Sustavi višestrukih klasifikatora temeljeni na upravljanom odabiru atributa u procjeni kreditnog rizika

Oreški, Goran

PODACI O DOKTORSKOM RADU

I. AUTOR

Ime i prezime	Goran Oreški
Datum i mjesto rođenja	19.10.1987. Bosanska Dubica
Naziv fakulteta i datum diplomiranja na VII/I stupnju	Fakultet organizacije i informatike, Varaždin, 2010.
Naziv fakulteta i datum diplomiranja na VII/II stupnju	-
Sadašnje zaposlenje	GO Studio d.o.o.

II. DOKTORSKI RAD

Naslov	Sustavi višestrukih klasifikatora temeljeni na upravljanom odabiru atributa u procjeni kreditnog rizika
Broj stranica, slika, tabela, priloga, bibliografskih podataka	146 stranica, 34 slike, 30 tablica, 2 priloga, 152 bibliografskih podataka
Znanstveno područje i polje iz kojeg je postignut doktorat znanosti	Društvene znanosti, informacijske i komunikacijske znanosti
Mentori ili voditelji rada	prof. dr. sc. Božidar Kliček
Fakultet na kojem je obranjen doktorski rad	Fakultet organizacije i informatike, Varaždin
Oznaka i redni broj rada	133

III. OCJENA I OBRANA

Datum sjednice Fakultetskog vijeća na kojoj je prihvaćena tema	16.09.2015.
Datum predaje rada	10.02.2016.
Datum sjednice Fakultetskog vijeća na kojoj je prihvaćena pozitivna ocjena rada	17.05.2016.
Sastav povjerenstva koje je rad ocijenilo	prof. dr. sc. Božidar Kliček izv. prof. dr. sc. Jasminka Dobša dr. sc. Dragan Gamberger
Datum obrane doktorskog rada	01.06.2016.
Sastav povjerenstva pred kojim je rad obranjen	prof. dr. sc. Božidar Kliček izv. prof. dr. sc. Jasminka Dobša dr. sc. Dragan Gamberger
Datum promocije	

Sveučilište u Zagrebu

FAKULTET ORGANIZACIJE I INFORMATIKE

GORAN OREŠKI

SUSTAVI VIŠESTRUKIH KLASIFIKATORA TEMELJENI NA UPRAVLJANOM ODABIRU ATRIBUTA U PROCJENI KREDITNOG RIZIKA

DOKTORSKI RAD

Varaždin, 2016.

University of Zagreb

FACULTY OF ORGANIZATION AND INFORMATICS

GORAN OREŠKI

MULTIPLE CLASSIFIER SYSTEMS BASED ON DIRECTED ATTRIBUTE SELECTION IN CREDIT RISK ASSESSMENT

DOCTORAL THESIS

Varaždin, 2016.

Sveučilište u Zagrebu

FAKULTET ORGANIZACIJE I INFORMATIKE

GORAN OREŠKI

SUSTAVI VIŠESTRUKIH KLASIFIKATORA TEMELJENI NA UPRAVLJANOM ODABIRU ATRIBUTA U PROCJENI KREDITNOG RIZIKA

DOKTORSKI RAD

Mentor:
Prof. dr. sc. Božidar Kliček

Varaždin, 2016.

University of Zagreb
FACULTY OF ORGANIZATION AND INFORMATICS

GORAN OREŠKI

MULTIPLE CLASSIFIER SYSTEMS BASED ON DIRECTED ATTRIBUTE SELECTION IN CREDIT RISK ASSESSMENT

DOCTORAL THESIS

Supervisor:
Prof. dr. sc. Božidar Kliček

Varaždin, 2016.

Mojoj obitelji.
(To my family.)

SAŽETAK

Kao nastavak prethodnih istraživanja autora, ova doktorska disertacija predstavlja sljedeći korak istraživanja problema klasifikacije kreditnog rizika. Utemeljena na opservaciji ponašanja koje intuitivno primjenjuje društvo u svakodnevnom životu, ideja kombiniranja glasova stručnjaka je dobila posebnu pozornost istraživačke zajednice na području klasifikacije podataka. Sve veći fokus istraživača ali i obećavajući pronalasci na području kombinacije klasifikatora usmjerili su interes autora prema tom području.

