

UNIVERZITET SINGIDUNUM
Fakultet za informatiku i menadžment
Danijelova 32, 11000 Beograd

INDUKTIVNO UČENJE RAZUMLJIVOG ZNANJA NA OSNOVU OSKUDNIH OBUČAVAJUĆIH SKUPOVA

Doktorska disertacija

Mentor:
prof. dr Milan Milosavljević

Kandidat:
mr Vladislav Miškovic, dipl. inž.

Beograd, april 2008. godine

Rezime

Osnovna hipoteza ovog rada je da je moguće znatno poboljšati performanse sistema za induktivno učenje pravila u slučaju oskudnih obučavajućih skupova *istovremeno* primenom više različitih metoda i tehnika korišćenja informacija iz prostora atributa: redukcijom dimenzionalnosti, analizom interakcije atributa, upotrebom kombinovanih klasifikatora, globalnih algoritama pretraživanja prostora hipoteza i globalnih funkcija ocene hipoteza.

Naučni doprinosi disertacije su iz oblasti induktivnog učenja (*inductive machine learning*) i sistema za istraživanje znanja (*data mining*):

- Pregled i analiza dosadašnjih istraživanja i rezultata u oblasti induktivnog učenja razumljivog znanja na osnovu oskudnih obučavajućih skupova.
- Pregled i analiza svojstava postojećih sistema, metoda i tehnika induktivnog učenja na ovoj vrsti problema.
- Analiza i predlog jednog rešenja za induktivno učenje razumljivog znanja u obliku pravila na osnovu oskudnih obučavajućih skupova, čija realizacija je izvršena kroz dogradnju sopstvenog sistema za istraživanje znanja *Empiric*.

Daje se pregled performansi dograđenog sistema za induktivno učenje na izabranim primerima oskudnih obučavajućih skupova, standardnim i nekim realnim primerima iz oblasti analize genskih ekspresija.

Abstract

The ultimate analysis supposition of this work is that it is possible to remarkably upgrade inductive rule learning systems performance in the case of sparse learning sets by *simultaneously* applying different methods and techniques of using information from attribute space: dimensionality reduction, interaction analysis, using combined classifiers, global search methods of hypothesis space and global hypothesis quality measures.

Contributions of the dissertation are to inductive machine learning and data mining systems:

- A survey and analysis of recent research and results in the field of inductive machine learning in the case of sparse learning sets.
- A survey and analysis of properties of current systems, methods and techniques for inductive machine learning against this sort of problems.
- An analysis and proposition of a solution for inductive data learning in the form of rules based on sparse learning sets, and its realization is done through an upgrade of an author's data mining system *Empiric*.

An upgraded inductive learning system performance overview for chosen examples of sparse learning sets, standard and some real examples from the domain of gene expression analysis are given.

Ključne reči

- Veštačka inteligencija
- Induktivno učenje
- Istraživanje podataka
- Oskudan skup podataka
- Redukcija dimenzionalnosti
- Selekcija atributa
- Estimacija
- Tačnost predviđanja
- Razumljivost pravila
- Algoritam
- Softver
- Implementacija

Key words

- Artificial intelligence
- Inductive learning
- Data mining
- Sparse dataset
- Dimensionality reduction
- Feature selection
- Estimation
- Predictive accuracy
- Comprehensibility
- Algorithm
- Software
- Implementation

SADRŽAJ

1. UVOD	- 1 -
1.1 PREDMET I CILJEVI ISTRAŽIVANJA	- 4 -
1.2 ZNAČAJ I AKTUELNOST ISTRAŽIVANJA	- 4 -
1.3 METODE ISTRAŽIVANJA	- 5 -
1.4 OSNOVNE I RADNE HIPOTEZE ISTRAŽIVANJA	- 6 -
1.5 OČEKIVANI REZULTATI ISTRAŽIVANJA I NAUČNI DOPRINOS	- 6 -
1.6 PREGLED SADRŽAJA PO POGLAVLJIMA	- 7 -
2. PROBLEM ISTRAŽIVANJA.....	- 9 -
2.1 MODEL INDUKTIVNOG UČENJA	- 9 -
2.1.1 Model učenja do granice.....	- 11 -
2.1.2 PAC model učenja.....	- 11 -
2.1.3 Model egzaktnog učenja.....	- 14 -
2.1.4 Modeli minimalnog kodiranja.....	- 14 -
2.1.5 Modeli učenja na osnovu Bayesove i klasične statističke teorije	- 15 -
2.1.6 Ostali pristupi modeliranju induktivnog učenja.....	- 16 -
2.1.6.1 Metod najbližeg suseda	- 16 -
2.1.6.1 Metod saport vektora.....	- 18 -
2.1.6.3 Metodi grupisanja (clustering)	- 18 -
2.2 ESTIMACIJA TAČNOSTI I OSKUDNI OBUČAVAJUĆI SKUPOVI.....	- 19 -
2.2.1 Estimacija tačnosti klasifikacije metodom testnog uzorka	- 19 -
2.2.2 Estimacija tačnosti klasifikacije metodom unakrsnog ocenjivanja (validacije).....	- 20 -
2.2.3 Estimacija tačnosti klasifikacije metodom početnog punjenja.....	- 20 -
2.2.4 Oskudni obučavajući skupovi.....	- 21 -
2.3 REDUKCIJA DIMENZIONALNOSTI I SELEKCIJA ATRIBUTA	- 22 -
2.4 INTERAKCIJE ATRIBUTA	- 23 -
2.5 GLOBALNE OCENE I GLOBALNE METODE PRETRAŽIVANJA PROSTORA HIPOTEZA	- 25 -
2.6 METOD KOMBINOVANJA KLASIFIKATORA	- 25 -
2.6.1 Usrednjavanje (averaging)	- 26 -
2.6.2 Natrpavanje (bagging).....	- 27 -
2.6.3 Isticanje (boosting).....	- 28 -
2.6.4 Slučajne šume (random forests).....	- 29 -
3. ANALIZA.....	- 31 -
3.1 ALGORITMI UČENJA RAZUMLJIVOG ZNANJA	- 31 -
3.1.1. Metode za indukciju razumljivih opisa.....	- 32 -
3.1.2 Razumljivost pravila u multimodelskim sistemima	- 33 -
3.2 SELEKCIJA ATRIBUTA U ALGORITMIMA UČENJA	- 34 -
3.2.1 Prethodna selekcija atributa	- 35 -
3.2.1.1 Pseudoslučajna selekcija (random subspace method).....	- 35 -
3.2.1.2 Selekcija filtriranjem (filter methods)	- 35 -
3.2.1.3 Selekcija estimacijom tačnosti na validacionom skupu (wrapper methods).....	- 36 -
3.2.2 Selekcija u toku generalizacije (embeeded methods).....	- 36 -
3.3 UPOTREBA INTERAKCIJA U SELEKCIJI ATRIBUTA	- 37 -
3.4 UPOTREBA GLOBALNIH OCENA KVALITETA PRAVILA I GLOBALNIH METODA PRETRAŽIVANJA.....	- 39 -
3.5 PRIMENA METODA KOMBINOVANIH KLASIFIKATORA	- 39 -
3.5.1 Metodi kombinovanja više klasifikatora	- 39 -
1. Metod glasanja (Voting).....	- 40 -
2. Bayesovske metode kombinovanja (Bayesian Ensemble Methods).....	- 41 -
3. Metod kombinovanja Behavior-Knowledge Space	- 41 -
4. Metod kombinovanja na osnovu teorije Dempster-Shafer	- 41 -
3.6 PREGLED POSTOJEĆIH SISTEMA ZA INDUKTIVNOG UČENJE RAZUMLJIVOG ZNANJA	- 42 -
4. REŠENJE.....	- 45 -
4.1 OPIS PODSISTEMA ZA INDUKTIVNO UČENJE SOPSTVENOG SISTEMA EMPIRIC	- 45 -
4.2 REALIZACIJA SELEKCIJE ATRIBUTA PRILIKOM UČENJA KONJUKTIVNIH PRAVILA	- 45 -
4.3 REALIZACIJA GLOBALNE OCENE KVALITETA PRAVILA	- 46 -
4.4 UČENJE KOMBINOVANIH KLASIFIKATORA I SELEKCIJA ATRIBUTA	- 47 -
4.4.1 Realizacija istovremene primene selekcije atributa i formiranja ansambla	- 47 -
4.5 MODIFIKACIJA OSNOVNOG ALGORITMA PRETRAŽIVANJA.....	- 47 -

5. PRIMERI PRIMENE	- 49 -
5.1 OPIS METODOLOGIJE IZVOĐENJA EKSPERIMENATA	- 49 -
5.1.1 Model eksperimenta	- 49 -
5.1.2 Estimacija tačnosti i razumljivosti	- 49 -
5.1.3 Redukcija dimenzionalnosti	- 50 -
5.1.3.1. Metod Filtriranja (<i>filter</i>)	- 51 -
5.1.3.2. Metod prethodnog učenja (<i>wrapper</i>)	- 51 -
5.1.3.3. Ugrađeni metodi (<i>embeeded</i>)	- 51 -
5.1.4 Upotrebljeni programi	- 52 -
5.1.4.1 Sistem Empiric	- 52 -
5.1.4.2 Sistem WEKA	- 55 -
5.1.4.3 Softver za Matlab	- 56 -
5.1.4.4 RandomForest (Breiman-Cutler)	- 56 -
5.2 OPIS IZABRANIH PROBLEMA INDUKTIVNOG UČENJA	- 57 -
5.2.1 Konstruisani merni problemi	- 57 -
5.2.2 UCI problemi, evidentno oskudni	- 58 -
5.2.3 Problemi klasifikacije slučajeva na osnovu genskih ekspresija	- 58 -
5.3 REZULTATI UČENJA I ESTIMACIJE PERFORMANSI NAUČENOG ZNANJA	- 59 -
5.3.1 Konstruisani merni problemi	- 59 -
5.3.2 Merni problemi iz baze podataka UCI	- 70 -
5.3.3 Problemi klasifikacije slučajeva na osnovu genskih ekspresija	- 81 -
5.4 DISKUSIJA REZULTATA	- 96 -
5.4.1 Tačnost predviđanja	- 96 -
Problemi učenja na osnovu genskih ekspresija	- 102 -
5.4.2 Razumljivost	- 102 -
6. ZAKLJUČCI	- 106 -
7. LITERATURA	- 107 -
8. PRILOZI	- 114 -
PRILOG 1: UPUTSTVO ZA KORIŠĆENJE SISTEMA <i>EMPIRIC</i>	- 115 -
PRILOG 2: PREGLED UPOTREBLJENOG SOFTVERA IZ SISTEMA <i>WEKA</i>	- 119 -
PRILOG 3: PREGLED UPOTREBLJENOG SOFTVERA ZA SISTEM <i>MATLAB</i>	- 122 -
PRILOG 4: SPECIFIKACIJA UPOTREBLJENIH OBUČAVAJUĆIH SKUPOVA	- 124 -
(A) Primeri veštački kreiranih problema učenja:	- 124 -
1. Problem Monks-2	- 124 -
2. Problem Waveform-40	- 124 -
3. Problem Breiman 1000	- 124 -
(B) Primeri oskudnih obučavajućih skupova iz baze podataka UCI	- 126 -
4. Breast Cancer LJ	- 126 -
5. Cleveland Heart Disease	- 126 -
6. Lung Cancer	- 126 -
(C) Primeri genskih ekspresija [<i>Pathwork Diagnostics,2007</i>]:	- 127 -
7. BC (175 x 22215 x 2)	- 127 -
8. RFS5 (441 x 22215 x 2)	- 127 -
9. NKI BC (182 x 22223 x 2)	- 127 -

Popis tabela

<i>Tabela 1: Pregled metoda estimacije tačnosti predviđanja.....</i>	<i>- 21 -</i>
<i>Tabela 2: Osnovne karakteristike konstruisanih problema učenja</i>	<i>- 58 -</i>
<i>Tabela 3: Osnovne karakteristike standardnih problema učenja iz baze UCI.....</i>	<i>- 58 -</i>
<i>Tabela 4: Osnovne karakteristike problema učenja na osnovu genskih ekspresija.....</i>	<i>- 59 -</i>
<i>Tabela 5: Tačnost predviđanja za problem Monks-2 (Empiric).....</i>	<i>- 60 -</i>
<i>Tabela 6: Tačnost predviđanja za problem Monks-2 (WEKA).....</i>	<i>- 61 -</i>
<i>Tabela 7: Tačnost predviđanja za problem Waveform-40 (Empiric).....</i>	<i>- 64 -</i>
<i>Tabela 8: Tačnost predviđanja za problem Waveform-40 (WEKA).....</i>	<i>- 65 -</i>
<i>Tabela 9: Tačnost predviđanja za problem Breiman-1K (Empiric)</i>	<i>- 68 -</i>
<i>Tabela 10: Tačnost predviđanja za problem Breiman-1K (WEKA)</i>	<i>- 69 -</i>
<i>Tabela 11: Tačnost predviđanja za problem Breast Cancer LJ (Empiric).....</i>	<i>- 72 -</i>
<i>Tabela 12: Tačnost predviđanja za problem Breast Cancer LJ (WEKA).....</i>	<i>- 73 -</i>
<i>Tabela 13: Tačnost predviđanja za problem Cleveland Heart (Empiric)</i>	<i>- 75 -</i>
<i>Tabela 14: Tačnost predviđanja za problem Cleveland Heart (WEKA)</i>	<i>- 76 -</i>
<i>Tabela 15: Tačnost predviđanja za problem Lung Cancer (Empiric)</i>	<i>- 79 -</i>
<i>Tabela 16: Tačnost predviđanja za problem Lung Cancer (WEKA)</i>	<i>- 80 -</i>
<i>Tabela 17: Tačnost predviđanja za problem BC (Empiric).....</i>	<i>- 84 -</i>
<i>Tabela 18: Tačnost predviđanja za problem BC (WEKA).....</i>	<i>- 85 -</i>
<i>Tabela 19: Tačnost predviđanja za problem RFS5 (Empiric)</i>	<i>- 89 -</i>
<i>Tabela 20: Tačnost predviđanja za problem RFS5 (WEKA)</i>	<i>- 90 -</i>
<i>Tabela 21: Tačnost predviđanja za problem NKI BC (Empiric).....</i>	<i>- 93 -</i>
<i>Tabela 22: Tačnost predviđanja za problem NKI BC (WEKA).....</i>	<i>- 94 -</i>
<i>Tabela 23: Povećanje tačnosti uz upotrebljene metode selekcije atributa</i>	<i>- 96 -</i>
<i>Tabela 24: Povećanje tačnosti primenom metoda učenja ansambala.....</i>	<i>- 97 -</i>
<i>Tabela 25: Povećanje tačnosti istovremenom primenom metoda selekcije atributa i učenja do 50 ansambala</i>	<i>- 100 -</i>
<i>Tabela 26: Poređenje najvećeg povećanja tačnosti istovremenom primenom metoda selekcije atributa i učenja do 50 ansambala.....</i>	<i>- 101 -</i>

Popis slika

Slika 1: Formalna definicija problema induktivnog učenja	10 -
Slika 2: Korelacija međusobno nezavisnih atributa s konceptom.....	24 -
Slika 3: Korelacija međusobno zavisnih atributa s konceptom	24 -
Slika 4: Odnos tačnosti predviđanja i složenosti naučenog koncepta.....	32 -
Slika 5: Metod selekcije atributa filtriranjem	36 -
Slika 6: Metod selekcije prethodnim učenjem.....	36 -
Slika 7: Selekcija atributa u sistemu Empiric	46 -
Slika 8: Realizacija globalne ocene kvaliteta pravila Relief	46 -
Slika 9: Metod dvostruke unakrsne validacije prilikom prethodne selekcije atributa..	49 -
Slika 10: Izgled ekrana sistema Empiric prilikom analize problema NKI BC.....	54 -
Slika 11: Pokretanje induktivnog učenja pravila (Empiric.Rules)	55 -
Slika 12: Postavljanje parametara induktivnog učenja pravila.....	55 -
Slika 13: Uticaj broja selektovanih atributa na tačnost predviđanja (Monks-2)	60 -
Slika 14: Prikaz tačnosti predviđanja ispitivanih metoda učenja i selekcije atributa iz sistema EMPIRIC za problem Monks-2.....	61 -
Slika 15: Prikaz tačnosti predviđanja ispitivanih metoda učenja i selekcije atributa iz sistema WEKA za problem Monks-2	62 -
Slika 16: Uticaj broja selektovanih atributa na tačnost predviđanja (Waveform-40)...	63 -
Slika 17: Raspodela ocena relevantnosti atributa metodima IngoGain, Relief i Simba (Waveform-40)	63 -
Slika 18: Prikaz tačnosti predviđanja ispitivanih metoda učenja i selekcije atributa iz sistema EMPIRIC za problem Waveform-40.....	64 -
Slika 19: Prikaz tačnosti predviđanja ispitivanih metoda učenja i selekcije atributa iz sistema WEKA za problem Waveform-40	65 -
Slika 20: Uticaj broja selektovanih atributa na tačnost predviđanja (Breiman-1K).....	66 -
Slika 21: Raspodela ocena relevantnosti atributa metodima IngoGain, Relief i Simba i prikaz međusobne korelacije (Breiman-1K).....	67 -
Slika 22: Prikaz tačnosti predviđanja ispitivanih metoda učenja i selekcije atributa iz sistema EMPIRIC za problem Breiman-1K	68 -
Slika 23: Prikaz tačnosti predviđanja ispitivanih metoda učenja i selekcije atributa iz sistema WEKA za problem Breiman-1K.....	69 -
Slika 24: Uticaj broja selektovanih atributa na tačnost predviđanja (Breast Cancer LJ)-	70 -
Slika 25: Raspodela ocena relevantnosti atributa metodima InfoGain, Relief i Simba (Breast Cancer LJ)	71 -
Slika 26: Stabilnost estimacije broja atributa (Breast Cancer LJ)	71 -

Slika 27: Prikaz tačnosti predviđanja ispitivanih metoda učenja i selekcije atributa iz sistema EMPIRIC za problem Breast Cancer LJ.....	- 72 -
Slika 28: Prikaz tačnosti predviđanja ispitivanih metoda učenja i selekcije atributa iz sistema WEKA za problem Breast Cancer LJ.....	- 73 -
Slika 29: Uticaj broja selektovanih atributa na tačnost predviđanja (Cleveland Heart)-	74 -
Slika 30: Raspodela ocena relevantnosti atributa metodima IngoGain, Relief i Simba (Cleveland Heart).....	- 74 -
Slika 31: Stabilnost estimacije broja atributa (Cleveland Heart).....	- 75 -
Slika 32: Prikaz tačnosti predviđanja ispitivanih metoda učenja i selekcije atributa iz sistema EMPIRIC za problem Cleveland Heart	- 76 -
Slika 33: Prikaz tačnosti predviđanja ispitivanih metoda učenja i selekcije atributa iz sistema WEKA za problem Cleveland Heart	- 77 -
Slika 34: Uticaj broja selektovanih atributa na tačnost predviđanja (Lung Cancer)	- 78 -
Slika 35: Raspodela ocena relevantnosti atributa metodima IngoGain, Relief i Simba (Lung Cancer).....	- 78 -
Slika 36: Stabilnost estimacije broja atributa (Lung Cancer).....	- 79 -
Slika 37: Prikaz tačnosti predviđanja ispitivanih metoda učenja i selekcije atributa iz sistema EMPIRIC za problem Lung Cancer	- 80 -
Slika 38: Prikaz tačnosti predviđanja ispitivanih metoda učenja i selekcije atributa iz sistema WEKA za problem Lung Cancer.....	- 81 -
Slika 39: Uticaj broja selektovanih atributa na tačnost predviđanja (Breast Cancer) ..	- 82 -
Slika 40: Raspodela ocena relevantnosti atributa metodima InfoGain, Relief i Simba i prikaz međusobne korelacije (Breast Cancer)	- 83 -
Slika 41: Poređenje raspodela ocena relevantnosti atributa metodima Relief i Simba (Breast Cancer)	- 83 -
Slika 42: Stabilnost estimacije broja atributa (Breast Cancer)	- 84 -
Slika 43: Prikaz tačnosti predviđanja ispitivanih metoda učenja i selekcije atributa iz sistema EMPIRIC za problem BC	- 85 -
Slika 44: Prikaz tačnosti predviđanja ispitivanih metoda učenja i selekcije atributa iz sistema WEKA za problem BC	- 86 -
Slika 45: Uticaj broja selektovanih atributa na tačnost predviđanja (RFS5).....	- 87 -
Slika 46: Raspodela ocena relevantnosti atributa metodima InfoGain, Relief i Simba i prikaz međusobne korelacije (Breast Cancer)	- 87 -
Slika 47: Poređenje raspodela ocena relevantnosti atributa metodima Relief i Simba (RFS5).....	- 88 -
Slika 48: Stabilnost estimacije broja atributa (RFS5).....	- 88 -
Slika 49: Prikaz tačnosti predviđanja ispitivanih metoda učenja i selekcije atributa iz sistema EMPIRIC za problem RFS5	- 89 -
Slika 50: Prikaz tačnosti predviđanja ispitivanih metoda učenja i selekcije atributa iz sistema WEKA za problem RFS5	- 90 -

Slika 51: Uticaj broja selektovanih atributa na tačnost predviđanja (NKI BC)	91 -
Slika 52: Raspodela ocena relevantnosti atributa metodima InfoGain, Relief i Simba i prikaz međusobne korelacije (NKI BC)	92 -
Slika 53: Poređenje raspodela ocena relevantnosti atributa metodima InfoGain, Relief i Simba (NKI BC)	92 -
Slika 54: Stabilnost estimacije broja atributa (NKI BC)	93 -
Slika 55: Prikaz tačnosti predviđanja ispitivanih metoda učenja i selekcije atributa iz sistema EMPIRIC za problem NKI BC	94 -
Slika 56: Prikaz tačnosti predviđanja ispitivanih metoda učenja i selekcije atributa iz sistema WEKA za problem NKI BC	95 -
Slika 57: Prikaz odnosa najveće tačnosti predviđanja za sve upotrebljene metode ...	101 -
Slika 58: Prikaz tačnosti predviđanja algoritma 1-NN za problem BC uz prethodnu selekciju metodom SIMBA	102 -
Slika 59: Primer naučenog skupa pravila za problem RFS5 (Empiric)	103 -
Slika 60: Primer naučenog stabla za problem RFS5 (WEKA.C45)	103 -
Slika 61: Primer naučenog skupa pravila za problem RFS5 (WEKA.C45Rules)	104 -
Slika 62: Primer naučenog skupa pravila za problem RFS5 (WEKA.RIPPER)	104 -

1. UVOD

Indukcija predstavlja izvođenje opštih zakonitosti na osnovu uvida u konkretne pojave, odnosno slučajeve. Predstavlja važan element ljudskog mišljenja, ali i naučnog metoda, procesa kojim se dolazi do *naučnih* saznanja o određenim pojavama. Naučni metod se sastoji od prikupljanja podataka putem posmatranja i eksperimentisanja, formulisanja pretpostavki ili hipoteza, te njihove provere u praksi putem eksperimenata.

Istinitost induktivnih tvrdnji se ne može formalno dokazati, već samo opovrgnuti, što su uočili još antički mislioci. Aristotel je razlikovao aproksimativno i egzaktno zaključivanje i razdvojio abduktivno, deduktivno i induktivno zaključivanje, kao i složene forme zaključivanja kao što je analogija.

U novije doba, značaj indukcije je istakao *Fransis Bacon* (1561-1626.), koji je kritikovao nauku svog vremena, smatrajući da predstavlja samo uređen sistem postojećeg znanja, da nije usmerena ka novim otkrićima i metodima otkrivanja. Takođe je ukazao na svojstvo ljudskog mišljenja da zanemaruje negativne primere i ukazao na opasnost od preterane generalizacije u induktivnom zaključivanju.

Bacon je bio uveren da se proces razumevanja može "izvršiti kao pomoću mašine" [3], ali je prve praktične korake ka tom cilju napravio *John Stuart Mill* (1806-1873.), koji je smatrao da je obezbeđenje modela i pravila induktivnog zaključivanja zadatak logike i predložio četiri poznata metoda indukcije:

1. *Metod podudaranja*: ako dva ili više primera fenomena koji istražujemo imaju samo jedan zajednički faktor, tada je faktor u kome se svi primeri podudaraju uzrok datog fenomena;
2. *Metod razlikovanja*: ako pozitivna instanca fenomena koji istražujemo i negativna instanca fenomena imaju iste sve okolnosti osim jedne, ta okolnost je efekat ili uzrok ili nužni deo uzroka fenomena koji razmatramo;
3. *Metod ostatka*: kada se iz fenomena ukloni svaki deo za koji se zna da je posledica nekih preduslova, tada je ostatak pojave posledica preostalih uzroka;
4. *Metod istovremenih varijacija*: ako jedan fenomen varira pravilno na neki način uvek kada drugi fenomen varira na neki poseban način, prvi je povezan sa drugim kroz neki lanac uzročnosti.

Prva tri zakona se odnose na nominalne (kategoričke) podatke, dok je četvrto pravilo primenjivo na podatke prikazane na numeričkoj skali. Četvrti zakon induktivnog mišljenja je u stvari opštiji i obuhvata prva tri kao specijalne slučajeve.

Neki savremeni filozofi, kao *Karl Popper* (1902-1994.), negirali su postojanje mehanizma sličnog indukciji u živom svetu, već smatraju da se znanje dobija stvaranjem pretpostavki i odbacivanjem onih koji nisu saglasne sa podacima, odnosno kao pogađanje i eliminisanje greške.

Filozofska razmatranja se uglavnom odnose na proces otkrivanja naučnih i univerzalnih istina i zanemaruju veliki deo ljudskog znanja iz svakodnevnog života. Ovo znanje i iskustvo je fragmentarno, neprecizno i često formulisano uz unapred poznate kontraprimere.

Savremeni filozof *Bertrand Russell* (1872-1970.) je smatrao da, ako se ne veruje u indukciju, ne veruje se ni u šta. Smatrao je da je induktivni princip je u suštini probabilistički i skrenuo pažnju na značaj statističkog zaključivanja.

Važan aspekt modeliranja učenja istakao je i filozof *Ludwig Wittgenstein* (1889-1951.), koji je ukazivao na značaj jednostavnosti i smatrao da je zadatak indukcije da prihvati kao istinitu najjednostavniju od postojećih zakonitosti, pošto postoji mnogo generalizacija konzistentnih sa iskustvom, kao što su npr. Ptolomejeva i (jednostavnija) Kopernikova kosmologija.

Ovaj *princip jednostavnosti* se u savremenoj literaturi često naziva "Okamova oštrica" (*Occam's razor*), po filozofu sholastičaru Vilijamu od Okama (*William of Ockham*, 1288-1347.) [2], [4]. On je suprotan antičkom *principu indiferentnosti* ili Epikurovom principu [1], [2], po kome treba ravnopravno tretirati sva objašnjenja neke pojave koja su u saglasnosti sa iskustvom, dok se ne pojave naučni razlozi da se usvoji jedno od njih.

Problem konstrukcije mehanizma koji bi mogao da kreira opšte zakonitosti na osnovu primera konkretizovan je pojavom računara, a kao jednako važan je postavljen i problem ocene valjanosti tih zakonitosti u praktičnim situacijama.

Kao odgovor na pitanje da li računar može da izvršava funkcije koje smatramo mišljenjem, *Alan Turing* (1912-1954.) je predložio eksperimentalni pristup, tzv. Tjuringov test: odgovor je potvrđan, ako čovek u komunikaciji s računarom i drugim čovekom nije u stanju da ih razlikuje.

Izučavanje i računarsko modeliranje procesa učenja je predmet proučavanja na području *mašinskog učenja*.

Prema [5] mašinsko učenje (*machine learning*) je skup procesa, gde spadaju: prikupljanje novog deklarativnog znanja, razvoj i usavršavanje motornih i saznavnih sposobnosti kroz praksu, strukturisanje postojećeg znanja i otkrivanje novih činjenica i teorija posmatranjem i aktivnim eksperimentisanjem.

Učenje se može posmatrati kroz dve osnovne forme:

- prikupljanje znanja (*knowledge acquisition*), koje predstavlja učenje nove, simboličke informacije, tako da se ona može efektivno primeniti (tako čovek uči teorijska znanja, npr. fiziku);
- uvežbavanje (*training*), koje podrazumeva poboljšavanje nekog stečenog znanja, mentalne ili motorne koordinacije, kroz praktično ponavljanje i korekciju odstupanja od željenog ponašanja (tako čovek uči različite veštine - vožnju bicikla ili sviranje na klaviru, pri čemu prvu fazu učenja predstavlja prikupljanje znanja).

Smatra se da učenje kod čoveka predstavlja mešavinu obeju formi, s tim da mentalne aktivnosti potenciraju prvu formu, a motorne aktivnosti u većoj meri drugu formu učenja.

Sistemi mašinskog učenja se najčešće dele prema odabranoj strategiji učenja, načinu predstavljanja znanja i području primene [5].

Podela prema *strategiji učenja* se odnosi na potrebnu količinu zaljučivanja sistema (u ulozi učenika) nad podacima koje dobija iz okruženja (u ulozi učitelja):

- učenje memorisanjem (*rote learning*), kada nema nikakvog zaključivanja ni transformacije znanja (klasično programiranje i upotreba baza podataka)

- učenje na osnovu rečenog (*learning by being told*), kada se deklarativno znanje transformiše u neku internu formu predstavljanja i integriše sa postojećim znanjem. Takvo znanje sistem zna da upotrebljava bez dodatnog programiranja.
- učenje po analogiji (*learning by analogy*), kada se zahteva veći obim zaključivanja, jer se novo znanje stiče transformacijom i proširivanjem postojećeg znanja u takvom obliku da se može upotrebiti za rešavanje novih problema, koji su u određenoj meri slični već rešenim.
- učenje na osnovu primera (*learning by examples*), koje zahteva induktivno zaključivanje. Analizom i generalizacijom rešenih primera i kontraprimera neke klase pojava (pojma, koncepta), dolazi se do pravila, teorije ili opisa pojma, koji objašnjava sve primere i nijedan kontraprimer. Ovakve metode učenja se dalje mogu klasifikovati prema izboru primera, izvoru i načinu upotrebe primera.
- učenje posmatranjem i samostalnim otkrivanjem (*learning by observation and discovery*), ili učenje bez učitelja, zahteva najveću količinu zaključivanja, jer sistem mora samostalno da otkriva nove i značajne klase objekata (pojmove, koncepte), postavlja hipoteze i proverava ih, te stvara teorije.

Pojam *induktivno mašinsko učenje* obuhvata poslednje dve strategije učenja: učenje na osnovu primera i učenje posmatranjem i samostalnim otkrivanjem.

Induktivno (mašinsko, automatizovano) učenje se može posmatrati kao proces u kome sistem poboljšava svoje performanse na datom zadatku bez dodatnog programiranja [6].

Postoji i podela sistema mašinskog učenja prema *vrsti naučenog znanja*. Empirijski je poznato da postoje znanja koja se ne mogu eksplicitno izraziti [6], pa se mogu posmatrati sistemi sa *eksplicitnim znanjem*, koje se predstavlja matematičkom logikom, produkcionim pravilima, frejmovima i sličnim načinima i sistemi sa *implicitnim (distribuiranim) znanjem*, kao što je znaje predstavljeno u obliku neuronskih mreža.

Istraživanje opisano u ovom radu se odnosi na oblast induktivnog učenja razumljivog znanja, odnosno učenja konceptata u različitim formama predstavljanja, najčešće u obliku stabala odlučivanja (*decision trees*), lista odlučivanja (*decision lists*) i produkcionih pravila (*production rules*).

Induktivno učenje razumljivog znanja ili učenje konceptata (hipoteza, modela, klasifikatora, pravila) bavi se indukcijom logičkih zakonitosti koje treba da budu razumljive čoveku. Najviši stepen razumljivosti omogućava korišćenje naučenog znanja bez pomoći računara [7].

Pošto je potrebna i ocena valjanosti naučenog znanja, često se skup primera koji služe za induktivno učenje deli na dva dela: obučavajući skup (*learning set*), koji se koristi za učenje i testni skup (*test set*), koji se koristi za testiranje naučenog znanja.

Osnovna mera uspešnosti naučenog (razumljivog) znanja je *tačnost predviđanja (predictive accuracy)*, koja predstavlja procenat uspešnosti klasifikacije novih, nerazmatranih primera korišćenjem i naučenih pravila.

U ovom radu će se smatrati da je obučavajući skup *oskudan (sparse)* ako induktivno naučeno znanje ne omogućava dovoljno uspešno razlikovanje novih primera [8].

Najčešći opšti razlog ove oskudnosti u procesu induktivnog učenja je visoka dimenzionalnost prostora hipoteza, odnosno prevelik broj atributa (osobina, svojstava) u modelu problema u odnosu na raspoloživ broj obučavajućih primera. Sličan problem u statistici je problem malog uzorka (*small sample problem*).

Dodatne informacije neophodne za poboljšanje performansi induktivno naučenog znanja mogu se pronaći analizom prostora atributa, koji mogu biti međusobno u različitim relacijama.

Predviđeni atributi u modelu problema mogu biti *irrelevantni* za sam problem učenja i tada se njihov veliki broj može posmatrati kao smetnja ili šum u podacima, koji prikriva stvarne zakonitosti koje treba ustanoviti metodama induktivnog učenja.

Redukcijom dimenzionalnosti prostora atributa, naročito eliminacijom irrelevantnih, kao i *redundantnih* atributa, načelno je moguće postići poboljšanje performansi naučenog znanja i kompenzovati oskudnost obučavajućeg skupa.

Drugi pristup poboljšanju performansi je upotreba efikasnijih algoritama učenja, primenom adekvatnijih metoda pretraživanja prostora hipoteza i različitih metoda učenja redundantnih skupova pravila, zajedno sa tehnikama njihovog kombinovanja u zadacima klasifikacije novih primera.

Rezultati različitih pristupa se mogu kombinovati pogodnim metodama, tako da se mogu postići značajna ukupna poboljšanja performansi.

1.1 PREDMET I CILJEVI ISTRAŽIVANJA

Predmet rada je realizacija sistema za induktivno učenje pravila na osnovu primera koji je pogodan za oskudne obučavajuće skupove.

U radu je trebalo rešiti problem identifikacije i implementacije pogodnih metoda i tehnika za poboljšanje performansi algoritama induktivnog učenja pravila za oskudne obučavajuće skupove, gde postojeći algoritmi ne daju praktično upotrebljive rezultate.

Cilj istraživanja je identifikacija, implementacija i eksperimentalna provera metoda i tehnika koje su posebno pogodne za induktivno učenje koncepata na osnovu veoma malog broja primera opisanih velikim brojem atributa, čija relevantnost i međusobna zavisnost nije unapred poznata.

1.2 ZNAČAJ I AKTUELNOST ISTRAŽIVANJA

Induktivno učenje se u mnogim savremenim primenama suočava sa problemom velike složenosti modela problema učenja, gde je broj atributa daleko veći od raspoloživog broja obučavajućih primera.

Jedna od tipičnih primena je učenje pravila za dijagnostiku različitih oboljenja na osnovu genskih ekspresija, gde je na raspolaganju veoma mali broj primera (slučajeva), reda desetine ili stotine, opisanih veoma velikim brojem obeležja (ekspresija gena), kojih može biti na desetine i stotine hiljada [9], [10].

Postojeći algoritmi učenja ne daju praktično upotrebljive rezultate, jer je ocena tačnosti predviđanja ovako naučenog znanja na novim primerima (slučajevima)

nezadovoljavajuća i nije dovoljna za eventualnu primenu ovako automatizovanih sistema u kliničkoj praksi.

Klasični algoritmi induktivnog učenja i različite postojeće metode poboljšanja performansi učenja za ovakve probleme (npr. prethodna selekcija atributa i kombinovani modeli) koriste se nezavisno ili se njihove interakcije na različitim problemima i metode primene još intenzivno istražuju.

U ovom istraživanju će se predložiti rešenja koja značajno unapređuju performanse učenja razumljivog znanja u sličnim primenama.

1.3 METODE ISTRAŽIVANJA

U radu su korišćene sledeće metode istraživanja:

- *metod deskripcije* - proučavanje i eksplicitno definisanje problema induktivnog učenja na osnovu malog broja primera u složenim domenima;
- *metod modelovanja* - definisanje više partikularnih modela rešenja problema i integrisanog predloga modela;
- *metod opservacije* - testiranje varijanti sistema na izabranim primerima (obučavajućim skupovima) i ocena pogodnosti primene na ovu vrstu problema.

U toku realizacije istraživanja izvršene su sledeće analize:

- *analiza sistema* induktivnog učenja na osnovu malog broja primera u primera u složenim domenima;
- *analiza metoda i tehnika* pogodnih za poboljšanje rezultata učenja klasičnih algoritama za induktivno učenje pravila na osnovu oskudnih obučavajućih skupova, između ostalog:
 - interakcije atributa radi upravljanja selekcijom u prostoru atributa;
 - multimodelskih metoda i tehnika učenja;
 - globalnog pretraživanja prostora hipoteza;
- *analiza uticaja* istovremene primene više metoda i tehnika, uz predlog perspektivnog rešenja;
- *komparativna analiza* performansi sistema induktivnog učenja na ovoj vrsti problema sa i bez dodatnih komponenti.

Upotrebljeni su sledeći instrumenti istraživanja:

- metode i tehnike induktivnog učenja razumljivog znanja: simboličke metode učenja stabala i pravila, kao i metode učenja kombinovanih klasifikatora;
- metode estimacije tačnosti naučenih klasifikatora: metoda testnog skupa, metoda unakrsne validacije;
- softverski alati:
 - sopstveni sistem za istraživanje znanja *Empiric*, sa sistemom za induktivno učenje pravila *Empiric.Rules* (magistarski rad autora, [11]) na kome se praktično programski realizuju i ispituju neke od predloženih metoda i tehnika.

- sistem za istraživanje znanja *WEKA* (*University of Waikato*) [4], sa velikim brojem implementiranih algoritama za induktivno učenje, analizu i redukciju dimenzionalnosti prostora atributa.
- sistem *Matlab* (*MathWorks*), sa bibliotekama i javno publikovanim programima [12],[13], [14] u kojima su implementirane brojne metode selekcije atributa i algoritmi induktivnog učenja.
- originalna realizacija algoritma *RandomForest* (*University of California at Berkeley*) autora *L. Breiman* i *A. Cutler* [15], [16], [17].

1.4 OSNOVNE I RADNE HIPOTEZE ISTRAŽIVANJA

Osnovna hipoteza ovog rada je da je moguće znatno poboljšati performanse sistema za induktivno učenje pravila u slučaju oskudnih obučavajućih skupova istovremenom primenom više različitih metoda i tehnika korišćenja informacija iz prostora atributa: redukcijom dimenzionalnosti, analizom interakcija atributa, upotrebom kombinovanih klasifikatora, globalnih algoritama pretraživanja prostora hipoteza i globalnih funkcija ocene hipoteza.

Upotrebljene radne hipoteze su:

1. Na osnovu postojećih rezultata na području induktivnog učenja na osnovu primera, moguće je identifikovati skup metoda čija istovremena primena značajno poboljšava rezultate induktivnog učenja pravila u slučaju oskudnih obučavajućih skupova.
2. Implementacijom identifikovanih metoda, zasebno ili u okviru sistema za induktivno učenje, moguće je značajno poboljšati tačnost postojećih metoda učenja na osnovu primera.
3. Postojeći sistemi za induktivno učenje pravila mogu se na više načina poboljšati, odnosno dopuniti predloženim metodima, čime se unapređuje njihova praktična primenljivost na oskudne obučavajuće skupove.
4. Predloženi metodi značajno poboljšavaju performanse učenja sistema za indukciju pravila na području genetike, čime se unapređuje učenje simboličkog, razumljivog znanja na ovom području.

1.5 OČEKIVANI REZULTATI ISTRAŽIVANJA I NAUČNI DOPRINOS

Osnovni cilj istraživanja je da potvrdi osnovnu hipotezu da je istovremenom primenom više različitih metoda i tehnika korišćenja postojećih informacija iz prostora atributa moguće znatno poboljšati performanse sistema za induktivno učenje pravila kada je obučavajući skup oskudan.

U ovoj disertaciji su očekivani sledeći rezultate:

1. Eksperimentalna potvrda da kombinovani algoritam formiranja ansambala, uz istovremene pseudoslučajne modifikacije prostora atributa, drastično smanjuje broj razmatranih atributa, uz istovremeno povećanje tačnosti, te je pogodan za

induktivno učenje konjuktivnih pravila i u slučaju oskudnih obučavajućih skupova.

2. Da će se heuristike za ocenu interakcije parova atributa u problemu učenja moći uspešno primeniti u selekciji atributa i, kombinovanjem sa pseudoslučajnom selekcijom atributa, u praksi poboljšati performanse naučenog znanja.
3. Realizacija algoritma globalnog pretraživanja *Relief* u okviru sopstvenog sistema *Empiric* i ispitivanje njegovog uticaja na učenje u slučaju oskudnih obučavajućih skupova za više različitih tipova atributa, uključujući strukturne attribute.
4. Da će u okviru sopstvenog sistema *Empiric* realizovati modifikaciju osnovnog usmerenog pretraživanja prostora hipoteza (*greedy search*), tako da koristi ocenu interakcije parova atributa za upravljanje pretraživanjem.

1.6 PREGLED SADRŽAJA PO POGLAVLJIMA

U uvodnom **poglavlju 1.** daje se kratak uvod u područje, opis najvažnijih pojmova i problema koji se istražuje, kao i prikaz mogućih rešenja.

U **poglavlju 2.** se daje detaljan pregled koncepata i tehnologija koji će se koristiti u samom radu. Definišu se osnovni pojmovi i detaljno opisuje problem induktivnog učenja na osnovu oskudnih obučavajućih skupova, kao i najvažniji dosadašnji rezultati.

Identifikuju se metode i tehnike kojima se mogu poboljšati performanse učenja, pre svega tačnost i razumljivost naučenog znanja.

U **poglavlju 3.** je izvršena analiza postojećih sistema za induktivno učenje i njihova pogodnost za induktivno učenje u slučaju oskudnih obučavajućih skupova.

Daje se pregled algoritama učenja razumljivog znanja, uloga selekcije relevantnih atributa u algoritmima učenja u slučaju oskudnosti obučavajućeg skupa i pregled postojećih rešenja. Razmatra se upotreba globalnih ocena kvaliteta pravila i globalnih metoda pretraživanja u algoritmima učenja. Diskutuje se primenjivost i ograničenja metoda kombinovanih klasifikatora u učenju razumljivog znanja.

Na kraju poglavlja se daje kraći pregled postojećih sistema za induktivnog učenje razumljivog znanja.

U **poglavlju 4.** se daje predlog i implementacija rešenja u obliku unapređenog sopstvenog sistema *Empiric*, koje koristi više razmatranih metoda i tehnika istovremeno.

Implementirani su metodi koji imaju sinergijski efekt i značajno doprinose poboljšanju rezultata učenja na osnovu oskudnih obučavajućih skupova.

Opisuje se podsistem za induktivno učenje sopstvenog sistema *Empiric* i realizacija unapređenih komponenti: selekcije atributa prilikom učenja konjuktivnih pravila, globalne ocene kvaliteta pravila, učenja kombinovanih klasifikatora, metoda selekcija atributa i modifikovanog osnovnog algoritma pretraživanja prostora hipoteza.

U **poglavlju 5.** se daju eksperimentalni rezultati primene unapređenog sistema *Empiric* na standardne merne (*benchmark*) probleme induktivnog učenja i realne

probleme iz oblasti genetike, koji su oskudni i za koje nijedna od poznatih metoda induktivnog učenja razumljivog znanja ne daje zadovoljavajuće rezultate.

Opisana je metodologija izvođenja eksperimenata, korišćeni problemi induktivnog učenja, prikazani rezultati učenja i estimacije performansi naučenog znanja. Na kraju poglavlja se diskutuju dobijeni rezultati i porede sa dostupnim podacima iz literature.

U **poglavlju 6.** se se rezimiraju nalazi i eksperimentalni rezultati istraživanja. Daje se kritički osvrt na ostvarene rezultate i doprinose, predlažu moguće oblasti primene uvedenih rešenja i mogući pravci daljeg razvoja.

2. PROBLEM ISTRAŽIVANJA

U ovom poglavlju se izlažu teoretske osnove metoda i tehnologija koji će se koristiti u samom radu. Definišu se osnovni pojmovi i detaljno opisuje problem induktivnog učenja na osnovu oskudnih obučavajućih skupova, kao i najvažniji dosadašnji rezultati.

Identifikuju se metode i tehnike kojima se mogu poboljšati performanse učenja, pre svega tačnost i razumljivost naučenog znanja.

2.1 MODEL INDUKTIVNOG UČENJA

Problem učenja je samo jedan deo opšte eksperimentalne procedure, koja se koristi u različitim područjima nauke i tehnike [3], [18]. Posebno je važno uočiti da učenje na osnovu podataka ima dva različita cilja: tačnost predviđanja (generalizaciju) i interpretaciju (objašnjenje).

U matematičkoj statistici, učenje na osnovu uzorka se naziva statistička estimacija. Na osnovu poznatih uzoraka ili obučavajućih podataka (aposteriornih informacija) vrši se estimacija nepoznate funkcije distribucije, koja će se koristiti za buduća predviđanja.

Sistem induktivnog učenja se može nalaziti u dva stanja:

1. učenje ili estimacija na osnovu obučavajućih uzoraka
2. upotreba ili predviđanje za buduće ili testne primere

Podrazumeva se da su i obučavajući i testni podaci iz iste nepoznate distribucije.

Najčešći problemi učenja su klasifikacija, regresija i estimacija funkcije gustine distribucije na osnovu uzoraka.

Opšta eksperimentalna metoda otkrivanja zakonitosti u podacima predviđa:

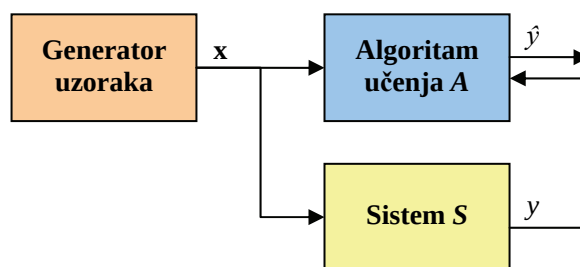
1. Iskazivanje problema
2. Formulaciju hipoteza
3. Dizajniranje eksperimenta i generisanje podataka
4. Prikupljanje i preprocesiranje podataka
5. Estimaciju modela
6. Interpretacija model i izvođenje zaključaka

Svi elementi eksperimentalnog metoda su važni i utiču na konačne zaključke. Nikakve formalne metode i algoritmi estimacije ne mogu da nadoknade nedovoljno informativne podatke ili neadekvatan model problema, koji može nepovoljno da utiče na zaključivanje na osnovu podataka ili ga čak onemogućiti.

Rezultat metoda je model visoke tačnosti (korak 5) i *objašnjenje* ustanovljenih zavisnosti (korak 6), koje je neophodno za dublje razumevanje pojave, ograničenja u primeni metoda, veze ulaznih i izlaznih promenljivih i donošenje odluka.

U ovom radu će se pažnja posvetiti pretežno učenju pravila klasifikacije, modela koji objašnjava (interpretira) eksperimentalne podatke.

Formalna definicija problema učenja, zajednička za različite pristupe i modele učenja [18], definiše učenje kao proces estimacije nepoznate ulazno-izlazne zavisnosti ili strukture posmatranog sistema S na osnovu ograničenog broja opservacija x .



Slika 1: Formalna definicija problema induktivnog učenja

Generator uzoraka modelira proces posmatranja neke pojave: uzorkuje nepoznatu distribuciju $p(\mathbf{x})$ i generiše slučajne vektore $\mathbf{x} \in \mathcal{R}^d$ (opservacije, obučavajuće primere) za algoritam učenja A , dok sistem S za svaki slučajni vektor $\mathbf{x}$ na osnovu fiksne nepoznate uslovne funkcije gustine distribucije $p(y|\mathbf{x})$ generiše izlaznu vrednost y .

Algoritam učenja A koristi konačan broj od n opservacija sistema S za formiranje što tačnije aproksimacije njegovih izlaza. Skup funkcija koje se mogu naučiti i način ocene najbolje aproksimacije definisan je unapred, na osnovu znanja o sistemu S i konstrukcijom algoritma A .

Obučavajući primeri koje generišu generator i sistem S su nezavisni i identično distribuirani u skladu sa (za kontinualne atribute) funkcijom gustine distribucije $p(\mathbf{x}, y) = p(\mathbf{x})p(y|\mathbf{x})$.

Mera tačnosti aproksimacije je nenegativna funkcija gubitka (loss) $L(y, f(\mathbf{x}, \omega))$. Očekivana vrednost gubitka se izražava funkcijom rizika (risk functional):

$$R(\omega) = \int L(y, f(\mathbf{x}, \omega)) p(\mathbf{x}, y) d\mathbf{x} dy$$

Proces učenja ima zadatak nalaženja takve vrednosti $f(\mathbf{x}, \omega_0)$ koja minimizuje R , što se obično ne može realizovati na osnovu konačnog broja primera i bez poznavanja distribucije $p(\mathbf{x}, y)$, već se dobija približno rešenje $f(\mathbf{x}, \omega_*)$.

Opšti model učenja obuhvata više klasa problema: klasifikaciju, regresiju, estimaciju gustine i grupisanje (klastering) i kvantizaciju vektora. Formalna razlika je u funkciji gubitka L i vrsti izlaza y . Zajedničko za sve probleme je minimizacija funkcionala rizika samo na osnovu obučavajućih primera.

Formulisanje matematičkih modela koji omogućavaju rigoroznu formalnu analizu i smisleno poređenje heuristika mašinskog učenja je predmet teorije mašinskog učenja (computational learning theory). Postoji više formalnih teorija ili modela mašinskog učenja [19], [20]:

- model učenja do granice (E.M. Gold, 1967),
- PAC model učenja (L. Valiant, 1984),
- model egzaktnog učenja (D. Angluin, 1988),
- modeli zasnovani na metodama minimalnog kodiranja (Wallace, 1987; Rissanen, 1987),
- modeli učenja na osnovu Bayesove i klasične statističke teorije (W.L. Buntine, 1992),
- ostali pristupi modeliranju induktivnog učenja ([19], [18]).

Ukratko će se opisati osnovni teoretski modeli, a nešto detaljnije će se izložiti najpopularniji teoretski model induktivnog učenja – PAC model (*Probably Approximately Correct Learning model*) i modeli induktivnog učenja koji se najviše koriste u učenju klasifikacija [18], [19], koji će se koristiti dalje u radu za izlaganje praktičnih problema i rezultata.

2.1.1 Model učenja do granice

Jednu od najstarijih teorija induktivnog zaključivanja razvio je M. Gold 1967. godine na osnovu teorije rekursivnih funkcija, odnosno teorije izračunljivosti [21]. Učenje se posmatra kao prikupljanje informacija o nepoznatom objektu, njihova obrada i formiranje opisa nepoznatog objekta. U idealnom slučaju, to je potpuni opis objekta.

Objekti u ovom modelu su izračunljive rekursivne funkcije (ili rekursivno prebrojivi jezici) pomoću kojih se može predstaviti proizvoljno algoritamsko ponašanje, odnosno algoritam učenja. Nepoznata funkcija f se zadaje preko svojih vrednosti $f: f(0), f(1), f(2), \dots$, a algoritam učenja generiše program h u univerzalnom programskom jeziku¹ koji izračunava nepoznatu funkciju f .

