predstavljaju bitne razlike u temeljnim vjerojatnosnim distribucijama.
Pošto obje tehnike imaju element traženja među njima ima mnogo zajedničkog. Štoviše, otkrivanje kontrasta se nadovezuje na neke dijelove asocijativnih pravila kako bi se unaprijedilo otkrivanje kontrasta. No, asocijativna pravila i kontrastni skupovi bitno se razlikuju i to u slijedećem. Otkrivanje kontrasta se, suprotno od asocijativnih pravila, bavi s više grupa i ima drugačije kriterije kod traženja. Stoga se algoritmi asocijativnih pravila ne mogu direktno primijeniti na traženje kontrastnih skupova. Npr. jedan od načina bi bio da se za svaku grupo posebno traže pravila i potom ta pravila uspoređuju. No, rudarenje svake grupe posebno dovodi do toga da se gube neke od mogućnost kresanja koje mogu znatno poboljšati učinkovitost. Jedan od alternativnih načina je i da se kodiraju grupe, to postane jedna varijable te se izgrade pravila na takvim podacima. No, time se ne dobivaju razlike među grupama i takve rezultate je teško interpretirati. Naime, dobiva se previše pravila da bi ih se uspoređivalo. A Davies i Billman tvrde da je takve rezultate teško interpretirati jer „asocijativna pravila se ne provode s ciljem traženja kontrasta“ (Davies i Billman, 1996.). To znači da ne koriste iste atribute za razdvajanje grupa. U ovakvim situacijama trebalo bi tražiti pravila koja se podudaraju. A i kad bi se našla takva pravila treba definirati koji statistički test

se koristi za utvrđivanje da li je razlika u podršci statistički značajna. Kod tehnika otkrivanja kontrasta to je jasno definirano i tu su u prednosti.

5. KARAKTERISTIKE SKUPA PODATAKA

Danas postoji širok raspon klasifikatora koji se koriste u brojnim aplikacijama. Nijedan klasifikator nije se pokazao nadmoćan u odnosu na sve ostale klasifikatore na svim klasifikacijskim zadacima te se proces odabira klasifikatora je još uvijek svodi na pokušaje i pogrešake. Michie i suradnici dokazuju da je optimalan klasifikator za neki zadatak određen karakteristikama skupa podataka koji se koristi (Michie et. al, 1994.). Isto prepoznaje i Sohn koji tvrdi: „performanse svakog algoritma su tijesno povezane s karakteristikama skupa podataka koji se obrađuje“ (Sohn, 1999). Stoga je razumijevanje odnosa između karakteristika podataka i učinkovitosti klasifikatora ključno za proces odabira klasifikatora. Temeljem tih spoznaja, u ovom istraživanju se pretpostavlja i da je odabir optimalne tehnike selekcije atributa također određen karakteristikama skupa podataka koji se koristi.

U ovom radu koriste se teorijska svojstva klasifikatora za identifikaciju karakteristika skupa podataka koji utječu na performanse klasifikatora i selekcije atributa. Ta svojstva skupa podataka vode do razvoja mjera koje opisuju odnos između karakteristika skupa podataka i performansi klasifikatora, a definirane su od strane Van der Walta.

Prethodna empirijska istraživanja su pokazala da izbor optimalnog klasifikatora ovisi o korištenim podacima (Michie et. al, 1994.). Van der Walt ispituje koje karakteristike podataka utječu na učinkovitost klasifikacije i razvija mjere za mjerenje tih karakteristika podataka. Ove mjere omogućuju definiciju odnosa između karakteristika podataka i učinkovitosti klasifikatora.

Mjere su grupirane u sljedeće kategorije: standardne mjere, mjere oskudnosti podataka, statističke mjere, mjere teorije informacija, mjere granice odluka, topološke mjere i mjere šuma. U tablici 2. prikazane su navedene mjere.

Tablica 2. Karakteristike podataka važne za klasifikaciju

Izvor: Van der Walt, 2008.

Karakteristika	***Mjera***
Standardne mjere (eng. *standard measures*)	
Dimenzionalnost (eng. *dimensionality*)	d
Broj instanci (eng. *number of samples*)	N
Broj klasa (eng. *number of classes*)	C
Mjere oskudnosti podataka (eng. *data sparseness meaures*)	
Omjer oskudnosti podataka *(eng. data sparseness ratio)*	DSR
Oskudnost podataka (eng. *data sparseness*)	DS
Statističke mjere (eng. *statistical measures*)	
Korelacija atributa (eng. *correlation of features*)	p
Multivarijantni normalitet (eng. *multivariate normality*)	MVN
Homogenost kovarijanci klasa (eng. *homogenity of class covariances*)	SDR
Mjere teorije informacija (eng. *information theoretic measures*)	
Unutarnja dimenzionalnost (eng. *intrinsic dimensionality)*	ID
Omjer unutarnje dimenzionalnosti (eng. *intrinsic dimensionality ratio*)	IDR
Mjere granice odluka (eng. *desicion boundary measures*)	
Linearna separabilnost (eng. *linear separability*)	L1
Varijacije u složenosti granica odluka (eng. *variation in desicion boundary complexity*)	L2
Složenost granica odluka (eng. *desicion*	DBC

boundary complexity)	
Topološke mjere (eng. *topology measures*)	
Broj slučajeva po grupi (eng. *number of samples per group*)	T2
Varijacije u SD atributa (eng. *variation in feature SD*)	T3
Varijacije u skali (eng. *variation in scale*)	T4
Mjere šuma (eng. *noise measures*)	
Ulazni šum (eng. *input noise*)	N1
Izlazni šum (eng. *output noise*)	N2
Šum atributa (eng. *feature noise*)	ID2

Ostatak ovog poglavlja opisuje kategorije mjera koje su korištene u ovom istraživanju, a izdvojene su temeljem kriterija da su prepoznate kao važne u području otkrivanja znanja u podacima te da su korištene u, osim Van der Waltovog, barem još jednom prethodnom istraživanju. Izdvojene grupe karakteristika podataka su korištene u slijedećim istraživanjima:

- standardne mjere: (Michie, Spiegelhalter i Taylor, 1994.. Sohn, 1999.)
- mjere oskudnosti podataka (Anand i Bharadwaj, 2011.)
- statističke mjere: (Michie, Spiegelhalter i Taylor, 1994., Sohn, 1999.)
- mjere teorije informacija (Sohn, 1999.)
- mjere šuma (Gamberger, Lavrač i Dzeroski, 2000., Wu i Zhu, 2004., Wu i Zhu, 2008.)

Od mjera šuma izostavljeni su ulazni i izlazni šum, a izdvojen je šum atributa iz razloga jer Wu i Zhan u komparaciji različitih vrsta šuma zaključuju da je utjecaj šuma atributa na točnost klasifikacije nedovoljno istražen i njegovo ispitivanje predlažu u smjernicama za buduće istraživanje (Wu i Zhu, 2004).

Iz daljnjeg istraživanja su isključene dvije grupe karakteristika podataka: mjere granice odluka i topološke mjere iz razloga jer nisu korištene u prethodnim istraživanjima.

5.1. Standardne mjere

Wan der Walt navodi tri temeljne, generičke mjere koje služe za normalizaciju nekih drugih mjera. Standardne mjere dane su u tablici 3.

Tablica 3. Standardne mjere

Mjera	**Karakteristika podataka**
d	Broj atributa
C	Broj klasa
N	Broj instanci

Broj instanci (broj slučajeva), tj. veličina uzorka može imati veliki utjecaj na odabir klasifikatora i vrlo često igra ključnu ulogu u izboru klasifikatora. Broj klasa u skupu podataka utječe na oskudnost podataka u klasi jer klasifikator uglavnom zahtijeva instance iz svake klase. Isto tako, broj instanci po klasi utječe na učinkovitost klasifikacije u velikoj mjeri, jer određuje količinu informacije dostupnu za treniranje modela.

5.2. Mjere oskudnosti podataka

U ovom poglavlju se istražuje odnos između dimenzionalnosti podataka i broja instanci potrebne za precizno modeliranje podataka. Ta veza nije trivijalna i zato Van der Walt definira mjere koje tumače relevantne faktore. U trećem odjeljku ovog poglavlja razvija se mjera za kvantificiranje da li je broj instanci u skupu podataka dovoljan da točno modelira skup; ova mjera će izmjeriti koliko je oskudan skup ako se uzimaju u obzir broj atributa, broj klasa i broj instanci.

Oskudnost podataka Van der Walt definira kroz odnos dimenzionalnosti i broja instanci koje su potrebne da bi se podaci precizno modelirali.

U tipičnom slučaju, odnos između dimenzionalnosti tj. broja atributa (*d*) i broja instanci (N) može biti linearni, kvadratni ili eksponencijalni, kao što će se pokazati u nastavku. Van der Walt koristi teorijska svojstva klasifikatora da opiše svaku od tri vrste odnosa.

Za testiranje da li postoji linearni odnos između *d* i *N* koristit će se test normalnosti podataka i korelacije između atributa. Broj parametara koji se mora procijeniti je *2dC + C*.

Za testiranje kvadratnog odnosa između *d* i *N* će se mjeriti homogenost matrica kovarijanci klasa, kao i normalnost. Ukupan broj parametara koji se stoga mora procijeniti je $d^2 + DC + C$.

Linearni odnos može se testirati primjenom testa normalnosti i izračunom korelacije. Kvadratni odnos može se testirati primjenom testa za normalnost i testa homogenosti matrice kovarijanci. Ako u skupu podataka nije prisutan ni linearni ni kvadratni odnos, vjerojatan je eksponencijalni odnos između *N* i *d*.

Nakon što se utvrdi odnos između *d* i *N* trebamo kvantificirati da li postoji dovoljno uzoraka u skupu za točno modeliranje podataka Za svaki od četiri spomenutih odnosa, definira se mjera(N_{min}), koja postavlja minimalni broj uzoraka koji su potrebni za precizno modeliranje.

Ako su podaci normalno distribuirani i nekolerirani, linearni odnos između *d* i *N* će postojati i minimalna veličina uzoraka koja će biti potrebna je:

$$N_{l(\min)} = 2dC + C$$

Ako su podaci normalno distribuirani, koreliraju i kovarijance su homogene, onda postoji kvadratni odnos između *d* i *N* i minimalan broj uzoraka koji su potrebni će biti:

$$N_{q1(\min)} = 2d^2 + dC + C$$

Ako su podaci normalno distribuirani, koreliraju i kovarijance nisu homogene, onda postoji kvadratni odnos između *d* i *N* i minimalni broj uzoraka koje su potrebne se računa prema formuli:

$$N_{q2(\min)} = Cd^2 + dC + C$$

Ako podaci nisu normalno distribuirani, pretpostavlja se eksponencijalni odnos između *d* i *N* i broj uzoraka koji su potrebni definira se na slijedeći način:

$$N_{e(\min)} = D_{steps}^{d}$$

Gdje je D_{steps}^{d} diskretni broj koraka po atributu (klasa).

Dalje se kvantificira kolika je veličina uzorka dovoljna za točno modeliranje podataka. Mjera se kvantificira na način da se određuje omjer između stvarnog broja uzoraka i minimalnog broj uzoraka koji su potrebni. Omjer oskudnosti podataka se definira kako slijedi:

$$DSR = \frac{N}{N_{min}}$$

Gdje N_{min}je minimalni broj instanci i N stvarni broj instanci u skupu podataka.

5.2.1. Mjera oskudnosti podataka

U ovom poglavlju će se kvantificirati kolika je veličina uzorka dovoljna za točno modeliranje podataka. Mjera se kvantificira na način da se određuje omjer između stvarnog broja uzoraka i minimalnog broj uzoraka koji su potrebni. Omjer oskudnosti podataka (DSR) se definira kako slijedi:

$$DSR = \frac{N}{N_{min}}$$

gdje N_{min}je minimalni broj uzoraka i *N* stvarni broj uzoraka u skupu podataka. Nadalje, definira se mjera koja ukazuje da li je veličina uzorka dovoljna izlučivanjem gore navedene jednadžbe, a naziva se oskudnost podataka (DS):

$$DS = \sqrt[d]{N}$$

gdje je *N* broj slučajeva u skupu podataka i *d* dimenzionalnost skupa podataka.

5.3. Statističke mjere

Korelacija je vrlo važno svojstvo u klasifikaciji. U ovom radu koristit će se slijedeća mjera za kvantificiranje korelacije između atributa u skupu podataka:

$$p = \frac{1}{T}\sum_{i=1}^{C}\sum_{j=1}^{d-1}\sum_{k=j+1}^{d}|p_{jk}|$$

Pri tome je $|p_{jk}|$ apsolutna vrijednost Pearsonova koeficijenta korelacije između atributa *j* i *k,* T je ukupan broj koeficijenata korelacija zajedno, *C* je broj klasa i *d* je broj atributa.

Mjera *p* je prosječna apsolutna vrijednost koeficijenta korelacije između svih parova atributa za obje klase. Ova mjera nam daje naznaku međuovisnosti između svih atributa i strogo je nula ako nijedan atribut ne korelira, odnosno strogo je 1 ako su svi atributi jednaki. Vrijednosti *p* koja je blizu 1 pokazuje da atributi visoko koreliraju i sugerira da postoji redundantnost u skupu podataka jer korelirani atributi daju slične informacije.

Geometrijska sredina je omjer između kovarijance matrice i pojedinih klasa kovarijance matrica, a može se koristiti za procjenu **homogenosti matrica kovarijanci**. Pojedine klase se mogu testirati na homogenost korištenjem Box's M testa, koji se koristi u ovom istraživanju. *M test* se definira kao:

$$M = \gamma \sum_{i=1}^{C} (n_i - 1) log|S_i^{-1} S|$$

gdje

$$\gamma = 1 - \frac{2d^2 + 3d - 1}{6(d+1)(C-1)} \left[\sum_i \frac{1}{n_i - 1} - \frac{1}{n - C} \right]$$

n_i je broj uzoraka u klasi *i*, *S* je udružena matrica kovarijanci i S_i^{-1}je inverzna matrica klase kovarijanci za klasu *i, D* i *C*.

5.4. Mjere teorije informacija

Uzajamna informacija između klasa i atributa, *M (C; X)*, može se koristiti kako bi se utvrdilo unutarnju dimenzionalnost skupa podataka. Ovdje će se mjeriti koliko atributa značajno ne doprinosi klasifikaciji mjerenjem važnosti atributa s njihovim vrijednostima M (C; X).

Kako bi se odredilo koliko je atributa potrebno za obuhvaćanje 90% zajedničke informacije između klasa i atributa, računa se funkcija kumulativne distribucije zajedničke informacije između klasa i atributa. *Unutarnju dimenzionalnost* definiramo kao broj atributa potrebnih za obuhvaćanje 90% zajedničke informacije između klase i atributa. Ova mjera se označava kao *ID*, a omjer između *ID* i prave dimenzionalnosti kao *IDR*. Ako je vrijednost *IDR*-a niska znači da postoje brojni suvišni atributi koji mogu biti uzrokovani visoko koreliranim atributima. Ako je vrijednost *IDR*-a viša, većina atributa sadrži značajnu količinu informacije za klasifikaciju i klasifikacijski problem je dobro opisan danim atributima.

5.5. Mjere šuma

Van der Walt govori o tri tipa šuma: ulazni šum, izlazni šum i šum atributa. Ulazni šum se definira kao preklapanje instanci klasa, izlazni šum kao netočno označene instance, a šum atributa kao postotak atributa koji ne doprinosi klasifikaciji.

Prethodno definirana unutrašnja dimenzionalnost može se koristiti za mjerenje proporcije atributa koji ne doprinose klasifikaciji. Van der Walt predlaže slijedeću mjeru kao mjeru šuma atributa:

$$ID2 = \frac{d - ID}{d}$$

Pri čemu je *d* dimenzionalnost podataka, a *ID* unutrašnja dimenzionalnost.

6. DEFINICIJA TEHNIKA OTKRIVANJA KONTRASTA ZA SELEKCIJU ATRIBUTA

Kako je pokazano u pregledu istraživanja selekcije atributa, tehnike selekcije atributa imaju istu temeljnu strukturu. Za njihovu definiciju potrebno je utvrditi:

(1) mjeru koja svakom atributu ili skupu atributa dodjeljuje vrijednost s obzirom na njegovu (njihovu) prediktivnu snagu,

(2) metoda za pronalaženje optimalnog podskupa atributa prema odabranom kriteriju. Metoda sadrži algoritam traženja i kriterij zaustavljanja kojim se određuje kada potraga staje.

Stoga su i tehnike otkrivanja kontrasta u selekciji atributa karakterizirane kao kombinacija slijedećih komponenata:

- **mjere vrednovanja** atributa koja dodjeljuje vrijednost svakom atributu
- **kriterija rezanja** za izbor broja atributa koji se selektiraju

6.1. Mjere vrednovanja za selekciju podskupa atributa

Arauzo-Azofra i suradnici opisuju pet mjera vrednovanja (Arauzo-Azofra, Aznarte, i Benitez, 2011):

- uzajamna informacija
- omjer dobiti
- gini indeks
- relief-f
- relevantnost

Uzajamna informacija (eng. *mutual information*) je još poznata pod nazivom informacijska dobit. Ova mjera kvantificira informaciju koju atribut daje o klasi. Dolazi od Shannonove

teorije informacija i definira se kao razlika između entropije klase i entropije klase pod uvjetom poznavanje vrednovanog atributa (Cover i Thomas, 1991)

$$I(C,F) = H(C) - H(C|F)$$

Omjer dobiti (eng. *omjer dobiti*) se definira kao omjer između informacijske dobiti i entropije atributa. Na taj način, ova mjera izbjegava favoriziranje atributa s više vrijednosti, što je slučaj kod prethodne mjere. Ovu mjeru je koristio Quinlan kod C4.5 algoritma (Quinlan, 1993).

$$omjer\ dobiti = \frac{I(F,C)}{H(F)}$$

Gini indeks (eng. *gini index*) odražava vjerojatnost da dvije slučajno odabrane instance pripadaju različitoj klasi. Koristio ga je Breiman za generiranje klasifikacijskih stabala. Mjera se definira na slijedeći način (Breiman, 1998):

$$gini\ indeks = \sum_{i,j \in C; i \neq j} p(i|F)p(j|F)$$

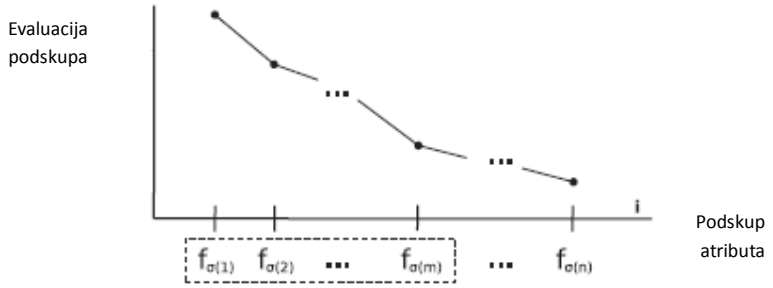
Relief-F je ekstenzija originalnog Relief algoritma (Kira i Rendell, 1992.), a razvio ga je Kononenko (1994.).Može raditi s diskretnim i kontinuiranim atributima, kao i s null vrijednostima. Iako vrednuje individualni atribut, Relief uzima u obzir veze između atributa. Zbog tog svojstva Relief daje dobre rezultate i često je korišten u selekciji atributa.

Relevantnost (eng. *relevance*) je mjera koja diskriminira između atributa na temelju njihova potencijala u formiranju pravila (Demsar i Zupan, 2004).

6.2. Kriteriji rezanja za selekciju podskupa atributa

Arauzo – Azofra i suradnici opisuju šest kriterija rezanja (Arauzo-Azofra, Aznarte, i Benitez, 2011):

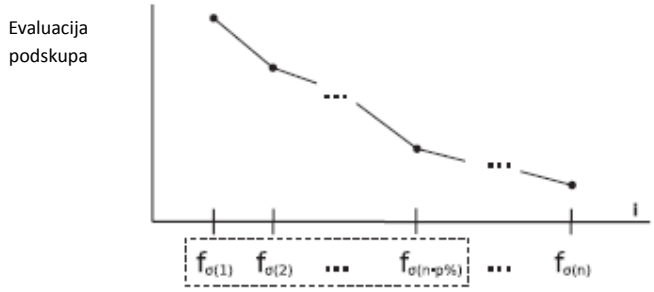
Fiksni broj atributa *(n)* – selektira se unaprijed, od korisnika, određeni broj atributa. Selektirani atributi su oni koji imaju najveću vrijednost funkcije evaluacije. Na slici u nastavku ovaj kriterij je prikazan grafički. Na osi *y* nalaze se vrijednosti mjere vrednovanje, a na osi *x* sortirani atributi. Selektirani atributi su označeni diskontinuiranim pravokutnikom.



Slika 9. Fiksni broj atributa kao kriterij rezanja

Izvor: Arauzo-Azofra, Aznarte i Benitez, 2011.

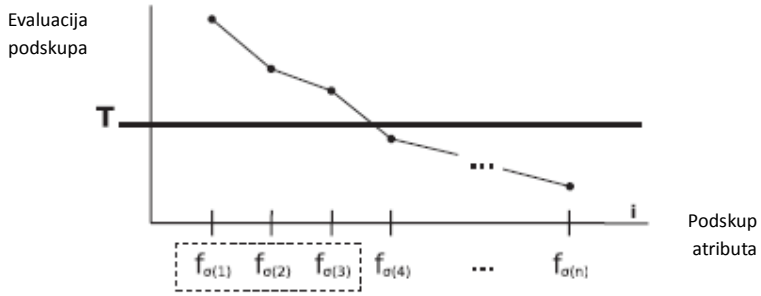
Udio (eng. *fraction*) selektira udio, koji se označava kao postotak od ukupnog broja atributa (slika 10).



Slika 10. Udio kao kriterij rezanja

Izvor: Arauzo-Azofra, Aznarte i Benitez, 2011.

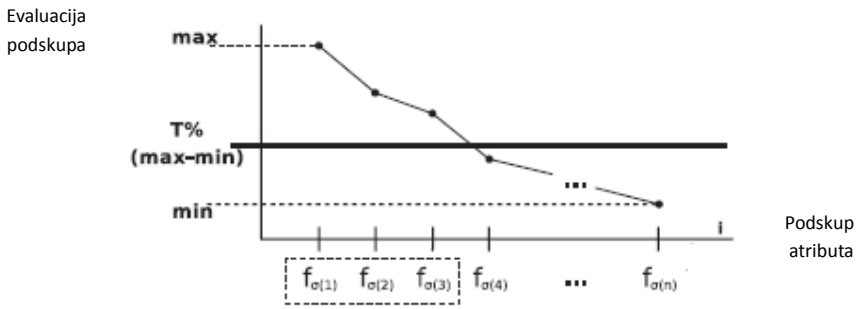
Prag (eng. *treshold*) selektira atribute s vrijednošću evaluacijske funkcije veće od praga definiranog od strane korisnika.



Slika 11. Prag kao kriterij rezanja

Izvor: Arauzo-Azofra, Aznarte i Benitez, 2011.

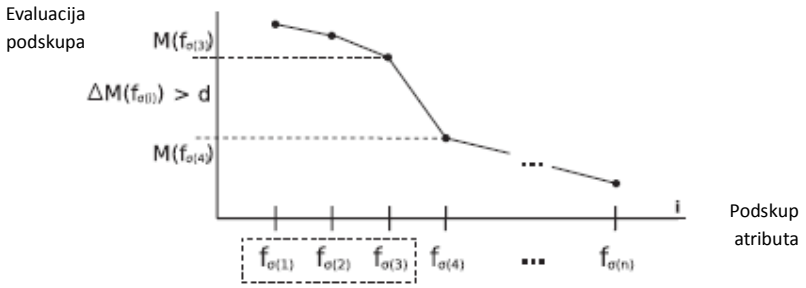
Prag izražen kao udio (eng. *threshold given as a fraction*) selektira atribute čija vrijednost funkcije vrednovanja je iznad određene granice, a ta granica se definira kao udio raspona evaluacijske funkcije (slika 12.).



Slika 12. Prag izražen kao udio kao kriterij rezanja

Izvor: Arauzo-Azofra, Aznarte i Benitez, 2011.

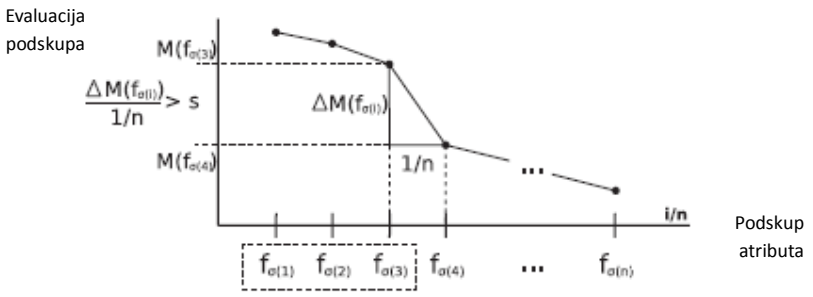
Razlika (eng. *difference*) selektira atribute, počevši od onog s najvećom vrijednosti evaluacijske funkcije i nastavlja dalje kroz sortiranu listu atributa sve dok razlika između dvije vrijednosti evaluacijskih funkcija nije veća od definirane razlike (slika 13.).



Slika 13. Razlika kao kriterij rezanja

Izvor: Arauzo-Azofra, Aznarte i Benitez, 2011.

Nagib (eng. *slope*) na sortiranoj listi atributa, selektira najbolje atribute sve dok nagib prema slijedećem atributu nije iznad definirane granice.



Slika 14. Nagib kao kriterij rezanja

Izvor: Arauzo-Azofra, Aznarte i Benitez, 2011.

6.3.Selekcija atributa tehnikama otkrivanja kontrasta

Ovo poglavlje definira tehnike otkrivanja kontrasta u selekciji atributa i naziva ih:

SfFS (eng. *Stucco for Feature Selection*) i

MOFS (eng. *Magnum Opus Feature Selection*).

Predložena metodologija koristi pretpostavku nezavisnosti atributa (eng. *feature independence assumption*). Ista pretpostavka se puno koristila u prethodnim istraživanjima zbog jednostavnosti, skalabilnosti i dobrih rezultata koje postiže u empirijskim istraživanjima te je vrlo efikasna u radu s velikim skupovima podataka (Yu i Liu, 2004.). Između ostalih, primjenjivali su je: Holz i Loew, 1994., Kudo i Sklansky, 1998., Blum i Langley, 1997, Guyon i Elisseeff, 2003 te Abe i suradnici, 2006..

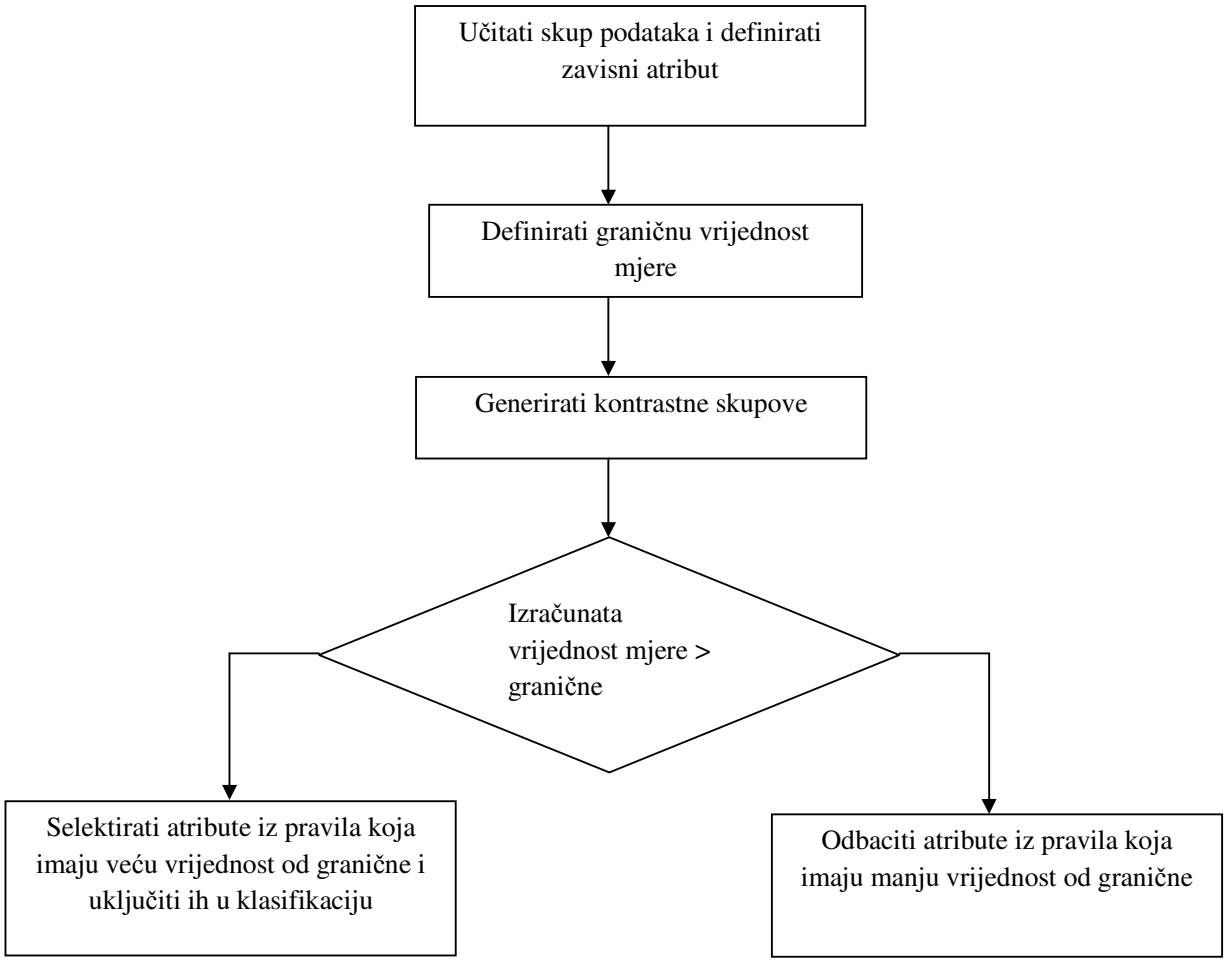
Korištenje ove pretpostavke znači da tehnike koje se predlažu u ovom radu koriste funkciju vrednovanja koja pridaje mjeru vrednovanja svakom atributu. Nakon što su atributi vrednovani, biraju se oni s najvećim vrijednostima. Za završetak procesa selekcije atributa definira se kriterij koji određuje gdje selekcija atributa staje.

Kroz evaluaciju u prošlom poglavlju definiranih pristupa Arauzo-Azofra, Aznarte i Benitez zaključuju da se ne može generalno preporučiti jedna mjera vrednovanja i jedan kriterij (Arauzo-Azofra, Aznarte, i Benitez, 2011). Stoga je napravljena analiza radova koji su citirali navedeni rad kako bi se vidjelo da li je u nekom kasnijem istraživanju napravljena evaluacija te da li se i ako da, koja mjera vrednovanja i/ili kriterij rezanja pokazuje superiornim u odnosu na ostale. Od osam radova koji citiraju Arauzo-Azofra, Aznarte i Beniteza u bazi Scopus, jedan dokazuje da je prag najučinkovitij kriterij rezanja (Ruiz et. al, 2012.) te se vođeno rezultatima tog istraživanja kod tehnika otkrivanja kontrasta u selekciji

atributa **prag** koristi kao kriterij. Kao mjera vrednovanja kod tehnika otkrivanja kontrasta koristi se **relevantnost** koja je definirana kao mjera koja diskriminira između atributa na temelju njihova potencijala u formiranju pravila (Demsar i Zupan, 2004). Razlog tome nalazi se u činjenici da su tehnike otkrivanja kontrasta, STUCCO i Magnum Opus, u svojoj biti definirane na način da kao rezultate daju pravila i mjere kvalitete pravila, odnosno mjeru koja razlikuje atribute s obzirom na njihov potencijal u definiranju pravila.
Tehnike otkrivanja kontrasta, za primjenu u selekciji atributa, se definiraju na slijedeći način:

mjera vrednovanja atributa je *relevantnost*, a kriterij rezanja je *prag* koji je definiran od strane korisnika.