Svrha istraživanja provedenih i opisanih u ovom radu je istražiti primjenjivost sustava višestrukih klasifikatora temeljnog na odabiru atributa na problem procjene kreditnog rizika građana. U skladu sa svrhom provedeno je više istraživanja koja zajednički predstavljaju jedan kompleksni pristup odabranom problemu. Glavni cilj ovog rada jest razviti brzu, robusnu tehniku za kombiniranje klasifikatora koja će na temelju upravljanog odabira atributa stvarati efikasne i kvalitetne sustave za ocjenu sposobnosti tražitelja kredita da vrati kredit na vrijeme i u skladu s ugovorenim uvjetima. Povrh navedenog, nova tehnika mora biti dovoljno jednostavna za laku implementaciju i široku primjenu u istraživačkoj zajednici uključujući i istraživače koji primarno ne istražuju navedeno područje.

Dva glavna elementa nove tehnike su: (1) odabir atributa kao strategija za postizanje raznolikosti odluka klasifikatora i (2) smanjivanje sustava kao način uključivanja samo bitnih klasifikatora koji doprinose kvaliteti sustava. Odabir atributa počiva na korištenju nekoliko različitih brzih tehnika koje rangiraju atribute po kvaliteti. Prilikom odabira tehnika, kako bi se osigurao odabir različitih atributa, bitno je voditi računa o mjerama koje se koriste prilikom rangiranja atributa. Tako odabrani podskupovi atributa koriste se za trening klasifikatora, koji na temelju različitih ulaza produciraju različite modele. U sljedećem koraku tehnika odabire samo one modele koji kombinirani mogu pozitivno utjecati na performanse sustava, temeljem odluka novog, u radu predloženog pohlepnog algoritma. Uključivanje smanjivanja sustava pozitivno utječe na efikasnost sustava i kvalitetu odluke.

Nova tehnika je kreirana na kreditnim skupovima podataka s ciljem testiranja postavljenih hipoteza doktorske disertacije. U istraživanju se uspoređuju rezultati nove tehnike u odnosu

na rezultate pojedinačnih klasifikatora koji su uključeni u konačni sustav, da bi se utvrdila opravdanost kombiniranja klasifikatora. Povrh toga, analizirane su odluke algoritma za smanjivanje i način odabira klasifikatora u sustav te odnos točnosti i Q statistike na treniranim sustavima. U slijedećem krugu istraživanja, rezultati tehnike su vrednovani pomoću tehnika Bagging i Boosting. Rezultati su uspoređivani pomoću četiri različite mjere performansi: točnosti, greške tipa I, greške tipa II i AUC mjere. Osim odabranih mjera uspoređena su i vremena potrebna za treniranje i test klasifikacijskih modela pomoću odabranih tehnika.

Rezultati pokazuju da se korištenjem nove tehnike mogu poboljšati rezultati klasifikacije podataka u odnosu na pojedinačne klasifikatore uključene u sustav. Dodatno, rezultati su kvalitetom usporedivi s najpopularnijim tehnikama, štoviše tri od četiri odabrane mjere pokazuju superiornost nove tehnike. U skladu s ciljem konstruiranja, nova tehnika ostvaruje najbolje rezultate na sustavima s manjim brojem članova i vremenski nije zahtjevna u usporedbi s tehnikama Bagging i Boosting. Ostvareni rezultati su obećavajući a predložena tehnika predstavlja dobru alternativu postojećim tehnikama za konstruiranje sustava višestrukih klasifikatora.

Ključne riječi: klasifikacija podataka, odabir atributa, sustavi višestrukih klasifikatora, smanjivanje sustava višestrukih klasifikatora, kreditni rizici, mudrost gomile

ABSTRACT

Following the previous author's researches, this doctoral dissertation is the next step in credit risk classification research. Based on observations of behavior that can be found in nature and society, the idea of combining experts' decisions has gained significant importance in research community, especially in the area of data classification. Increasing focus of researchers as well as promising findings have directed author's interest to the mentioned research area.

The purpose of researches, conducted and elaborated in this dissertation is to investigate the application of multiple classifier systems based on attribute selection on credit risk assessment. In accordance with the purpose, several researches have been conducted, that jointly represent a complex approach to the selected problem. The main goal of this paper is to develop fast and robust technique for combining classifiers, based on directed attribute selection, which will be able to create efficient and accurate systems for credit risk assessment in retail. The aforementioned technique must be sufficiently simple for easy implementation and wide application by the research community, including researchers that are not primarily focused on this field.