U ovom modelu se podrazumeva raspoloživost *neograničenog* broja primera za učenje. Broj hipoteza koje sistem može da postavi i ispravi takođe nije ograničen, a opis se dobija nakon dovoljnog broja primera, pa se ovaj model učenja naziva i *učenje do granice* (*learning in the limit*) [21].

2.1.2 PAC model učenja

Teorija učenja do granice bavi eventualnom identifikacijom *potpuno istinite* hipoteze. Cilj učenja prema teoriji *L. Valianta* je otkrivanje *razumno tačne* hipoteze za razumno vreme [19].

Naziv "PAC učenje" je nastao od namene modela, u kome uspešno učenje nepoznatog koncepta povlači nalaženje, sa visokom pouzdanošću, hipoteze koja predstavlja njegovu dobru aproksimaciju.

PAC model formalizuje opšti zadatak učenja koncepta, tako da to bude praktično izvedivo (*feasible*) učenje, kod koga je zavisnost vremena izvršavanja i potrebnog broja primera da se koncept nauči polinomska. U ovom modelu učenja se ne zahteva nalaženje potpuno preciznog opisa nepoznatog koncepta, već se samo traži da greška aproksimacije bude po volji mala i to sa određenom verovatnoćom.

Podrazumeva se da se primeri slučajno uzimaju iz prostora primera u skladu sa nekom distribucijom verovatnoća. Greška aproksimacije se izražava kroz verovatnoću da neki primer pripada skupu primera za koje se ciljani i naučeni koncept razlikuju.

Osnovni rezultati PAC modela induktivnog učenja na osnovu primera

PAC model mašinskog učenja je uveo 1984. godine *L.G. Valiant* [8] radi proučavanja računске složenosti algoritama mašinskog učenja, pri čemu je koristio pojednostavljene pojmove iz statističkog prepoznavanja uzoraka i teorije odlučivanja, kao i teorije izračunljivosti, odnosno algoritamske kompleksnosti. Proučavao je

¹ Odnosno, programski jezik dovoljno univerzalan da se u njemu implementira Turingova mašina

izračunljive klase problema mašinskog učenja, odnosno one za koje postoje algoritmi polinomske složenosti koji ih rešavaju.

U originalnom PAC modelu, prostor primera, kao i prostor hipoteza ili koncepata je $\{0,1\}^n$, odnosno sastoji se od niza vrednosti binarnih (*boolean*) atributa. Smisao aproksimacije se definiše uz pretpostavku da postoji neka distribucija D definisana nad prostorom primera $\{0,1\}^n$, koja svakom primeru pridružuje neku verovatnoću.

Greška hipoteze h , kad je poznat traženi koncept f , definiše se kao verovatnoća da se hipoteza h i koncept f razlikuju na primeru slučajno izabranom u skladu sa distribucijom D :

$$\varepsilon_h = \sum_{x \in \Delta} D(x)$$

gde Δ označava simetričnu diferenciju.

Hipoteza h je dobra aproksimacija nepoznatog koncepta f ako je greška ε_h dovoljno mala. Najjednostavniji način dobijanja dobre hipoteze je traženje nezavisnih slučajnih primera traženog koncepta f , od kojih se svaki sastoji od primera koja je izabran slučajno u skladu sa distribucijom D i označava se kao pozitivan, ako je primer koncepta f , odnosno kao negativan, ako nije primer koncepta. Učenje i testiranje koriste istu distribuciju primera, odnosno ne razmatra se "šum" u podacima. Algoritam učenja je tada izračunljiva procedura koja uzima uzorak traženog koncepta f , koji se sastoji od niza nezavisnih slučajnih primera koncepta f i vraća kao rezultat hipotezu.

Neka je za svaki $n \geq 1$ skup C_n traženih koncepata nad prostorom primera $\{0,1\}^n$ i neka je $C = \{C_n\}_{n \geq 1}$. Neka su H_n , za $n \geq 1$ i $\mathbf{H}$ definisani na sličan način. Tada se *moгуćnost PAC učenja* (*PAC learnability*) može definisati na sledeći način:

Klasa koncepata $\mathbf{C}$ se može PAC naučiti u prostoru hipoteza $\mathbf{H}$ ako postoji algoritam učenja koji se izvršava za polinomsko vreme i polinom $p(\cdot, \cdot, \cdot)$, takav da za sve $n \geq 1$, svi traženi koncepti $c \in C_n$, sve distribucije verovatnoća D nad prostorom primera $\{0,1\}^n$ i svaki ε i δ , gde su $0 < \varepsilon, \delta < 1$, ako se algoritmu A zadaje najmanje $p(n, 1/\varepsilon, 1/\delta)$ nezavisnih slučajnih primera koncepta c uzetih na osnovu distribucije D , onda sa verovatnoćom od najmanje $1/\delta$, algoritam A vraća hipotezu $h \in H_n$ sa greškom $\varepsilon_h \leq \varepsilon$.

Polinom p najmanjeg reda se naziva kompleksnost uzorka (*sample complexity*) algoritma učenja A .

Definicija ističe da algoritam učenja mora biti efikasan (polinomske složenosti) i mora sa velikom verovatnoćom da daje dobru aproksimaciju nepoznatog koncepta, uz razuman broj slučajno izabranih obučavajućih primera.

Ako se prostor hipoteza $\mathbf{H}$ poklapa sa klasom koncepata $\mathbf{C}$, klasu koncepata je moguće *pravilno* (*properly*) PAC naučiti. Ovaj zahtev nije neophodan kad se samo traži efikasan način predviđanja novih primera, bez obzira na oblik naučene hipoteze.

Neki od poznatijih rezultata PAC teorije o mogućnosti pravilnog PAC učenja su:

- konjunktivni koncepti (logički izrazi) se mogu pravilno PAC naučiti. Klasa koncepata u obliku disjunkcije dva konjunktivna izraza se *ne može* pravilno PAC naučiti, kao ni klasa egzistencijalnih konjunktivnih koncepata u strukturnim prostorima primera sa dva objekta;

- pojedinačni koncepti u obliku perceptrona, odnosno linearnih *threshold* funkcija, mogu se pravilno PAC naučiti. Klasa koncepata u obliku konjukcije ili disjunkcije dva koncepta u obliku perceptrona *ne može* se pravilno PAC naučiti, kao ni mreže u više nivoa sa dve sakrivene jedinice.
- klase disjunktivnih i konjuktivnih Bulovih formula *konačne dužine* k , kao i liste odlučivanja konačne dužine mogu pravilno PAC naučiti za svako fiksirano k . Nije poznato da li se klase svih disjunktivnih i konjuktivnih Bool-ovih formula i svih lista odlučivanja mogu pravilno PAC naučiti.

Dokaz mogućnosti PAC učenja podrazumeva:

- dokazivanje da postoji efikasan algoritam koji nalazi hipotezu u konkretnom prostoru hipoteza i koji je konzistentan (tačan za sve primere) sa zadanim uzorkom bilo kog koncepta iz ciljne klase koncepata, i
- dokazivanje da je kompleksnost uzorka svakog takvog algoritma polinomska.

Za veće prostore hipoteza je potreban veći uzorak m primera radi nalaženja konzistentne hipoteze sa zadanom verovatnoćom bar $1/\delta$.

Ako se koriste samo diskretni atributi, gornja granica broja primera (tzv. *Blumerova granica*) se može izraziti preko kardinalnosti prostora hipoteza $|H_n|$:

$$m \leq \frac{1}{\varepsilon} \left(\ln |H_n| + \ln \frac{1}{\delta} \right)$$

Kompletni prostor hipoteza nad n propozicionih simbola ima kardinalnost $|H_n| = 2^{2^n}$ (2^n primera, a svaki se može označiti kao istinit ili lažan), a prostor od k disjunkcija konjuktivnih izraza proizvoljne dužine ima $|H_n| \leq (3^n)^k$.

Osim kardinalnosti prostora hipoteza, razmatra se i tzv. VC-dimenzija (*Vapnik-Chervonenkis dimension*) prostora hipoteza. Označava se sa $VC\text{-dim}(H)$ i definiše kao maksimalni broj d primera koji se mogu označiti kao pozitivni i negativni primeri na svih 2^d mogućih načina, tako da je svako označavanje konzistentno sa nekim hipotezama u H .

Pomoću ove veličine se može oceniti gornja granica kompleksnosti, odnosno najmanja potrebna veličina uzorka. Za $C_n \subseteq H_n$ i $n \geq 1$ se može pokazati da svaki konzistentan algoritam za učenje klase C pomoću prostora hipoteza H ima gornju granicu kompleksnosti uzorka

$$m \leq \frac{1}{\varepsilon(1-\sqrt{\varepsilon})} \left(2 \cdot VC\text{-dim}(H_n) \ln \frac{6}{\varepsilon} + \ln \frac{2}{\delta} \right)$$

Model PAC učenja se može jednostavno proširiti na prostore primera sa drugim tipovima atributa, uključujući attribute koji opisuju strukturalna svojstva primera (npr. za opis scene). Moguće je definisati primere kao nizove nad fiksnim alfabetom i ispitivati mogućnost učenja konačnih automata, beskontekstnih gramatika i sličnih klasa koncepata.

Važna modifikacija PAC modela je slabi (*weak*) PAC model. Pretpostavka tzv. jakog (*strong*) PAC modela je mogućnost postizanja po volji male klasifikacijske greške, uz dovoljan broj primera. Slabi PAC model učenja postavlja slabiji zahtev pred

algoritam učenja, jer je dovoljno da klasifikacijska greška koju garantuje algoritam učenja bude samo nešto manja od greške slučajne klasifikacije.

Zahtev jakog PAC modela za postizanje po volji male greške klasifikacije može se realizovati i kombinacijom slabog PAC algoritma (koji zadovoljava uslove slabog PAC modela) i posebnog algoritma kombinovanja, kojim se kombinuju rezultati više aktiviranja slabog algoritma nad različitim uzorcima obučavajućih primera. Najpoznatije metode kombinovanja su isticanje (*boosting*) i natrpavanje (*bagging*).

Praktično realizovani uspešni sistemi mašinskog učenja, kao što su programi za indukciju stabala odlučivanja *C4.5* i *CART*, pravila *C45RULES* i *RIPPER*, kao i neuronskih mreža *Backprop*, ne zadovoljavaju kriterijume PAC modela učenja [20]. Na taj način, PAC model učenja ne omogućava u praksi smisleno poređenje različitih algoritama učenja.

Postoji više proširenja PAC modela radi uključivanja realnih problema u analizu: agnostički (*agnostic*) PAC model više ne pretpostavlja da postoji traženi koncept, dok PAC model sa šumom (*noise model*) pretpostavlja prisustvo velike količine slučajnog šuma ili malog dela namernog šuma u podacima.

Nedostatak PAC modela mašinskog učenja je egzistencijalna priroda rezultata koje daje: *mogućnost* ili *nemogućnost* učenja određenih klasa koncepata, koji malo utiču na praktičnu realizaciju sistema induktivnog učenja.

2.1.3 Model egzaktnog učenja

Model egzaktnog učenja ili model upita pred algoritam učenja postavlja zahtev da proizvede hipotezu koja je identična ciljnoj, odnosno klasifikuje sve primere na identičan način kao i nepoznati ciljni koncept. Algoritam učenja može da koristi *dodatne izvore* informacija, osim obučavajućeg skupa, u obliku *upita o klasifikaciji* pojedinih primera, *upita o pripadnosti* i *upita o ekvivalentnosti*, odnosno proveriti da li je naučeni koncept već jednak ciljnom.

Jedan od prvih rezultata ovog modela je algoritam *D. Angluin* za učenje monotonih DNF (*disjunctive normal form*) formula pomoću upita o pripadnosti i ekvivalentnosti. Može se modifikovati za učenje koncepata korišćenjem samo upita o pripadnosti, ako se reše dodatni računski problemi za koje su poznati efikasni algoritmi, što je povezano sa praktičnim problemima na području istraživanja podataka (problem otkrivanja asocijacija, nalaženja svih ključeva u relaciji i sl.) [20].

2.1.4 Modeli minimalnog kodiranja

Prema MDL principu (*Minimum Description Length Principle*), za objašnjenje podataka je najbolja ona teorija koja minimizuje sumu dužine opisa same teorije i dužinu podataka zapisanih uz pomoć te teorije (u bitima) [1].

Učenje koncepata se može posmatrati kao kompresija informacija, pošto koncept (teorija ili hipoteza) opisuje (*describes*) ili objašnjava (*explains*) veliki broj primera iz obučavajućeg skupa. U opštem slučaju ne postoji jedinstvena hipoteza koja objašnjava obučavajući skup primera. Prema MDL principu, najbolji je onaj opis koji minimizuje sumu dužine opisa koncepta i dužinu zapisa svih primera pomoću tog koncepta.

Jedan od poznatih principa u izboru optimalne hipoteze ili opisa koncepta je tzv. princip "Okamove oštrice", koji nalaže izbor najjednostavnije hipoteze, odnosno one koja minimizuje informaciju potrebnu za njen zapis

$$\arg \min_{h \in H} I(h)$$

gde je $H = \{h_i\}$ skup hipoteza, a $I(h)$ je količina informacije potrebna za kodiranje hipoteze $h \in H$.

Doslovna primena principa izbora najjednostavnije hipoteze generiše suviše specijalizovane teorije, kojima se opisuju i brojna odstupanja podataka u primerima od tražene zakonitosti.

U prisustvu šuma, određena nekonzistentnost i nekompletnost opisa koncepta je neophodna [22]. MDL princip ocenjuje ovakve hipoteze i bira one koje minimizuju količinu informacije potrebne za zapis same hipoteze i zapis obučavajućeg skupa primera pomoću te hipoteze:

$$\arg \min_{h \in H} [I(h) + I(S_t|h)]$$

gde je S_t obučavajući skup primera relativan za hipotezu, a $I(S_t|h)$ je količina informacije potrebna za kodiranje obučavajućeg skupa primera S_t pomoću hipoteze h . Ovaj član predstavlja količinu informacija potrebnih za predstavljanje izuzetaka od teorije, pošto su primeri koje teorija korektno klasifikuje već sadržani u opisu same teorije.

U odsustvu šuma, MDL princip se za nekonzistentne hipoteze svodi na princip "Okamove oštrice", jer za njih važi $I(S_t|h) = \infty$, pa se za (kraće) konzistentne hipoteze generiše isti poredak kao i pomoću izraza $I(h)$.

U praksi se pokazalo da, za mali broj primera, MDL mera favorizuje suviše generalne koncepte, koji imaju loše performanse na novim primerima. Ograničenja opisnog jezika takođe mogu imati kao posledicu nemogućnost generisanja opisa koji dobro objašnjava obučavajuće primere.

Npr. u [23] MDL princip je kombinovan sa tačnošću na obučavajućem skupu u programu za učenje predikatnih, kao i propozicionih koncepata, odnosno:

$$\arg \min_{h \in H: I(h) + I(S_t|h) < I(S_t)} I(S_t|h)$$

gde je $I(S_t)$ količina informacija potrebna za kodiranje obučavajućeg skupa S_t . Za propozicione koncepte je dobijena ista tačnost predviđanja, uz značajno smanjenje veličine opisa.

U [24] i [25] razmatra se upotreba modela minimalnog kodiranja, odnosno MDL ocene kompleksnosti naučenog znanja, a u [26] upotreba MDL ocene kompleksnosti pravila kao posredne ocene razumljivosti, u praktičnom problemu kvantifikacije efekta uvođenja strukturnih atributa u propozicioni model problema induktivnog učenja na razumljivost naučenih pravila.

2.1.5 Modeli učenja na osnovu Bayesove i klasične statističke teorije

Prema [19], modeli Bayesovske statističke teorije predstavljaju dovoljan teoretski aparat za modeliranje svih bitnih aspekata problema mašinskog učenja koncepata, kao

što su izračunljivost (kompleksnost) algoritama učenja, ocena greške učenja, predubedenje (*bias*), pretraživanje prostora hipoteza, problem detaljisanja (*overfitting*) i prisustvo "šuma" u primerima.

Model se zasniva na Bayesovoj teoriji verovatnoće koja uslovnu verovatnoću $P(A|B)$ definiše kao stepen uverenosti u tvrdnju A na osnovu raspoložive činjenice B . Teorija je pogodna za mašinsko učenje jer je jednostavna, pošto se zasniva na dva aksioma:

$$P(A|C) + P(\bar{A}|C) = 1$$

$$P(AB|C) = P(B|C)P(A|BC) = P(A|C)P(B|AC)$$

Njihova direktna posledica je Bayesova teorema:

$$P(A|BC) = \frac{P(A|C)P(B|AC)}{P(B|C)}$$

Bayesova teorema se može razmatrati i kao algoritam induktivnog učenja koncepata. Neka je C_t predznanje u trenutku t , uključujući prostor primera I i prostor hipoteza H definisanih nad I , a A_t tvrdnja da je hipoteza H_t ciljani koncept, a B_t novi primer, sa oznakom pripadnosti klasi (da li predstavlja pozitivan primer koncepta ili ne).

Učenje se sastoji u računanju $P(A_t|B_tC_t)$ na osnovu Bayesove teoreme i postavljanju vrednosti $P(A_t|C_{t+1}) = P(A_t|B_tC_t)$ za svaki novi primer. U svakom trenutku ciljani koncept je hipoteza H^* sa najvećom verovatnoćom $P(A^*|C_t)$.

Posebna pogodnost je što Bays-ova teorija kroz kvantifikaciju istinitosti tvrdnji i operacije nad njima predstavlja generalizaciju matematičke logike.

Takođe je pokazano da za mnoge probleme koji se tretiraju drugim teorijama mašinskog učenja postoji dualna interpretacija u okviru Bayesovske statističke teorije.

2.1.6 Ostali pristupi modeliranju induktivnog učenja

U [19] se kao teorije induktivnog učenja navode još i modeli učenja iz klasične matematičke statistike i prepoznavanja uzoraka: diskriminantna analiza, estimacija osnovne gustine (*kernel density estimation*), metoda najbližeg suseda (*nearest neighbour*), selekcija i ekstrakcija atributa (*feature selection and extraction*), kao i metode razbijanja (*partitioning methods*).

Na ovom mestu će se još ukratko neformalno opisati modeli induktivnog učenja koji se zasnivaju na konceptima rastojanja (*distance*) i margine (*margin*):

1. metod najbližeg suseda (*nearest neighbour*) iz grupe metoda učenja na osnovu instanci (*instance learning*)
2. metod saport vektora (*support vector machines*, SVM) i
3. metodi učenja klasifikacija - hijerarhijskog i partitivnog grupisanja (*clustering*).

Selekcija atributa će se opisati posebno u poglavlju 2.3.

2.1.6.1 METOD NAJBLIŽEG SUSEDA

Metoda najbližeg suseda je jedna od najstarijih i najjednostavnijih metoda učenja klasifikacionih pravila u oblasti prepoznavanja uzoraka (*pattern recognition*).

Pravilo najbližeg suseda (1-Nearest Neighbor rule, 1-NN) je metod klasifikacije kod koga se klasifikacija novih primera vrši direktno na osnovu rastojanja koja se računaju u odnosu na testni primer. Klasa primera se određuje prema nasličnijem objektu iz obučavajućeg skupa. Najsličniji objekt je onaj koji je najmanje različit, odnosno najbliži testnom primeru. U ovom pristupu nema posebne faze obučavanja klasifikatora.

Pravilo k najbližih suseda (k -Nearest Neighbor rule, k -NN) se odnosi na metod klasifikacije kod koga se klasifikacija novih primera vrši na osnovu najfrekventnije klase u skupu od k najbližih suseda. Vrednost k se može optimalno odrediti na osnovu primera iz obučavajućeg skupa T nekom od metoda estimacije tačnosti klasifikacije, npr. *leave-one-out* (tačka 2.2 ovog rada).

Najjednostavnija tehnika određivanja stepena sličnosti primera x i y opisanih pomoću d atributa (numeričkih i nominalnih) je računanje Euklidovog rastojanja u d -dimenzionalnom prostoru:

$$similarity = -\sqrt{\sum_{i=1}^d f(x_i, y_i)}$$

gde je

$$f(x_i, y_i) = \begin{cases} (x_i - y_i)^2 & \text{za numeričke atribute} \\ 1 & \text{za nominalne atribute, ako je } x_i \neq y_i \\ 0 & \text{za nominalne atribute, ako je } x_i = y_i \end{cases}$$

Sam proces učenja (osim prvog primera, koji se samo zapamti) podrazumeva prethodnu probnu klasifikaciju novog obučavajućeg primera, tako što se računa sličnost sa svim prethodno zapamćenim primerima i novom primeru dodeli klasa najsličnijeg zapamćenog primera.

Ako postojeći primeri klasifikuju obučavajući primer ispravno, povećava se brojčak uspešnosti klasifikacije prethodno zapamćenih primera, a zatim se novi primer pamti sa srednjom početnom vrednošću uspešnosti.

Ako se postojeći primer ne klasifikuje ispravno, brojčak uspešnosti postojećih primera se smanjuje, a novi primer se pamti sa srednjom početnom vrednošću uspešnosti.

Prilikom klasifikacije se koriste samo *dobri* primeri, sa visokom vrednošću brojača uspešnosti, dok se *neutralni* primeri sa srednjom vrednošću brojača koriste samo ako više nema nijednog dobrog. Kada brojčak uspešnosti nekog primera padne ispod određene granice, primer se briše iz memorije.

Dobra strana ovog metoda učenja je nezavisnost od poznavanja *statističkih svojstava* obučavajućeg skupa.

Učenje se svodi na pamćenje tipičnih primera, a klasifikacija na svrstavanje u klasu kojoj pripada najsličniji prethodno zapamćen (naučen) primer. Ovakav pristup ise naziva metod ili princip najbližeg suseda (*nearest-neighbor principle*) [7].

Poznati algoritmi učenja osnovu instanci (*instance based learning*) su *kNN* i *IBL*, čija implementacija postoji u sistemu *WEKA* [5] Kao i druge metode učenja, metod je osetljiv na oskudnost obučavajućeg skupa, koja nastaje zbog nedovoljnog broja obučavajućih primera.

2.1.6.1 METOD SAPORT VEKTORA

Metod saport vektora (*support vector machine*, SVM) je algoritam učenja koji pronalazi optimalnu hiperpovršinu koja razdvaja obučavajuće primere sa maksimalnom marginom. Saport vektori su *granični primeri* najbliži hiperpovršini maksimalne margine [5].

Margina (*margin*) je rastojanje između hiperpovršine i njoj najbližeg obučavajućeg primera. Svi primeri jedne klase se mogu zatvoriti u konveksni omotač (*convex hull*), hiperpoligon koji se dobija međusobnim povezivanjem svih primera jedne klase. Kada su klase linearno separabilne, ovi poligoni se ne preklapaju.

Hiperpovršina maksimalne margine je maksimalno udaljena od konveksnih omotača i pod pravim uglom preseca najkraću liniju koja ih spaja. Algoritam SVM bira mali skup kritičnih graničnih primera (saport vektora) svake od klasa i gradi linearnu funkciju koja ih najbolje razdvaja.

SVM za izgradnju nelinearnih granica razdvajanja klasa u hiperprostoru koristi *linearne modele*, pomoću nelinearnih transformacija (*kernel trick*): transformiše prostor primera u novi prostor, u kome formira linearnu funkciju razdvajanja, koja povratkom u originalni prostor postaje nelinearna.

2.1.6.3 METODI GRUPISANJA (CLUSTERING)

Metodi grupisanja se koriste kada je cilj induktivnog učenja definisanje novih pojmova, odnosno *formiranje* novih koncepata na osnovu primera. Rezultat učenja je grupisanje (*clustering*) skupa obučavajućih primera u podskupove. Grupisanje se može biti partitivno i hijerarhijsko, u zavisnosti od disjunktnosti formiranih podskupova primera.

Metodi grupisanja se zasnivaju na konceptu međusobne sličnosti primera, koja se može izraziti preko distance primera $dist(\mathbf{x}, \mathbf{y})$. Kreirane su brojne mere sličnosti, odnosno rastojanja, za rešavanje praktičnih problema u različitim oblastima

empirijskih istraživanja, npr. [4],[18],[41] Euklidova $dist(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^d (x_i - y_i)^2}$,

mera Minkovskog $dist(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^d |x_i - y_i|^r \right)^{1/r}$ i Manhattan (*city block*)

$dist(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^d |x_i - y_i|$. Koriste se i neke mere iz statistike, npr. korelacija

$$dist(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^d (x_i - \bar{x}_i)(y_i - \bar{y}_i)}{\sqrt{\sum_{i=1}^d (x_i - \bar{x}_i)^2 \sum_{i=1}^d (y_i - \bar{y}_i)^2}}.$$

Interpretacija ili *razumljiv opis* dobijenih grupisanja se može dobiti metodima učenja koncepata, koji se razmatraju u ovom radu.

2.2 ESTIMACIJA TAČNOSTI I OSKUDNI OBUČAVAJUĆI SKUPOVI

Da bi bilo korisno, indukovano znanje mora biti tačno i razumljivo. Tačnost predviđanja (*predictive accuracy*) je osnovni pokazatelj performansi većine sistema induktivnog učenja. Predstavlja procenat uspešnosti klasifikacije novih, nerazmatranih primera korišćenjem naučenih pravila.

Pouzdana ocena ili estimacija tačnosti predviđanja potrebna je zbog predviđanja performansi klasifikatora u budućnosti, izbora najpogodnijeg klasifikatora iz nekog skupa (*model selection*) i ocene ponašanja kombinacije više različitih klasifikatora.

Prema [27] model ocenjivanja tačnosti polazi od prostora neklasifikovanih primera $\mathbf{V}$ i skupa mogućih klasa $\mathbf{Y}$. Prostor klasifikovanih primera $\mathbf{X} = \mathbf{V} \times \mathbf{Y}$, a $\mathbf{D} = \{x_1, \dots, x_n\}$ je skup od n klasifikovanih primera $x_i = \langle v_i \in \mathbf{V}, y_i \in \mathbf{Y} \rangle$. Klasifikator (*classifier*) $\mathbf{C}$ je funkcija koja vrši pridruživanje klase $y \in \mathbf{Y}$ neklasifikovanom primeru $v \in \mathbf{V}$. Algoritam indukcije (*inducer*) $\mathbf{I}$ preslikava zadani prostor neklasifikovanih primera $\mathbf{D}$ u klasifikator $\mathbf{C}$, odnosno gradi klasifikator.

Oznaka $\mathbf{I}(\mathbf{D}, v)$ označava klasu koju primeru v dodeljuje klasifikator koji gradi algoritam indukcije $\mathbf{I}$ na skupu primera $\mathbf{D}$. Tačnost klasifikatora $\mathbf{C}$ se onda definiše kao verovatnoća ispravne klasifikacije slučajno izabranog primera $\langle v, y \rangle \in \mathbf{X}$, odnosno $acc = P(\mathbf{C}(v) = y)$, pri čemu se podrazumeva ista distribucija primera kao kod obučavajućeg skupa.

Za ocenu klasifikacijske tačnosti skupa pravila na primerima koji će se pojaviti u budućnosti koriste se različite metode ocenjivanja, odnosno estimacije [27]:

- metod estimacije testnog uzorka (*holdout method*)
- metod unakrsnog ocenjivanja (*cross-validation*)
- metod početnog punjenja (*bootstrap*)

2.2.1 Estimacija tačnosti klasifikacije metodom testnog uzorka

Metoda deli primere na dva međusobno nezavisna podskupa - obučavajući skup i testni ili *holdout* skup primera. Obično se 2/3 primera koristi za učenje, a 1/3 za ocenu tačnosti predviđanja. Nakon učenja, skup pravila se proverava kroz klasifikaciju primera iz testnog skupa i tačnost predviđanja se ocenjuje kao

$$acc_h = \frac{1}{h} \sum_{\langle v_i, y_i \rangle \in D_h} \delta(\mathbf{I}(D_t, v_i), y_i)$$

gde je D skup primera, D_h je testni, a D_t je obučavajući skup ($D \setminus D_h$), dok je $\delta(i, j) = 1$ ako je $i = j$, inače je 0.

Estimacija tačnosti predviđanja je slučajni broj koji zavisi od podele primera na obučavajući i testni skup. Postupak se može ponaviti k puta (*random subsampling*) i tačnost se može oceniti kao srednja vrednost. Standardna devijacija se dobija kao standardna devijacija pojedinačnih ocena.

Tačnost predviđanja naučenog skupa pravila povećava se brojem primera u obučavajućem skupu. Pošto za učenje nisu upotrebljeni svi primeri, ovaj metod daje pesimističku ocenu tačnosti predviđanja i neefikasno koristi raspoloživi skup primera. Izdvajanjem većeg broja primera za testiranje, povećava se pristrasnost estimacije, dok se smanjivanjem skupa primera proširuju granice intervala poverenja.

2.2.2 Estimacija tačnosti klasifikacije metodom unakrsnog ocenjivanja (validacije)

Metoda unakrsnog ocenjivanja ili rotaciona estimacija slučajno deli skup primera D na k međusobno isključivih podskupova primera $D_1, D_2, \dots, D_k$ približno iste veličine. Postupak učenja i ocenjivanja se ponavlja k puta, svaki put koristeći po jedan podskup D_i kao testni skup. Ukupna tačnost se računa kao odnos broja ispravnih klasifikacija i ukupnog broja primera u skupu D .

$$acc_{cv} = \frac{1}{h} \sum_{\langle v_i, y_i \rangle \in D} \delta(I(D \setminus D_i, v_i), y_i)$$

Estimacija tačnosti predviđanja unakrsnim ocenjivanjem je slučajni broj koji zavisi od podele primera na podskupove. Kompletana unakrsna validacija je srednja vrednost svih $\binom{m}{m/k}$ mogućnosti izbora m/k primera od ukupno m primera.

Za zadani algoritam učenja se kaže da je *stabilan* na zadanom skupu primera ako indukuje skupove pravila koji daju ista predviđanja za različite perturbacije obučavajućeg skupa. Za stabilne algoritme induktivnog učenja, estimacija tačnosti predviđanja unakrsnom validacijom je nepristrasna i iznosi $acc_{cv} \cdot (1 - acc_{cv})/n$, gde je n ukupan broj primera u skupu D , dok su varijance pojedinih ocena istovetne, odnosno ne zavise od k .

2.2.3 Estimacija tačnosti klasifikacije metodom početnog punjenja

Metode početnog punjenja predstavljaju familiju metoda za estimaciju tačnosti predviđanja. Za zadani skup od n primera *bootstrap* uzorak se formira slučajnim izborom n primera uniformno iz skupa primera, sa zamenom. Pošto je slučajni uzorak formiran sa zamenom, verovatnoća da bilo koji zadani primer nije izabran nakon n uzoraka je $(1 - 1/n)^n \approx e^{-1} \approx 0,368$. Estimacija tačnosti ϵ_0 se izvodi korišćenjem *bootstrap* uzorka za učenje i ostatka primera za testiranje. Za zadani broj *bootstrap* uzoraka b , ocenu tačnosti ϵ_0 ; *bootstrap* uzorka i i ocenu tačnosti resubstitucije na celom skupu primera acc_s (tj. tačnost na obučavajućem skupu), definiše se .632 *bootstrap* estimacija kao:

$$acc_{boot} = \frac{1}{b} \sum_{i=1}^b (0,632 \cdot \epsilon_{0_i} + 0,368 \cdot acc_s)$$

Varijansa estimacije se određuje računanjem varijanse estimacija pojedinih uzoraka. Kao i za unakrsnu validaciju, pretpostavlja se stabilnost algoritma induktivnog učenja na zadanom skupu primera ("svet *bootstrap*-a" je verna aproksimacija stvarnog sveta).

Uporedni pregled osnovnih karakteristika metoda estimacije tačnosti predviđanja dat je u Tabeli 1 [28].

Metoda estimacije	Broj primera		Broj inter. iteracija	Broj iteracija	Varijabilnost metoda	Objektivnost
	za učenje	za testiranje				
Resubstitucija (R)	n	n	1	1	vrlo visoka	vrlo optimistički
Hold-out (H)	$(0,6-0,8) \cdot n$	$(0,2-0,4) \cdot n$	1	1	visoka	pesimistički
Random subsampling (H^I)	$(0,6-0,8) \cdot n$	$(0,2-0,4) \cdot n$	1	$I \ll n, O(10)$	srednje visoka	pesimistički
K-fold CV(K)	$n(k-1)/k$	n/k	$k (\sim 10)$	1	srednje visoka	skoro objektiv
I K-folds CV (K^I)	$n(k-1)/k$	n/k	$k (\sim 10)$	$I \ll n, O(10)$	srednja	skoro objektiv
Leave-one-out (L)	$n-1$	1	n	1	srednje visoka	skoro objektiv
.632 Bootstrap (B)	n^*	n^*	1	$I(50-200)$	niska	neznatno optimistički

Tabela 1: Pregled metoda estimacije tačnosti predviđanja

Tek kad je tačnost predviđanja (ocenjena na neki od izloženih načina) zadovoljavajuća, razmatraju se i drugi mogući faktori kvaliteta naučenih pravila, kao što su razumljivost opisa (*comprehensibility*), jednostavnost izračunavanja (*computational simplicity*), usredsređenost (*focusing*) na posebno istaknute attribute i slično.

U stvarnim primenama, podaci su obično zašumljeni (sadrže greške) i nekonzistentni (sadrže kontradikcije), pa se nužno u skupu naučenih pravila pojavljuju prividne zakonitosti, koje smanjuju klasifikacijsku tačnost na novim primerima.

2.2.4 Oskudni obučavajući skupovi

Induktivno učenje se vrši tako da se model problema kreira tako da uključuje sve raspoložive informacije o nekoj pojavi, a obučavajući skup se formira na osnovu konačnog broja poznatih ili razmatranih slučajeva ili instanci.

Primeri problema sa velikim brojem atributa i veoma malim brojem primera su problemi analize genskih ekspresija [9], [10], [29], strukture farmaceutskih supstanci [30], analiza teksta [31] i prepoznavanje fizionomija [14].

Oskudni podaci (*sparse data*) se definišu u [32] kao slučaj kada obučavajući skup predstavlja veoma mali podskup prostora primera. Prema [33], to su podaci kod kojih je odnos broja opservacija (primera) i broja promenljivih (atributa) mali.

Analogno, prema [34] skup podataka je oskudan (*sparse dataset*) ako većina tačaka iz prostora atributa nije zastupljena u primerima. Prema [35] oskudan skup primera (*sparse example set*) je skup u kojem se javlja problem nedostatka relevantnih, odnosno krucijalnih pozitivnih primera.

U [36] se razmatraju oskudni skupovi primera u kontekstu baza podataka (*sparse itemsets/databases*) i navodi se da su stvarni skupovi primera "gusti" (*dense*), odnosno "teški" (*hard*). U [38] se takođe razmatra "težina" skupa primera (*dataset*

difficulty) i konstatuje da tačnost predviđanja ne treba da bude suviše visoka nakon uvida u samo mali broj primera iz obučavajućeg skupa.

Veličina prostora primera, koja je Dekartov proizvod domena atributa, tipično je više redova veličine veća od prostora primera. Ovaj odnos se može smanjiti ograničavanjem broja atributa, tako da podaci više nisu oskudni (*sparse*).

U [39] prikazana je klasifikacija konkretnih obučavajućih skupova na oskudne (*sparse*) i bogate (*dense*). Vidi se da su neki skupovi sa samo 100 obučavajućih primera opisanih sa 10.000 atributa svrstani u bogate, dok su skupovi sa 300-800 primera sa 20.000-100.000 atributa označeni kao oskudni. Kriterijum oskudnosti je povezan sa procenom težine problema učenja (*task difficulty*).

U teoriji mašinskog učenja [8] razmatra se problem nalaženja konzistentne hipoteze u prostoru svih mogućih hipoteza (konzistentna hipoteza je saglasna sa svakim primerom iz obučavajućeg skupa, a algoritam koji nalazi konzistentne hipoteze je konzistentan algoritam). U ovom radu će se usvojiti definicija oskudnosti obučavajućeg skupa iz [8], koja se zasniva na tačnosti predviđanja naučenog znanja:

Definicija Obučavajući skup je oskudan (*sparse*) ako induktivno naučeno znanje *ne omogućava dovoljno uspešno razlikovanje novih primera*.

Uspešnost razlikovanja novih primera zavisi od konkretnog problema i određuje se *empirijski*. Npr. u medicini je potrebno postići istu ili veću tačnost predviđanja nego lekari-specijalisti, što je veoma subjektivna ocena, koja se razvojem medicine menja.

Estimacija tačnosti naučenog znanja veoma zavisi i od metoda estimacije. U ovom radu se koristi metod unakrsne validacije, koji se smatra "skoro objektivnim" [28]. Kada se koriste metodi selekcije atributa, vrši se "dvostruka krosvalidacija" [37], odnosno krosvalidacija rezultata algoritma učenja se vrši unutar jednog krosvalidacionog koraka selekcije atributa.

U [40] se navode dva osnovna razloga oskudnosti obučavajućeg skupa: (1) nedovoljan broj primera za učenje i (2) prevelika složenost razmatranog problema.

U takvom slučaju su i uobičajene pretpostavke postojećih metoda pojednostavljivanja naučenog znanja neodgovarajuće i one same imaju negativne posledice na tačnost predviđanja naučenog znanja.

2.3 REDUKCIJA DIMENZIONALNOSTI I SELEKCIJA ATRIBUTA

Algoritmi induktivnog učenja koncepata generišu skup pravila koji istovremeno maksimizuje tačnost predviđanja i minimizuje dužinu zapisa naučenog koncepta. Dužina zapisa je proporcionalna broju elementarnih selektora, odnosno broja atributa. Na taj način algoritmi učenja koncepata u toku procesa učenja vrše *selekciju* atributa nekom *ugrađenom* metodom.

Zašto je onda *prethodna selekcija* atributa ipak neophodna u induktivnom učenju koncepata?

Pozitivni efekti prethodne selekcije atributa su:

- smanjenje efekta visoke dimenzionalnosti [18], čime se popravljaju performanse kada se raspolaze ograničenim brojem primera (za istu preciznost, kod povećanja

broja dimenzija za k , neophodno je n^{d+k} instanci-tačaka, što je povećanje za faktor n^k ;

- povećanje kvaliteta generalizacije, odnosno tačnosti predviđanja na novim primerima, jer je manja verovatnoća preteranog podešavanja prema obučavajućim podacima (*overfitting*), posebno u prisustvu šuma;
- povećanje razumljivosti naučenog znanja;
- veliko smanjenje vremena računanja.

Svojstvena dimenzionalnost (*intrinsic dimension*) skupa podataka je minimalni skup nezavisnih promenljivih potrebnih za njihovo generisanje. Opštije, skup podataka sa d dimenzija ima svojstvenu dimenzionalnost $m < d$ ako se celokupni podaci nalaze u m -dimenzionalnom prostoru S .

Estimacija svojstvene dimenzionalnosti kao donje granice broja promenljivih neophodnih za opis skupa podataka veoma je važna u statistici i prepoznavanju uzoraka (npr. za vizualizaciju i određivanje nekih parametara učenja).

Poznate tehnike estimacije svojstvene dimenzionalnosti su *projekcione tehnike* (linearne i nelinearne) i *estimacija topološke dimenzije* podataka.

Jedna od opštih tehnika pretprocesiranja skupova podataka visoke dimenzionalnosti je selekcija atributa (*attribute selection*, *feature selection*), čiji je cilj izbor relativno malog podskupa atributa koji će, osim smanjenja dimenzionalnosti, uklanjanjem irelevantnih i redundantnih atributa poboljšati performanse učenja, pre svega tačnost i razumljivost naučenog modela.

Primena različitih tehnika selekcije atributa je neophodna npr. u bioinformatiči, gde se koriste modeli visoke dimenzionalnosti u analizi sekvenci, analizi genskih ekspresija i spektralnoj analizi.

Selekcija atributa je posebno značajna tehnika redukcije dimenzionalnosti, jer čuva originalno značenje atributa, koje je razumljivo čoveku.

Tehnike transformacije prostora atributa (npr. analiza glavnih komponenti, *principal component analysis* [41], [18]) i tehnike kompresije na osnovu teorije informacija menjaju originalni model problema uvodeći nove attribute koji nemaju razumljivu interpretaciju u kontekstu problema koji se razmatra.

2.4 INTERAKCIJE ATRIBUTA

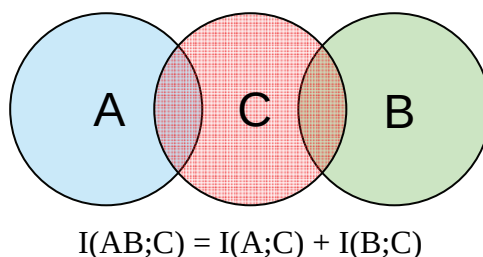
Interakcija atributa ima veliki značaj u mašinskom učenju, jer atributi u realnim konceptima i bazama podataka uglavnom nisu nezavisni.

Jedan deo atributa u modelu problema često nije u korelaciji sa konceptom i nema isti značaj prilikom klasifikacije novih primera, odnosno primene koncepta. Veliki broj irelevantnih atributa u modelu obično dovodi do preterane detaljizacije (*overfitting*) i loših performansi naučenog koncepta [42], [43].

Algoritmi učenja imaju ugrađene neke mehanizme pojednostavljivanja suviše detaljno naučenog koncepta, ali oni ne mogu dovoljno uspešno da razreše sve situacije, posebno kod problema sa veoma velikim brojem atributa, reda desetina i stotina hiljada.

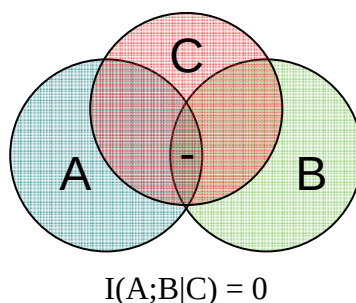
Jedan od pristupa je da se vrši smanjenje dimenzionalnosti problema prethodnom selekcijom potencijalno relevantnih atributa.

Prethodna selekcija podskupa relevantnih atributa zavisi od (1) njihove direktne korelacije sa konceptom, odnosno klasifikacionim atributom i (2) od njihovih međusobnih interakcija, preko kojih grupa atributa može biti u jakoj korelaciji sa konceptom, iako svaki pojedinačno nije u značajnijoj direktnoj korelaciji s konceptom, Sl. 2 [45].



Slika 2: Korelacija međusobno nezavisnih atributa s konceptom

Atributi A i B na slici su međusobno nezavisni. Oba atributa daju informaciju o klasifikaciji C, ali nemaju ništa zajedničko. Mere značajnosti atributa koje istovremeno mere samo značaj pojedinačnih atributa, kao što je npr. *Information gain*, pretpostavljaju ovakvu situaciju.



Slika 3: Korelacija međusobno zavisnih atributa s konceptom

Atributi A i B na slici 2 su međusobno nezavisni u odnosu na klasifikaciju C. Sve što atributi A i B imaju zajedničko je deo informacije o klasifikaciji C.

Ovakvu pretpostavku o uslovnoj nezavisnosti atributa imaju metod *Naive Bayes* i Bayesovske mreže.

Stabla odlučivanja su neefikasna ako atributi nisu uslovno nezavisni u odnosu na oznaku klase. Kada je $I(A;C|B)=0$, atribut B je *irelevantan* za predviđanje C, što je osnova selekcije atributa filtriranjem.

Pošto pojedinačno uklanjanje atributa koji su u grupnoj korelaciji s konceptom može značajno smanjiti performanse naučenog koncepta, neophodno je izvršiti analizu korelacija i identifikovati značajne grupne korelacije.

Neke od metoda analize i upotrebe grupnih korelacija u selekciji atributa date su u [43], [44], [13], [45] i [46], [47].

2.5 GLOBALNE OCENE I GLOBALNE METODE PRETRAŽIVANJA PROSTORA HIPOTEZA

Klasični algoritmi učenja razumljivog znanja, stabala i pravila koriste ocenjivanje hipoteza kao osnovni mehanizam induktivnog učenja stabala i pravila. Element ocenjivanja je i ocena diskriminabilnosti pojedinačnih atributa.

Za ocenjivanje se koriste različite mere kvaliteta, koje ocenjuju informativnost ili entropiju pojedinačnih atributa nezavisno [48], [49], [50], pošto pretpostavljaju njihovu međusobnu statističku nezavisnost.

Za klasu problema gde postoji relativna ili stroga zavisnost atributa, pretpostavke o nezavisnosti više ne važe. Potrebne su mere kvaliteta koje uzimaju u obzir više atributa istovremeno, tako da mogu da detektuju i ocene i ovu vrstu zakonitosti.

Jedan od pristupa rešenju ovog problema je upotreba mera relevantnosti atributa, kao što su *Focus* [51] i *Relief* [52]. Ove mere prilikom estimacije upotrebljavaju više atributa istovremeno i izbegavaju efekt kratkovidosti, odnosno dostizanje lokalnih umesto globalnih maksimuma.

Algoritam *Relief* ocenjuje atribut prema tome kako njegove vrednosti iz obučavajućeg skupa razlikuju primere koji su međusobno slični, odnosno bliski i vrši aproksimaciju razlike verovatnoća

$$F(A) = \frac{P(\text{različita vrednost atributa } A \mid \text{najbliži primer iz različite klase})}{P(\text{različita vrednost atributa } A \mid \text{najbliži primer iz iste klase})}$$

Za svaki primer x , algoritam traži u obučavajućem skupu dva najbliža suseda, jedan iz iste, a drugi iz ostalih klasa (najbliži "pogodak" x_{hit} i najbliži "promasaj" x_{miss}) i računa sumu međusobnih rastojanja ovih vrednosti posmatranog atributa A .

Nakon zadanog broja od n primera, za koje se računa suma rastojanja vrednosti za taj atribut, njihova srednja vrednost predstavlja ocenu atributa A :

$$F(A) = \frac{1}{n} \sum_{i=1}^n -\text{difference}(x, x_{hit}) + \text{difference}(x, x_{miss})$$

Distanca za različite vrednosti diskretnog atributa je 1, a za iste vrednosti je 0, dok je za kontinualne attribute rastojanje razlika samih vrednosti, normalizovana na interval [0..1].

2.6 METOD KOMBINOVANJA KLASIFIKATORA

Učenje redundantnog znanja se koristi u problemima gde je količina informacija u podacima mala (*sparse*), pa se poboljšanje rezultata klasifikacije može postići učenjem više različitih opisa koncepta i kombinovanjem njihovih klasifikacija.

Metod se zasniva na višestrukom ponavljanju procesa učenja, svaki put nad drugačijom particijom skupa obučavajućih primera, skupa atributa ili oba skupa istovremeno, kao i različitih algoritama ili strategija učenja. Svi naučeni klasifikatori formiraju konačni složeni klasifikator (višestruki model ili ansambl (*ensemble*)) i svi se koriste se za klasifikovanje novih primera.

Grupa metoda koje manipulišu primerima ponavljaju proces učenja nad različitim (ne nužno disjunktним) particijama skupa obučavajućih primera. Metodi koje koriste manipulacije skupom atributa za učenje elemenata ansambla koriste različite podskupove atributa.

Metodi koji koriste različite algoritme induktivnog učenja, kombinuju rezultate učenja bitno različitih algoritama ili *strategija* mašinskog učenja (npr., neuronskih mreža, algoritama učenja pravila ili stabala odlučivanja i metoda prepoznavanja uzoraka).

Najpoznatiji su metodi koje koriste manipulisanje primerima: "usrednjavanje" (*averaging*) pomoću Bayesove teorije odlučivanja [19], [53], [54], "natrpavanje" (*bagging*) [55] i "isticanje" (*boosting*) [56], [57], [58].

Rezultati pojedinih metoda se mogu kombinovati na različite načine, pa se mogu se postići velika ukupna poboljšanja performansi [59]. Prema [60], ako se na svakom koraku generišu bitno različiti klasifikatori i ako komponente klasifikatora *nezavisno proizvode svoje greške*, ako je srednja greška klasifikacije *na primeru* manja od 50% očekivana greška *na tom primeru* se može smanjiti do 0 kada broj kombinovanih klasifikatora raste u beskonačnost.

Ako elementi ansambla nisu potpuno nezavisni, stepen usaglašenosti modela u ansamblu se meri kao srednja vrednost uzajamne korelacije ϕ_e svih modela u ansamblu $\mathfrak{F} = \{ \hat{f}_1 \dots \hat{f}_T \}$ [61]:

$$\phi_e(\mathfrak{F}) = \frac{1}{T(T-1)} \sum_{i=1}^T \sum_{j \neq i}^T P(\hat{f}_i(x) = \hat{f}_j(x), \hat{f}_i(x) = f(x))$$

gde $\hat{f}_i(x) = y$ označava da model i klasifikuje primer x u klasu y , a $f(x)$ je ispravna klasa primera x .

Ako je uzajamna korelacija ϕ_e mala, višestruki modeli bitno smanjuju klasifikacijsku grešku.

Tačnost i nesaglasnost elemenata su važne osobine ansambla: *idealno kombinovani klasifikator se sastoji od pojedinačnih klasifikatora visoke tačnosti predviđanja, koji su maksimalno nesaglasni.*

2.6.1 Usrednjavanje (*averaging*)

U ovom pristupu se učenjem u različitim uslovima izgradi više modela i koristi istovremeno za klasifikovanje novih primera. Postoje različiti pristupi izgradnji optimalnog skupa modela namenjenih usrednjavanju, npr. za modele u obliku stabala odlučivanja ili skupova pravila [19], [54], [61].

Novi primeri se klasifikuju u klasu koja se određuje uprosečavanjem rezultata klasifikacije svakog pojedinačnog modela. Klasifikacija svakog modela se uzima u skladu sa nekim težinskim faktorom, koji se određuje na različite načine, a treba da meri koliko dobro pojedinačni model opisuje obučavajuće primere.

Prema Bayesovoj teoriji verovatnoće novi primeri se klasifikuju u klasu sa maksimalnom očekivanom klasifikacijskom tačnošću. Proces Bajesovog usrednjavanja (*Bayes averaging*) podrazumeva računanje ovog očekivanja nad svim mogućim modelima koje razmatra algoritam učenja.

U praksi se računanje vrši samo na osnovu malog skupa najverovatnijih modela. Novi (testni) primer x se klasifikuje u klasu y koja ima najveću očekivanu aposteriornu verovatnoću

$$E[P(y|x, H)] = \sum_{h \in H} P(y|x, h) \cdot P(h|\mathbf{x}, \mathbf{y})$$

- $\mathbf{x}$ je skup od n obučavajućih primera x , čije su oznake klasa y u vektoru $\mathbf{y}$;
- H je skup indukovanih modela, a h je model iz H , odnosno skup indukovanih pravila;
- $P(h|\mathbf{x}, \mathbf{y})$ je *aposteriorna verovatnoća modela h* za zadani obučavajući skup primera i njihove vrednosti oznake klasa;
- $P(y|x, h)$ je *aposteriorna verovatnoća klase* za zadani testni primer x i model h i može se shvatiti kao stepen do kog h potvrđuje klasu y za primer x .

Ukoliko skup H nije konačan ili prebrojiv skup, umesto sume treba računati integral nad svim kontinualnim promenljivama ili koristiti aproksimaciju [19], [54].