Predložene tehnike selekcije atributa koriste pristup filtra i provode se u nekoliko koraka. Slika 15. opisuje postupak predložene metodologije.



Slika 15. Dijagram tijeka selekcije atributa tehnikama otkrivanja kontrasta

Tehnike otkrivanja kontrasta mjere koliko pojedini atributi doprinose razlici između dviju klasa i kao rezultat provođenja daju kontrastne skupove.

Mjera koja se koristi u selekciji atributa kod Magnum Opusa je *utjecaj*. Utjecaj je mjera koja je inicijalno postavljena u Magnum Opusu, a predlaže se za korištenje u selekciji atributa iz slijedećeg razloga. Temelji se na stupnju do kojeg se promatrana zajednička pojavljivanja prethodnika i sljedbenika razlikuju od očekivanog broja pojavljivanja kada bi prethodnik i sljedbenik bili neovisni, tj. *utjecaj* pokazuje veličinu te kao takav identificira one prethodnike (atribute) koji najviše doprinose razlici. Što je i zadaća selekcije atributa.

Magnum Opus algoritam za svako dobiveno pravilo računa vrijednost mjere *utjecaja* prema jednadžbi opisanoj u poglavlju 4.3. i uz to statističku značajnost. Svi atributi koji su s lijeve strane onih pravila koji imaju statistički značajnu ($p<0.05$) vrijednost *utjecaja* većeg od definirane korisničke vrijednosti odabiru se u podskup.

Pseudokod algoritma *Magnum Opus u selekciji atributa, MOFS,* dan je u nastavku.

Ulaz: Skup podataka $S = \{A_1, A_2, \dots A_n\}$

$Utjecaj_{min}$ //minimalna vrijednost mjere *utjecaj* definirana od korisnika

m// broj generiranih kontrastnih skupova

p// statistička značajnost pravila

i, atribut //pomoćne varijable

Izlaz: Selektirani podskup atributa *P*

1) $P_0 = generirati_kontrastne_skupove\ (S)$ //generira kontrastne skupove
2) FOR (i=1; i<= m; i++){

   IF ($p < 0.05\ AND\ utjecaj \geq utjecaj_{min}$)

   THEN Atribut= LHS

   Dodati Atribut u *P*

   }

STUCCO algoritam pronalazi kontrastne skupovi koji se zovu *devijacije*. Devijacija je kontrastni skup koji je značajan i velik. Kontrastni skup za koji se najmanje dvije grupe razlikuju u podršci je *značajan*. Za utvrđivanje značajnosti provodi se *hi-kvadrat* test s nul hipotezom da je podrška kontrastnog skupa jednaka među grupama. Kod izračuna *hi kvadrat* testa provjerava se vrijednost distribucije. Vrijednost mora biti manja od definirane granične vrijednosti statističke značajnosti (obično *p=0.05* Kontrastni skup za koji je maksimalna razlika između podrška veća od vrijednosti *mindev* (minimalna devijacija) je *velik*. U postupku selekcije atributa odabrani su oni atributi koji se nalaze s lijeve strane kontrastnog skupa koji je značajan i velik.

Na slijedećoj slici dan je pseudokod algoritma STUCCO u selekciji atributa, **SfFS** (eng. **S**tucco **f**or **F**eature **S**election).

Slika 17. Pseudokod SfFS-a

Ulaz: Skup podataka $S = \{A_1, A_2, \dots A_n\}$

min_dev //minimalna vrijednost *devijacije* definirana od korisnika

m// broj generiranih kontrastnih skupova

p// statistička značajnost pravila

i, atribut //pomoćne varijable

Izlaz: Selektirani podskup atributa *P*

3) $P_0 = generirati_kontrastne_skupove\ (S)$ //generira kontrastne skupove oblika *LHS->RHS*
4) FOR (i=1; i<= m; i++){

IF ($p < 0.05\ AND\ devijacija \geq$ min $_dev$)

THEN Atribut= LHS

Dodati Atribut u *P*

}

U nastavku rada definirane tehnike se primjenjuju u selekciji atributa i uspoređuju s postojećim tehnikama.

7. EMPIRIJSKO ISTRAŽIVANJE

Cilj je ovog poglavlja vrednovati tehnike otkrivanja kontrasta u smislu točnosti klasifikacije i vremena potrebnog za selektiranje atributa kako bi se utvrdilo koliko dobro tehnike otkrivanja kontrasta rade selekciju atributa. Pri tome treba uzeti u obzir da se ove tehnike koriste u pripremnoj fazi procesa otkrivanja znanja u podacima koja je osmišljena u svrhu što kvalitetnije naknadne analize podataka i klasifikacije. Stoga se kvaliteta ovih tehnika predobrade mora istražiti posredno, rezultatima njihove učinkovitosti u klasifikaciji. Međutim, poznato je i argumentirano u ranijim poglavljima ovog rada da učinkovitost klasifikacije ne ovisi samo o korištenim tehnikama pripreme podataka i/ili primjenjenom klasifikatoru već i o karakteristikama skupa podataka nad kojima se vrši klasifikacija.

Kako bi se testirale hipoteze rada provedeno je istraživanje i napravljene su usporedbe na skupovima podataka iz četiri javno dostupno repozitorija. Za svaki skup podataka izračunate su karakteristike. U istraživanju se koristi poznata procedura validacije tehnika selekcije atributa primjenom realnih skupova podataka. Izabrani su stvarni skupovi podataka koji predstavljaju referentne skupove. Izabrani podskup se testira koliko je točan upotrebom dvaju klasifikatora. Kako bi se evaluirale performanse tehnika selekcije atributa, točnost klasifikatora treniranih na atributima selektiranih od strane tih tehnika će se međusobno uspoređivati, kao i vrijeme potrebno za provođenje selekcije atributa-

Postoji mnogo klasifikatora, svaki s određenim prednostima i nedostacima. Ne postoji jedan klasifikator koji dalje najbolje rezultate za sve probleme. U ovom istraživanju koristit će se dva koja pripadaju različitim pristupima jer uče na različite načine, a njihov izbor argumentiran je u poglavlju 2. To su: neuronske mreže i diskriminacijska analiza.

Ovo poglavlje u nastavku daje empirijsku usporedbu reprezentativnih tehnika selekcije atributa s tehnikama otkrivanja kontrasta koje se prvi put koriste u svrhu selekcije atributa, a organizirano je na slijedeći način: prvi dio opisuje i karakterizira skupove podataka korištene u eksperimentu, drugi dio opisuje provedbu klasifikacije neuronskim mrežama, treći dio provedbu klasifikacije diskriminacijskom analizom, a četvrti dio odnosi se na mjerenje vremena potrebnog za provedbu selekcije atributa.

7.1. Karakterizacija skupova podataka

Eksperimenti su provedeni na 128 skupova podataka pronađenih u četiri repozitorija, i to slijedeća:

- UCI Machine Learning Repository
- StatLib - Carnegie Mellon University
- Sociology Data Set Server of Saint Joseph`s University in Philadelphia
- Feature selection datasets at Arizona State University /

Skupovi se razliku po svojim karakteristikama, a karakterizirani su s: brojem atributa, brojem instanci, stupnjem oskudnosti skupa podataka, koeficijentom korelacije, normalnošću, homogenošću, omjerom unutarnje dimenzionalnosti i šumom atributa. Kod svakog skupa zavisni atribut ima dvije klase.

Prostor traženja se nastoji ograničiti kako bi se uopće moglo provesti istraživanje te se ovih sedam uglavnom kvantitativnih veličina (mjera za karakteristike skupa podataka) kategorizira u dvije kategorije. Kako bi se obuhvatile sve kombinacije karakteristika skupova podataka, potrebno je pronaći 128 skupova podataka (2^7 kombinacija, pošto je 7 karakteristika svaka s dvije kategorije).

Tijek istraživanja je slijedeći. Za svaki skup podatka izračunava se sedam mjera karakteristika skupa podataka na način koji će se u nastavku prikazati na jednom skupu podataka, *vote*. U gore navedenim repozitorijima pronalazi se 128 skupova podataka koji obuhvaćaju sve kombinacije različitih karakteristika. Nad svakim skupom provodi se selekcija atributa primjenom sedam tehnika selekcije atributa. Izabrani podskupovi se testiraju upotrebom dvaju klasifikatora koji imaju različite pristupe u učenju: neuronskim mrežama i diskriminacijskom analizom. Kako bi se evaluirale performanse tehnika selekcije atributa, vrijeme provođenja selekcije atributa i točnost klasifikatora treniranih na atributima selektiranih od strane tih tehnika, će se međusobno uspoređivati.

U prvom koraku, provjeri karakteristika skupa podataka, izračunata je vrijednost za svaku od sedam karakteristika skupa podataka. Svaka izračunata vrijednost klasificirana je u jednu od dvije kategorije. Kategorije su određene empirijski s obzirom na skupove podataka koji su korišteni u istraživanju.

Kategorije za svaku od karakteristika skupa podataka su slijedeće:

broj atributa: mali, veliki,
broj instanci: mali, veliki,
mjera oskudnosti podataka: mala, velika
korelacija: da, ne,
normalnost: da, ne,
homogenost: da, ne,
omjer unutarnje dimenzionalnosti: mali, veliki,
šum atributa: mali, veliki.

Ove kategorije određene su na slijedeći način. Iz dostupnih repozitorija nasumično je odabrano 7 skupova podataka i izračunate su im vrijednosti karakteristika. Granica kategorija dobivena je koristeći formulu za medijan koja glasi ovako:

$$r = int(\frac{N}{2})$$

$$Me = x_{r+1}$$

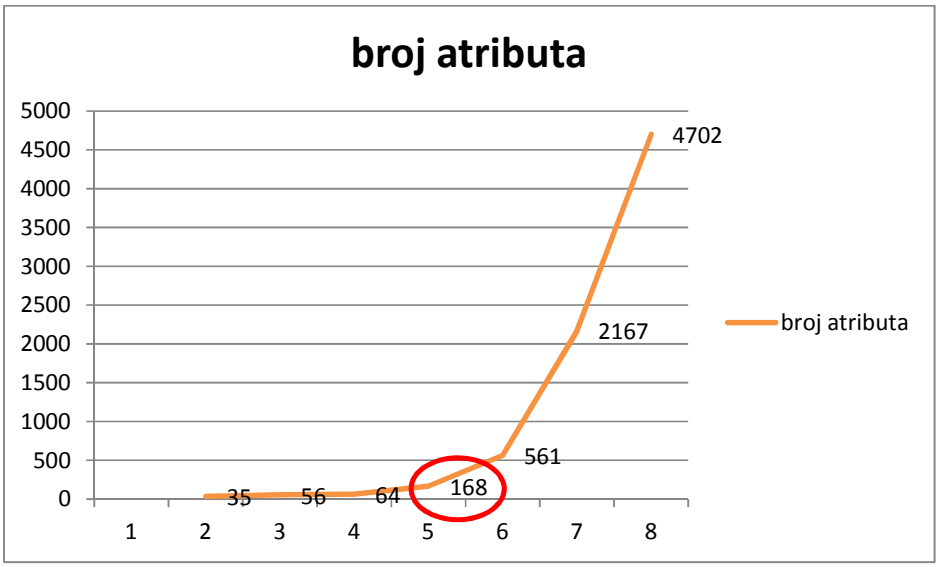
pri čemu je N broj skupova podataka, a x_{r+1} vrijednost broja atributa na mjestu r+1, kada su brojevi atributa sortirani od najmanjeg do najvećeg.
U ovom slučaju, gdje se medijan računa na 7 skupova podataka, N=7.

$$r = int\left(\frac{7}{2}\right) = 4$$

Dobivene vrijednosti su sortirane i prikazane grafički. Za svaku karakteristiku podataka izračunat je medijan kao četvrta vrijednost u nizu. Time je podskup podijeljen na dva dijela: polovica skupova u podskupu ima vrijednost karakteristike podataka manju ili jednaku medijanu, a pola veću ili jednaku medijanu. Kako bi se spriječila situacija da se npr. skup podataka s *n* atributa okarakterizira kao mali, a skup podataka s *n+1* atributom kao veliki postavlja se *sigurnosna zona* (eng. *buffer zone*) oko medijana iz kojeg se ne koriste vrijednosti niti za jednu od dvije definirane kategorije. Zona se postavlja u rasponu od *±10%* dobivene vrijednosti medijana.

Za primjer karakteristike *broj atributa* medijan je 168 (slika 14.)

Da se spriječi situacija da se skup podataka s 168 atributa okarakterizira kao mali, a onaj s 169 kao veliki, postavljena je *sigurnosna zona* (eng. *buffer zone*) od 16 atributa, što znači da su skupovi s manje od 150 atributa karakterizirani kao mali, a oni od 185 atributa i više kao veliki.



Slika 18. Određivanje kategorija za karakteristiku *broj atributa*

Postupak određivanje kategorije (mali ili veliki) za karakteristiku *broj atributa* dan je slijedećim pseudokodom:

Neka je *d* broj atributa u skupu podataka.

ZA SVAKI *i=1* do *i=128*

Provjeri d_I skupa podataka.

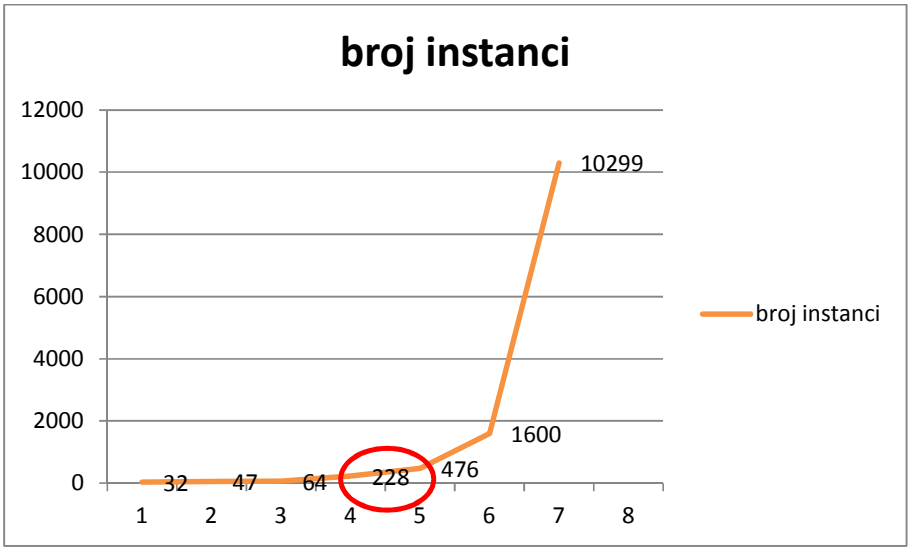
AKO JE $d_I \leq 150$, TADA *broj atributa* = mali,

INAČE AKO JE $d_I \geq 185$ TADA *broj atributa* = veliki

Opravdanost ovog pristupa potkrepljuje se slijedećim podacima. Odabrani podskup od 7 sedam skupova podataka je reprezentativan jer ima aritmetičku sredinu broja atributa 1108, a

standardnu devijaciju 1759, dok cijeli skup od 128 skupova podataka ima artitmetičku sredinu broja atributa 1264, a standardnu devijaciju 3435.

Analogno opisanom pristupu, na isti način definirane su kategorije za karakteristiku *broj slučajeva* (slika 15.). Izračunati medijan kod ove karakteristike je 228. postavljena je sigurnosna zona od 20 instanci, što znači da su skupovi s manje od 208 instanci i manje karakterizirani kao mali, a oni od 248 instanci i više kao veliki.



Slika 19. Određivanje kategorija za karakteristiku *broj slučajeva*

Postupak određivanje kategorije (mali ili veliki) za karakteristiku *broj slučajeva* dan je slijedećim pseudokodom:

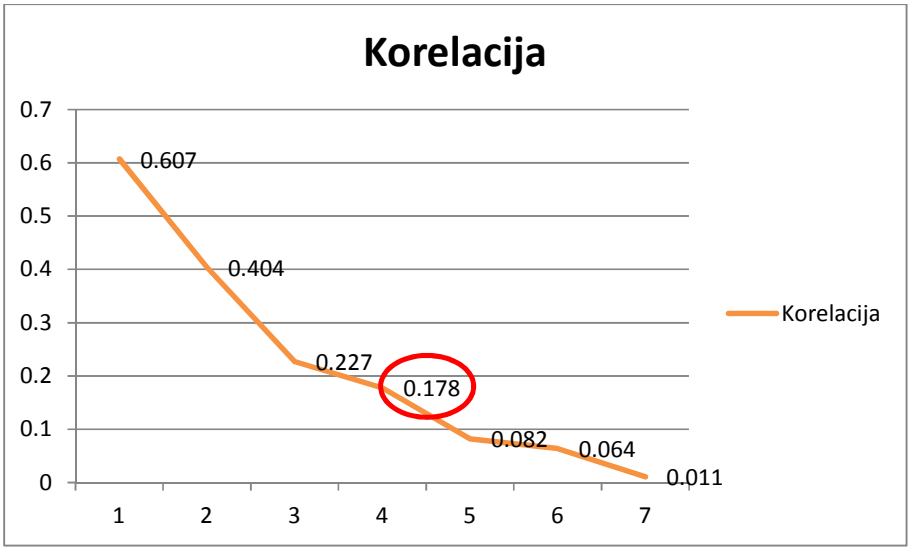
Neka je *N* broj slučajeva u skupu podataka.
ZA SVAKI *i=1* do *i=128*
  Provjeri N_i skupa podataka.
    AKO JE $N_I \leq 208$, TADA *broj slučajeva* = mali,
      INAČE AKO JE $N_I \geq 248$ TADA *broj slučajeva* = veliki

Opravdanost ovog pristupa i kod ove karakteristike potkrepljuje se podacima o aritmetičkoj sredini i standardnoj devijaciji cijelog skupa od 128 skupova podataka te odabranog

podskupa od 7 skupova podataka. Odabrani podskup od 7 sedam skupova podataka je reprezentativan jer ima aritmetičku sredinu 1821, a standardnu devijaciju 3779, dok cijeli skup od 128 skupova podataka ima artitmetičku sredinu broja slučajeva 2105, a standardnu devijaciju 4157.

U nastavku se nalaze vrijednosti koeficijenata korelacije i omjera unutarnje dimenzionalnosti za ovih sedam skupova na temelju kojih su postavljene granice. I kod ovih karakteristika postavljena je sigurnosna zona od $\pm 10\%$ kod definiranja kategorija *mali-veliki.*

Izračunati medijan za karakteristiku korelacija je 0.178. Postavljena je sigurnosna zona od ±0.018.



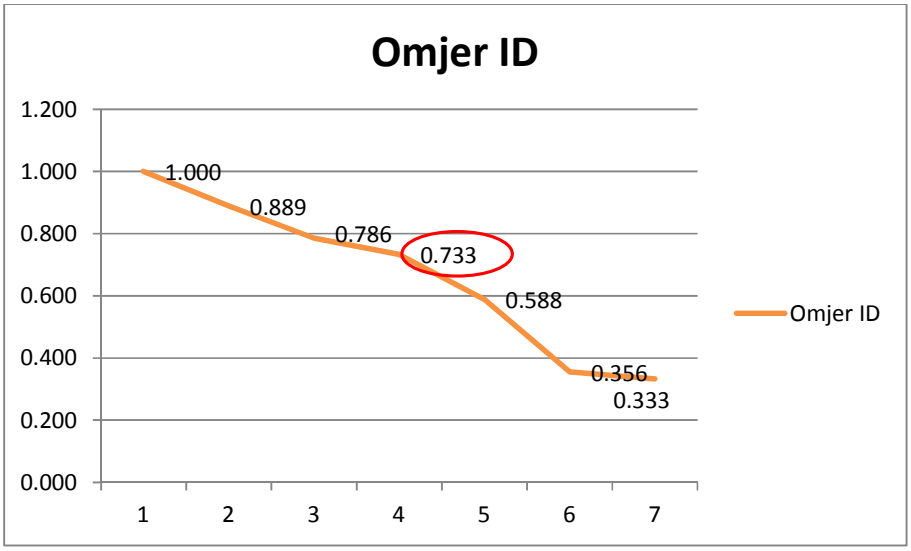
Slika 20. Određivanje kategorija za karakteristiku *korelacija*

Postupak određivanje kategorije (mala ili velika) za karakteristiku *korelacija* dan je slijedećim pseudokodom:

Neka je r korelacija skupa podataka.
ZA SVAKI $i=1$ do $i=128$
  Provjeri r_i skupa podataka.
    AKO JE $r_i \leq 0,16$ TADA *korelacija* = mala,
      INAČE AKO JE $r_i \geq 0.196$ TADA *korelacija* = velika.

Opravdanost ovog pristupa i kod ove karakteristike potkrepljuje se podacima o aritmetičkoj sredini i standardnoj devijaciji cijelog skupa od 128 skupova podataka te odabranog podskupa od 7 skupova podataka. Odabrani podskup od 7 sedam skupova podataka je reprezentativan jer ima prosječnu korelaciju 0.225, a standardnu devijaciju 0.213, dok cijeli skup od 128 skupova podataka ima prosjčnu korelacije 0.245, a standardnu devijaciju 0.195.

Izračunati medijan za karakteristiku omjer unutarnje dimenzionalnosti je 0.733. Postavljena je sigurnosna zona od ±0.07.



Slika 21. Određivanje kategorija za karakteristiku *omjer unutarnje dimenzionalnosti*

Postupak određivanje kategorije (mala ili velika) za karakteristiku *unutarnja dimenzionalnost* dan je slijedećim pseudokodom:

Neka je *ID* unutarnja dimenzionalnost skupa podataka.
ZA SVAKI *i=1* do *i=128*
Provjeri ID_i skupa podataka

AKO JE $ID_i \leq 0{,}663$ TADA *unutarnja dimenzionalnost* = mala,

INAČE AKO JE $ID_i \geq 0.803$ TADA *unutarnja dimenzionalnost* = velika.

Opravdanost ovog pristupa i kod ove karakteristike potkrepljuje se podacima o aritmetičkoj sredini i standardnoj devijaciji cijelog skupa od 128 skupova podataka te odabranog podskupa od 7 skupova podataka. Odabrani podskup od 7 sedam skupova podataka je reprezentativan jer ima aritemtičku sredinu 0.669, a standardnu devijaciju 0.256, dok cijeli skup od 128 skupova podataka ima artitmetičku sredinu unutarnje dimenzionalnosti 0.623 a standardnu devijaciju 0.3165.

Pošto se šum atributa dobiva tako da se od 1 oduzme omjer unutarnje dimenzionalnosti, skupovi podataka koji imaju mali omjer unutarnje dimenzionalnosti imaju velik šum i obrnuto.

Za karakteristiku normalnost provodi se Kolmogorov Smirnov test, a za karakteristiku homogenost kovarijanci Box's test. Na temelju rezultata provedbe tih testova određuje se da li skup ima normalnu distribuciju ili nema te da li su kovarijance homogene ili ne.
Karakteristika oskudnost ima dvije kategorije, mala i velika. Oskudnost skupa podataka je mala ako je stvarno broj instanci veći ili jednak od stvarnog broja instanci. Oskudnost skupa podataka je velika ako je stvarni broj instanci manji od potrebnog broja instanci.

Pošto izračun vrijednosti pojedinih mjera zahtijeva značajno računanje, način izračuna pokazat će se samo na jednom skupu kako bi se ilustrirale glavne ideje sadržane u pojedinim mjerama. Na isti način napravljen je izračun za sve ostale skupove podataka. U nastavku se nalazi izračun svake karakteristika za jedan skup podataka, *vote*.

Skup podataka: vote

Izvor: UCI Machine Learning Repository

Standardne mjere

Dimenzionalnost: 17 (mali broj atributa)

Broj instanci: 435 (velik broj instanci)

Mjere oskudnosti podataka

Primjenom Kolmogorov Smirnov testa provjeravana je normalnost distribucije. Kod skupa podataka *vote* nijedan atribut u skupu nema normalnu distribuciju. Iz toga zaključujemo da postoji eksponencijalni odnos između atributa u skupu podataka te se potreban broj instanci računa kao:

$$2^{17} = 131072$$

Pošto je minimalni broj instanci koji su potrebni za modeliranje (131 073) veći od stvarnog broja instanci (435), zaključuje se kako ne postoji dovoljno instanci za precizno modeliranje, i vrijednost karakteristike oskudnost je VELIKA.

Statističke mjere

Korelacija

Ukupna suma korelacija među svim atributima je 55,04, a ukupan broj korelacija je 136. Kada se te dvije vrijednosti podjele dobiva se korelacija skupa

$$\frac{55{,}04}{136} = 0{,}404$$

Kako je dobivena vrijednost korelacije skupa podataka veća od 0.196 korelacija se karakterizira kao VELIKA.

Normalnost podataka

Nijedan atribut u skupu nema normalnu distribuciju. Stoga je vrijednost karakteristike normalnost NE.

Homogenost kovarijanci

Box`s M test jednakosti matrica kovarijanci provjerava pretpostavku homogenosti kovarijanci među grupama uzimajuću kao kriterij granicu od $p < 0.001$ kao granicu statističke značajnosti rezultata. U ovom slučaju, test je statistički značajan što ukazuje na to da postoje značajne razlike između matrica kovarijanci. Stoga je zaključak da kovarijance **nisu homogene**. U slučaju gdje Box`s M test nije značajan, kovarijance su homogene.

Tablica 4. Rezultati Box`s M testa

Box's M		594.173
F	Approx.	4.051
	df1	136
	df2	157196.532
	Sig.	.000

Mjere teorije informacija

Kao mjere teorije informacija koriste se unutarnja dimenzionalnost *(ID*) i omjer unutarnje dimenzionalnosti izveden iz unutarnje dimenzionalnosti. Za slučaj skupa podataka *vote* unutarnja dimenzionalnost iznosi 10, što znači da je 10 atributa potrebno za obuhvaćanje 90% zajedničke informacije između klasa i atributa. Omjer unutarnje dimenzionalnosti između *ID* i prave dimenzionalnosti kao *IDR*. Ako je vrijednost IDR-a niska znači da postoje brojni suvišni atributi koji mogu biti uzrokovane visoko koreliranim atributima. Takva vrijednost sugerira da se transformacija svojstvenih vrijednosti treba uzeti u obzir. Pošto je u ovom slučaju je vrijednost IDRa veća:

$$IDR = \frac{10}{17}$$

većina atributa sadrži značajnu količinu informacije za klasifikaciju i klasifikacijski problem je dobri opisan atributima.

Mjere šuma

Šum atributa je mjera koja govori koliki udio atributa ne doprinosi klasifikaciji. U ovom slučaju je to manje od polovice atributa:

$$ID2 = \frac{17 - 10}{17} = \frac{7}{17} = 0{,}411$$

Udio atributa koji ne doprinose klasifikaciji je 0,411.

Na isti način su izračunate karakteristike za preostalih 127 skupova (s 128 skupova obuhvaćene su sve moguće kombinacije karakteristika skupapodataka).

Tablica 5. daje sumarne, kvalitativne podatke o svakom skupu, a tablica 1 u prilogu rada pokazuje izračunate kvantitativne vrijednosti za svaku karakteristiku koja to zahtijeva.

Tablica 5. Karakterizacija skupova podataka

Broj atributa	Broj instanci	Oskudnost	Korelacija	Normalnost	Homogenost	Omjer ID	Šum atributa	Skup
mali	mali	mala	ne	da	da	mali	veliki	*Pittsburgh+Bridges*
mali	mali	mala	ne	da	da	veliki	mali	*Trains*
mali	mali	mala	ne	da	ne	mali	veliki	*Balloons*
mali	mali	mala	ne	da	ne	veliki	mali	*Titanic*
mali	mali	mala	ne	ne	da	mali	veliki	*broadway*
mali	mali	mala	ne	ne	da	veliki	mali	*assessment*
mali	mali	mala	ne	ne	ne	mali	veliki	*Soybean+Small*
mali	mali	mala	ne	ne	ne	veliki	mali	*molecular biology promoters*
mali	mali	mala	da	da	da	mali	veliki	*Spectf*
mali	mali	mala	da	da	da	veliki	mali	*japansolvent*
mali	mali	mala	da	da	ne	mali	veliki	*Post-Operative+Patient*
mali	mali	mala	da	da	ne	veliki	mali	*hepatitis*
mali	mali	mala	da	ne	ne	mali	veliki	*election*
mali	mali	mala	da	ne	ne	veliki	mali	*Lung Cancer*
mali	mali	mala	da	ne	da	mali	veliki	*sponge*
mali	mali	mala	da	ne	da	veliki	mali	*creditscore*
mali	mali	velika	ne	da	da	mali	veliki	*bankruptcy*
mali	mali	velika	ne	da	da	veliki	mali	*gviolence*
mali	mali	velika	ne	da	ne	mali	veliki	*Labor+Relations*

Broj atributa	Broj instanci	Oskudnost	Korelacija	Normalnost	Homogenost	Omjer ID	Šum atributa	Skup
mali	mali	velika	ne	da	ne	veliki	mali	*Acute+Inflammations*
mali	mali	velika	ne	ne	da	mali	veliki	*runshoes*
mali	mali	velika	ne	ne	da	veliki	mali	*Cyyoung9302*
mali	mali	velika	ne	ne	ne	mali	veliki	*impeach*
mali	mali	velika	ne	ne	ne	veliki	mali	*fraud*
mali	mali	velika	da	da	da	mali	veliki	*Campus Climate 2011 SJU*
mali	mali	velika	da	da	da	veliki	mali	*homerun*
mali	mali	velika	da	da	ne	mali	veliki	*sonar*
mali	mali	velika	da	da	ne	veliki	mali	*bondrate*
mali	mali	velika	da	ne	ne	mali	veliki	*ICPSR 3009*
mali	mali	velika	da	ne	ne	veliki	mali	*gsssexsurvey*
mali	mali	velika	da	ne	da	mali	veliki	*uktrainacc*
mali	mali	velika	da	ne	da	veliki	mali	*ncaa*
mali	veliki	mala	ne	da	da	mali	veliki	*credit*
mali	veliki	mala	ne	da	da	veliki	mali	*weights*
mali	veliki	mala	ne	da	ne	mali	veliki	*ICPSR 2743*
mali	veliki	mala	ne	da	ne	veliki	mali	*city*
mali	veliki	mala	ne	ne	da	mali	veliki	*supreme*
mali	veliki	mala	ne	ne	da	veliki	mali	*ICPSR 2751*
mali	veliki	mala	ne	ne	ne	mali	veliki	*blood-transfusion/*
mali	veliki	mala	ne	ne	ne	veliki	mali	*authorship*
mali	veliki	mala	da	da	da	mali	veliki	*ICPSR 2867*

Broj atributa	Broj instanci	Oskudnost	Korelacija	Normalnost	Homogenost	Omjer ID	Šum atributa	Skup
mali	veliki	mala	da	da	da	veliki	mali	*ICPSR 2480*
mali	veliki	mala	da	da	ne	mali	veliki	*halloffame*
mali	veliki	mala	da	da	ne	veliki	mali	*CPS_85_Wages*
mali	veliki	mala	da	ne	ne	mali	veliki	*Physical+Activity+Monitoring*
mali	veliki	mala	da	ne	ne	veliki	mali	*marketing*
mali	veliki	mala	da	ne	da	mali	veliki	*binge*
mali	veliki	mala	da	ne	da	veliki	mali	*ionosphere*
mali	veliki	velika	ne	da	da	mali	veliki	*ICPSR 2859*
mali	veliki	velika	ne	da	da	veliki	mali	*Mushroom*
mali	veliki	velika	ne	da	ne	mali	veliki	*ICPSR 2039*
mali	veliki	velika	ne	da	ne	veliki	mali	*Thyroid+Disease*
mali	veliki	velika	ne	ne	da	mali	veliki	*sick*
mali	veliki	velika	ne	ne	da	veliki	mali	*One-hundred+plant+species+leaves+data+set*
mali	veliki	velika	ne	ne	ne	mali	veliki	*Kr-Vs-Kp*
mali	veliki	velika	ne	ne	ne	veliki	mali	*tic-tac-toe*
mali	veliki	velika	da	da	da	mali	veliki	*abgss98*
mali	veliki	velika	da	da	da	veliki	mali	*ICPSR 2686*
mali	veliki	velika	da	da	ne	mali	veliki	*ICPSR 2155*
mali	veliki	velika	da	da	ne	veliki	mali	*heart-statlog*
mali	veliki	velika	da	ne	ne	mali	veliki	*spambase*