Two key elements of the new technique are: (1) attribute selection used as strategy for training diverse classifiers and (2) ensemble thinning used to include only those classifiers that contribute to overall system quality. Attribute selection in this context refers to the implementation of several different fast techniques which rank attributes by their quality. In order to ensure selection of different attributes, it is necessary to consider techniques based on different evaluation criteria for attribute ranking. Subsets of attributes, selected in such manner, are used in training process of classifiers, thus ensuring difference in produced models. In the next step technique selects only those models which when combined together, positively contribute to performances of ensemble. The selection is conducted using new, in this paper proposed, greedy algorithm for ensemble thinning. Including ensemble thinning in new technique increases efficiency and quality of decisions.

The new technique has been tested on credit data sets in accordance with defined research hypothesis of this doctoral dissertation. In presented research the results obtained using new technique are compared to results of individual classifiers included in the final ensemble, in order to justify combining action. Additionally, decisions made by algorithm for ensemble thinning are analyzed as well as relationship between Q statistics and ensemble accuracy. In following research, the results of the new technique are evaluated by techniques Bagging and Boosting. Results are evaluated with four different performance measures: accuracy, error type I, error type II and AUC. Moreover, time necessary for training and testing of models are measured and compared in research.

Results show significant improvement of classification performance compared to individual classifiers as a direct result of the new technique. Furthermore, quality of obtained results can be compared with results of most popular techniques; moreover three out of four performance measures show superiority of the new technique. In accordance with the design, the new technique performs best on ensembles with small number of members and it is not time consuming compared to Bagging and Boosting.

Keywords: data classification, attribute selection, multiple classifier systems, ensemble thinning, credit risk, wisdom of the crowds

SADRŽAJ

POPIS SLIKA

POPIS TABLICA

POPIS KRATICA

SVK	sustav višestrukih klasifikatora
DFSE	sustav višestrukih klasifikatora s upravljanim odabirom atributa (engl. *Directed Feature Selection Ensemble*)
COP	smanjivanje sustava višestrukih klasifikatora temeljem mjere izgleda klasifikatora *(*engl. *Classification Odds based Pruning)*
CO	mjera izgled klasifikatora *(*engl. *Classification Odds)*
AUC	površina ispod krivulje *(*engl. *Area Under Curve)*
SVM	tehnika potpornih vektora *(*engl. *Support Vector Machine)*
GA-NN	algoritam za selekciju atributa i klasifikaciju (engl. *Genetic Algorithm with Neural Networks*)
HGA-NN	algoritam za selekciju atributa i klasifikaciju (engl. *Hybrid Genetic Algorithm with Neural Networks*)
ROC	krivulja (engl. *Receiver Operating Characteristic*)
MRMR	minimalna redundancija maksimalna važnost *(*engl. *Minimal-Redundancy-Maximal-Relevance)*
HRK	hrvatska kuna
CPU	središnja jedinica za obradu *(*engl. *Central Processing Unit)*
RAM	radna memorija *(*engl. *Random Access Memory)*
NN	neuronska mreža *(*engl. *Neural Network)*
SO	stabla odluke
PV	potporni vektori
LR	logistička regresija
ANOVA	analiza varijance *(*engl. *ANalysis Of VAriance)*
RI	Ripper
RE	reljefna tehnika
GI	Gini indeks
MN	mjera nesigurnosti
KO	korelacija
MARS	multivarijantna adaptivna regresija *(*engl. *Multivariate Adaptive Regression Splines)*
CV	unakrsna validacija *(*engl. *Cross Validation)*
TPR	osjetljivost *(*engl. *True Positive Rate)*
FPR	opadanje *(*engl. *False Positive Rate)*

1

Uvod

U uvodu doktorske disertacije je iznesen plan istraživanja koji uključuje detaljan opis ciljeva, postavljene hipoteze i metodologiju istraživanja.

1.1 Uvodna razmatranja

Suvremeni bankarski sustav kroz interakciju s klijentima i primjenu naprednih informacijskih sustava stvara i pohranjuje velike količine podataka potrebnih za pružanje financijskih usluga. Postoje podaci koji su iznimno bitni, frekventno korišteni i preduvjet u svakodnevnom poslovanju, ali i oni drugi koji se vrlo rijetko koriste kako u transakcijskim tako i u izvještajnim procesima. Veliki opseg prikupljenih podataka utječe na pojavu skrivenog znanja, koje postaje vidljivo primjenom tehnika dubinske analize podataka (engl. *data mining*). Obzirom na prethodno spomenutu količinu podataka ali i na njihovu dimenzionalnost prirodno se nameće potreba da se računala koriste u procesu otkrivanja znanja (Gamberger, 2011).