Svaki model h (skup pravila) glasa za klasifikaciju novog primera x u neku klasu sa određenom verovatnoćom, uz zadani obučavajući skup $\mathbf{x}, \mathbf{y}$. Primena Bayesovog pravila daje izraz za *aposteriornu verovatnoću modela h* na zadanom skupu primera

$$P(h|\mathbf{x}, \mathbf{y}) = \frac{P(\mathbf{x}, \mathbf{y}|h) \cdot P(h)}{P(\mathbf{x}, \mathbf{y})} \sim P(\mathbf{x}, \mathbf{y}|h) \cdot P(h)$$

Imenilac je zanemaren jer je konstantan za zadani obučavajući skup (tačna vrednost verovatnoće nije potrebna, već samo relativna). $P(h)$ je apriorna verovatnoća, uz pretpostavku o uniformnoj distribuciji modela. Verovatnoća $P(\mathbf{x}, \mathbf{y}|h)$ se naziva *vredostojnošću tvrdnje (likelihood of evidence)* za zadani model. Uz pretpostavku da su svi modeli jednakoverovatni, moguće je računati ovu verovatnoću kao

$$P(\mathbf{x}, \mathbf{y}|h) = \prod_{i=1}^n P(x_i, y_i|h)$$

gde je n broj obučavajućih primera, a verovatnoće $P(x_i, y_i|h)$ se ocenjuju empirijski na osnovu relativnog broja primera klase y_i koje pravilo ispravno klasifikuje. $P(x_i, y_i|h) \approx n_{ci}/n$, gde je n_{ci} je broj ispravno klasifikovanih u klasu y_i , a n ukupan broj obučavajućih primera koje pravilo klasifikuje u klasu y_i .

Novi primer se klasifikuje u najverovatniju klasu, odnosno klasu sa najvećom očekivanom aposteriornom verovatnoćom $E[P(y|x, H)]$.

2.6.2 Natrpavanje (bagging)

Prema [55] metoda "natrpavanja" (*bagging*, od *bootstrap aggregating*) je metoda za generisanje više verzija klasifikatora koji se koriste kao jedinstvena celina, preko mehanizma glasanja. Više klasifikatora se generiše perturbacijom obučavajućeg skupa primera metodom *bootstrap*-ovanja.

Svaki obučavajući skup je nezavisan uzorak primera, odnosno neki primeri su izostavljeni, dok se neki ponavljaju. Kao i kod drugih metoda ansambala, postupak je pogodan za agregaciju rezultata rada "nestabilnih" algoritama, odnosno algoritama

kod kojih male promene u obučavajućem skupu izazivaju velike promene u naučenom skupu pravila. Algoritmi poput neuronskih mreža i algoritama učenja pravila su "nestabilni", dok su metode k najbližih suseda (k -nearest neighbour, skr. kNN) "stabilne".

Postupak natrpavanja se može opisati kao [55]:

Bagging

Atributni model problema zadan preko skupa atributa X i skupa primera $\mathbf{x}$

Ulaz: $\mathbf{x}_t \cup \mathbf{x}_h = \mathbf{x}$ - obučavajući i testni skup primera

A - algoritam učenja

T - broj klasifikatora u ansamblu

1. Za zadani broj klasifikatora T u ansamblu
 - 1.1. Nezavisno generiši obučavajući skup $\mathbf{x}_i \leftarrow \mathbf{x}_t$ (sa ponavljanjem primera)
 - 1.2. Generiši skup pravila za zadani obučavajući skup $\mathbf{x}_i$ (A).
2. Na testnom skupu $\mathbf{x}_h$ izmeri tačnost ansambla (većinskim glasanjem t klasifikatora)

Za učenje razumljivog znanja, obično se preporučuje da vrednost T broja klasifikatora (skupova pravila) u jednom ansamblu bude 10 [55], [62], pošto u praktičnim primenama najveći deo poboljšanja performansi nastaje do tog broja generisanih klasifikatora, a neophodno povećanje obima računanja za jedan red veličine radi dobijanja značajnog poboljšanja performansi je psihološki prihvatljivo.

2.6.3 Isticanje (boosting)

Tehnika "isticanja" [57] ne uzima za svaki element ansambla novi nezavisni slučajni uzorak iz originalnog obučavajućeg skupa, već se svakom primeru dodeljuje težina, koja meri značaj primera u obučavajućem skupu. Nakon učenja svakog klasifikatora iz ansambla, na osnovu njegovih performansi na testnom skupu se ažuriraju težine, tako da se značaj pogrešno klasifikovanih primera povećava, a ispravno klasifikovanih smanjuje.

Konačni kombinovani klasifikator klasifikuje nove primere glasanjem, pri čemu glas svakog pojedinačnog klasifikatora ima težinu koja je funkcija njegove tačnosti na testnom skupu.

Na početku je težina svakog primera x jednaka i iznosi $w_x^1 = 1/n$, gde je n broj primera u obučavajućem skupu. Na svakom koraku t se generiše novi klasifikator, u skladu sa distribucijom primera koju definiše vektor w^t . Greška ε^t novog klasifikatora se meri u skladu sa težinama w_x^t i izračunava se kao suma težina pogrešno klasifikovanih primera.

Kada je $\varepsilon^t > 0,5$, algoritam okončava i daje ansambl sa $T-1$ klasifikatora. Ako klasifikator na koraku t ispravno klasifikuje sve primere, $\varepsilon^t = 0$ i algoritam okončava sa t klasifikatora u ansamblu. Inače, generiše se vektor težina za sledeći korak algoritma w^{t+1} , tako što se težine ispravno klasifikovanih primera množe faktorom $\beta = \varepsilon^t / (1 - \varepsilon^t) < 1$, nakon čega se vrši renormalizacija, tako da suma svih težina u obučavajućem skupu bude 1.

Kombinovani klasifikator ili ansambl izračunava klasifikaciju tako da sumira glasove svih pojedinačnih elemenata, tako da svaki ima težinu $\log(1/\beta^t)$.

U [57] je dokazano da, uz uslov da je uvek $\varepsilon^t < 0,5$ greška konačnog klasifikatora za zadani obučavajući skup i originalnu distribuciju w^1 teži nuli eksponencijalno sa porastom broja klasifikatora u ansamblu.

Takođe je pokazano da je klasifikacijska greška ansambla *na obučavajućem* skupu najviše jednaka vrednosti najtačnijeg klasifikatora, a obično je značajno ispod te vrednosti. Ovakve performanse se ne mogu garantovati i za nove primere.

Jedan od najpoznatijih algoritama isticanja je *AdaBoost*, koji se može opisati na sledeći način [57]:

AdaBoost

Atributni model problema zadan preko skupa atributa A i skupa primera x

Ulaz: $\mathbf{x} = \mathbf{x}_t \cup \mathbf{x}_h$ - skup primera (obučavajući i testni skup klasifikovanih primera)

D - distribucija primera

A - algoritam učenja

T - broj klasifikatora u ansamblu

1. Inicijalizuj težinski vektor $w_i^1 = D(i)$ za $i=1, \dots, n$

2. Za klasifikator $t = 1, \dots, T$

2.1. Postavi vektor distribucije $p^t = \frac{w^t}{\sum_{i=1}^N w_i^t}$

2.2. Generiši skup pravila R_t za zadani obučavajući skup $\mathbf{x}_t$ i distribuciju p^t pomoću algoritma učenja pravila A

2.3. Na testnom skupu $\mathbf{x}_h$ izmeri tačnost klasifikatora prema izrazu

$$\varepsilon^t = \sum_{i=1}^N p^t \cdot |R_t(i) - c(i)|$$

2.4 Postavi $\beta_t = \varepsilon^t / (1 - \varepsilon^t)$

2.5 Postavi nove težine u težinskom vektoru

$$w_i^{t+1} = w_i^t \cdot \beta_t^{1 - |R_t(i) - c(i)|}$$

3. Kombinovani klasifikator klasifikuje nove primere koristeći izraz

$$R(i) = \begin{cases} 1, & \sum_{t=1}^m \left(\log \frac{1}{\beta_t} \right) R_t(i) \geq \frac{1}{2} \sum_{t=1}^m \log \frac{1}{\beta_t} \\ 0, & \text{inace} \end{cases}$$

Postoji više varijanti ovog poznatog algoritma, koji je efikasan, jednostavan, ima dobro formalno proverena svojstva i samo jedan parametar T - broj generisanih klasifikatora (broj iteracija).

U [15] je dokazano da se *AdaBoost* može smatrati jednim predstavnikom iz klase algoritama za učenje slučajnih šuma *Random Forest*.

2.6.4. Slučajne šume (random forests)

Indukcija metodom slučajnih šuma (*random forest induction*, [15]) je jedna od metoda za generisanje kompozitnih klasifikatora koji se sastoje od stabala odlučivanja. Algoritam učenja generiše skup klasifikatora koje koristi za klasifikaciju novih primera nekom od metoda glasanja na osnovu odluka pojedinačnih stabala.

Neophodan i dovoljan uslov da kompozitni klasifikator bude tačniji od svakog od pojedinačnih klasifikatora je da su oni *tačni* i međusobno različiti, odnosno *raznovrsni* [63]. Tačan klasifikator je onaj čija je tačnost predviđanja na novim primerima veća od slučajnog pogađanja. Raznovrsni klasifikatori su oni koji na novim primerima greše na različit način.

Metodi generisanja kompozitnih klasifikatora *bagging* [55] i *boosting* [57] generišu raznovrsnost klasifikatora manipulacijom obučavajućeg skupa.

Raznovrsnost se može povećati manipulisanjem atributima ili uvođenjem slučajnosti u algoritam učenja [64].

Metod slučajnih šuma (*random forests*) povećava raznovrsnost klasifikatora istovremeno na dva načina: (1) resampliranjem podataka i (2) selekcijom podskupova atributa resampliranjem prilikom indukcije stabala.

Procedura se može opisati na sledeći način:

RandomForest

For $i = 1$ to T do:

- formirati skup podataka D_i resampliranjem skupa D s ponavljanjem
- naučiti stablo odlučivanja T_i na osnovu D_i koristeći slučajne podskupove skupa atributa (*randomly restricted feature sets*)

Klasifikovati nove primere koristeći skup od T stabala.

Osnovni algoritam se zasniva na metodu *bagging*-a, koja je dopunjena manipulacijom skupom atributa (deo istaknut je kurzivom). Prilikom izgradnje stabla, na svakom čvoru se bira najbolji atribut iz fiksnog, slučajno odabranog podskupa skupa atributa X .

U proceduri *bagging*-a, skup atributa se ne menja u više prolaza procedure indukcije. U proceduri *RandomForest*, na svakom čvoru se razmatra različit slučajni podskup atributa dimenzija $f(|X|)$ i najbolji atribut se bira za čvor stabla odlučivanja. Slučajni izbor podskupa atributa obezbeđuje povećanje varijabilnosti naučenih ansambala.

3. ANALIZA

U ovom poglavlju se daje pregled postojećih sistema za induktivno učenje i analizira njihova pogodnost za induktivno učenje u uslovima oskudnosti obučavajućeg skupa.

Daje se pregled algoritama učenja razumljivog znanja, uloga i metode selekcije (relevantnih) atributa u algoritmima učenja i pregled postojećih rešenja.

Razmatra se upotreba globalnih ocena kvaliteta pravila i globalnih metoda pretraživanja u algoritmima učenja. Diskutuje se primenjivost i ograničenja metoda kombinovanih klasifikatora u učenju razumljivog znanja.

Na kraju poglavlja se daje kraći pregled postojećih sistema za induktivnog učenje razumljivog znanja.

3.1 ALGORITMI UČENJA RAZUMLJIVOG ZNANJA

Razvoj nekih primena induktivnog mašinskog učenja, pre svega u izgradnji sistema zasnovanih na znanju i otkrivanju znanja u podacima (*data mining*) dovelo je do naglašenog zahteva za razumljivošću naučenog znanja.

Prema stepenu zadovoljenja principa razumljivosti, definisana su tri nivoa ili kriterijuma [7]:

- *slab kriterijum*: sistem koristi primere za generisanje znanja koje poboljšava njegove performanse na novim podacima
- *jak kriterijum*: zadovoljava slab kriterijum i može da saopšti svoje interne procese u obliku koji razume korisnik
- *ultra jak kriterijum*: zadovoljava slab i jak kriterijum, a sistem može da saopšti svoje interne procese u obliku koji čovek može da upotrebi i bez pomoći računara.

Da bi bilo korisno, znanje mora biti i tačno i razumljivo. Tačnost predviđanja se može direktno povezati sa složenošću opisa naučenog koncepta. Teoretske analize *najgoreg slučaja* predviđaju da minimizacija opisa (npr. veličine stabla odlučivanja ili broj selektora u skupu pravila) povećava tačnost predviđanja, odnosno smanjuje grešku klasifikacije na novim primerima. U opštem (ili *srednjem*) slučaju performanse skupa naučenih pravila se ne mogu tako predviđati [65].

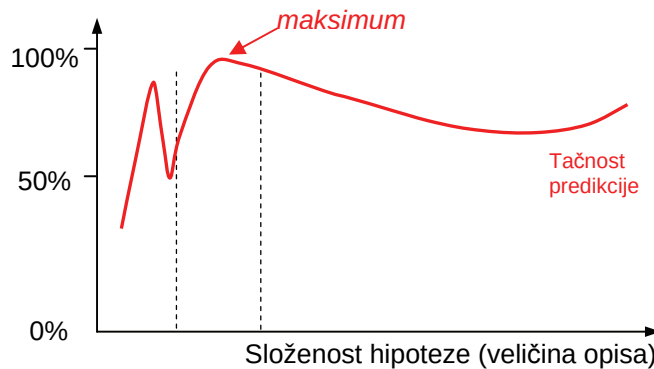
Princip jednostavnosti objašnjenja je važan u nauci još od antičkih vremena [1]. U teoretskim analizama ovaj princip je nazvan "Okamovom oštricom" (*Occam's razor*) [1], [65].

U [65] je na nekoliko konstruisanih i stvarnih problema eksperimentalno proveravana teza da najkraće otkrivene hipoteze konzistentne sa podacima (koje korektno klasifikuju sve primere iz obučavajućeg skupa) imaju i najmanju klasifikacijsku grešku na novim podacima (najveću tačnost predviđanja). Eksperimentalno je ustanovljeno da:

- složeni opisi (npr. velika stabla odlučivanja) obično imaju veliku klasifikacijsku grešku,
- najjednostavniji opisi (npr. minimalna stabla) nemaju istovremeno i najmanju klasifikacijsku grešku,

- najmanju klasifikacijsku grešku imaju opisi nešto složeniji od minimalnih.

Kvalitativni prikaz rezultata iz [65] je kao na Slici 3. Vidi se da za veoma jednostavne opise tačnost predviđanja varira, zatim dostiže maksimum za relativno malu složenost opisa, nakon čega monotonno pada i, za veoma složene opise koncepata, ponovo raste.



Slika 4: Odnos tačnosti predviđanja i složenosti naučenog koncepta

Jedna od ključnih komponenti sistema induktivnog učenja je način predstavljanja znanja- primera i koncepata (zadanih, naučenih ili radnih hipoteza).

Osim prirodnog jezika, u sistemima koji zadovoljavaju ultra jak kriterijum se najčešće koriste stabla i liste odlučivanja, produkciona pravila i različiti formalizmi koji se zasnivaju na matematičkoj logici.

Jedno neformalno praktično uputstvo za formiranje razumljivih logičkih izraza dato je u [5]: osnovne komponente opisa koncepta treba da budu izrazi koji sadrže samo nekoliko uslova u konjukciji (npr., manje od 5), svega nekoliko uslova u disjunkciji, najviše jedan nivo ugneždenosti, najviše jednu implikaciju i najviše dva kvantifikatora, a rekurzija se ne upotrebljava.

3.1.1. Metode za indukciju razumljivih opisa

Razumljivi opisi se mogu dobiti upravljanjem procesom pretraživanja prostora svih mogućih hipoteza, odnosno izborom jednog od mnogo mogućih opisa koji su konzistentni sa podacima [5], [7].

Jedan od faktora za koji se smatra da utiče na razumljivost naučenog znanja je jednostavnost (*simplicity*) opisa koncepta. Jedna od kvantifikacija ovog kriterijuma je tzv. MDL princip (*minimum description length principle*), po kome je smisao generalizacije kompresija informacija. Najbolja generalizacija je ona koja minimizuje sumu dužine opisa teorije (u bitima) i dužine opisa zadanih primera (u bitima) kodiranih uz pomoć te teorije.

Kada se ne raspolaže dovoljnim brojem primera, ova ocena favorizuje suviše jednostavna pravila (opise), koja imaju malu tačnost predviđanja na novim primerima, odnosno teže preteranoj generalizaciji.

MDL princip se, u odsustvu šuma u podacima, svodi na princip "Okamove oštrice". U prisustvu šuma, određena nekonzistentnost i nekompletnost je neophodna [22].

Razumljivost se ne može svesti na jednostavnost, odnosno konciznost opisa koncepta [66]. Neki od faktora razumljivosti su vezani za psihološke aspekte učenja, kao što je, npr. saglasnost sa prethodnim znanjem [67], [68].

U praksi su se rezultati algoritama učenja za neke probleme, uprkos velike tačnosti i dobrih dijagnoza, pokazali neupotrebljivim, zbog neslaganja sa postojećim znanjem, odnosno narušavanjem očekivanja stručnjaka na razmatranom području.

Jedno rešenje je uključivanje posebnog skupa pravila u predznanje o problemu, kojim se ograničava i usmerava rad algoritma induktivnog učenja, tako da se određene klase opisa izuzimaju iz razmatranja.

U [22] se razmatraju kriterijumi za izbor najboljeg pravila u poznatijim algoritmima induktivnog učenja koncepta.

U slučaju kada se zna da u podacima nema šuma, klasični algoritmi induktivnog učenja (npr. ID3, CN2, AQ18) koriste različite varijante ocena pravila koje se zasnivaju na dva osnovna kriterijuma - kompletnosti (*completeness*) i konzistentnosti (*consistency*) pravila.

U stvarnim primenama, podaci su obično zašumljeni (sadrže greške) i nekonzistentni (sadrže kontradikcije), pa se nužno pojavljuju prividne zakonitosti. Konzistentnost i kompletnost skupa pravila tada nisu najvažniji kriterijumi, već to zavisi od problema koji se rešava. Traži se određeni balans između konzistentnosti i kompletnosti, uz poštovanje i ostalih kriterijuma kvaliteta pravila.

Primer je istraživanje podataka (*data mining*), gde se, osim čvrstih zakonitosti, generiše i veliki broj prividnih. Jedan od pristupa rešavanju ovog problema je odustajanje od kompletnosti i oslanjanje samo na konzistentna pravila [22].

Eksperimentalni rezultati primene algoritama učenja su relevantni samo za konkretne i srodne probleme. Prema [69], ne mogu se automatski očekivati iste performanse na drugim problemima, zbog "zakona očuvanja" generalizacionih performansi, po kome:

- ne postoji apsolutno najbolji algoritam učenja za sve probleme,
- algoritam učenja može biti bolji od drugog u jednoj situaciji samo na račun gubitka performansi u nekim drugim situacijama, i
- srednja klasifikacijska tačnost algoritma u odnosu na sve *matematički moguće* probleme je konstantna i ne zavisi od algoritma.

Zbog toga se generalizacione performanse algoritama učenja (pre svega tačnost predviđanja) mere i poboljšavaju u odnosu na određene stvarne probleme, u nameri da se gubitak performansi odrazi na problemima koji se nikad neće javiti u praksi.

3.1.2 Razumljivost pravila u multimodelskim sistemima

Višestruki modeli ili ansambli kombinovanjem klasifikatora poboljšavaju tačnost predviđanja i stabilnost rezultata učenja, ali na račun gubitka razumljivosti naučenog skupa pravila. Samo nekoliko desetina čak i veoma jednostavnih skupova pravila ili stabala prevazilaze ljudske mogućnosti razumevanja i koncentracije [58].

Jedan od pristupa rešavanju gubitka razumljivosti u multimodelskim sistemima je formiranje *metamodela*, koji predstavlja dovoljno dobru aproksimaciju kombinovanog modela, a bitno je jednostavniji [58].

Algoritam učenja koristi obučavajući skup primera za podelu prostora primera u regione koji pripadaju samo po jednoj klasi i može se upotrebiti za modeliranje podele na regione koju definišu kombinovani klasifikatori, učenjem na osnovu slučajno generisanih primera, čije klase određuju ansamblu.

Prilikom generisanja meta-obučavajućeg skupa, koristi se ista distribucija ili njena što tačnija ocena, ako distribucija nije poznata.

Kada osnovni algoritam učenja proizvodi razumljive modele, očekuje se da će se i na osnovu novog skupa primera generisati razumljiv model, iako se očekuje da je kombinovani model nešto složeniji.

3.2 SELEKCIJA ATRIBUTA U ALGORITMIMA UČENJA

Selekcija atributa (*feature selection*, *attribute selection*) je oblast koja se razvijala u okviru prepoznavanja uzoraka [39], matematičke statistike, otkrivanja znanja [4], mašinskog učenja [7], [18], posebno neuronskih mreža i mnogim drugim područjima.

Osnovni zadatak selekcije atributa je redukcija dimenzionalnosti prostora atributa i uklanjanje redundantnih, irelevantnih i zašumljenih podataka, čime se ubrzava rad algoritama učenja, poboljšava kvalitet podataka i povećava tačnost naučenog znanja.

Problem selekcije atributa se može formalno opisati ako se pretpostavi postojanje originalnog prostora atributa $\mathbf{X}$ kardinalnosti d , selektovanog prostora $\bar{\mathbf{X}} \subseteq \mathbf{X}$ kardinalnosti $\bar{d}$ i kriterijuma selekcije $J(\bar{\mathbf{X}})$, koji ocenjuje kvalitet prostora atributa. Selekcija atributa je nalaženje takvog podprostora atributa $\bar{\mathbf{X}} \subseteq \mathbf{X}$ za koji je vrednost kriterijuma $J()$ maksimalna

$$J(\bar{\mathbf{X}}) = \max_{\mathbf{Z} \subseteq \mathbf{X}, |\mathbf{Z}|=\bar{d}} J(\mathbf{Z})$$

Ispitivanje svih mogućih podprostora podrazumeva razmatranje mogućih $\binom{d}{\bar{d}}$ kombinacija. Zbog kombinatorne složenosti problema selekcije, koriste se različiti pristupi nalaženja suboptimalnog rešenja:

1. traži se suboptimalni podskup od $\bar{d}$ selektovanih atributa
2. traži se najmanja dimenzionalnost prostora atributa $\bar{d}$ za koju se postižu performanse iznad zadanog praga (npr. zadana tačnost predviđanja) i
3. traži se suboptimalni podskup atributa po kombinovanom kriterijumu, koji predstavlja kompromis između što manjeg broja selektovanih atributa i što veće tačnosti predviđanja.

Prisustvo irelevantnih i redundantnih atributa negativno utiče na performanse induktivnog učenja [70].

Optimalan podskup atributa sadrži sve relevantne attribute, dok se redundantni i irelevantni atributi izuzimaju iz razmatranja.

U praksi, slabo relevantni redundantni atributi sadrže informacije koje mogu da utiču na poboljšanje klasifikacionih performansi.

U [71] je predložena klasifikacija atributa u četiri disjunktne klase:

- (1) irelevantni (*irrelevant*) atributi
- (2) slabo relevantni redundantni (*weakly relevant and redundant*) atributi
- (3) slabo relevantni neredundantni (*weakly relevant but non-redundant*) atributi i
- (4) jako relevantni (*strongly relevant*) atributi.

Optimalni skup atributa za učenje čine atributi klase (3) i (4), odnosno neredundantni atributi (strogo relevantni i slabo relevantni).

3.2.1 Prethodna selekcija atributa

Metode prethodne selekcije pogodnog podskupa atributa u zadacima induktivnog učenja dele se u tri osnovne grupe, prema međusobnom odnosu selekcije atributa i algoritma učenja: metode filtriranja (*filter methods*), metode prethodnog učenja (*wrapper methods*) i ugrađene metode (*embedded methods*).

3.2.1.1 PSEUDOSLUČAJNA SELEKCIJA (RANDOM SUBSPACE METHOD)

Metod pseudoslučajnog izbora podskupa atributa iz kompletnog, potencijalno veoma velikog skupa, originalno se koristi u algoritmima učenja slučajnih stabala (*Random Trees* [72]) i slučajnih šuma (*Random Forests* [15]).

Metod *u srednjem* čuva relativne odnose međusobnih rastojanja primera iz originalnog prostora, posebno ako se redukcija dimenzionalnosti vrši istovremeno sa učenjem kombinovanih klasifikatora [73], [74], [4].

Iako redukcija dimenzionalnosti metodom kao što je analiza glavnih komponenti za većinu algoritama učenja daje veću tačnost predviđanja nego slučajne projekcije, eksperimenti pokazuju da se povećanjem broja dimenzija, odnosno broja atributa M , ta razlika smanjuje [74], [4].

Osim smanjenja računске kompleksnosti i potencijalno povećanje razumljivosti redukcijom broja atributa, metod slučajnih podprostora je posebno pogodan za učenje razumljivog znanja, jer je *direktno* primenjiv na probleme koji u modelu predstavljanja koriste različite tipove atributa.

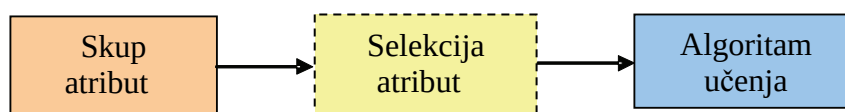
Primena ove metode u učenju ansambala daje poboljšanje performansi učenja, jer povećava međusobnu različitost elemenata ansambla.

Metod slučajnih šuma [15] je jedna od varijanti ovog metoda je, u kojoj se selekcija metodom slučajnog izbora podskupa atributa vrši istovremeno sa izgradnjom osnovnih klasifikatora u obliku stabala odlučivanja.

3.2.1.2 SELEKCIJA FILTRIRANJEM (FILTER METHODS)

Podskup atributa se bira nezavisno od algoritma učenja, na osnovu neke ocene koja rangira sve attribute. To može biti npr. koeficijent korelacije vrednosti atributa sa vrednostima klasifikacionog atributa (klase) [70].

Model selekcije atributa filtriranjem prikazan je na Sl. 5.

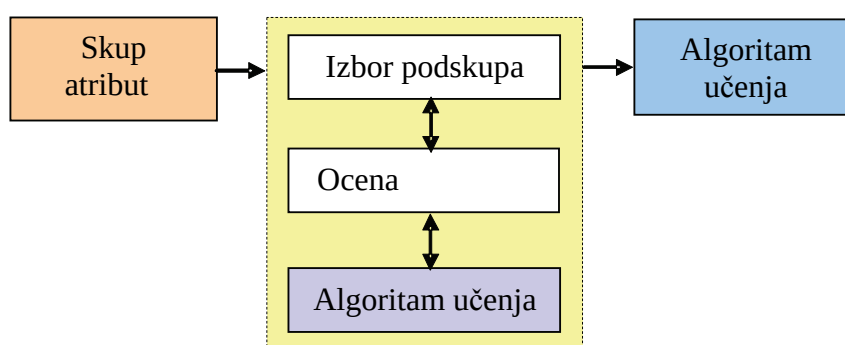


Slika 5: Metod selekcije atributa filtriranjem

3.2.1.3 SELEKCIJA ESTIMACIJOM TAČNOSTI NA VALIDACIONOM SKUPU (WRAPPER METHODS)

Podskup atributa se bira prema estimaciji tačnosti predviđanja koju daje izabrani klasifikator nakon učenja pravila za svaki razmatrani podskup. Učenje pravila se vrši nakon selekcije najbolje ocenjenog podskupa, na osnovu tako redukovano modela.

Iscrpno ispitivanje svih mogućih podskupova je prihvatljivo samo za mali broj atributa, pošto je složenost takvog postupka iz klase složenosti NP [39]. Zbog toga se koriste poznate opšte metode pretraživanja: *best-first*, *branch-and-bound*, *simulated annealing* i genetički algoritmi [38].



Slika 6: Metod selekcije prethodnim učenjem

Za evaluaciju se koriste poznate metode procene tačnosti predviđanja, npr. metod testnog skupa ili unakrsna validacija.

Kao algoritam indukcije se obično koriste algoritmi učenja stabala i pravila, naivni Bayesov klasifikator, linearne diskriminantne funkcije (*least-square linear predictors*) i saport vektori (*support vector machines*, SVM).

3.2.2 Selekcija u toku generalizacije (embeded methods)

Za razliku od selekcije probnim učenjem, koja selekciju atributa posmatra kao spoljašnji sloj procesa indukcije, ugrađene metode selekcije predstavljaju deo osnovnog algoritma indukcije.

Tipični predstavnici ovog pristupa su algoritmi za induktivno učenje stabala odlučivanja i produkcionih pravila (npr. ID3, C4.5 [62], CART [48], AQ-20 [22] i RIPPER). Algoritmi učenja stabala, koji kreiraju stablo od korena prema listovima i algoritmi učenja pravila, koji obično kreiraju konjuktivna pravila dodavanjem jednostavnih logičkih izraza sa samo jednim atributom, prilikom kreiranja novog

čvora ili jednostavnog izraza, koriste funkcije za ocenu i izbor najpogodnijeg atributa za dodavanje u strukturu.

Kada stablo ili skup pravila obuhvataju sve slučajeve iz obučavajućeg skupa, postupak se prekida. Upotrebljeni atributi se smatraju relevantnima, a ostali se izostavljaju iz daljeg razmatranja.

Osim modela sekvencijalne selekcije atributa, postoje težinski modeli, gde se koriste težinske ocene.

U [15] je pokazano kako se algoritam učenja ansambala u obliku slučajnih šuma *Random Forest* može upotrebiti za ocenu važnosti (*importance*) atributa metodom zašumljavanja. Metoda se sastoji u permutovanju vrednosti atributa i učenju slučajnih stabala pre i posle ove promene.

Poređenjem rezultata se dobija mera relevantnosti atributa. Ocena značaja jednog atributa je procentualno smanjenje tačnosti predviđanja slučajne šume nastalo usled njegovog zašumljavanja, dok vrednosti ostalih $M-1$ atributa ostaju nepromenjene. Ova jednostavna estimacija ocenjuje relevantnost, ali ne i redundantnost atributa.

Neke metode učenja neuronskih mreža takođe mogu istovremeno vršiti izbor relevantnih atributa, npr. metoda *Optimal Brain Damage* [75].

3.3 UPOTREBA INTERAKCIJA U SELEKCIJI ATRIBUTA

Složenost analize grupnih korelacija proističe iz velikog broja kombinacija atributa čije interakcije treba razmotriti ($O(2^N)$, gde je N broj atributa u modelu) [51].

Zbog velike složenosti, obično se pribegava aproksimaciji, npr. tako što se izvrši samo delimična analiza korelacije pojedinačnih atributa s klasom ($O(N)$) ili se analiziraju samo neke od mogućih kombinacija (interakcije dužine 2 ili 3 atributa).

Prema [44] interakcije se mogu otkriti pomoću svojstva ireducibilnosti, jer atribut gubi relevantnost kada se uklone atributi koji su s njim u interakciji. U [43] se kao pogodna heuristika za otkrivanje interakcija nudi mera dobitka interakcije (*interaction gain*), pomoću koje se mogu ocenjivati interakcije atributa sa klasom (2-way) i dva atributa s klasom (3-way). U [44] se statistička mera značajnosti koristi za ocenu i prikaz značajnih interakcija u formi grafa interakcija.

Prepoznavanje prisustva interakcije atributa radi prethodne selekcije (*feature selection*) u sistemu *WEKA* [4] se može izvršiti pomoću više metoda:

- *FCBF* [76]

Metod selekcije atributa gde se u dva koraka iz kompletnog skupa atributa eliminišu (1) irelevantni i (2) redundantni atributi.

(1) Prvi korak je uklanjanje irelevantnih atributa. Mera relevantnosti je simetrična neizvesnost (*symmetrical uncertainty*) SU , koja se definiše na osnovu entropije atributa H kao:

$$SU(X, Y) = 2 \cdot \left[\frac{IG(X | Y)}{H(X) + H(Y)} \right]$$

gde su:

$$IG(X | Y) = H(X) - H(X | Y)$$

$$H(X | Y) = \sum_j P(y_j) \sum_i P(x_i | y_j) \log_2 P(x_i | y_j)$$

(2) Na drugom koraku se pomoću heurističke procedure (forma pretraživanja usmerenog grafa, čiji su čvorovi atributi) iz skupa *relevantnih* atributa uklanjanju oni za koje se oceni da su *redundantni*.

Mera redundantnosti je stepen korelacije atributa X sa klasom C i drugim preostalim atributima. Atribut X_i se smatra redundantnim u odnosu na atributa X_j ako važe izrazi:

$$SU(X_i, C) \geq SU(X_j, C) \text{ i } SU(X_i, X_j) \geq SU(X_j, C)$$

Kada se za dva atributa utvrdi da su međusobno redundantni, uklanja se atribut koji je manje relevantan u odnosu na klasu.

Procedura koju koristi FCBF se može ukratko opisati:

Fast Corelation Based Filter

1. Izračunati SU za sve attribute i sortirati ih prema opadajućoj vrednosti SU . Ukloniti *irrelevantne* attribute prema unapred definisanom pragu t_{SU}
2. Počev od prvog atributa, eliminisati sve attribute koji su *redundantni* u odnosu na njega. Ponoviti korak 2 za svaki od preostalih atributa.

- *ReliefF* [50]

Originalni algoritam *Relief* je ukratko opisan u tački 2.5 rada. U [50] je opisana njegova proširena verzija, koja je dobila naziv *ReliefF*.

Algoritam ima dva bitna poboljšanja:

- radi pouzdanije ocene u prisustvu šuma u obučavajućem skupu, umesto udaljenosti do najbližeg i najdaljeg suseda, koristi se prosečna udaljenost do k primera;
- proširena je definicija funkcija rastojanja za slučaj izostavljenih vrednosti u primerima i rešen problem učenja više klasa.

Algoritam ocenjuje i rangira svaki atribut globalnom funkcijom ocene $[-1..1]$.

- *INTERACT* [77]

U radu se ispituje verzija algoritma selekcije atributa na osnovu međusobnih interakcija *INTERACT* [77], koji je razvijen za sistem WEKA.

Rezultati primene ovog algoritma u [77] pokazuju da algoritam značajno smanjuje broj atributa i održava ili poboljšava tačnost predviđanja otkrivanjem interakcija atributa.

U sistemu WEKA postoji realizacija algoritma FOCUS [51], koji sistematski pretražuje prostor atributa proveravajući sve podskupove (dužine $d'=1,..,d$ atributa), dok ne pronade minimalni podskup atributa, koji klasifikuje obučavajuće primere u podskupove primera samo jedne klase. U radu se neće se koristiti zbog velike računске složenosti, zbog koje nije pogodan za veliki broj atributa.

3.4 UPOTREBA GLOBALNIH OCENA KVALITETA PRAVILA I GLOBALNIH METODA PRETRAŽIVANJA

Klasični algoritmi učenja stabala i pravila često podrazumevaju međusobnu statističku nezavisnost atributa, jer za upravljenje pretraživanjem koriste funkcije ocene kvaliteta hipoteza zasnovane na oceni informativnosti ili entropiji *pojedinačnih* atributa.

Za klasu problema gde postoji relativna međuzavisnost ili stroga zavisnost atributa, kao što su "problemi pariteta" (npr. *Monks-2* [78]), ove pretpostavke ne važe. Potrebne su mere kvaliteta koje uzimaju u obzir više atributa istovremeno, tako da mogu da detektuju i ovu vrstu zakonitosti.

Metodi prethodne selekcije atributa *Focus* [52] i *Relief* [50] za ocenu relevantnosti atributa koriste više atributa istovremeno. Ocena je povezana sa stepenom uzajamne nezavisnosti atributa.

Jedan od primera uspešne primene ovih mera je opisan u [79], gde je demonstrirano značajno poboljšanje tačnosti predviđanja algoritma za učenje stabala odlučivanja *Assistant-86* za neke poznate probleme mašinskog učenja. Algoritam je tako izmenjen, da je umesto klasične mere informativnosti atributa, koja ocenjuje svaki atribut nezavisno od drugih, ugrađena varijanta globalne mere kvaliteta, koja u osnovi predstavlja algoritam selekcije relevantnih atributa, nazvan *ReliefF*.

Osnovna ideja originalnog algoritma *Relief* je ocenjivanje atributa prema tome kako njegove vrednosti razlikuju primere koji su međusobno slični, odnosno bliski u prostoru primera.

U razvijenoj varijanti algoritma *ReliefF*, radi pouzdanije ocene u prisustvu šuma u obučavajućem skupu, računa se prosečno rastojanje do zadanog broja k najbližih primera. Takođe je proširena funkcija *difference* za slučaj izostavljenih vrednosti u primerima i rešen problem učenja više klasa.

3.5 PRIMENA METODA KOMBINOVANIH KLASIFIKATORA

Metod se zasniva na višestrukom ponavljanju procesa učenja na osnovu primera, svaki put nad drugačijom particijom skupa obučavajućih primera, skupa atributa ili oba skupa istovremeno. Svi naučeni klasifikatori formiraju konačni složeni klasifikator (višestruki model) i svi se koriste se za klasifikovanje novih primera.

Grupa metoda koje manipulišu primerima ponavljaju proces učenja nad različitim (ne nužno disjunktним) particijama skupa obučavajućih primera. Metodi koje koriste manipulacije skupom atributa za učenje elemenata ansambla koriste različite podskupove atributa.

Metod ansambala je pogodan za učenje pomoću algoritama koji pokazuju svojstvo nestabilnosti, odnosno daju drugačije rezultate učenja za male varijacije obučavajućeg skupa.

3.5.1 Metodi kombinovanja više klasifikatora

Za određivanje ukupne klasifikacije na osnovu pojedinačnih odluka elemenata ansambla koriste se različite tehnike kombinovanja, npr. linearni zbir mišljenja (*linear*

opinion pools), izbor najbolje ocenjene odluke (*winner-take-all*) i probabilističko zaključivanje (*evidential reasoning*).

U prvom modelu, odluka se dobija linearnom kombinacijom odluka pojedinačnih klasifikatora. U drugom modelu, prihvata se odluka jednog, najbolje ocenjenog klasifikatora. U trećem modelu, svaki klasifikator daje težinu svojoj odluci, a konačna odluka se dobija na osnovu neke teorije probabilističkog zaključivanja (npr. teorije Dempster-Shafer).

Kombinovanje klasifikatora podrazumeva kombinovanje ne samo odluka, već i ocena pouzdanosti samih odluka, obično nekom od statističkih metoda, kao što su usrednjavanje [55],[61], izbor jedne najbolje ocenjene odluke ili na osnovu Bayesove teorije [80].

Neophodan uslov za primenjivost ovih metoda je *homogenost* upotrebljenih mera (verovatnoće, rastojanja) ili njihova *konverzija* na istu skalu merenja. Kada se koriste različiti klasifikatori, moguće je da takve ocene i nisu na raspolaganju, pa se za heterogene klasifikatore koriste samo informacije koje su zajedničke za sve klasifikatore, odnosno same klasifikacije.

1. METOD GLASANJA (VOTING)

Metod glasanja ili većinske klasifikacije se sastoji u brojanju odluka pojedinačnih klasifikatora i izboru odluke sa najvećim brojem saglasnih klasifikatora. Veoma je jednostavan za realizaciju i u primeni ne zahteva predznanje o svojstvima klasifikatora.

Nedostatak pristupa je što istu težinu imaju glasovi klasifikatora različite kompetentnosti, odnosno ocene tačnosti.

Predložene su različite varijante metoda glasanja:

- *konsenzus* – kombinovani klasifikator donosi odluku da primer x pripada klasi C_j ako i samo ako je to odluka svih klasifikatora u ansamblu, inače ne donosi sopstvenu odluku za primer x .
- *konsenzus izuzev uzdržanih* – kombinovani klasifikator donosi odluku da primer x pripada klasi C_j ako i samo ako nijedan klasifikator u ansamblu ne klasifikuje primer x u bilo koju drugu klasu različitu od C_j (npr. jer ne donose nikakvu odluku), inače ne donosi sopstvenu odluku za primer x .
- *prosta većina* – koristi se pravilo (proste) većine, gde kombinovani klasifikator donosi odluku da primer x pripada klasi C_j ako i samo ako zadana većina klasifikatora u ansamblu (određena proporcija, npr. više od polovine) klasifikuje primer x u klasu C_j .
- *kvalifikovana većina* – koristi se pravilo većine, gde kombinovani klasifikator donosi odluku da primer x pripada klasi C_j ako i samo ako zadana većina klasifikatora (značajno veća od broja klasifikatora koji primer klasifikuje u druge klase) klasifikuje primer x u klasu C_j .

U [63] je dokazano da, ako svi klasifikatori u ansamblu imaju grešku manju od 50% (tj. tačnost veću od 50%), može se očekivati da tačnost ansambla raste s povećanjem broja klasifikatora u ansamblu.

U dokazu se pretpostavlja da su klasifikatori međusobno nezavisni i da imaju istu tačnost. Zaključak zavisi od činjenice da, ako pojedinačni klasifikatori imaju određenu verovatnoću donošenja ispravne klasifikacije $(1-p)$, verovatnoća da će većinska odluka biti pogrešna je tada

$$\sum_{k=D/2}^D \binom{D}{k} p^k (1-p)^{D-k}$$

gde je D broj pojedinačnih klasifikatora u ansamblu.

2. BAYESOVSKIE METODE KOMBINOVANJA (BAYESIAN ENSEMBLE METHODS)

Metodi glasanja daju istu težinu odlukama svakog elementa ansambla. Koristeći Bayesovu teoremu, moguće je u proces kombinovanja uključiti i pokazatelj uspešnosti elemenata ansambla u prethodnim klasifikacijama.

$$belief(C_i) = \frac{P(e_1(x) = j_1, e_2(x) = j_2, \dots, e_d(x) = j_d \mid x \in C_i) P(x \in C_i)}{P(e_1(x) = j_1, e_2(x) = j_2, \dots, e_d(x) = j_d)}, i = 1, \dots, d$$

Koristi se aproksimacija verovatnoća relativnim frekvencijama.

3. METOD KOMBINOVANJA BEHAVIOR-KNOWLEDGE SPACE

Neophodan uslov i ograničenje za primenu Bayesovskih metoda kombinovanja je međusobna nezavisnost klasifikatora.

Metod *Behavior-Knowledge Space* se ne zasniva na nezavisnosti klasifikatora. Proces odlučivanja se razmatra u d -dimenzionalnom prostoru, gde svaka dimenzija odgovara jednom elementu ansambla. Svaki element donosi jednu od $c+1$ odluka (c klasa i odbijanje donošenja odluke). Presek pojedinačnih odluka elemenata je jedna tačka u prostoru.

Stepen uverenosti u odluku o klasifikaciji primera x u klasu C_i ($x \in C_i$) zasniva se na kombinovanju vrovatnoća pripadnosti pojedinim klasama, $e_k = j_k$; $k = 1..d$ prema izrazu:

$$belief(C_i) = \frac{P(e_1(x) = j_1, e_2(x) = j_2, \dots, e_d(x) = j_d, x \in C_i)}{P(e_1(x) = j_1, e_2(x) = j_2, \dots, e_d(x) = j_d)}$$

u skladu sa izrazom koji se koristi i u Bayesovskom pristupu.

4. METOD KOMBINOVANJA NA OSNOVU TEORIJE DEMPSTER-SHAFER

Primena Bayesove teoreme podrazumeva korišćenje mera uverenosti koje imaju svojstva verovatnoće. Teorija probablističkog odlučivanja Dempstera i Schafera, koja koristi opštije mere stepena uverenosti u neku tvrdnju ne mora da zadovoljava svojstvo aditivnosti.

3.6 PREGLED POSTOJEĆIH SISTEMA ZA INDUKTIVNOG UČENJE RAZUMLJIVOG ZNANJA

Algoritmi induktivnog mašinskog učenja koncepata postoje kao samostalni programi i kao delovi složenijih okruženja i sistema za projektovanje sistema zasnovanih na znanju i istraživanje znanja (*data mining, knowledge discovery*), kao što su *WEKA* [4], *Empiric* [11] i *CLOP* [12].

Prema složenosti koncepata koje su u stanju da nauče, odnosno izražajnoj snazi jezika za opis koncepata, uglavnom se dele na sisteme za *propoziciono* učenje (predikatni račun 0-tog reda) i sisteme za *predikatno* učenje (predikatni račun prvog i, izuzetno, viših redova). Ovakvi koncepti se opisuju izrazima sa rekurzijom i omogućavaju jednostavniji opis složenih objekata.

Najpoznatiji programi za *propoziciono* učenje su:

- **AQ11** - Prema [7], AQ-učenje se zasniva na ideji progresivnog pokrivanja skupa primera za učenje uzastopnim generisanjem novih klasifikacionih pravila. Razvijena je cela familija metoda izvedenih na osnovu algoritma koji je prvi publikovao *R. Michalski* 1969. godine i adaptirao u svrhu mašinskog učenja 1973.

Suština algoritma pokrivanja je u traženju takvog skupa logičkih pravila (konjukcije parova atribut-vrednost ili, u opštem slučaju, proizvoljnih predikata) koji pokrivaju sve pozitivne i nijedan negativan primer koncepta, odnosno klase.

Algoritam AQ generalizuje, korak po korak, opise izabranih pojedinačnih pozitivnih primera zvanih semena (*seeds*). Ovaj postupak omogućava logičko preplitanje pravila, onda kad je to poželjno.

- **ID3** - ID3 je nastao na osnovu algoritma CLS (*Concept Learning System*), kog su objavili istraživači na području psihologije saznanja *Hunt, Marine* i *Stone* 1966. godine. Osnovna ideja algoritma CLS se zasniva na ideji da ljudi grade nove (diskretne) koncepte na osnovu nestrukturiranih podataka tako što razbijaju pojedinačne podatke i koncepte u podskupove sa sličnim karakteristikama.

Krajem sedamdesetih godina, *J. R. Quinlan* je upotrebio ovaj model formiranja koncepta za razvoj programa **ID3** (*Iterative Dichotomizer 3*). Naziv je nastao iz uočene tendencije ljudske spoznaje da formira dihotomije (*dichotomy*), odnosno da primere jedne klase deli na dve razdvojene podgrupe, prema tome da li zadovoljavaju kriterijum iz čvora stabla ili ne. Rekurzivna primena ovog postupka daje stablo učenja, čiji su listovi klase u koje se, nakon ispitivanja vrednosti odgovarajućih svojstava, klasifikuju novi primeri.

Upotrebom heuristike za ocenu i izbor kriterijuma, koja se zasniva na teoriji informacija i opisanom algoritmu selekcije primera, dobijen je ID3, koji se pokazao kao efikasan i uspešan sistem u mnogim područjima primene, te je postao jedan od standarda među sistemima induktivnog mašinskog učenja.

Osnovni algoritam je vremenom doživeo brojna poboljšanja i predstavlja osnovu određenog broja drugih uspešnih sistema induktivnog učenja (*Asisstant-86, C4.5*), a pristupačan je i preko biblioteke objektnih programa *MLC++* i u okviru drugih sistema za induktivno učenje i inteligentnu analizu podataka (*WEKA* [4]).

- **Assistant** – prof. *Ivan Bratko* sa grupom saradnika je razvio sistem **Assistant** [81], reimplementaciju sistema ID3 sa nekoliko važnih poboljšanja: uvedena je binarizacija kontinualnih atributa, potkresivanje nepouzdatih delova stabla, rad sa nepotpunim podacima i primena naivnog Bayesovog klasifikatora kada se pojavi vrednost koje nije bilo u obučavajućem skupu.

Sistem je reimplementiran u još dve varijante: **Assistant-I**, kod koga je poboljšana ocena vrednosti verovanoća u stablu preko tzv. *m*-estimacije i **Assistant-R**, kod koga je poboljšana ocena verovanoća *m*-estimacijom i primenjena sopstvena verzija jednog kriterijuma za razbijanje skupa primera globalnom ocenom pod nazivom *ReliefF* [50], [79], [82].

- **C4.5** - Algoritam C4.5 za indukciju stabala odlučivanja je razvijen na osnovu algoritma ID3, sa više značajnih poboljšanja u odnosu na osnovni algoritam: usavršen rad sa kontinualnim atributima i ispuštenim vrednostima atributa, nova ocena kvaliteta razbijanja skupa primera (*gain ratio*) i pojednostavljivanje naučenog stabla, radi povećanja tačnosti klasifikacije novih primera.

Raspoloživ je kao samostalan program, kao objektni modul (biblioteka MLC++) i u okviru drugih sistema za induktivno učenje i inteligentnu analizu podataka (WEKA [4]).

- **CART** - Algoritam CART je, osim za zadatke klasifikacije, gde se predviđaju vrednosti diskretnih atributa, namenjen za zadatke regresije, odnosno može da predviđa vrednosti kontinualnih atributa.

Osnovni algoritam se zasniva na dihotomizaciji, odnosno binarnom rekurzivnom razbijanju (*binary recursive partitioning*) skupa primera. Metod se sastoji od pravila za razbijanje skupa primera, ocene da li je stablo kompletirano i pravila pridruživanja terminalnih čvorova stabla klasama (za zadatak klasifikacije), odnosno predviđenoj vrednosti (za zadatak regresije).

Sistem CART dozvoljava upotrebu diskretnih i kontinualnih atributa za opis primera, kao i ispuštene vrednosti u primerima.

CART je raspoloživ je kao samostalan program, kao objektni modul (biblioteka MLC++) i kao deo drugih sistema za induktivno učenje i inteligentnu analizu podataka (WEKA [4]).

- **CN2** - Prema [83], ovaj sistem za indukciju pravila na osnovu primera je kreiran radi proširenja prostora pravila koja se mogu naučiti i poboljšanja svojstava učenja osnovnog algoritma pokrivanja familije AQ u prisustvu šuma.

Izlaz algoritma je lista odlučivanja, odnosno uređena lista if-then pravila. Pojednostavljivanje pravila je ugrađeno u algoritam pretraživanja, tako da nije potrebno njihovo naknadno pojednostavljivanje, kao kod AQ serije algoritama.

Otpornost na prisustvo šuma je rešeno po ugledu na *ID3*, dok je rad sa kontinualnim atributima je rešen kao kod sistema *Assistant* (dinamička diskretizacija binarizacijom).

Kriterijum za kvalitet naučenih pravila je informativnost (negativna entropija), dok se značajnost elementa pravila u zvezdi određuje statistički, aproksimacijom χ^2 statistike.

Neki od poznatijih programa za predikatno učenje, odnosno induktivno učenje logičkih koncepata su:

- **FOIL** - Prema [84], FOIL je sistem za učenje složenih koncepata, koji se mogu opisati predikatima prvog reda, odnosno dijalektom programskog jezika *Prolog*. Sistem pripada klasi algoritama *separate-and-conquer*, a predikate formira specijalizacijom ili *top-down* načinom.

FOIL je ograničen na učenje opisa pojedinačnog koncepta na osnovu primera i kontraprimera, odnosno primeri mogu pripadati samo jednoj od dveju klasa. Algoritam koristi pristup pokrivanja pozitivnih primera pravilima, kao i algoritam AQ, ali u unutrašnjoj petlji koristi drugi, iterativni metod za formiranje takvog pravila.

- **FOCL** - Sistem FOCL je razvijen proširenjem sistema FOIL tako da može da koristi predznanje o problemu u obliku pravila [85], [86], kao osnova algoritma multistrategijskog učenja.

Algoritmu FOIL je dograđena komponenta za učenje na osnovu objašnjenja (*explanation based learning*), koja deduktivno izvodi operaciono znanje u obliku Hornovih klauzula na osnovu primera koncepta i (eventualno) predznanja o opisu koncepta.

- **HYDRA** - Sistem HYDRA je razvijen radi otklanjanja nekih bitnih nedostataka sistema FOCL, pre svega osetljivost na problem malih disjunkta [87], pojavu da pravila koja perfektno pokrivaju vrlo mali broj primera iz skupa za učenje imaju slabe performanse na novim primerima i prouzrokuju veliki broj pogrešnih klasifikacija.