Broj atributa	Broj instanci	Oskudnost	Korelacija	Normalnost	Homogenost	Omjer ID	Šum atributa	Skup
mali	veliki	velika	da	ne	ne	veliki	mali	*vote*
mali	veliki	velika	da	ne	da	mali	veliki	*Hill-Valley*
mali	veliki	velika	da	ne	da	veliki	mali	*hepatitis*
veliki	mali	mala	ne	da	da	mali	veliki	*ICPSR 4291*
veliki	mali	mala	ne	da	da	veliki	mali	*ICPSR 4582*
veliki	mali	mala	ne	da	ne	mali	veliki	*ICPSR 9595*
veliki	mali	mala	ne	da	ne	veliki	mali	*ICPSR 21600 2*
veliki	mali	mala	ne	ne	da	mali	veliki	*ICPSR 21600 3*
veliki	mali	mala	ne	ne	da	veliki	mali	*ICPSR 28641 2*
veliki	mali	mala	ne	ne	ne	mali	veliki	*ICPSR 6542*
veliki	mali	mala	ne	ne	ne	veliki	mali	*ICPSR 4367*
veliki	mali	mala	da	da	da	mali	veliki	*ICPSR 4572 02*
veliki	mali	mala	da	da	da	veliki	mali	*ICPSR 21600 4*
veliki	mali	mala	da	da	ne	mali	veliki	*DBWorld+e-mails*
veliki	mali	mala	da	da	ne	veliki	mali	*ICPSR 6135*
veliki	mali	mala	da	ne	ne	mali	veliki	*ICPSR 4537 8th form 1*
veliki	mali	mala	da	ne	ne	veliki	mali	*ICPSR 4275*
veliki	mali	mala	da	ne	da	mali	veliki	*GLI-85*
veliki	mali	mala	da	ne	da	veliki	mali	*ICPSR 21600 1*
veliki	mali	velika	ne	da	da	mali	veliki	*ICPSR 4566 02*
veliki	mali	velika	ne	da	da	veliki	mali	*ICPSR 8255*
veliki	mali	velika	ne	da	ne	mali	veliki	*ICPSR 28641*

Broj atributa	Broj instanci	Oskudnost	Korelacija	Normalnost	Homogenost	Omjer ID	Šum atributa	Skup
veliki	mali	velika	ne	da	ne	veliki	mali	*SMK-CAN-187*
veliki	mali	velika	ne	ne	da	mali	veliki	*ICPSR 23041 2*
veliki	mali	velika	ne	ne	da	veliki	mali	*ICPSR 6480*
veliki	mali	velika	ne	ne	ne	mali	veliki	*ICPSR 4537 10th form 2*
veliki	mali	velika	ne	ne	ne	veliki	mali	*ICPSR 4138*
veliki	mali	velika	da	da	da	mali	veliki	*ICPSR 23041*
veliki	mali	velika	da	da	da	veliki	mali	*ICPSR 4690*
veliki	mali	velika	da	da	ne	mali	veliki	*ICPSR 4372*
veliki	mali	velika	da	da	ne	veliki	mali	*ICPSR 20022*
veliki	mali	velika	da	ne	ne	mali	veliki	*ICPSR 6484*
veliki	mali	velika	da	ne	ne	veliki	mali	*ICPSR 4566 01*
veliki	mali	velika	da	ne	da	mali	veliki	*ICPSR 6693*
veliki	mali	velika	da	ne	da	veliki	mali	*ICPSR 4572 01*
veliki	veliki	mala	ne	da	da	mali	veliki	*Dorothea*
veliki	veliki	mala	ne	da	da	veliki	mali	*Human+Activity+Recognition+Using+Smartphones*
veliki	veliki	mala	ne	da	ne	mali	veliki	*ICPSR 31221*
veliki	veliki	mala	ne	da	ne	veliki	mali	*ICPSR 3669*
veliki	veliki	mala	ne	ne	da	mali	veliki	*ICPSR 2743 Person Level Data*

Broj atributa	Broj instanci	Oskudnost	Korelacija	Normalnost	Homogenost	Omjer ID	Šum atributa	Skup
veliki	veliki	mala	ne	ne	da	veliki	mali	*ICPSR 2258*
veliki	veliki	mala	ne	ne	ne	mali	veliki	*Madelon*
veliki	veliki	mala	ne	ne	ne	veliki	mali	*adult*
veliki	veliki	mala	da	da	da	mali	veliki	*ICPSR 31202 5*
veliki	veliki	mala	da	da	da	veliki	mali	*ICPSR 2857*
veliki	veliki	mala	da	da	ne	mali	veliki	*ICPSR 2346*
veliki	veliki	mala	da	da	ne	veliki	mali	*PEMS-SF*
veliki	veliki	mala	da	ne	ne	mali	veliki	*Dexter*
veliki	veliki	mala	da	ne	ne	veliki	mali	*ICPSR 2686 Caregiver Data*
veliki	veliki	mala	da	ne	da	mali	veliki	*ICPSR 3534*
veliki	veliki	mala	da	ne	da	veliki	mali	*ICPSR 2535*
veliki	veliki	velika	ne	da	da	mali	veliki	*ICPSR 2149*
veliki	veliki	velika	ne	da	da	veliki	mali	*Semeion+Handwritten+Digit*
veliki	veliki	velika	ne	da	ne	mali	veliki	*ICPSR 3548*
veliki	veliki	velika	ne	da	ne	veliki	mali	*Gisette*
veliki	veliki	velika	ne	ne	da	mali	veliki	*ICPSR 2295*
veliki	veliki	velika	ne	ne	da	veliki	mali	*ICPSR 2743*
veliki	veliki	velika	ne	ne	ne	mali	veliki	*ICPSR 2163*
veliki	veliki	velika	ne	ne	ne	veliki	mali	*SECOM*
veliki	veliki	velika	da	da	da	mali	veliki	*ICPSR 3789*
veliki	veliki	velika	da	da	da	veliki	mali	*ICPSR 2833*
veliki	veliki	velika	da	da	ne	mali	veliki	*Arcene*
veliki	veliki	velika	da	da	ne	veliki	mali	*ICPSR 2566*
veliki	veliki	velika	da	ne	ne	mali	veliki	*ICPSR 31202 4*

Broj atributa	Broj instanci	Oskudnost	Korelacija	Normalnost	Homogenost	Omjer ID	Šum atributa	Skup
veliki	veliki	velika	da	ne	ne	veliki	mali	*ICPSR 2039 2*
veliki	veliki	velika	da	ne	da	mali	veliki	*ICPSR 3151*
veliki	veliki	velika	da	ne	da	veliki	mali	

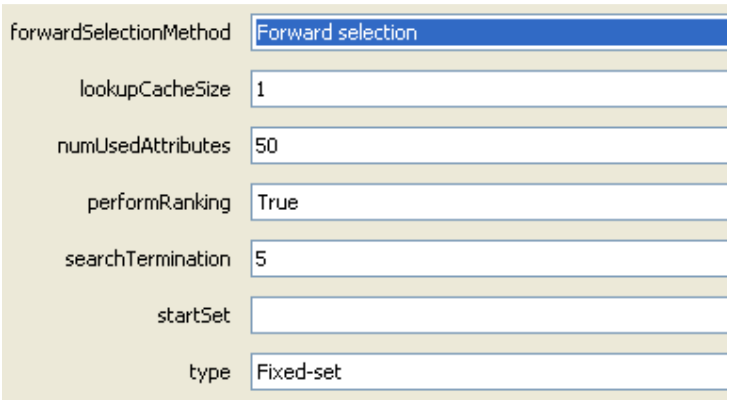
7.2. Selekcija atributa

U selekciji atributa primijenjene su tehnike: informacijska dobit, omjer dobiti, Relief, linearni odabir unaprijed, tehnika glasovanja, Stucco i Magnum Opus.
Prve četiri tehnike primijenjene su u alatu Weka 3.6., Magnum Opus u alatu Magnum Opus, a STUCCO je implementiran u Javi.

Iz početnog skupa od 17 atributa svakom tehnikom izdvojena su 4 atributa jer pod tim uvjetima postignuta najtočnija klasifikacija. U narednim podpoglavljima opisan je način provedbe i rezultati svake pojedine tehnike selekcije atributa.

7.2.1. Selekcija atributa dosad poznatim tehnikama

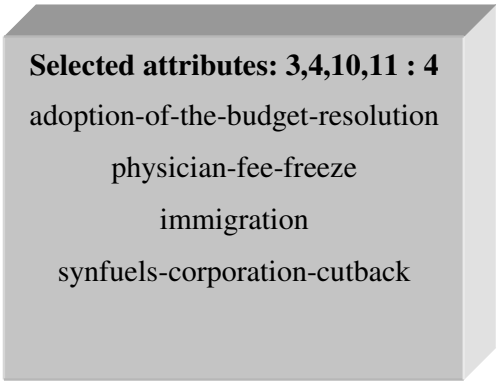
Postavke pod kojima je provedena tehnika linearni korak unaprijed pokazane su na slici 22.



Slika 22. Postavke tehnike linearni korak unaprijed

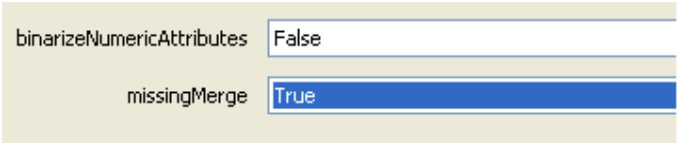
Prvi parametar definira smjer pretrage. Parametar broj korištenih atributa (eng. *numUsedAttributes*) definira koliko atributa se uzima u obzir u procesu traženja. Opcija izvođenje rangiranja (eng. *performRanking*) omogućuje rangiranje, a opcija *startSet* omogućuje postavljanje inicijalnog skupa atributa.

Pod ovim postavkama tehnika *Linearni odabir unaprijed* iz početnog skupa od 17 atributa selektirala ih je 4.



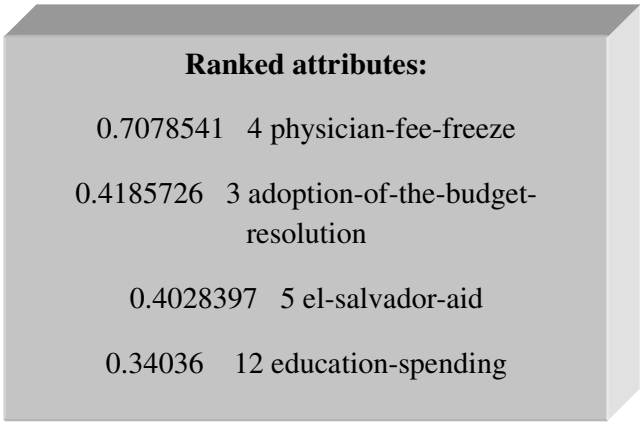
Slika 23. Atributi selektirani tehnikom *linearni korak unaprijed*

Tehnika informacijska dobit provedena je na postavkama prikazanim na slici 24.



Slika 24. Postavke tehnike informacijska dobit

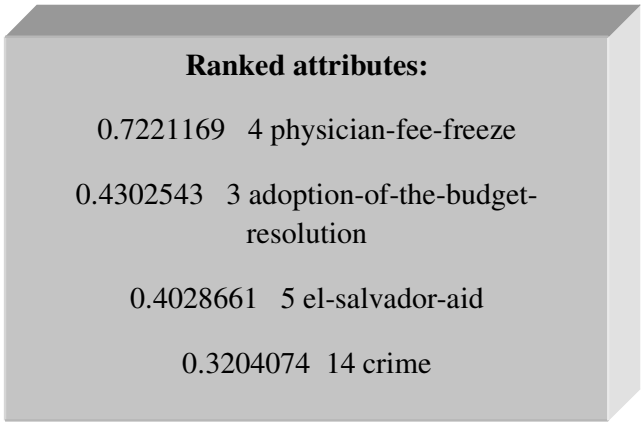
Dvije opcije prikazane na slici govore da se ne radi binarizacija numeričkih atributa te da se nedostajuće vrijednosti ne tretiraju kao zasebna vrijednost. Rezultati rangiranja atributa ovom tehnikom dani su na slici.



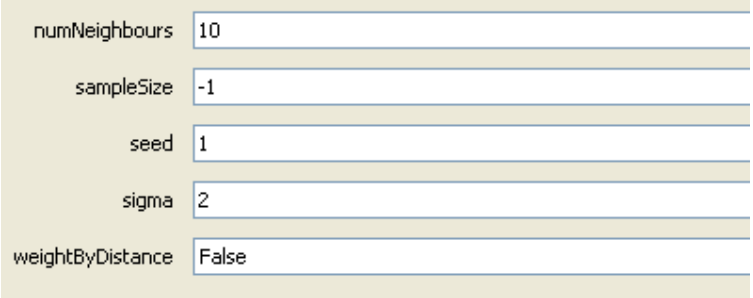
Slika 25. Rezultati provedbe informacijske dobiti

Tehnika omjer dobiti provedena je po istim postavkama kao i informacijska dobit i dala je slične rezultate, prikazane na slici 26.

Slika 26. Rezultati selekcije tehnikom omjer dobiti

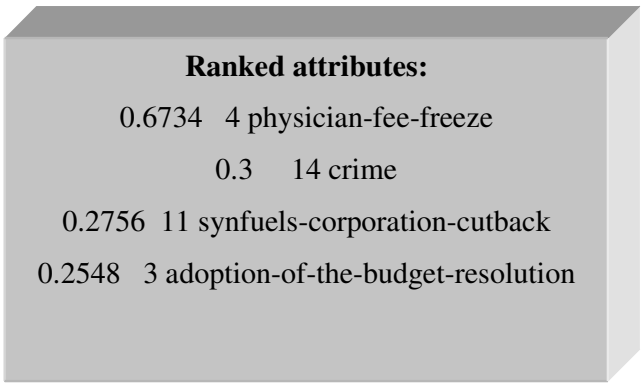


Relief tehnika provedena je u Weki s postavkama prikazanim na slici 27.



Slika 27. Postavke Relief tehnike

Prvi parametar definira broj najbližih susjeda za procjenu atributa. Vrijednost je postavljena na 10, kako je i inicijalno definirano u alatu. Vrijednost -1 kod parametra veličina uzorka (eng.sample size) ukazuje da će se sve instance koristiti za procjenu atributa. Ostale vrijednosti ostavljene su kako je inicijalno u alatu.
Dobiveni rezultati dani su na slici 28.

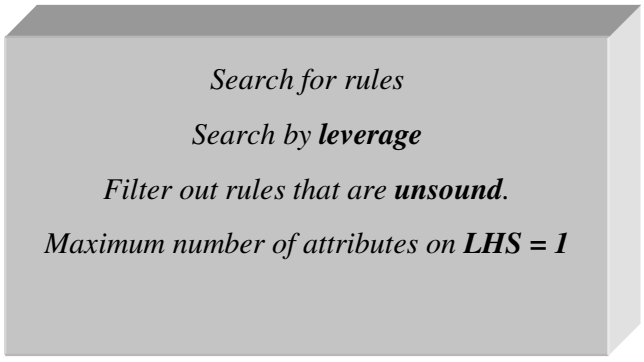


Slika 28. Rezultati Relief tehnike

Prva četiri njbolje rangirana atributa biti će ulaz u klasifikacijski model.

7.2.2. Selekcija atributa tehnikom Magnum Opus

Tehnika Magnum Opus FS primijenjena je na način opisan u poglavlju 5. Postavke su dane na slici 29.

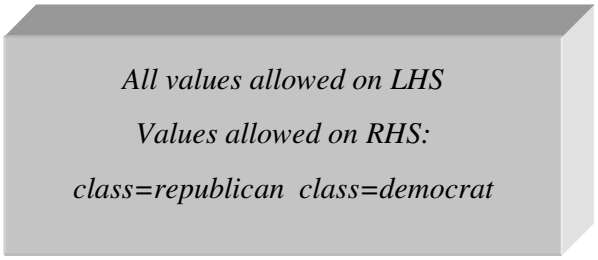


Slika 29. Postavke primjene *Magnum Opus FS*

Definirane postavke utvrđuju slijedeće aspekte. Mjera kvalitete pravila je utjecaj, te se atributi rangiraju s obzirom na vrijednost te mjere. Kao filter koristi se opcija *unsound* koja izdvaja samo statistički značajna pravila koja imaju vrijednost $p<0,05$. Zadnja postavka definira da se s lijeve strane pravila nalazi samo jedan atribut. To je kod primjene Magnum

Opusa za selekciju atributa izrazito važno jer se ne uzimanjem u obzir više atributa s lijeva strane zaobilazi interakcija atributa. U selekciji atributa Magnum Opusom kroz pravila želimo dobiti utjecaj pojedinog atributa na atribut klase, ali ne i utjecaj grupa atributa na atribut klase.

Stoga su s lijeve strane pravila dopuštene sve vrijednosti (svi ulazni atributi), a s desne strane samo samo atribut klase (u ovom slučaju atribut klase s vrijednostima: *republican* i *democrat*). To je i prikazano i na slici 30.



Slika 30. Dopuštene vrijednosti u pravilima

Kao rezultat izvršavanja Magnum Opus, pod definiranim postavkama, dao je 10 statistički značajnih pravila koja su dana u nastavku.

Found 10 rules

physician-fee-freeze=y -> class=republican

[Coverage=0.407 (177); Support=0.375 (163); Strength=0.921; Lift=2.38;
***Leverage=0.2176** (94.6); p=4.55E-095]*

physician-fee-freeze=n -> class=democrat

[Coverage=0.568 (247); Support=0.563 (245); Strength=0.992; Lift=1.62;
***Leverage=0.2147** (93.4); p=2.01E-092]*

adoption-of-the-budget-resolution=n -> class=republican

[Coverage=0.393 (171); Support=0.326 (142); Strength=0.830; Lift=2.15;
***Leverage=0.1746** (76.0); p=8.38E-057]*

adoption-of-the-budget-resolution=y -> class=democrat

[Coverage=0.582 (253); Support=0.531 (231); Strength=0.913; Lift=1.49;
***Leverage=0.1740** (75.7); p=9.85E-056]*

el-salvador-aid=y -> class=republican

[Coverage=0.487 (212); Support=0.361 (157); Strength=0.741; Lift=1.92;
Leverage=0.1727 (75.1); p=1.06E-055]

el-salvador-aid=n -> class=democrat

[Coverage=0.478 (208); Support=0.460 (200); Strength=0.962; Lift=1.57;
Leverage=0.1663 (72.3); p=7.17E-053]

education-spending=n -> class=democrat

[Coverage=0.536 (233); Support=0.490 (213); Strength=0.914; Lift=1.49;
Leverage=0.1609 (70.0); p=5.49E-047]

aid-to-nicaraguan-contras=y -> class=democrat

[Coverage=0.556 (242); Support=0.501 (218); Strength=0.901; Lift=1.47;
Leverage=0.1597 (69.5); p=4.21E-046]

education-spending=y -> class=republican

[Coverage=0.393 (171); Support=0.310 (135); Strength=0.789; Lift=2.04;
Leverage=0.1585 (69.0); p=3.73E-046]

mx-missile=n -> class=republican

[Coverage=0.474 (206); Support=0.336 (146); Strength=0.709; Lift=1.84;
Leverage=0.1527 (66.4); p=3.84E-042]

Pravila su sortirana su po veličini mjere utjecaja. Kako bi se detaljno opisali rezultati, u nastavku je izdvojeno jedno pravila i objašnjene su vrijednosti dobivene uz pravilo.

physician-fee-freeze=y -> class=republican

[Coverage=0.407 (177); Support=0.375 (163); Strength=0.921; Lift=2.38;
***Leverage=0.2176** (94.6); p=4.55E-095]*

Prva linija u pravilu daje kontrastni skup. Vrijednosti u zagradi su mjere kvalitete pravila (od *coverage* do *leverage*) te statistička značajnost pravila dana kroz vrijednost *p*. Mjera utjecaja je podebljana jer su na temelju te mjere vrednuju atributi u selekciji atributa.

Kao rezultat provedbe selekcije atributa primjenom tehnike Magnum Opus izdvojeni su atributi koji se sa svakom svojom vrijednošću nalaze s lijeve strane statistički značajnih pravila. Za skup podataka *vote* to su slijedeća četiri atributa:

- *physician-fee-freeze*
- *adoption-of-the-budget-resolution*
- *el-salvador-aid*
- *education-spendin*

7.2.3. Selekcija atributa tehnikom STUCCO

Za primjenu STUCCO algoritma podaci su u *.csv* formatu učitani u bazu i na njima je proveden algoritam sa slijedećim zadanim vrijednostima:

$$minDev = 0.1$$

$$alpha = 0.05$$

$$surprisingThreshold = 0.2$$

Pod ovim postavkama STUCCO traži kontrastne skupove koji su statistički značajni na razini $p<0,05$ i za koje je minimalna razlika u podršci 0.1.

Ovakve vrijednosti bile su zadane i u istraživanju Webba i suradnika u kojem je STUCCO evaluiran i pokazale su se učinkovitim (Webb, Butler i Newlands, 2003.)

Dobiveni kontrasni skupovi su u nastavku, a prva četiri od njih su značajni i veliki:

==== Node: SUPERFUND_RIGHT_TO_SUE = y;

Contingency table:

	republican	democrat
T:	4	4
F:	0	1
P:	1,00000	0,800000

==== Node: EDUCATION_SPENDING = y;

Contingency table:

	republican	democrat
T:	3	0
F:	1	5
P:	0,750000	0,00000

==== Node: CRIME = y;

Contingency table:

	republican	democrat
T:	4	4
F:	0	1
P:	1,00000	0,800000

==== Node: WATER_PROJECT_COST_SHARING = y;

Contingency table:

	republican	democrat
T:	4	5
F:	0	0
P:	1,00000	1,00000

==== Node: ADOPTION_OF_THE_BUDGET_RESOLUT = y;

Contingency table:

	republican	democrat
T:	0	4
F:	4	1
P:	0,00000	0,800000

==== Node: ADOPTION_OF_THE_BUDGET_RESOLUT = n;

Contingency table:

	republican	democrat
T:	4	1
F:	0	4
P:	1,00000	0,200000

==== Node: HANDICAPPED_INFANTS = n;

Contingency table:

	republican	democrat
T:	4	3
F:	0	2
P:	1,00000	0,600000

==== Node: PHYSICIAN_FEE_FREEZE = y;

Contingency table:

	republican	democrat
T:	4	1
F:	0	4
P:	1,00000	0,200000

==== Node: HANDICAPPED_INFANTS = n;

Contingency table:

	republican	democrat
T:	4	3
F:	0	2
P:	1,00000	0,600000

==== Node: PHYSICIAN_FEE_FREEZE = y;

Contingency table:

	republican	democrat
T:	4	1
F:	0	4
P:	1,00000	0,200000

Kao rezultat selekcije izdvajaju se slijedeći atributi:

SUPERFUND_RIGHT_TO_SUE

EDUCATION_SPENDING

CRIME

WATER_PROJECT_COST_SHARING

Selektirani atributi koriste se u daljem tijeku procesa otkrivanja znanja u podacima.

7.3. Klasifikacija

Vrednovanje algoritma temeljni je aspekt strojnog učenja. U ovom istraživanju performanse tehnika selekcije atributa evaluiraju se kroz točnost klasifikacije. Tehnike se uspoređuju tako da se na atributima selektiranim kroz sedam tehnika provodi klasifikacija primjenom neuronskih mreža (postupak opisan u ovom podpoglavlju) i diskriminacijska analize (opisano u slijedećem poglavlju). Postupak će se detaljno prikazati na primjeru skupa podataka *vote*, a za ostale skupove će se prikazati i interpretirati rezultati dobiveni po istom principu kao i za skup *vote*.

Za svaki skup od četiri atributa provedena je klasifikacija neuronskim mrežama. Pri tome je korišteno unakrsno vrednovanje (eng. *k-fold cross validation*). Točnost neuronske mreže za svaku od tehnika selekcije atributa prikazana je u tablici u nastavku. Kako je i objašnjeno u poglavlju 2., u ovom istraživanju koristi se metoda unakrsnog vrednovanja, a tablica daje srednju točnost klasifikacije. Učenje se provodi neuronskom mrežom s tri sloja: jednim ulaznim, jednim srednjim i jednim izlaznim slojem. Empirijski je dokazano da je neuronska mreža s jednim skrivenim slojem sposobna obraditi svaki skup podataka koji se nalazi na ulazu (Heaton, 2011.). Broj neurona u skrivenom sloju jednak je aritmetičkoj sredini neurona u ulaznom i izlaznom sloju. Znači, broj neurona u skrivenom sloju izračunava se na slijedeći način (Heaton, 2011.):

$$\frac{broj\ neurona\ na\ ulazu + broj\ neurona\ na\ izlazu}{2}$$

Klasifikacija neuronskim mrežama provedena je u *trial* verziji alata SAS JMP verzija 7. Ova verzija JMP-a implementira neuronsku mrežu širenja unutrag koja je opisana u drugom poglavlju.

Model neuronske mreže za skup podataka *vote* sastoji se od četiri neurona u ulaznom sloju, tri neurona u srednjem sloju te jednog neurona u izlaznom sloju. Točnost klasifikacije za pojedine tehnike selekcije dana je u tablici 6.

Tablica 6. Srednja točnost neuronske mreže

Tehnika selekcije	Točnost
Informacijska dobit	94,14
Omjer dobiti	94,21
Relief	94,33
Linearni odabir unaprijed	96,32
Voting	94,25
Magnum Opus	95,41
STUCCO	**97,79**

Kako bi se dobio odgovor na pitanje da li su razlike u točnosti između tehnika statistički značajne provodi se Friedman test. Prije provedbe testa utvrđeno je da su zadovoljeni uvjeti primjene testa opisani u drugom poglavlju. Friedman testom testira se nul hipoteza da nema razlike u rezultatima koje daju pojedine tehnike selekcije atributa. Tablica koja testira statistiku Friedman testa nalazi se u nastavku, a govori da li postoji statistički signifikantna razlika u točnosti između tehnika selekcije atributa. Vrijednosti koje daju su: vrijednost statistike testa *(hi kvadrat*), broj stupnjeva slobode (*f*) i razinu statističke značajnosti (*Asymp. Sig.).*

Tablica 7. Friedman test

N	7
Hi - kvadrat	32,071
df	7
Asymp. Sig.	,000

U statističkom vrednovanju za skup podataka *vote* zaključujemo da postoji statistički značajna razlika između točnosti tehnika selekcije atributa (*hi kvadrat= 32,071, p=0,0002*) .
Rezultati testa su pokazali slijedeće:

- Friedman statistika jednakosti preformansi tehnika iznosi 32,07 i ima *p* vrijednost od 0,0002

Ovi rezultati odbacuju nul hipotezu i indiciraju da postoji razlika u točnosti klasifikatora i samim time u performansama tehnika selekcije atributa. U tablici 8. tehnike selekcije su rangirane. Rangiranje tehnika je izvedeno s obzirom na postignute točnosti klasifikacije, a primjenom principa Friedman testa opisanog u poglavlju 2.3.2.

Tablica 8. Rangiranje tehnika selekcije za skup podataka *vote*

	Rang
Informacijska dobit	3,67
Omjer dobiti	7,00
Relief	6,00
Linearni odabir unaprijed	1,83
Tehnika glasovanja	4,67
MagnumOpus	3,33
STUCCO	1,50

Iz tablice se zaključuje kako je na skupu podataka vote tehnika STUCCO provela selekciju atributa koja je rezultirala najtočnijom klasifikacijom (najmanji rang 1,50), a tehnika linearni odabir unaprijed je druga najbrža (rang 1,83). Da je razlika između točnosti ove dvije tehnike statistička značajna vidljivo je iz tablice 7. Koja indicira statističku značajnost razlika u preformansama tehnika.

Ovi rezultati odbacuju nul hipotezu i ukazuju na to da postoji razlika u učinku tehnika selekcije atributa za ovaj skup podataka, a interpretiraju se na slijedeći način. Stucco daje najbolje rezultate, a slijedi ga tehnika linearni odabir unaprijed. Rangiranje tehnika selekcije atributa na temelju rezultata Friedman testa za svaki skup podataka dano je tablično u slijedećem poglavlju.

Dalje se provodi klasifikacija diskriminacijskom analizom. Pošto je diskiriminacijska analiza napredna statistička metoda koja ima neke pretpostavke (objašnjene u drugom poglavlju) koje skup podataka mora zadovoljiti da bi se mogla primijeniti, klasifikacija diskriminacijskom analizom je provedena samo na onim skupovima koji su zadovoljili te pretpostavke. Od 128 skupova podataka uključenih u istraživanje, 32 skupa su zadovoljila pretpostavke diskriminacijske analize. Rezultati rangiranja tehnika na skupovima koji zadovoljavaju pretpostavke dana je u tablici u slijedećem poglavlju i interpretirana. Skup *vote* nije zadovoljio pretpostavku (nema normalnu distribuciju) te stoga na tom skupu nije provedena diskriminacijska analiza.

7.4. Vrijeme provođenja selekcije atributa

Vrijeme provođenja selekcije atributa definira se kao vrijeme procesora potrebno da se provede selekcija. Sve analize provedene su na računalu s procesorom Intel(R)Atom(TM) CPU N450 1.67 GHz koji ima 32 bitni operacijski sustav.

Vrijeme se u ovom radu izražava u sekundama. Vrijeme provođenja selekcije atributa na skupu podataka *vote* dan je u tablici 9.

Tablica 9. Vrijeme provođenja selekcije atributa na skupu podataka

Tehnika selekcije	Vrijeme (sekunde)
Magnum Opus	2
Omjer dobiti	3
Relief	4
Informacijska dobit	6
Linearni odabir unaprijed	8
STUCCO	12

Najbrže je selekciju na ovom skupu podataka proveo Magnum Opus, slijede ga omjer dobiti i Relief. Friedman testom je utvrđeno da je razlika u vremenu provođenja statistički značajna (na što indiciraju rezultati tablice 11). Tablica 10. daje rangove tehnika selekcije atributa za skup podataka *vote*. Rangovi su izračunati na temelju vremena potrebnog za provođenje selekcije atributa prikazanog u tablici 9.

Tablica 10. Rangiranje tehnika na skupu *vote* prema brzini

Tehnika selekcije	Rang
MagnumOpus	1,50
Omjer Dobiti	2,00
Relief	2,90
Informacijska Dobit	3,90
Linearni Odabir Unaprijed	5,20
STUCCO	5,50

Rangiranje tehnika prikazano ukazuje da je Magnum Opus najbrža tehnika, a STUCCO najsporija. Omjer dobiti je druga najbrža tehnika, a slijedi ju Relief tehnika. Razlika u

vremenu provođenja selekcije izmeeđu tehnika je statistička značajna jer je $p<0,05$ (tablica 11.).

Tablica 11. Statistika Friedman testa za brzinu

N	5
Chi-Square	20,088
df	5
Asymp. Sig.	,001

Tablica 11. indicira postojanje statistički značajne razlike u rangovima pojedinih tehnika. Možemo zaključiti da postoji statistički značajna razlika između tehnika selekcije atributa s obzirom na vrijeme koje provede selekciju atributa ($p=0.01$)

8. REZULTATI

"*If you torture the data for long enough,*
in the end they will confess."
Ronald H. Coase

Eksperimenti objašnjeni u prethodnom poglavlju generirali su veliku količinu rezultata. Prikladna sumarna analiza tih rezultata je potrebna u svrhu interpretacije i prihvaćanja ili odbacivanja postavljenih hipoteza istraživanja, a to je dano u ovom poglavlju. Za pomoć u interpretaciji rezultata, generirano je mnogo tablica i slika koje kompariraju tehnike selekcije atributa. Najvažnije su prikazane u ovom poglavlju.

Istraživanje prezentirano u ovom radu se fokusiralo na usporedbu točnosti i vremena provedbe selekcije atributa primjenom tehnika otkrivanja kontrasta i ostalih, dosad najčešće korištenih, tehnika selekcije atributa. Različite tehnike djeluju različito na skupovima podataka koji se razlikuju u karakteristikama, a postupak evaluacije pokazan je u prethodnom poglavlju na jednom skupu podataka. Ovo poglavlje sintetizira rezultate dobivene na 128 skupova podataka različitih karakteristika i na temelju toga donosi zaključke s ciljem boljeg razumijevanja tehnika selekcije atributa.