Spoznaja novih znanja može doprinijeti efikasnosti poslovanja, a posebno aktivnim bankarskim poslovima i procesu odobravanja kredita. Za razliku od pravnih osoba koje su zakonskim odredbama primorane podnositi financijska izvješća na godišnjoj razini, fizičke osobe nemaju takvih obveza stoga je predviđanje njihovog financijskog stanja veći izazov. Međutim, iako je prisutan nedostatak financijskih izvješća za fizičke osobe, banke na dnevnoj razini, kroz različite kanale, prikupljaju podatke koji klijenta mogu jednako dobro opisati. Bilo da je klijent u poslovnici napravio transakciju, provukao kreditnu karticu ili koristio Internet bankarstvo, banka je zabilježila trag njegove aktivnosti. Osim same aktivnosti klijenta, bilježi se i prati mjesto, vrijeme te ostali podaci koji jednako mogu skrivati znanje o predmetnom klijentu. Svi prikupljeni podaci čine skup na temelju kojeg se može predviđati kreditna sposobnost klijenta u budućnosti s određenom pouzdanosti, točnosti.

Primjenom metoda strojnog učenja (engl. *machine learning*), koje su dio računalnog područja umjetne inteligencije (engl. *artificial intelligence*), u svrhu dubinske analize podataka, unutar prikupljenih podataka mogu se prepoznati uzorci koji odvajaju dobre klijente od loših s aspekta kreditne sposobnosti. Istraživači diljem svijeta razvijaju različite modele kako bi odgovorili na potrebe financijskih institucija i pronašli dobar način predviđanja kreditne rizičnosti klijenata iz dostupnih podataka (Abdou et al., 2008; Khashman, 2010). Korištenje takvih modela nije ograničeno samo na primjenu u financijskom sektoru već može biti korisno i u mnogim drugim realnim domenama (Ansari et al., 2013; Belciug i Gorunescu, 2013). U ovoj doktorskoj disertaciji će se istražiti opravdanost korištenja sustava višestrukih klasifikatora (engl. *Multiple Classifier Systems*) temeljenih na upravljanom odabiru atributa u procjeni kreditnog rizika. Istraživanje će povezati teoriju mudrosti gomile (Surowiecki, 2004)

s metodama strojnog učenja temeljenih na odabiru onih atributa koji daju kvalitetnu osnovu za postizanje dobrih rezultata klasifikacije. U radu će biti predložen način kombiniranja pojedinačnih klasifikatora u sustav višestrukih klasifikatora koji poboljšava točnost klasifikacije pojedinačnih klasifikatora. Predloženo istraživanje predstavlja logički nastavak dosadašnjih istraživanja autora (Oreski at al., 2012; Oreski i Oreski, 2014).

1.2 Motivacija

Financijska kriza, koja je započela 2007. godine, i pojačani oprez prema riziku koji se javio u bankama bili su primarni motivi ulaska u istraživanje primjene umjetne inteligencije u bankarstvu (Oreski at al., 2012). U skladu s inicijalnom motivacijom, cilj ranijih istraživanja je bio kreiranje vlastite hibridne tehnike zasnovane na neuronskim mrežama i pametnom odabiru atributa kao način utvrđivanja do koje mjere podaci banaka mogu služiti u predviđanjima kreditne sposobnosti klijenata (Oreski at al., 2012). Kao rezultat istraživanja predstavljena je tehnika GA-NN koja je u slijedećem koraku poboljšana uključivanjem ekspertnih znanja i inkrementalne faze u genetski algoritam, te je stvorena HGA-NN tehnika (Oreski i Oreski, 2014). Model generiran tehnikom HGA-NN na „njemačkom skupu podataka“ je dao najveću točnost klasifikacije u dotada autorima, u najboljoj vjeri, poznatoj objavljenoj literaturi.