Uvedena su unapređenja: (1) mogućnost učenja više od dve klase istovremeno (više koncepata) (2) mera pouzdanosti pravila, koja se koristi prilikom klasifikacije novih primera i (3) posebna mera kvaliteta pravila za izbor najbolje specijalizacije, koja je otpornija na *overfitting*, pojavu nepotrebno detaljnog opisa koncepta, koji proizvodi male dijunkte, posebno u problemima sa velikim procentom šuma u podacima.

Razvijena je i verzija algoritma HYDRA-MM (od *multiple model*), za učenje višestrukih opisa koncepata, čime je popravljena tačnost predviđanja i stabilnost rezultata algoritma učenja.

Metode i sistemi iz ove kategorije se još razvijaju, a veliki broj implementiranih varijanti ovih metoda se nalazi u javno dostupnim alatima upotrebljenim u ovom radu.

4. REŠENJE

U ovom poglavlju se daje predlog jednog rešenja i njegova implementacija u okviru proširenja i unapređenja sopstvenog sistema za otkrivanje znanja *Empiric* [11]. Unapređenje obuhvata više razmotrenih metoda i tehnika istovremeno. Implementirane su metode koje imaju sinergijski efekt i značajno doprinose poboljšanjima rezultata učenja na osnovu oskudnih obučavajućih skupova.

Opisano je unapređenje podsistema za induktivno učenje *Empiric.Rules* i realizacija unapređenih komponenti: (1) selekcije atributa prilikom učenja konjuktivnih pravila, (2) globalnog ocenjivanja kvaliteta pravila, (3) učenja kombinovanih klasifikatora uz različite metode selekcije atributa i (4) modifikovanog osnovnog algoritma pretraživanja prostora hipoteza.

4.1 OPIS PODSISTEMA ZA INDUKTIVNO UČENJE SOPSTVENOG SISTEMA EMPIRIC

Algoritam učenja pravila pripada klasi *separate-and-conquer*, a osnovni oblik algoritma učenja je metod postepenog pokrivanja obučavajćeg skupa, jednu po jednu klasu (*set-covering* ili *separate-and-conquer*):

```
procedure Empiric.Rules
  Rules = {}
  for each class C
    pos = { x ∈ S | class(x) = C }
    neg = { x ∈ S | class(x) ≠ C }
    attr = { x1, .. , xd }
    repeat
      R = BestRule( C, pos, neg )
      pos = pos \ { x ∈ pos | R covers x }
      Rules = Rules ∪ { R }
    until pos={} or R=nil
end
```

Osim osnovnog algoritma učenja *Empiric.Rules*, za implementaciju razmatranih metoda su važni i algoritam klasifikacije ili primene naučenog skupa pravila za zadani skup primera *Empiric.TestRules* [11] i algoritam estimacije tačnosti predviđanja naučenih pravila *Empiric.LearnRules* [11], koji formira skup primera za učenje i testiranje i izračunava tačnost predviđanja na osnovu rezultata algoritma testiranja pravila *TestRules*.

4.2 REALIZACIJA SELEKCIJE ATRIBUTA PRILIKOM UČENJA KONJUKTIVNIH PRAVILA

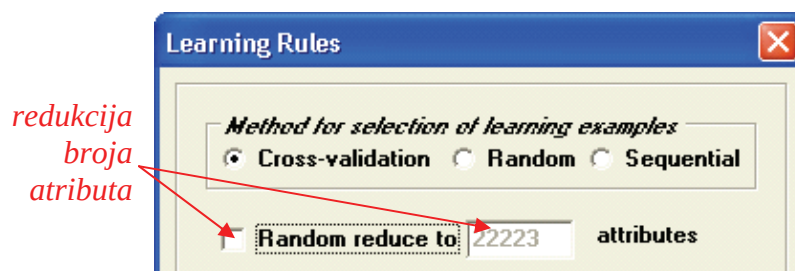
Selekcija atributa u sistemu *Empiric* ugrađena je u algoritam induktivnog učenja pravila u obliku metoda slučajnih podprostora [71], [72], [73], koji je ugrađen u osnovni algoritam učenja:

```

procedure Empiric.Rules
  Rules = {}
  for each class C
    pos = { x ∈ S | class(x) = C }
    neg = { x ∈ S | class(x) ≠ C }
    attr = RandomSubset({ x1, .. , xd }, d')
    repeat
      R = BestRule( C, pos, neg )
      pos = pos \ { x ∈ pos | R covers x }
      Rules = Rules ∪ { R }
    until pos={} or R=nil
  end

```

Može se uključiti prilikom pokretanja algoritma učenja, pri čemu je neophodno zadati redukovani broj atributa d (obično je $d \ll D$), Slika 7.



Slika 7: Selekcija atributa u sistemu Empiric

4.3 REALIZACIJA GLOBALNE OCENE KVALITETA PRAVILA

U podsistem za induktivno učenje pravila sistema *Empiric* prethodno su ugrađene sledeće mere za ocenu kvaliteta pravila [11]:

- Informativnost atributa (*information gain*, ID3)
- Relativna infomativnost atributa (*gain ratio*, C4)
- Entropija (*entropy*, CN2)
- Gini indeks (*gini*, CART)
- Logička pouzdanost (*logical sufficiency content*, HYDRA)
- Mera kvaliteta Kaufman-Michalski (*Q-measure*, AQ18)

U ovom radu je u algoritam učenja *Empiric.Rules* dograđena nova, globalna mera kvaliteta *Relief*. Umesto izbora atributa na osnovu diskriminabilnosti samo jednog atributa, sada se kroz ocenu *Relief* uključuje i uticaj ostalih atributa, Sl. 8.

```

Relief
for j=1 to d do F[j] := 0.0;
for i := 1 to k do
  slučajni izbor pozitivnog primera xi;
  pronalaženje najbližeg suseda primera xi - iz iste klase xhit i ostalih klasa xmiss;
  for j:= 1 to d do
    F[j]= F[j] - distance(j, xi, xhit)/k + distance(j, xi, xmiss)/k

```

Slika 8: Realizacija globalne ocene kvaliteta pravila Relief

Mera je ugrađena u osnovni algoritam izgradnje konjuktivnog izraza, koji opisuje podskup primera jedne klase nasuprot svim ostalima.

Pogodnost *set-covering* klase algoritama što se uvek razmatra slučaj dve klase - pokrivanje primera tekuće klase u odnosu na sve ostale primere, pa se može koristiti osnovna varijanta algoritma *Relief*.

4.4 UČENJE KOMBINOVANIH KLASIFIKATORA I SELEKCIJA ATRIBUTA

U sistemi *Empiric.Rules* je ugrađeno učenje ansambala metodom *bagging* [11]. Sada omogućeno učenje ansambala istovremeno sa selekcijom atributa metodom slučajnih podprostora, čime je povećana međusobna nesaglasnost osnovnih klasifikatora (skupova pravila).

Metod pseudoslučajnog izbora podskupa atributa iz kompletnog, potencijalno veoma velikog skupa, originalno je primenjen u algoritmima učenja stabala (*Random Trees* [72] i *Random Forests* [15]) koji pripadaju klasi *divide-and-conquer*.

U ovom radu je upotrebljen sličan način kombinovanja (metod *bagging*), ali u algoritmu učenja produkcionih pravila, koji se zasniva na metodu pokrivanja skupa primera (*set-covering method*) i pripada klasi *separate-and-conquer*.

4.4.1 REALIZACIJA ISTOVREMENE PRIMENE SELEKCIJE ATRIBUTA I FORMIRANJA ANSAMBLA

Osnova za kombinovanje klasifikatora je postojeći kod za većinsko glasanje, koji se u sistemu *Empiric.Rules* koristi za klasifikaciju testnih primera, prilikom estimacije tačnosti osnovnih i kombinovanih skupova pravila dobijenih metodom *bagging*.

Ugrađen je novi algoritam formiranja pojedinačnih klasifikatora, koji manipuliše atributima tako što formira (potencijalno preklapajuće) pseudoslučajne podskupove atributa metodom slučajnih podprostora (*random subspace method*) primenjenom na algoritam učenja produkcionih pravila.

Proces formiranja i proces kombinovanja klasifikatora se posmatraju odvojeno. Kao i kod metoda slučajnih šuma, prvo se manipuliše podacima tako što se formiraju obučavajući i testni skup, a zatim se nad svakim od njih formiraju slučajni podmodeli, izvrši učenje i formiraju ansambli nad istim obučavajućim skupovima.

Na taj način proces učenja skupa pravila metodom pokrivanja skupa (*set-covering*) za opis svakog podskupa primera koristi drugi slučajni podprostor.

Izbor obučavajućih primera se vrši na neki od postojećih načina, slučajnim ili sekvencijalnim izborom testnog skupa ili formiranjem kros-validacionih disjunktih podskupova.

4.5 MODIFIKACIJA OSNOVNOG ALGORITMA PRETRAŽIVANJA

Osnovni algoritam učenja je brz, ali nedovoljno efikasan heuristički algoritam usmerenog pretraživanja prostora mogućih hipoteza [11].

Osnovni nedostatak ove klase algoritama je često prevelika redukcija prostora razmatranih hipoteza zbog praćenja samo jednog gradijenta kriterijumske funkcije ocene kvaliteta hipoteze i okončavanje postupka nalaženjem lokalnog maksimuma [88].

Manja poboljšanja se mogu postići upotrebom globalnih funkcija ocene, koje koriste više od jednog atributa istovremeno i povećavaju verovatnoću nalaženja globalnog maksimuma funkcije ocene kvaliteta hipoteze, odnosno najbolje moguće hipoteze prema izabranom kriterijumu [79], [89].

Dodatno proširenje broja razmatranih hipoteza i povećanje verovatnoće nalaženja globalnog maksimuma kriterijumske funkcije se može postići modifikacijom osnovnog algoritma pretraživanja, ako se značajne ne poveća računaska složenost.

Prvi sledeći algoritam pretraživanja po računskoj složenosti je *BeamSearch* [90], koji razmatra više od jedne (samo trenutno najbolje) alternative na svakom čvoru stabla pretraživanja prostora hipoteza.

U sistemu *Empiric.Rules*, ostavljena je mogućnost izbora algoritma pretraživanja, kao parametra algoritma učenja.

5. PRIMERI PRIMENE

U ovom poglavlju se daju eksperimentalni rezultati primene unapređenog sistema *Empiric* na više standardnih problema induktivnog učenja iz baze UCI (*University of California at Irvine*) i tri realna problema iz oblasti analize genskih ekspresija, koji su oskudni i za koje nijedna od poznatih metoda induktivnog učenja razumljivog znanja ne daje zadovoljavajuće rezultate.

Opisana je metodologija izvođenja eksperimenata i upotrebljeni problemi induktivnog učenja. Prikazani su rezultati induktivnog učenja i estimacije najvažnijih performansi naučenog znanja.

Na kraju poglavlja je diskusija rezultata učenja i poređenje sa dostupnim podacima iz literature.

5.1 OPIS METODOLOGIJE IZVOĐENJA EKSPERIMENATA

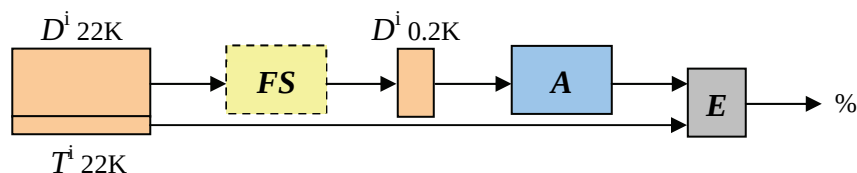
Upotrebljene su metode induktivnog učenja stabala odlučivanja (*decision trees*) i produkcionih pravila (*production rules*), kao i pomoćne metode za redukciju dimenzionalnosti u uslovima oskudnosti obučavajućih skupova, koje su implementirane u okviru sistema *WEKA* [4], *EMPIRIC* [11] i *Matlab* programskim bibliotekama *SPIDER* i *CLOP* [12].

5.1.1 Model eksperimenta

Koristi se klasičan model induktivnog učenja na osnovu svih raspoloživih obučavajućih primera istovremeno (*batch learning*).

5.1.2 Estimacija tačnosti i razumljivosti

Induktivno učenje je izvršeno uz estimaciju tačnosti naučenog znanja metodom dvostruke unakrsne validacije [9]. Unakrsna validacija, primenjena na problem induktivnog učenja u uslovima oskudnosti, gde se obično koristi i neka od metoda prethodne selekcije atributa (*feature selection*) *FS*, podrazumeva da se na svakom koraku unakrsne validacije pre bilo kakve akcije nad primerima D odvoji testni skup primera T^i , koji ni na koji način neće učestvovati u učenju pravila čija će se tačnost ocenjivati, Sl. 9.



Slika 9: Metod dvostruke unakrsne validacije prilikom prethodne selekcije atributa

Nakon toga se algoritmi prethodne selekcije atributa *FS* i induktivnog učenja stabala/pravila *A* sekvencijalno primenjuju na generisani podskup D^i osnovnog

obučavajućeg skupa D i za sve druge estimacije se koristi nova petlja unakrsne validacije koja polazi od podskupa primera D^i kao osnovnog.

Konačna estimacija tačnosti naučenog znanja ovom metodom se dobija se testiranjem nad testnim skupom primera T^i , koji nisu upotrebljavani u procesu selekcije atributa i učenja pravila [9], [4].

U svakom prolazu se 90% ukupnog broja primera koristi za učenje (sa i bez redukcije dimenzionalnosti), a preostalih 10% samo za estimaciju tačnosti.

Postupak se, za primer učenja genskih ekspresija D i algoritam učenja A , ukratko može opisati:

```
for i=1 to 10 do  
    slučajni izbor  $0.9D_{22K}^i$ ;  
    redukcija dimenzionalnosti  $[0.9D_{200}^i]$ ;  
    induktivno učenje( $A, 0.9D_{22K}^i$  ili  $0.9D_{200}^i$ );  
    estimacija tačnosti( $A, 0.1D_{22K}^i$ );
```

Estimacija tačnosti predviđanja se zasniva na (pseudo)slučajnoj podeli raspoloživog skupa primera na disjunktne podskupove.

U sistemu *WEKA* se prilikom estimacije tačnosti metodom unakrsne validacije pseudoslučajna podela na primere za učenje i primere za testiranje uvek izvrši na *isti* način, tako da je i estimacija tačnosti *jedinstvena*. Sistem *WEKA* naučeno znanje, odnosno model, uvek generiše upotrebom svih raspoloživih primera iz obučavajućeg skupa, tako da se u ponovljenim eksperimentima dobija isto rešenje, pa je estimacija složenosti takođe *jedinstvena*.

Sistem *Empiric* omogućava generisanje konačnog rešenja na osnovu celog obučavajućeg skupa, ali je sama pseudoslučajna podela skupa primera za potrebe unakrsne validacije *različita* u svakom eksperimentu, tako da *estimacija tačnosti*, kao i složenosti *nije jedinstvena*.

Estimacija tačnosti u slučaju originalnog Breiman-Cutler programa *RandomForest* izvršena je po metodologiji estimacije greške ugrađenoj u sam program (*out-of-bag estimate*²), koja je po navodima autora dokazano objektivna [15], [17] tako da nije potrebna posebna estimacija pomoću posebnog testnog skupa ili unakrsne validacije.

Pošto se uslovi i algoritmi estimacije tačnosti predviđanja u sistemima *WEKA*, *Empiric* i originalnom Breiman-Cutler programu *RandomForest* razlikuju, dobijene ocene se neće direktno upoređivati, već će se prikazivati zasebno.

5.1.3 Redukcija dimenzionalnosti

Svi problemi genskih ekspresija su ispitani na celom skupu od $M \cong 22K$ atributa i na redukovanom skupu od $d=200$ atributa (oko 1% ukupnog broja).

² Prema ovoj metodologiji, svako stablo se indukuje nad 2/3 slučajno odbranih primera iz obučavajućeg skupa, a preostala 1/3 primera se koristi samo za estimaciju greške. Tako je svaki primer u oko 1/3 slučajeva izostavljen iz učenja i koristi se samo za testiranje i estimaciju greške naučenog stabla. Prosečna proporcija pogrešne klasifikacije svih primera za svaku od klasa je *oob estimacija* tačnosti predviđanja.

Ovaj broj atributa, koji se koristi za sve metode selekcije atributa, određen je eksperimentalno. Početna najniža vrednost je izabrana na osnovu heuristike iz [16], [17], gde se koristi vrednost $d = \sqrt{D} = \sqrt{22.215} \approx 149$.

Redukcija dimenzionalnosti je izvršena pomoću više metoda selekcije atributa raspoloživih u sistemima WEKA i EMPIRIC:

1. Metodom filtriranja (*filter methods*)
2. Metodom prethodnog učenja (*wrapper methods*)
3. Ugrađenim metodama u algoritam učenja Empiric.Rules (*embeeded methods*)

5.1.3.1. METOD FILTRIRANJA (*FILTER*)

Selekcija atributa pomoću različitih tehnika filtriranja, koje vrše rangiranje atributa prema oceni značajnosti, u sistemu WEKA je izvršena tako da su atributi ocenjeni i rangirani na trenutnom obučavajućem skupu D_{22K}^i , nakon čega je izvršena selekcija 200 najbolje rangiranih atributa.

Za ocenjivanje značajnosti (ranga) atributa korišćeni su sledeći metodi:

1. Random subspace [71], [72], metod *slučajnog izbora*
2. ReliefF [50], metod *Ranker*
3. FCBF [76], metod *Symmetrycal Uncertainty Attribute Set Eval* [76]
4. INTERACT [77], metod *Symmetrycal Uncertainty Attribute Set Eval* [76]

Metod slučajnih podprostora je upotrebljen u okviru sopstvenog sistema *Empiric*.

5.1.3.2. METOD PRETHODNOG UČENJA (*WRAPPER*)

Selekcija podskupa atributa na osnovu rezultata učenja pogodnim, obično znatno jednostavnijim algoritmom učenja, izvršena je u sistemima *Empiric* i WEKA.

U izvršenim eksperimentima ovaj metod je korišćen radi početne estimacije broja atributa za metod filtriranja, tako što je izvršeno probno učenje pomoću klasičnih algoritama učenja u sistemima WEKA (*C4.5*, *C4.5RULES*, *RIPPER*), *Empiric* (*Empiric.Rules*) i *Feature Selection Tool* [13], [14] na celom i tekućem krosvalidacionom skupu primera.

Analiza upotrebljenih atributa u naučenim listama, stablima i pravilima, kao i dijagramima programa *Feature Selection Tool* osnov je za estimaciju parametara selekcije atributa u poglavlju 5.3.

Upotreba ovog metoda za izbor konkretnog podskupa atributa pomoću složenijih algoritama učenja računski je suviše složena za realne probleme sa velikim brojem atributa, kao što su npr. problemi učenja na osnovu genskih ekspresija.

5.1.3.3. UGRAĐENI METODI (*EMBEEDED*)

Selekcija atributa u sistemu *Empiric* je izvršena ugrađenim metodom slučajnih podprostora, koji se može koristiti istovremeno sa sa metodom za učenje ansambala *bagging*.

Na svakom koraku formiranja pojedinačnih pravila slučajno se bira relativno mali podskup atributa, na osnovu kojeg se metodom pokrivanja obučavajućeg skupa (*set-covering*) formira konjuktivno pravilo samo od izabranih atributa. Npr. u problemima učenja genskih ekspresija je izabrano ograničenje od maksimalno po 200 *pseudoslučajno* izabranih atributa na svakom koraku formiranja pravila.

Ova tehnika omogućava poboljšanje tačnosti predviđanja naučenih pravila, kao i potencijalno povećanje razumljivosti zbog jednostavnijih pravila naučenih na osnovu manjeg skupa atributa.

Pravila će biti jednostavnija ako se indukuju dovoljno tačni skupovi pravila, odnosno ako ne dođe do prevelikog smanjenja tačnosti osnovnih modela usled detaljizacije (*overfitting*) ili generalizacije (*underfitting*), koja se ne može nadoknaditi relativno malim brojem elemenata ansambla koji je u ovom prihvatljiv (10-50).

5.1.4 Upotrebljeni programi

Prema opisanoj metodologiji ispitane su raspoložive simboličke metode induktivnog učenja iz sistema *WEKA* (učenje pravila i stabala), sistema *Empiric* (učenje pravila, odnosno liste odlučivanja) i više javno dostupnih softverskih alata za otkrivanje znanja (selekcija atributa i induktivno učenje) u obliku biblioteka programa za okruženje *Matlab*.

U radu je korišćen sledeći softver:

- sopstveni sistem za otkrivanje znanja *Empiric*, sa podsistemom za učenje pravila *Empiric.Rules* [91], [11],
- sistem za otkrivanje znanja otvorenog Java koda *WEKA* [4], sa više metoda induktivnog učenja razumljivog znanja: stabala metodom *C4.5* i pravila metodom *C4.5RULES* i *RIPPER*. Korišćeno je više metoda selekcije atributa filtriranjem () i nekoliko metoda učenja kombinovanih klasifikatora (*Random Forest*, *Bagging*, *AdaBoost* i *MultiBoost*) na osnovu stabala i pravila.

Za kontrolu rezultata je upotrebljena i implementacija metoda učenja Bayesovog klasifikatora (*NaiveBayes*).

- metod redukcije dimenzionalnosti i selekcije atributa *SIMBA* razvijen u kodu sistema *Matlab* [13], [14], kao i neki metodi iz biblioteka programa za sistem *Matlab SPIDER* i *CLOP* [12].
- *Matlab* softver za estimaciju optimalnog broja atributa *Feature Selection Tool* [13], [14].
- originalna verzija programa *RandomForest* (u jeziku FORTRAN) autora Lea Breiman-a i Adele Cutler sa univerziteta Berkeley [15], [17].

5.1.4.1 SISTEM EMPIRIC

Sistem za induktivno učenje na osnovu primera *Empiric* je nastao kao alat u zadacima istraživanja podataka (*data mining*), odnosno otkrivanja znanja (*knowledge discovery*).

Sistem *Empiric* se sastoji od više podsistema, koji obezbeđuju minimalni skup alata za inteligentnu analizu podataka metodama induktivnog mašinskog učenja:

- Editor primera, namenjen za unos i ažuriranje modela problema i primera;
- Vizualizator, realizovan algoritmom za nelinearnu projekciju prostora primera (*nonlinear mapping*) u dve dimenzije, radi prikaza strukture skupa primera geometrijskim rasporedom tačaka na površini. Omogućava praćenje rada algoritama za generisanje klasifikacija.
- Podsystem za induktivno učenje bez učitelja (*unsupervised learning*):
 - generator jednostavnih klasifikacija, realizovan algoritmom partitivnog grupisanja (*partitional clustering*);
 - generator hijerarhije klasa, realizovan algoritmom hijerarhijskog grupisanja (*hierarhical clustering*);
- Podsystem za induktivno učenje pravila na osnovu primera (*induction of conjunctive rules*);

Svi podsystemi su prilagođeni jedinstvenom modelu predstavljanja znanja, tzv. atributnom modelu sa više tipova atributa. Sistem omogućava korišćenje *kontinualnih* i *diskretnih* atributa. Diskretni atributi su ugrađeni u tri varijante: *nominalni* (neuređen skup vrednosti), *linearni* (uređen skup vrednosti) i *strukturni* (parcijalno uređen skup vrednosti).

Osnovne karakteristike sistema *Empiric* u celini su:

- za predstavljanje primera i naučenog znanja koristi uniformni atributni model primera sa više tipova atributa neophodnih za istraživanje znanja u različitim oblastima ljudske delatnosti;
- Prihvata i razmatra ispuštene, odnosno nepoznate vrednosti u primerima za sve funkcije sistema;
- Vizuelno prikazuje klasifikaciju obučavajućih primera radi boljeg uvida u strukturu i rad generatora klasifikacija;
- Omogućava odabir primera za učenje i testiranje pravila (slučajno i sekvencijalno) i postavljanje parametara testiranja;
- Kod učenja pravila vrši automatsku dinamičku diskretizaciju kontinualnih atributa sopstvenim metodom rekurzivne podele intervala, tako da se u svim funkcijama sistema mogu direktno i ravnopravno koristiti numerički i nenumerički podaci;
- Generisano znanje predstavlja u obliku konjunktivni pravila, sa internim disjunkcijama u elementarnim iskazima (selektorima). Svaka klasa se u skupu pravila predstavlja jednim ili logičkom disjunkcijom više pravila;
- Omogućava povećanje tačnosti u slučaju oskudnih obučavajućih skupova metodom selekcije atributa (metod pseudoslučajne selekcije) i metodom rangiranja ugrađenim algoritmom *Relief*.
- Omogućava povećanje tačnosti predviđanja učenjem ansambala metodom *bagging*.

U podsystem za induktivno učenje pravila sistema *Empiric* je ugrađeno više različitih lokalnih mera za ocenu kvaliteta pravila [11]: informativnost atributa (*information gain*), relativna informativnost (*gain ratio*), entropija (*entropy*), Gini indeks (*gini*

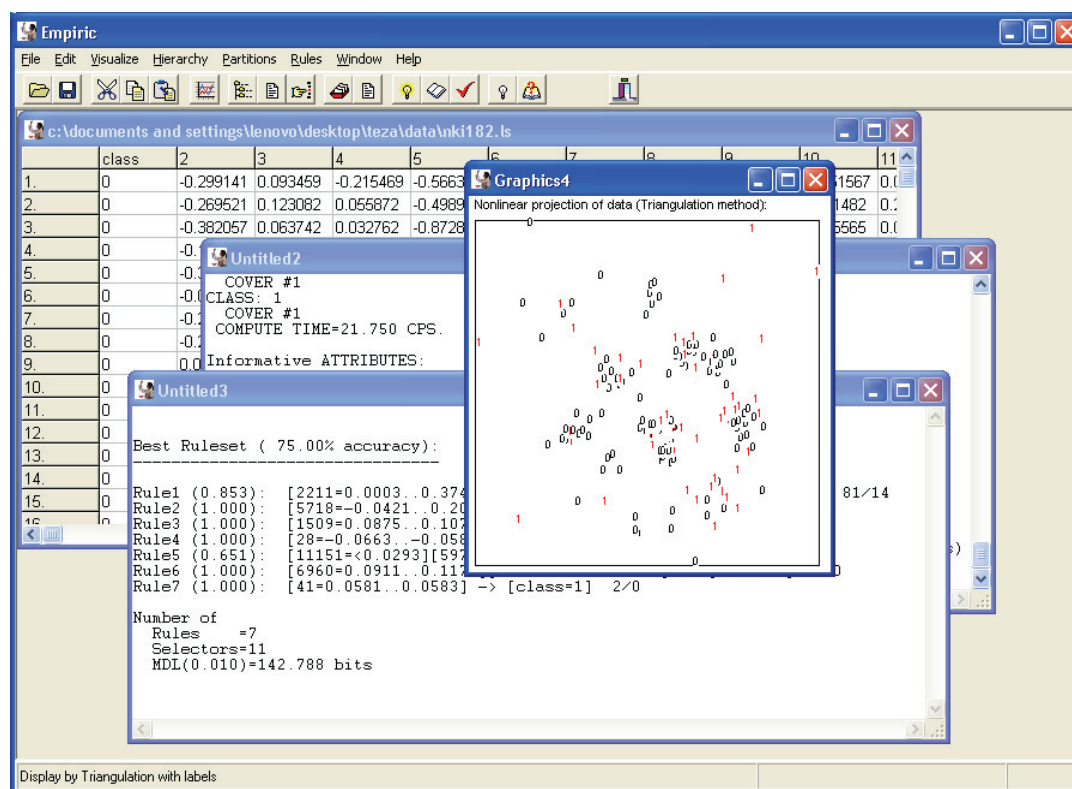
index), logička pouzdanost pravila (*logical sufficiency content*) i mera kvaliteta Kaufmana-Michalskog (*Q measure*)³.

Svaka od ovih mera se može upotrebiti za generisanje pojedinačnih ili višestrukih modela, odnosno jednostavnih ili višestrukih skupova pravila, sa i bez prethodne selekcije atributa.

U algoritam učenja je ugrađena automatska dinamička diskretizacija kontinualnih atributa sopstvenom metodom rekurzivne podele intervala [11].

Prvi prototip sistema za inteligentnu analizu podataka pod nazivom *Empiric* autor je razvio na Katedri za računarsku tehniku Visokih vojnotehničkih škola KoV JNA u Zagrebu 1989-1991. godine (verzija za DOS). Novu verziju sistema za Windows okruženje, u kojoj su dograđena poboljšanja osnovnog algoritma za induktivno učenje pravila (više ocena kvaliteta pravila, pojednostavljivanje pravila i učenje kombinovanih modela ili ansambala) urađena je kao deo izrade magistarskog rada [11]:

Radi ilustracije realizacije grafičkog interfejsa, na Slika 10 je dat primer izgleda ekrana sistema *Empiric* prilikom analize problema **NKI 182**.

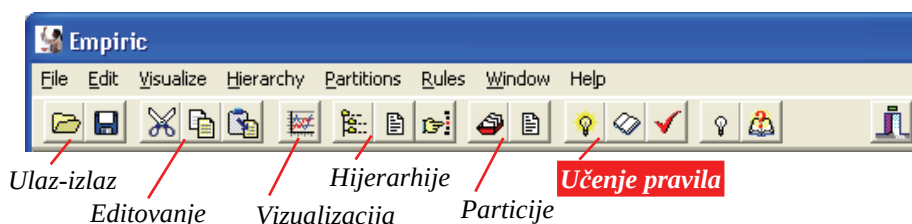


Slika 10: Izgled ekrana sistema *Empiric* prilikom analize problema NKI BC

Na Slika 10 se vidi kako podsistem za vizualizaciju prikazuje strukturu primera ovog problema, koji su razvrstani u dve klase, na prikazu označene sa **0** i **1**.

Algoritam učenja pravila se pokreće iz osnovnog menija sistema **Rules** | **Learn** ili aktiviranjem tastera , Slika 11:

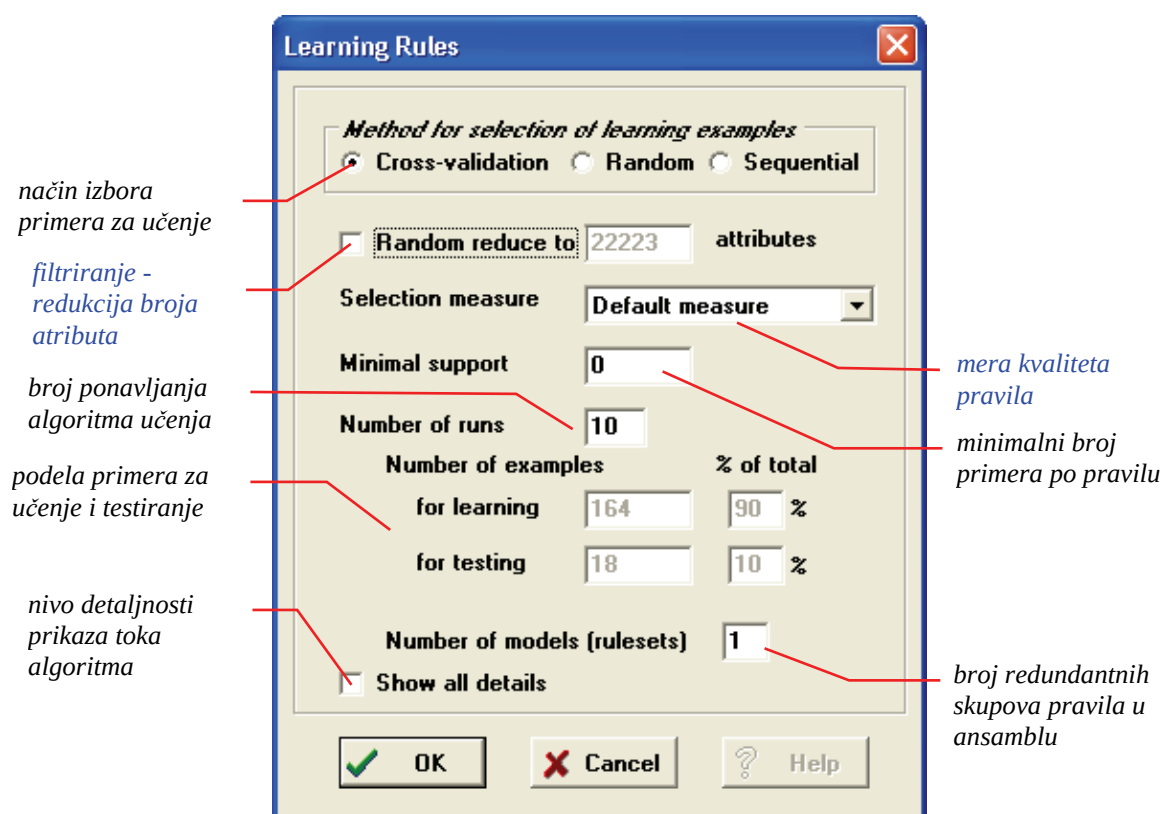
³ u ovom radu je dograđena i globalna mera *Relief*



Slika 11: Pokretanje induktivnog učenja pravila (Empiric.Rules)

Model predstavljanja znanja je tzv. atributni model sa više tipova atributa. Sistem razlikuje kontinualne i tri vrste diskretnih atributa: *nominalne* (neuređen skup vrednosti), *linearne* (totalno uređen skup vrednosti) i *strukturne* (parcijalno uređen skup vrednosti).

U sistemu *Empiric* meru kvaliteta pravila korisnik bira interaktivno, prilikom aktiviranja algoritma učenja pravila, zajedno sa načinom izdvajanja primera za učenje i brojem primera za učenje, odnosno testiranje naučenih pravila. Na Slika 12 je prikazan izgled dijaloga u kome se zadaju svi bitni parametri učenja pravila.



Slika 12: Postavljanje parametara induktivnog učenja pravila

Uputstvo za upotrebu sistema *Empiric.Rules* dato je u *Prilogu 1* ovog rada.

5.1.4.2 SISTEM WEKA

Sistem za istraživanje znanja otvorenog koda [4], razvijen u jeziku Java.

Obezbeđuje veoma veliki broj različitih metoda za istraživanje podataka: za pretprocesiranje, učenje klasifikacija, generisanje klasifikacija (*clustering*), selekciju atributa i vizualizaciju primera i rezultata učenja.

U ovom radu se iz verzija sistema 3.5.7 koriste:

- algoritmi učenja stabala i pravila C4.5, C4.5Rules i RIPPER (koji su u sistemu nazvani *J48*, *PART* i *JRip*) i metod učenja slučajnih šuma *RandomForest*.
- metodi učenja ansambala *Bagging*, *AdaBoost.M1* i *MultiBoost.AB*, koji su u preliminarnim eksperimentima pokazali uspešnost u poboljšanju generalizacionih performansi učenja na osnovu oskudnih obučavajućih skupova.
- metodi selekcije atributa, koji su u prethodnim eksperimentima pokazali najbolje rezultate u poboljšanju generalizacionih performansi oskudnih skupova: *Fast Corelation Based Filter* (FCBF) [76] i *Relief* [50]. Sistem je dopunjen i novim metodom *INTERACT* [77].

Svi metodi se kombinuju pomoću ugrađenog Java objekta *AttributeSelectedClassifier*, koji omogućava selekciju atributa samo na tekućem krosvalidacionom podskupu obučavajućih primera.

Kratki prikaz sistema je dat u *Prilogu 2* ovog rada.

5.1.4.3 SOFTVER ZA MATLAB

Korišćeni su paketi softvera razvijenih za okruženje *Matlab*:

- *SIMBA* [13], [14]

Simba je algoritam za prethodnu selekciju atributa koji maksimizuje funkciju ocene kvaliteta skupa atributa zasnovanu na konceptu margine [13], [14].

U radu se koristi za detaljnije poređenje estimacije tačnosti predviđanja i optimalnog broja atributa metoda koji su korišćeni u radu.

- Feature Selection Tool [13], [14]

Program za prethodnu selekciju atributa, koji se u radu koristi za procenu efekata prethodne selekcije atributa metodom probnog učenja pomoću više metoda selekcije i algoritama učenja.

U radu se za estimaciju koristi samo algoritam 1-NN i metode selekcije *InfoGain*, *Relief* i *Simba*.

Kratki opis sistema je u *Prilogu 3* ovog rada.

5.1.4.4 RANDOMFOREST (BREIMAN-CUTLER)

Originalna implementacija višefunkcionalnog programa za indektivno učenje slučajnih šuma autora *L. Breimana* i *A. Cutler*, čiji je izvorni kod u jeziku FORTRAN javno dostupan [17].

U ovom radu se koristi samo radi poređenja, zbog drugačije metode estimacije tačnosti.

5.2 OPIS IZABRANIH PROBLEMA INDUKTIVNOG UČENJA

Razmotrena su po tri konstruisana problema sa karakteristikama oskudnosti (*Monks-2* [78], *Waveform-40* [48], *Breiman1K* [15]), tri standardna problema iz baze podataka University of California at Irvine [92] (*Breast Cancer LJ*, *Cleveland Heart Disease* i *Lung Cancer*) i tri aktuelna problema analize genskih ekspresija [29], [9], [93] korišćenih u istraživanjima iz oblasti dijagnostike i prognostike raka dojke.

5.2.1 Konstruisani merni problemi

1. Problem *Monks-2* [78] je problem klasifikacije objekata (osoba) u vidokrugu zamišljenog robota na osnovu 6 nominalnih atributa, definisanih kao:

Head-shape	∈ {round, square, octagon}
Body-shape	∈ {round, square, octagon}
Is-smiling	∈ {yes, no}
Holding	∈ {sword, balloon, flag}
Jacket-colour	∈ {red, yellow, green, blue}
Has-tie	∈ {yes, no}

Ukupan broj mogućih primera je $3 \cdot 3 \cdot 2 \cdot 3 \cdot 4 \cdot 2 = 432$. Problem je konstruisan radi poređenja različitih algoritama učenja, a upotrebljena verzija *Monks-2* opisuje koncept "tačno dva od šest atributa imaju *prvu* od mogućih vrednosti", koji se ne može kompaktno opisati propozicionim logičkim izrazima (CNF ili DNF). Tačan propozicioni opis koncepta se sastoji od disjunkcije svih konjuktivnih izraza u kojima se *dva* atributa pojavljuju sa svojim prvim vrednostima, a *svi ostali* atributi bez prve vrednosti iz domena.

Ovo je težak problem zbog toga što su svi atributi *relevantni* i postoje međusobne interakcije atributa. Pošto se za učenje ne koriste svi obučavajući primeri, algoritmi učenja propozicionih opisa (stabala i pravila) daju približna rešenja.

2. Problem *Waveform-40* [48] je konstruisan radi provere funkcionisanja algoritama učenja u prisustvu šuma, odnosno *irrelevantnih* atributa.

Skup podataka ima 21 relevantan atribut, dobijen kombinovanjem dva od tri osnovna signala (funkcije) sa dodatkom nezavisnog šuma i dodatnih 19 irrelevantnih atributa, čije vrednosti su generisane potpuno slučajno.

3. Problem *Breiman-1K* [15] je veštački konstruisan kao primer problema sa velikim brojem "slabih" atributa, kod koga se ne može selektovati mala grupa atributa koja bi dobro razdvajala klase. Ovakvi problemi su teški za učenje uobičajenim metodima, kao što su neuronske mreže i stabla klasifikacije.

Obučavajući skup ima 1.000 primera razvrstanih u 10 klasa i opisanih sa 1.000 nezavisnih binarnih atributa. Način formiranja detaljnije je opisan u *Prilogu 4*.

Svojstva konstruisanih problema data su u Tabeli 2.

Problem	Broj primera	Broj atributa		Broj klasa	% većinske klase	Ispuštene vrednosti
		Diskret.	Kontinual.			
Monks-2	432	6	-	2	67,13%	ne
Waveform-40	5.000	-	40	3	33,33%	ne
Breiman-1K	1.000	-	1.000	10	11,00%	ne

Tabela 2: Osnovne karakteristike konstruisanih problema učenja

5.2.2 UCI problemi, evidentno oskudni

Za ilustraciju je izabrano nekoliko problema manjeg obima iz baze podataka University of California at Irvine [92], koji nemaju veliki broj atributa u modelu, ali naučeno znanje ima malu tačnost predviđanja [40], [88].

Problem	Broj primera	Broj atributa		Broj klasa	% većinske klase	Ispuštene vrednosti
		Diskret.	Kontinual.			
Breast Cancer LJ	286	9	-	2	70,28%	da
Cleveland Heart	303	7	6	5	54,46%	da
Lung Cancer	32	56	-	3	40,63%	da

Tabela 3: Osnovne karakteristike standardnih problema učenja iz baze UCI

Sva tri problema su iz medicinske domene i odnose se na:

- prognostiku raka dojke (*Breast Cancer, Ljubljana*)
Problem je učenje pravila za klasifikovanje slučajeva za koje postoji mogućnost ponovnog pojavljivanja tumora dojke nakon operacije, na osnovu opštih podataka o pacijentu i kliničkih nalaza, (podaci Instituta za onkologiju Univerzitetskog medicinskog centra u Ljubljani, [92]).
- dijagnostiku srčanih oboljenja (*Cleveland Hart Disease*)
Problem je učenje pravila za dijagnostiku oboljenja srca na osnovu opštih podataka o pacijentu i kliničkih nalaza. Slučajevi se klasifikuju kao zdravi ili u jednu od 4 kategorije oboljenja srca.
- dijagnostiku raka pluća (*Lung Cancer*)
Klasifikacija slučajeva raka pluća u jednu od tri kategorije na osnovu medicinskih podataka koje koriste u praksi i lekari specijalisti.

5.2.3 Problemi klasifikacije slučajeva na osnovu genskih ekspresija

Razmotrena su tri stvarna problema iz oblasti analize genskih ekspresija [29], [9], [93] u dijagnostici i prognostici raka dojke:

- *Breast Cancer* (BC), problem učenja pravila dijagnostike prisustva raka dojke na osnovu genskih ekspresija
- *Relapse Free Survival 5* (RFS5), problem učenja pravila za određivanja rizika ponovnog oboljevanja od raka dojke u periodu od 5 godina
- *Netherland Cancer Institute Breast Cancer* (NKI BC), takođe problem učenja pravila dijagnostike prisustva raka dojke na osnovu genskih ekspresija u verziji instituta NKI.

Osnovna svojstva tri razmatrana probleme binarne klasifikacije (*Breast Cancer* (BC), *Relapse Free Survival 5* (RFS5) i *NKI BC*) dati su u Tabeli 4.

Problem	Broj primera	Broj atributa		Broj klasa	% većinske klase	Ispuštene vrednosti
		Diskret.	Kontinual.			
BC	175	-	22.215	2	66,29%	ne
RFS5	441	-	22.215	2	54,20%	ne
NKI BC	182	-	22.223	2	62,64%	ne

Tabela 4: Osnovne karakteristike problema učenja na osnovu genskih ekspresija

Iz Tabele 4. se vidi da je broj primera u navedenim problemima koji pripadaju većinskoj klasi 54,20-66,29%, što se može posmatrati kao tačnost predviđanja većinskog klasifikatora, koji uvek klasifikuje primere u većinsku klasu.

Zadatak programa za induktivno učenje je da nauče pravila koja imaju veću tačnost predviđanja od većinskog klasifikatora.

5.3 REZULTATI UČENJA I ESTIMACIJE PERFORMANSI NAUČENOG ZNANJA

Za induktivno učenje pravila klasifikacije i estimaciju tačnosti naučenog znanja korišćene su metode induktivnog učenja razumljivih koncepata (klasične ili simboličke metode), metode kombinovanih klasifikatora i metode prethodne selekcije atributa.

Radi poređenja same estimacije tačnosti sa rezultatima drugih metoda dati su i neki rezultati učenja metodom *WEKA.NaiveBayes* (po istoj metodologiji estimacije) i originalnim programom *RandomForest* autora *L. Breiman* i *A. Cutler* [15].

5.3.1 Konstruisani merni problemi

1.Problem Monks-2

Problem ima poznata svojstva (tačka 5.2 ovog rada) i koristi se za proveru primenjene metodologije estimacije optimalnog broja atributa za metode selekcije atributa koji će se koristiti za ostale probleme.

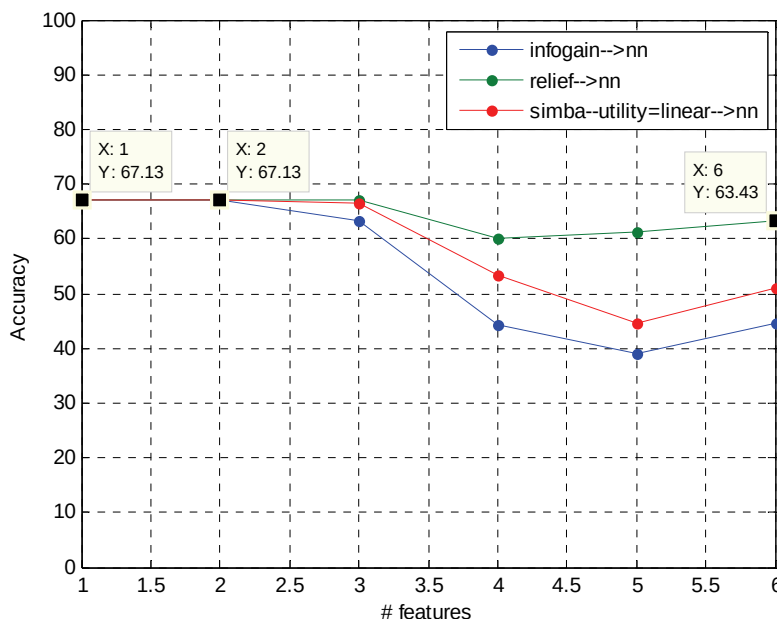
Metode estimacije treba da pokažu da kompletan skup atributa (svih $n=6$) najbolje razlikuje primere, odnosno da su svi atributi relevantni i da se tačnost predviđanja naučenog znanja smanjuje za sve varijante redukcije njihovog broja na n atributa ($1 \leq n \leq 5$).

Upotrebljeni alat je zasnovan na rezultatima učenja 1-NN klasifikatora, a estimaciju vrši metodom unakrsne validacije.

Dijagram na Slika 13 ocenjuje da podskup od jednog atributa ($n=1$) u proseku najbolje klasifikuje obučavajuće primere, dok je tačnost predviđanja za vrednosti $2 \leq n \leq 6$ manja i poraste tek za konačnu vrednost $n=6$.

Ugrađena zakonitost "tačno dva od šest atributa imaju prvu od mogućih vrednosti" se u propozicionoj formi može izraziti kao disjunkcija svih konjuktivnih izraza u kojima dva od šest atributa imaju prvu vrednost iz domena, a preostala četiri imaju jednu ili

više vrednosti, od kojih *nijedna* nije prva vrednost iz domena i teško se može naučiti na osnovu samo dela obučavajućih primera koji se koristi u krosvalidaciji.



Slika 13: Uticaj broja selektovanih atributa na tačnost predviđanja (Monks-2)

U Tabeli 5. je prikazana tačnost previđanja sistema *Empiric.Rules* u učenju pojedinačnih skupova pravila i uticaj selekcije atributa metodom slučajnih podprostora za tri ocene kvaliteta pravila.

Za osnovnu meru kvaliteta je prikazan uticaj učenja ansambala i istovremene primene selekcije atributa i učenja ansambala (10 i 50 elemenata).

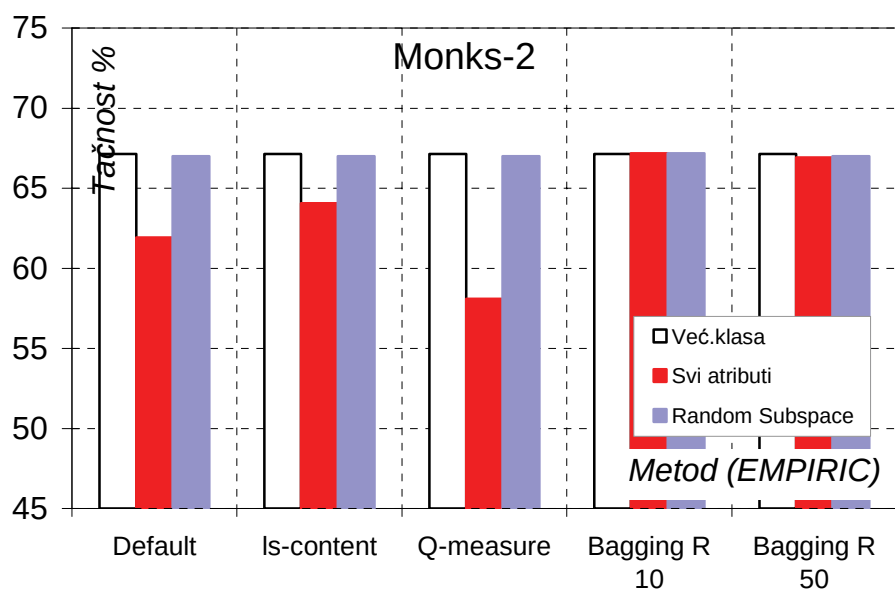
Dataset: Monks-2 (432 primera, 6 atributa, 2 klase, većinska 67,13%)

Algoritam učenja		Svi atributi	Sluč. podp.
		6	2
EMPIRIC.Rules			
<i>Default</i>		61,92%±4,10%	67,00%±0,41%
<i>ls-content</i>		64,06%±6,51%	67,00%±0,41%
<i>Q-measure</i>		58,09%±4,11%	67,00%±0,41%
<i>Bagging.Default</i>	10	67,17%±1,39%	67,18%±0,57%
	50	66,92%±1,00%	67,00%±0,41%

Tabela 5: Tačnost predviđanja za problem Monks-2 (Empiric)

Grafički prikaz rezultata je na Sl. 14. Vidi se da za ovaj složeni primer neadekvatnog modela problema metod slučajne redukcije atributa daje povećanje tačnosti predviđanja kod učenja pojedinačnih skupova pravila.

Tačnost ansambala je ista sa i bez selekcije atributa, jer se u učenje uključuje veći broj primera, koji je neophodan da se ispravno nauči kompletan, neadekvatno opisan koncept.



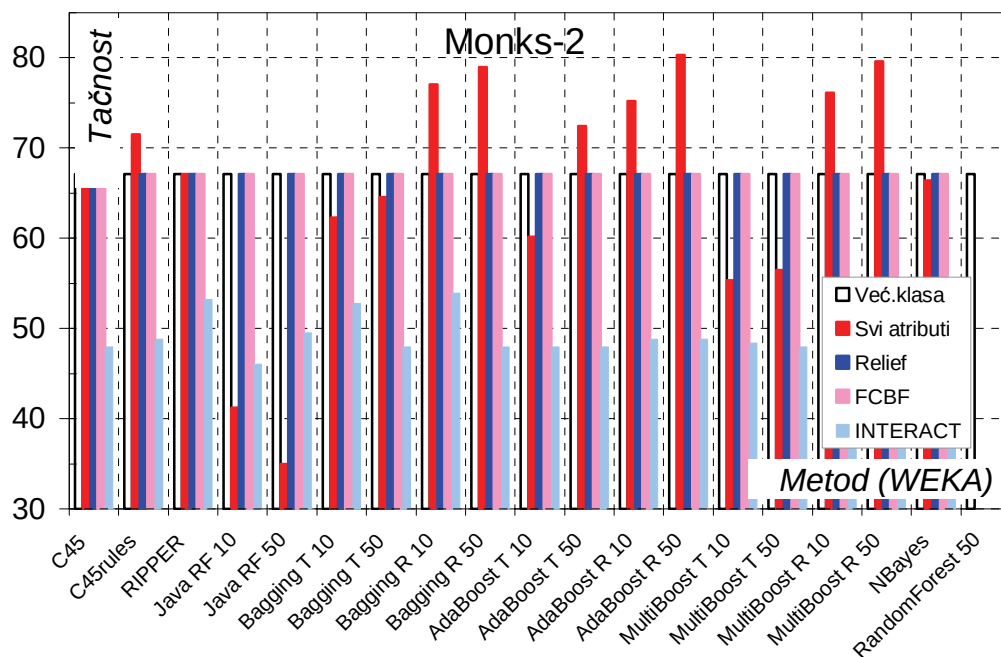
Slika 14: Prikaz tačnosti predviđanja ispitivanih metoda učenja i selekcije atributa iz sistema EMPIRIC za problem Monks-2

Prikaz estimacije tačnosti predviđanja iz sistema WEKA je u Tabeli 6 i na Sl. 15.