Rezultati su prezentirani u tri dijela. U prvom dijelu (poglavlje 8.1.) daje se komparacija tehnika selekcija atributa gdje je kriterij točnost klasifikacije, a kao klasifikator korištene su neuronske mreže. U drugom dijelu (poglavlje 8.2.) uspoređuje se točnost tehnika selekcija atributa, a kao klasifikator korištene su neuronske mreže. Treći dio (poglavlje 8.3.) prikazuje rezultate iz perspektive vremena potrebnog za provedbu selekcije atributa te uspoređuje vremena različitih tehnika selekcije atributa.

Na temelju rezultata poglavlja 8.1 i 8.2 prihvaća se ili odbija prva hipoteza (H1:), a na temelju rezultata poglavlja 8.3. donose se zaključci vezani za drugu hipotezu.

8.1. Usporedba točnosti tehnika selekcije atributa - klasifikator neuronske mreže

Mnogo je različitih tehnika selekcija atributa razvijeno dosad. Ako primjenjujemo neku od tehnika na određenom zadatku, potrebno je definirati koja je tehnika najbolje prikladna za koji od problema. U ovom poglavlju se povezuju performanse tehnika selekcija atributa s karakteristikama skupova podataka, a rezultati su prikazani na slijedeći način. Najprije se daju rezultati provedbe Friedman testa i prikazuju se tablično na način da se za svaki skup podataka rangiraju tri tehnike selekcije atributa s najboljim rezultatima. Nakon toga se rezultati prikazuju u obliku stabla odlučivanja. U trećem dijelu se generiraju pravila, a u četvrtom se daje diskusija rezultata,. Proces generiranja skupova pravila koji povezuju ova dva koncepta (karakteristike podataka i tehnike selecije atributa) naziva se učenje meta razine (engl. *meta-level learning*).

Prethodna istraživanja su pokazala da se skupove podataka može karakterizirati pomoću određenih elemenata kao što su: broj atributa, broj instanci, šum atributa, oskudnost podataka te ostali statistički pokazatelji. Osnovna ideja ovog rada leži u povezivanju karakteristika skupa podataka i algoritama selekcije atributa. Ako se u procesu otkrivanja znanja u podacima za selekciju atributa uzme algoritam koji najviše odgovara karakteristikama skupa, povećava se vjerojatnost dobivanja korisnih rezultata. Znanje o tome koji algoritam se koristi u kojoj situaciji prikazano je u obliku pravila koja govore ako dani skup podataka ima određene karakteristike tada se primjenjuje određeni algoritam. Pri tome se koristi princip Brazdila i suradnika koji su znanje prikazivali prema slijedećoj shemi:

„AKO skup podataka ima karakteristike C_1, C_2,...:

TADA

Koristiti algoritam L_1.“

(Brazdil, Gama i Henery, 1994.)

što znači da atributi selektirani tehnikom L_1 postižu znatno veću točnost u odnosu na ostale tehnike selekcije atributa.

Ovo podpoglavlje daje rezultate klasifikacije primjenom neuronskih mreža na temelju kojih donosi zaključke o prikladnosti tehnika selekcije atributa za pojedine karakteristike skupa podataka.

Od 128 skupova podataka na kojima je provedena klasifikacija neuronskim mrežama za 82,03% skupova (105 skupa podataka) tehnike otkrivanja kontrasta dale su statistički značajnije točniju klasifikacije u odnosu na ostale tehnike selekcije atributa.

U 17,97% slučajeva (23 skupa podataka) tehnike otkrivanja kontrasta su dale lošije rezultate (manja točnost klasifikacije) od ostalih ili nisu statistički značajno bolje od ostalih.

- Na 23 skupa podataka tehnike otkrivanja kontrasta nisu bolje, i to u slijedećim situacijama:
    - Na 12 skupova podataka najbolji je *Relief*
    - Na 4 skupa podataka najbolji je *Informacijska dobit*
    - Na 2 skupa je najbolja tehnika *Linearni odabir unaprijed*
    - Na 5 skupova podataka tehnike otkrivanja kontrasta jesu bolje, ali razlika u točnosti između njih i ostalih tehnika nije statistički značajna

Tablica 12. prikazuje tri najbolje rangirane tehnike selekcije atributa za svaki od 128 skupa podataka.

Tablica 12. Rangiranje tehnika selekcije atributa u klasifikaciji neuronskim mrežama

Skup	**Rang tehnike selekcije atributa**
Pittsburgh+Bridges	1. Relief 2. Magnum Opus 3.STUCCO
Trains	1. STUCCO 2. Relief 3. Magnum Opus
Balloons	1. STUCCO 2. Relief 3. Omjer dobiti
Titanic	1. STUCCO 2. Informacijska dobit 3. Magnum Opus

Skup	Rang tehnike selekcije atributa
broadway	1. STUCCO 2. Relief 3. Magnum Opus
assessment	1. Magnum Opus 2. 2. STUCCO 3. Omjer dobiti
Soybean+Small	1. STUCCO 2. Omjer dobiti 3. Magnum Opus
molecular biology promoters	1. STUCCO 2. Informacijska dobit 3. Magnum Opus
Spectf	1. Relief 2. STUCCO 3. Magnum Opus
japansolvent	1. STUCCO 2. Voting 3. Magnum Opus
Post-Operative+Patient	1. STUCCO 2. Relief 3. Omjer dobiti
hepatitis	1. STUCCO 2. LFS 3. Magnum Opus
election	1. Relief 2. Magnum Opus 3. STUCCO
Lung Cancer	1. STUCCO 2. Magnum Opus 3. Relief
sponge	1. STUCCO 2. Relief 3. Voting
creditscore	1. Relief 2. Magnum Opus 3. STUCCO
bankruptcy	1. STUCCO 2. Informacijska dobit 3. Magnum Opus
gviolence	1. STUCCO 2. Omjer dobiti 3. Voting
Labor+Relations	1. Relief 2. Magnum Opus 3. STUCCO
Acute+Inflammations	1. STUCCO 2. Magnum Opus 3. LFS

Skup	Rang tehnike selekcije atributa
runshoes	1. STUCCO 2. Informacijska dobit 3. Magnum Opus
Cyyoung9302	1. STUCCO 2. Omjer dobiti 3. Voting
impeach	1. STUCCO 2. LFS 3. Informacijska dobit
fraud	1. STUCCO 2. Magnum Opus 3. Informacijska dobit
Campus Climate 2011 SJU	1. Magnum Opus 2. Informacijska dobit 3. Omjer dobiti
homerun	1. STUCCO 2. Magnum Opus 3. LFS
sonar	1. Relief 2. Magnum Opus 3. STUCCO
bondrate	1. STUCCO 2. Magnum Opus 3. Voting
ICPSR 3009	1. STUCCO 2. Relief 3. Magnum Opus
gsssexsurvey	1. STUCCO 2. Magnum Opus 3. STUCCO
uktrainacc	1. Relief 2. Magnum Opus 3. STUCCO
ncaa	1. STUCCO 2. Relief 3. LFS
credit	1. STUCCO 2. Relief 3. Magnum Opus
weights	1. InfoGain 2. Magnum Opus 3. STUCCO
ICPSR 2743	1. STUCCO 2. Voting 3. Relief
city	1. STUCCO 2. Magnum Opus 3. Voting

Skup	Rang tehnike selekcije atributa
supreme	1. Relief 2. Omjer dobiti 3. LFS
ICPSR 2751	1. STUCCO 2. Informacijska dobit 3. Magnum Opus
blood-transfusion	1. InfoGain 2. Relief 3. Magnum Opus
authorship	1. STUCCO 2. Omjer dobiti 3. Informacijska dobit
ICPSR 2867	1. STUCCO 2. Magnum Opus 3. LFS
ICPSR 2480	1. STUCCO 2. Magnum Opus 3. LFS
halloffame	1. STUCCO 2. Relief 3. Voting
CPS_85_Wages	1. Relief 2. Magnum Opus 3. STUCCO
Physical+Activity+Monitoring	1. STUCCO 2. Relief 3. Magnum Opus
marketing	1. STUCCO 2. LFS 3. Magnum Opus
binge	1. STUCCO 2. Relief 3. LFS
ionosphere	1. Relief 2. Magnum Opus 3. STUCCO
ICPSR 2859	1. InfoGain 2. Omjer dobiti 3. LFS
Mushroom	1. STUCCO 2. Informacijska dobit 3. Magnum Opus
ICPSR 2039	1. STUCCO 2. Relief 3. Informacijska dobit
Thyroid+Disease	1. STUCCO 2. Magnum Opus 3. LFS

Skup	Rang tehnike selekcije atributa
sick	1. STUCCO 2. Magnum Opus 3. Relief
One-hundred+plant+species+leaves+data+set	1. STUCCO 2. Omjer dobiti 3. Informacijska dobit
Kr-Vs-Kp	1. Relief 2. Omjer dobiti 3. Magnum Opus
tic-tac-toe	1. InfoGain 2. Omjer dobiti 3. STUCCO
abgss98	1. STUCCO 2. Magnum Opus 3. Relief
ICPSR 2686	1. STUCCO 2. Relief 3. Voting
ICPSR 2155	1. STUCCO 2. Voting 3. Magnum Opus
heart-statlog	1. STUCCO 2. Omjer dobiti 3. Voting
spambase	1. STUCCO 2. Magnum Opus 3. LFS
marketing	1. STUCCO 2. LFS 3. Magnum Opus
Hill-Valley	1. Relief 2. Magnum Opus 3. Informacijska dobit
hepatitis	1. STUCCO 2. Omjer dobiti 3. Informacijska dobit
ICPSR 4291	1. STUCCO 2. Relief 3. Omjer dobiti
ICPSR 4582	1. STUCCO 2. Informacijska dobit 3. Magnum Opus
ICPSR 9595	1. STUCCO 2. Relief 3. Omjer dobiti
ICPSR 21600 2	1. STUCCO 2. Magnum Opus 3. Voting

Skup	Rang tehnike selekcije atributa
ICPSR 21600 3	1. STUCCO 2. Informacijska dobit 3. Magnum Opus
ICPSR 28641 2	1. STUCCO 2. Omjer dobiti 3. Informacijska dobit
ICPSR 6542	1. STUCCO 2. Magnum Opus 3. Voting
ICPSR 4367	1. STUCCO 2. Magnum Opus 3. LFS
ICPSR 4572 02	1. STUCCO 2. Relief 3. LFS
ICPSR 21600 4	1. STUCCO 2. Magnum Opus 3. Relief
DBWorld+e-mails	1. STUCCO 2. Magnum Opus 3. Informacijska dobit
ICPSR 6135	1. STUCCO 2. Omjer dobiti 3. Magnum Opus
ICPSR 4537 8th form 1	1. STUCCO 2. Relief 3. Informacijska dobit
ICPSR 4275	1. STUCCO 2. Magnum Opus 3. Voting
GLI-85	1. STUCCO 2. Relief 3. LFS
ICPSR 21600 1	1. STUCCO 2. Magnum Opus 3. Relief
ICPSR 4566 02	1. STUCCO 2. LFS 3. Magnum Opus
ICPSR 8255	1. STUCCO 2. Magnum Opus 3. Voting
ICPSR 28641	1. STUCCO 2. Magnum Opus 3. LFS
SMK-CAN-187	1. STUCCO 2. Omjer dobiti 3. Magnum Opus

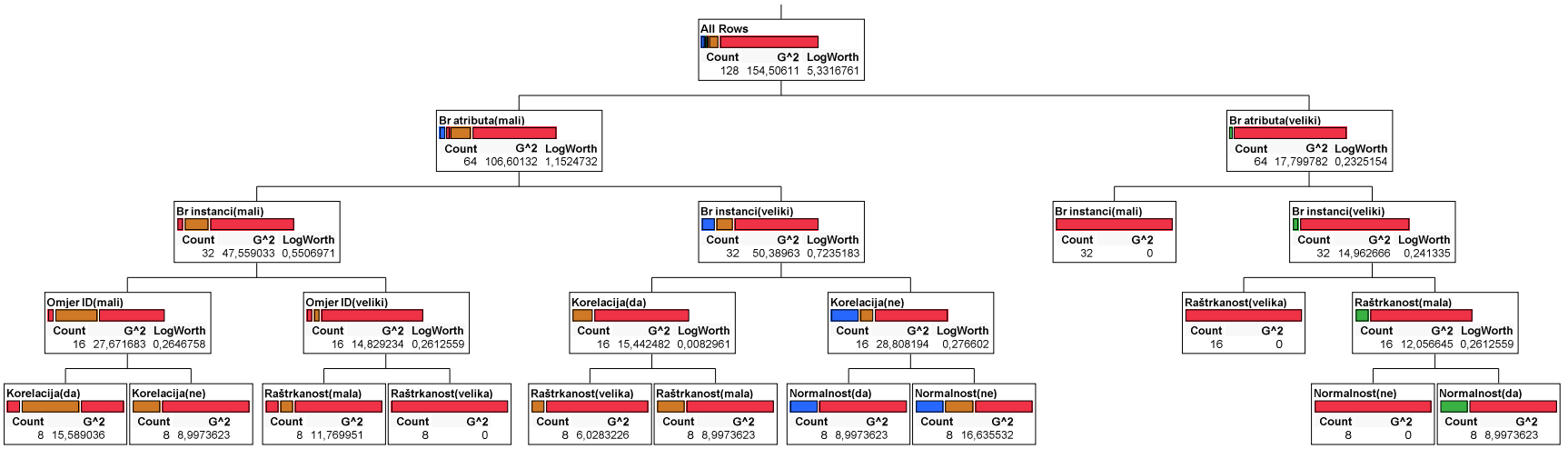
Skup	Rang tehnike selekcije atributa
ICPSR 23041 2	1. STUCCO 2. Magnum Opus 3. Omjer dobiti
ICPSR 6480	1. STUCCO 2. Informacijska dobit 3. Magnum Opus
ICPSR 4537 10th form 2	1. STUCCO 2. Magnum Opus 3. LFS
ICPSR 4138	1. STUCCO 2. Omjer dobiti 3. Magnum Opus
ICPSR 23041	1. STUCCO 2. Relief 3. Magnum Opus
ICPSR 4690	1. STUCCO 2. Magnum Opus 3. Relief
ICPSR 4372	1. STUCCO 2. Magnum Opus 3. Omjer dobiti
ICPSR 20022	1. STUCCO 2. Magnum Opus 3. LFS
ICPSR 6484	1. STUCCO 2. Voting 3. Magnum Opus
ICPSR 4566 01	1. STUCCO 2. Magnum Opus 3. Informacijska dobit
ICPSR 6693	1. STUCCO 2. Magnum Opus 3. Relief
ICPSR 4572 01	1. STUCCO 2. Magnum Opus 3. Omjer dobiti
Dorothea	1. STUCCO 2. Relief 3. Informacijska dobit
Human+Activity+Recognition +Using+Smartphones	1. STUCCO 2. Magnum Opus 3. Voting
ICPSR 31221	1. STUCCO 2. Relief 3. Magnum Opus
ICPSR 3669	1. LFS 2. Omjer dobiti 3. Informacijska dobit

Skup	Rang tehnike selekcije atributa
ICPSR 2743 Person Level Data	1. STUCCO 2. Relief 3. Voting
ICPSR 2258	1. STUCCO 2. Magnum Opus 3. Voting
Madelon	1. STUCCO 2. Relief 3. Omjer dobiti
adult	1. STUCCO 2. Magnum Opus 3. LFS
ICPSR 31202 5	1. STUCCO 2. Relief 3. LFS
ICPSR 2857	1. STUCCO 2. Magnum Opus 3. Relief
ICPSR 2346	1. LFS 2. Magnum Opus 3. Omjer dobiti
PEMS-SF	1. STUCCO 2. Informacijska dobit 3. Voting
Dexter	1. STUCCO 2. Relief 3. Omjer dobiti
ICPSR 2686 Caregiver Data	1. STUCCO 2. Magnum Opus 3. Voting
ICPSR 3534	1. STUCCO 2. Relief 3. Magnum Opus
ICPSR 2535	1. STUCCO 2. Informacijska dobit 3. Omjer dobiti
ICPSR 2149	1. STUCCO 2. Informacijska dobit 3. Omjer dobiti
Semeion+Handwritten+Digit	1. STUCCO 2. Magnum Opus 3. Voting
ICPSR 3548	1. STUCCO 2. Magnum Opus 3. Relief
Gisette	1. STUCCO 2. Omjer dobiti 3. Informacijska dobit

Skup	Rang tehnike selekcije atributa
ICPSR 2295	1. STUCCO 2. Magnum Opus 3. LFS
ICPSR 2743	1. STUCCO 2. Magnum Opus 3. Relief
ICPSR 2163	1. STUCCO 2. Omjer dobiti 3. Magnum Opus
SECOM	1. STUCCO 2. Voting 3. Informacijska dobit
ICPSR 3789	1. STUCCO 2. Relief 3. Magnum Opus
ICPSR 2833	1. STUCCO 2. Omjer dobiti 3. Voting
Arcene	1. STUCCO 2. Magnum Opus 3. Relief
ICPSR 2566	1. STUCCO 2. Informacijska dobit 3. Omjer dobiti
ICPSR 31202 4	1. STUCCO 2. Magnum Opus 3. Informacijska dobit
ICPSR 2039 2	1. STUCCO 2. Magnum Opus 3. LFS
ICPSR 3151	1. STUCCO 2. Magnum Opus 3. Omjer dobiti
ICPSR 6627	1. STUCCO 2. Magnum Opus 3. LFS

Iz tablice je vidljivo da na najvećem broju skupova podataka tehnike otkrivanja kontrasta postižu statistički značajnije točniju klasifikaciju u odnosu na druge tehnike selekcije atributa. Dobiveni rezultati prikazani su i kroz stablo odlučivanja kako bi se identificirale zakonitosti i utvrdilo za koje karakteristike podataka tehnike otkrivanja kontrasta nisu superiorne i koje su karakteristike najbitnije za proces klasifikacije. Stablo je prikazano na slici 31.

Slika 31. Ovisnost točnosti klasifikacije neuronskim mrežama o karakteristikama skupa podataka



	Informacijska dobit
	Linearni odabir unaprijed
	Relief
	Tehnika otkrivanja kontrasta (STUCCO, Magnum Opus)

Iz stabla odlučivanja može se očitati da su tehnike otkrivanja kontrasta superiorne u slučajevima s velikim brojem atributa. Od svih slučajeva s velikim brojem atributa jedino gdje pokazuju lošije performanse je u slučaju velikog broja instanci, male oskudnosti i normalne distribucije. U tim situacijama *Linearni odabir unaprijed* daje bolje rezultate.

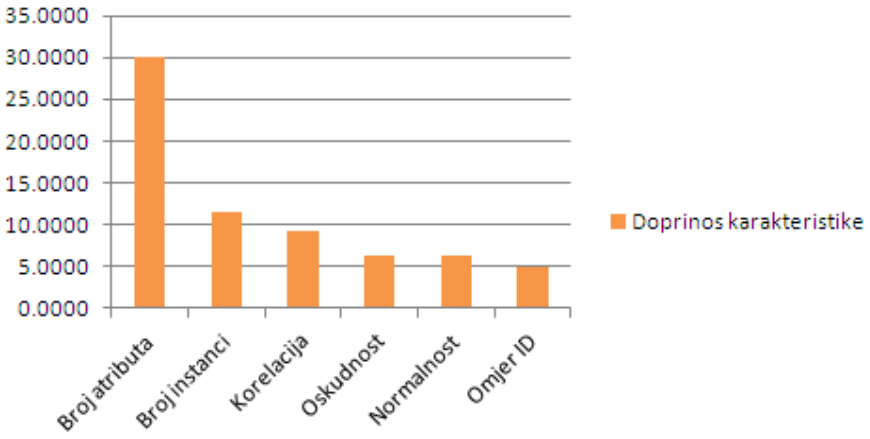
Relief postiže veću točnost klasifikacije jedino na 12 skupova s malim brojem atributa. Karakteristika 9 of tih 12 skupova je da imaju mali omjer unutarnje dimenzionalnosti, a karakteristično za preostala 3 skupa je da imaju malu oskudnost. Informacijska dobit na 4 skupa podataka s malim brojem atributa postiže veću točnost klasifikacije od tehnika otkrivanja kontrasta. Druga karakteristika zajednička za ova četiri skupa je mala korelacija. Na 2 skupa podataka koji imaju velik broj atributa i velik broj instanci je najbolja tehnika *Linearni odabir unaprijed.*

Kada se pogleda distribucija skupova podataka na kojima tehnike otkrivanja kontrasta nisu dale najbolje rezultate dolazi se do slijedećih zaključaka.

Prosječan broj atributa je 97 (SD 200) što je znatno manje u odnosu na prosječan broj atributa svih 128 skupova koji su sudjelovali u analizi (1264 atributa). Prosječan broj instanci skupova na kojima tehnike otkrivanja kontrasta nisu dale najbolje rezultate je 9741 što je mnogo više u odnosu na prosječan broj instanci svih 128 skupova (2105 instanci). Prosječna korelacija ovih 18 skupova (0.229) je približna ista prosječnoj korelaciji svih 128 skupova (0.245), kao i omjer unutarnje dimenzionalnosti (0.623 omjer unutarnje dimenzionalnosti svih skupova, a 0.549 omjer unutarnje dimenzionalnosti 18 skupova na kojima tehnike otkrivanja kontrasta nisu superiorne).

Jedan od doprinosa ovog rada je i opsežno istraživanje u prostoru karakteristika skupa podataka. Postavlja se pitanje koje karakteristike skupa podataka su bitne za proces otkrivanja znanja u podacima. Značajnost karakteristika se može očitati i iz stabla odlučivanja kroz opciju doprinos stupca (eng. *column contribution*). Doprinos predstavlja mjeru u kojoj pojedini stupac tj. atribut (u ovom slučaju karakteristika podataka) daje modelu informacije za klasifikaciju vrijednosti zavisnog atributa. Doprinosi karakteristika podataka grafički su prikazani na slici 32.

Slika 32. Doprinos karakteristika kod klasifikacije neuronskim mrežama



Kod klasifikacije neuronskim mrežama, broj atributa se pokazuje kao daleko najvažnija karakteristika. Slijede ga broj instanci i korelacija, ali sa značajnije manjim doprinosom. Za primjetiti je da kod klasifikacije neuronskim mrežama još tri karakteristike skupa podataka doprinose razlikovanju tehnika selekcije atributa (oskudnost, normalnost i omjer unutarnje dimenzionalnosti), dok homogenost matrica kovarijanci nema nikakav značaj.

Na temelju tablice 12. izdvojeni su skupovi podataka nad kojima tehnike otkrivanja kontrasta u selekciji atributa nisu rezultirale najtočnijom klasifikacijom neuronskim mrežama. Ti rezultati spojeni su s opisom karakteristika skupa podataka iz tablice 5. i na temelju rezultata tih dviju tablica složena su pravila koja govore na skupovima podataka kojih karakteristika ne treba koristiti tehnike otkrivanja kontrasta u selekciji atributa.

Kao zaključak podpoglavlja 8.1. tvrdi se kako se u klasifikaciji neuronskim mrežama predlaže primjena tehnika otkrivanja kontrasta u selekciji atributa za sve karakteristike podataka, osim za slijedeće situacije:

- *AKO* skup podataka ima karakteristike

  Broj atributa: **mali** *Broj instanci: mali Oskudnost: mala Korelacija: ne Normalnost: da Homogenost: da Omjer ID: mali Šum atributa: veliki*

TADA koristiti algoritam *Relief* .

- *AKO* skup podataka ima karakteristike

  Broj atributa: mali Broj instanci: mali Oskudnost mala Korelacija: da Normalnost: da Homogenost: da Omjer ID: veliki Šum atributa: mali

  TADA koristiti algoritam *Relief*.

- AKO skup podataka ima karakteristike

  Broj atributa: mali Broj instanci: mali Oskudnost mala Korelacija: da Normalnost: ne Homogenost: ne Omjer ID: mali Šum atributa: veliki

  TADA koristiti algoritam *Relief*.

- AKO skup podataka ima karakteristike

  Broj atributa: mali Broj instanci: mali Oskudnost mala Korelacija: da Normalnost: ne Homogenost: da Omjer ID: veliki Šum atributa: mali

  TADA koristiti algoritam *Relief*.

- AKO skup podataka ima karakteristike

  Broj atributa: mali Broj instanci: mali Oskudnost velika Korelacija: ne Normalnost: da Homogenost: ne Omjer ID: mali Šum atributa: veliki

  TADA koristiti algoritam *Relief*.

- AKO skup podataka ima karakteristike

  Broj atributa: mali Broj instanci: mali Oskudnost velika Korelacija: da Normalnost: da Homogenost: ne Omjer ID: mali Šum atributa: veliki

  TADA koristiti algoritam *Relief*.

- AKO skup podataka ima karakteristike

  Broj atributa: mali Broj instanci: mali Oskudnost velika Korelacija: da Normalnost: ne Homogenost: da Omjer ID: mali Šum atributa: veliki

  TADA koristiti algoritam *Relief*.

- AKO skup podataka ima karakteristike

  Broj atributa: mali Broj instanci: veliki Oskudnost mala Korelacija: ne Normalnost: da Homogenost: da Omjer ID: veliki Šum atributa: mali

  TADA koristiti algoritam *Informacijska dobit*.

- AKO skup podataka ima karakteristike

  Broj atributa: mali Broj instanci: veliki Oskudnost mala Korelacija: ne Normalnost: ne Homogenost: da Omjer ID: mali Šum atributa: veliki

  TADA koristiti algoritam *Relief*.

- AKO skup podataka ima karakteristike

- *Broj atributa: mali Broj instanci: veliki Oskudnost mala Korelacija: ne Normalnost: ne Homogenost: ne Omjer ID: mali Šum atributa: veliki*

  TADA koristiti algoritam *Informacijska dobit*.

- AKO skup podataka ima karakteristike

  Broj atributa: mali Broj instanci: veliki Oskudnost mala Korelacija: da Normalnost: da Homogenost: ne Omjer ID: veliki Šum atributa: mali

  TADA koristiti algoritam *Relief*.

- AKO skup podataka ima karakteristike

  Broj atributa: mali Broj instanci: veliki Oskudnost mala Korelacija: da Normalnost: ne Homogenost: da Omjer ID: veliki Šum atributa: mali

  TADA koristiti algoritam *Relief*.

- AKO skup podataka ima karakteristike

  Broj atributa: mali Broj instanci: veliki Oskudnost velika Korelacija: ne Normalnost: da Homogenost: da Omjer ID: mali Šum atributa: veliki

  TADA koristiti algoritam *Informacijska dobit*.

- AKO skup podataka ima karakteristike

  Broj atributa mali Broj instanci: veliki Oskudnost velika Korelacija: ne Normalnost: ne Homogenost: ne Omjer ID: mali Šum atributa: veliki

  TADA koristiti algoritam *Relief*.

- AKO skup podataka ima karakteristike

  Broj atributa mali Broj instanci: veliki Oskudnost velika Korelacija: ne Normalnost: ne Homogenost: ne Omjer ID: veliki Šum atributa: mali

  TADA koristiti algoritam *Informacijska dobit*.

- AKO skup podataka ima karakteristike

  Broj atributa mali Broj instanci: veliki Oskudnost velika Korelacija: da Normalnost: ne Homogenost: da Omjer ID: mali Šum atributa: veliki

  TADA koristiti algoritam *Relief*.

- AKO skup podataka ima karakteristike

Broj atributa veliki Broj instanci: veliki Oskudnost mala Korelacija: ne Normalnost: da Homogenost: ne Omjer ID: veliki Šum atributa: mali

TADA koristiti algoritam *Linearni odabir unaprijed*.

- AKO skup podataka ima karakteristike

  Broj atributa veliki Broj instanci: veliki Oskudnost mala Korelacija: da Normalnost: da Homogenost: ne Omjer ID: mali Šum atributa: veliki

  TADA koristiti algoritam *Linearni odabir unaprijed*.

8.2. Usporedba tehnika selekcije atributa – klasifikator diskriminacijska analiza

Tehnike otkrivanja kontrasta u selekciji atributa definirane su kao tehnike filtra. Iako tehnike omotača ponekad rezultiraju točnijom klasifikacijom, one koriste samo jedan klasifikator u odabiru relevantnih atributa. Problem kod upotrebe samo jednog klasifikatora može predstavljati činjenica da je priroda svakog klasifikatora različita. Različiti klasifikatori imaju različit utjecaj na selekciju atributa. Npr. jedna vrsta klasifikatora može biti više (ili manje) pogodna za odabir atributa nego druga vrsta klasifikatora. Uzrok tome može biti u činjenici da se karakteristike jednog klasifikatora podudaraju s karakteristikama skupa podataka i/ili korištene tehnike selekcije atributa. (Chrysostomou, 2008.)

Stoga se u ovom istraživanju ispituje i utjecaj tehnika selekcije atributa i karakteristika podataka na različite klasifikatore. Ovo podpoglavlje prikazuje rezultate klasifikacije dobivene diskriminacijskom analizom. Diskriminacijska analiza je provedena na 32 skupa podataka koji su zadovoljili pretpostavke opisane u poglavlju 2. Rangiranje tehnika selekcije dano je u tablici 13.

Tablica 13. Rangiranje tehnika selekcije u klasifikaciji diskriminacijskom analizom

Skup	Rang tehnika selekcije
Pittsburgh+Bridges	1. STUCCO 2. Relief 3. Voting
Trains	1. STUCCO 2. Magnum Opus 3. Relief
Spectf	1. STUCCO 2. Relief 3. Informacijska dobit
japansolvent	1. STUCCO 2. Magnum Opus 3. Informacijska dobit
bankruptcy	1. STUCCO 2. Omjer dobiti 3. Informacijska dobit
gviolence	1. STUCCO 2. Voting 3. Magnum Opus
Campus Climate 2011 SJU	1. Relief 2. STUCCO 3. Omjer dobiti
homerun	1. STUCCO 2. Magnum Opus 3. LFS
credit	1. STUCCO 2. Magnum Opus 3. Relief
weights	1. Relief 2. Magnum Opus 3. Informacijska dobit
ICPSR 2867	1. STUCCO 2. LFS 3. Magnum Opus
ICPSR 2480	1. Magnum Opus 2. STUCCO 3. LFS
ICPSR 2859	1. Magnum Opus 2. LFS 3. Omjer dobiti
Mushroom	1. STUCCO 2. Informacijska dobit 3. Magnum Opus
abgss98	1. STUCCO 2. Relief 3. Magnum Opus

Skup	Rang tehnike selekcije
ICPSR 2686	1. Magnum Opus 2. STUCCO 3. Relief
ICPSR 4291	1. Relief 2. Omjer dobiti 3. STUCCO
ICPSR 4582	1. Informacijska dobit 2. STUCCO 3. Magnum Opus
ICPSR 4572 02	1. STUCCO 2. LFS 3. Relief
ICPSR 21600 4	1. STUCCO 2. Magnum Opus 3. Voting
ICPSR 4566 02	1. LFS 2. STUCCO 3. Magnum OPUS
ICPSR 8255	1. STUCCO 2. Magnum Opus 3. Relief
ICPSR 23041	1. Magnum Opus 2. Relief 3. STUCCO
ICPSR 4690	1. STUCCO 2. Magnum Opus 3. Relief
Dorothea	1. Magnum Opus 2. STUCCO 3. Informacijska dobit
Human+Activity+Recognition+Using+Smartphones	1. STUCCO 2. Magnum Opus 3. Voting
ICPSR 31202 5	1. STUCCO 2. LFS 3. Relief
ICPSR 2857	1. Magnum Opus 2. STUCCO 3. Relief
ICPSR 2149	1. STUCCO 2. Magnum Opus 3. Voting
Semeion+Handwritten+Digit	1. STUCCO 2. LFS 3. Voting
ICPSR 3789	1. Magnum Opus 2. Relief 3. STUCCO

Skup	Rang tehnike selekcije
ICPSR 2833	1. STUCCO 2. Relief 3. Magnum Opus

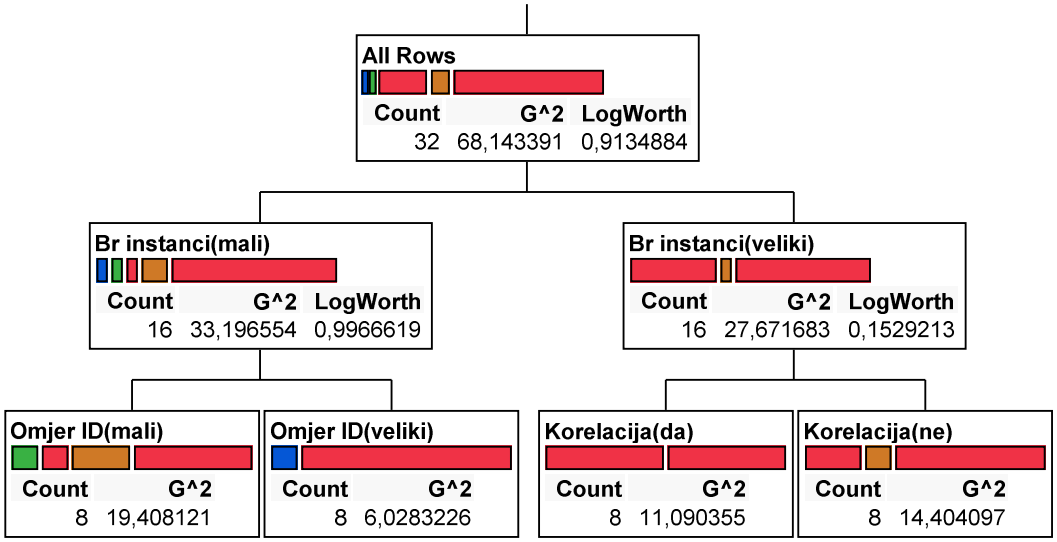
Od 32 skupa podataka koji su zadovoljili pretpostavke i nad kojima je provedna klasifikacija diskriminacijskom analizom, na 78,12% (25 skupova podatka) tehnike otkrivanja kontrasta dale su bolje rezultate od ostalih.