Uzimajući u obzir opisana istraživanja u kojima je autor sudjelovao i postignute zaključke kao i sve više objavljenih radova (Finlay, 2011; Wang i Ma, 2012; Wang et al., 2011) na tom području vezanih uz sustave višestrukih klasifikatora (u daljnjem tekstu SVK), koji opravdavaju kombiniranje klasifikatora s istim ciljem, daljnje istraživanje autora je usmjereno upravo u tom smjeru. SVK u osnovi se temelji na teoriji mudrosti gomile (Surowiecki, 2004) tj. na tvrdnji da su odluke donesene od strane grupe u prosjeku bolje od odluke koju može donijeti pojedinačno njezin jedan član. Autor Dietterich (2000.) daje matematički dokaz koji potvrđuje navedenu teoriju. Vjerojatnost da će SVK koji sadrži 21 hipotezu, od kojih svaka ima stopu pogreške od 0.3, pogrešno klasificirati neki primjer tj. vjerojatnost slučaja da će 11 i više hipoteza istodobno krivo klasificirati neku pojavu iznosi približno 0.026, što je značajnije manje od stope greške individualnih hipoteza. Osim navedenog rada postoji cijeli niz radova koji teorijski (Dietterich, 2000; Ranawana i Palade, 2006) i empirijski (Finlay, 2011; Maclin i Optiz, 2011) dokazuju opravdanost korištenja SVK.

Prilikom konstruiranja SVK postoji nekoliko pristupa (Dietterich, 2000; Ranawana i Palade, 2006). Jedan od pristupa je manipulacija ulaznih atributa koji zbog ranijih neuspjelih pokušaja konstruiranja efikasnih SVK nije doživio uspjeh u istraživačkoj zajednici. Problem se pojavio prilikom primjene pristupa na skupovima s relativno malom dimenzionalnosti te korištenjem slučajnog odabira atributa. Zaključak istraživanja je da je navedeni pristup neprikladan za probleme s malom dimenzionalnosti.

Motivacija za ovo istraživanje se temelji upravo na činjenici koja je dokazana u prethodnim istraživanjima autora (Oreski at al., 2012; Oreski i Oreski, 2014), da je odabir atributa ključan u procesu klasifikacije kreditnog rizika te da upravljani odabir atributa daje bolje rezultate klasifikacije od slučajnog odabira. Motivaciju za istraživanje predstavlja hipoteza da korištenje višestrukih filtarskih tehnika za odabir atributa unutar SVK-a može doprinijeti točnosti klasifikacije te potaknuti interes istraživača u taj pristup konstruiranja SVK-a.

1.3 Svrha i ciljevi istraživanja

Svrha doktorske disertacije je istražiti primjenjivost sustava višestrukih klasifikatora temeljnog na upravljanom odabiru atributa na problem procjene kreditnog rizika građana. Odabir atributa je zasnovan na tehnikama koje pridjeljuju rang atributima kao suprotnost pristupu slučajnog odabira. Odabirom brzih tehnika koje počivaju na deterministčkim algoritmima direktno se upravlja i utječe na odabir atributa koji se koriste u treningu klasifikatora, stoga se govori o upravljanom odabiru atributa. Kako se iz prijašnjih istraživanja autora može zaključiti da se izdvajanjem korisnih atributa iz skupa podataka može pozitivno utjecati na kvalitetu klasifikacije podataka, slijedeći korak je istražiti da li se kombiniranjem takvih klasifikatora dodatno može poboljšati rezultat. U istraživanju primjenjivosti SVK-a na problem procjene kreditnog rizika građana bit će prezentiran nov način kombiniranja klasifikatora, takav koji postiže najbolje performanse s manjim brojem članova i na taj način povećava efikasnost sustava. U tom kontekstu, iznimno je bitno odrediti optimalan broj klasifikatora, članova sustava i odabrati one koji su najkompatibilniji za spajanje.

U skladu s definiranom svrhom istraživanja koja spaja odabir atributa kao tehniku predprocesiranja podataka s novim pristupom kombiniranja klasifikatora postavljen je i glavni

cilj istraživanja: razviti brzu, robusnu tehniku za kombiniranje klasifikatora koja će na temelju upravljanog odabira atributa stvarati efikasne i kvalitetne sustave za ocjenu sposobnosti tražitelja kredita da vrati kredit u skladu s preuzetom ugovornom obvezom.