Dataset: Monks-2 (432 primera, 6 atributa, 2 klase, većinska 67,13%)

Algoritam učenja		Svi atributi	FCBF	RELIEF	INTERACT
		6	2	2	2
WEKA					
C4.5		67,13%	67,13%	67,13%	47,92%
C4.5Rules		71,53%	67,13%	67,13%	48,84%
RIPPER		67,13%	67,13%	67,13%	53,24%
RandomForest	10	41,20%	67,13%	67,13%	46,06%
	50	34,95%	67,13%	67,13%	49,54%
Bagging-C4.5	10	62,27%	67,13%	67,13%	52,78%
	50	64,58%	67,13%	67,13%	47,92%
Bagging-C4.5rules	10	77,08%	67,13%	67,13%	53,94%
	50	78,94%	67,13%	67,13%	47,92%
AdaBoostM1-C4.5	10	60,19%	67,13%	67,13%	47,92%
	50	72,45%	67,13%	67,13%	47,92%
AdaBoostM1-C4.5rules	10	75,23%	67,13%	67,13%	48,84%
	50	80,32%	67,13%	67,13%	48,84%
MultiBoostAB-C4 5	10	55,32%	67,13%	67,13%	48,38%
	50	56,48%	67,13%	67,13%	47,92%
MultiBoostAB-C4.5rules	10	76,16%	67,13%	67,13%	47,45%
	50	79,63%	67,13%	67,13%	48,84%
NaiveBayes		66,44%	67,13%	67,13%	49,54%
RandomForest (Breiman-Cutler)					
RandomForest	50				

Tabela 6: Tačnost predviđanja za problem Monks-2 (WEKA)



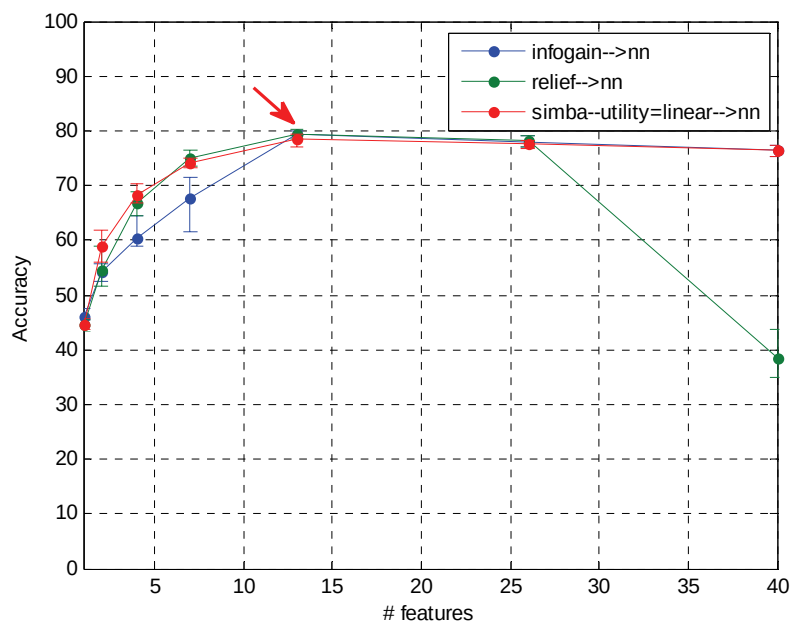
Slika 15: Prikaz tačnosti predviđanja ispitivanih metoda učenja i selekcije atributa iz sistema WEKA za problem Monks-2

Vidi se da je tačnost predviđanja u gotovo svim eksperimentima za metode sistema WEKA približno ista za sve podskupove od 2 atributa i ne prevazilazi tačnost apriorne većinske klasifikacije.

Tačnost kombinovanih modela pokazuje poboljšanja za ceo skup atributa i veći broj elemenata ansambla, jer se u proces učenja uključuje sve veći broj primera, koji su neophodni da se ispravno nauči kompletan, neadekvatno opisan koncept.

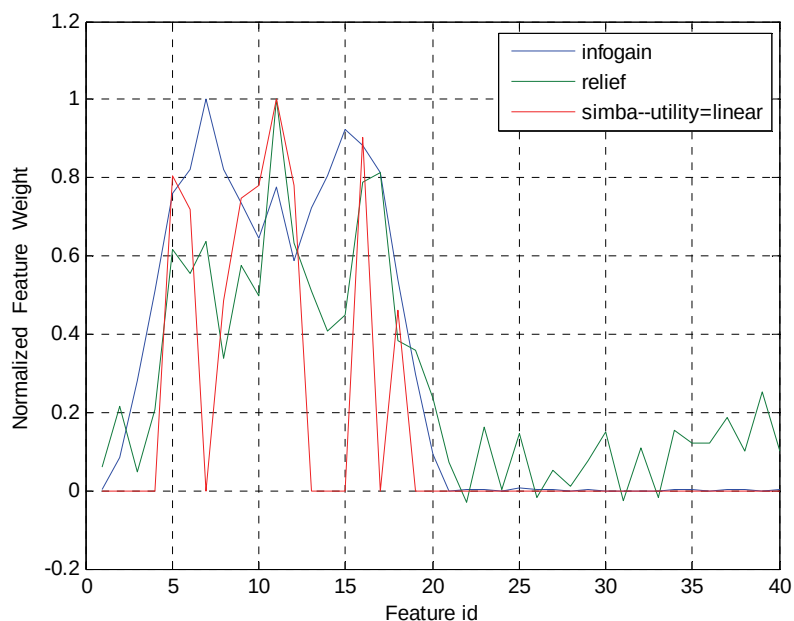
2. Problem Waveform 40

Problem *Waveform-40* takođe ima unapred poznata svojstva (tačka 5.2 rada) i u ovom radu će poslužiti kao jedan od primera za verifikaciju primenjene metodologije estimacije optimalnog broja atributa za korišćene metode selekcije.



Slika 16: Uticaj broja selektovanih atributa na tačnost predviđanja (Waveform-40)

Iz dijagrama na Sl. 16 se vidi da prema svim metodama estimacije tačnost raste i dostiže maksimum za 13 atributa, a nakon toga opada. U Tabeli 7 su prikazani rezultati estimacije tačnosti predviđanja za ovaj broj selektovanih atributa.



Slika 17: Raspodela ocena relevantnosti atributa metodima IngoGain, Relief i Simba (Waveform-40)

Prema dijagramu na Sl. 17, irelevantnost slučajno generisanih atributa od 22 do 40 je ispravno ocenjena (estimacije relevantnosti *InfoGain*=0, *SIMBA*=0, *Relief*≈0).

U Tabeli 7. je prikazana tačnost previđanja sistema *Empiric.Rules* u učenju pojedinačnih skupova pravila i uticaj selekcije atributa metodom slučajnih podprostora za tri ocene kvaliteta pravila.

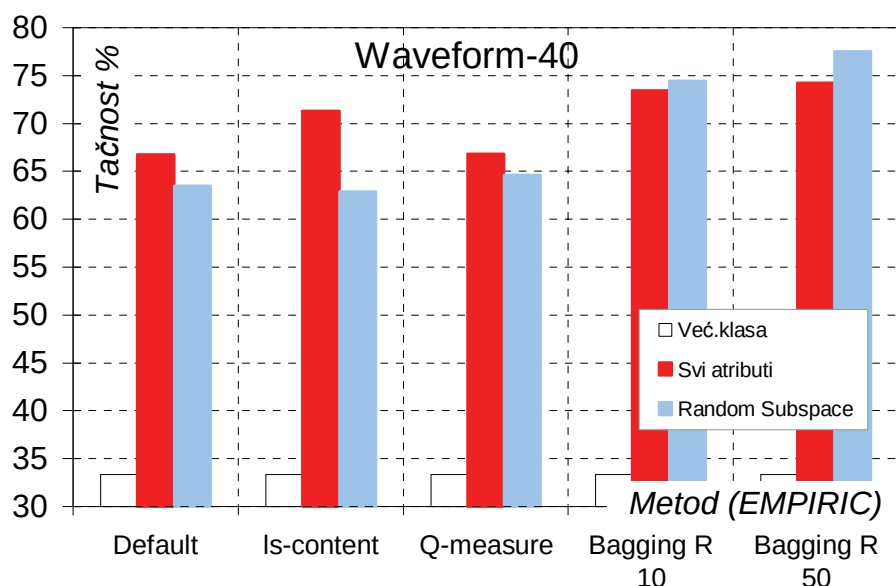
Za osnovnu meru kvaliteta je prikazan uticaj učenja ansambala i istovremene primene selekcije atributa i učenja ansambala (10 i 50 elemenata).

Dataset: Waveform-40 (5.000 primera, 40 atributa, 3 klase)

Algoritam učenja		Svi atributi	Sluč. podr.
		40	13
EMPIRIC.Rules			
Default		66,77%±2,28%	63,49%±2,45%
Is-content		71,34%±2,14%	62,90%±2,59%
Q-measure		66,82%±2,75%	64,63%±1,64%
Bagging.Default	10	73,47%±1,47%	74,42%±3,02%
	50	74,27%±1,45%	77,54%±2,02%

Tabela 7: Tačnost predviđanja za problem Waveform-40 (Empiric)

Grafički prikaz rezultata je na Sl. 18.



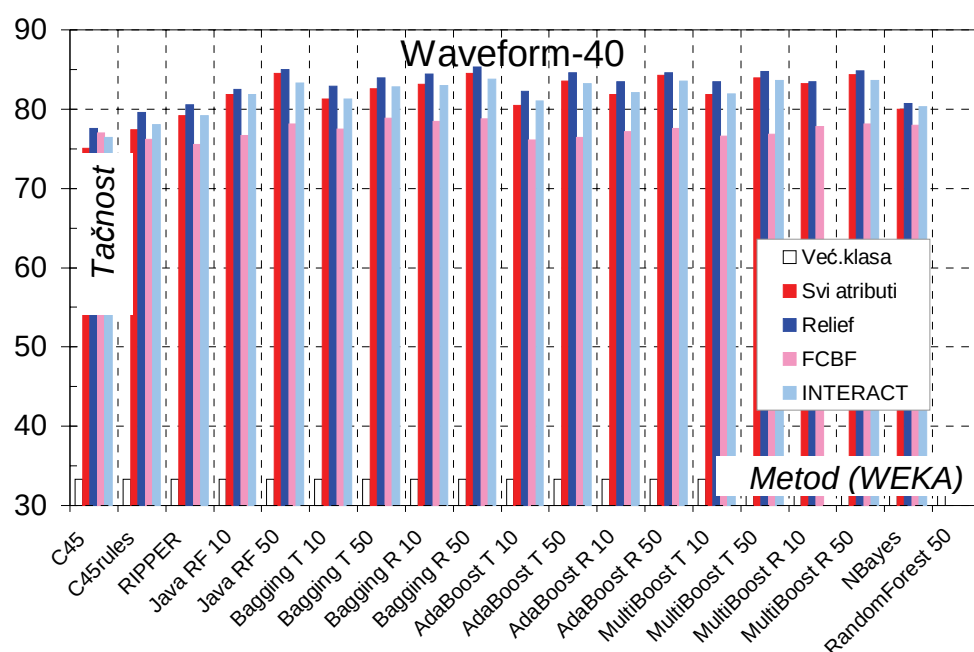
Slika 18: Prikaz tačnosti predviđanja ispitivanih metoda učenja i selekcije atributa iz sistema EMPIRIC za problem Waveform-40

Prikaz estimacije tačnosti predviđanja ispitivanih metoda učenja i selekcije atributa iz sistema WEKA je u Tabeli 8 i na Sl. 19.

Dataset: Waveform-40 (5.000 primera, 40 atributa, 3 klase)

Algoritam učenja		Svi atributi	FCBF	RELIEF	INTERACT
		40	13	13	13
WEKA					
C4.5		75,08%	76,96	77,56	76,42
C4.5Rules		77,42%	76,22	79,56	78,02
RIPPER		79,20%	75,58	80,52	79,16
RandomForest	10	81,82%	76,66	82,48	81,88
	50	84,54%	78,14	84,98	83,30
Bagging-C4.5	10	81,30%	77,52	82,86	81,30
	50	82,54%	78,88	83,94	82,78
Bagging-C4.5rules	10	83,14%	78,46	84,40	82,96
	50	84,54%	78,76	85,34	83,80
AdaBoostM1-C4.5	10	80,48%	76,08	82,28	81,02
	50	83,56%	76,46	84,62	83,22
AdaBoostM1-C4.5rules	10	81,84%	77,18	83,48	82,12
	50	84,24%	77,54	84,58	83,58
MultiBoostAB-C4 5	10	81,84%	76,58	83,42	81,90
	50	83,92%	76,86	84,78	83,62
MultiBoostAB-C4.5rules	10	83,24%	77,82	83,42	81,82
	50	84,34%	78,10	84,86	83,64
NaiveBayes		80,00	77,96	80,72	80,28
RandomForest (Breiman-Cutler)					
RandomForest	50				

Tabela 8: Tačnost predviđanja za problem Waveform-40 (WEKA)

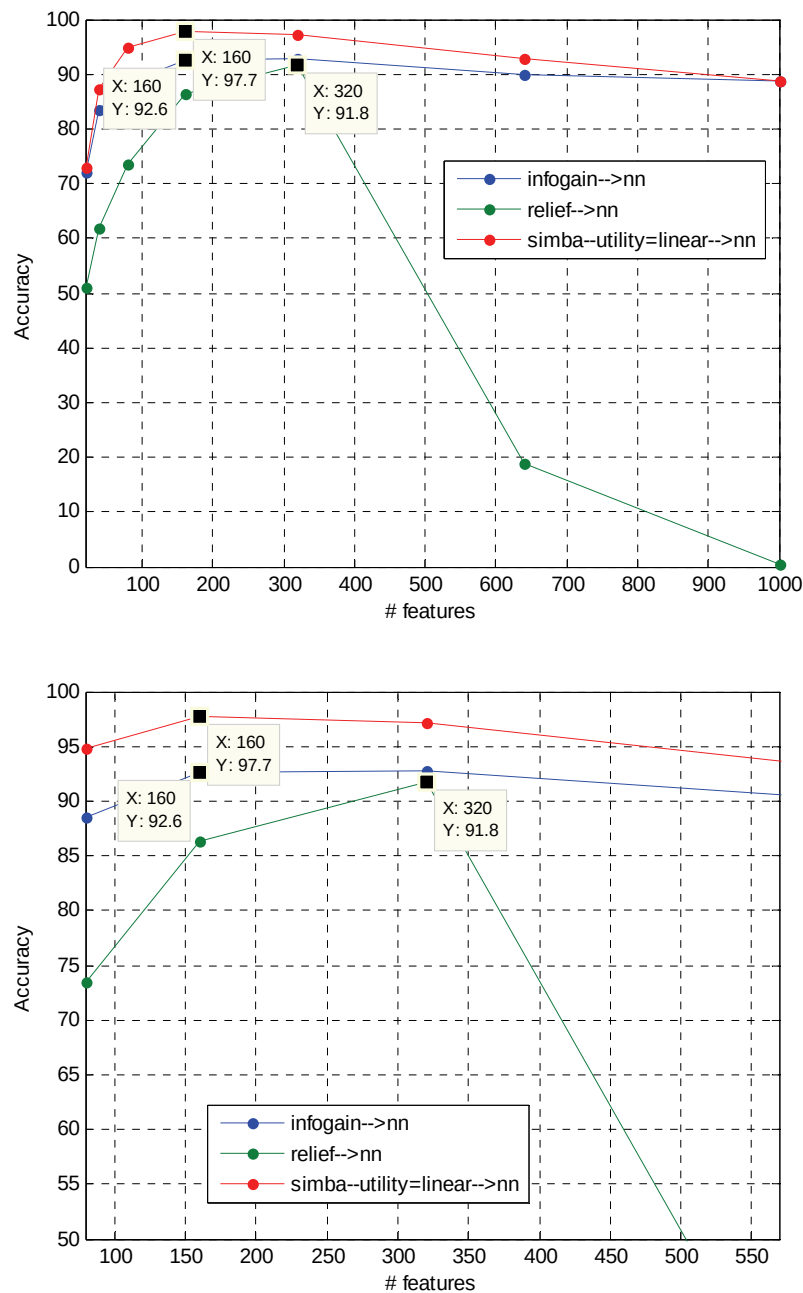


Slika 19: Prikaz tačnosti predviđanja ispitivanih metoda učenja i selekcije atributa iz sistema WEKA za problem Waveform-40

Vidi se da je metodi selekcije atributa *Relief* najuspešniji i da u svim eksperimentima (za osnovne i višestruke modele) prevazilazi tačnost celog skupa atributa. metod *INTERACT* je takođe uspešan u delu eksperimenata, dok metod *FCBF* na ovom primeru nije dao poboljšanja tačnosti predviđanja.

3. Problem Breiman 1K

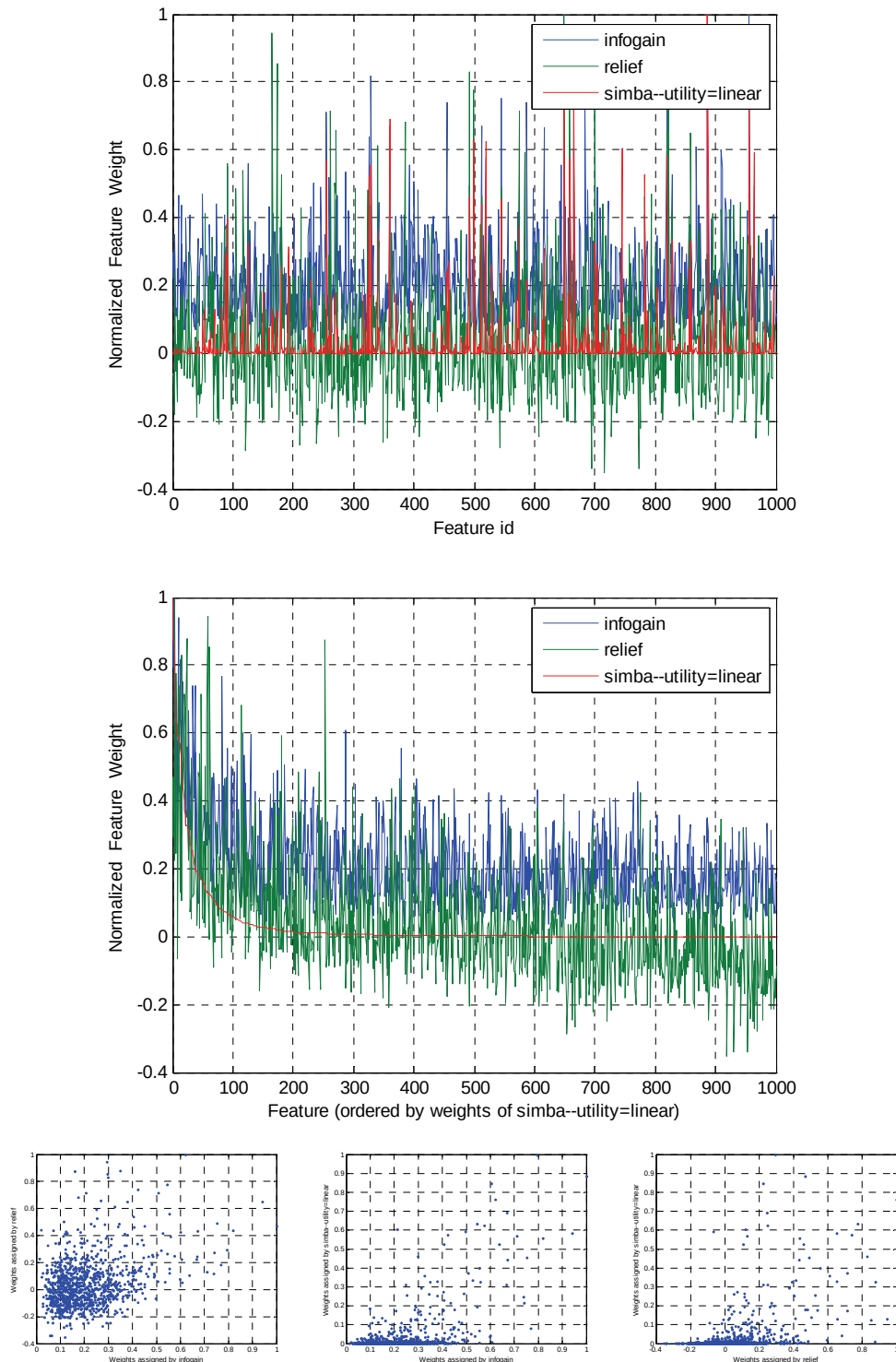
Problem ima unapred poznata svojstva (tačka 5.2 rada) i u ovom radu će poslužiti kao testni primer za verifikaciju primenjene metodologije estimacije optimalnog broja atributa u modelu i efekata primenjenih metoda učenja na obučavajućem skupu koji pokazuje neka svojstva oskudnosti, iako ima relativno veliki broj primera u odnosu na broj atributa.



Slika 20: Uticaj broja selektovanih atributa na tačnost predviđanja (Breiman-1K)

Na dijagramu na Sl. 20 se vidi da po svim metodama estimacije tačnost predviđanja raste i dostiže maksimum za 320 atributa, a nakon toga značajno pada. To što znači da se može izabrati podskup atributa koji je tačniji od celog skupa.

Na Sl. 21 je prikaz raspodele ocena relevantnosti svih 1.000 atributa metodama *InfoGain*, *Relief* i *Simba*. Vidi se koje attribute svaki od metoda ocenjuje kao značajne, kao i neke međusobne korelacije ovih ocena.



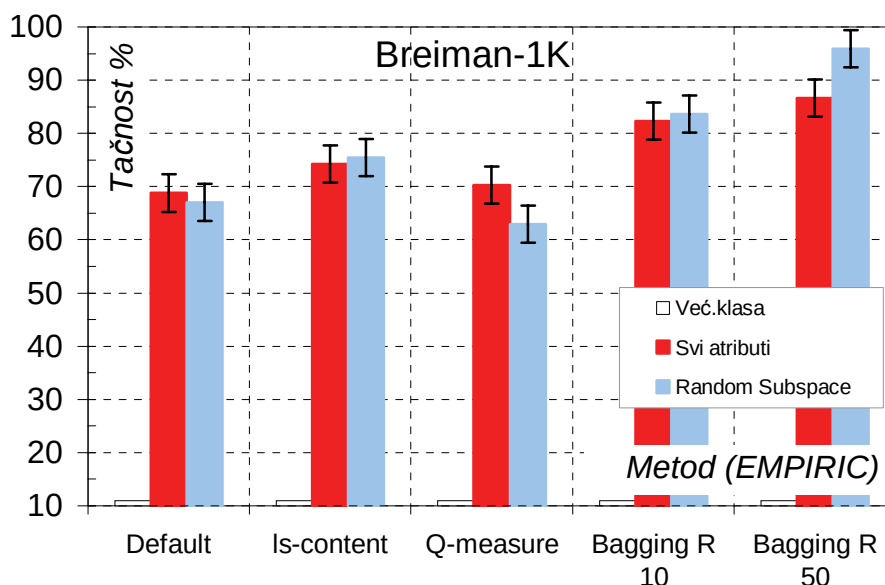
Slika 21: Raspodela ocena relevantnosti atributa metodima IngoGain, Relief i Simba i prikaz međusobne korelacije (Breiman-1K)

Rezultati učenja u sistemu *Empiric.Rules* prikazani su u Tabeli 9 i na Sl. 22. Tačnost predviđanja je približno ista za podskup od 320 atributa i ceo skup, osim metode učenja višestrukih modela sa većim brojem osnovnih klasifikatora.

Dataset: Breiman-1K (1.000 primera, 1.000 atributa, 10 klasa)

Algoritam učenja		Svi atributi	Sluč. podpr.
		1.000	320
EMPIRIC.Rules			
<i>Default</i>		68,78%±3,81%	67,01%±3,95%
<i>Is-content</i>		74,24%±3,10%	75,42%±3,55%
<i>Q-measure</i>		70,32%±4,59%	62,96%±2,54%
<i>Bagging.Default</i>	10	82,27%±3,54%	83,61%±2,48%
	50	86,62%±5,46%	95,93%±2,20%

Tabela 9: Tačnost predviđanja za problem Breiman-1K (Empiric)



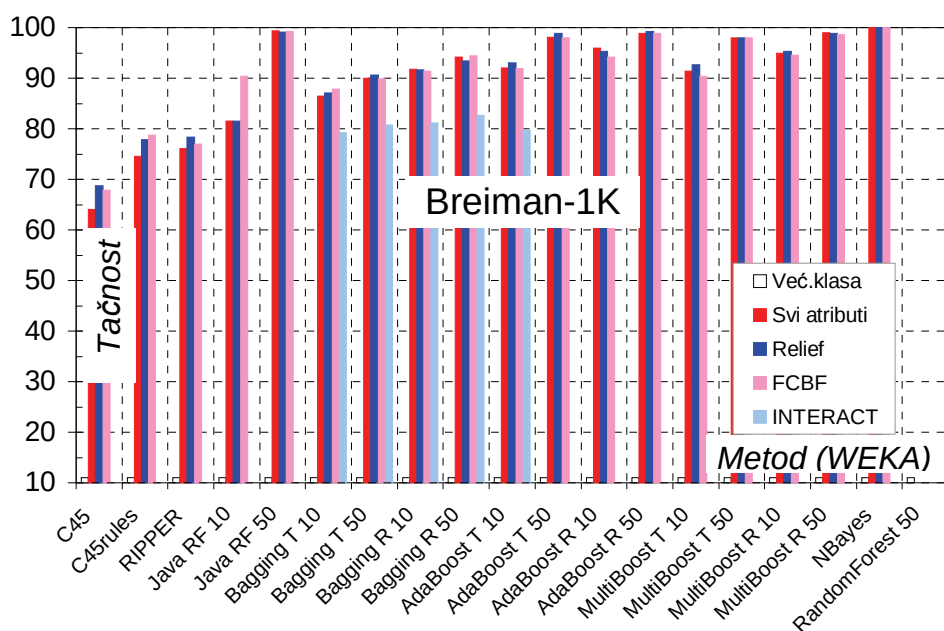
Slika 22: Prikaz tačnosti predviđanja ispitivanih metoda učenja i selekcije atributa iz sistema EMPIRIC za problem Breiman-1K

Prikaz estimacije tačnosti predviđanja ispitivanih metoda učenja i selekcije atributa iz sistema WEKA je u Tabeli 10 i na Sl. 23.

Dataset: Breiman-1K (1.000 primera, 1.000 atributa, 10 klasa)

Algoritam učenja		Svi atributi	FCBF	RELIEF	INTERACT
		1.000	320	320	320
WEKA					
C4.5		64,10	67,80	68,70	72,10
C4.5Rules		74,60	78,70	77,80	76,60
RIPPER		76,10	77,00	78,40	73,00
RandomForest	10	81,50	90,40	81,50	80,50
	50	99,40	99,20	99,10	80,50
Bagging-C4.5	10	86,40	87,90	87,10	79,30
	50	90,00	89,90	90,60	80,70
Bagging-C4.5rules	10	91,80	91,40	91,70	81,20
	50	94,20	94,40	93,40	82,60
AdaBoostM1-C4.5	10	92,00	91,90	93,10	79,80
	50	98,10	98,00	98,80	81,00
AdaBoostM1-C4.5rules	10	96,00	94,20	95,30	81,40
	50	98,90	98,90	99,20	82,50
MultiBoostAB-C4 5	10	91,40	90,40	92,70	79,90
	50	98,00	98,00	98,00	81,50
MultiBoostAB-C4.5rules	10	94,90	94,60	95,30	80,90
	50	99,00	98,60	98,90	82,40
NaiveBayes		100,00	100,00	100,00	86,60
RandomForest (Breiman-Cutler)					
RandomForest	50				

Tabela 10: Tačnost predviđanja za problem Breiman-1K (WEKA)



Slika 23: Prikaz tačnosti predviđanja ispitivanih metoda učenja i selekcije atributa iz sistema WEKA za problem Breiman-1K

Vidi se da je tačnost predviđanja u gotovo svim eksperimentima za metode sistema *WEKA* približno ista za podskup od 320 atributa i ceo skup atributa, osim za osnovne algoritme učenja.

Iako je obučavajući skup formiran s namerom da svi atributi budu relevantni, kada su u pitanju osnovne metode učenja *C4.5*, *C4.5Rules* i *RIPPER* u eksperimentima se ipak javlja ista i veća estimacija tačnosti za podskup selektovan metodima *Relief* i *FCBF*, nego za potpun skup od 1.000 atributa.

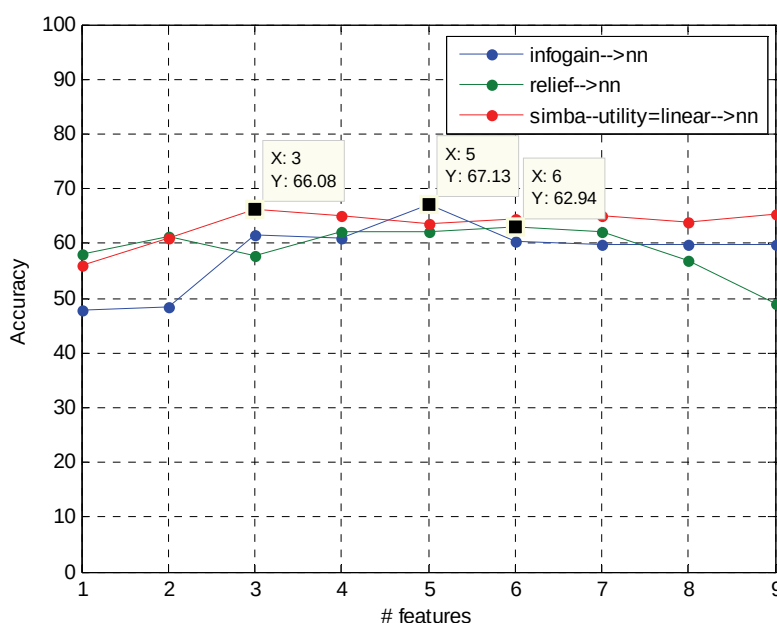
Postojanje podskupova atributa istih diskriminatorskih svojstava koje ima potpun skup potvrđuje i metod *NaiveBayes*, gde je ista estimacija tačnosti predviđanja od 100%.

5.3.2 Merni problemi iz baze podataka UCI

Ovo su standardni problemi oskudnih obučavajućih skupova, koji su formirani na osnovu stvarnih podataka iz istraživanja različitih problema medicinske dijagnostike, gde se obično raspolaze relativno malim konačnim skupom rešenih slučajeva (klasifikovanih primera).

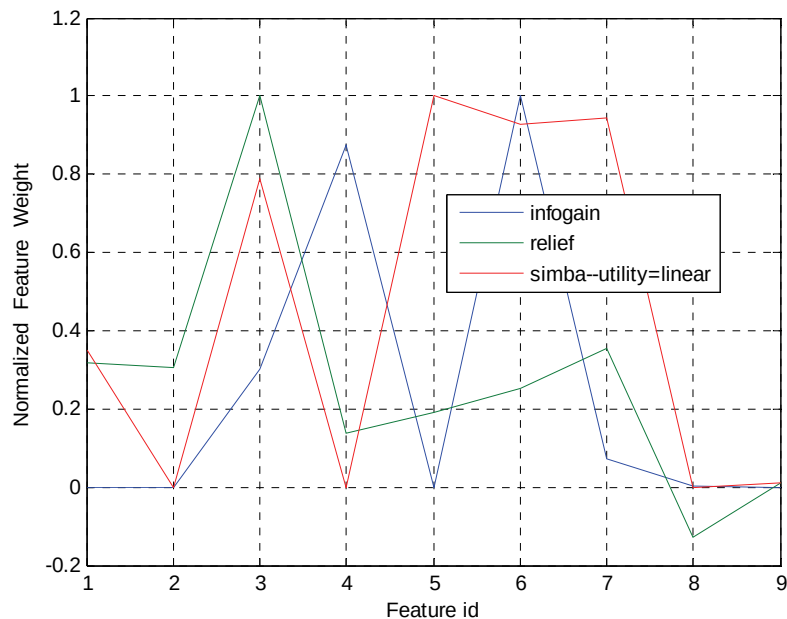
4. Problem Breast Cancer LJ

Prikaz na Sl. 24 daje različite procene optimalnog broja atributa svake od upotrebljenih metoda rangiranja. Estimacija *InfoGain* najveću tačnost predviđa za podskup od $d=5$ atributa, metod *Simba* $d=3$ atributa, a metod *Relief* za podskup od $d=6$ atributa.



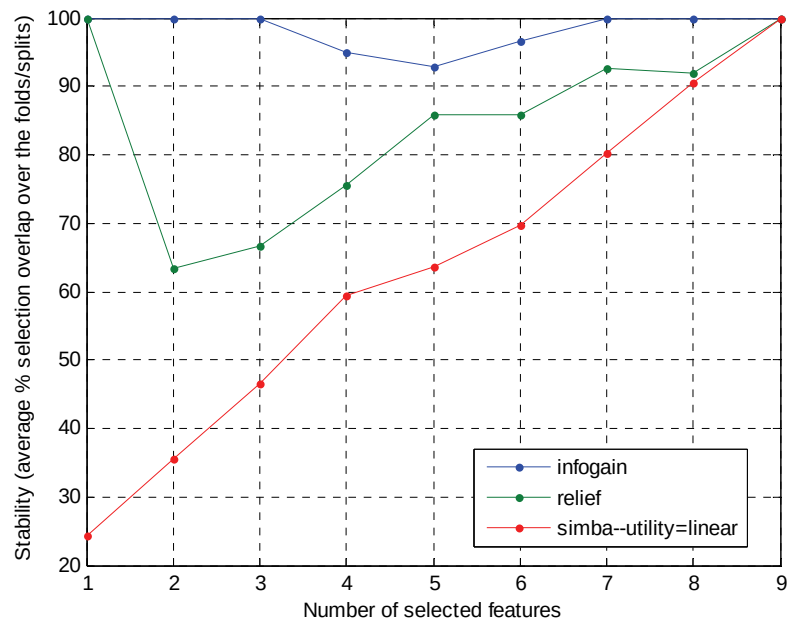
Slika 24: Uticaj broja selektovanih atributa na tačnost predviđanja (Breast Cancer LJ)

Na Sl. 25 je prikaz raspodele ocena relevantnosti svih $d=9$ atributa metodama *InfoGain*, *Relief* i *Simba*. Vidi se koje attribute svaki od metoda ocenjuje kao značajne.



Slika 25: Raspodela ocena relevantnosti atributa metodima *InfoGain*, *Relief* i *Simba* (*Breast Cancer LJ*)

Na Sl. 26 se vidi da najbolja ocena metoda *InfoGain* ($d=5$) nije stabilna, dok je npr. ocena ocena tačnosti za $d=3$ stabilna. Najbolje ocene druga dva metoda takođe nisu stabilne, jer dostižu 100% tek za svih $d=9$ atributa.



Slika 26: Stabilnost estimacije broja atributa (*Breast Cancer LJ*)

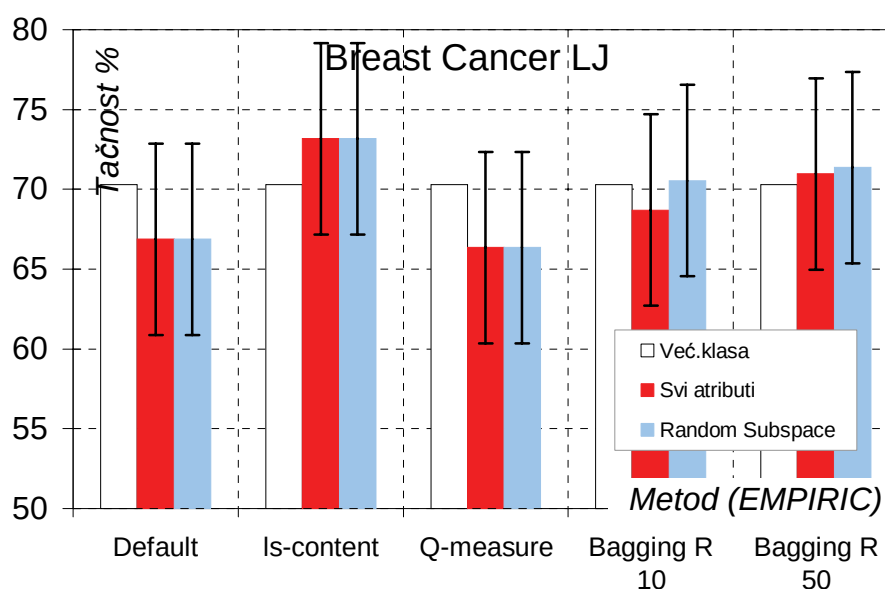
U Tabeli 11. je prikazana tačnost previđanja sistema *Empiric.Rules* u učenju pojedinačnih skupova pravila i ansambala, kao i uticaj selekcije atributa metodom slučajnih podprostora za tri ocene kvaliteta pravila.

Dataset: Breast Cancer LJ (286 primera, 9 atributa, 2 klase)

Algoritam učenja		Svi atributi	Sluč. podpr.
		9	6
EMPIRIC.Rules			
Default		66,87%±8,13	66,77%±8,13
ls-content		73,17%±6,44	73,17%±6,44
Q-measure		66,35%±6,24	66,35%±6,24
Bagging.Default	10	68,70%±5,90	70,54%±6,70
	50	70,97%±5,26	71,37%±9,86

Tabela 11: Tačnost predviđanja za problem Breast Cancer LJ (Empiric)

Rezultati estimacije iz Tabele 11. grafički su prikazani na Sl. 27, za skup metoda koje podržava sistem *EMPIRIC*. Vidi se da je najveća tačnost za meru ls-content, kao i da tačnost monotono raste kada se koristi selekcija atributa prilikom učenja ansambala.



Slika 27: Prikaz tačnosti predviđanja ispitivanih metoda učenja i selekcije atributa iz sistema *EMPIRIC* za problem Breast Cancer LJ

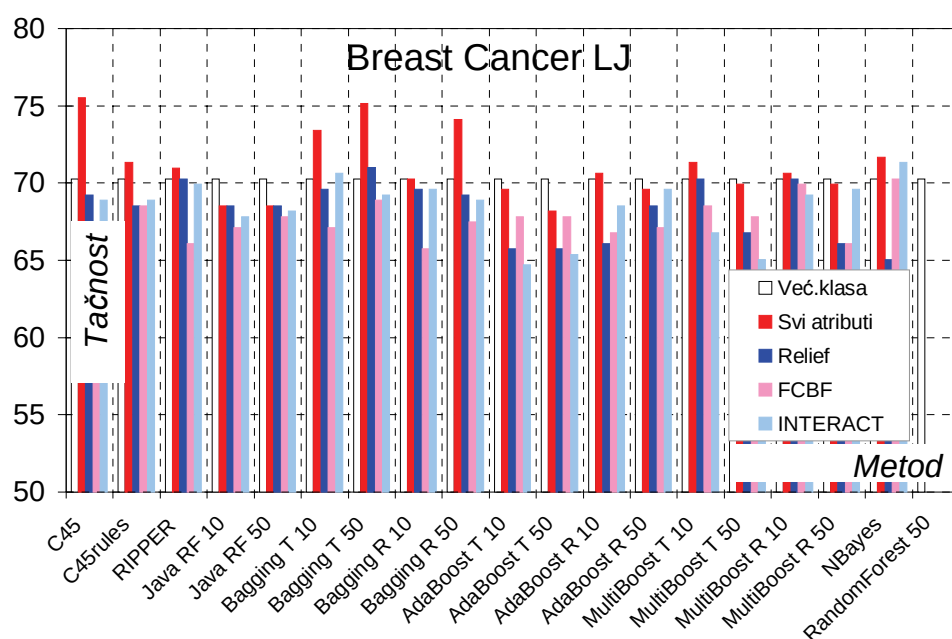
U Tabeli 12. su prikazani rezultati estimacije tačnosti predviđanja metoda iz sistema *WEKA* i *RandomForest* dobijeni primenom dvostruke krosvalidacione šeme [37].

Rezultati estimacije iz Tabele 12. grafički su prikazani na Sl. 28, za skup metoda koje podržavaju sistemi *WEKA* i *RandomForest*.

Dataset: Breast Cancer LJ (286 primera, 9 atributa, 2 klase)

Algoritam učenja		Svi atributi	FCBF	RELIEF	INTERACT
		9	6	6	6
WEKA					
C4.5		75,52	67,13	69,23	68,88
C4.5Rules		71,33	68,53	68,53	68,88
RIPPER		70,98	66,08	70,28	69,93
RandomForest	10	68,53	67,13	68,53	67,83
	50	68,53	67,83	68,53	68,18
Bagging-C4.5	10	73,43	67,13	69,58	70,63
	50	75,17	68,88	70,99	69,23
Bagging-C4.5rules	10	70,28	65,73	69,58	69,58
	50	74,13	67,48	69,23	68,88
AdaBoostM1-C4.5	10	69,58	67,83	65,73	64,69
	50	68,18	67,83	65,73	65,38
AdaBoostM1-C4.5rules	10	70,63	66,78	66,08	68,53
	50	69,58	67,13	68,53	69,58
MultiBoostAB-C4 5	10	71,33	68,53	70,28	66,78
	50	69,93	67,83	66,78	65,04
MultiBoostAB-C4.5rules	10	70,63	69,93	70,28	69,23
	50	69,93	66,08	66,08	69,58
NaiveBayes		71,68	70,28	65,04	71,33
RandomForest (Breiman-Cutler)					
RandomForest	50				

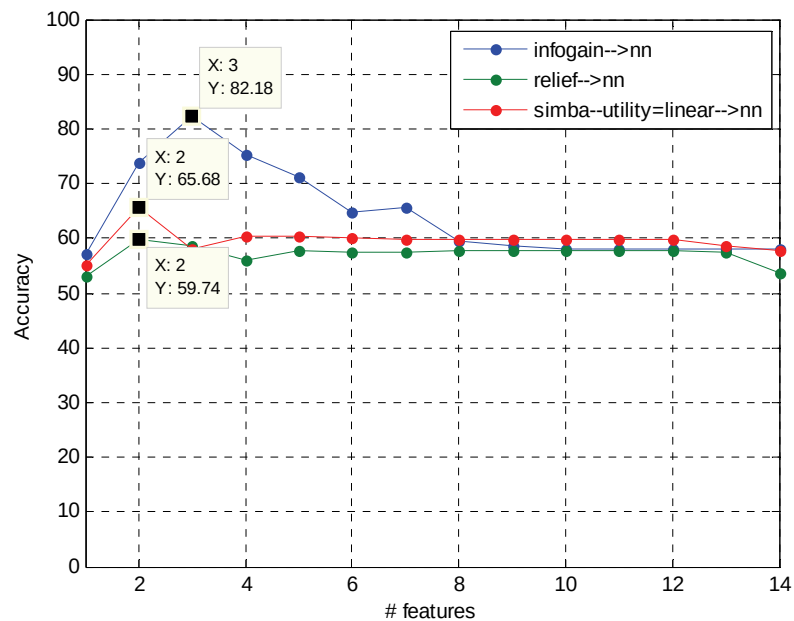
Tabela 12: Tačnost predviđanja za problem Breast Cancer LJ (WEKA)



Slika 28: Prikaz tačnosti predviđanja ispitivanih metoda učenja i selekcije atributa iz sistema WEKA za problem Breast Cancer LJ

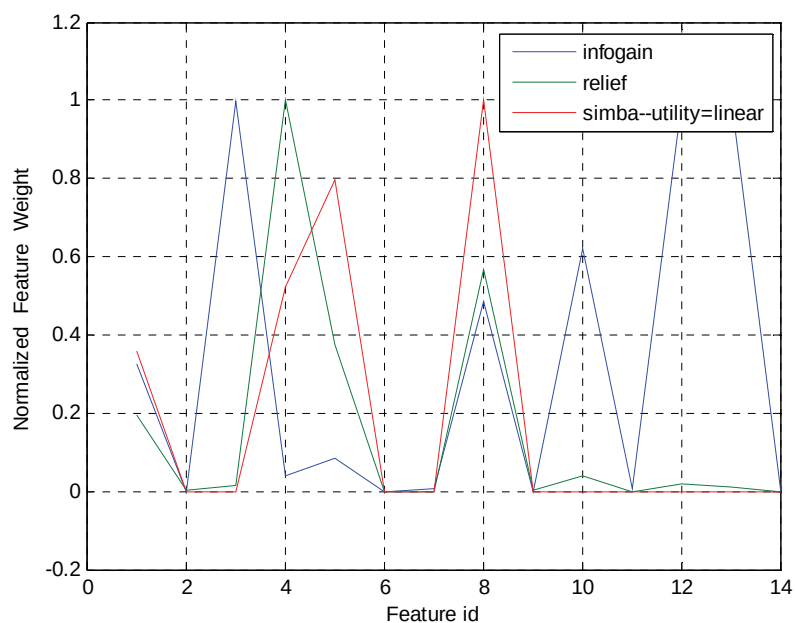
5. Problem Cleveland Heart

Na Sl. 29 su prikazane estimacije optimalnog broja atributa. Ocena *InfoGain* najveću tačnost predviđa za podskup od $d=3$ atributa, a metodi *Simba* i *Relief* za $d=2$ atributa.



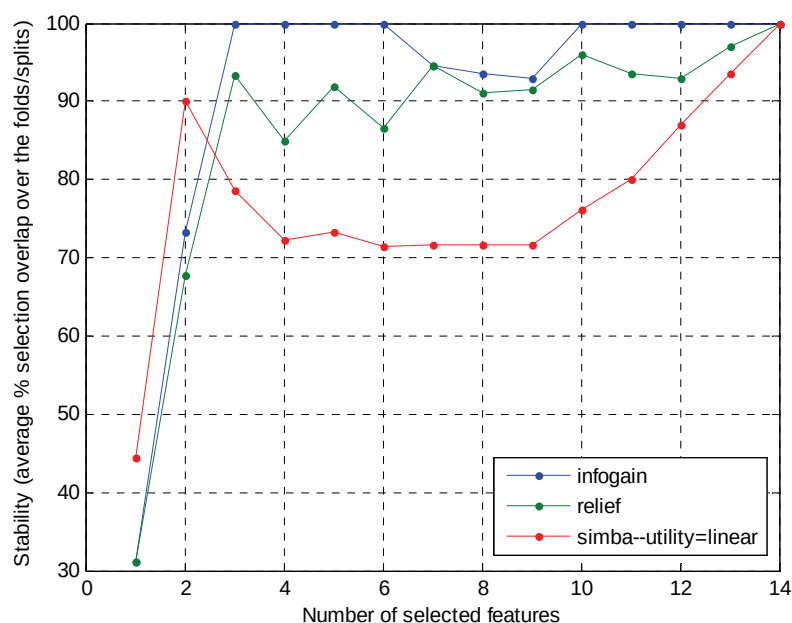
Slika 29: Uticaj broja selektovanih atributa na tačnost predviđanja (Cleveland Heart)

Na Sl. 30 se mogu identifikovati relevantni atributi prema svakoj od tri upotrebljene ocene relevantnosti.



Slika 30: Raspodela ocena relevantnosti atributa metodima IngoGain, Relief i Simba (Cleveland Heart)

Stabilnost estimacije je prikazana na Sl. 31. Za sva tri metoda u početku veoma brzo raste do $d=3$ atributa, gde za *Infogain* dostiže maksimum, a za ostale metode tek za ceo skup od $d=14$ atributa.



Slika 31: Stabilnost estimacije broja atributa (Cleveland Heart)

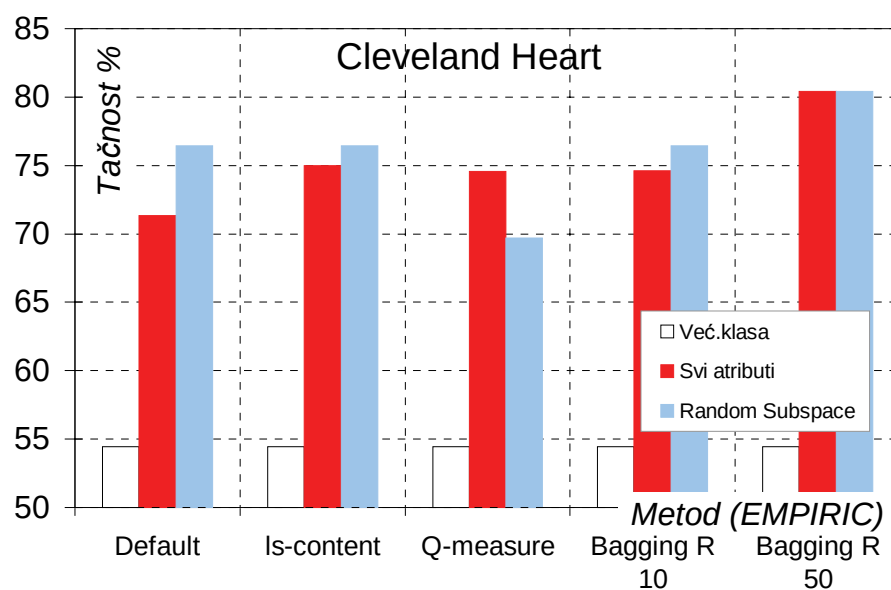
U Tabeli 13. je prikazana tačnost predviđanja sistema *Empiric.Rules* u učenju pojedinačnih skupova pravila za tri ocene kvaliteta pravila, kao i uticaj učenja ansambala selekcije atributa metodom slučajnih podprostora.

Dataset: Cleveland Heart Disease (303 primera, 13 atributa, 5 klasa)

Algoritam učenja	Svi atributi	
	13	RELIEF 10
EMPIRIC.Rules		
<i>Default</i>	71,35%±6,34	76,46%±4,26
<i>ls-content</i>	74,97%±6,17	76,46%±5,30
<i>Q-measure</i>	74,58%±5,72	69,72%±6,44
<i>Bagging.Default</i>	10	74,62%±9,23
	50	80,40%±6,28
		76,43%±6,20
		80,41%±6,16

Tabela 13: Tačnost predviđanja za problem Cleveland Heart (Empiric)

Rezultati estimacije iz Tabele 13. grafički su prikazani na Sl. 32, za skup metoda koje podržava sistem *EMPIRIC*. Vidi se da se za ovaj primer najveća tačnost dobija selekcijom atributa prilikom učenja ansambala.