U 21,88% (7 skupova) tehnike otkrivanja kontrasta su dalje lošije (manja točnost klasifikacije) rezultate od ostalih ili nisu statistički značajno bolje od ostalih. Na ovih 7 skupova slijedeći su rezultati:

- Na 3 skupa podataka najbolji je *Relief*
- Na 1 skupu je najbolji *Informacijska dobit*
- Na 1 skupu je najbolji *Linearni odabir unaprijed*
- Na 2 skupa tehnike otkrivanja kontrasta jesu bolje, ali razlika u točnosti između njih i ostalih tehnika nije statistički značajna

Dobiveni rezultati prikazani su i pomoću stabla odlučivanja iz kojeg se jasnije vidi kakav utjecaj karakteristike podataka imaju na različite tehnike selekcije atributa. Stablo je prikazano na slici 33.

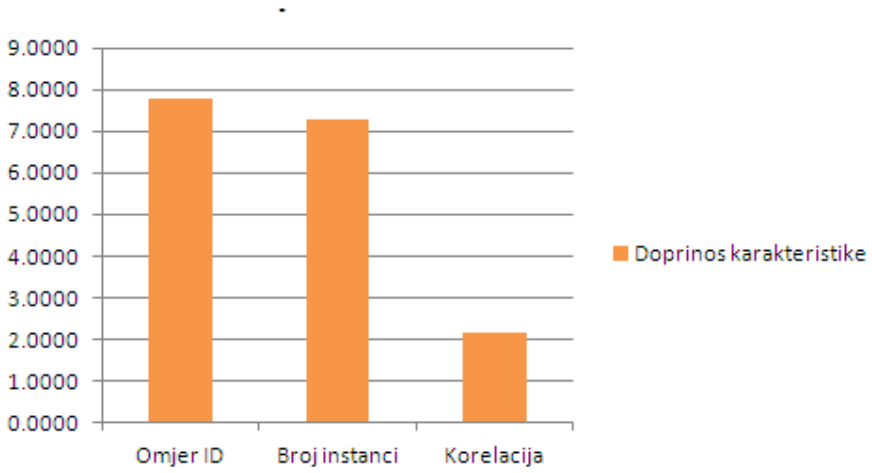
Slika 33. Ovisnost točnosti klasifikacije diskriminacijskom analizom o karakteristikama skupa podataka



	Informacijska dobit
	Linearni odabir unaprijed
	Relief
	Tehnika otkrivanja kontrasta (STUCCO, Magnum Opus)

Na 5 skupova podataka tehnike otkrivanja kontrasta nisu rezultirale najtočnijom klasifikacijom. Karakteristika svih 5 skupova je normalnost skupa podataka i homogenost kovarijanci klasa. Tehnike otkrivanja kontrasta rezultiraju točnijom klasifikacijom gotovo na svim skupovima podataka s velikim brojem instanci osim na jednom gdje *Relief* rezultira najtočnijom klasifikacijom. Sveukupno, *Relief* algoritam se pokazao najboljim na 3 skupa podataka. Od toga 2 skupa imaju mali omjer unutarnje dimenzionalnosti i to oni skupovi koji imaju mali broj instanci. Na jednom skupu podataka informacijska dobit daje najbolje rezultate, a na drugom linearni odabir unaprijed. Karakteristika oba skupa je veliki broj atributa i mali broj instanci te mala korelacija. Razlikuju se u oskudnosti (informacijska dobit je bolja gdje je oskudnost skupa manja, a linearni odabir unaprijed gdje je oskudnost velika) te u šumu (informacijska dobit daje bolju klasifikaciju na skupu s malim šumom, a linearni odabir unaprijed na skupu s velikim šumom).

Ovih 5 skupova na kojima klasifikacija diskriminacijskom analizom nije rezultirala najtočnijom klasifikacijom za tehnike otkrivanja kontrasta ima manji prosječan broj atributa.



Slika 34. Doprinos karakteristika podataka kod klasifikacije diskriminacijskom analizom

Kod klasifikacije diskriminacijskom analizom tri karakteristike skupa podataka se izdvajaju kao važne za klasifikaciju, i to redom: omjer unutarnje dimenzionalnosti, broj instanci i korelacija (slika 34.). Treba primjetiti kako se kod klasifikacije neuronskim mrežama najvažnijom karakteristikom pokazala broj atributa koja u klasifikaciji diskriminacijskom analizom nije prepoznata kao važna.

Kao zaključak podpoglavlja 8.2. tvrdi se kako se u klasifikaciji diskriminacijskom analizom, predlaže primjena tehnika otkrivanja kontrasta u selekciji atributa za sve karakteristike podataka, osim za slijedeće situacije:

AKO skup podataka ima karakteristike

Broj atributa: mali Broj instanci: mali Oskudnost: velika Korelacija: da Normalnost: da Homogenost: da Omjer ID: mali Šum atributa: veliki

TADA koristiti algoritam *Relief* .

AKO skup podataka ima karakteristike

Broj atributa: mali Broj instanci: veliki Oskudnost: mali Korelacija: ne Normalnost: da Homogenost: da Omjer ID: veliki Šum atributa: mali

TADA koristiti algoritam *Relief* .

AKO skup podataka ima karakteristike

Broj atributa: veliki Broj instanci: mali Oskudnost: mala Korelacija: ne Normalnost: da Homogenost: da Omjer ID: mali Šum atributa: veliki

TADA koristiti algoritam *Relief* .

AKO skup podataka ima karakteristike

Broj atributa: veliki Broj instanci: mali Oskudnost: mala Korelacija: ne Normalnost: da Homogenost: da Omjer ID: veliki Šum atributa: mali

TADA koristiti algoritam *Omjer dobiti* .

AKO skup podataka ima karakteristike

Broj atributa: veliki Broj instanci: mali Oskudnost: velika Korelacija: ne Normalnost: da Homogenost: da Omjer ID: mali Šum atributa: veliki

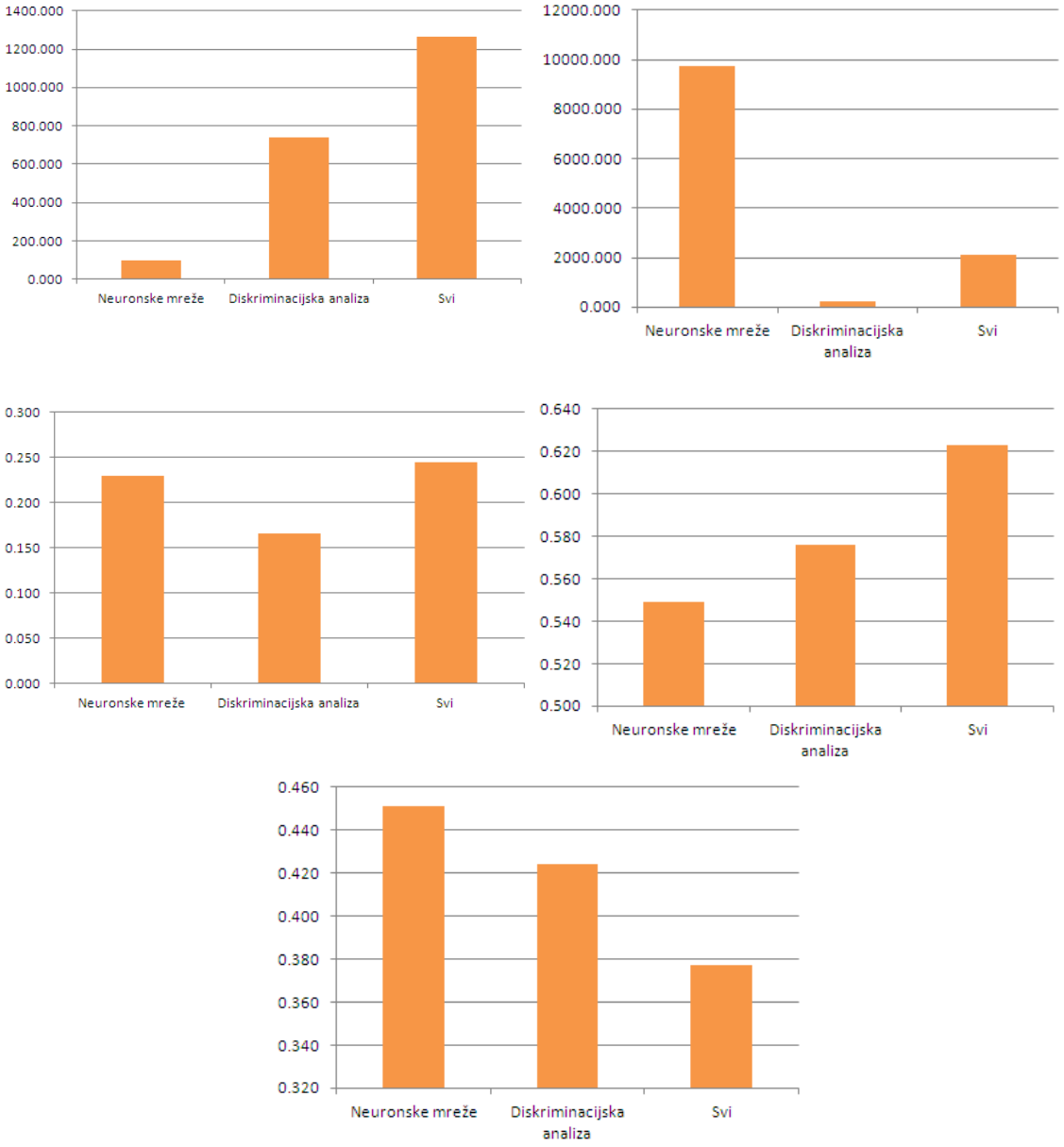
TADA koristiti algoritam *Linearni odabir unaprijed* .

8.3. Usporedba rezultata dobivenih neuronskim mrežama i diskriminacijskom analizom

Samo na jednom skupu podataka (skup podataka *weights*) tehnike otkrivanja kontrasta nisu rezultirale najtočnijom klasifikacijom na oba klasifikatora, neuronske mreže i diskriminacijsku analizu. Karakteristika toga skupa je mali broj atributa i velik broj instanci te mala oskudnost i mali šum, normalnost i homogenost skupa podataka. Zanimljivo je primjetiti da je kod klasifikacije neuronskim mrežama na tom skupu natočnije rezultate dala informacijska dobit u selekciji atributa, a kod klasifikacije diskriminacijskom analizom tehnika *Relief*.

Nadalje, izračunate su prosječne vrijednosti karakteristika skupova podataka nad kojima tehnike otkrivanja kontrasta nisu dale najtočniju klasifikaciju neuronskim mrežama (skupova)

i diskriminacijskom analizom (skupova) te su uspoređene s prosječnom vrijednošću karakteristika svih 128 skupova koji su sudjelovali u analizi. Grafički prikaz usporedbe aritmetičkih sredina dan je na slici 35.



Slika 35. Prosječne vrijednosti karakteristika

Prosječna vrijednost karakteristika broj atributa prikazane na slici 35a za skupove podataka na kojima tehnike otkrivanja kontrasta nisu dale najtočniju klasifikaciju neuronskim mrežama

znatno je manja u odnosu na prosječan broj atributa svih 128 skupova iz analize. Prosječna vrijednost broja atributa za skupove na kojima klasifikacija diskriminacijskom analizom nije dala najtočniju klasifikaciju za tehnike otkrivanja kontrasta znatno je veća nego za neuronske mreže, ali manja od prosječne vrijednosti svih 128 skupova.

Slijedom navedenog, zaključuje se da upotreba tehnika otkrivanja kontrasta u selekciji atributa rezultira točnijom klasifikacijom za skupove podataka s velikim brojem atributa, u slučaju oba klasifikatora.

Prosječne vrijednosti za karakteristiku broj instanci dane su na slici 35b. Kod ove karakteristike rezultati neuronskih mreža i diskriminacijske analize se znatno razlikuju. Naime, u klasifikaciji neuronskim mrežama tehnike otkrivanja kontrasta nisu dale dobre rezultate na skupovima podataka s većim brojem instanci u odnosu na prosjek, dok u klasifikaciji diskriminacijskom analizom nisu dale dobre rezultate na skupovima sa znatno manjim brojem instanci u odnosu na prosjek. Ovdje se primjećuje kako pojedine karakteristike podataka imaju različito djeluju na različite klasifikatore.

Slika 35c daje pregled prosječnih vrijednosti za korelaciju. Kod ove karakteristike skupa podataka prosječna vrijednost skupova na kojima primjena tehnika otkrivanja kontrasta u selekciji atributa nije rezultirala najtočnijom klasifikacijom ne odstupa znatno od prsječne vrijednosti svih skupova. U slučaju oba klasifikatora prosječna korelacija malo je manja od prosječne vrijednosti s napomenom kako je niža kod klasifikacije diskriminacijskom analizom.

Prosječne vrijednosti karakteristika omjer unutarnje dimenzionalnosti i šum atributa prikazane su na slikama 35d i 35c. Tehnike otkrivanja kontrasta dale su slabije rezultate na skupovima podataka koji imaju omjer unutarnje dimezionalnosti manji od prosjeka, a šum atributa veći od prosjeka. Možemo zaključiti da daju dobre rezultate na skupovima podataka s velikim šumom.

8.4. Usporedba tehnika selekcije atributa – vrijeme provedbe selekcija atributa

Koristeći samo točnost klasifikacije kao jedini kriterije usporedbe, u praksi nije nužno optimalno jer je često značajka sporijih tehnika selekcije atributa veća točnost klasifikacije, a to ponekad rezultira većim troškovima. Stoga je potrebno napraviti neku vrstu kompromisa između vremena izvođenja tehnika selekcije atributa i točnosti klasifikacije.

Ovo poglavlje prikazuje evaluaciju tehnika selekcije atributa s obzirom na vrijeme izvođenja. Rezultati rangiranja tri najbrže tehnike za svaki od 128 skupova dani su u tablici 14. Vrijeme izvođenja tehnika selekcije atributa za svaku tehniku danu je u prilogu ovog rada, a izraženo je u sekundama. Pod vremenom izvođenja smatra se vrijeme procesora potrebno za provedbu selekcije atributa.

Tablica 14. Rangiranje tehnika selekcije atributa s obzirom na brzinu izvođenja

Skup	Rang tehnika selekcije
Pittsburgh+Bridges	1. Magnum Opus 2. Informacijska dobit 3. Omjer dobiti
Trains	1. Magnum Opus 2. Relief 3. Omjer dobiti
Balloons	1. Magnum Opus 2. Informacijska dobit 3. Relief
Titanic	1. Omjer dobiti 2. Magnum Opus 3. Informacijska dobit
broadway	1. Omjer dobiti 2. Relief 3. Magnum Opus
assessment	1. Magnum Opus 2. Omjer dobiti 3. Relief
Soybean+Small	1. Magnum Opus 2. Relief 3. Omjer dobiti

Skup	Rang tehnika selekcije
molecular biology promoters	1. Omjer dobiti 2. Informacijska dobit 3. Relief
Spectf	1. Magnum Opus 2. Relief 3. Omjer dobiti
japansolvent	1. Magnum Opus 2. Informacijska dobit 3. Relief
Post-Operative+Patient	1. Omjer dobiti 2. Informacijska dobit 3. Magnum Opus
hepatitis	1. Magnum Opus 2. Informacijska dobit 3. Relief
election	1. Informacijska dobit 2. Magnum Opus 3. Omjer dobiti
Lung Cancer	1. Magnum Opus 2. Informacijska dobit 3. Omjer dobiti
sponge	1. Magnum Opus 2. Omjer dobiti 3. Informacijska dobit
creditscore	1. Omjer dobiti 2. Magnum Opus 3. Informacijska dobit
bankruptcy	1. Magnum Opus 2. Omjer dobiti 3. Informacijska dobit
gviolence	1. Magnum Opus 2. Informacijska dobit 3. Omjer dobiti
Labor+Relations	1. Magnum Opus 2. Omjer dobiti 3. Informacijska dobit
Acute+Inflammations	1. Omjer dobiti 2. Magnum Opus 3. Informacijska dobit
runshoes	1. Magnum Opus 2. Omjer dobiti 3. Informacijska dobit
Cyyoung9302	1. Informacijska dobit 2. Magnum Opus 3. Omjer dobiti
impeach	1. Informacijska dobit 2. Magnum Opus 3. Omjer dobiti

Skup	Rang tehnika selekcije
fraud	1. Informacijska dobit 2. Magnum Opus 3. Omjer dobiti
Campus Climate 2011 SJU	1. Magnum Opus 2. Informacijska dobit 3. Omjer dobiti
homerun	1. Informacijska dobit 2. Magnum Opus 3. Relief
sonar	1. Magnum Opus 2. Relief 3. Informacijska dobit
bondrate	1. Informacijska dobit 2. Magnum Opus 3. Omjer dobiti
ICPSR 3009	1. Informacijska dobit 2. Magnum Opus 3. Relief
gsssexsurvey	1. Magnum Opus 2. Relief 3. Informacijska dobit
uktrainacc	1. Magnum Opus 2. Informacijska dobit 3. Omjer dobiti
ncaa	1. Informacijska dobit 2. Magnum Opus 3. Relief
credit	1. Magnum Opus 2. Informacijska dobit 3. Omjer dobiti
weights	1. Magnum Opus 2. Relief 3. Informacijska dobit
ICPSR 2743	1. Informacijska dobit 2. Magnum Opus 3. Relief
city	1. Magnum Opus 2. Informacijska dobit 3. Omjer dobiti
supreme	1. Magnum Opus 2. Relief 3. Informacijska dobit
ICPSR 2751	1. Informacijska dobit 2. Magnum Opus 3. Relief
blood-transfusion/	1. Magnum Opus 2. Relief 3. Informacijska dobit

Skup	Rang tehnika selekcije
authorship	1. Informacijska dobit 2. Magnum Opus 3. Omjer dobiti
ICPSR 2867	1. Informacijska dobit 2. Magnum Opus 3. Relief
ICPSR 2480	1. Informacijska dobit 2. Magnum Opus 3. Relief
halloffame	1. Magnum Opus 2. Informacijska dobit 3. Relief
CPS_85_Wages	1. Magnum Opus 2. Relief 3. Informacijska dobit
Physical+Activity+Monitoring	1. Magnum Opus 2. Relief 3. Informacijska dobit
marketing	1. Magnum Opus 2. Informacijska dobit 3. Relief
binge	1. Magnum Opus 2. Relief 3. Informacijska dobit
ionosphere	1. Magnum Opus 2. Relief 3. Informacijska dobit
ICPSR 2859	1. Magnum Opus 2. Informacijska dobit 3. Omjer dobiti
Mushroom	1. Magnum Opus 2. Relief 3. Informacijska dobit
ICPSR 2039	1. Informacijska dobit 2. Magnum Opus 3. Omjer dobiti
Thyroid+Disease	1. Magnum Opus 2. Omjer dobiti 3. Informacijska dobit
sick	1. Magnum Opus 2. Informacijska dobit 3. Omjer dobiti
One-hundred+plant+species+leaves+data+set	1. Informacijska dobit 2. Magnum Opus 3. Omjer dobiti
Kr-Vs-Kp	1. Magnum Opus 2. Relief 3. Informacijska dobit

Skup	Rang tehnika selekcije
tic-tac-toe	1. Magnum Opus 2. Informacijska dobit 3. Omjer dobiti
abgss98	1. Informacijska dobit 2. Magnum Opus 3. Omjer dobiti
ICPSR 2686	1. Informacijska dobit 2. Magnum Opus 3. Relief
ICPSR 2155	1. Informacijska dobit 2. Magnum Opus 3. Omjer dobiti
heart-statlog	1. Magnum Opus 2. Relief 3. Informacijska dobit
spambase	1. Magnum Opus 2. Relief 3. Informacijska dobit
marketing	1. Informacijska dobit 2. Magnum Opus 3. Omjer dobiti
Hill-Valley	1. Informacijska dobit 2. Magnum Opus 3. Relief
hepatitis	1. Informacijska dobit 2. Magnum Opus 3. Relief
ICPSR 4291	1. Magnum Opus 2. Relief 3. Omjer dobiti
ICPSR 4582	1. Magnum Opus 2. Informacijska dobit 3. Omjer dobiti
ICPSR 9595	1. Magnum Opus 2. Relief 3. Omjer dobiti
ICPSR 21600 2	1. Magnum Opus 2. Relief 3. Omjer dobiti
ICPSR 21600 3	1. Magnum Opus 2. Informacijska dobit 3. Omjer dobiti
ICPSR 28641 2	1. Magnum Opus 2. Relief 3. Omjer dobiti
ICPSR 6542	1. Magnum Opus 2. Informacijska dobit 3. Omjer dobiti

Skup	Rang tehnika selekcije
ICPSR 4367	1. Magnum Opus 2. Informacijska dobit 3. Relief
ICPSR 4572 02	1. Magnum Opus 2. Informacijska dobit 3. Omjer dobiti
ICPSR 21600 4	1. Magnum Opus 2. Relief 3. Omjer dobiti
DBWorld+e-mails	1. Magnum Opus 2. Informacijska dobit 3. Omjer dobiti
ICPSR 6135	1. Magnum Opus 2. Relief 3. Omjer dobiti
ICPSR 4537 8th form 1	1. Magnum Opus 2. Omjer dobiti 3. Informacijska dobit
ICPSR 4275	1. Magnum Opus 2. Relief 3. Omjer dobiti
GLI-85	1. Magnum Opus 2. Informacijska dobit 3. Omjer dobiti
ICPSR 21600 1	1. Magnum Opus 2. Informacijska dobit 3. Omjer dobiti
ICPSR 4566 02	1. Magnum Opus 2. Omjer dobiti 3. Informacijska dobit
ICPSR 8255	1. Magnum Opus 2. Relief 3. Omjer dobiti
ICPSR 28641	1. Magnum Opus 2. Omjer dobiti 3. Informacijska dobit
SMK-CAN-187	1. Magnum Opus 2. Informacijska dobit 3. Omjer dobiti
ICPSR 23041 2	1. Magnum Opus 2. Relief 3. Omjer dobiti
ICPSR 6480	1. Magnum Opus 2. Omjer dobiti 3. Informacijska dobit
ICPSR 4537 10th form 2	1. Magnum Opus 2. Relief 3. Omjer dobiti

Skup	Rang tehnika selekcije
ICPSR 4138	1. Magnum Opus 2. Informacijska dobit 3. Omjer dobiti
ICPSR 23041	1. Magnum Opus 2. Omjer dobiti 3. Informacijska dobit
ICPSR 4690	1. Magnum Opus 2. Informacijska dobit 3. Omjer dobiti
ICPSR 4372	1. Magnum Opus 2. Relief 3. Omjer dobiti
ICPSR 20022	1. Magnum Opus 2. Informacijska dobit 3. Omjer dobiti
ICPSR 6484	1. Magnum Opus 2. Omjer dobiti 3. Informacijska dobit
ICPSR 4566 01	1. Magnum Opus 2. Informacijska dobit 3. Omjer dobiti
ICPSR 6693	1. Magnum Opus 2. Informacijska dobit 3. Omjer dobiti
ICPSR 4572 01	1. Magnum Opus 2. Omjer dobiti 3. Informacijska dobit
Dorothea	1. Magnum Opus 2. Informacijska dobit 3. Omjer dobiti
Human+Activity+Recognition+Using+Smartphones	1. Magnum Opus 2. Omjer dobiti 3. Informacijska dobit
ICPSR 31221	1. Magnum Opus 2. Relief 3. Omjer dobiti
ICPSR 3669	1. Magnum Opus 2. Omjer dobiti 3. Informacijska dobit
ICPSR 2743 Person Level Data	1. Magnum Opus 2. Informacijska dobit 3. Omjer dobiti
ICPSR 2258	1. Magnum Opus 2. Omjer dobiti 3. Informacijska dobit
Madelon	1. Magnum Opus 2. Omjer dobiti 3. Informacijska dobit

Skup	Rang tehnika selekcije
adult	1. Magnum Opus 2. Informacijska dobit 3. Omjer dobiti
ICPSR 31202 5	1. Magnum Opus 2. Informacijska dobit 3. Omjer dobiti
ICPSR 2857	1. Magnum Opus 2. Omjer dobiti 3. Informacijska dobit
ICPSR 2346	1. Magnum Opus 2. Informacijska dobit 3. Omjer dobiti
PEMS-SF	1. Magnum Opus 2. Omjer dobiti 3. Informacijska dobit
Dexter	1. Magnum Opus 2. Informacijska dobit 3. Omjer dobiti
ICPSR 2686 Caregiver Data	1. Magnum Opus 2. Omjer dobiti 3. Informacijska dobit
ICPSR 3534	1. Magnum Opus 2. Omjer dobiti 3. Informacijska dobit
ICPSR 2535	1. Magnum Opus 2. Informacijska dobit 3. Omjer dobiti
ICPSR 2149	1. Omjer dobiti 2. Relief 3. Magnum Opus
Semeion+Handwritten+Digit	1. Relief 2. Informacijska dobit 3. Omjer dobiti
ICPSR 3548	1. Magnum Opus 2. Informacijska dobit 3. Omjer dobiti
Gisette	1. Omjer dobiti 2. Magnum Opus 3. Relief
ICPSR 2295	1. Magnum Opus 2. Informacijska dobit 3. Omjer dobiti
ICPSR 2743	1. Relief 2. Informacijska dobit 3. Magnum Opus
ICPSR 2163	1. Magnum Opus 2. Relief 3. Omjer dobiti

Skup	Rang tehnika selekcije
SECOM	1. Magnum Opus 2. Informacijska dobit 3. Relief
ICPSR 3789	1. Omjer dobiti 2. Relief 3. Magnum Opus
ICPSR 2833	1. Magnum Opus 2. Relief 3. Omjer dobiti
Arcene	1. Magnum Opus 2. Informacijska dobit 3. Omjer dobiti
ICPSR 2566	1. Magnum Opus 2. Relief 3. Informacijska dobit
ICPSR 31202 4	1. Magnum Opus 2. Informacijska dobit 3. Omjer dobiti
ICPSR 2039 2	1. Relief 2. Magnum Opus 3. Informacijska dobit
ICPSR 3151	1. Magnum Opus 2. Omjer dobiti 3. Informacijska dobit
ICPSR 6627	1. Magnum Opus 2. Relief 3. Informacijska dobit

Od 128 skupova podataka, u 39,06% slučajeva tehnike otkrivanja kontrasta su dalje lošije rezultate (duže vrijeme izvođenja) od ostalih ili nisu statistčki značajno bolje od ostalih.

- Na 20 skupova podataka Informacijska dobit najbrže provede selekciju atributa
- Na 9 skupova podataka Omjer dobiti je najbrže proveo selekciju atributa
- Na 3 skupa podataka Relief je najbrže proveo selekciju atributa
- Na 18 skupova podataka Magnum Opus je bio najbrži, ali razlika u brzini u odnosu na ostale tehnike nije statistični značajna

Analiza karakteristika skupova podataka na kojima tehnike otkrivanja kontrasta nisu bile najbrže u selekciji atributa pokazuje daje prosječnu vrijednost broja atributa od 259. To je znatno manje od prosječnog broja atributa svih 128 skupova što vodi do zaključka da tehnike

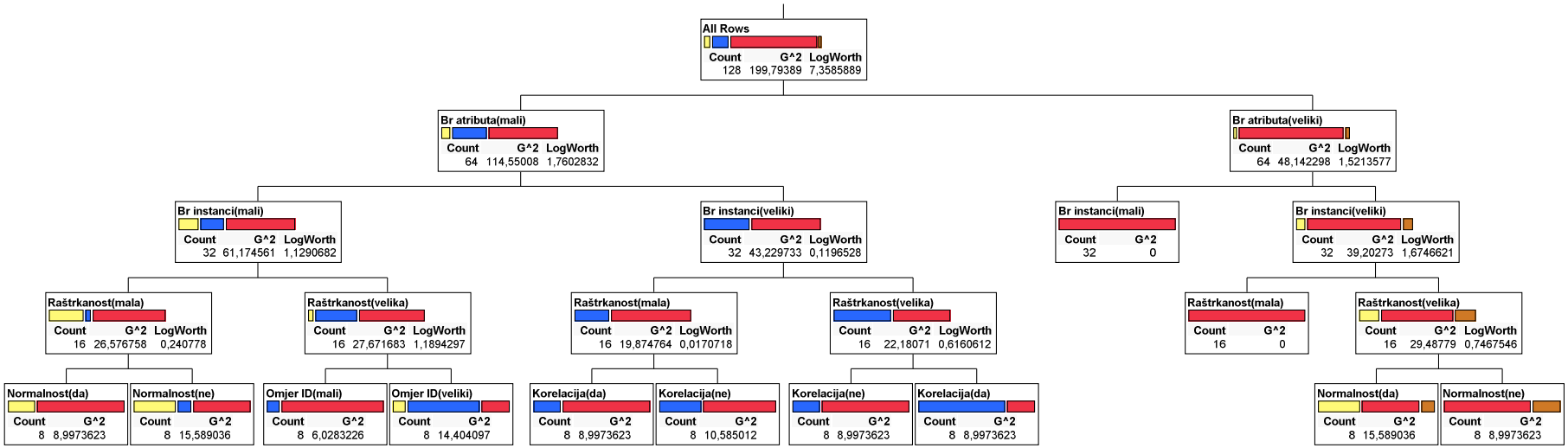
otkrivanja kontrasta rade selekciju atributa brže od ostalih tehnika na skupovima podataka s velikim brojem atributa. Ovi rezultati imaju posebnu težinu u današnje vrijeme kada govorimo o „velikim podacima“ (eng. *big data).* Prosječan broj instanci (2902) skupova podataka na kojima tehnike otkrivanja kontrasta nisu bile najbrže je blizak prosjeku svih 128 skupova podataka (2105). Isti rezultati dobiveni su i analizom ostalih karakteristika podataka. Prosječna korelacija, omjer unutarnje dimenzionalnosti i šum atributa skupova podataka na kojima tehnike otkrivanja kontrasta nisu bile najbrže ne odstupa znatno od prosjeka svih skupova (tablica 15).