Opći cilj istraživanja predstavlja širi pogled na istraživanje koji je detaljnije razrađen kroz znanstvene ciljeve istraživanja:

- Razviti mjeru za kvantifikaciju korisnosti uključivanja novih klasifikatora u sustav višestrukih klasifikatora, temeljenu na vrednovanju težine primjera.
- Kreirati robusnu tehniku za konstruiranje sustava višestrukih klasifikatora temeljenu na upravljanom odabiru atributa, koja daje:
    a. bolje rezultate od pojedinačnih klasifikatora uključenih u sustav
    b. jednako dobre ili bolje rezultate klasifikacije kod procjene kreditnog rizika nego najpopularnije tehnike Bagging i Boosting.
- Utvrditi zavisnost između mjere raznolikosti iskazane Q statistikom i mjere točnosti.

Osim znanstvenih ciljeva istraživanja, očekuje se da će rad kroz rezultate istraživanja dati primjenjiv društveni doprinos. Razvoj učinkovitog SVK-a omogućio bi primjenu istog u poslovanju banaka što bi unaprijedilo korištenje tehnika dubinske analize podataka u bankarstvu. Prednosti koje donosi automatizacija odobravanja kredita mogu biti od velike koristi bankama ali i pojednostaviti i olakšati postupak pojedincima koji su tražitelji kredita. Obzirom da se radi o brzoj tehnici, kojom se nastoji smanjiti vrijeme potrebno za trening u odnosu na druge tehnike za konstruiranje sustava višestrukih klasifikatora, ušteda vremena se također može ubrojiti kao jedan od doprinosa.

1.4 Hipoteze istraživanja

U skladu s prethodno navedenim znanstvenim ciljevima istraživanja, definirane su slijedeće hipoteze istraživanja.

H1: Sustav višestrukih klasifikatora koji je temeljen na odabiru različitih podskupova atributa pomoću filtarskih tehnika te konstruiran na temelju u ovom radu predloženog algoritma za smanjivanje sustava će postizati statistički značajno veću točnost klasifikacije od pojedinačnih klasifikatora uključenih u sustav na razini statističke značajnosti $p \leq 0,05$.

H2: Sustav višestrukih klasifikatora koji je temeljen na odabiru različitih podskupova atributa pomoću filtarskih tehnika te konstruiran na temelju u ovom radu predloženog algoritma za smanjivanje sustava će postizati statistički jednake ili bolje rezultate u odnosu na najpopularnije tehnike, Bagging i Boosting primijenjene na originalnim skupovima podataka (njemačkom i hrvatskom) sa svim karakteristikama.

Hipoteza H1 se odnosi na prvi i djelomično na drugi znanstveni cilj istraživanja, a njezino prihvaćanje predstavlja potvrdu uspješne realizacije tih ciljeva. Da bi se hipoteza mogla prihvatiti potrebno je u prvom koraku stvoriti preduvjete za njezino razmatranje. Prvi preduvjet je konstrukcija algoritma za smanjivanje sustava višestrukih klasifikatora, s ciljem odstranjivanja nepotrebnih klasifikatora unutar sustava. Predloženi algoritam će se zasnivati na novom pristupu kombiniranja klasifikatora koji se temelji na vrednovanju težine primjera. To je novi koncept kojim se odabiru klasifikatori s najvećim potencijalom za kombiniranje odluka. U slijedećem koraku je potrebno konstruirati tehniku koja će iskoristiti potencijal novog algoritma zajedno s odabirom atributa i stvoriti sustav koji postiže bolje performanse od vlastitih članova. Hipoteza definira i statističke uvjete pod kojima će se prihvaćanje te hipoteze smatrati potvrdom kvalitete ostvarenih rezultata.

Hipoteza H2 se direktno odnosi na drugi cilj istraživanja, a njezino prihvaćanje interpretirat će se kao potvrda uspješne realizacije tog cilja. Nakon stvaranja uspješnog sustava višestrukih klasifikatora koji zadovoljava temeljni razlog kombiniranja, tj. postizanje boljih rezultata od pojedinih klasifikatora, potrebno je vrednovati postignute rezultate u odnosu na druge dostupne tehnike. U tu svrhu su korištene dvije najzastupljenije tehnike za kombiniranje klasifikatora, Bagging i Boosting. Da bi se postigla usporedba tehnika s njihovim punim potencijalom u hipotezi se naglašava da se prilikom treniranja tehnika odabranih za usporedbu neće provoditi odabir atributa. Ostvareni rezultati će se vrednovati pomoću točnosti klasifikacije.