Slika 32: Prikaz tačnosti predviđanja ispitivanih metoda učenja i selekcije atributa iz sistema EMPIRIC za problem Cleveland Heart

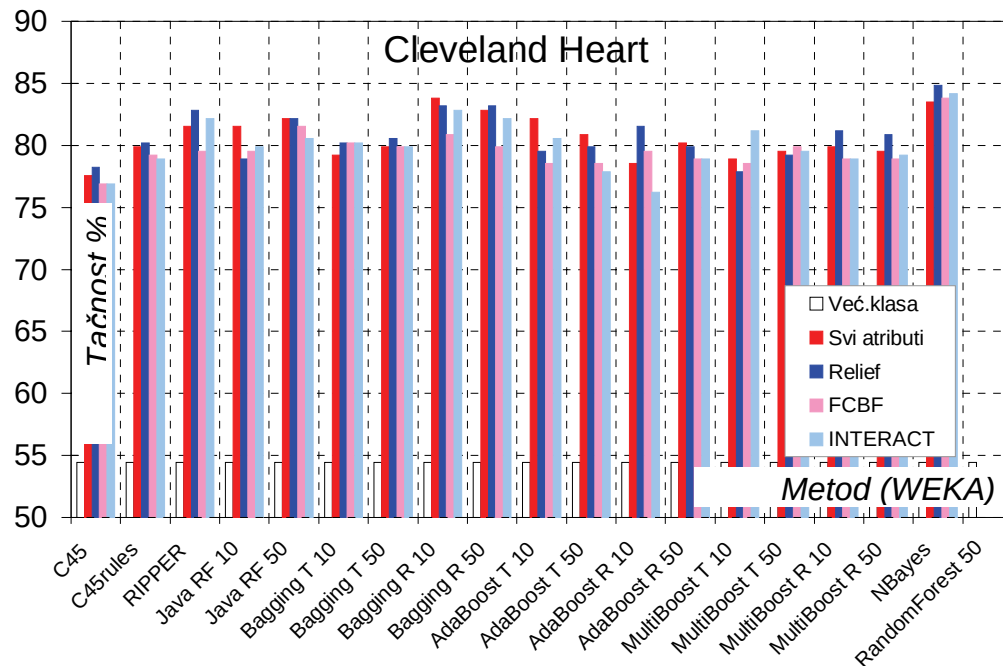
Prikaz estimacije tačnosti predviđanja iz sistema WEKA je u Tabeli 14.

Dataset: Cleveland Heart Disease (303 primera, 13 atributa, 5 klasa)

Algoritam učenja		Svi atributi	FCBF	RELIEF	INTERACT
		13	10	10	10
WEKA					
C4.5		77,56	76,90	78,22	76,90
C4.5Rules		79,88	79,21	80,20	78,88
RIPPER		81,52	79,54	82,84	82,18
RandomForest	10	81,52	79,54	78,88	79,87
	50	82,18	81,52	82,18	80,53
Bagging-C4.5	10	79,21	80,20	80,20	80,20
	50	79,87	79,87	80,53	79,87
Bagging-C4.5rules	10	83,83	80,86	83,17	82,84
	50	82,84	79,87	83,17	82,18
AdaBoostM1-C4.5	10	82,18	78,55	79,54	80,53
	50	80,86	78,55	79,87	77,89
AdaBoostM1-C4.5rules	10	78,55	79,54	81,52	76,24
	50	80,20	78,88	79,87	78,88
MultiBoostAB-C4 5	10	78,88	78,55	77,89	81,19
	50	79,54	79,87	79,21	79,54
MultiBoostAB-C4.5rules	10	79,87	78,88	81,19	78,88
	50	79,54	78,88	80,86	79,21
NaiveBayes		83,50	83,83	84,82	84,16
RandomForest (Breiman-Cutler)					
RandomForest	50				

Tabela 14: Tačnost predviđanja za problem Cleveland Heart (WEKA)

Rezultati estimacije iz Tabele 14. grafički su prikazani na Sl. 33, za skup metoda koje podržavaju sistemi *WEKA* i *RandomForest*.



Slika 33: Prikaz tačnosti predviđanja ispitivanih metoda učenja i selekcije atributa iz sistema *WEKA* za problem *Cleveland Heart*

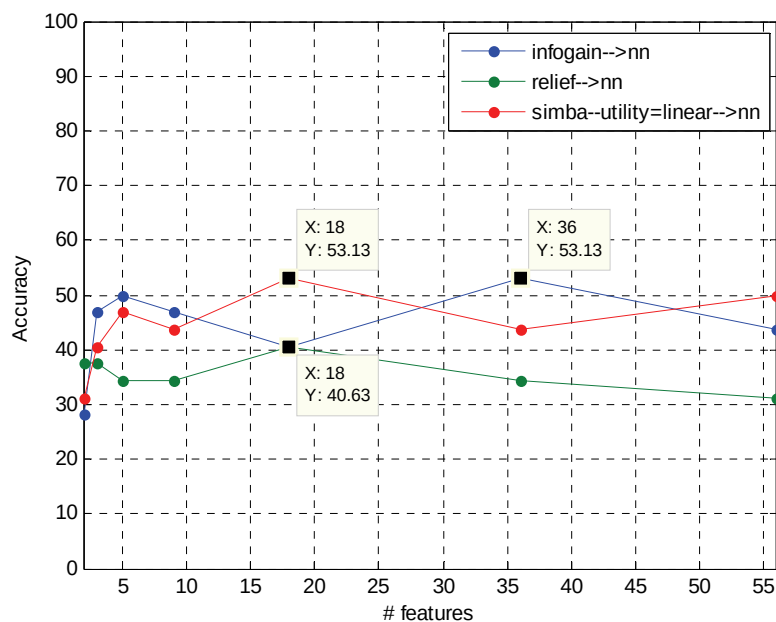
Vidi se da je uspešnost metoda selekcije različita, iako je *Relief* i dalje relativno najuspešniji. Metod ansambala *bagging* je najuspešniji i na celom i na redukovanom skupu atributa.

6. Problem Lung Cancer

Na Sl. 34 su prikazane estimacije optimalnog broja atributa za problem *Lung Cancer*, koji jedini iz ove grupe ima veći broj atributa nego primera: $n/d=32/56=0,57$.

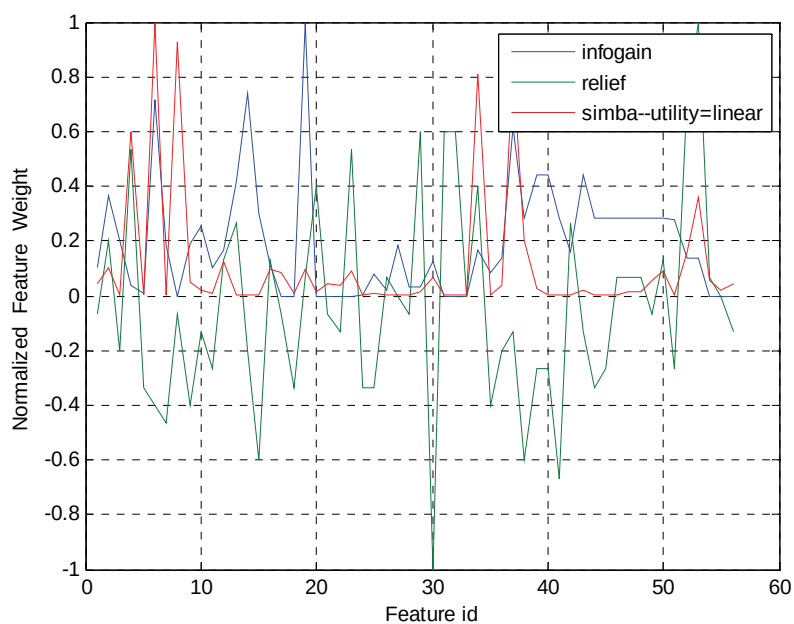
Ocena *InfoGain* najveću tačnost predviđa za podskup od $d=36$ atributa, a metodi *Simba* i *Relief* za podskup od $d=18$ atributa (pri čemu je najveća estimacija tačnosti samo 53,13%).

Takođe su pomoću sve tri metod dobro ocenjeni i manji podskupovi do $d=5$ atributa, ali sa nešto manjom tačnošću predviđanja. Kada bi se tražila izuzetno koncizna rešenja, razmotrili bi se i ovi znatno manji podskupovi.



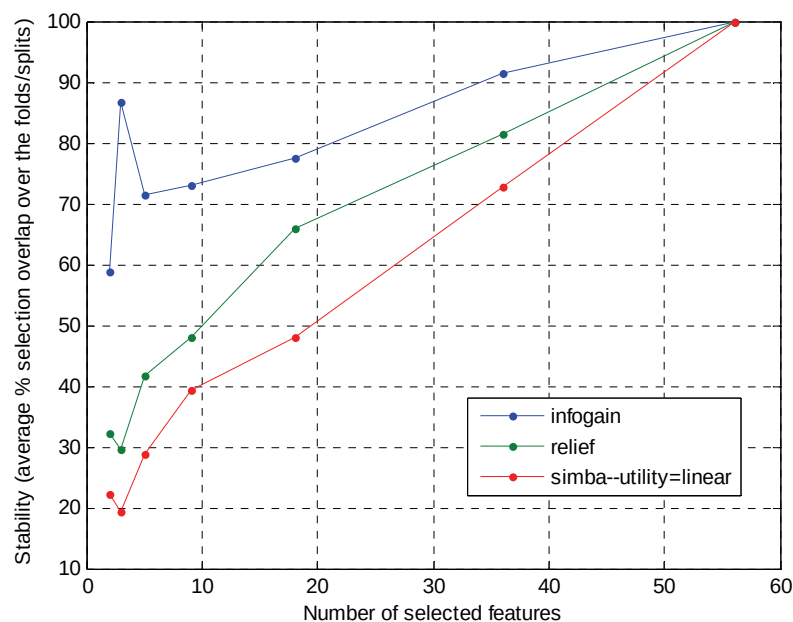
Slika 34: Uticaj broja selektovanih atributa na tačnost predviđanja (Lung Cancer)

Raspodela ocena pokazuje, za veliku većinu atributa, veoma malu saglasnost primenjenih metoda estimacije relevantnosti.



Slika 35: Raspodela ocena relevantnosti atributa metodima IngoGain, Relief i Simba (Lung Cancer)

Stabilnost estimacije je prikazana na Sl. 36. Za metod InfoGain u početku je malo raste (estimacija $d=5$), dok za ostale metode kratko opada. Nakon toga stabilnost monotono raste, ali maksimum dostiže tek za ceo skup atributa.



Slika 36: Stabilnost estimacije broja atributa (Lung Cancer)

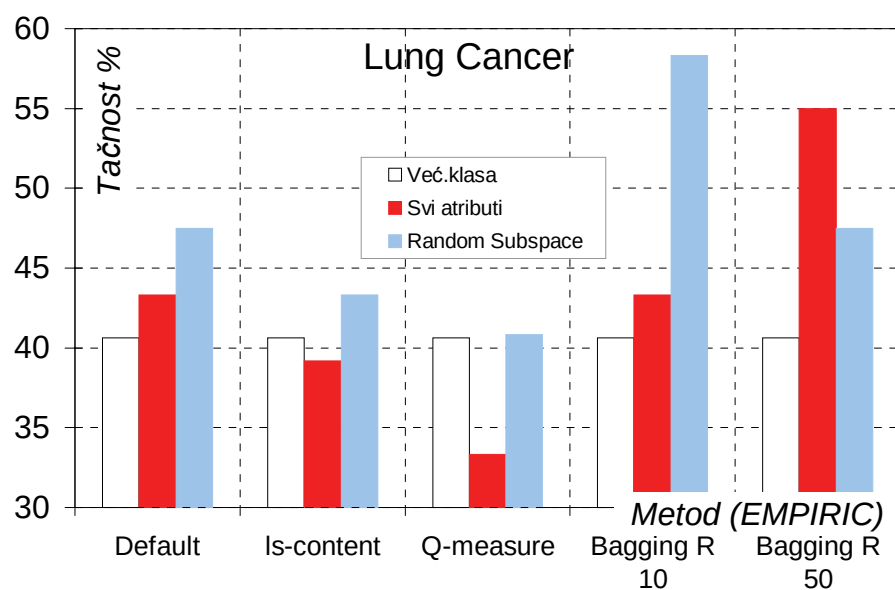
U Tabeli 15. je prikazana tačnost previđanja sistema *Empiric.Rules* u učenju pojedinačnih skupova pravila za tri ocene kvaliteta pravila, kao i uticaj učenja ansambala selekcije atributa metodom slučajnih podprostora.

Dataset: Lung Cancer (32 primera, 56 atributa, 3 klase)

Algoritam učenja		Svi atributi	RELIEF
		56	12
EMPIRIC.Rules			
Default		43,33%±30,00	47,50%±30,87
ls-content		39,17%±23,58	43,33%±24,94
Q-measure		33,33%±25,82	40,83%±25,94
Bagging.Default	10	43,33%±20,00	58,33%±20,07
	50	55,00%±29,86	47,50%±26,37

Tabela 15: Tačnost predviđanja za problem Lung Cancer (Empiric)

Rezultati estimacije iz Tabele 15. grafički su prikazani na Sl. 37 za skup metoda koje podržava sistem *EMPIRIC*. Vidi se da se ubedljivo najveća tačnost dobija selekcijom atributa prilikom učenja manjih ansambala.



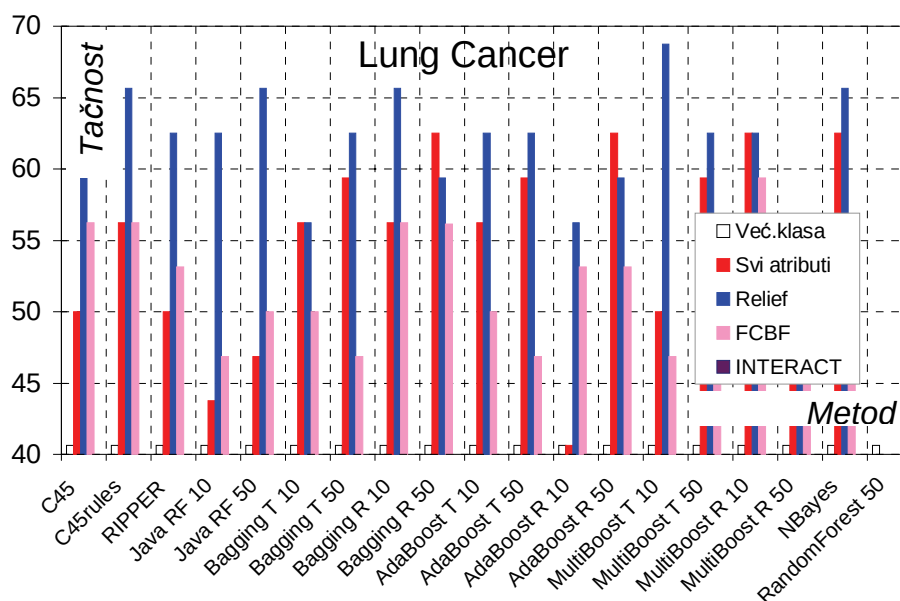
Slika 37: Prikaz tačnosti predviđanja ispitivanih metoda učenja i selekcije atributa iz sistema EMPIRIC za problem Lung Cancer

Dataset: Lung Cancer (32 primera, 56 atributa, 3 klase)

Algoritam učenja		Svi atributi	FCBF	RELIEF	INTERACT
		56	12	12	12
WEKA					
C4.5		50,00	56,25	59,38	-
C4.5Rules		56,25	56,25	65,63	-
RIPPER		50,00	53,13	62,50	-
RandomForest	10	43,75	46,88	62,50	-
	50	46,88	50,00	65,63	-
Bagging-C4.5	10	56,25	50,00	56,25	-
	50	59,38	46,88	62,50	-
Bagging-C4.5rules	10	56,25	56,25	65,63	-
	50	62,50	56,13	59,38	-
AdaBoostM1-C4.5	10	56,25	50,00	62,50	-
	50	59,38	46,88	62,50	-
AdaBoostM1-C4.5rules	10	40,63	53,13	56,25	-
	50	62,50	53,13	59,38	-
MultiBoostAB-C4 5	10	50,00	46,88	68,75	-
	50	59,38	50,00	62,50	-
MultiBoostAB-C4.5rules	10	62,50	59,38	62,50	-
	50	56,25	53,13	56,25	-
NaiveBayes		62,50	50,00	65,63	-
RandomForest (Breiman-Cutler)					
RandomForest	50				

Tabela 16: Tačnost predviđanja za problem Lung Cancer (WEKA)

Rezultati estimacije iz Tabele 16. grafički su prikazani na Sl. 38 za skup metoda koje podržavaju sistemi *WEKA* i sistem *RandomForest*.



Slika 38: Prikaz tačnosti predviđanja ispitivanih metoda učenja i selekcije atributa iz sistema *WEKA* za problem *Lung Cancer*

5.3.3 Problemi klasifikacije slučajeva na osnovu genskih ekspresija

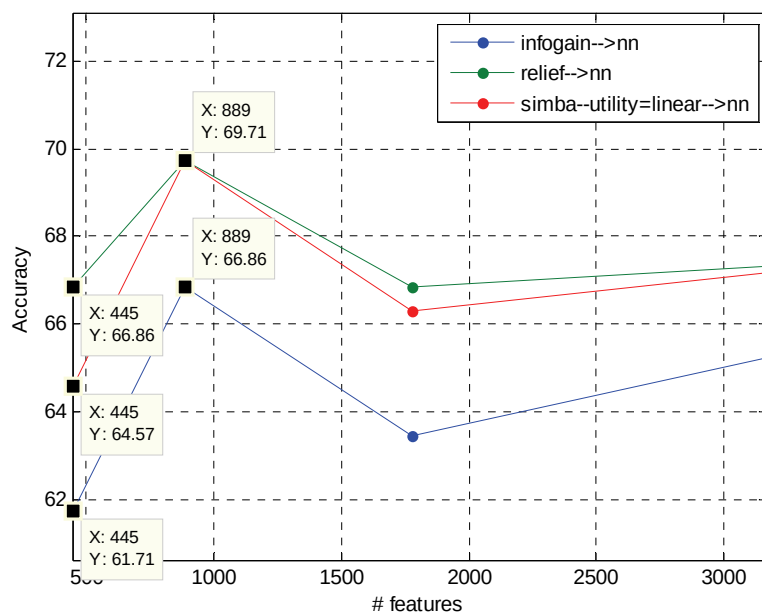
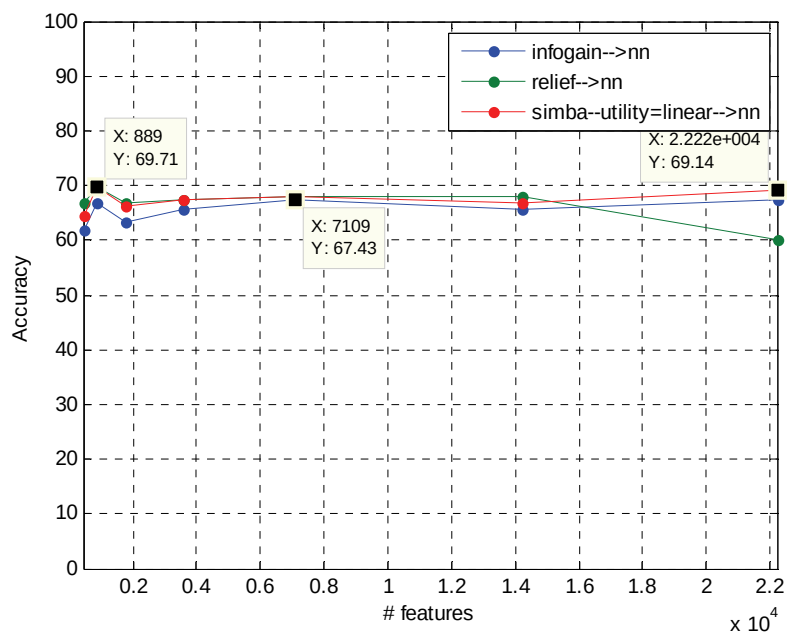
Za tri razmatrana problema genskih ekspresija - *Breast Cancer* (BC), *Relapse Free Survival 5* (RFS5) i *NKI Breast Cancer* (NKI BC) - koji predstavljaju primere binarne klasifikacije, dobijeni su sledeći rezultati.

7. Problem BC

Prikaz uticaja broja selektovanih atributa na tačnost predviđanja pomoću alata *Feature Selection Tool* (Matlab) za primer *Breast Cancer*, je na Sl. 39. Najveća tačnost se previđa za relativno mali broj od 889 atributa.

Prema estimaciji metodom *InfoGain* i *Simba*, za broj atributa veći od 1.800, tačnost predviđanja jednostavnog klasifikatora zasnovanog na poređenju instanci (1-NN) je manja, a povećava se tek za veoma velike vrednosti, preko 14.000 atributa. Prema estimaciji metodom *Relief*, za velike vrednosti d tačnost predviđanja još više opada.

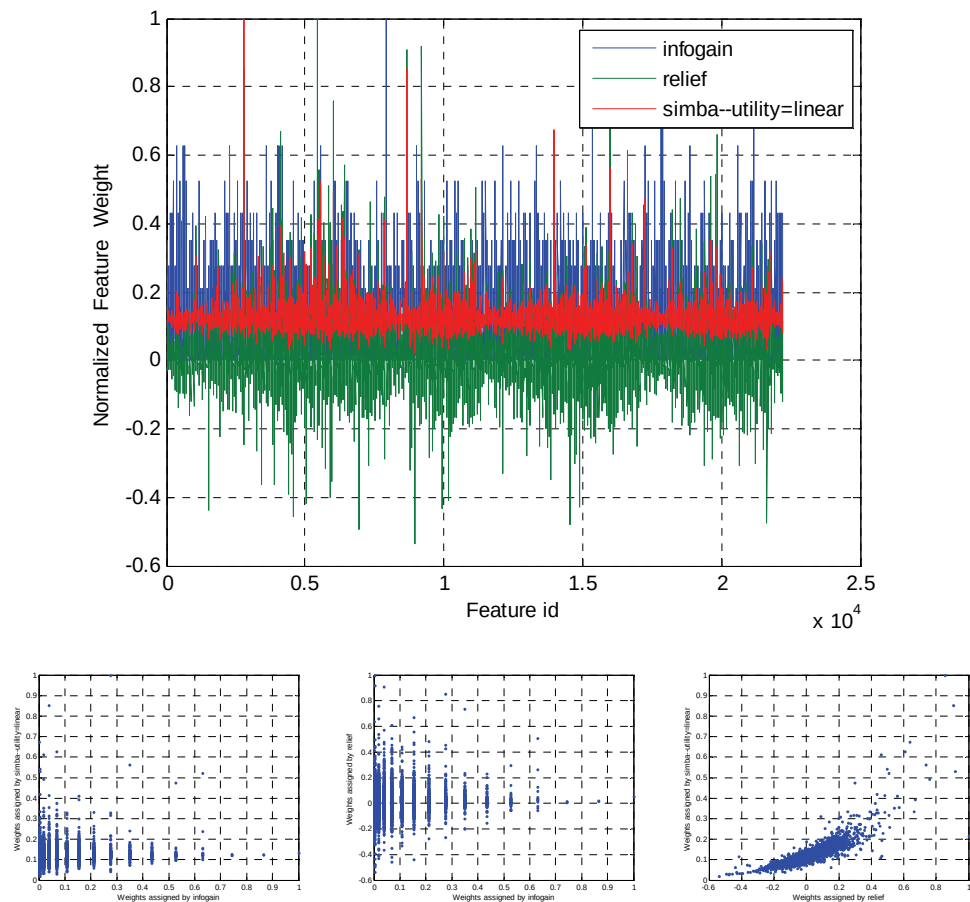
Preciznije očitavanje daje najveću ocenu tačnosti za vrednosti $d=889$ (*Simba* i *Relief*) i $d=7.109$ (*Info Gain*).



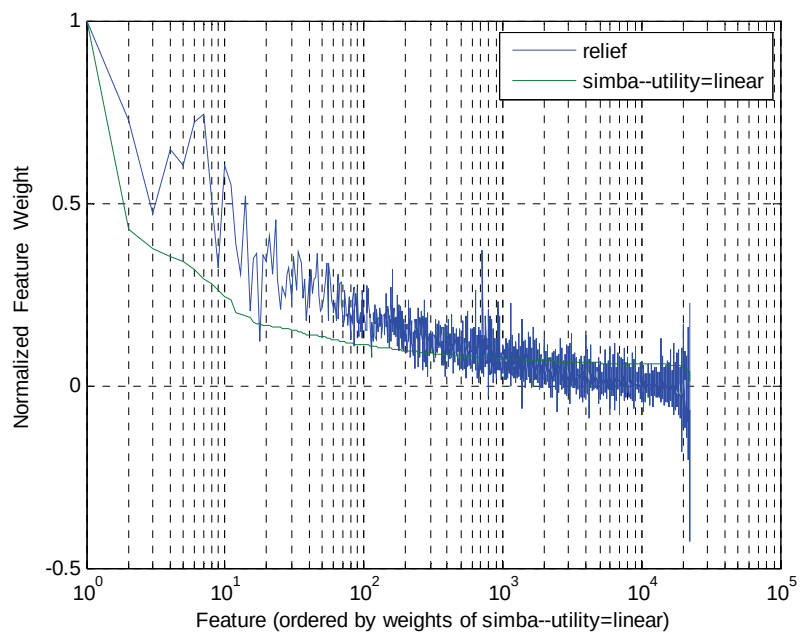
Slika 39: Uticaj broja selektovanih atributa na tačnost predviđanja (Breast Cancer)

Raspodela relevantnosti 22.215 atributa i prikaz međusobne korelacije nekih ocena prikazani su na Sl. 40. Vidi se da korelacija postoji između ocena *Simba* i *Relief*.

Na Sl. 41 se vidi da već za $d > 100$ metodi *Relief* i *Simba* daju nisku ocenu relevantnosti velike većine od ukupno 22.215 atributa.

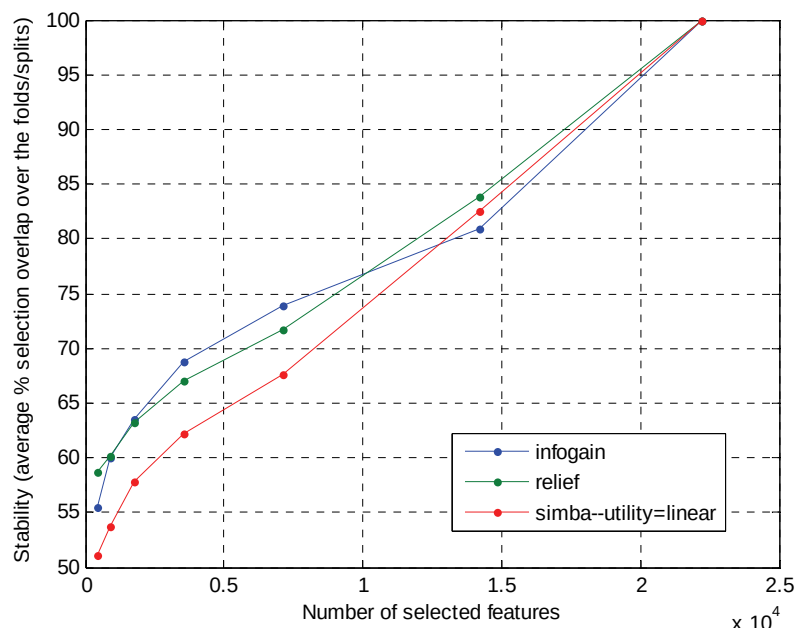


Slika 40: Raspodela ocena relevantnosti atributa metodima InfoGain, Relief i Simba i prikaz međusobne korelacije (Breast Cancer)



Slika 41: Poređenje raspodela ocena relevantnosti atributa metodima Relief i Simba (Breast Cancer)

Na Sl. 42. je prikazana stabilnost estimacije za sva tri metoda. U početku veoma brzo raste, ali maksimum dostiže tek za ceo skup atributa.



Slika 42: Stabilnost estimacije broja atributa (Breast Cancer)

Sama estimacija vrednosti tačnosti predviđanja je od značaja samo za inicijalnu procenu oskudnosti, jer 1-NN metod učenja ne daje objašnjenja klasifikacije, odnosno razumljive rezultate. Direktna primena samostalnih metoda učenja razumljivog znanja (stabala i pravila) na oskudne obučavajuće skupove obično daje nižu vrednost ocene tačnosti predviđanja naučenog znanja od ocene koja se dobije učenjem uz dodatne zahteve razumljivosti.

U praksi se pažljivim izborom metode učenja i pomoćnim tehnikama može naučiti i razumljivo znanje sa zadovoljavajućom estimacijom tačnosti predviđanja.

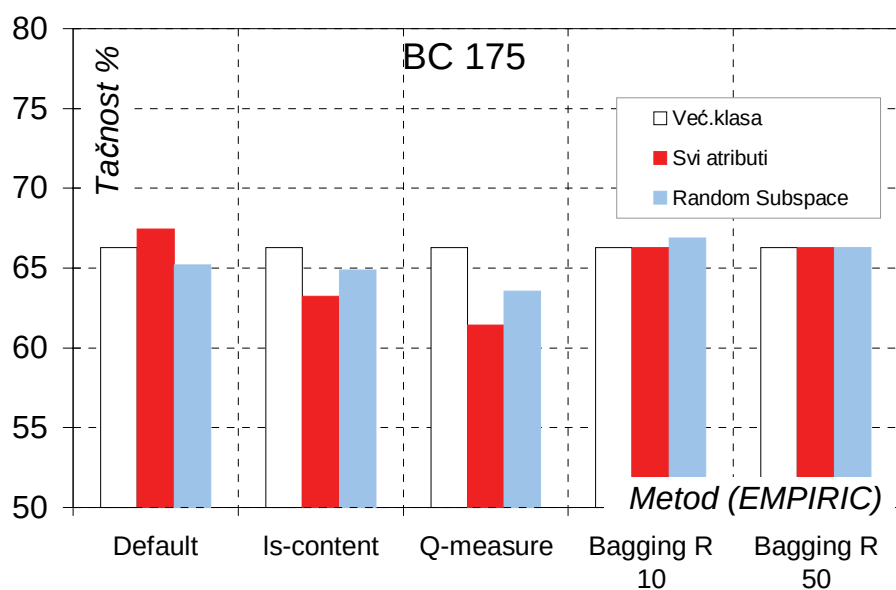
U Tabeli 17. su prikazani rezultati estimacije tačnosti predviđanja dobijeni metodom dvostruke krosvalidacije [37] (za svaki filtrirani model se vrši nezavisna estimacija tačnosti 10-strukom unakrsnom validacijom).

Dataset: BC (175 primera, 22.215 atributa, 2 klase)

Algoritam učenja		Svi atributi	Sluč. podr.
		200	200
EMPIRIC.Rules			
Default		67,43%±5,54	65,19% ±7,70
ls-content		63,23%±7,66	64,89%±6,75
Q-measure		61,44%±5,07	63,55%±11,08
Bagging.Default	10	66,30%±1,51	66,87%±2,30
	50	66,29%±2,01	66,30%±1,51

Tabela 17: Tačnost predviđanja za problem BC (Empiric)

Rezultati estimacije iz Tabele 17. grafički su prikazani na Sl. 43, za skup metoda koje podržava sistem EMPIRIC. Vidi se da se najveća tačnost dobija na celom skupu ili selekcijom atributa prilikom učenja manjih ansambala.



Slika 43: Prikaz tačnosti predviđanja ispitivanih metoda učenja i selekcije atributa iz sistema EMPIRIC za problem BC

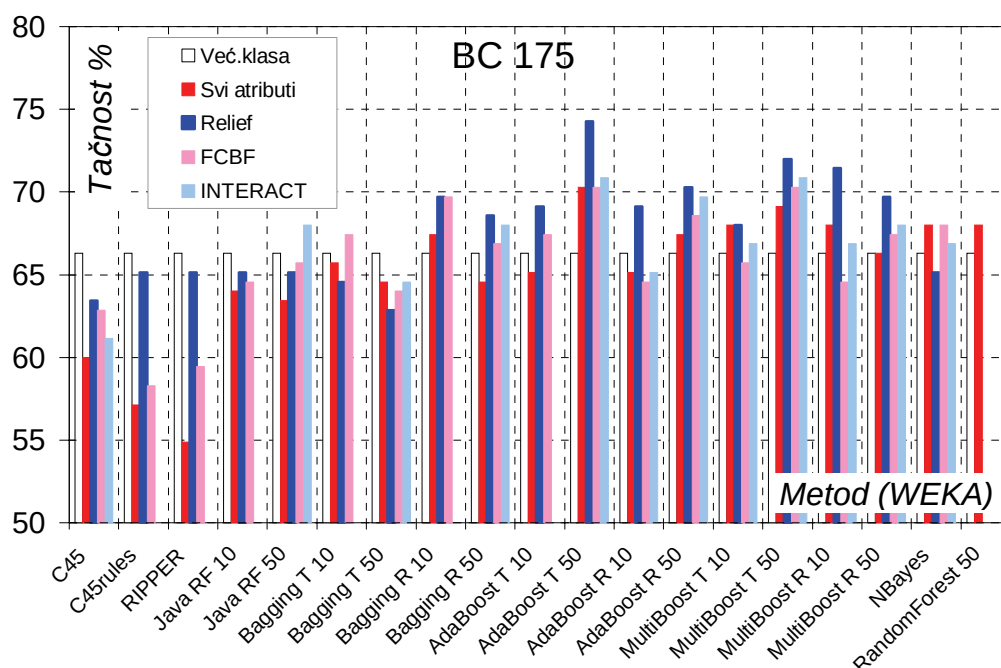
Prikaz estimacije tačnosti predviđanja metoda iz sistema WEKA je u Tabeli 18.

Dataset: BC (175 primera, 22.215 atributa, 2 klase)

Algoritam učenja		Svi atributi 22.215	FCBF 200	RELIEF 200	INTERACT 200
WEKA					
C4.5		60,00%	62,86%	63,43%	61,14%
C4.5Rules		57,14%	58,29%	65,14%	60,57%
RIPPER		54,86%	59,43%	65,14%	
RandomForest	10	64,00%	64,57%	65,14%	
	50	63,43%	65,71%	65,14%	68,00%
Bagging-C4.5	10	65,71%	67,43%	64,57%	
	50	64,57%	64,00%	62,86%	64,57%
Bagging-C4.5rules	10	67,43%	69,71%	69,71%	
	50	64,57%	66,86%	68,57%	68,00%
AdaBoostM1-C4.5	10	65,14%	67,43%	69,14%	
	50	70,29%	70,29%	74,29%	70,86%
AdaBoostM1-C4.5rules	10	65,14%	64,57%	69,14%	65,14%
	50	67,43%	68,57%	70,29%	69,71%
MultiBoostAB-C4 5	10	68,00%	65,71%	68,00%	66,86%
	50	69,14%	70,29%	72,00%	70,86%
MultiBoostAB-C4.5rules	10	68,00%	64,57%	71,43%	66,86%
	50	66,29%	67,43%	69,71%	68,00%
NaiveBayes		68,00%	68,00%	65,14%	66,86%
RandomForest (Breiman-Cutler)					
RandomForest	50	68,00%			

Tabela 18: Tačnost predviđanja za problem BC (WEKA)

Rezultati estimacije tačnosti predviđanja metoda iz sistema *WEKA* i *RandomForest* dobijeni primenom dvostruke krosvalidacione šeme [37] iz Tabele 18. grafički su prikazani na Sl. 44.



Slika 44: Prikaz tačnosti predviđanja ispitivanih metoda učenja i selekcije atributa iz sistema *WEKA* za problem *BC*

Rezultati eksperimenata potvrđuju da je obučavajući skup oskudan, pošto je za sve korišćene metode učenja razumljivog znanja ocena tačnosti predviđanja je niža od Bayesovog klasifikatora (68,00%), pa čak i od verovatnoće klasifikacije većinskog klasifikatora (66,29%).

Upotreba višestrukih klasifikatora povećava tačnost predviđanja, pri čemu su metode *AdaBoost* i *MultiBoost* uspješnije od metoda *Bagging*.

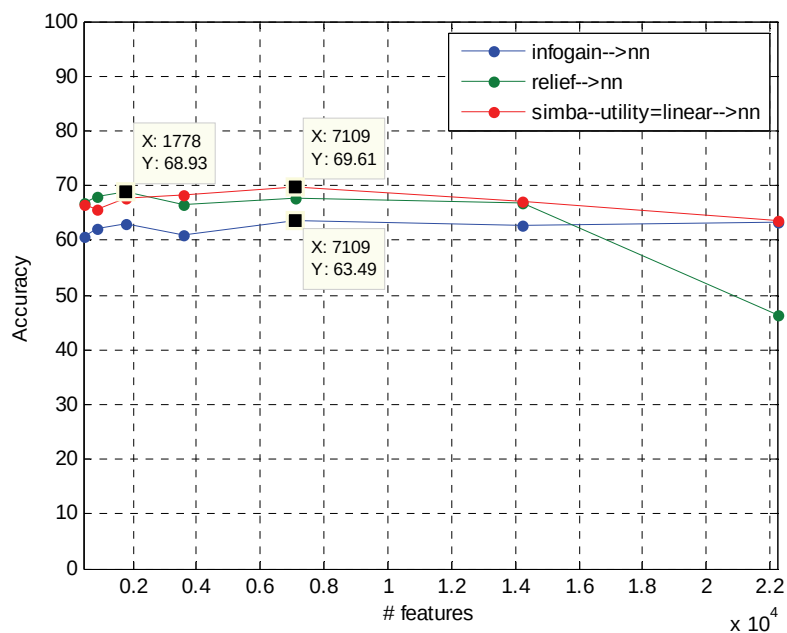
Upotreba selekcije atributa daje takođe poboljšanja, pri čemu se pokazuje se da je u najvećem broju eksperimenata metoda selekcije *Relief* superiorna u odnosu na *FCBF* i *INTERACT*.

Istovremena upotreba obe metode daje najbolje rezultate, Slika 44. Najveća ocena tačnosti predviđanja se postiže kada se koriste metode kombinovanja *AdaBoost* i *MultiBoost*, uz selekciju atributa metodom *Relief*.

8. Problem RFS5

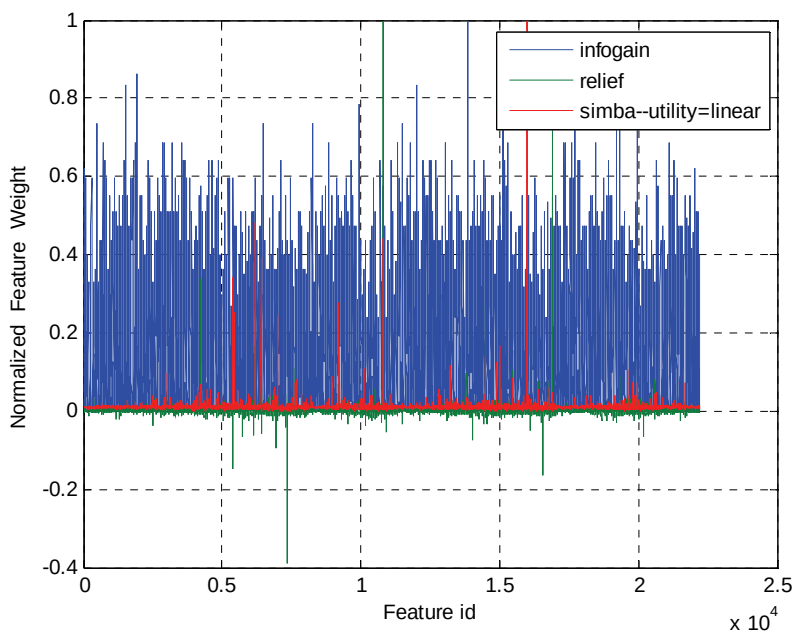
Prikaz uticaja broja selektovanih atributa na tačnost predviđanja pomoću alata *Feature Selection Tool (Matlab)* za primer *RFS5*, uz parametre vidljive sa Sl. 45, pokazuje da se najbolji rezultati učenja očekuju za vrednosti $d=1.778$ (*Relief*) i $d=7.09$ (*Infogain* i *Simba*).

Za velike vrednosti $d>14.000$, prema estimaciji metodom *Info Gain* ostaje skoro ista, a prema *Simba* i *Relief* značajno pada.



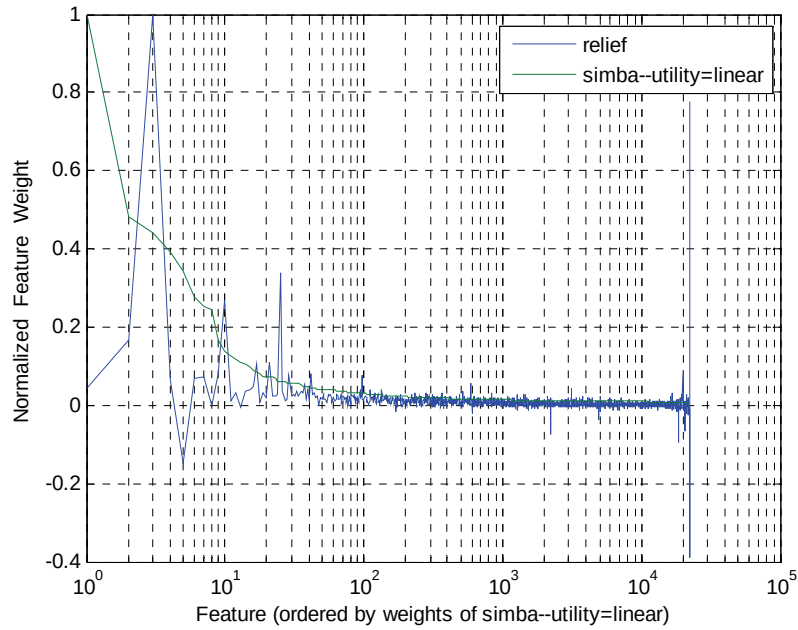
Slika 45: Uticaj broja selektovanih atributa na tačnost predviđanja (RFS5)

Ukupna raspodela ocena je prikazana na Sl. 46, gde se vidi se ocene veoma razlikuju, naročito ocena *InfoGain* u odnosu na ocene *Simba* i *Relief*.



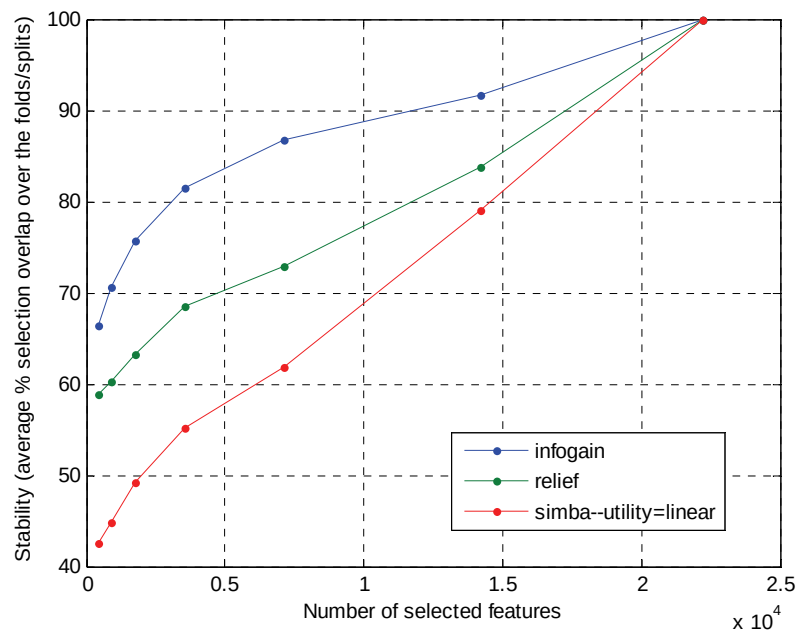
Slika 46: Raspodela ocena relevantnosti atributa metodima *InfoGain*, *Relief* i *Simba* i prikaz međusobne korelacije (Breast Cancer)

Precizniji prikaz raspodela relativno saglasnih ocena *Relief* i *Simba* je na Sl. 47.



Slika 47: Poređenje raspodela ocena relevantnosti atributa metodima Relief i Simba (RFS5)

Stabilnost estimacije je prikazana na Sl. 48. Za sva tri metoda u početku raste brže, ali maksimum dostiže tek za ceo skup atributa.



Slika 48: Stabilnost estimacije broja atributa (RFS5)

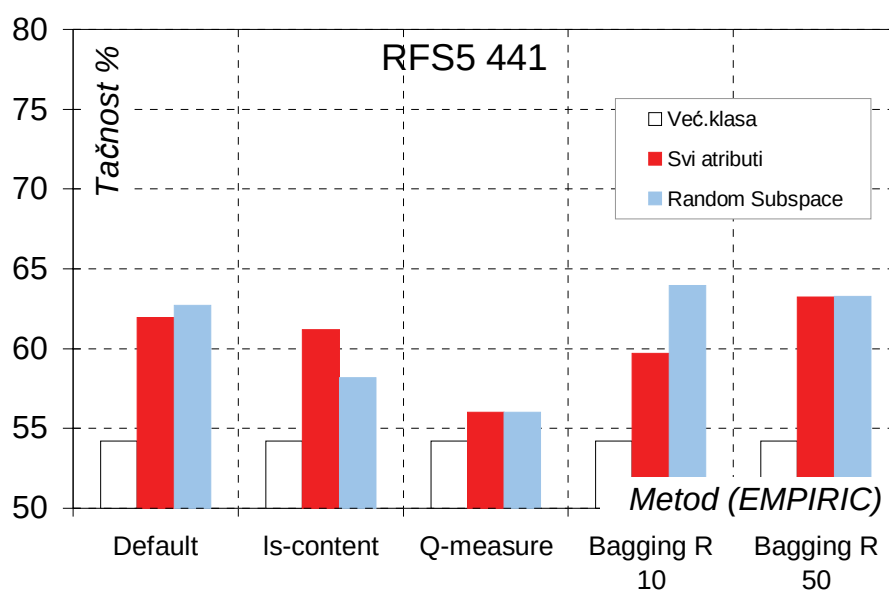
U Tabeli 19. su prikazani rezultati estimacije tačnosti metoda iz sistema *Empiric* predviđanja dobijeni primenom dvostruke krosvalidacione šeme [37].

Dataset: RFS5 (441 primer, 22.215 atributa, 2 klase)

Algoritam učenja		Svi atributi	RELIEF
		22.215	200
EMPIRIC.Rules			
Default		61,95%±6,68	62,72%±7,02
Is-content		61,20%±3,36	58,18%±6,61
Q-measure		56,02%±7,71	56,02%±7,71
Bagging.Default	10	59,69%±5,59	63,94%±4,39
	50	63,24%±4,03	63,26%±4,70

Tabela 19: Tačnost predviđanja za problem RFS5 (Empiric)

Rezultati estimacije iz Tabele 19. grafički su prikazani na Sl. 49, za skup metoda koje podržava sistem *EMPIRIC*.



Slika 49: Prikaz tačnosti predviđanja ispitivanih metoda učenja i selekcije atributa iz sistema *EMPIRIC* za problem RFS5

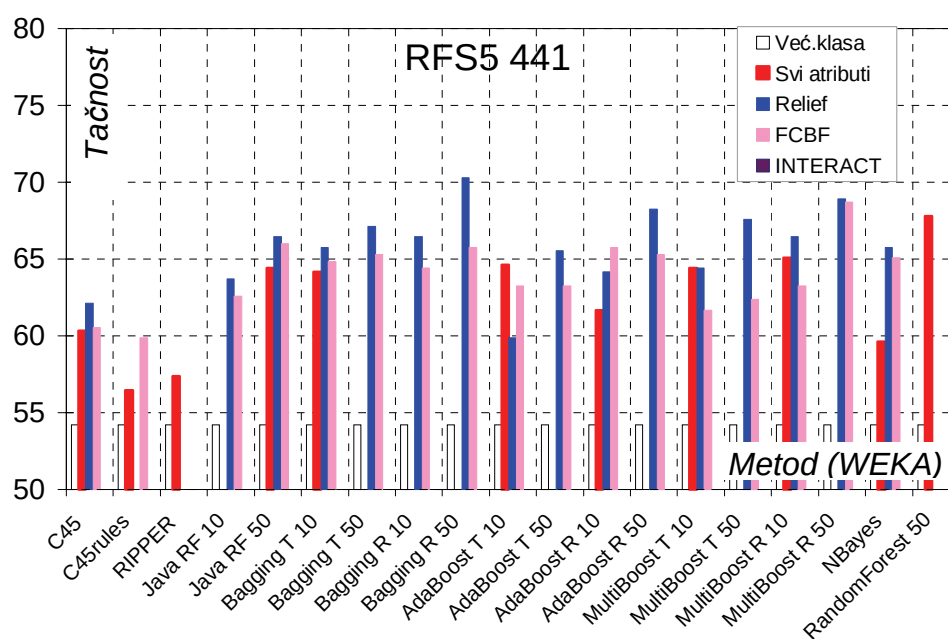
U Tabeli 20. su prikazani rezultati estimacije tačnosti predviđanja metoda iz sistema *WEKA* i *RandomForest* dobijeni primenom dvostruke krosvalidacione šeme [37].

Rezultati estimacije iz Tabele 20. grafički su prikazani na Sl. 50, za skup metoda koje podržavaju sistemi *WEKA* i *RandomForest*.

Dataset: RFS5 (441 primer, 22.215 atributa, 2 klase)

Algoritam učenja		Svi atributi	FCBF	RELIEF	INTERACT
		22.215	200	200	200
WEKA					
C4.5		60,32%	60,54%	62,13%	-
C4.5Rules		56,46%	59,86%		-
RIPPER		57,37%			-
RandomForest	10		62,59%	-	-
	50	64,40%	65,99%	-	-
Bagging-C4.5	10	64,17%	64,85%	65,76%	-
	50		65,31%	67,12%	-
Bagging-C4.5rules	10		64,40%	66,44%	-
	50		65,76%	70,29%	-
AdaBoostM1-C4.5	10	64,63%	63,27%	59,86%	-
	50		63,27%	65,53%	-
AdaBoostM1-C4.5rules	10	61,68%	65,76%	64,17%	-
	50		65,31%	68,25%	-
MultiBoostAB-C4 5	10	64,40%	61,68%	64,40%	-
	50		62,36%	67,57%	-
MultiBoostAB-C4.5rules	10	65,08%	63,27%	66,44%	-
	50		68,71%	68,93%	-
NaiveBayes		59,64%	65,08%	65,76%	-
RandomForest (Breiman-Cutler)					
RandomForest	50	67,80%			-

Tabela 20: Tačnost predviđanja za problem RFS5 (WEKA)

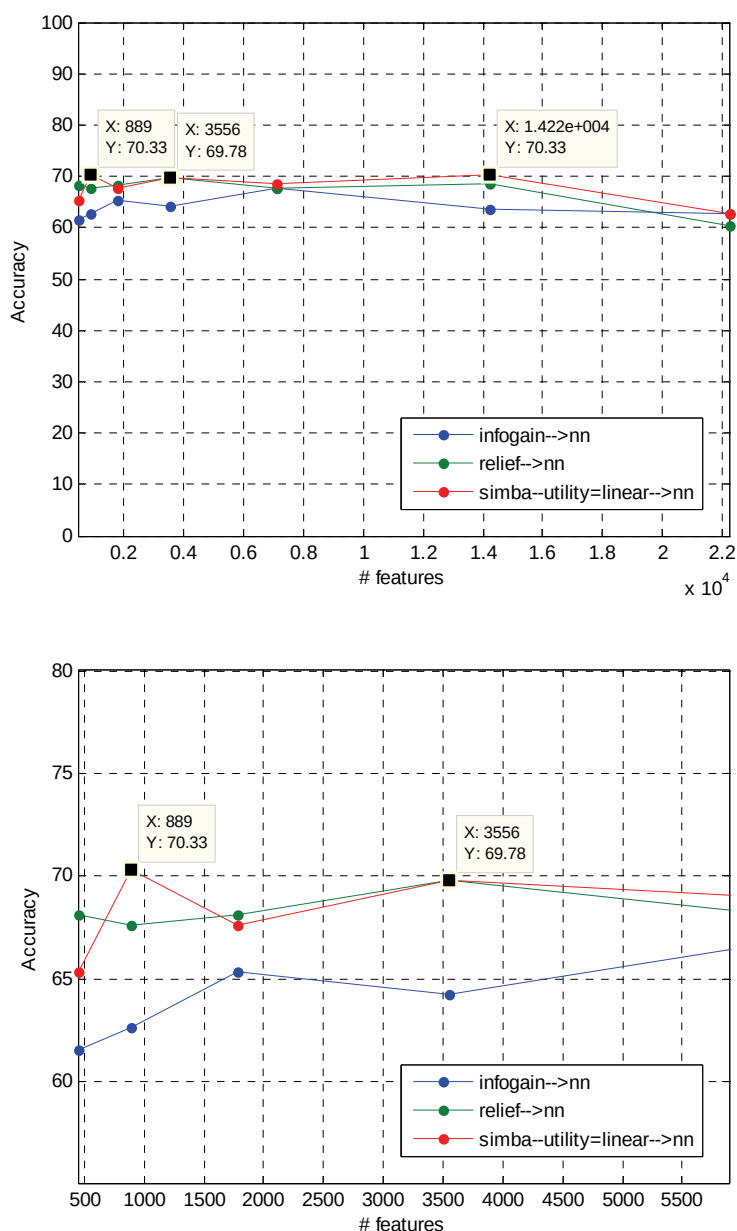


Slika 50: Prikaz tačnosti predviđanja ispitivanih metoda učenja i selekcije atributa iz sistema WEKA za problem RFS5

Vidi se da je metod *Relief* i za ovaj primer najuspešniji u selekciji atributa, a uz istovremenu primenu učenja asambala metodom *bagging* postiže najveću tačnost predviđanja za ovaj primer.