Tablica 15. Povezanost karakteristika skupa podataka i vremena provođenja selekcije atributa

	Svi skupovi (128)	Skupovi na kojima tehnike otkrivanja kontrasta nisu najbrže (32)
Broj atributa	1264	258
Broj instanci	2105	2902
Korelacija	0.245	0.255
Omjer ID	0.623	0.689
Šum atributa	0.377	0.310

S ciljem identificiranja povezanosti karakteristika podataka i performansi tehnika selekcije atributa u smislu vremena potrebnog da se provede selekcija, napravljeno je stablo odlučivanja prikazano na slici 36. Stablo grafički prikazuje karakteristike skupova na kojim tehnike otkrivanja kontrasta najbrže provode selekciju atributa i identificira najvažnije karakteristike podataka za ovaj aspekt evaluacije tehnika.

Slika 36. Ovisnost brzine izvođenja selekcije atributa o karakteristikama skupa podataka



Legenda:

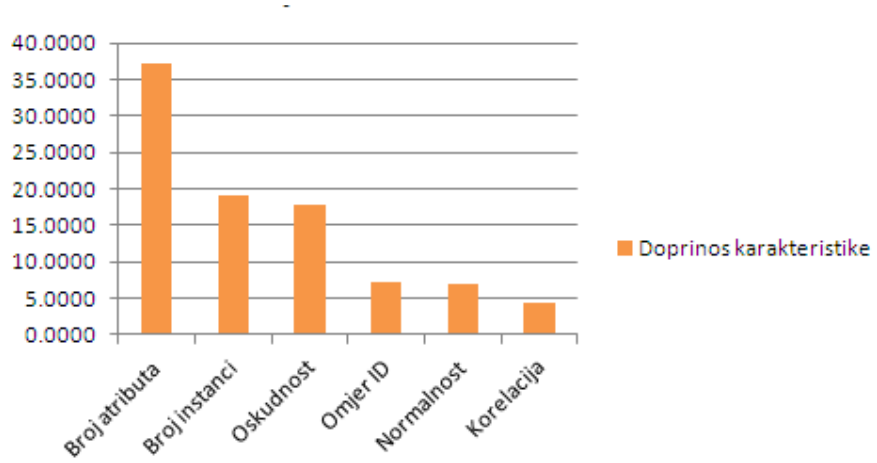
	Informacijska dobit
	Omjer dobiti
	Linearni odabir unaprijed
	Relief
	Tehnika otkrivanja kontrasta (STUCCO, Magnum Opus)

Informacijska dobit najbrža je tehnika na 15% analiziranih skupova (20 skupova). Karakteristika svih skupova na kojima je informacijska dobit najbrže rpovela selekciju atributa je mali broj atributa. Omjer dobiti je najbrža tehnika na 9 skupova. Karakteristika 6 od tih 9 skupova je mali broj atributa i mali broj instanci, dok je karakteristika preostala 3 skupa podataka veliki broj atributa i veliki broj instanci. Relief algoritam najbrže je proveo selekciju atributa na 3 skupa podataka. Karakteristike ta 3 skupa su velik broj atributa, velik broj instanci, veliki oskudnost i veliki omjer unutarnje dimenzionalnosti.

Važno je za uočiti da je od 32 skupa podataka na kojima tehnike otkrivanja kontrasta nisu najbrže provele selekciju atributa samo 6 njih ima veliki broj atributa. Jasno je da tehnike otkrivanja kontrasta dobro rade na skupovima podataka s velikim brojem atributa. Ovaj zaključak ima posebnu težinu kada se zna da količina dostupnih podataka svakim danom raste.

Analiza utjecaja karakteristika skupa podataka na brzinu izvođenja tehnika selekcije atributa dana je na slici 37.

Slika 37. Doprinos karakteristika skupa podataka klasifikaciji tehnika s obzirom na vrijeme provođenja selekcije atributa



Karakteristika skupa podataka koja najviše doprinosi razvrstavanju tehnika selekcija atributa s obzirom na vrijeme koje im je potrebno da provedu selekciju atributa je broj atributa. Kako je vidljivo na grafu, utjecaj karakteristike broj atributa je daleko najveći u usporedbi s ostalim. Slijede je dvije karakteristike koje imaju podjednak doprinos: broj instanci i

oskudnost. Jedina karakteristika skupa podataka koja se ne pokazuje relevantnom za brzinu izvođenja selekcije atributa je šum atributa.

Kao zaključak podpoglavlja 8.4. navodi se slijedeće: tehnike otkrivanja kontrasta najbrže provode selekciju atributa za sve kombinacije karakteristika skupa podataka, osim za slijedeće:

AKO skup podataka ima karakteristike

Broj atributa: mali Broj instanci: mali Oskudnost: mala Korelacija: ne Normalnost: da Homogenost: ne Omjer ID: veliki Šum atributa: mali

TADA koristiti algoritam *Omjer dobiti* .

AKO skup podataka ima karakteristike

Broj atributa: mali Broj instanci: mali Oskudnost: mala Korelacija: ne Normalnost: ne Homogenost: da Omjer ID: mali Šum atributa: veliki

TADA koristiti algoritam *Omjer dobiti* .

AKO skup podataka ima karakteristike

Broj atributa: mali Broj instanci: mali Oskudnost: mala Korelacija: ne Normalnost: ne Homogenost: ne Omjer ID: veliki Šum atributa: mali

TADA koristiti algoritam *Omjer dobiti* .

AKO skup podataka ima karakteristike

Broj atributa: mali Broj instanci: mali Oskudnost: mala Korelacija: da Normalnost: da Homogenost: ne Omjer ID: mali Šum atributa: veliki

TADA koristiti algoritam *Omjer dobiti* .

AKO skup podataka ima karakteristike

Broj atributa: mali Broj instanci: mali Oskudnost: mala Korelacija: da Normalnost: ne Homogenost: ne Omjer ID: mali Šum atributa: veliki

TADA koristiti algoritam *Informacijska dobit* .

AKO skup podataka ima karakteristike

Broj atributa: mali Broj instanci: mali Oskudnost: mala Korelacija: da Normalnost: ne Homogenost: da Omjer ID: veliki Šum atributa: mali

TADA koristiti algoritam *Omjer dobiti* .

AKO skup podataka ima karakteristike

Broj atributa: mali Broj instanci: mali Oskudnost: velika Korelacija: ne Normalnost: da Homogenost: ne Omjer ID: veliki Šum atributa: mali

TADA koristiti algoritam *Omjer dobiti* .

AKO skup podataka ima karakteristike

Broj atributa: mali Broj instanci: mali Oskudnost: velika Korelacija: ne Normalnost: ne Homogenost: da Omjer ID: veliki Šum atributa: mali

TADA koristiti algoritam *Informacijska dobit* .

AKO skup podataka ima karakteristike

Broj atributa: mali Broj instanci: mali Oskudnost: velika Korelacija: ne Normalnost: ne Homogenost: ne Omjer ID: mali Šum atributa: veliki

TADA koristiti algoritam *Informacijska dobit* .

AKO skup podataka ima karakteristike

Broj atributa: mali Broj instanci: mali Oskudnost: velika Korelacija: ne Normalnost: ne Homogenost: ne Omjer ID: veliki Šum atributa: mali

TADA koristiti algoritam *Informacijska dobit* .

AKO skup podataka ima karakteristike

Broj atributa: mali Broj instanci: mali Oskudnost: velika Korelacija: da Normalnost: da Homogenost: da Omjer ID: veliki Šum atributa: mali

TADA koristiti algoritam *Informacijska dobit* .

AKO skup podataka ima karakteristike

Broj atributa: mali Broj instanci: mali Oskudnost: velika Korelacija: da Normalnost: da Homogenost: ne Omjer ID: veliki Šum atributa: mali

TADA koristiti algoritam *Informacijska dobit* .

AKO skup podataka ima karakteristike

Broj atributa: mali Broj instanci: mali Oskudnost: velika Korelacija: da Normalnost: ne Homogenost: da Omjer ID: veliki Šum atributa: mali

TADA koristiti algoritam *Informacijska dobit*

AKO skup podataka ima karakteristike

Broj atributa: mali Broj instanci: veliki Oskudnost: mala Korelacija: ne Normalnost: da Homogenost: ne Omjer ID: mali Šum atributa: veliki

TADA koristiti algoritam *Informacijska dobit*

AKO skup podataka ima karakteristike

Broj atributa: mali Broj instanci: veliki Oskudnost: mala Korelacija: ne Normalnost: ne Homogenost: da Omjer ID: veliki Šum atributa: mali

TADA koristiti algoritam *Informacijska dobit*

AKO skup podataka ima karakteristike

Broj atributa: mali Broj instanci: veliki Oskudnost: mala Korelacija: ne Normalnost: ne Homogenost: ne Omjer ID: veliki Šum atributa: mali

TADA koristiti algoritam *Informacijska dobit*

AKO skup podataka ima karakteristike

Broj atributa: mali Broj instanci: veliki Oskudnost: mala Korelacija: da Normalnost: da Homogenost: da Omjer ID: mali Šum atributa: veliki

TADA koristiti algoritam *Informacijska dobit*

AKO skup podataka ima karakteristike

Broj atributa: mali Broj instanci: veliki Oskudnost: mala Korelacija: da Normalnost: da Homogenost: da Omjer ID: veliki Šum atributa: mali

TADA koristiti algoritam *Informacijska dobit*

AKO skup podataka ima karakteristike

Broj atributa: mali Broj instanci: veliki Oskudnost: velika Korelacija: ne Normalnost: da Homogenost: ne Omjer ID: mali Šum atributa: veliki

TADA koristiti algoritam *Informacijska dobit*

AKO skup podataka ima karakteristike

Broj atributa: mali Broj instanci: veliki Oskudnost: velika Korelacija: ne Normalnost: ne Homogenost: da Omjer ID: veliki Šum atributa: mali

TADA koristiti algoritam *Informacijska dobit*

AKO skup podataka ima karakteristike

Broj atributa: mali Broj instanci: veliki Oskudnost: velika Korelacija: da Normalnost: da Homogenost: da Omjer ID: mali Šum atributa: veliki

TADA koristiti algoritam *Informacijska dobit*

AKO skup podataka ima karakteristike

Broj atributa: mali Broj instanci: veliki Oskudnost: velika Korelacija: da Normalnost: da Homogenost: da Omjer ID: veliki Šum atributa: mali

TADA koristiti algoritam *Informacijska dobit*

AKO skup podataka ima karakteristike

Broj atributa: mali Broj instanci: veliki Oskudnost: velika Korelacija: da Normalnost: da Homogenost: ne Omjer ID: mali Šum atributa: veliki

TADA koristiti algoritam *Informacijska dobit*

AKO skup podataka ima karakteristike

Broj atributa: mali Broj instanci: veliki Oskudnost: velika Korelacija: da Normalnost: ne Homogenost: ne Omjer ID: veliki Šum atributa: mali

TADA koristiti algoritam *Informacijska dobit*

AKO skup podataka ima karakteristike

Broj atributa: mali Broj instanci: veliki Oskudnost: velika Korelacija: da Normalnost: ne Homogenost: da Omjer ID: mali Šum atributa: veliki

TADA koristiti algoritam *Informacijska dobit*

AKO skup podataka ima karakteristike

Broj atributa: mali Broj instanci: veliki Oskudnost: velika Korelacija: da Normalnost: ne Homogenost: da Omjer ID: veliki Šum atributa: mali

TADA koristiti algoritam *Informacijska dobit*

AKO skup podataka ima karakteristike

Broj atributa: veliki Broj instanci: veliki Oskudnost: velika Korelacija: ne Normalnost: da Homogenost: da Omjer ID: mali Šum atributa: veliki

TADA koristiti algoritam *Omjer dobiti*

AKO skup podataka ima karakteristike

Broj atributa: veliki Broj instanci: veliki Oskudnost: velika Korelacija: ne Normalnost: da Homogenost: da Omjer ID: veliki Šum atributa: mali

TADA koristiti algoritam *Relief*

AKO skup podataka ima karakteristike

Broj atributa: veliki Broj instanci: veliki Oskudnost: velika Korelacija: ne Normalnost: da Homogenost: ne Omjer ID: veliki Šum atributa: mali

TADA koristiti algoritam *Omjer dobiti*

AKO skup podataka ima karakteristike

Broj atributa: veliki Broj instanci: veliki Oskudnost: velika Korelacija: ne Normalnost: ne Homogenost: da Omjer ID: veliki Šum atributa: mali

TADA koristiti algoritam *Relief*

AKO skup podataka ima karakteristike

Broj atributa: veliki Broj instanci: veliki Oskudnost: velika Korelacija: da Normalnost: da Homogenost: da Omjer ID: mali Šum atributa: veliki

TADA koristiti algoritam *Omjer dobiti*

AKO skup podataka ima karakteristike

Broj atributa: veliki Broj instanci: veliki Oskudnost: velika Korelacija: da Normalnost: ne Homogenost: ne Omjer ID: veliki Šum atributa: mali

TADA koristiti algoritam *Relief*

8.5. Ograničenja i preporuke za buduća istraživanja

Ovaj rad definira i implementira nove tehnike za selekciju atributa, tehnike otkrivanja kontrasta (SfFS i MOFS). Velik broj eksperimenata je proveden (cca 2000 analiza) s ciljem komparacije predloženih tehnika s postojećem tehnikama selekcije atributa iz perspektive dva kriterija: točnost klasifikacije i vrijeme provođenja selekcije atributa. Pritom je poseban naglasak stavljen na karakteristike skupova podataka i identifikaciju karakteristika skupa podataka na kojima su tehnike otkrivanja kontrasta točnije i/ili brže u odnosu na postojeće tehnike.

Cilj tehnika otkrivanja kontrasta je kvantificirati i opisati razlike između dvije grupe. Umjesto da se grupe uspoređuju direktno, pristup tehnika otkrivanja kontrasta prvo uči zakonitosti na jednoj grupi, zatim na drugoj, te ih uspoređuje. Prednost ovog pristupa je da se kompleksnost skupa podataka reducira, dok je s druge strane, sačuvana informacija sadržana u cijelom skupu. Iz tog razloga tehnike otkrivanja kontrasta dobro rade u slučajevima velikog broja atributa i velikog broja instanci te velike raštkranosti.

Kao takvo istraživanje daje slijedeće doprinose u području rudarenja podataka i selekcije atributa:

- Predlaže inovativne tehnike za primjenu u svrhu selekcije atributa, tehnike otkrivanja kontrasta (nazvane SfFS i MOFS). Kroz empirijsko istraživanje se dokazuje da ove tehnike brzo odabiru točne podskupove s relevantnim atributima.

- U provedenom istraživanju upotrijebilo se sedam tehnika selekcije atributa i dva klasifikatora da se istraži utjecaj karakteristika skupa podataka na proces selekcije atributa i točnost klasifikacije. Rezultati su pokazali da odabir tehnike selekcije atributa ovisi o karakteristikama skupa podataka na kojem s primjenjuje kao i o odabiru klasifikatora. Kao takvi, rezultati istraživanja vode do zaključka koje su karakteristike skupa podataka važne za proces klasifikacije.

- Istraživanje je izrazito opsežno u pogledu broja korištenih skupova podataka (128), karakteristika skupa podataka (7), tehnika selekcije atributa (7) i klasifikatora (2).

Unatoč nabrojanim prednostima, postoje neka ograničenja koja treba uzeti u obzir prilikom interpretacije rezultata ovog istraživanja. Kroz diskusiju ograničenja istraživanja daju se i smjernice za buduća istraživanja. Prvo, tehnike otkrivanja kontrasta u selekciji atributa su definirane s pretpostavkom nezavisnosti atributa. Iako taj pristup ima ranije nabrojane prednosti s jedne strane, predstavlja ograničenje s druge strane jer su vrlo često atributi u međusobnoj interakciji.

Tehnike su evaluirane samo na skupovima podataka s dvije klase. U budućim istraživanjim to se može proširiti radeći evaluaciju na skupovima podataka s više klasa. U empirijskom istraživanju uzeto je u obzir sedam karakteristika podataka. Iako to predstavlja znatno opsežnije istraživanje u odnosu na prethodna, najnovija istraživanja prepoznaju juš neke karakteristike podataka koje bi trebalo razmotriti u budućim istraživanjima, a to su: neravnoteža vrijednosti zavisnog atributa (eng. *class imbalance*), kojeg identificiraju npr. Longadge i suradnici (Longadge Dongre i Malik, 2013.) te promjena skupa podataka (eng. *dataset shift)* koju ističu Moreno-Torres i suradnici (Moreno-Torres et. al., 2012.). Nadalje, s ciljem smanjenja prostora traženja napravljena je diskretizacija karakteristika skupa podataka. Postavlja se pitanje da li bi rezultati bili drugačiji da je diskretizacija napravljena drugačije?

Još jedan aspekt koji je, zbog kompleksnosti istraživanja, izostavljen u ovoj disertaciji je domena iz koje dolazi skup podataka. Vezano na to, nisu ispitani troškovi krive klasifikacije.

U ovom radu tehnike otkrivanja kontrasta primijenjene su u selekciji atributa u svrhu klasifikacije i dale su dobre rezultate. To predstavlja vrlo solidno polazište za daljnja istraživanja u kojima se tehnike otkrivanja kontrasta mogu primijeniti u svrhu regresije, a posebno selekcije atributa u rudarenja teksta, gdje se slična tehnika *Odds Ratio*, već uspješno primjenuje.

U ovom istraživanju tehnike selekcije atributa evaluirane su s obzirom na dva kriterija: točnost klasifikacije i vrijeme provođenja selekcije. Rezultati su pokazali da je jedna tehnika otkrivanja kontrasta većinom točnija u selekciji atributa (STUCCO), dok je druga većinom brža (Magnum Opus). U budućim istraživanjima može se postaviti pitanje da li je moguće npr. optimirati STUCCO algoritma da radi brže, a da pritom ne izgubi na točnosti.

9. ZAKLJUČAK

U zaključnom poglavlju rada diskutira se realizacija postavljenih ciljeva istraživanja i potvrđuju se hipoteze istraživanja.

Glavni cilj istraživanja bio je primijeniti tehnike otkrivanja kontrasta, STUCCO i Magnum Opus, u selekciji atributa i identificirati za koje karakteristike skupa podataka primjena tih tehnika poboljšava točnost i skraćuje vrijeme klasifikacije u odnosu na dosad najčešće korištene tehnike selekcije atributa.

Kako bi se ostvario glavni cilj provedeno je nekoliko aktivnosti koje su dovele do realizacije slijedećih podciljeva istraživanja:

1. Analizirane su postojeće tehnike selekcije atributa i utvrđeno je koje su tehnike najčešće korištene u prethodnim istraživanjima
    - U sklopu provedenog istraživanja izvršena je sistematizacija tehnika selekcije atributa. Svrha sistematizacija bila je analiza postojećeg stanja te utvrđivanje koje su se tehnike u prethodnim istraživanjima najviše koristile. Ovaj dio predstavljen je u trećem poglavlju.

2. Dan je pregled područja otkrivanja kontrasta: definirani su ključni pojmovi i identificirane glavne karakteristike tehnika otkrivanja kontrasta
    - U četvrtom poglavlju opisane su tehnike otkrivanja kontrasta, posebno STUCCO i Magnum Opus. Metode su dosad primjenjivanje u nekoliko područja što je prikazano i objašanjeno u poglavlju 4.3. Temeljni pojmovi ovog područja su objašnjeni i prvi put prevedni na hrvatski jezik.
3. Definirane su tehnike STUCCO i Magnum Opus kao tehnike selekcije atributa
    - U šestom poglavlju ostvaren je jedan od glavnih ciljeva rada: tehnike otkrivanja kontrasta su prilagođene zadaći selekcije atributa te je defniran postupak provedbe selekcije atributa primjenom tehnika STUCCO i Magnum Opus. Tehnike otkrivanja kontrasta definirane su za selekciju atributa jer je njihova temeljna zadaća razumijevanje razlika između grupa i karakterizira ih činjenica da smanjuju kompleksnost podataka, a da pritom čuvaju većina

informacija iz originalnog skupa podataka, što je i ideja selekcije atributa. Tehnike su definirane s pretpostavkom nezavisnosti atributa.

4. Primjenjene su tehnike STUCCO i Magnum Opus u selekciji atributa na 128 skupova podataka koji se razlikuju u karakteristikama
    - U petom poglavlju opisane su karakteristike skupa podataka važne za klasifikaciju, a identificirane od strane Van der Walta. U nekoliko baza podataka pronađeno je 128 skupova podataka koji se razlikuju u karakteristikama i nad njima je provedena selekcija atributa primjenom tehnika otkrivanja kontrasta (postupak je opisan u poglavlju 7),
5. Uspoređene su tehnike STUCCO i Magnum Opus s dosad najčešće korištenim tehnikama selekcije atributa na način da su atributi dobiveni selekcijom svakom od tehnika primjenjeni na dva algoritma učenja i uspoređena je točnost klasifikacije i brzina izvođenja selekcije atributa
    - Nad istim skupovima podataka provedena je selekcija atributa primjenom još pet tehnika (Informacijska dobit, Omjer dobiti, LFS, Relief i Voting). Vrijeme provođenja selekcije atributa i točnost klasifikacije primjenom atributa selektiranih tehnikama otkrivanja kontrasta i ostalih tehnika su komparirane i testirana je statistička značajnost razlike u vremenu provođenja i točnosti klasifikacije (opisano u poglavlju 8).
6. Identificirano je za koje karakteristika skupa podataka tehnike otkrivanja kontrasta kao tehnike za selekciju atributa daju bolje rezultate od dosad najčešće korištenih tehnika selekcije atributa
    - Vrednovanje rezultata je provedeno primjenom Friedman testa kojim su rangirane tehnike selekcije atributa i kojim je utvrđeno da li postoji statistički signifikantna razlika u razlici točnosti klasifikacije i vremena provođenja selekcije atributa.Rezultati su pokazali da za veliku većinu kombinacija karakteristika podataka tehnike otkrivanje kontrasta brže i točnije provode selekciju atributa. Kroz analizu ovih rezultata došlo se i do zaključka kako pojedine tehnike djeluju na kojim karakteristikama skupova podataka te koje su karakteristike skupa podataka važne za zadaću klasifikacije i cjelokupni proces otkrivanja znanja u poda

Odgovor na hipoteze istraživanja

H1: Tehnike otkrivanja kontrasta za određene karakteristike podataka brže provode selekciju atributa od dosad šire korištenih tehnika selekcije atributa.

Od 128 skupova, na 60,94% karakteristika skupova podataka tehnike otkrivanja kontrasta su provele statistički značajno brže selekciju atributa od ostalih korištenih tehnika selekcije atributa. Karakteristike skupova na kojima su tehnike otkrivanja kontrasta brže u selekciji atributa su velik broj atributa i velik broj instanci.

- ***Potvrđuje se prva hipoteza istraživanja***

H2: Primjenom otkrivanja kontrasta u selekciji atributa za određene karakteristike podataka postiže se točnija klasifikacija nego dosad šire korištenim tehnikama selekcije atributa.

Klasifikacija je provedena primjenom neuronskih mreža i diskriminacijske analize. U klasifikaciji neuronskim mrežama na 82,03% kombinacija karakteristika skupova tehnike otkrivanja kontrasta u selekciji atributa su rezultirale statistički značajnijom točnijom klasifikacijom u odnosu na ostale tehnike selekcije atributa. Karakteristike tih skupova su veći broj atributa i manji broj instanci.

U klasifikaciji diskriminacijskom analizom na 78,12% kombinacija karakteristika podataka tehnike otkrivanja kontrasta dale su točniju klasifikaciju od ostalih korištenih tehnika. Karakteristike tih skupova podataka su normalnost skupa podataka i homogenost kovarijanci klasa. Tehnike otkrivanja kontrasta rezultiraju točnijom klasifikacijom gotovo na svim skupovima podataka s velikim brojem instanci.

- ***Potvrđuje se druga hipoteza istraživanja***

LITERATURA

1. Abe, N., Kudo, M., Toyama, J., Shimbo, M., Classifier-independent feature selection on the basis of divergence criterion, Pattern Analysis & Applications, 9 (2), 2006., str., 127.-137.
2. Ali, K. M., Pazzani, J., Error reduction through learning multiple descriptions. Machine Learning, 24, 1996., str. 173-202.
3. Alibeigi, M., Hashemi, S., Hamzeh, A., Unsupervised Feature Selection Using Feature Density Functions, International Journal of Electrical and Electronics Engineering 3:7, 2009., str. 394-399.
4. Anand, D., Bharadwaj, K.K., Utilizing various sparsity measures for enhancing accuracy of collaborative recommender systems based on local and global similarities, Expert Systems with Applications, 38 (5), 2011., str. 5101-5109.
5. Arauzo-Azofra, A., Aznarte, J.L., Benitez, J.M., Emprical study of feature selection methods based on individual feature evaluation for classification problems, Expert systems with applications, 38, 2011., str. 8170-8177.
6. Arauzo-Azofra, A., Benitez, J.M., Castro, J.L., A feature set measure based on Relief, RASC, 2004., str.104 – 109.
7. Azevedo A.; Santos M. F., KDD, SEMMA and CRISP – DM : a parallel overview. *In Proceedings of the IADIS European Conference on Data Mining 2008,* 2008., str. 182-185.
8. Bay, S.D., Pazzani, M.J., Detecting change in categorical data: Mining contrast sets, KDD 1999., Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York: ACM, 1999., str. 302-306.
9. Blum, A.L., Langley, P., Selection of relevant features and examples in machine learning, Artificial Intelligence, 97 (1-2), 1997., str. 245-271.
10. Bonev, B.I., Feature selection based on information theory, Phd Thesis, University of Alicante, 2010.
11. Boettcher, M., Contrast and change mining, John Wiley & Sons, Inc. WIREs Data Mining Knowl Discov May/June 2011 Volume 1, str. 215–230 DOI: 10.1002/widm.27
12. Brazdil, P., Gama, J., Henery, B., Characterizing the applicability of classification

algorithms using meta-level learning, In Lecture Notes in Artificial Intelligence, European Conference on Machine Learning, Catania, Italy, April 1994., Proceeding, str.83-102.

13. Breiman, L., Bagging predictors, Machine Learning, 24, 1996., 123-140.
14. Breiman, L. (Ed.),1998., Classification and regression trees. Chapman & Hall.
15. Cadenas, J.M., Garrido, C.M., Martinez, R., Feature subset selection Filter-Wrapper based on low quality data, Expert systems with applications, 40, 2013, str. 6241-6252.
16. Chan, P., Stolfo, S., Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection. Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining. AAAI, 1998., str. 121-125.
17. Chawla, N., Moore, T. E., Bowyer, K. W., Hall, L. O., Springer, C., Kegelmeyer, P., Bagging is a small-data-set phenomenon, In International Conference on Computer Vision and Pattern Recognition (CVPR), 2001., str. 684-689.
18. Conley, D., Get fuzzy, 2001.
19. Cover, T. M., & Thomas, J. A. (1991). Elements of information theory. Wiley Interscience.
20. Chrysostomou, K.A., The Role of Classifiers in Feature Selection: Number vs Nature, School of Information Systems, Computing and Mathematics Brune University, Doctoral thesis, 2008.
21. Chrysostomou, K., Wrapper Feature Selection. In J. Wang (Ed.), Encyclopedia of Data Warehousing and Mining, Second Edition, 2009, str. 2103-2108., Hershey, PA: Information Science Reference. doi:10.4018/978-1-60566-010-3.ch322
22. Čehovin, L., Bosnić, Z., Empirical evaluation of feature selection methods in classification, Intelligent data analysis, 14, 2010., str. 265-281.
23. Dash, M., Liu,H., Feature Selection for Classification, An International Journal of Intelligent Data Analysis, vol. 1, no. 3,1. 1997., str.131-156.
24. Dietterich, T. G., Machine Learning Research: Four Current Directions, *AI Magazine* 18(4), 1997., str. 97–136.
25. Demšar, J., Statistical Comparisons of Classifiers over Multiple Data Sets, Journal of Machine Learning Research, 7, 2006., str. 1–30.
26. Dong, G., Bailey, J., Contrast data mining: Concepts, algorithms and applications, CRC Press. Taylor & Francis Group, 2013.

27. Drugan, M.D., Wiering, M.,A, Feature selection for Bayesian network classifiers using the MDL-FS score, Internation journal of approximate reasoning, 51, 2010, str. 695-717.
28. De Veaux, R., Predictive Analytics:Modeling the World, OR/MS Seminar, 2005., dostupno na: student.som.umass.edu/informs/slides/Predictive.pdf, zadnji pristup: 15.07.2013.
29. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., From data mining to knowledge discovery in databases, AI magazine, 17 (3), 1996.., str. 37-54.
30. *Feature selection datasets at Arizona State University*, dostupno na: / http://featureselection.asu.edu/datasets.php, pristupano: 20.01.2013.
31. Gamberger, D., Lavrač, N., Dzeroski, S., Noise detection and elimination in data preprocessing: experiments in medical domains, Applied artificial intelligence, 14, 2000., str. 205-223.
32. Ganchev, T., Zervas, P., Fakotakis, N., Kokkinakis, G., Benchmarking feature selection techniques on the speaker verification task, Proc. of the Fifth International Symposium on Communication Systems, Networks and Digital Signal Processing, CSNDSP'06, 2006., str. 123-128.
33. Garson, G. David, 2008. "Discriminant Function Analysis", from Statnotes: Topics in Multivariate Analysis, http://www2.chass.ncsu.edu/garson/pa765/statnote.htm, pristupano [04.04.2013.]
34. Geng, X., Liu, T.Y., Qin, T., Li, h., Feature selection for ranking, SIGIR:Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, 2007., str. 407- 414.
35. Gutkin, M., Feature selection methods for classification of gene expression profiles, Master thesis, School of Computer Science, Tel-Aviv University, 2008.
36. Guyon, I., Elisseeff, A., An introduction to variable and feature selection. Journal of Machine Learning Research, 3, str. 1157–1182, 2003.
37. Gütlein, M., Frank, E., Hall, M., Karwath, A., Large-scale attribute selection using wrappers, CIDM 2009: 332-339
38. Hand, D.J., Manila, H., Smyth, P., Principles of data mining, Cambridge, MA:MIT Press, 2001.

39. Hall, L. O., Bowyer, K. W., Kegelmeyer, P., Moore, T. E., Chao, C., Distributed learning on very large data sets. Proceedings of the Sixth International ACM SIGKDD, 2000., str. 141-147.
40. Hall, M.A., Holmes, G., Benchmarking attribute selection techniques for discrete class data mining, IEEE transactions on knowledge and data engineering, 15 (3), 2003., str. 1-16.
41. Hand, D. J. 1981. Discrimination and Classification.Chichester, U.K.: Wiley.
42. Haury A-C, Gestraud P, Vert J-P., The Influence of Feature Selection Methods on Accuracy, Stability and Interpretability of Molecular Signatures. PLoS ONE 6(12): e28210. doi:10.1371/journal.pone.0028210, 2011., str. 1-16.
43. Heaton, J., Introduction to Neural Networks for Java, 2nd Edition, 2011.
44. HO, T., The random subspace method for constructing decision forests. IEEE Trans. Pattern Analysis and Machine Intelligence, 20 (8), 1998., str. 832-844.
45. Holz HJ, Loew MH, Relative feature importance: a classifier-independent approach to feature selection. In: Gelsema ES, Kanal LN (eds) Pattern Recognition in Practice, vol IV, Elsevier, 1994., str. 473–487
46. Jain, A., Zongker, D., Feature selection: Evaluation, application and small sample performance. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(2), 1997., str. 153–158.
47. Janecek, A., Efficient feature reduction and classification methods, Doctoral dissertation Universitat Wien, 2009.
48. Japkowicz, N., Shah, M., Evaluating learning algorithms: A classification perspective, Cambridge University Press, New York, 2011.
49. John, G.H., Kohavi, R., Pfleger, K., Irrelevant features and the subset selection problem, Machine Learning: Proceedings of the Eleventh International Conference, edited by William W. Cohen and Haym Hirsh, 1994., str. 121-129.
50. Kira, K., L. A. Rendell, A practical approach to feature selection'. In: D.Sleeman and P.Edwards (eds.): *Machine Learning: Proceedings of International Conference (ICML'92)*., 1992., str. 249–256.
51. Koller, D., Sahami, M., Toward optimal feature selection, Proceedings of the Thirteenth International Conference on Machine Learning, 1996., str. 284-292.
52. Kohavi, R., John, G. H.. Wrappers for feature subset selection. Artificial Intelligence, 97(1-2), 1997., str. 273–324.