Svako i najmanje poboljšanje konstruiranog sustava u odnosu na performanse pojedinačnih algoritama i tehnika korištenih za usporedbu koje se može statistički dokazati, smatrat će se zadovoljavajućim obzirom da je i najmanji postotak poboljšanja rezultata od velikog značaja bankarskim institucijama. Kroz istraživanja vezana za hipoteze H1 i H2 će se adresirati i treći cilj, a to je utvrđivanje odnosa između mjere raznolikosti i točnosti klasifikatora. Potencijalna veza može biti iznimno korisna prilikom konstruiranja novih sustava. Iako takva veza na drugim domenama na kojima su provedena istraživanja nije potvrđena, u ovoj disertaciji će se rasvijetliti njihov odnos na predmetnoj domeni.

1.5 Metodologija istraživanja

Doktorska disertacija se sastoji od više istraživanja. Prikazana su logičkim slijedom tako da prate ciljeve istraživanja te kao takva čine jednu cjelinu, premda se mogu promatrati i kao samostalne cjeline.

U disertaciji će se koristiti generirani i stvarni skupovi podataka iz domene kreditnog rizika. Razlog zašto istraživanje koristi generirane (sintetičke) skupove podataka jest proučavanje ponašanja predloženog algoritma za smanjivanje SVK-a ukoliko se neka od karakteristika npr. kao što je mjera raznolikosti mijenja. Provođenje takvih simulacija na stvarnim podacima je teško ili gotovo nemoguće, stoga je potrebna simulacijska rutina koja kroz parametre prima karakteristike klasifikatora, a na izlazu daje matricu odluka. U okviru istraživanja je poboljšan algoritam za generiranje teorijskih skupova podataka s unaprijed definiranim vrijednostima točnosti i raznolikosti, što daje znanstveni doprinos daljnjim teorijskim i praktičnim istraživanjima SVK-a.

Stvarni skupovi podataka podrazumijevaju hrvatski i njemački skup kreditnih podataka. Priprema hrvatskog skupa podataka, će se bazirati na već prikupljenom skupu podatka u jednoj hrvatskoj kreditnoj instituciji koji je korišten u ranijim istraživanjima (Oreski at al., 2012; Oreski i Oreski, 2014). Sirovi podaci koji su prikupljeni već su prethodno pročišćeni, tj. otklonjene su nekonzistentnosti i ekstremne vrijednosti u svrhu prethodnih istraživanja. Njemački skup podataka će biti preuzet iz nekog od svjetskih repozitorija za strojno učenje. U doktorskoj disertaciji će skupovi biti opisani deskriptivnom statistikom.

U radu će se kreirati nova robusna tehnika DFSE (engl. *Directed Feature Selection Ensemble*) za konstruiranje SVK-a kroz nekoliko koraka. Tehnika će se bazirati na filtarskim tehnikama za odabir atributa. U istraživanju će se koristiti pet različitih filtarskih tehnika: korelacija, Gini indeks, omjer informacijske dobiti, reljefna metoda i mjera nesigurnosti. U prvom koraku za svaku odabranu tehniku će se kreirati više klasifikacijskih modela, koji se razlikuju po broju odabranih atributa. Prilikom odabira klasifikatora testirat će se dva pristupa: homogeni i heterogeni. U homogenom sustavu, klasifikator će biti neuronska mreža s parametrima s kojim je u prijašnjim istraživanjima ostvarila najbolje rezultate, a u heterogenom koristit će se više algoritama koji će biti odabrani slučajnim odabirom. Na opisani način bit će trenirano N različitih klasifikatora.

Nakon treniranja, u drugom koraku konstrukcije slijedi analiza generiranih modela. Napravit će se analiza performansi klasifikacije pomoću mjere točnosti. Cilj analize je

isključiti one modele koji ne postignu dovoljnu razinu točnosti za sudjelovanje u SVK, ukoliko takvi postoje. Uvjet da neki model uđe u razmatranje za SVK jest da ima veću točnost od 50% (ne smije biti slučajni klasifikator) (Zhang i Ma, 2012). Za performanse sustava osim točnosti je bitna i mjera raznolikosti (neovisnosti) pojedinih modela unutar sustava. Da bi se utvrdila raznolikost koju je donijela strategija redukcije dimenzionalnosti računat će se mjera Q statistike (engl. *the Q statistics*) koja predstavlja jednu od uparenih mjera za izračun zavisnosti klasifikatora. U istraživanju će se pratiti mjera Q statistike i njezin odnos prema točnosti klasifikacije.