9. Problem NKI BC

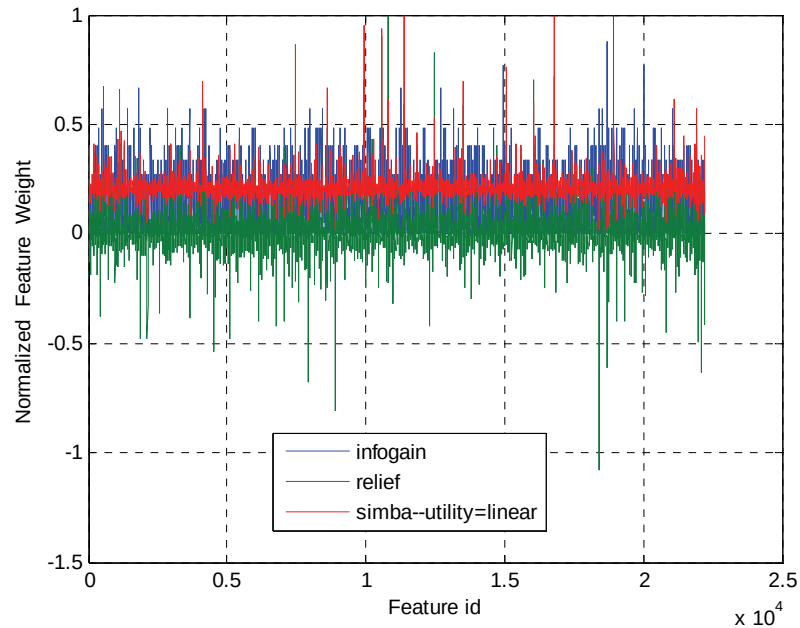
Prikaz uticaja broja selektovanih atributa na tačnost predviđanja pomoću alata *Feature Selection Tool (Matlab)* za primer *NKI BC*, uz parametre vidljive sa Sl. 51, pokazuje da se najbolji rezultati učenja očekuju za različite vrednosti $d=889$ (*Simba*), $d=3.556$ (*Relief*) i $d=7.109$ (*InfoGain*), dok za veće vrednosti $d>14.220$ tačnost značajno opada za sve metode estimacije.



Slika 51: Uticaj broja selektovanih atributa na tačnost predviđanja (NKI BC)

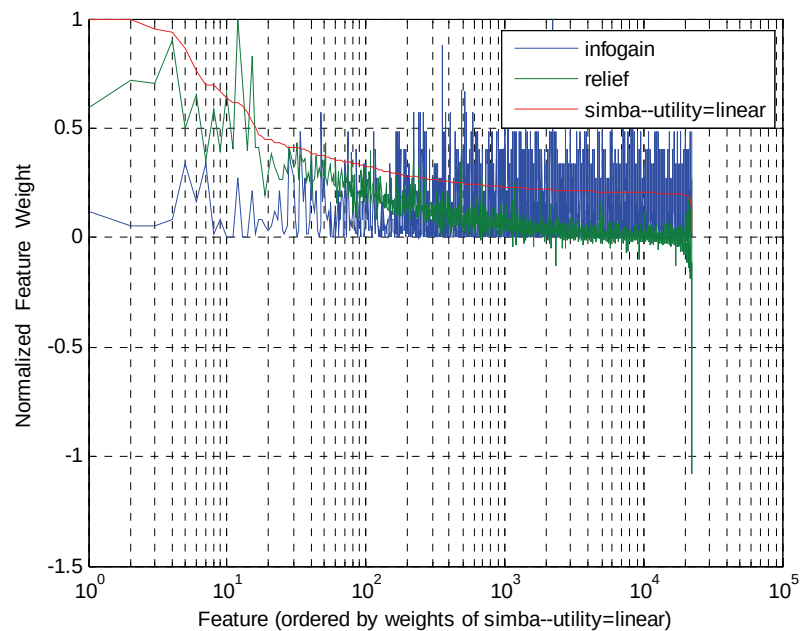
Ukupna raspodela ocena je prikazana na Sl. 46, gde se vidi se ocene veoma razlikuju, naročito ocena *InfoGain* u odnosu na ocene *Simba* i *Relief*.

Ukupna raspodela ocena je prikazana na Sl. 52, gde se vidi se sve ocene razlikuju na celom skupu atributa.



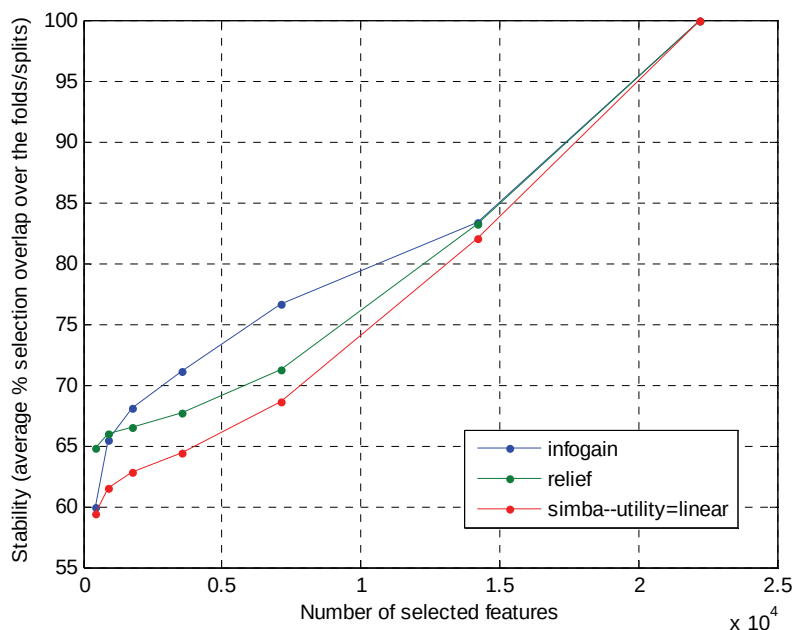
Slika 52: Raspodela ocena relevantnosti atributa metodima InfoGain, Relief i Simba i prikaz međusobne korelacije (NKI BC)

Na Sl. 53 se vidi da tek nakon veoma velikog broja atributa ($d > 20.000$) ocene relevantnosti metoda Simba i Relief postaju veoma male.



Slika 53: Poređenje raspodela ocena relevantnosti atributa metodima InfoGain, Relief i Simba (NKI BC)

Stabilnost estimacije je prikazana na Sl. 54. Za sva tri metoda u početku veoma brzo raste, ali maksimum dostiže tek za ceo skup atributa.



Slika 54: Stabilnost estimacije broja atributa (NKI BC)

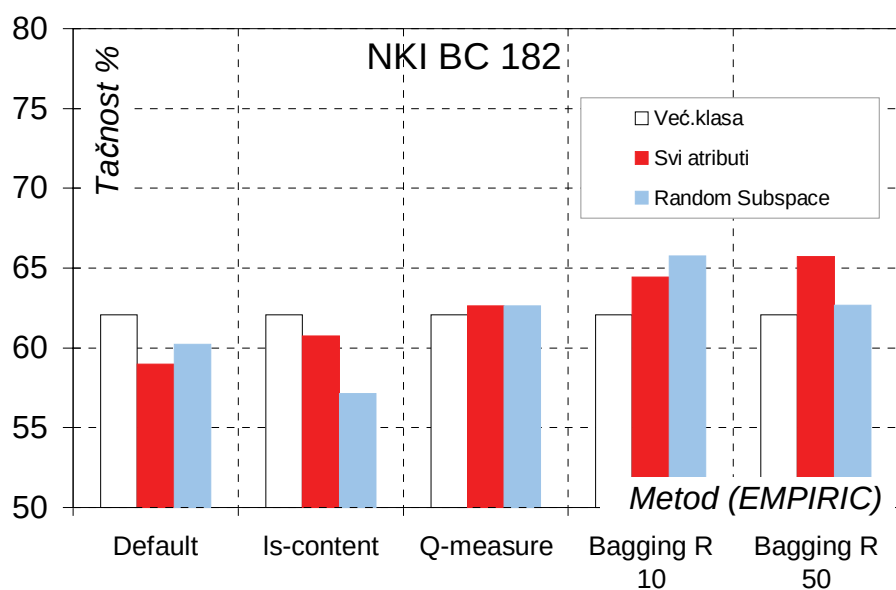
U Tabeli 21. su prikazani rezultati estimacije tačnosti metoda iz sistema *Empiric* predviđanja dobijeni primenom dvostruke krosvalidacione šeme [37].

Dataset: NKI BC (182 primera, 22.223 atributa, 2 klase)

Algoritam učenja		Svi atributi	RELIEF
		22.215	200
EMPIRIC.Rules			
Default		58,99%±9,14	60,24%±8,67
ls-content		60,76%±8,66	57,14%±8,12
Q-measure		62,62%±8,37	62,62%±8,37
Bagging.Default	10	64,44%±2,77	65,75%±3,53
	50	65,71%±4,76	62,67%±3,93

Tabela 21: Tačnost predviđanja za problem NKI BC (Empiric)

Rezultati estimacije iz Tabele 22. grafički su prikazani na Sl. 55, za skup metoda koje podržava sistem *EMPIRIC*. Vidi se da se najveća tačnost dobija selekcijom atributa i učenjem manjih ansambala.



Slika 55: Prikaz tačnosti predviđanja ispitivanih metoda učenja i selekcije atributa iz sistema EMPIRIC za problem NKI BC

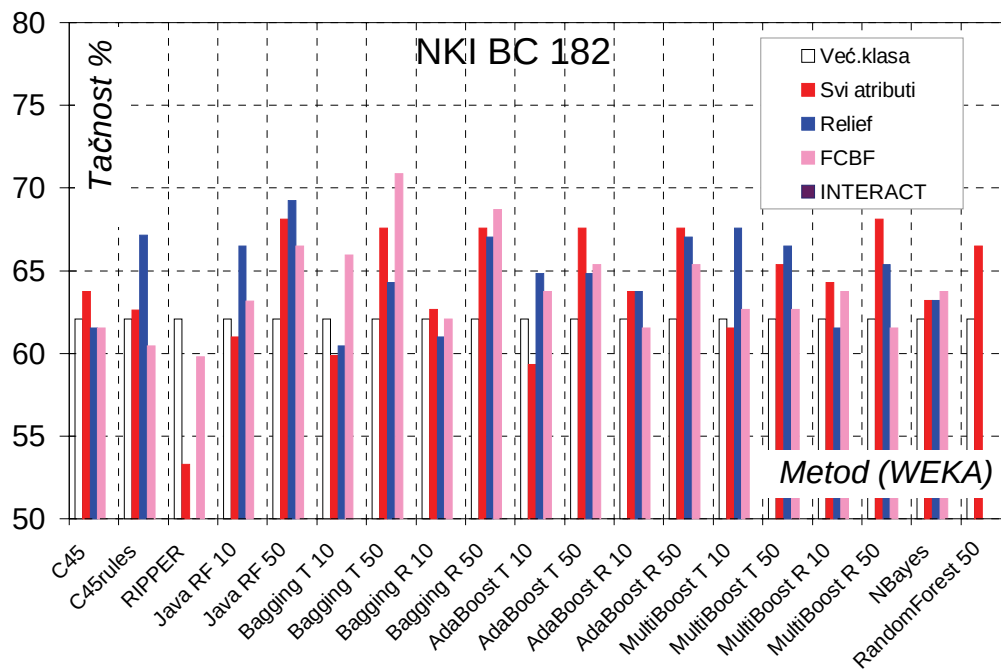
U Tabeli 22. su rezultati estimacije tačnosti sistema WEKA i RandomForest.

Dataset: NKI BC (182 primera, 22.223 atributa, 2 klase)

Algoritam učenja		Svi atributi 22.215	FCBF 200	RELIEF 200	INTERACT 200
WEKA					
C4.5		63,74%	61,54%	61,54%	-
C4.5Rules		62,63%	60,44%	67,14%	-
RIPPER		53,30%	59,79%	46,70%	-
RandomForest	10	60,99%	63,18%	-	-
	50	68,13%	66,48%	-	-
Bagging-C4.5	10	59,89%	65,93%	-	-
	50	67,58%	70,88%	-	-
Bagging-C4.5rules	10	62,64%	62,09%	-	-
	50	67,58%	68,68%	-	-
AdaBoostM1-C4.5	10	59,34%	63,74%	-	-
	50	67,58%	65,38%	-	-
AdaBoostM1-C4.5rules	10	63,74%	61,54%	63,74%	-
	50	67,58%	65,38%	67,03%	-
MultiBoostAB-C4 5	10	61,54%	62,64%	67,58%	-
	50	65,38%	62,64%	66,48%	-
MultiBoostAB-C4.5rules	10	64,29%	63,74%	61,54%	-
	50	68,13%	61,54%	65,38%	-
NaiveBayes		63,19%	63,74%	63,19%	-
RandomForest (Breiman-Cutler)					
RandomForest	50	66,48%	-	-	-

Tabela 22: Tačnost predviđanja za problem NKI BC (WEKA)

Rezultati estimacije tačnosti predviđanja metoda iz sistema *WEKA* i *RandomForest* dobijeni primenom dvostruke krosvalidacione šeme [37] iz Tabele 22. grafički su prikazani na Sl. 56.



Slika 56: Prikaz tačnosti predviđanja ispitivanih metoda učenja i selekcije atributa iz sistema *WEKA* za problem *NKI BC*

Za ovaj primer je veoma različita uspešnost pojedinih metoda: veoma velika povećanja tačnosti su dobijena selekcijom metodom relief (učenje pravila *C45Rules*, *Java RandomForest* i *MultiBoost*), ali je za metod *Bagging* uspešnija selekcija pomoću *FCBF*.

Za ovaj primer, selekcija atributa nije dala efekte prilikom učenja ansambala pravila. Metodi *AdaBoost* i *MultiBoost* najbolje rezultate su pokazali kada se koristio ceo skup primera.

5.4 DISKUSIJA REZULTATA

U ovoj tački se daje sumarni pregled eksperimentalnih rezultata i zaključak o uticaju istovremene primene izabranih metoda prethodne selekcije i učenja malih simboličkih ansambala.

5.4.1 Tačnost predviđanja

Najveće izmerene vrednosti za tačnost predviđanja algoritama učenja stabala i pravila uz upotrebljene metode selekcije atributa prikazane su u Tabeli 23:

Problem	% većin. klase	Svi atributi	Metod selekcije			
			RSS	FCBF	INTERACT	Relief
Monks-2	67,13	71,53	67,00	67,13	53,24	67,13
Waveform-40	33,33	79,20	64,60	76,96	79,16	80,52
Breiman-1K	11,00	76,10	75,42	78,70	76,60	78,40
Breast Cancer LJ	70,28	75,52	73,17	68,53	69,93	70,28
Cleveland Heart	54,46	81,52	76,46	79,54	82,18	82,84
Lung Cancer	40,63	56,25	47,50	56,25	-	65,63
BC	66,29	60,00	68,53	62,86	61,14	65,14
RFS5	54,20	60,32	62,72	60,54	-	62,13
NKI BC	62,64	63,74	62,62	61,54	-	67,14

Tabela 23: Povećanje tačnosti uz upotrebljene metode selekcije atributa

Vidi se da u dva slučaja (problemi Monks-2 i Breast Cancer LJ) upotreba samo selekcije atributa (metodima *RSSM*, *FCBF*, *INTERACT* i *Relief*) smanjila tačnost predviđanja u odnosu na ceo skup atributa.

Za ostale primere, selekcija atributa je povećala tačnost predviđanja bar za jednu od metoda selekcije. Najuspešniji metod je *Relief* (uspešan u 7 od 9 primera), zatim *FCBF* (4 od 9), *INTERACT* (3 od 9) i *Random Subspace* (2 od 9).

Ukupni uticaj učenja ansambala na tačnost predviđanja (uz umeren broj od 50 osnovnih elemenata) prikazan je u Tabeli 24:

Problem	% već. klase	Algoritam	Metod učenja ansambala (50)			
			Osnovni	Bagging	AdaBoost	MultiBoost
Monks-2	67,13	<i>Empiric.Default</i>	61,92	66,92	-	-
		<i>Weka.C45</i>	67,13	64,58	72,45	56,48
		<i>Weka.C45Rules</i>	71,53	78,94	80,32	79,63
Waveform-40	33,33	<i>Empiric.Default</i>	66,77	74,27	-	-
		<i>Weka.C45</i>	75,08	82,54	83,56	83,92
		<i>Weka.C45Rules</i>	77,42	84,54	84,24	84,34
Breiman-1K	11,00	<i>Empiric.Default</i>	68,78	86,62	-	-
		<i>Weka.C45</i>	64,10	90,00	98,10	98,00
		<i>Weka.C45Rules</i>	74,60	94,20	98,90	99,00
Breast Cancer LJ	70,28	<i>Empiric.Default</i>	66,87	70,97	-	-
		<i>Weka.C45</i>	75,52	75,17	68,18	69,93
		<i>Weka.C45Rules</i>	71,33	74,13	69,58	69,93
Cleveland Heart	54,46	<i>Empiric.Default</i>	71,35	80,40	-	-
		<i>Weka.C45</i>	77,56	79,87	80,86	79,54
		<i>Weka.C45Rules</i>	79,88	82,84	80,20	79,54
Lung Cancer	40,63	<i>Empiric.Default</i>	43,33	59,38	-	-
		<i>Weka.C45</i>	50,00	62,50	62,50	62,50
		<i>Weka.C45Rules</i>	56,25	59,38	59,38	62,50
BC	66,29	<i>Empiric.Default</i>	68,53	66,29	-	-
		<i>Weka.C45</i>	60,00	64,57	70,29	69,14
		<i>Weka.C45Rules</i>	57,14	64,57	67,43	66,29
RFS5	54,20	<i>Empiric.Default</i>	61,95	63,24	-	-
		<i>Weka.C45</i>	60,32	64,17	64,63	64,40
		<i>Weka.C45Rules</i>	56,46		61,68	65,08
NKI BC	62,64	<i>Empiric.Default</i>	58,99	65,71	-	-
		<i>Weka.C45</i>	63,74	67,58	67,58	65,38
		<i>Weka.C45Rules</i>	62,63	67,58	67,58	68,13

Tabela 24: Povećanje tačnosti primenom metoda učenja ansambala

Vidi se da je učenje ansambala povećalo tačnost predviđanja, osim u dva slučaja. Za probleme *Breast Cancer LJ* i *BC* osnovni algoritmi učenja *Weka.C45* i *Empiric.Default* samostalno postižu najveću tačnost, odnosno njihova interna metoda selekcije atributa je efikasnija ako se koristi bez prethodne selekcije atributa.

Povećanje tačnosti je različito za svaki problem i metod učenja ansambala. Metod *Bagging* je u ovom skupu merenja uspešan u najvećem broju slučajeva, zatim *AdaBoost* i *MultiBoost*.

Detaljni rezultati *istovremene* primene metoda selekcije i metoda učenja relativno malih ansambala (uz relativno mali broj od 50 elemenata) prikazani su u Tabeli 25:

Problem	Algoritam	Metod ansambala	Metod selekcije				
			Svi atributi	Sluč. podp.	FCBF	Interact	Relief
1. Monks-2	Empiric.Default	-	61,92	67,00	-	-	-
		Bagging	66,92	67,00	-	-	-
	Weka.C4.5	-	67,13	-	67,13	47,92	67,13
		Bagging	64,58	-	67,13	47,92	67,13
		AdaBoost	72,45	-	67,13	47,92	67,13
		MultiBoost	56,48	-	67,13	47,92	67,13
	Weka.C4.5rules	-	71,53	-	67,13	48,84	67,13
		Bagging	78,94	-	67,13	47,92	67,13
		AdaBoost	80,32	-	67,13	48,84	67,13
		MultiBoost	79,63	-	67,13	48,84	67,13
	Weka.Random Forest		34,95	-	67,13	49,54	67,13
2. Waveform-40	Empiric.Default	-	66,77	63,49	-	-	-
		Bagging	74,27	77,54	-	-	-
	Weka.C4.5	-	75,08	-	76,96	76,42	77,56
		Bagging	82,54	-	78,88	82,78	83,94
		AdaBoost	83,56	-	76,46	83,22	84,62
		MultiBoost	83,92	-	76,86	83,62	84,78
	Weka.C4.5rules	-	77,42	-	76,22	78,02	79,56
		Bagging	84,54	-	78,76	83,80	85,34
		AdaBoost	84,24	-	77,54	83,58	84,58
		MultiBoost	84,34	-	78,10	83,64	84,86
	Weka.Random Forest		84,54	-	-	83,30	84,98
3. Breiman-1K	Empiric.Default	-	68,78	67,01	-	-	-
		Bagging	86,62	95,93	-	-	-
	Weka.C4.5	-	64,10	-	67,80	72,10	68,70
		Bagging	90,00	-	89,90	80,70	90,60
		AdaBoost	98,10	-	98,00	81,00	98,80
		MultiBoost	98,00	-	98,00	81,50	98,00
	Weka.C4.5rules	-	74,60	-	78,70	76,60	77,80
		Bagging	94,20	-	94,40	82,60	93,40
		AdaBoost	98,90	-	98,90	82,50	99,20
		MultiBoost	99,00	-	98,60	82,40	98,90
	Weka.Random Forest		99,40	-	99,20	80,50	99,10

Problem	Algoritam	Metod ansambala	Metod selekcije				
			Svi atributi	Sluč. podp.	FCBF	Interact	Relief
4. Breast Cancer LJ	Empiric.Default	-	66,87	66,77	-	-	-
		Bagging	70,97	71,37	-	-	-
	Weka.C4.5	-	75,52	-	67,13	68,88	69,23
		Bagging	75,17	-	68,88	69,23	70,99
		AdaBoost	68,18	-	67,83	65,38	65,73
		MultiBoost	69,93	-	67,83	65,04	66,78
	Weka.C4.5rules	-	71,33	-	68,53	68,88	68,53
		Bagging	74,13	-	67,48	68,88	69,23
		AdaBoost	69,58	-	67,13	69,58	68,53
		MultiBoost	69,93	-	66,08	69,58	66,08
	Weka.Random Forest		68,53	-	67,83	68,18	68,53
5. Cleveland Heart	Empiric.Default	-	71,35	76,46	-	-	-
		Bagging	80,40	80,41	-	-	-
	Weka.C4.5	-	77,56	-	76,90	76,90	78,22
		Bagging	79,87	-	79,87	79,87	80,53
		AdaBoost	80,86	-	78,55	77,89	79,87
		MultiBoost	79,54	-	79,87	79,54	79,21
	Weka.C4.5rules	-	79,88	-	79,21	78,88	80,20
		Bagging	82,84	-	79,87	82,18	83,17
		AdaBoost	80,20	-	78,88	78,88	79,87
		MultiBoost	79,54	-	78,88	79,21	80,86
	Weka.Random Forest		82,18	-	-	80,53	82,18
6. Lung Cancer	Empiric.Default	-	43,33	47,50	-	-	-
		Bagging	55,00	47,50	-	-	-
	Weka.C4.5	-	50,00	-	56,25	-	59,38
		Bagging	59,38	-	46,88	-	62,50
		AdaBoost	59,38	-	46,88	-	62,50
		MultiBoost	59,38	-	50,00	-	62,50
	Weka.C4.5rules	-	56,25	-	56,25	-	65,63
		Bagging	62,50	-	56,13	-	59,38
		AdaBoost	62,50	-	53,13	-	59,38
		MultiBoost	56,25	-	53,13	-	56,25
	Weka.Random Forest		46,88	-	50,00	-	65,63

Problem	Algoritam	Metod ansambala	Metod selekcije				
			Svi atributi	Sluč. podp.	FCBF	Interact	Relief
7. BC	Empiric.Default	-	68,53	-	-	-	-
		Bagging	66,29	66,30	-	-	-
	Weka.C4.5	-	60,00	-	62,86	61,14	63,43
		Bagging	64,57	-	64,00	64,57	62,86
		AdaBoost	70,29	-	70,29	70,86	74,29
		MultiBoost	69,14	-	70,29	70,86	72,00
	Weka.C4.5rules	-	57,14	-	58,29	60,57	65,14
		Bagging	64,57	-	66,86	68,00	68,57
		AdaBoost	67,43	-	68,57	69,71	70,29
		MultiBoost	66,29	-	67,43	68,00	69,71
	Weka.Random Forest		63,43	-	65,71	68,00	65,14
8. RFS5	Empiric.Default	-	61,95	62,72	-	-	-
		Bagging	63,24	63,26	-	-	-
	Weka.C4.5	-	60,32	-	60,54	-	62,13
		Bagging	-	-	65,31	-	67,12
		AdaBoost	-	-	63,27	-	65,53
		MultiBoost	-	-	62,36	-	67,57
	Weka.C4.5rules	-	56,46	-	59,86	-	-
		Bagging	-	-	65,31	-	67,12
		AdaBoost	-	-	63,27	-	65,53
		MultiBoost	-	-	62,36	-	67,57
	Weka.Random Forest		64,40	-	65,99	-	-
9. NKI BC	Empiric.Default	-	58,99	60,24	-	-	-
		Bagging	65,71	62,67	-	-	-
	Weka.C4.5	-	63,74	-	61,54	-	61,54
		Bagging	67,58	-	70,88	-	64,29
		AdaBoost	67,58	-	65,38	-	64,84
		MultiBoost	65,38	-	62,64	-	67,58
	Weka.C4.5rules	-	62,63	-	60,44	-	67,14
		Bagging	67,58	-	70,88	-	64,29
		AdaBoost	67,58	-	65,38	-	64,84
		MultiBoost	65,38	-	62,64	-	66,48
	Weka.Random Forest		68,13	-	66,48	-	-

Tabela 25: Povećanje tačnosti istovremenom primenom metoda selekcije atributa i učenja do 50 ansambala

Pregled odnosa najveće uspešnosti svih razmotrenih metoda da je u Tabeli 26:

Problem	% većin. klase	Osnovni algoritam	Metodi selekcije	Metodi ansambala	Istovremena primena
Monks-2	67,13	71,53	67,13	80,32	80,32
Waveform-40	33,33	79,20	80,52	84,54	84,98
Breiman-1K	11,00	76,10	78,70	99,00	99,20
Breast Cancer LJ	70,28	75,52	70,28	75,17	71,37
Cleveland Heart	54,46	81,52	82,84	82,84	83,17
Lung Cancer	40,63	56,25	65,63	62,50	65,63
BC	66,29	60,00	68,53	70,29	70,29
RFS5	54,20	60,32	62,72	65,08	67,57
NKI BC	62,64	63,74	67,14	68,13	70,88

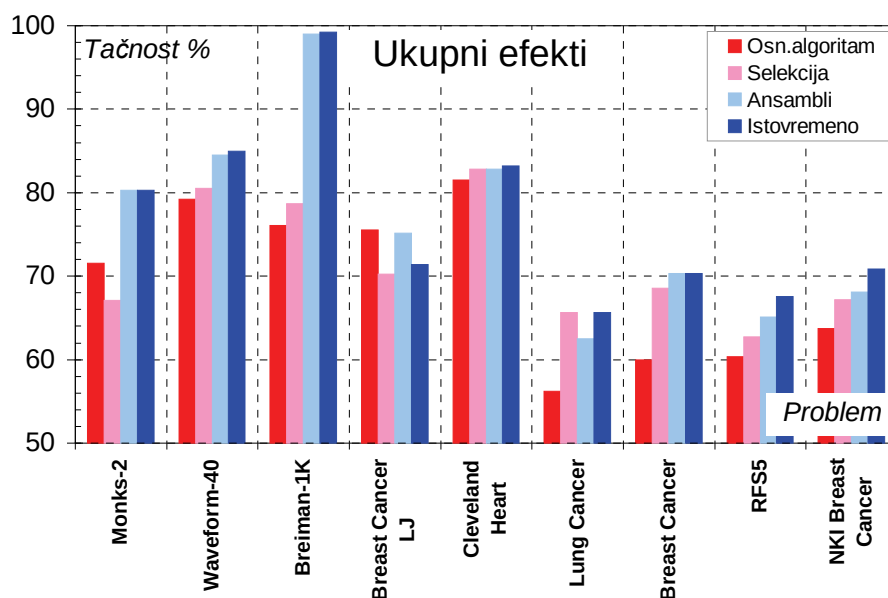
Tabela 26: Poređenje najvećeg povećanja tačnosti istovremenom primenom metoda selekcije atributa i učenja do 50 ansambala

Istovremena primena izabranih metoda je znatno poboljšala performanse sistema za induktivno učenje pravila za sve primere, izuzev problema *Breast Cancer Ljubljana*.

Poznati algoritma učenja C4.5, koji koristi ocenu informativnosti (*InformationGain*) za upravljanje generalizacijom, pokazao se najboljim upravo za ovaj problem učenja, Tabela 25.

Za ostale metode učenja na ovom problemu se dobijaju znatna poboljšanja kada se metod selekcije ili metod ansambala koriste pojedinačno, dok njihova *istovremena primena ne daje značajna poboljšanja za*, Tabela 25 i Tabela 26.

Prikaz ukupnih efekata svih upotrebljenih metoda dat je na Sl. 57.

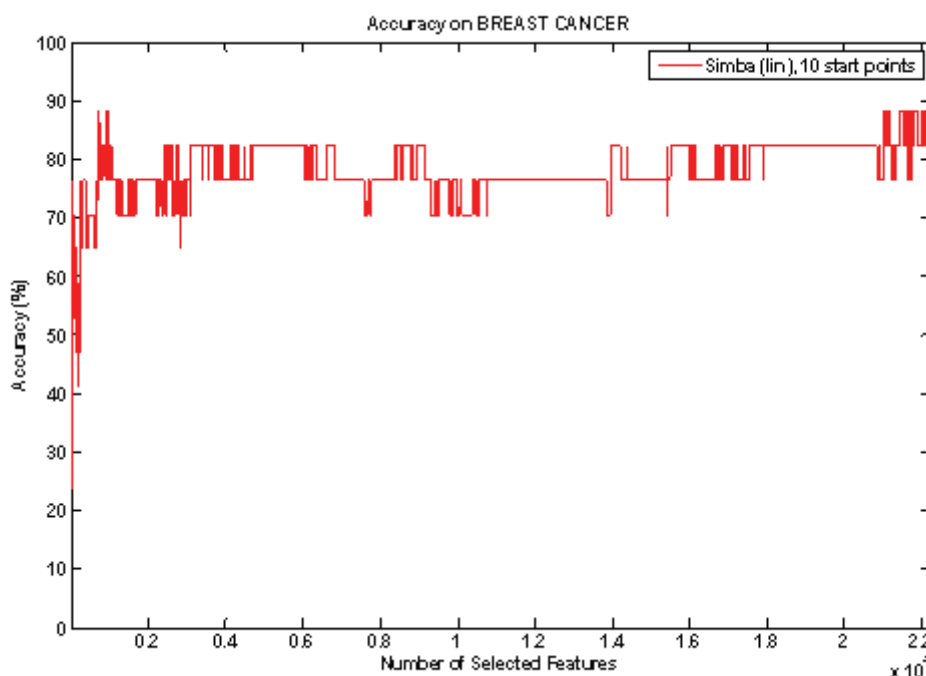


Slika 57: Prikaz odnosa najveće tačnosti predviđanja za sve upotrebljene metode

Najveća tačnost za probleme učenja na osnovu genskih ekspresija i ansamble od 50 elemenata je 74,29% za problem BC, 67,57% za problem RFS5 (veća je za ansamble od 10 elemenata, 70,29%) i 70,88% za problem NKI BC. Ova tačnost se dobija rigoroznom eksperimentalnom procedurom učenja i estimacije greške [37] i još uvek nije dovoljna za ozbiljne praktične primene.

Navedeni rezultati su ostali u okviru estimacije mogućnosti prethodne selekcije atributa koja je ustanovljena metodom *SIMBA*, zasnovanom na konceptu margine [13], [14] za konkretne obučavajuće skupove. Estimacija ovom metodom pokazuje da li je moguće je identifikovati pogodan podskup informativnih atributa, za koji i veoma jednostavni algoritmi učenja postižu povećanje tačnosti predviđanja.

Npr. za problem BC metod 1-NN [18], uz prethodnu selekciju algoritmom *SIMBA*, postiže tačnost veću od 75% za 100 do 1.500 selektovanih atributa, Sl. 58.



Slika 58: Prikaz tačnosti predviđanja algoritma 1-NN za problem BC uz prethodnu selekciju metodom *SIMBA*

Nešto manju tačnost 1-NN metoda za približno isti broj atributa predviđa i *Feature Selection Tool*, Sl. 39.

Predviđena tačnost predviđanja na osnovu ovih estimacija je dostignuta u rezultatima izvedenih eksperimenata, s tim da su u ovom radu upotrebljene samo metode učenja koje daju razumljive rezultate, za razliku od 1-NN i drugih sličnih metoda, koji mogu pokazati i veću tačnost, ali bez razumljivog objašnjenja.

5.4.2 Razumljivost

Razumljivost naučenih stabala i pravila je teško objektivno ocenjivati [1], [24], [25], [26], [66], [67], [68], ali je prirodno očekivati da kraći opisi budu razumljiviji od

obimnijih opisa (stabla sa manjim brojem čvorova, skupovi pravila sa manjim brojem jednostavnijih pravila).

Razumljivost je posebno značajna za probleme učenja genskih ekspresija, najobimnije realne probleme sa svojstvima oskudnosti koji se koriste u ovom radu.

Primeri znanja naučenog pomoću algoritama *Empiric.Rules.Default*, *WEKA.C45*, *WEKA.C45Rules* i *WEKA.RIPPER* za najobimniji problem RFS5 dati su na Sl. 59-62.

```
Rule1 (0,738): [220173_at=-0,0597..0,6559][220039_s_at=-2,0791..-0,5183] -> [Class=0] 158/56
Rule2 (1,000): [213562_s_at=0,2526..0,4203][219453_at=-0,1245..0,1515][121_at=0,2944..0,8117] ->
[Class=0] 20/0
Rule3 (1,000): [217741_s_at=0,9648..1,0138][204501_at=-0,6430..0,0461] -> [Class=0] 14/0
Rule4 (1,000): [209392_at=0,5006..0,5273][200016_x_at=1,7886..1,9979] -> [Class=0] 9/0
Rule5 (1,000): [211609_x_at=1,1832..1,2010][1255_g_at=-1,2358..-0,3931] -> [Class=0] 8/0
Rule6 (1,000): [201634_s_at=0,3471..0,3606] -> [Class=0] 4/0
Rule7 (1,000): [200671_s_at=-0,7737..-0,7572] -> [Class=0] 3/0
Rule8 (1,000): [1007_s_at=1,0814] -> [Class=0] 1/0
Rule9 (0,675): [214464_at=-1,1012..0,2418][213684_s_at=0,0558..0,6767] -> [Class=1] 139/67
Rule10 (1,000): [212547_at=0,5717..0,6427][204455_at=-0,0538..1,0996] -> [Class=1] 14/0
Rule11 (1,000): [206947_at=-0,2610..-0,2409][205037_at=0,2966..0,7399] -> [Class=1] 11/0
Rule12 (1,000): [202664_at=0,8594..0,9610][214851_at=-0,6889..-0,3193] -> [Class=1] 10/0
Rule13 (1,000): [209026_x_at=1,2889..1,3213][1294_at=0,1433..0,4336] -> [Class=1] 5/0
Rule14 (1,000): [202238_s_at=1,2457..1,2858] -> [Class=1] 4/0

Number of
Rules      =14
Selectors=25
MDL(0,010)=384,168 bits
```

Slika 59: Primer naučenog skupa pravila za problem RFS5 (*Empiric*)

```
J48 pruned tree
-----
216555_at <= 0.01287
|
| 215978_x_at <= 0.584703: 0 (20.0)
| | 215978_x_at > 0.584703
| | | 204156_at <= 0.422166
| | | | 200953_s_at <= 0.999618
| | | | | 219869_s_at <= 0.660908
| | | | | | 217229_at <= -1.869942
| | | | | | | 201728_s_at <= 0.116936: 0 (3.0)
| | | | | | | | 201728_s_at > 0.116936
| | | | | | | | | 221657_s_at <= 0.401678: 1 (32.0)
| | | | | | | | | 221657_s_at > 0.401678: 0 (2.0)
| | | | | | | | | 217229_at > -1.869942
| | | | | | | | | | 219369_s_at <= -0.80586: 0 (17.0)
| | | | | | | | | | 219369_s_at > -0.80586
| | | | | | | | | | | 221785_at <= -1.068339: 1 (13.0)
| | | | | | | | | | | 221785_at > -1.068339
| | | | | | | | | | | | 203010_at <= 0.6046
| | | | | | | | | | | | | 204230_s_at <= 0.076878
| | | | | | | | | | | | | | 221099_at <= -0.774281: 0 (16.0)
| | | | | | | | | | | | | | 221099_at > -0.774281
| | | | | | | | | | | | | | | 220196_at <= -1.299159: 1 (14.0)
| | | | | | | | | | | | | | | 220196_at > -1.299159
| | | | | | | | | | | | | | | | 210113_s_at <= -1.149681: 0 (15.0)
| | | | | | | | | | | | | | | | 210113_s_at > -1.149681
| | | | | | | | | | | | | | | | | 208776_at <= 0.251415
| | | | | | | | | | | | | | | | | | 202992_at <= -0.732586: 0 (10.0)
| | | | | | | | | | | | | | | | | | 202992_at > -0.732586
| | | | | | | | | | | | | | | | | | | 207298_at <= -0.884548: 1 (23.0)
| | | | | | | | | | | | | | | | | | | 207298_at > -0.884548
| | | | | | | | | | | | | | | | | | | | 219230_at <= -0.525746: 1 (18.0)
| | | | | | | | | | | | | | | | | | | | 219230_at > -0.525746
| | | | | | | | | | | | | | | | | | | | | 220964_s_at <= 0.706046: 1 (12.0)
| | | | | | | | | | | | | | | | | | | | | 220964_s_at > 0.706046
| | | | | | | | | | | | | | | | | | | | | | 203873_at <= -0.694168: 1 (8.0)
| | | | | | | | | | | | | | | | | | | | | | 203873_at > -0.694168
| | | | | | | | | | | | | | | | | | | | | | | 205598_at <= -0.423186: 1 (10.0/1.0)
| | | | | | | | | | | | | | | | | | | | | | | 205598_at > -0.423186
| | | | | | | | | | | | | | | | | | | | | | | | 202456_s_at <= -0.009709: 0
| | | | | | | | | | | | | | | | | | | | | | | | | 202456_s_at > -0.009709: 1 (4.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | 204230_s_at > 0.076878: 0 (20.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | 203010_at > 0.6046: 1 (12.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | 219869_s_at > 0.660908: 0 (16.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | 200953_s_at > 0.999618: 1 (25.0/1.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | 204156_at > 0.422166: 1 (31.0/1.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | 216555_at > 0.01287: 0 (60.0/1.0)

Number of Leaves :      24
Size of the tree :      47
```

Slika 60: Primer naučenog stabla za problem RFS5 (*WEKA.C45*)

```

PART decision list
-----

216555_at > 0.01287: 0 (60.0/1.0)

215978_x_at > 0.584703 AND
204156_at <= 0.422166 AND
200953_s_at <= 0.999618 AND
219869_s_at <= 0.668908 AND
217229_at <= -1.869942 AND
201728_s_at > 0.116936 AND
221657_s_at <= 0.401678: 1 (32.0)

204156_at > 0.422166: 1 (31.0/1.0)

200953_s_at > 0.999618: 1 (25.0/1.0)

207176_s_at <= -0.183604: 1 (17.0/1.0)

220096_at > 0.100127 AND
200018_at > 1.898022: 1 (15.0)

202257_s_at > 0.826548 AND
201253_s_at <= 1.456523: 1 (19.0)

212953_x_at <= 0.809469 AND
201218_at <= 1.129682: 1 (11.0)

207504_at <= 0.195194 AND
209556_at <= 0.423106 AND
204270_at > -0.18837 AND
216674_at <= -0.477459 AND
216422_at > -0.75606 AND
220656_at > -2.106392 AND
219464_at <= 0.237691: 0 (139.0/1.0)

208270_s_at <= 1.13568 AND
208974_x_at > 0.927554 AND
207873_x_at <= 0.145339 AND
218898_at > -0.113927 AND
216139_s_at > -1.307294 AND
212159_x_at <= 0.999859: 1 (54.0/1.0)

: 0 (38.0)

Number of Rules :      11

```

Slika 61: Primer naučenog skupa pravila za problem RFS5 (WEKA.C45Rules)

```

JRIP rules:
=====

{203013_at >= 0.574033} and {220295_x_at <= -0.21259} => class=1 (61.0/6.0)
{211964_at <= 0.92563} and {206773_at <= -0.308441} => class=1 (49.0/8.0)
{210358_x_at >= 0.518107} and {218391_at <= 0.873324} and {1773_at >= -0.255042} => class=1 (41.0/8.0)
{206496_at >= -0.282137} and {217158_at >= -0.212959} => class=1 (26.0/5.0)
{201572_x_at >= 0.806074} and {201907_x_at <= -0.228367} => class=1 (31.0/11.0)
{201153_s_at >= 1.102134} and {202099_s_at >= -0.146055} => class=1 (16.0/3.0)
{204549_at >= 0.30407} and {208493_at >= -1.086198} and {201673_s_at <= 0.557974} => class=1 (13.0/3.0)
{214402_s_at >= -0.035448} and {211846_s_at >= -0.283851} and {1316_at <= -0.12128} => class=1 (5.0/0.0)
{205575_at >= -0.30043} and {200057_s_at <= 1.373881} => class=1 (3.0/0.0)
=> class=0 (196.0/1.0)

Number of Rules : 10

```

Slika 62: Primer naučenog skupa pravila za problem RFS5 (WEKA.RIPPER)

Upotrebljeni algoritmi induktivnog učenja za probleme genskih ekspresija, opisane desetinama hiljada atributa, tipično indukuju skupove pravila reda *samo nekoliko desetina* osnovnih elemenata (selektora), odnosno 10-50 pravila.

Ovako kratka rešenja, koja koriste svega nekoliko desetina gena, dobijaju se jer je osnovni mehanizam algoritma indukcije svake od razmatranih metoda učenja ugrađen je i neki metod pojednostavljivanja (*simplification, pruning*), kojim se poboljšavaju generalizacione performanse i dobijaju znatno kraći i tačniji opisi (pravila, liste ili stabla odlučivanja).

Na taj način već i osnovni algoritmi učenja tipično daju kratka rešenja, odnosno potencijalno razumljivo znanje.

Primeri na Sl. 59-62. ilustruju oblik i složenost naučenog znanja u nepovoljnijem slučaju, kada se koristi ceo skup atributa.

Pravila klasifikacije se odnose na nazive i vrednosti ekspresije izabranih atributa (gena) [9], [10], [29]. Dodatni podaci predstavljaju broj obučavajućih primera (pozitivnih i negativnih) koji su u procesu učenja poslužili za formiranje konkretnog pravila ili čvora stabla odlučivanja.

Cilj ovog rada je bio da se poveća tačnost predviđanja bez značajnijeg gubitka razumljivosti naučenog znanja.

Za visok stepen razumljivosti je još potrebno zadovoljiti i određene psihološke zahteve [66], pre svega saglasnost sa postojećim znanjem.

Ovaj zahtev je zadovoljen izborom takvih metoda selekcije i učenja koji koriste samo attribute iz originalnog modela problema, za koje se pretpostavlja da su formulisani u skladu sa postojećim znanjem [29].

U propozicionom modelima predstavljanja znanja, koji se koriste u ovom radu, model problema formira čovek izborom početnog skupa atributa, koji treba da reprezentuje relevantne *postojeće* pojmove iz razmatranog područja [67].

6. ZAKLJUČCI

Osnovni cilj sprovedenog istraživanja je potvrda hipoteze da je istovremenom primenom više različitih metoda i tehnika korišćenja postojećih informacija iz prostora atributa moguće znatno poboljšati performanse sistema za induktivno učenje pravila kada je obučavajući skup oskudan.

Cilj je postignut kroz empirijsko istraživanje, koje je potvrdilo da se tačnost naučenog znanja na osnovu oskudnih obučavajućih skupova može značajno povećati pomoću više različitih metoda, bez značajnijeg gubitka razumljivosti.

Osnovni doprinosi ove disertacije su sledeći rezultati:

1. Eksperimentalno je potvrđeno da kombinovani algoritam formiranja ansambala, uz istovremene pseudoslučajne modifikacije prostora atributa, drastično smanjuje broj razmatranih atributa i istovremeno značajno povećava tačnost naučenog znanja, te je pogodan za induktivno učenje konjuktivnih pravila i u slučaju oskudnih obučavajućih skupova.
2. Eksperimentalno je potvrđeno da se heuristike za ocenu interakcija *parova* atributa u problemu učenja mogu uspešno primeniti u selekciji atributa i, kombinovanjem sa pseudoslučajnom selekcijom atributa, u praksi dodatno poboljšati performanse naučenog znanja.
3. Značajno je unapređen sopstveni sistem za otkrivanje znanja *Empiric* ugradnjom algoritma globalnog ocenjivanja hipoteza *Relief*, što je demonstrirano poboljšanjima rezultata učenja u slučaju izabranih oskudnih obučavajućih skupova za više različitih tipova atributa, uključujući. strukturne attribute.
4. Izvršeno je dodatno unapređenje sistema *Empiric* izmenom osnovnog usmerenog pretraživanja prostora hipoteza (*greedy search*), tako da koristi i ocenu interakcije parova atributa za upravljanje pretraživanjem.

Dalji rad na ovom području treba usmeriti na usavršavanje korišćenih i razvoj novih ugrađenih metoda zasnovanih na konceptu margine, koji se mogu se integrisati u postojeće sisteme za induktivno učenje koncepata.

7. LITERATURA

- [1] Li, M. and P. Vitányi P., "Theories of Learning", in *An Introduction to Kolmogorov Complexity and Its Applications*, Text and Monographs in Computer Science. Springer-Verlag, 1993.
- [2] Hutter M., "On the existence and convergence of Computable Universal priors", In Gavalda R., Jantke K.P, Takomoto E. (eds), *Proc. of 14th International Conference on Algorithmic Learning Theory*, 2842, pp. 298-312, Springer, 2003.
- [3] Forsyth R. (ed), *Machine Learning: Principles and technics*, London: Chapman and Hall, 1989.
- [4] Witten I.H., Frank E., *Data Mining: Practical machine learning tools and techniques*, 2nd edition, Morgan Kaufmann, San Francisco, 2005.
- [5] Michalski, R., Carbonell, J., Mitchell T. (Eds.), *Machine learning: An artificial intelligence approach (Vol. I)*, San Francisco, CA: Morgan Kaufmann, 1983.
- [6] Hart, A., "Machine induction as a form of knowledge acquisition in knowledge engineering", in Forsyth, R. (ed), *Machine Learning: Principles and techniques*, Chapman and Hall, London, 1989.
- [7] Kubat M., Bratko I., Michalski R.S., "A Review of Machine Learning methods", in Michalski, R.S., Bratko I., Kubat M. (eds), *Machine Learning and Data Mining: Methods and Applications*, New York, NY: John Wiley & Sons Ltd, 1996.
- [8] Haussler D., "Probably approximately correct learning", In *Proc. of the 8th National Conference on Artificial Intelligence*, pages 1101-1108, Morgan Kaufmann, 1990.
- [9] Milosavljević M., Buturović LJ., "Prepoznavanje i ekstrakcija obeležja genskih ekspresija", *Zbornik radova 50. Konferencije za ETRAN*, tom III, pp. 177-179, Beograd, 6-8. juna 2006.
- [10] Piatetsky-Shapiro G., Tamayo P., "Microarray data mining: facing the challenges", *SIGKDD Explorations*, Vol. 5 (2), pp. 1-5, 2003.
- [11] Mišković V., *Jedna klasa algoritama za induktivno učenje*, Magistarski rad, Elektrotehnički fakultet, Univerzitet u Beogradu, Beograd, maj 2002.
- [12] Saffari A., Guyon I., *Quick Start Guide for CLOP*, Institute for Computer Graphics and Vision, Graz University of Technology and Clopinet, 2006.
- [13] Gilad-Bachrach R., Navot A., Tishby N., "Margin based feature selection – theory and algorithms", In *Proc. 21st ICML*, 2004.
- [14] Gilad-Bachrach R., Navot A., Tishby N., "Large margin principles for feature selection", In I. Guyon, S. Gunn, M. Nikravesh and L. Zadeh (eds.), *Feature extraction, foundations and applications*, Springer-Verlag, 2006.
- [15] Breiman L., "Random Forests", *Machine Learning*, 45, pp. 5–32, 2001.
- [16] Breiman L., "RF/tools A Class of Two-eyed Algorithms", *SIAM Workshop*, may 2003.