53. Kohavi, R., Sommerfield, D., Feature subset selection using the wrapper method: overfitting and dynamic search space topology, Proceedings of the 1st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1995, str. 192-197
54. Kralj Novak, P., Lavrač, N., Gamberger, D., and Krstačič, A., CSM-SD: Methodology for contrast set mining through subgroup discovery. Journal of Biomedical Informatics, 42(1), 2009,.113–122.
55. Kudo M, Sklansky J (1998) Classifier-independent feature selection for two-stage feature selection. In: Amin A, Dori D, Pudil P, Freeman H (eds) Proceedings of the Joint IAPR International Workshops on SSPR'98 and SPR'98, str. 548–55
56. Kusonmano, K., Netzer, M., Pfeifer. B., Baumgartner, C., Liedl, K.R., Graber A.,Evaluation of the impact of dataset characteristics for classification problems in biological applications, ICBB 2009, International Conference on Bioinformatics and Biomedicine, Venice, Italy, str. 966.-970.
57. Kwok, S. W., Carter, C., Multiple decision trees. Uncertainty in Artificial Intelligence, 4, 1990., str. 327-335.
58. Lahiri, R., Comparison of data mining and statistical techniques for classification model, The department of information systems and decision sciences, Jadavpur University India, Thesis, 2006.
59. Lavanya, D., Usha Rani, K., Analysis of feature selection with classification: breast cancer datasets, Indian Journal of Computer Science and Engineering (IJCSE), 2 (5), 2011., str. 756-763
60. Liu, H., Motoda, H., Setiono, R., Zhao, Z., Feature Selection: An Ever Evolving Frontier in Data Mining, Proceedings of the Fourth International Workshop on Feature Selection in Data Mining June 21st, 2010, Hyderabad, India, str. 4-13.
61. Liu, H., Motoda, H., Yu, L., Feature selection with selective sampling, Proceeding ICML '02 Proceedings of the Nineteenth International Conference on Machine Learning, 2002, str. 395-402
62. Liu, H., Yu, L., Toward Integrating Feature Selection Algorithms for Classification and Clustering, IEEE Transactions on Knowledge and Data Engineering, 17 (4), 2005., str. 491-502.
63. Loekito, E., Bailey, J., Mining influential attributes that capture class and group contrast behavior, CIKM, 2008, str. 971-980.

64. Longadge, R., Dongre, S.S., Malik, L., Class Imbalance Problem in Data Mining: Review, International Journal of Computer Science and Network, 2 (1), 2013., str. 1-6
65. Luengo, J., Garcia, S., Herrera, F., A Study on the Use of Statistical Tests for Experimentation with Neural Networks, Expert Systems with Applications, 36 (4), 2009., str. 7798.-7808.
66. Lutu, P.E.N.,2010, Dataset selection for aggregate model implementation in predictive data mining, PhD thesis, University of Pretoria, Pretoria, pristupano 29.04.2013. < http://upetd.up.ac.za/thesis/available/etd-11152010-203041/ >
67. Marbán O.; Segovia J.; Menasalvas E.; Fernández – Baizán C., Toward data mining engineering: A software engineering approach. *Information Systems*, 2009., str.87–107.
68. McKinsey Global Institute, Big data: The next frontier for innovation, competition, and productivity, dostupno na: www.mckinsey.com/.../Big%20Data/MGI_big_data_full_report.ashx, zadnji pristup: 20.06.2012.
69. Michie, D., Spiegelhalter, D.J., Taylor, C.C., Machine Learning, Neural and Statistical Classification, 1994.
70. Mladenić. D., Feature selection for dimensionality reduction, In Lecture Notes in Computer Science Volume 3940, 2006, str. 84-102.
71. Mladenić. D., Grobelnik, M., Feature selection for unbalanced class distribution and Naive Bayes, Proceedings of the Sixteenth International Conference on Machine Learning (ICML), 1999. str. 258-267.
72. Moreno-Torres. Moreno-Torres, J.G., Raeder, T., Alaiz-Rodriguez, R., Chawla, N.V., Herrera, V., A unifying view on dataset shift in classification, Pattern Recognition, 45, 2012., str. 521–530
73. Novaković, J., Strbac, P., Bulatović, D., Toward optimal feature selection using ranking methods and classification algorithms, Yugoslav Journal of Operations Research, 21 (1), 2011, str. 119-135.
74. Ooi, C. H., Chetty, M., Teng, S. W., Differential prioritization in feature selection and classifier aggregation for multiclass microarray datasets. Data Mining and Knowledge Discovery,14, 2007., str. 329-366.
75. Oreški, S., Oreški, D., Oreški, G., Hybrid System with Genetic Algorithm and Artificial Neural Networks and its Application to Retail Credit Risk Assessment, Expert systems with applications, 39 (16), 2012., str. 12605–12617.

76. Polat, K., Gunes, S., A new feature selection method on classification of medical datasets: Kernel F-score feature selection, Expert Systems with Applications, 36, 2009., str. 10367–10373.
77. Pudil, P., Novovičova, J., Kittler, J., Floating search methods in feature selection, Pattern Recognition Letters, 15, 1994., str. 1119-1125
78. Ruiz, R., Riquelme, J.C., Aguilar-Ruiz, J.S., Garcia-Torres, M., Fast feature selection aimed at high-dimensional data via hybrid-sequential-ranked searches, Expert Systems with Applications 39, 2012, str. 11094–11102.
79. Quinlan, J. R., C4.5: Programs for machine learning, Morgan Kaufman, 1994.
80. Ramaswami, M., Bhaskaran, R., A Study on Feature Selection Techniques in Educational Data Mining, Journal of computing 1(1), 2009., str. 7- 11.
81. Robnik-Šikonja, M., Kononenko, I., Theoretical and Empirical Analysis of ReliefF and RReliefF, Machine Learning Journal, 53, 2003., str. 23-69.
82. Salzberg, S.L., On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach, Data Mining and Knowledge Discovery, 1, 1997., str. 317–327.
83. Satsanagi, A., Data Mining Using Contrast-sets: A Comparative Study, Master thesis, University of Alberta, 2011.
84. Silva, L.O.L.A., Koga, M.L., Cugnasca, C.E., Costa, A.H.R., Comparative assessment of feature selection and classification techniques for visual inspection of pot plant seedlings, Compters and electronics in agriculture, 97, 2013, str. 47-55.
85. *Sociology Data Set Server of Saint Joseph`s University in Philadelphia*, dostupno na:http://sociology-data.sju.edu/, pristupano: 14.12.2012.
86. Sohn, S.Y., Meta analysis of classification algorithms for pattern recognition, IEEE transactions on pattern analysis and machine intelligence, 21 (11), 1999, str. 1137-1144.
87. *StatLib - Carnegie Mellon University,* dostupno na: http://lib.stat.cmu.edu/, pristupano: 10.12.2012.
88. Sumathi, S., Sivanandam, S.N., Introduction to data mining and its application, Studies in Computational Intelligence, Vol. 29 , 2006, XXII, Springer, Berlin
89. Sun, X., Liu, Y., Li, J., Zhu, J., Chen, H., Liu, X., Feature evaluation and selection with cooperative game theory, Pattern recognition, 45, 2012., str. 2992-3002.
90. Tu, J.V., Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes, J Clin Epidemiol, 49 (11), 1996., str. 1225-1231.

91. *UCI Machine Learning Repository,* dostupno na: http://archive.ics.uci.edu/ml/datasets.html, zadnji pristup: 29.10.2013.
92. Van der Walt, C.M., Data measures that characterise classification problems, Master's Dissertation, dostupno na: http://upetd.up.ac.za/thesis/available/etd-08292008-162648/, zadnji pristup: 21.09.2013.
93. Verikas, A., Bacauskiene, M., Feature selection with neural networks, Pattern Recognition Letters, 23, 2002., str. 1323-1335.
94. Webb, G. I., Efficient Search for Association Rules. In R. Ramakrishnan and S. Stolfo (Eds.), Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2000) Boston, MA. New York: The Association for Computing Machinery, 2000., str. 99-107.
95. Webb, G. I., Butler, S., Newlands, D., On detecting differences between groups, KDD 2003., Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York: ACM., str. 739 – 745.
96. Weiss, S. I., and Kulikowski, C. 1991. Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Networks, Machine Learning, and Expert Systems. San Francisco, Calif.:Morgan Kaufmann.
97. Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Feature selection for SVMs, Advances in Neural Information Processing Systems 13, 2001, str. 668 - 674
98. Wu, X., Zhu, X., Artificial Intelligence Review, 22, 2004., str.177–210.
99. Wu, X., Zhu, X., Mining with noise knowledge: error-aware data mining, IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, 38 (4), 2008., str. 917-932.
100. Yang, Y., A., Pedersen, J., 1997. A comparative study of feature selection in text categorization. In: ICML '97: Proceedings of the Fourteenth International Conference on Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., str. 412–420.
101. Yu, L., Liu, H., Efficient Feature Selection via Analysis of Relevance and Redundancy, Journal of Machine Learning Research, 5, 2004, str. 1205–1224.
102. Zahedi, F., Intelligent Systems for Business, Expert Systems With Neural Networks, Wodsworth Publishing Inc., 1993.
103. Zekić, M., Neural Network Applications in Stock Market Predictions – A Methodology Analysis, Proceedings of the 9th International Conference on

Information and Intelligent Systems ‘98, Eds. Aurer, B., Logožar, R., Varaždin, September 23-25, 1998., str. 255-263.

PRILOZI

Tablica 16. Kvantitativne vrijednosti karakteristika skupa podataka

Broj atributa	Broj instanci	Oskudnost	Korelacija	Normalnost	Homogenost	Omjer ID	Šum atributa	Skup
13	108	mala	0.057	da	da	0.308	0.692	*Pittsburgh+Bridges*
32	10	mala	0.048	da	da	0.938	0.063	*Trains*
4	16	mala	0.033	da	ne	0.250	0.750	*Balloons*
5	12	mala	0.084	da	ne	0.800	0.200	*Titanic*
10	95	mala	0.111	ne	da	0.300	0.700	*broadway*
16	14	mala	0.082	ne	da	0.875	0.125	*assessment*
35	47	mala	0.006	ne	ne	0.600	0.400	*Soybean+Small*
57	106	mala	0.144	ne	ne	0.982	0.018	*molecular biology promoters*
45	80	mala	0.409	da	da	0.644	0.356	*Spectf*
10	52	mala	0.201	da	da	0.900	0.100	*japansolvent*
8	90	mala	0.285	da	ne	0.625	0.375	*Post-Operative+Patient*

Broj atributa	Broj instanci	Oskudnost	Korelacija	Normalnost	Homogenost	Omjer ID	Šum atributa	Skup
19	155	mala	0.688	da	ne	0.947	0.053	*hepatitis*
16	67	mala	0.561	ne	ne	0.125	0.875	*election*
56	32	mala	0.434	ne	ne	0.839	0.161	*Lung Cancer*
42	77	mala	0.288	ne	da	0.429	0.571	*sponge*
7	100	mala	0.337	ne	da	0.857	0.143	*creditscore*
7	50	velika	0.117	da	da	0.143	0.857	*bankruptcy*
10	74	velika	0.025	da	da	0.900	0.100	*gviolence*
16	57	velika	0.046	da	ne	0.313	0.688	*Labor+Relations*
6	120	velika	0.142	da	ne	1.000	0.000	*Acute+Inflammations*
11	60	velika	0.126	ne	da	0.364	0.636	*runshoes*
11	92	velika	0.068	ne	da	0.909	0.091	*Cyyoung9302*
11	100	velika	0.142	ne	ne	0.182	0.818	*impeach*
12	42	velika	0.041	ne	ne	0.917	0.083	*fraud*
114	138	velika	0.680	da	da	0.333	0.667	*Campus Climate 2011 SJU*
28	162	velika	0.576	da	da	0.964	0.036	*homerun*

Broj atributa	Broj instanci	Oskudnost	Korelacija	Normalnost	Homogenost	Omjer ID	Šum atributa	Skup
60	208	velika	0.685	da	ne	0.317	0.683	*sonar*
12	57	velika	0.657	da	ne	0.917	0.083	*bondrate*
100	153	velika	0.549	ne	ne	0.410	0.590	*ICPSR 3009*
10	159	velika	0.496	ne	ne	0.900	0.100	*gsssexsurvey*
17	32	velika	0.324	ne	da	0.412	0.588	*uktrainacc*
20	120	velika	0.311	ne	da	0.950	0.050	*ncaa*
15	690	mala	0.077	da	da	0.333	0.667	*credit*
16	550	mala	0.023	da	da	0.875	0.125	*weights*
65	5216	mala	0.073	da	ne	0.477	0.523	*ICPSR 2743*
97	1049	mala	0.121	da	ne	0.887	0.113	*city*
8	4052	mala	0.082	ne	da	0.250	0.750	*supreme*
108	15779	mala	0.079	ne	da	0.843	0.157	*ICPSR 2751*
5	748	mala	0.148	ne	ne	0.200	0.800	*blood-transfusion/*
71	842	mala	0.105	ne	ne	0.901	0.099	*authorship*
129	1037	mala	0.639	da	da	0.147	0.853	*ICPSR 2867*

Broj atributa	Broj instanci	Oskudnost	Korelacija	Normalnost	Homogenost	Omjer ID	Šum atributa	Skup
53	416	mala	0.520	da	da	0.849	0.151	*ICPSR 2480*
18	1340	mala	0.359	da	ne	0.278	0.722	*halloffame*
11	534	mala	0.238	da	ne	1.000	0.000	*CPS_85_Wages*
52	3850505	mala	0.346	ne	ne	0.077	0.923	*Physical+Activity+Monitoring*
33	364	mala	0.323	ne	ne	0.970	0.030	*marketing*
48	1399	mala	0.525	ne	da	0.604	0.396	*binge*
34	351	mala	0.386	ne	da	0.941	0.059	*ionosphere*
101	1342	velika	0.125	da	da	0.644	0.356	*ICPSR 2859*
22	8124	velika	0.160	da	da	0.955	0.045	*Mushroom*
109	5866	velika	0.045	da	ne	0.532	0.468	*ICPSR 2039*
21	7200	velika	0.018	da	ne	0.810	0.190	*Thyroid+Disease*
30	3772	velika	0.096	ne	da	0.367	0.633	*sick*
64	1600	velika	0.030	ne	da	0.859	0.141	*One-hundred+plant+species+leaves+data+set*

Broj atributa	Broj instanci	Oskudnost	Korelacija	Normalnost	Homogenost	Omjer ID	Šum atributa	Skup
37	3196	velika	0.120	ne	ne	0.649	0.351	*Kr-Vs-Kp*
9	958	velika	0.101	ne	ne	0.889	0.111	*tic-tac-toe*
52	2831	velika	0.552	da	da	0.192	0.808	*abgss98*
81	864	velika	0.294	da	da	0.914	0.086	*ICPSR 2686*
49	4657	velika	0.554	da	ne	0.102	0.898	*ICPSR 2155*
14	270	velika	0.347	da	ne	1.000	0.000	*heart-statlog*
58	4601	velika	0.411	ne	ne	0.379	0.621	*spambase*
17	435	velika	0.495	ne	ne	0.824	0.176	*Vote*
101	606	velika	0.335	ne	da	0.356	0.644	*Hill-Valley*
20	3196	velika	0.216	ne	da	0.900	0.100	*hepatitis*
483	132	mala	0.011	da	da	0.532	0.468	*ICPSR 4291*
453	160	mala	0.036	da	da	0.819	0.181	*ICPSR 4582*
495	148	mala	0.099	da	ne	0.101	0.899	*ICPSR 9595*
1851	96	mala	0.144	da	ne	0.844	0.156	*ICPSR 21600 2*
2532	34	mala	0.124	ne	da	0.457	0.543	*ICPSR 21600 3*

Broj atributa	Broj instanci	Oskudnost	Korelacija	Normalnost	Homogenost	Omjer ID	Šum atributa	Skup
269	81	mala	0.115	ne	da	0.989	0.011	*ICPSR 28641 2*
1769	118	mala	0.154	ne	ne	0.110	0.890	*ICPSR 6542*
397	197	mala	0.078	ne	ne	0.947	0.053	*ICPSR 4367*
2984	20	mala	0.358	da	da	0.351	0.649	*ICPSR 4572 02*
2799	71	mala	0.457	da	da	0.823	0.177	*ICPSR 21600 4*
4702	64	mala	0.284	da	ne	0.542	0.458	*DBWorld+e-mails*
305	90	mala	0.566	da	ne	0.987	0.013	*ICPSR 6135*
288	95	mala	0.334	ne	ne	0.125	0.875	*ICPSR 4537 8th form 1*
907	171	mala	0.263	ne	ne	0.954	0.046	*ICPSR 4275*
22283	85	mala	0.504	ne	da	0.305	0.695	*GLI-85*
922	135	mala	0.418	ne	da	0.977	0.023	*ICPSR 21600 1*
2646	189	velika	0.079	da	da	0.321	0.679	*ICPSR 4566 02*
541	44	velika	0.067	da	da	0.874	0.126	*ICPSR 8255*
1345	129	velika	0.002	da	ne	0.394	0.606	*ICPSR 28641*
19993	187	velika	0.027	da	ne	0.932	0.068	*SMK-CAN-187*

Broj atributa	Broj instanci	Oskudnost	Korelacija	Normalnost	Homogenost	Omjer ID	Šum atributa	Skup
9120	56	velika	0.035	ne	da	0.224	0.776	*ICPSR 23041 2*
361	38	velika	0.059	ne	da	0.873	0.127	*ICPSR 6480*
292	43	velika	0.064	ne	ne	0.620	0.380	*ICPSR 4537 10th form 2*
2235	84	velika	0.121	ne	ne	0.955	0.045	*ICPSR 4138*
5413	110	velika	0.601	da	da	0.398	0.602	*ICPSR 23041*
3502	38	velika	0.276	da	da	0.955	0.045	*ICPSR 4690*
563	102	velika	0.587	da	ne	0.632	0.368	*ICPSR 4372*
618	17	velika	0.498	da	ne	0.934	0.066	*ICPSR 20022*
306	63	velika	0.356	ne	ne	0.212	0.788	*ICPSR 6484*
2646	189	velika	0.411	ne	ne	0.990	0.010	*ICPSR 4566 01*
2960	85	velika	0.324	ne	da	0.273	0.727	*ICPSR 6693*
2984	20	velika	0.284	ne	da	0.963	0.037	*ICPSR 4572 01*
1950	100000	mala	0.085	da	da	0.083	0.917	*Dorothea*
561	10299	mala	0.036	da	da	0.838	0.162	*Human+Activity+Recognition+Using+Smartphones*

Broj atributa	Broj instanci	Oskudnost	Korelacija	Normalnost	Homogenost	Omjer ID	Šum atributa	Skup
365	1822	mala	0.016	da	ne	0.162	0.838	*ICPSR 31221*
506	158865	mala	0.007	da	ne	0.862	0.138	*ICPSR 3669*
269	18513	mala	0.025	ne	da	0.431	0.569	*ICPSR 2743 Person Level Data*
1044	2965	mala	0.089	ne	da	0.836	0.164	*ICPSR 2258*
500	4400	mala	0.065	ne	ne	0.306	0.694	*Madelon*
2167	16281	mala	0.119	ne	ne	0.932	0.068	*adult*
824	6857	mala	0.279	da	da	0.278	0.722	*ICPSR 31202 5*
1066	962	mala	0.485	da	da	0.997	0.003	*ICPSR 2857*
751	907	mala	0.239	da	ne	0.609	0.391	*ICPSR 2346*
138672	440	mala	0.554	da	ne	0.971	0.029	*PEMS-SF*
20000	2600	mala	0.410	ne	ne	0.248	0.752	*Dexter*
819	864	mala	0.297	ne	ne	0.990	0.010	*ICPSR 2686 Caregiver Data*
219	2991	mala	0.343	ne	da	0.292	0.708	*ICPSR 3534*
686	8915	mala	0.321	ne	da	0.822	0.178	*ICPSR 2535*
837	20791	velika	0.148	da	da	0.427	0.573	*ICPSR 2149*

Broj atributa	Broj instanci	Oskudnost	Korelacija	Normalnost	Homogenost	Omjer ID	Šum atributa	Skup
256	1593	velika	0.094	da	da	0.973	0.027	*Semeion+Handwritten+Digit*
274	1754	velika	0.157	da	ne	0.208	0.792	*ICPSR 3548*
5000	13500	velika	0.078	da	ne	0.809	0.191	*Gisette*
1205	1825	velika	0.064	ne	da	0.009	0.991	*ICPSR 2295*
652	5216	velika	0.073	ne	da	0.845	0.155	*ICPSR 2743*
4367	2032	velika	0.028	ne	ne	0.055	0.945	*ICPSR 2163*
591	1567	velika	0.011	ne	ne	0.929	0.071	*SECOM*
285	1484	velika	0.258	da	da	0.509	0.491	*ICPSR 3789*
1120	4022	velika	0.470	da	da	0.929	0.071	*ICPSR 2833*
10000	900	velika	0.342	da	ne	0.014	0.986	*Arcene*
1852	7999	velika	0.284	da	ne	0.884	0.116	*ICPSR 2566*
617	6857	velika	0.416	ne	ne	0.360	0.640	*ICPSR 31202 4*
109	5866	velika	0.292	ne	ne	0.945	0.055	*ICPSR 2039 2*
311	94716	velika	0.576	ne	da	0.540	0.460	*ICPSR 3151*
201	731	velika	0.333	ne	da	0.836	0.164	*ICPSR 6627*

Tablica 17.Točnost neuronske mreže

	SfFS	*MOFS*	*Informacijska dobit*	*Omjer dobiti*	*Linearni odabir unaprijed*	*Relief*	*Tehnika glasovanja*
Pittsburgh+Bridges	64.54	65.39	63.22	63.35	62.18	65.48	63.55
Trains	85.94	84.32	76.74	76.98	78.77	84.41	76.49
Balloons	91.23	88.03	88.12	88.65	88.10	89.76	88.15
Titanic	97.88	96.17	96.24	96.00	94.43	94.44	94.97
broadway	81.25	77.72	71.56	70.85	75.24	78.65	72.23
assessment	92.04	92.35	91.24	91.75	90.12	90.56	90.07
Soybean+Small	90.01	87.46	85.66	88.35	85.04	84.45	85.99
molecular biology promoters	95.42	91.35	93.66	89.34	88.24	90.86	90.91
Spectf	90.01	88.32	87.12	88.21	88.75	90.34	87.75
japansolvent	87.43	83.64	80.04	80.98	81.16	81.19	84.35
Post-Operative+Patient	91.82	89.55	89.76	91.06	89.21	91.11	89.44
hepatitis	84.35	81.03	80.05	79.54	81.24	79.22	78.16
election	86.00	86.15	84.87	84.44	85.10	86.34	85.12
Lung Cancer	76.69	76.6	73.34	74.01	73.11	75.43	72.23
sponge	83.76	76.34	75.61	75.22	74.18	79.45	79.25
creditscore	88.00	88.15	87.16	87.54	79.17	88.81	79.63

	SfFS	*MOFS*	*Informacijska dobit*	*Omjer dobiti*	*Linearni odabir unaprijed*	*Relief*	*Tehnika glasovanja*
bankruptcy	92.44	91.88	92.00	91.54	90.05	90.12	91.00
gviolence	91.13	88.10	88.21	89.65	87.05	87.19	88.54
Labor+Relations	75.18	76.12	73.21	73.38	75.02	76.54	74.45
Acute+Inflammations	100.00	99.55	99.00	99.00	99.15	98.55	98.76
runshoes	97.98	96.65	97.43	96.44	95.58	96.14	95.74
Cyyoung9302	96.54	95.13	95.21	95.89	94.44	94.87	95.27
impeach	93.86	93.01	93.34	93.18	93.62	92.16	92.09
fraud	94.47	94.20	93.84	93.35	92.45	92.66	92.17
Campus Climate 2011 SJU	81.79	86.15	84.32	83.96	81.13	80.22	82.74
homerun	85.77	85.29	83.11	83.36	85.03	84.38	84.29
sonar	76.34	77.01	74.41	75.54	74.38	78.43	73.31
bondrate	89.64	88.97	88.21	88.18	88.06	88.24	88.52
ICPSR 3009	69.85	68.52	66.25	65.19	63.21	68.76	62.73
gsssexsurvey	83.65	82.87	81.37	81.94	82.08	82.19	82.05
uktrainacc	83.61	85.28	83.19	82.12	81.55	86.17	82.76
ncaa	97.77	93.11	93.35	92.82	94.44	95.12	91.74
credit	86.24	82.94	81.47	78.87	80.00	83.75	81.22

	SfFS	*MOFS*	*Informacijska dobit*	*Omjer dobiti*	*Linearni odabir unaprijed*	*Relief*	*Tehnika glasovanja*
weights	85.04	85.22	85.55	84.12	84.38	84.63	84.95
ICPSR 2743	71.93	63.99	62.81	62.34	62.16	64.77	66.28
city	81.93	80.04	76.45	76.59	78.13	78.22	78.86
supreme	81.48	81.98	82.14	83.22	82.19	85.43	80.06
ICPSR 2751	68.99	65.41	65.32	64.31	61.47	61.91	62.18
blood-transfusion/	68.75	70.44	72.19	68.14	68.21	71.51	68.43
authorship	55.51	52.31	53.81	54.12	52.25	52.10	52.06
ICPSR 2867	71.23	70.15	67.15	66.81	70.01	68.31	66.54
ICPSR 2480	78.95	77.45	76.19	75.92	76.36	75.11	73.21
halloffame	54.32	51.21	50.67	50.89	51.39	52.18	52.11
CPS_85_Wages	84.39	85.22	82.56	82.00	83.97	85.64	81.59
Physical+Activity+Monitoring	45.95	43.17	42.19	42.55	41.16	43.39	41.29
Marketing	96.66	93.33	90.05	89.15	94.25	92.15	90.28
binge	80.09	77.81	76.41	76.98	78.42	78.56	77.95
ionosphere	91.78	92.15	89.45	90.62	88.13	93.86	88.72
ICPSR 2859	75.18	75.27	78.81	77.12	76.16	73.91	73.22
Mushroom	98.87	97.12	97.61	96.89	96.72	96.11	95.25

	SfFS	*MOFS*	*Informacijska dobit*	*Omjer dobiti*	*Linearni odabir unaprijed*	*Relief*	*Tehnika glasovanja*
ICPSR 2039	83.32	79.21	81.12	80.16	80.52	82.45	81.02
Thyroid+Disease	76.54	74.23	73.81	73.13	74.12	72.87	72.24
sick	98.88	97.64	93.15	90.54	91.87	95.18	91.24
One-hundred+plant+species+leaves+data+set	88.19	82.16	83.38	85.22	81.52	80.71	78.83
Kr-Vs-Kp	85.45	86.31	84.97	87.76	81.27	89.25	83.22
tic-tac-toe	95.21	91.47	98.34	96.68	90.47	92.15	93.15
abgss98	84.38	83.21	81.51	80.89	80.66	83.01	82.18
ICPSR 2686	89.01	83.22	85.01	84.31	85.22	87.24	85.97
ICPSR 2155	68.31	66.57	63.39	65.19	65.11	64.32	66.98
heart-statlog	96.71	94.15	95.12	96.16	94.56	93.78	95.57
spambase	93.84	90.22	86.15	82.00	84.29	86.39	85.27
Vote	97.79	95.41	94.14	94.21	96.32	94.33	94.25
Hill-Valley	80.51	82.65	82.15	81.27	80.09	83.41	79.14
hepatitis	87.63	85.24	86.79	87.18	84.78	84.36	85.88
ICPSR 4291	91.12	87.15	86.42	88.76	83.22	90.03	82.72
ICPSR 4582	89.21	87.10	88.51	86.42	84.23	82.95	85.73
ICPSR 9595	93.44	91.28	91.76	92.81	90.55	93.29	90.16

	SfFS	*MOFS*	*Informacijska dobit*	*Omjer dobiti*	*Linearni odabir unaprijed*	*Relief*	*Tehnika glasovanja*
ICPSR 21600 2	85.47	83.19	78.85	73.25	68.96	77.61	81.17
ICPSR 21600 3	86.91	83.98	85.12	80.67	79.14	78.87	79.51
ICPSR 28641 2	67.52	64.57	65.17	65.89	63.48	63.29	63.26
ICPSR 6542	81.26	80.35	75.23	77.36	76.82	79.18	80.11
ICPSR 4367	92.34	91.81	88.91	87.36	90.55	89.45	90.13
ICPSR 4572 02	57.86	54.38	52.99	52.17	55.75	56.62	54.91
ICPSR 21600 4	84.75	84.19	78.97	78.16	81.93	83.79	80.17
DBWorld+e-mails	66.61	64.83	63.95	63.18	60.71	58.95	61.71
ICPSR 6135	97.16	95.13	94.62	95.88	92.82	93.67	91.77
ICPSR 4537 8th form 1	83.25	80.03	81.77	81.00	78.82	82.71	76.92
ICPSR 4275	76.68	75.68	73.92	74.51	72.13	71.06	75.10
GLI-85	54.31	50.56	49.80	48.37	51.19	52.98	51.02
ICPSR 21600 1	90.56	89.93	85.43	85.16	84.33	87.66	87.14
ICPSR 4566 02	83.87	80.71	76.54	78.22	81.65	75.12	75.43
ICPSR 8255	87.16	86.91	80.55	80.19	83.44	83.21	84.78
ICPSR 28641	66.13	65.71	64.44	62.65	65.19	63.18	61.36
SMK-CAN-187	56.77	52.11	51.91	53.45	48.76	46.51	50.73

	SfFS	*MOFS*	*Informacijska dobit*	*Omjer dobiti*	*Linearni odabir unaprijed*	*Relief*	*Tehnika glasovanja*
ICPSR 23041 2	75.56	74.32	73.11	74.12	72.23	71.61	73.16
ICPSR 6480	93.33	92.71	93.12	92.18	91.15	91.10	90.77
ICPSR 4537 10th form 2	97.78	97.14	94.22	92.59	95.81	95.66	93.81
ICPSR 4138	83.38	81.10	80.05	81.22	80.26	80.41	80.67
ICPSR 23041	66.61	62.33	61.15	60.22	58.71	64.21	59.75
ICPSR 4690	51.22	50.03	48.81	45.55	47.64	49.75	46.81
ICPSR 4372	77.76	75.48	72.38	75.22	74.11	74.47	73.88
ICPSR 20022	45.76	44.39	41.23	40.99	44.12	44.03	42.22
ICPSR 6484	84.35	82.17	80.71	81.14	80.23	81.69	82.65
ICPSR 4566 01	77.67	75.64	75.12	74.44	74.12	72.25	71.98
ICPSR 6693	72.32	72.11	68.54	67.34	66.12	71.86	69.95
ICPSR 4572 01	69.67	68.17	65.31	66.53	63.81	61.85	64.23
Dorothea	94.86	85.47	88.92	82.39	86.65	90.34	86.53
Human+Activity+Recognition+Using+Smartphones	82.39	81.85	79.93	79.41	77.76	76.65	80.31
ICPSR 31221	77.52	75.13	72.76	72.31	73.94	75.84	74.75
ICPSR 3669	77.56	77.86	80.53	81.65	82.35	78.84	80.21
ICPSR 2743 Person Level Data	64.74	58.86	57.73	57.42	56.68	63.12	60.22