- [17] Breiman L., Cutler A., "RFtools for Predicting and Understanding Data", *Interface '04 Workshop*, 2004.
- [18] Cherkassky V., Mulier F. M., *Learning from Data: Concepts, Theory, and Methods*, 2 edition, John Wiley - IEEE Press, 2007.
- [19] Buntine, W., *A Theory of Learning Classification Rules*, PhD thesis, School of Computing Science, University of Technology, Sydney, November 1992.
- [20] Domingo, C., *Concept Learning: Theory, Applications and Related Problems*, PhD thesis, Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya, November 1998.
- [21] Ambainis, A., "Probabilistic Inductive Inference: A Survey", ???, 1999.
- [22] Kaufman, K.A. and Michalski, R.S., "Learning in an Inconsistent World: Rule Selection in AQ18," *Reports of the Machine Learning and Inference Laboratory*, MLI 99-2, George Mason University, Fairfax, VA, May, 1999.
- [23] Kovačič M., *Stochastic Inductive Logic Programming*, PhD, Faculty of Electrical Engineering and Computer Science, University of Ljubljana, Ljubljana, 1994.
- [24] Cunningham S.J., Humphrey M., Witten I.H., "Understanding what machine learning produces, Part I: representations and their comprehensibility", *Working Paper 96/21*, Computer Science Department, University of Waikato, 1996.
- [25] Pfahringer B., *Practical Uses of the Minimum Description Length Principle in Inductive Learning*, fur Med. Kybernetik u. AI, Technische Universität Wien, Dissertation, 1995.
- [26] Mišković V., "Uticaj upotrebe strukturnih atributa na kompleksnost induktivno naučenih propozicionih koncepata (2)", *Zbornik radova 51. Konferencije za ETRAN*, Herceg Novi - Igalo, 4-8. juna 2007.
- [27] Kohavi R., "A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection", in *Proc. of International Joint Conference on Artificial Intelligence*, 1995.
- [28] Reich Y., Barai S.V., "Evaluating Machine Learning Models for Engineering Problems", in *Artificial Intelligence in Engineering*, 1999.
- [29] Golub T.R., Slonim D.K., Tamayo P., Huard C., Gaasenbeek M., Mesirov J.P., Coller H., Loh M.L., Downing J.R., Caligiuri M.A., Bloomfield C.D., Lander E.S., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring", *Science*, No.286, Vol. 531–537, 1999.
- [30] Weston J., Perez-Cruz F., Bousquet O., Chapelle O., Eliseef A., Scholkopf B., "Feature Selection and Transduction for Prediction of Molecular Bioactivity for Drug Design", *Bioninformatics*, 19 (6), 764-771, 2003.
- [31] Bekkerman, El-Yaniv R., Tishby N., Winter Y., "Distributional Word Clusters vs. Word Text Categorization", *Journal of Machine Learning Research*, 3, 1183-1208, 2003.
- [32] Kloesgen W., Zytrow J., "Machine Discovery Terminology", In Fayyad U.M., Uthurusamy R. (eds), *Knowledge Discovery in Databases: Papers from the*

- AAAI Workshop, pp 463-, Technical Report WS-94-03, American Association for Artificial Intelligence, Menlo Park, California, 1994.
[<http://www.kdnuggets.com/publications/index.html>]
- [33] Murphy K, "Learning Bayes net structure from sparse data sets", *Technical report*, Comp. Sci. Div., UC Berkeley, 2001.
 - [34] Barbara D., DuMouchel W., Faloutsos C., Haas P. J., Hellerstein J. M., Ioannidis Y., Jagadish H.V, Johnson T., Ng R., Poosala V., Ross K. A., Sevcik K. C., "TheNewJerseyDataReduction Report", Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, Vol. 20, No. 4, December 1997.
 - [35] Jorge A., Brazdil P. B., , KDD 2000, Boston, MA USA
 - [36] Zaki, M.J., "Generating non-redundant association rules", In *Proceedings of the Sixth ACM-SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY: ACM, 34-43, 2000.
 - [37] Milosavljević M., Buturović LJ., "Analiza jedne klase metoda za diskriminatornu selekciju genskih ekspresija", Zbornik radova 51. Konferencije za ETRAN, Herceg Novi – Igalo, 4-8. juna 2007.
 - [38] Kohavi R., John G. H., "Wrappers for Feature Subset Selection", AIJ special issue on relevance, May 1997.
 - [39] Guyon I., Gunn S., Nikravesh M., Zadeh L. (editors), *Feature extraction, foundations and applications*, Springer, 2006.
 - [40] Schaffer C., "Sparse Data and the Effect of Overfitting Avoidance in Decision Tree Induction", *National Conference on Artificial Intelligence*, pp. 147-152, 1992.
 - [41] Jain A. K., Dubes R. C., *Algorithms for Clustering Data*, Englewood Cliffs: Prentice Hall, 1988.
 - [42] Guyon I., Elisseeff A., "An Introduction to Variable and Feature Selection", *Journal of Machine Learning Research*, 3, pp. 1157-1182, 2003.
 - [43] Jakulin A., Bratko I., "Analyzing attribute dependencies", In *Proc. of PKDD*, 2003.
 - [44] Jakulin A., Bratko I., "Testing the significance of attribute interactions ", In *Proc. of ICML*, 2004.
 - [45] Jakulin A., *Learning Based on Attribute Interactions*, PhD thesis, University of Ljubljana, 2005.
 - [46] Zagoruiko N.G, Kutnenko O.A., Borisova I.A., "Vybor informativnogo podprostranstva priznakov (Algoritm GRAD)", *Doklady 12-y vserosiyskoy konferencii Matematicheskie metody raspoznavania obrazov*, pp. 106-109, Moskva, 2005.
 - [47] Alves A., Zagoruiko N., Okun O., Kutnenko O., Borisova I., "Predictive analysis of gene expression data from human SAGE libraries", *Proc. of Workshop on Discovery Challenge (in conjunction with the 16th ECML and 9th PKDD)*, pp. 60-71, Porto, Portugal, 2005.

- [48] Breiman L., Friedman J.H., Olshen R.A., Stone C.J., *Classification and Regression Trees*, Wadsworth, Belmont, 1984.
- [49] Quinlan, J. R., "Induction of decision trees", *Machine Learning*, Volume 1, pages 81-106, 1986.
- [50] Kononenko, I., "Estimating attributes: Analysis and extensions of RELIEF", in *De Raedt, L. and Bergadano, F., editors, Proc. European Conf. on Machine Learning*, pp. 171-182, Catania, Italy, SpringerVerlag, 1994.
- [51] Almuallim H., Dietterich T.G., "Learning with many irrelevant features", *Proceedings of the 9th National Conference on Artificial Intelligence (AAAI-91)*, pp. 547-552, 1991.
- [52] Kira K., Rendell L. A., "A Practical Approach to Feature Selection", *Proceedings of 9th International Workshop on Machine Learning*, 249-256, 1992.
- [53] Ali K., Brunk C., Pazzani M., "On Learning Multiple Descriptions of a Concept", in *Proc. on Sixth International Conference on Tools with Artificial Intelligence*, IEEE Computer Society Press, pp. 476-483, 1994.
- [54] Domingos P., "Bayesian Model Averaging In Rule Induction", *Preliminary Papers of the Sixth International Workshop on Artificial Intelligence and Statistics*, (pp. 157-164), Ft. Lauderdale, FL: Society for Artificial Intelligence and Statistics, 1997.
- [55] Breiman, L., "Bagging predictors", Technical report No. 421, Department of statistics, University of California, Berkeley, CA 94720, September 1994.
- [56] Freund, Y., "Boosting a weak learning algorithm by majority", *Information and Computation*, Vol. 121, No. 2, pp. 256-285, 1995.
- [57] Freund, Y., Schapire, R.E., "A decision-theoretic generalization of on-line learning and an application to boosting", in *Proc. of the Second European Conference on Computational learning Theory*, March 1995.
- [58] Domingos P., "Knowledge Discovery Via Multiple Models", *Intelligent Data Analysis*, 2(3), 1998.
- [59] Kukar M., "Multistrategy attribute learning", in *Proc. of 3rd electrotechnical and computer science conference ERK '94*, Portoroz, Slovenia , 1994. (in Slovene)
- [60] Opitz D., Maclin R., "Popular Ensemble Methods: An Empirical Study", *Journal of Artificial Intelligence Research*, Vol. 11, pp. 169-198, 1999.
- [61] Ali, K., Pazzani M., "Error Reduction through Learning Multiple Descriptions", *Machine Learning*, 24, pp. 173-202, 1996.
- [62] Quinlan J. R., "Bagging, Boosting and C4.5", in *Proc. of AAAI-96 Fourteenth national Conference on Artificial Intelligence*, Portland, OR, AAAI Press, Menlo Park, CA, 1996.
- [63] Hansen L. K., Salamon P., "Neural network ensembles", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10):993-1001, October 1990.
- [64] Dietterich T., "Ensemble Methods in Machine Learning", *Lecture Notes in Computer Science*, Vol. 1857, pp. 1-15, 2000.

- [65] Murphy P. M., Pazzani M. J., "Exploring the Decision Forest: An Empirical Investigation of Occam's Razor in Decision Tree Induction", *Journal of Artificial Intelligence Research*, Volume 1, pp. 257-275, 1994.
- [66] Pazzani, M. J., "Knowledge discovery from data?", *IEEE Intelligent Systems*, pp 10-13, March-April 2000.
- [67] Pazzani, M., Mani, S., Shankle, W. R., "Beyond concise and colorful: learning intelligible rules", *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, Newport Beach, CA. AAAI Press, 235-238, 1997.
- [68] Pazzani, M., Mani, S. Shankle, W. R., "Comprehensible knowledge-discovery in databases", In M. G. Shafto and P. Langley (Ed.), *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society*, pp. 596-601. Lawrence Erlbaum, 1997.
- [69] Schaffer, C., "A Conservation Law for Generalization Performance", in *Proceedings of the Twelfth International Conference on Machine Learning*, pp. 259-265, New Brunswick, NJ: Morgan Kaufmann, 1994.
- [70] John G.H., Kohavi R., Pfleger K., "Irrelevant features and the subset selection problem", In *Proceedings of the Eleventh International Conference on Machine learning*, pages 121–129, New Brunswick, NJ, 1994. Morgan Kaufmann.
- [71] Yu L, Liu H, "Efficient Feature Selection Via Analysis of Relevance and Redundancy", *Journal of Machine Learning Research*, 5 (Oct), pp. 1205–1224, 2004.
- [72] Ho T.K., "Random Decision Forests", *Proc. of the 3rd International Conference on Document Analysis and Recognition*, pp. 278-282, Montreal, Canada, 1995.
- [73] Ho T.K., "The Random Subspace Method for Constructing Decision Forests", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20 (8), pp. 832-844, 1998.
- [74] Fradkin, D., D. Madigan, "Experiments with random projections for machine learning", In L. Getoor, T. E. Senator, P. Domingos, and C. Faloutsos, editors, *Proceedings of the Ninth International Conference on Knowledge Discovery and Data Mining*, pp. 517–522, New York: ACM, Washington DC, 2003.
- [75] LeCun Y., Denker J., Solla S., Howard R. E., Jackel L. D., "Optimal brain damage", In D. S. Touretzky, editor, *Advances in Neural Information Processing Systems II*, San Mateo, CA, Morgan Kauffman, 1990.
- [76] Yu L., Liu H., "Feature selection for high dimensional data: a fast correlation-based filter solution", In *ICML*, 2003.
- [77] Zhao Z., Liu H., "Searching for Interacting Features", *Proc. IJCAI07*, pp. 1156-1161, 2007.
- [78] Thrun S.D. et al., "The MONK's Problems: A Performance Comparison of Different Learning Algorithms", *Technical Report CS-CMU-91-197*, Carnegie Mellon University, December 1991.

- [79] Kononenko I., "On Biases in Estimating Multi-Valued Attributes", in *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pp. 1034-1040, 1995.
- [80] Bahler D., Navarro L., *Combining heterogeneous sets of classifiers: Theoretical and experimental comparison of methods*, 2000.
- [81] Kononenko I., Bratko I., Roškar E., "Sistem za induktivno učenje Asistent", *Informatika*, vol 10, No 1, pp 43-52, 1986.
- [82] Kononenko I., Šimec E., "Induction of decision trees using RELIEFF", in Kruse R., Viertl R., Della Riccia G. (eds): *CISM Lecture Notes*, Springer Verlag, 1995.
- [83] Clark, P., Niblett, T., "The CN2 induction algorithm", *Machine Learning*, 3, 261-284, 1989.
- [84] Quinlan J.R., Cameron-Jones R.M., "Induction of Logic Programs: FOIL and Related Systems", in *New Generation Computing*, Vol 13, pp. 287-312, 1995.
- [85] Ali K., Brunk C., Pazzani M., "On Learning Multiple Descriptions of a Concept", in *Proc. on Sixth International Conference on Tools with Artificial Intelligence*, IEEE Computer Society Press, pp. 476-483, 1994.
- [86] Ali K. M., Pazzani M. J., "Reducing the Small Disjuncts Problem by Learning Probabilistic Concept Descriptions", in T. Petsche, S. Judd and S. Hanson (ed.s), *Computational Learning Theory and Natural Learning Systems*, Vol. 3, pp. 179-195. MIT Press, 1995.
- [87] Michalski, R. S., Kaufman A., "Data Mining and Knowledge Discovery: A Review of Issues and Multistrategy Approach", in Michalski R. S., Bratko I. and Kubat M.(eds), *Machine Learning and Data Mining: Methods and Applications*, John Wiley & Sons, 1997.
- [88] Kononeko I., Šimec E., Robnik-Šikonja M., "Overcoming the Myopia of Inductive Learning Algorithms with RELIEFF", *Applied Intelligence*, 7(1), pp. 39-55, 1997.
- [89] Robnik-Sikonja M., Kononenko I., "Theoretical and empirical analysis of Relief and ReliefF", *Machine Learning*, 53:23–69, 2003.
- [90] Jones M. T., *Artificial Intelligence: A Systems Approach*, Infinity Science Press LLC, 2008.
- [91] Mišković V., "Sistem za induktivno učenje propozicionih pravila Empiric.Rules", *Zbornik radova INFOFEST 2001*, 143-150, Budva, septembar 2001.
- [92] Blake C.L., Merz C.J., UCI Repository of machine learning databases [<http://www.ics.uci.edu/~mllearn/MLRepository.html>], Irvine, CA: University of California, Department of Information and Computer Science, 1998.
- [93] Pathwork Diagnostics, <http://www.pathworkdx.com>, 2007.
- [94] Mišković V., "Princip razumljivosti u induktivnom učenju koncepata", *Zbornik radova XLV Konferencije ETRAN-a*, Sveska III, 123-126, Bukovička Banja, jun 2001.
- [95] Losee R. M., "A Discipline Independent Definition of Information", *Journal of the American Society for Information Science*, 48 (3), pp. 254-269, 1997.

- [96] Cunningham P., "Dimension Reduction", *Technical Report UCD-CSI-2007-7*, University College, Dublin, 2007.
- [97] Freund Y., Schapire R., "Experiments with a new boosting algorithm", in *Machine Learning: Proceedings of the Thirteenth International Conference*, pp 148-156, 1996.
- [98] Breiman L., Cutler A.,
www.stat.berkeley.edu/~breiman/RandomForests/cc_software.htm, 2004.
- [99] Hall M. A., "Correlation-based feature selection for discrete and numeric class machine learning", In *ICML*, 2000.
- [100] Ramón Díaz-Uriarte, and Sara Alvarez de Andrés: "Gene selection and classification of microarray data using random forest", *BMC Bioinformatics*, 7:3, 2006.

8. PRILOZI

Prilog 1: Uputstvo za korišćenje sistema *Empiric*

Prilog 2: Pregled upotrebljenog softvera za sistem *WEKA*

Prilog 3: Pregled upotrebljenog softvera za sistem *Matlab*

Prilog 4: Specifikacija upotrebljenih obučavajućih skupova

PRILOG 1: UPUTSTVO ZA KORIŠĆENJE SISTEMA *EMPIRIC*

Sistem *Empiric* je alat za inteligentnu analizu podataka (*data mining, knowledge discovery*), odnosno otkrivanje skrivenih zakonitosti u podacima koji se nalaze u bazama podataka.

U odnosu na poznate sisteme za otkrivanje znanja metodama induktivnog učenja *Empiric* je manji po obimu i sistemskim zahtevima. Realne probleme može da rešava i na prosečnim desktop računarskim konfiguracijama.

Sistem obezbeđuje osnovni skup alata za inteligentnu analizu podataka metodama induktivnog učenja. Sastoji se od više celina/podsistema:

1. *Editor primera* je namenjen za unos i ažuriranje modela problema i obučavajućih primera;
2. *Vizualizator* prikazuje strukturu obučavajućeg skupa geometrijskim rasporedom tačaka na površini. Neophodan za praćenje rada podsistema za induktivno učenje bez učitelja.
3. Podsystem za *induktivno učenje bez učitelja*, koji se sastoji od:
 - generatora jednostavnih klasifikacija (*partitionial clustering*);
 - generatora hijerarhijske klase (*hierarhycal clustering*);
4. Podsystema za *induktivno učenje pravila* na osnovu primera (*induction of conjunctive rules*);

Svi podsystemi su prilagođeni jedinstvenom modelu predstavljanja znanja, tzv. atributnom modelu sa *kontinualnim* i više tipova *diskretnih* atributa.

Diskretni atributi su ugrađeni u tri varijante: *nominalni* (neuređen skup vrednosti), *linearni* (uređen skup vrednosti) i *strukturni* (parcijalno uređen skup vrednosti).

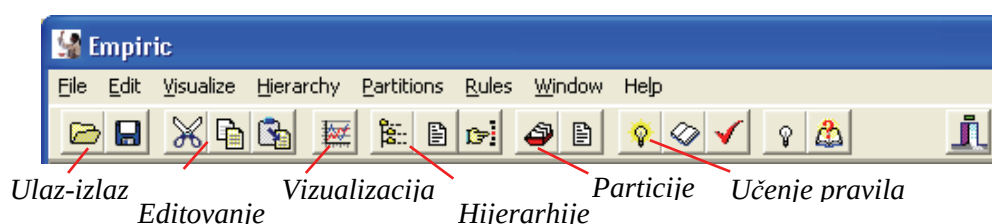
U sistemu *Empiric* meru kvaliteta pravila korisnik bira interaktivno, prilikom pokretanja algoritma učenja pravila, zajedno sa načinom izdvajanja i brojem primera za učenje, odnosno testiranje naučenih pravila.

U podsystem za induktivno učenje pravila sistema *Empiric* su ugrađene mere za ocenu kvaliteta pravila (hipoteze) [Miškovic,2002]:

- Informativnost atributa (*information gain*)
- Relativna informativnost atributa (*gain ratio*)
- Entropija (*entropy*)
- Gini indeks (*gini*)
- Logička pouzdanost (*logical sufficiency content*)
- Mera kvaliteta Kaufman-Michalski (*Q-measure*)
- Globalna mera Relief (*Relief, ReliefF*)

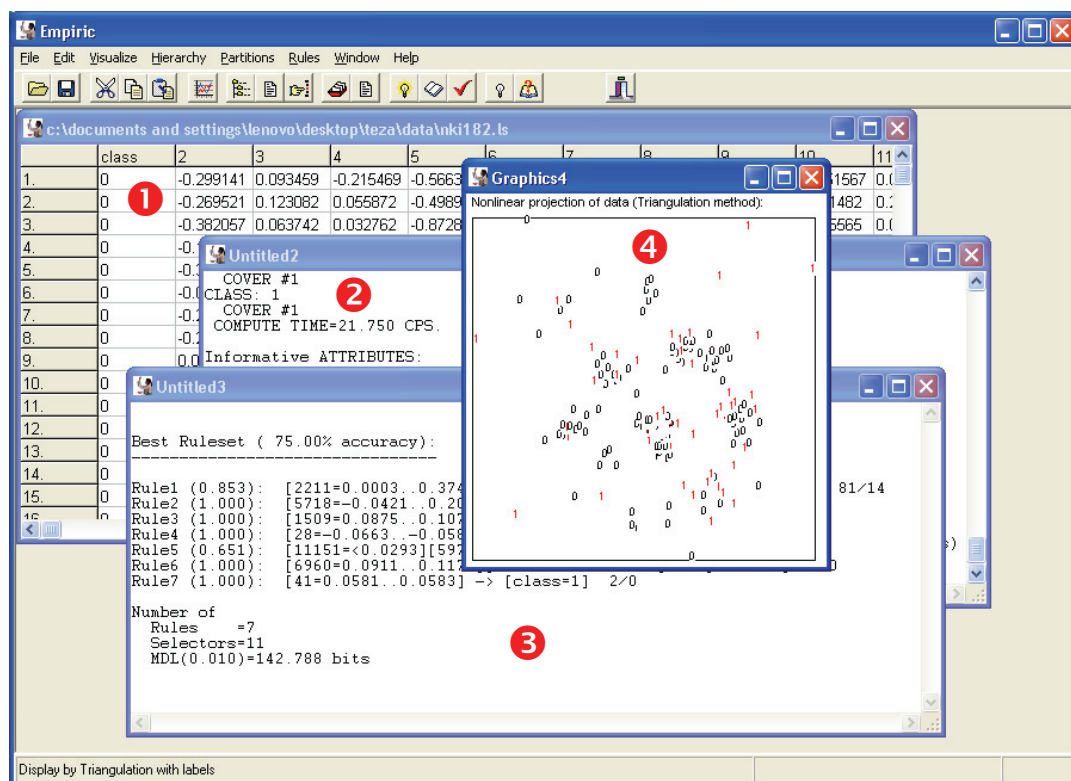
Svaka od ovih mera se može upotrebiti za generisanje pojedinačnih ili višestrukih modela, odnosno jednostavnih ili višestrukih skupova pravila.

Osnovni meni sistema *Empiric* je prikazan na Sl. 1.1.



Slika 1.1: Osnovni meni sistema *Empiric*

Na Sl. 1.2 se vidi kako podsistem za vizualizaciju prikazuje strukturu primera jednog problema NKI182, koji su razvrstani u dve klase (0 i 1).



Slika 1.2: Radni prozori sistema *Empiric* prilikom analize problema

Obučavajući primeri su u preglednoj tabeli ①, dok su u ostalim prozorima prikazani tok izvršavanja algoritma induktivnog učenja pravila ② i rezultati učenja, odnosno sama pravila ③.

Na Sl. 1.2 se u grafičkom prozoru ④ vidi kako podsistem za vizualizaciju prikazuje strukturu obučavajućih primera problema NKI BC (182 primera x 22.224 atributa).

Unos podataka se vrši izborom iz menija **File / Open** ili aktiviranjem tastera.

Priprema podataka

U demo verziji, priprema podataka problema induktivnog učenja vrši se izvan sistema, npr. u nekom editoru tekstualnih fajlova (*Notepad*, *WordPad*, *MS Word* ili neki od specijalizovanih tekst-editora, npr. *UltraEdit*). Opis problema i svi primeri se čuvaju u tekstualnoj datoteci, koja ima sufiks **.ls** (od *learning set*).

Podaci se pripremaju u skladu sa internim formatom, koji je ilustrovan na slici:

ukupan broj atributa	ukupan broj klasa	tip atributa (1-nominalni, 2-linearni, 3-kontinualni, 4-strukturni)	broj vrednosti nominalnog atributa	domen nominalnog atributa (nepoznata vrednost je ...)
22223	2	1	2	0 1
class	1	2	0	1
2	3	-1000000	1000000	1000001
3	3	-1000000	1000000	1000001
...				
22221	3	-1000000	1000000	1000001
22222	3	-1000000	1000000	1000001
22223	3	-1000000	1000000	1000001
22224	3	-1000000	1000000	1000001
0				
0	-0.299141	0.093459	-0.215469	-0.566347 -0.595818 -0.195499 ...
1	0.234519	-0.064074	0.160706	-0.242352 -0.122788 0.025056 ...
...				

Slika 1.3: Primer formata ulaznih podataka sistema Empiric (problem NKI BC)

Problem učenja se zadaje u dva dela:

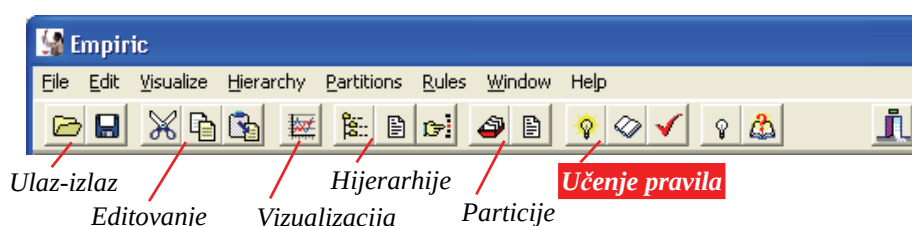
1. opisom modela problema definisanjem *naziva*, *tipa* i *domena* svih atributa i
2. navođenjem obučavajućih primera.

Za opis modela problema se u prvom redu fajla navodi broj atributa i broj klasa, nakon čega sledi opis svakog atributa, u jednom ili više redova. Prvi atribut je uvek nominalnog tipa i definiše naziv i moguće vrednosti *klasifikacije*. Ostali atributi mogu biti diskretnog tipa (oznaka 1,2,4) i kontinualni (oznaka 3).

Nakon toga se navode obučavajući primeri, koji predstavljaju nizove vrednosti atributa, u skladu sa tipom i redosledom koji je definisan u modelu problema.

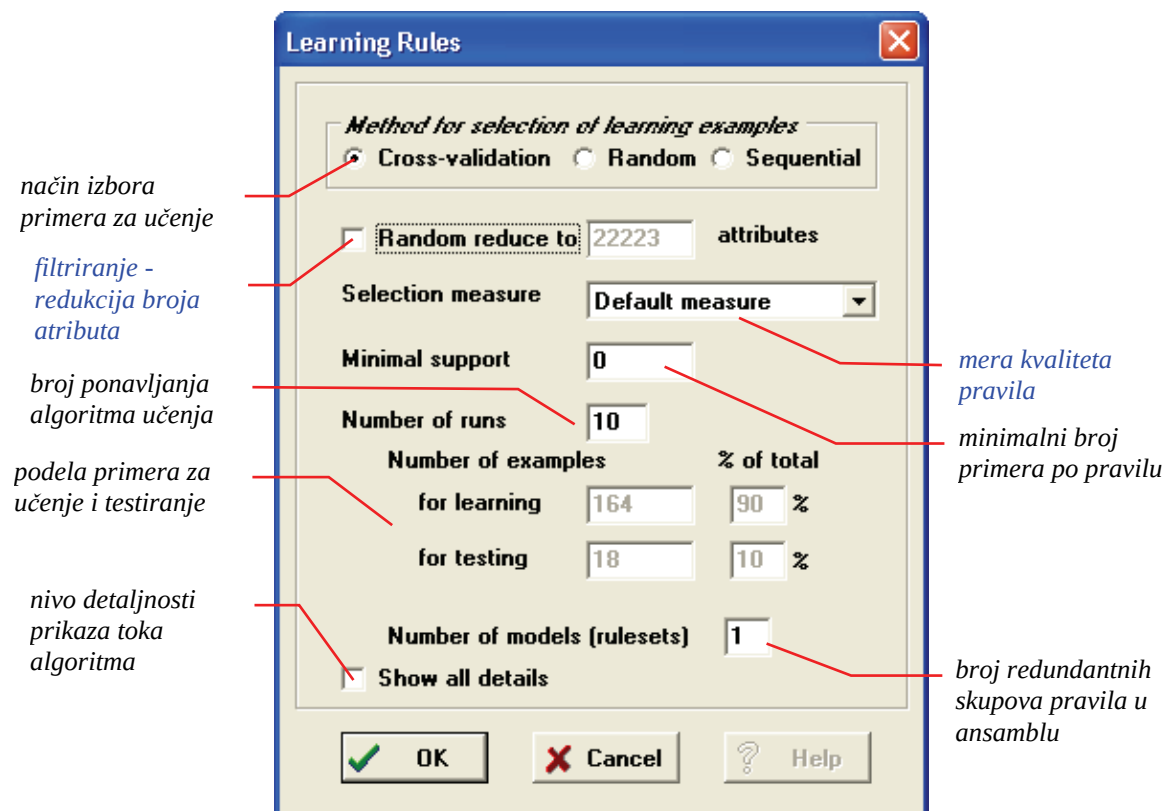
Induktivno učenje

Algoritam učenja pravila se pokreće iz osnovnog menija sistema **Rules** | **Learn** ili aktiviranjem tastera , Sl. 3:



Slika 1.4: Pokretanje induktivnog učenja pravila Empiric.Rules

Na Sl. 1.4 je prikazan izgled dijaloga u kome se zadaju svi bitni parametri učenja pravila. Učenje se pokreće tasterom **OK**, a rad programa se može pratiti u posebnom prozoru.



Slika 1.5: Postavljanje parametara induktivnog učenja pravila

Naučena pravila se predstavljaju kao produkciona pravila ili implikacije, uz koje se daje i ocena njihove pouzdanosti kroz tačnost na obučavajućem skupu i broj pozitivnih i negativnih primera iz kojih su izvedena, npr. kao na Slici 6.

```
Rule1(1,000):[216555_at=>0,0089][200020_at=0,7074..1,1471]->[Class=0] 55/0
```

Slika 1.6: Primer naučenog pravila

Istorijat sistema

Prvi prototip sistema za inteligentnu analizu podataka pod nazivom *Empiric* autor je razvio na Katedri za računarsku tehniku Visokih vojnotehničkih škola KoV JNA u Zagrebu 1989-1991. godine (verzija za DOS).

Nova verzija sistema je razvijena delom kao praktični deo izrade magistarske teze na Elektrotehničkom fakultetu u Beogradu 2002. godine. Osim novog interfejsa za rad u Windows okruženju, dograđena su bitna poboljšanja osnovnog algoritma induktivnog učenja pravila: upotreba više ocena kvaliteta pravila, upravljanje pojednostavljivanjem pravila i učenje višestrukih modela ili ansambala.

U ovom radu je sistem dograđen algoritmom selekcije atributa metodom slučajnih podprostora i globalnom ocenom *Relief*, koji je ugrađen u osnovni algoritam učenja i može se koristiti istovremeno sa ugrađenim algoritmom učenja ansambala.

PRILOG 2: PREGLED UPOTREBLJENOG SOFTVERA IZ SISTEMA WEKA

U radu je korišćena javno dostupna razvojna verzija sistema za istraživanje znanja WEKA 3.5.7 za operativni sistem Windows.

WEKA 3.5.7 (Windows)

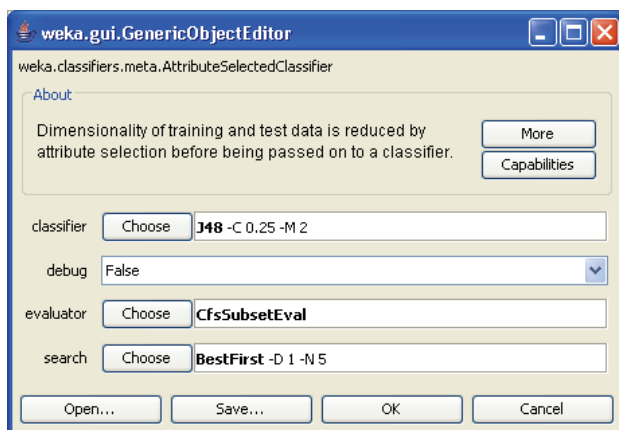
<http://www.cs.waikato.ac.nz/ml/weka/>

Upotrebljeni su algoritmi učenja razumljivog znanja:

- **J48** Java implementacija algoritma C4.5, jednog od najpoznatijih algoritama za induktivno učenje stabala odlučivanja;
- **PART** Java implementacija algoritma C4.5Rules, jednog od najpoznatijih algoritama za induktivno učenje produkcionih pravila;
- **JRip** Java implementacija algoritma RIPPER, poznatog i efikasnog algoritma za induktivno učenje produkcionih pravila;
- **RandomForest** specifična Java implementaciju metoda za induktivno učenje slučajnih šuma, ansambala koji se sastoje od stabala odlučivanja kao pojedinačnih klasifikatora;
- **Bagging** metod induktivnog učenja ansambala koji se sastoje od proizvoljnih klasifikatora;
- **AdaBoost.M1** metod induktivnog učenja ansambala koji se sastoje od proizvoljnih klasifikatora;
- **MultiBoost.AB** metod induktivnog učenja ansambala koji se sastoje od proizvoljnih klasifikatora;

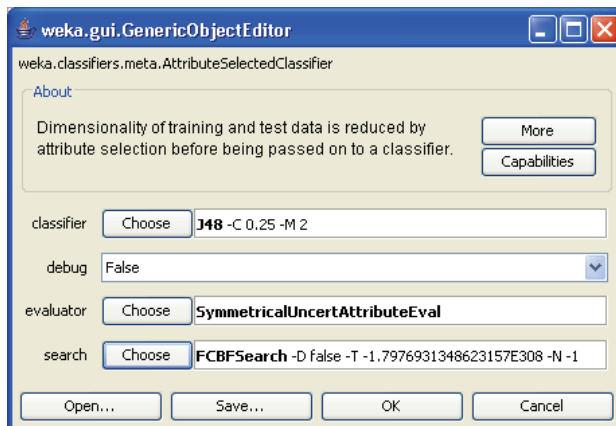
Selekcija atributa filtriranjem se u sistemu WEKA koristi uz pomoć posebnog objekta *AttributeSelectedClassifier*, koji služi za primenu algoritma učenja na filtrirani podskup atributa tekućeg obučavajućeg skupa, npr. slučajnog podskupa primera koji se generiše u jednom prolazu unakrsne validacije.

Parametri objekta od značaja su: osnovni algoritam učenja (*classifier*), algoritam pretraživanja prostora atributa (*search*) i algoritam ocene atributa (*evaluator*):

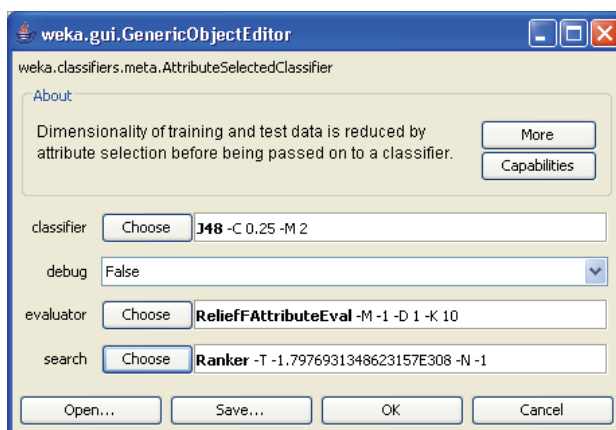


Upotrebljeni su algoritmi selekcije atributa:

- **FCBF** (*Fast Correlation Based Feature Selection*) se postavlja izborom algoritma evaluacije **SymmetricalUncertAttributeEval**, a kao parametar pretraživanja se postavi **FCBFSearch**.



- **Relief**, koji se postavlja izborom vrednosti parametra **ReliefAttributeEval**, a kao algoritam evaluacije se izabere **Ranker**.



Starija verzija sistema (WEKA 3.4) opisana je u knjizi:

Witten I.H., Frank E., *Data Mining: Practical machine learning tools and techniques*, 2nd edition, Morgan Kaufmann, San Francisco, 2005.

Novija verzija (WEKA 3.5) se može preuzeti sa adrese:

<http://prdownloads.sourceforge.net/weka/weka-3-5-7jre.exe>

INTERACT

Dodatak sistemu WEKA je algoritam selekcije atributa *INTERACT*, opisan na adresi <http://www.public.asu.edu/~huanliu/INTERACT/INTERACTsoftware.html>

Algoritam *INTERACT* rangira i selektuje relevantne attribute na osnovu njihovih međusobnih interakcija. Eksperimentalno je pokazano da znatno smanjuje broj atributa i održava ili poboljšava tačnost predviđanja kada postoje interakcije atributa.

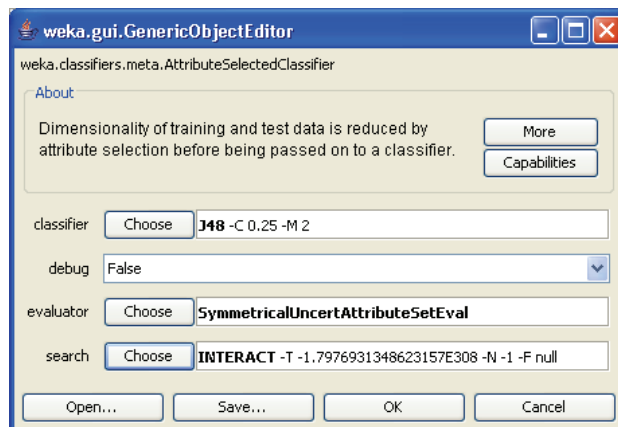
Napomena:

Složenost postupka selekcije optimalnog podskupa od N atributa je $O(2^N)$ [51]. Složenost algoritma *INTERACT* je u najgorem slučaju $O(N^2M)$, a u srednjem $O(NM(1-\alpha^{N+1})/(1-\alpha))$.

Za ubrzanje rada upotrebljena posebna *hash* tabela za izbegavanje ponovljnih prolaza kroz obučavajući skup.

Algoritam nije deo Java paketa *WEKA* i zahteva posebnu manuelnu instalaciju u objektni sistem *WEKA*.

Nakon instalacije, koristi se pomoću objekta *AttributeSelectedClassifier*, izborom algoritma evaluacije **SymmetricalUncertAttributeSetEval** i algoritma pretraživanja **INTERACT**.



Softver se može preuzeti sa adrese:

<http://www.public.asu.edu/%7Ehuanliu/INTERACT/INTERACT-java.zip>

PRILOG 3: PREGLED UPOTREBLJENOG SOFTVERA ZA SISTEM MATLAB

SIMBA

Simba je algoritam za prethodnu selekciju atributa koji maksimizuje funkciju ocene kvaliteta skupa atributa zasnovanu na konceptu margine [13], [14].

Softver se može preuzeti sa adrese:

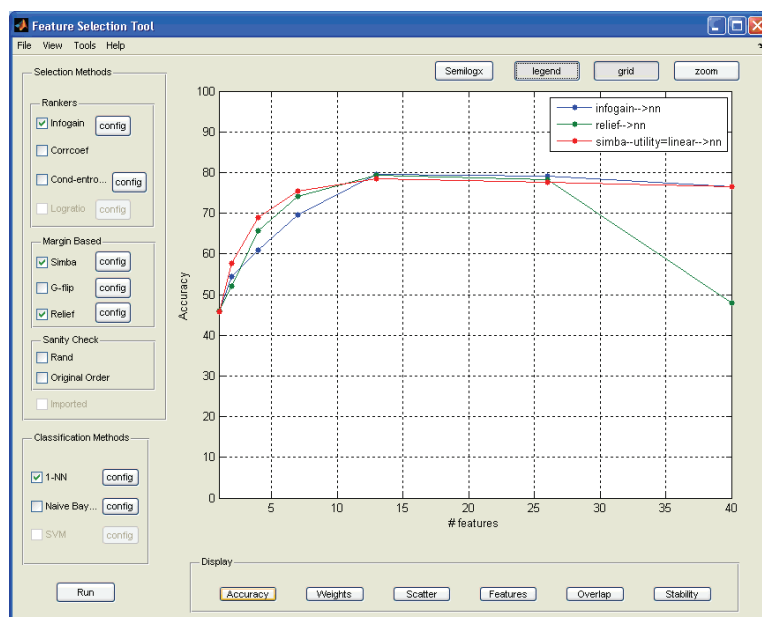
http://www.cs.huji.ac.il/labs/learning/Papers/Simba_04_11_2004.zip

FEATURE SELECTION TOOL

http://www.cs.huji.ac.il/~anavot/feature_selection_tool/fst.htm

Za korišćenje alata *Feature Selection Tool* potrebna je verzija *Matlab* 7.0 ili novija. Nije potrebna posebna instalacija, već je dovoljno uneti putanju foldera na koji je izdvojen kod u listu putanja sistema *Matlab*.

Za pokretanje programa je dovoljno u komandnu liniju uneti naziv **feature_selection_tool**, nakon čega se pojavljuje grafički interfejs:



U meniju **Help/Getting Started** se nalazi uputstvo za upotrebu alata, u kome se nalazi uputstvo za pripremu podataka, objašnjenje mogućnosti i uputstvo za upotrebu sistema.

Priprema podataka za *Feature Selection Tool* se može izvršiti konverzijom tekstualnih fajlova u format **.mat** na sledeći način:

1. Izvrši se prekodiranje nenumeričkih atributa i primera u celobrojne ekvivalente.
2. Ukloni se tekst definicije atributa i ostavi samo matrica vrednosti primera. Atribut koji definiše klasifikaciju je na prvom ili poslednjem mestu.

3. Pokrene se sistem *Matlab* i pripreme se matrica vrednosti X i matrica klasifikacije Y pomoću procedure:

a) ako je klasa je *poslednji* atribut u nizu
`traindata = textread('BC175.txt');`
`X = traindata(:,1:(end-1));`
`Y = traindata(:,end);`

b) ako je klasa je *prvi* atribut u nizu
`traindata = textread('RFS5.txt');`
`X = traindata(:,2:end);`
`Y = traindata(:,1);`

4. Dobiju se matrice X i Y, koje se sačuvaju na ***.mat** fajlu (u *Workspace* prozoru sistema *Matlab* se izvrši selekcija obe matrice, pa se sačuvaju funkcijom **Save/Save As**).

Softver se može preuzeti sa adrese:

http://www.cs.huji.ac.il/~anavot/feature_selection_tool/fst_download_page.html

PRILOG 4: SPECIFIKACIJA UPOTREBLJENIH OBUČAVAJUĆIH SKUPOVA

(A) Primeri veštački kreiranih problema učenja:

1. PROBLEM MONKS-2

Problem *Monks-2* [78] je problem klasifikacije objekata (osoba) u vidokrugu zamišljenog robota na osnovu 6 nominalnih atributa, definisanih kao:

Head-shape	∈ {round, square, octagon}
Body-shape	∈ {round, square, octagon}
Is-smiling	∈ {yes, no}
Holding	∈ {sword, balloon, flag}
Jacket-colour	∈ {red, yellow, green, blue}
Has-tie	∈ {yes, no}

Ukupan broj mogućih primera je $3 \cdot 3 \cdot 2 \cdot 3 \cdot 4 \cdot 2 = 432$. Problem je konstruisan radi poređenja različitih algoritama učenja, a upotrebljena verzija Monks-2 opisuje koncept "tačno dva od šest atributa imaju *prvu* od mogućih vrednosti", koji se ne može kompaktno opisati propozicionim logičkim izrazima (CNF ili DNF). Tačan propozicioni opis koncepta se sastoji od disjunkcije svih konjuktivnih izraza u kojima se *dva* atributa pojavljuju sa svojim prvim vrednostima, a *svi ostali* atributi bez prve vrednosti iz domena.

2. PROBLEM WAVEFORM-40

Problem *Waveform-40* je konstruisan radi provere funkcionisanja algoritama učenja u prisustvu šuma, odnosno *irrelevantnih* atributa.

Skup podataka ima 21 relevantan atribut, dobijen kombinovanjem dva od tri osnovna signala (funkcije) sa dodatkom nezavisnog šuma i dodatnih 19 irrelevantnih atributa, čije vrednosti su generisane potpuno slučajno.

3. PROBLEM BREIMAN 1000

Problem je konstruisan kao primer problema učenja sa veoma velikim brojevima slaborelevantnih atributa, kao što su problemi, npr. u medicinskoj dijagnostici i pretraživanju dokumenata. Kod nekih od ovih problema se ne može izdvojiti mala grupa atributa koja dobro razdvaja klase, već se najveća tačnost predviđanja postiže tek kada se koriste svi atributi.

Ovakvi problemi su teški za uobičajene programe učenja, kao što su neuronske mreže i stabla klasifikacije.

Problem je opisan sa 1.000 binarnih atributa, a primeri su razvrstani u 10 klasa.

Generisan je programom na slici (*random* je generator slučajnih brojeva po uniformnoj distribuciji):

```

do j=1,10
  do k=1,1000
    p(j,k)=.2*random+.01
  end do
end do
do j=1,10
  do i=1, nint(400*random) !nint=nearest integer
    k=nint(1000*random)
    p(j,k)=p(j,k)+.4*random
  end do
end do
do n=1,N
  j=nint(10*random)
  do m=1,1000
    if (rnd<p(j,m) )then
      x(m,n)=1
    else
      x(m,n)=0
    end if
    y(n)=j ! y(n) is the class label of the nth example
  end do
end do

```

Program generiše skup verovatnoća $\{ p(j,m) \}$, gde je j oznaka klase, a m redni broj atributa. Vrednost primera klase j je string od M binarnih promenljivih, gde m -ta promenljiva ima verovatnoću $p(j,m)$ da je jednaka jedinici.

Za generisanje obučavajućeg skupa u ovom radu je generisan skup od $N=1.000$ primera, koji je upotrebljen za učenje i estimaciju tačnosti predviđanja metodom unakrsne validacije (nije posebno generisan testni skup od 4.000 primera, kao u [15], jer se ne koristi estimacija tačnosti metodom testnog skupa).

Iz koda je vidljivo da svaka klasa ima visoku verovatnoću na određenim mestima. Ali ovih mesta za sve klase ukupno ima oko 2.000, tako da postoji značajno preklapanje. Uz pretpostavku da su poznate sve $\{ p(j,k) \}$, Bayesova greška izračunata za konkretnu $\{ p(j,k) \}$ je 1.0%.

(B) Primeri oskudnih obučavajućih skupova iz baze podataka UCI

4. BREAST CANCER LJ

Problem je učenje pravila za klasifikaciju obolelih od raka dojke, na osnovu opštih podataka o pacijentu i kliničkih nalaza, u dve klase: one za koje postoji mogućnost ponovnog pojavljivanja tumora dojke nakon operacije i one za koje ne postoji.

Podaci su formirani u Institutu za onkologiju Univerzitetskog medicinskog centra u Ljubljani [92].

5. CLEVELAND HART DISEASE

Problem je učenje pravila za dijagnostiku oboljenja srca na osnovu opštih podataka o pacijentu i kliničkih nalaza. Slučajevi se klasifikuju kao zdravi ili u jednu od 4 kategorije oboljenja srca.

6. LUNG CANCER

Problem je učenje pravila klasifikacije slučajeva raka pluća u jednu od tri kategorije, na osnovu medicinskih podataka koje koriste u praksi i lekari specijalisti.

(C) Primeri genskih ekspresija [Pathwork Diagnostics,2007]:

Sva tri problema analize genskih ekspresija se odnose na problem dijagnostike oboljenja raka dojke na osnovu genskih ekspresija.

Podaci su formirani uzimanjem uzoraka tkiva od zdravih i obolelih pacijenata, njihovom obradom i nanošenjem na odgovarajući DNK mikročip, koji ima više hiljada oligonukleotskih tačaka za predstavljanje gena [37].

Oligonukleotidska tačka je ekvivalent odgovarajućeg gena i ima sposobnost da veže za sebe posebnu informacionu iRNK. Nakon formiranja mikročipa, uzima se uzorak iRNK iz ćelije koja se proučava uz dodatak fluorescentne primese. Potom se mikročip izlaže delovanju fluorescentne iRNK.

Mikročip mikroskopski skenira i izmeri svetlosni intenzitet (ekspresija) svake tačke koja odgovara nekom genu.

7. BC (175 x 22215 x 2)

Primeri ekspresija gena 175 slučajeva zdravih i obolelih pacijenata, prikazanih pomoću skupa od 22.215 gena.

8. RFS5 (441 x 22215 x 2)

Primeri ekspresija gena za 441 slučaj potpuno izlečenih i ponovo obolelih od raka dojke u roku od 5 godina nakon terapije, prikazanih pomoću 22.215 gena.

9. NKI BC (182 x 22223 x 2)

Primeri ekspresija gena 182 slučajeva zdravih i obolelih pacijenata, prikazanih pomoću 22.223 gena, koji su formirani u holandskom nacionalnom institutu za rak (*Netherland Cancer Institute*).

Biografija

Mišković (Andrije) Vladislav rođen je 1. februara 1957. godine u Bačkoj Palanci, gde je završio osnovnu školu i gimnaziju prirodno-matematičkog smera 1976. godine. Diplomirao je 1981. godine na Tehničkoj vojnoj akademiji u Zagrebu u prvoj generaciji smera računarske tehnike sa temom iz matematičkih principa programiranja "Monitor strukture stack u jeziku Diktran".

Magistarsku tezu iz oblasti veštačke inteligencije pod nazivom "Jedna klasa algoritama za induktivno učenje" odbranio je 2002. godine na Elektrotehničkom fakultetu u Beogradu.

Nakon diplomiranja, u periodu 1981-1983. godine radi kao sistem inženjer u elektronskom računskom centru Tehničkog školskog centra kopnene vojske u Zagrebu na poslovima održavanja sistemskog softvera i uvođenju novih računarskih i softverskih sistema.

Izabran je za asistenta na Katedri za računarsku tehniku Visokih vojnotehničkih škola kopnene vojske u Zagrebu za predmet *Operativni sistemi* 1983. godine. Osim vežbi iz predmeta *Operativni sistemi* i *Računarsko upravljanje*, poverena su mu predavanja i vežbe iz predmeta *Modeliranje programskih sistema* i *Informatika sa računarskom tehnikom*.

Sa grupom nastavnika Katedre za računarsku tehniku čestvuje u uvođenju i realizaciji nastave predmeta *Tehnologija održavanja računarskih sistema* (deo za održavanje sistemskog softvera) i *Sistemi veštačke inteligencije* (deo za ekspertne sisteme i automatizovano učenje).

Osim nastave iz poverenih predmeta, sarađuje na izradi softvera u zvaničnim projektima usavršavanja sistema tehničkog održavanja kopnene vojske (projekt "Izrada normativa rezervnih delova, TU SSNO 1983/84), kao i u više internih softverskih projekata usmerenih na usavršavanje nastavnog procesa i opremanje računarskih laboratorija neophodnim softverom (posebno za nastavu iz predmeta *Sistemi veštačke inteligencije*, 1985-1989).

Godine 1990. izabran je za predavača na Katedri za računarsku tehniku (tada) Visokih vojnotehničkih škola kopnene vojske u Zagrebu za predmet *Operativni sistemi*.

Od 1992. godine radi u Tehničkoj upravi Generalštaba Vojske Jugoslavije, gde je zadužen za Tehnički informacioni sistem kopnene vojske. Uvodi u operativnu upotrebu grafički orijentisane mrežne operativne sisteme, nove razvojne alate i produktivnije aplikativne programe.

Godine 1994. prelazi u novoosnovani Centar za informatičku podršku pozadine, gde radi kao projektant na informatičkoj podršci i projektovanju novih komponenti informacionog sistema pozadine Vojske Jugoslavije.

Od 2002. godine je načelnik je Odseka za projektovanje u Odeljenju za informatičku podršku logistike Sektora logistike Generalštaba VJ, gde radi na modernizaciji računarske mreže Sektora i modernizaciji, razvoju i uvođenju informacionih sistema i novih tehnologija za potrebe logistike Vojske Jugoslavije.

Od 2003. godine sarađuje u nastavi na Fakultetu za poslovnu informatiku na predmetima *Multimedija*, *Projektovanje informacionih sistema* i *Poslovni informacioni sistemi*.

U periodu 2005.-2007. godine načelnik je Centra za Bolnički informacioni sistem Vojnomedicinske akademije u Beogradu. Rukovodi izradom projekta nove računarske mreže VMA, radom i modernizacijom Bolničkog informacionog sistema, kao i razvojem i uvođenjem novih softverskih podsistema.

Od januara 2008. godine je stalno zaposlen na Univerzitetu "Singidunum".

Osim osnovnog obrazovanja i usavršavanja u struci, pohađao je didaktičko-metodički kurs za nastavnike Visokih Vojnotehničkih škola kopnene vojske u Zagrebu, završio specijalističku obuku za održavanje računara, operativnih sistema i softvera firmi *Honeywell* i *Control Data Corporation*, kao i obuku za izradu elektronskih nastavnih materijala u sistemu za edukaciju uz pomoć računara *PLATO* firme *Control Data*.

Učesnik je seminara iz veštačke inteligencije koji su 1985. godine organizovali Institut "Jože Štefan" i Elektrotehnički fakultet u Ljubljani.

Od 2001. godine redovno učestvuje na konferencijama Društva za ETRAN radovima iz oblasti veštačke inteligencije.