	SfFS	*MOFS*	*Informacijska dobit*	*Omjer dobiti*	*Linearni odabir unaprijed*	*Relief*	*Tehnika glasovanja*
ICPSR 2258	84.35	81.26	80.22	80.58	79.91	78.14	80.76
Madelon	74.29	67.17	64.19	69.28	64.24	71.72	68.33
adult	77.76	73.12	68.86	70.91	71.12	68.17	67.57
ICPSR 31202 5	81.27	72.55	71.19	71.00	75.45	78.72	70.64
ICPSR 2857	95.12	93.19	91.02	89.91	90.81	91.10	89.16
ICPSR 2346	63.69	66.82	63.22	64.53	67.51	61.86	61.98
PEMS-SF	88.95	84.41	87.17	85.19	84.23	86.16	86.91
Dexter	90.01	83.38	85.14	86.24	82.17	87.63	84.57
ICPSR 2686 Caregiver Data	85.39	83.18	80.53	80.22	82.06	81.74	82.11
ICPSR 3534	92.13	89.91	87.12	87.77	88.76	90.65	89.15
ICPSR 2535	94.44	89.68	93.12	92.00	90.74	90.13	91.19
ICPSR 2149	87.17	82.18	84.98	84.31	81.19	83.37	81.97
Semeion+Handwritten+Digit	81.92	79.81	76.18	74.36	76.58	74.91	78.87
ICPSR 3548	73.78	73.61	72.81	72.67	72.35	73.25	72.98
Gisette	92.64	88.45	87.82	90.01	87.76	87.98	89.62
ICPSR 2295	94.67	92.89	90.90	90.56	91.10	90.44	90.75
ICPSR 2743	97.18	95.98	92.87	91.52	92.13	94.81	94.19

	SfFS	*MOFS*	*Informacijska dobit*	*Omjer dobiti*	*Linearni odabir unaprijed*	*Relief*	*Tehnika glasovanja*
ICPSR 2163	76.59	73.97	70.66	74.78	73.21	72.19	71.74
SECOM	85.63	84.53	84.89	83.47	83.74	84.12	85.12
ICPSR 3789	87.49	86.11	85.21	84.78	83.96	86.39	85.43
ICPSR 2833	82.34	77.95	78.86	81.19	77.87	79.54	80.71
Arcene	66.29	64.11	59.45	58.89	61.54	62.85	60.63
ICPSR 2566	71.08	65.19	69.89	67.81	66.62	64.78	64.19
ICPSR 31202 4	82.19	80.05	80.00	79.61	79.20	77.93	77.41
ICPSR 2039 2	92.36	92.12	90.64	89.52	92.10	91.76	91.27
ICPSR 3151	86.49	84.44	83.88	83.95	81.17	82.87	80.75
ICPSR 6627	65.97	63.19	61.36	60.52	63.07	62.45	62.14

Tablica 18. Točnost diskriminacijska analiza

	SfFS	*MOFS*	*Informacijska dobit*	*Omjer dobiti*	*Linearni odabir unaprijed*	*Relief*	*Tehnika glasovanja*
Pittsburgh+Bridges	70.29	68.52	66.61	66.18	68.87	69.65	69.18
Trains	81.72	80.14	75.55	75.23	76.45	78.65	77.52
Spectf	88.59	86.25	86.34	86.11	85.57	87.45	84.22
japansolvent	88.87	86.29	84.99	84.13	83.67	82.84	82.04
bankruptcy	93.54	90.42	90.59	92.88	90.11	89.96	89.35
gviolence	91.67	90.56	87.65	88.01	90.00	88.56	91.19
Campus Climate 2011 SJU	90.38	89.76	89.45	90.24	88.90	90.55	89.75
homerun	81.67	80.45	78.86	78.35	80.24	79.45	79.89
credit	85.89	83.27	81.97	82.29	80.32	83.06	80.76
weights	84.39	87.22	84.78	83.86	83.62	87.69	84.12
ICPSR 2867	71.23	69.13	67.18	67.26	69.84	69.11	67.43
ICPSR 2480	78.28	78.95	73.86	73.32	74.91	71.94	71.35
ICPSR 2859	76.15	78.81	77.07	77.64	78.40	75.35	75.21
Mushroom	98.87	96.12	96.54	96.01	95.22	95.47	94.36
abgss98	84.38	84.03	81.72	81.13	83.52	84.22	83.24
ICPSR 2686	87.27	89.01	85.63	85.24	84.77	86.35	84.31

	SfFS	*MOFS*	*Informacijska dobit*	*Omjer dobiti*	*Linearni odabir unaprijed*	*Relief*	*Tehnika glasovanja*
ICPSR 4291	90.14	88.42	87.56	90.55	86.25	91.12	86.13
ICPSR 4582	88.15	87.67	89.21	87.12	86.22	87.10	86.34
ICPSR 4572 02	57.86	52.16	54.24	54.78	55.29	55.11	54.33
ICPSR 21600 4	84.75	81.56	77.22	75.46	76.35	77.54	80.00
ICPSR 4566 02	81.26	81.12	80.31	76.32	83.87	79.24	78.55
ICPSR 8255	87.16	86.55	83.99	81.35	82.71	84.32	83.46
ICPSR 23041	63.94	66.61	63.27	61.34	61.98	65.28	62.67
ICPSR 4690	51.22	50.16	47.49	44.97	44.21	48.87	46.81
Dorothea	91.52	94.86	90.16	88.17	83.74	85.39	89.64
Human+Activity+Recognition+Using+Smartphones	82.39	81.25	77.68	76.18	75.33	77.26	79.54
ICPSR 31202 5	81.27	78.51	76.50	75.99	80.06	78.97	77.96
ICPSR 2857	93.98	95.12	92.85	91.76	91.24	93.27	92.32
ICPSR 2149	87.17	84.68	82.85	81.86	83.73	82.34	84.12
Semeion+Handwritten+Digit	81.92	76.28	77.56	75.76	80.77	80.12	76.95
ICPSR 3789	85.93	87.49	82.67	84.28	82.19	87.12	84.16
ICPSR 2833	82.34	78.54	76.58	77.49	75.58	79.56	77.28

Tablica 19. Vrijeme provođenja tehnika selekcije atributa

	SfFS	*MOFS*	*Informacijska dobit*	*Omjer dobiti*	*Linearni odabir unaprijed*	*Relief*
Pittsburgh+Bridges	0.95	0.49	0.55	0.57	0.84	0.61
Trains	1.10	0.84	0.92	0.89	0.96	0.87
Balloons	0.53	0.18	0.21	0.26	0.29	0.25
Titanic	0.59	0.25	0.27	0.21	0.38	0.30
broadway	1.87	0.99	1.04	0.96	1.32	0.98
assessment	1.12	0.53	0.59	0.56	0.71	0.58
Soybean+Small	14.68	9.7	10.31	10.05	11.24	9.9
molecular biology promoters	16.53	10.00	8.7	8.1	11.66	9.1
Spectf	8.96	3.46	3.85	3.81	4.14	3.79
japansolvent	4.33	0.73	0.81	0.94	1.17	0.92
Post-Operative+Patient	3.97	1.34	1.15	1.03	1.67	1.53
hepatitis	6.53	1.89	1.94	1.97	2.54	2.14
election	4.85	1.90	1.85	1.94	2.31	1.98
Lung Cancer	5.11	2.44	2.56	2.63	2.94	2.79
sponge	4.94	3.31	3.59	3.54	3.95	3.97
creditscore	2.36	0.64	0.66	0.53	0.82	0.71

	SfFS	*MOFS*	*Informacijska dobit*	*Omjer dobiti*	*Linearni odabir unaprijed*	*Relief*
bankruptcy	2.54	0.43	0.55	0.48	0.73	0.59
gviolence	3.76	0.94	0.98	0.99	1.10	1.11
Labor+Relations	3.87	1.10	1.23	1.10	1.36	1.28
Acute+Inflammations	1.77	0.67	0.69	0.55	0.82	0.75
runshoes	2.53	0.85	0.92	0.89	1.23	1.12
Cyyoung9302	4.73	1.31	1.25	1.42	1.64	1.68
impeach	3.88	1.81	1.79	1.89	2.03	2.06
fraud	2.57	0.73	0.71	0.80	0.99	0.86
Campus Climate 2011 SJU	79.25	56.23	58.26	58.99	62.34	60.18
homerun	8.35	5.12	4.43	5.62	5.97	5.51
sonar	74.55	52.35	55.64	56.97	59.89	54.18
bondrate	3.34	0.92	0.88	0.98	1.22	1.00
ICPSR 3009	126.37	98.25	96.11	99.25	99.99	99.13
gsssexsurvey	3,68	1.23	1.45	1.54	1.97	1.39
uktrainacc	2,67	0,89	0,95	0,97	1,35	1,11
ncaa	5,04	2,28	2,10	2,31	3,02	2,53
credit	18.67	11.21	11.96	11.98	12.31	13.14

	SfFS	*MOFS*	*Informacijska dobit*	*Omjer dobiti*	*Linearni odabir unaprijed*	*Relief*
weights	27.78	21.08	22.25	22.34	23.91	22.15
ICPSR 2743	98.74	79,11	79,02	79,73	81,26	79,56
city	128,37	110.00	110,36	110,98	111,28	113,89
supreme	83,47	73,39	76,68	76,99	79,28	75,84
ICPSR 2751	150,02	138,93	136,64	141,13	145,64	140,28
blood-transfusion/	6,94	4,80	5,56	5,75	6,02	5,21
authorship	130,26	125,00	124,30	125,63	127,41	125,89
ICPSR 2867	181,52	175,00	174,45	177,63	179,63	177,56
ICPSR 2480	80,00	64,12	63,05	66,35	68,00	65,20
halloffame	99,84	95,50	97,03	97,10	98,86	97,05
CPS_85_Wages	6,70	3,80	3,99	4,50	4,68	3,95
Physical+Activity+Monitoring	74,29	55,00	61,46	62,33	65,84	59,65
marketing	13,13	10,03	10,25	10,86	11,97	10,35
binge	179,84	166,50	169,67	169,73	175,82	168,96
ionosphere	61,20	42,10	45,82	45,97	49,30	43,50
ICPSR 2859	76,51	55,50	61,25	61,98	68,41	58,62
Mushroom	67,24	51,35	56,81	56,92	58,11	55,32

	SfFS	*MOFS*	*Informacijska dobit*	*Omjer dobiti*	*Linearni odabir unaprijed*	*Relief*
ICPSR 2039	95,64	83,10	81,35	84,12	86,48	85,87
Thyroid+Disease	89,72	76,68	79,94	79,35	83,64	79,98
sick	142,31	119,71	123,56	124,03	129,38	126,72
One-hundred+plant+species+leaves+data+set	101,01	97,13	95,25	98,46	99,98	99,30
Kr-Vs-Kp	109,68	89,26	92,43	92,69	95,73	91,07
tic-tac-toe	25,34	18,16	19,32	19,84	22,68	20,05
abgss98	146,34	124,50	123,06	125,63	128,93	128,97
ICPSR 2686	163,82	145,16	143,20	149,38	152,47	148,64
ICPSR 2155	119.62	93.57	91.25	93.97	97.63	94.78
heart-statlog	9.95	4.76	6.22	6,91	7.82	5.49
spambase	146.99	114,21	120.82	123,78	128.44	118.93
Vote	12.09	2.01	6.19	3.22	8.63	4.81
Hill-Valley	155.55	124.89	121.56	127.88	132.23	126.73
hepatitis	138.94	102.29	96.11	99.75	104.73	98.94
ICPSR 4291	159.73	126.00	133.77	132.49	135.81	129.48
ICPSR 4582	278.61	225.66	229.15	231.54	245.96	238.74
ICPSR 9595	198.96	173.94	178.65	180.09	185.44	176.09

	SfFS	*MOFS*	*Informacijska dobit*	*Omjer dobiti*	*Linearni odabir unaprijed*	*Relief*
ICPSR 21600 2	311.26	254.38	265.86	261.54	271.29	257.12
ICPSR 21600 3	289.99	236.67	241.22	245.87	253.90	257.84
ICPSR 28641 2	127.50	98.77	99.97	99.84	102.39	99.25
ICPSR 6542	188.22	151.42	154.68	156.75	163.70	159.06
ICPSR 4367	175.64	129.28	133.38	138.33	142.49	136.74
ICPSR 4572 02	213.95	183.24	186.94	187.36	194.47	189.15
ICPSR 21600 4	283.55	252..28	259.90	259.07	265.73	256.68
DBWorld+e-mails	380.68	311.36	317.38	319.64	330.14	320.29
ICPSR 6135	111.76	88.26	96.59	93.39	98.99	90.05
ICPSR 4537 8th form 1	198.36	171.29	171.88	171.53	173.27	172.78
ICPSR 4275	222.26	164.48	175.54	171.33	188.44	168.26
GLI-85	397.64	349.82	355.76	358.19	378.21	366.39
ICPSR 21600 1	184.40	125.28	131.19	135.41	142.37	138.04
ICPSR 4566 02	200.05	177.23	180.22	179.01	190.29	188.10
ICPSR 8255	163.49	117.22	127.37	125.59	130.05	121.97
ICPSR 28641	270.11	231.44	242.13	240.85	255.17	249.56
SMK-CAN-187	397.30	346.11	355.29	358.84	372.10	366.14

	SfFS	*MOFS*	*Informacijska dobit*	*Omjer dobiti*	*Linearni odabir unaprijed*	*Relief*
ICPSR 23041 2	344.83	285.55	295.50	294.10	301.10	291.76
ICPSR 6480	92.19	81.11	86.20	85.62	88.30	88.85
ICPSR 4537 10th form 2	96.74	73.76	79.05	78.86	83.21	77.19
ICPSR 4138	256.21	220.09	228.17	229.41	238.27	239.14
ICPSR 23041	299.95	266.09	272.84	271.10	291.38	289.07
ICPSR 4690	347.94	301.76	304.55	305.29	321.80	311.42
ICPSR 4372	133.28	107.73	114.64	111.37	118.43	109.96
ICPSR 20022	49.23	39.27	41.12	42.00	46.69	44.41
ICPSR 6484	86.39	73.39	75.21	74.55	79.64	77.90
ICPSR 4566 01	222.16	192.27	195.59	196.65	199.94	198.85
ICPSR 6693	244.96	206.03	209.10	211.36	225.54	220.51
ICPSR 4572 01	255.37	233.49	236.73	235.14	241.25	238.85
Dorothea	253.85	211.18	215.46	216.38	229.44	219.05
Human+Activity+Recognition+Using+Smartphones	197.49	161.66	166.88	165.59	177.18	169.99
ICPSR 31221	171.46	148.77	156.39	156.60	159.97	154.49
ICPSR 3669	210.10	184.69	184.49	188.83	195.67	191.27
ICPSR 2743 Person Level Data	131.26	119.11	119.50	119.87	123.44	120.70

	SfFS	*MOFS*	*Informacijska dobit*	*Omjer dobiti*	*Linearni odabir unaprijed*	*Relief*
ICPSR 2258	231.54	168.94	175.66	171.15	187.95	184.32
Madelon	166.74	135.39	141.27	139.96	149.04	144.55
adult	157.49	123.45	128.75	129.16	139.65	131.28
ICPSR 31202 5	177.73	142.28	145.54	149.79	159.83	155.54
ICPSR 2857	244.65	201.11	214.58	209.74	222.37	223.14
ICPSR 2346	196.37	163.27	165.19	166.25	173.14	171.19
PEMS-SF	758.43	686.17	693.27	687.54	704.55	696.12
Dexter	401.16	357.12	364.29	368.11	381.19	372.54
ICPSR 2686 Caregiver Data	220.15	172.20	178.86	176.63	188.29	184.73
ICPSR 3534	135.83	99.01	104.55	102.39	111.27	108.37
ICPSR 2535	167.38	137.04	139.26	142.27	149.94	145.84
ICPSR 2149	214.19	193.28	195.53	187.74	198.99	191.39
Semeion+Handwritten+Digit	102.28	95.40	90.20	91.50	99.59	89.06
ICPSR 3548	102.25	93.22	95.40	95.65	98.43	97.73
Gisette	152.64	108.25	111.16	105.19	134.49	110.25
ICPSR 2295	255.50	213.47	218.64	219.17	231.87	225.96
ICPSR 2743	151.29	130.59	128.85	132.47	139.65	122.14

	SfFS	*MOFS*	*Informacijska dobit*	*Omjer dobiti*	*Linearni odabir unaprijed*	*Relief*
ICPSR 2163	193.10	175.42	183.76	181.22	186.70	178.85
SECOM	168.54	120.56	125.76	129.98	143.27	129.90
ICPSR 3789	101.34	98.01	99.15	94.12	99.87	97.86
ICPSR 2833	288.51	263.41	269.15	268.64	276.38	266.74
Arcene	364.86	341.75	345.89	346.12	349.99	349.97
ICPSR 2566	131.47	119.24	122.34	124.42	128.75	119.56
ICPSR 31202 4	91.27	77.58	79.97	80.17	85.64	85.94
ICPSR 2039 2	122.35	115.42	116.69	117.74	119.70	111.64
ICPSR 3151	175.68	122.79	128.18	126.64	136.12	134.48
ICPSR 6627	181.29	151.29	157.49	159.12	163.45	156.75

Tablica 20. Adrese skupova podataka

Skup	Izvor
Pittsburgh+Bridges	http://archive.ics.uci.edu/ml/datasets/Pittsburgh+Bridges
Trains	http://archive.ics.uci.edu/ml/datasets/Trains
Balloons	http://archive.ics.uci.edu/ml/datasets/Balloons
Titanic	http://lib.stat.cmu.edu/modules.php?op=modload&name=Downloads&file=index&req=getit&lid=880
broadway	http://lib.stat.cmu.edu/modules.php?op=modload&name=Downloads&file=index&req=getit&lid=880
assessment	http://lib.stat.cmu.edu/modules.php?op=modload&name=Downloads&file=index&req=getit&lid=880
Soybean+Small	http://archive.ics.uci.edu/ml/datasets/Soybean+%28Small%29
molecular biology promoters	http://repository.seasr.org/Datasets/UCI/arff/molecular-biology_promoters.arff
Spectf	http://archive.ics.uci.edu/ml/machine-learning-databases/spect/SPECTF.train
japansolvent	http://lib.stat.cmu.edu/modules.php?op=modload&name=Downloads&file=index&req=getit&lid=880
Post-Operative+Patient	http://archive.ics.uci.edu/ml/datasets/Post-Operative+Patient
hepatitis	http://archive.ics.uci.edu/ml/datasets/Hepatitis
election	http://lib.stat.cmu.edu/modules.php?op=modload&name=Downloads&file=index&req=getit&lid=880
Lung Cancer	http://archive.ics.uci.edu/ml/datasets/Lung+Cancer
sponge	http://archive.ics.uci.edu/ml/datasets/Sponge
creditscore	http://lib.stat.cmu.edu/modules.php?op=modload&name=Downloads&file=index&req=getit&lid=880
bankruptcy	http://lib.stat.cmu.edu/modules.php?op=modload&name=Downloads&file=index&req=getit&lid=880
gviolence	http://lib.stat.cmu.edu/modules.php?op=modload&name=Downloads&file=index&req=getit&lid=880
Labor+Relations	http://archive.ics.uci.edu/ml/datasets/Labor+Relations
Acute+Inflammations	http://archive.ics.uci.edu/ml/datasets/Acute+Inflammations
runshoes	http://lib.stat.cmu.edu/modules.php?op=modload&name=Downloads&file=index&req=getit&lid=880
Cyyoung9302	http://lib.stat.cmu.edu/modules.php?op=modload&name=Downloads&file=index&req=getit&lid=880
impeach	http://lib.stat.cmu.edu/modules.php?op=modload&name=Downloads&file=index&req=getit&lid=880

Skup	Izvor
fraud	http://lib.stat.cmu.edu/modules.php?op=modload&name=Downloads&file=index&req=getit&lid=880
Campus Climate 2011 SJU	http://sociology-data.sju.edu/#mac
homerun	http://lib.stat.cmu.edu/modules.php?op=modload&name=Downloads&file=index&req=getit&lid=880
sonar	http://repository.seasr.org/Datasets/UCI/arff/sonar.arff
bondrate	http://lib.stat.cmu.edu/modules.php?op=modload&name=Downloads&file=index&req=getit&lid=880
ICPSR 3009	http://sociology-data.sju.edu/#mac
gsssexsurvey	http://lib.stat.cmu.edu/modules.php?op=modload&name=Downloads&file=index&req=getit&lid=880
uktrainacc	http://lib.stat.cmu.edu/modules.php?op=modload&name=Downloads&file=index&req=getit&lid=880
ncaa	http://lib.stat.cmu.edu/modules.php?op=modload&name=Downloads&file=index&req=getit&lid=880
credit	http://archive.ics.uci.edu/ml/datasets/Credit+Approval
weights	http://www.blackwellpublishing.com/medicine/bmj/medstats/contents.asp
ICPSR 2743	http://sociology-data.sju.edu/#mac
city	http://sociology-data.sju.edu/#mac
supreme	http://lib.stat.cmu.edu/modules.php?op=modload&name=Downloads&file=index&req=getit&lid=880
ICPSR 2751	http://sociology-data.sju.edu/#mac
blood-transfusion/	http://archive.ics.uci.edu/ml/machine-learning-databases/blood-transfusion/
authorship	http://lib.stat.cmu.edu/modules.php?op=modload&name=Downloads&file=index&req=getit&lid=880
ICPSR 2867	http://sociology-data.sju.edu/#mac
ICPSR 2480	http://sociology-data.sju.edu/#mac
halloffame	http://lib.stat.cmu.edu/modules.php?op=modload&name=Downloads&file=index&req=getit&lid=880
CPS_85_Wages	http://lib.stat.cmu.edu/modules.php?op=modload&name=Downloads&file=index&req=getit&lid=880
Physical+Activity+Monitoring	http://archive.ics.uci.edu/ml/datasets/PAMAP2+Physical+Activity+Monitoring
vote	http://repository.seasr.org/Datasets/UCI/arff/vote.arff
binge	http://sociology-data.sju.edu/#mac
ionosphere	http://repository.seasr.org/Datasets/UCI/arff/ionosphere.arff
ICPSR 2859	http://sociology-data.sju.edu/#mac
Mushroom	http://archive.ics.uci.edu/ml/datasets/Mushroom
ICPSR 2039	http://sociology-data.sju.edu/#mac
Thyroid+Disease	http://archive.ics.uci.edu/ml/datasets/Thyroid+Disease
sick	http://repository.seasr.org/Datasets/UCI/arff/sick.arff

Skup	Izvor
One-hundred+plant+species+leaves+data+set	http://archive.ics.uci.edu/ml/datasets/One-hundred+plant+species+leaves+data+set
Kr-Vs-Kp	http://repository.seasr.org/Datasets/UCI/arff/kr-vs-kp.arff
tic-tac-toe	http://repository.seasr.org/Datasets/UCI/arff/tic-tac-toe.arff
abgss98	http://sociology-data.sju.edu/#mac
ICPSR 2686	http://sociology-data.sju.edu/#mac
ICPSR 2155	http://sociology-data.sju.edu/#mac
heart-statlog	http://archive.ics.uci.edu/ml/datasets/Statlog+%28Heart%29
spambase	http://archive.ics.uci.edu/ml/datasets/Spambase
marketing	http://lib.stat.cmu.edu/modules.php?op=modload&name=Downloads&file=index&req=getit&lid=880
Hill-Valley	http://archive.ics.uci.edu/ml/datasets/Hill-Valley
hepatitis	http://archive.ics.uci.edu/ml/datasets/hepatitis
ICPSR 4291	http://sociology-data.sju.edu/#mac
ICPSR 4582	http://sociology-data.sju.edu/#mac
ICPSR 9595	http://sociology-data.sju.edu/#mac
ICPSR 21600 2	http://sociology-data.sju.edu/#mac
ICPSR 21600 3	http://sociology-data.sju.edu/#mac
ICPSR 28641 2	http://sociology-data.sju.edu/#mac
ICPSR 6542	http://sociology-data.sju.edu/#mac
ICPSR 4367	http://sociology-data.sju.edu/#mac
ICPSR 4572 02	http://sociology-data.sju.edu/#mac
ICPSR 21600 4	http://sociology-data.sju.edu/#mac
DBWorld+e-mails	http://archive.ics.uci.edu/ml/datasets/DBWorld+e-mails
ICPSR 6135	http://sociology-data.sju.edu/#mac
ICPSR 4537 8th form 1	http://sociology-data.sju.edu/#mac
ICPSR 4275	http://sociology-data.sju.edu/#mac
GLI-85	http://featureselection.asu.edu/datasets.php
ICPSR 21600 1	http://sociology-data.sju.edu/#mac
ICPSR 4566 02	http://sociology-data.sju.edu/#mac
ICPSR 8255	http://sociology-data.sju.edu/#mac
ICPSR 28641	http://sociology-data.sju.edu/#mac
SMK-CAN-187	http://featureselection.asu.edu/datasets.php
ICPSR 23041 2	http://sociology-data.sju.edu/#mac
ICPSR 6480	http://sociology-data.sju.edu/#mac
ICPSR 4537 10th form 2	http://sociology-data.sju.edu/#mac

ICPSR 4138	http://sociology-data.sju.edu/#mac
ICPSR 23041	http://sociology-data.sju.edu/#mac
Skup	**Izvor**
ICPSR 4690	http://sociology-data.sju.edu/#mac
ICPSR 4372	http://sociology-data.sju.edu/#mac
ICPSR 20022	http://sociology-data.sju.edu/#mac
ICPSR 6484	http://sociology-data.sju.edu/#mac
ICPSR 4566 01	http://sociology-data.sju.edu/#mac
ICPSR 6693	http://sociology-data.sju.edu/#mac
ICPSR 4572 01	http://sociology-data.sju.edu/#mac
Dorothea	http://www.nipsfsc.ecs.soton.ac.uk/datasets/
Human+Activity+Recognition+Using+Smartphones	http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones
ICPSR 31221	http://sociology-data.sju.edu/#mac
ICPSR 3669	http://sociology-data.sju.edu/#mac
ICPSR 2743 Person Level Data	http://sociology-data.sju.edu/#mac
ICPSR 2258	http://sociology-data.sju.edu/#mac
Madelon	http://www.nipsfsc.ecs.soton.ac.uk/datasets/
adult	http://orange.biolab.si/datasets.psp#datasets
ICPSR 31202 5	http://sociology-data.sju.edu/#mac
ICPSR 2857	http://sociology-data.sju.edu/#mac
ICPSR 2346	http://sociology-data.sju.edu/#mac
PEMS-SF	http://archive.ics.uci.edu/ml/datasets/PEMS-SF
Dexter	http://www.nipsfsc.ecs.soton.ac.uk/datasets/
ICPSR 2686 Caregiver Data	http://sociology-data.sju.edu/#mac
ICPSR 3534	http://sociology-data.sju.edu/#mac
ICPSR 2535	http://sociology-data.sju.edu/#mac
ICPSR 2149	http://sociology-data.sju.edu/#mac
Semeion+Handwritten+Digit	http://archive.ics.uci.edu/ml/datasets/Semeion+Handwritten+Digit
ICPSR 3548	http://sociology-data.sju.edu/#mac
Gisette	http://www.nipsfsc.ecs.soton.ac.uk/datasets/
ICPSR 2295	http://sociology-data.sju.edu/#mac
ICPSR 2743	http://sociology-data.sju.edu/#mac
ICPSR 2163	http://sociology-data.sju.edu/#mac
SECOM	http://archive.ics.uci.edu/ml/datasets/SECOM
ICPSR 3789	http://sociology-data.sju.edu/#mac
ICPSR 2833	http://sociology-data.sju.edu/#mac
Arcene	http://www.nipsfsc.ecs.soton.ac.uk/datasets/
ICPSR 2566	http://sociology-data.sju.edu/#mac
ICPSR 31202 4	http://sociology-data.sju.edu/#mac
ICPSR 2039 2	http://sociology-data.sju.edu/#mac

ICPSR 3151	http://sociology-data.sju.edu/#mac
ICPSR 6627	http://sociology-data.sju.edu/#mac

Dijana Oreški rođena je 1986.g. u Varaždinu. U osnovnoj i srednjoj školi uspješno je sudjelovala na natjecanjima iz matematike, kemije i geografije. Opću gimnaziju završava 2004.g. s prosjekom ocjena 5,0. Iste godine upisuje Fakultet organizacije i informatike (FOI). Tijekom studija radila je kao demonstratorica na četiri kolegija. Godine 2007. radi na FP6 projektu "*We-Go: Enhancing Western Balkan eGovernment Expertise*". Peterostruka je dobitnica Dekanove nagrade:

- **2005.-2008. – četiri nagrade Dekana FOI-a za postignuća tijekom studija, uključujući najveći prosjek ocjena**
- 2008. - priznanje dekana FOI-a za trud i izvrsnost u radu koja je rezultirala i nominacijom za Rektorovu nagradu.

Diplomirala je 2008. godine s temom "Prediktori uspjeha studenata na *Fakultetu organizacije i informatike*" kod prof. dr. sc. Blaženke Divjak te stekla titulu magistra informatike. Od 2009.g. zaposlena je na FOI-u u svojstvu asistenta na Katedri za razvoj informacijskih sustava. Iste godine upisuje poslijediplomski doktorski studij "Informacijske znanosti" na FOI-u. Područje interesa je fokusirano na razvoj tehnika umjetne inteligencije te njihovu primjenu u obrazovanju, upravljanju odnosima s klijentima i za javno dobro. Radi kao istraživač na projektima: *Adaptibilnost visokotehnoloških organizacija* i *Inovativne tehnike rudarenja podataka za analizu složenih društvenih istraživanja.*

U slobodno vrijeme volontira kao predavač u školi informatike za starije osobe te je uključena u rad organizacija koje brinu o poboljšanju uvjeta života mladih. Radila je na dva projekta financirana sredstvima Europske komisije, program *Youth in Action*:

- *Being young in Europe, How to start?*
- *Ivanščica za bolje sutra.*

Popis radova

1. Oreški, S., **Oreški, D.,** Oreški, G., Hybrid System with Genetic Algorithm and Artificial Neural Networks and its Application to Retail Credit Risk Assessment, Expert Systems with Applications 39 (16), 2012., str.12605-12617.
2. Divjak B., **Oreški D.,** Prediction of Academic Performance Using Discriminant Analysis, Proceedings of the ITI 2009, 31st International Conference on INFORMATION TECHNOLOGY INTERFACES, str. 225 -230.
3. **Oreški D**., Peharda P., Application of Factor Analysis In Course Evaluation, Proceedings of the ITI 2008, 30th International Conference on INFORMATION TECHNOLOGY INTERFACES, str. 551 -556.
4. Bambir, D., Horvat, J., **Oreški, D.,** Human Resource Information System in Croatian companies // Proceedings of the 33rd International Conference on Information Technology Interfaces / Lužar-Stiffler, Vesna ; Jarec, Iva ; Bekić, Zoran (ur.). Zagreb: University Computing Centre, University of Zagreb, 2011., str. 415-420.
5. Horvat, J., **Oreški, D.,** Bambir, D., Gender Differences in the Internet Usage among Postgraduate Students // Proceedings of the 33rd International Conference on Information Technology Interfaces / Lužar-Stiffler, Vesna ; Jarec, Iva ; Bekić, Zoran (ur.).Zagreb : University Computing Centre, University of Zagreb, 2011. Str. 281-286.
6. Kliček, B., **Oreški, D.,** Begičević, N.,Temporal Recommender Systems // Recent researches in applied computer and applied computational science / Chen, S ; Mastorakis, Nikos ; Rivas-Echeverria, Francklin ; Mladenov, Valeri (ur.).World Scientific and Egineering Academy and Society (WSEAS) Press, 2011., str. 248-253.
7. **Oreški, D.,**Impact of Data Characteristics on Feature Selection Techniques Performance. // Research papers Faculty of Materials Science and Technology Slovak University of Technology in Trnava. 21, 2013., str. 84-89.
8. **Oreški, D**., Strategy development by using SWOT - AHP. // TEM JOURNAL - Technology, Education, Management, Informatics, 4, 2012., str. 283-288.
9. Kliček, B., **Oreški, D**., Divjak, B., Determining individual learning strategies for students in higher education using neural networks. // International Journal of Arts and Sciences. 3, 2010, 18, str. 22-40.
10. Plantak Vukovac, D., **Oreški, D.,** Active and Collaborative Learning at the University Blended Learning Course // ICERI 2012 Proceedings (5th International Conference of Education, Research and Innovation) / Gómez Chova, L., López Martínez, A., Candel Torres, I. (ur.).Madrid, Spain : International Association of Technology, Education and Development (IATED), 2012. 2221-2230
11. Topolko, K., **Oreški, D.,** Horvat, J., Multi-criteria Modeling of Postgraduate Students Bank Selection // Proceedings of the 23rd Central European Conference on Information and Intelligent Systems / Hunjak, Tihomir ; Lovrenčić, Sandra ; Tomičić, Igor (ur.). Varaždin, 2012, str. 219-226.