Nakon što su klasifikatori trenirani i analizirani, u trećem koraku će se izbaciti oni koji ne doprinose kvaliteti rezultata. Spomenuti korak u literaturi se naziva smanjivanje sustava višestrukih klasifikatora (engl. *ensemble pruning*) i jedan je od predmeta istraživanja u ovoj disertaciji. Smanjivanje SVK je važno iz dva razloga: efikasnosti i performansi predviđanja. U tu svrhu, u prvom istraživanju doktorske disertacije, će se razviti vlastiti algoritam za smanjivanje SVK koji će za metriku odabira koristiti novu mjeru pod nazivom koeficijent utjecaja. Novi algoritam treba biti brz i računalno ne zahtjevan te treba pružiti dobre rezultate u vidu efikasnosti i performansi predviđanja. Razlika u odnosu na prethodne algoritme dostupne u literaturi jest nova mjera koja efikasno bira smjer odabira rješenja.

Nakon konstrukcije tehnike DFSE ista će se primijeniti na odabranim skupovima kreditnih podataka. Evaluacija rezultata klasifikacije provodit će se unakrsnom validacijom s 10 preklapanja, koja predstavlja standard u sličnim istraživanjima. U slijedećim istraživanjima disertacije, prvo će se usporediti performanse sustava s pojedinačnim klasifikatorima uključenim u sustav. Cilj je istražiti da li je sustav donio poboljšanje u odnosu na pojedinačne članove. A potom, da bi se utvrdilo koliko su ti rezultati značajni, bit će uspoređeni s rezultatima dobivenim korištenjem popularnih tehnika Bagging i Boosting na hrvatskom i njemačkom skupu podataka. Postignuti rezultati će se također usporediti i s ostvarenim rezultatima iz dostupne literature.

Za statističku provjeru rezultata u istraživanjima će se koristiti parametarski (t-test, ANOVA) i neparametarski testovi (Wilcoxonov test uparenih parova, Friedmanov test).

1.6 Struktura rada

Ostatak rada organiziran je na sljedeći način. U sljedećem poglavlju dan je uvod u sustave višestrukih klasifikatora. Osim definicije naglašeni su i glavni faktori koji čine takav sustav uspješnim. Ukazano je na njihove prednosti koje opravdavaju kombiniranje klasifikatora kao i moguće strategije konstruiranja sustava. Cilj ovog poglavlja je dati opći uvod i definirati osnovne pojmove koji se koriste u istraživanjima koja slijede u disertaciji.

Poglavlje broj 3 opisuje istraživanje u kojem je konstruiran novi algoritam za smanjivanje sustava višestrukih klasifikatora. U okviru poglavlja je predstavljeno poboljšanje algoritma za generiranje sintetičkih skupova podataka, koji je korišten za stvaranje skupova podataka korištenih u istraživanju. Dan je opis novog algoritma COP (engl. *Classification Odds based Pruning*) za smanjivanje SVK-a i njegove jezgre, mjere CO (engl. *Classification Odds*). Provedeno je vrlo iscrpno istraživanje kvalitete novog algoritma na različitim skupovima podataka koji su generirani u rasponu vrijednosti točnosti i raznolikosti odluka kakve se mogu očekivati na stvarnim podacima.

U poglavlju 4 opisana je konstrukcija nove DFSE tehnike, na temelju odabira atributa i COP algoritma. Predstavljena je robusna tehnika, široko primjenjiva i zbog svoje jednostavnosti prikladna za korištenje u širokom krugu istraživanja na području strojnog učenja. Performanse DFSE tehnike su testirane na hrvatskom i njemačkom skupu podataka te su uspoređene s pojedinačnim klasifikatorima unutar sustava. Istražena su dva različita pristupa pri odabiru klasifikacijskih algoritama, tako da je na jednom skupu korišten homogeni, a na drugom heterogeni pristup. U sklopu istraživanja napravljena je i analiza odnosa Q statistike kao mjere raznolikosti i točnosti klasifikacije.

U poglavlju 5 su vrednovani rezultati DFSE tehnike u odnosu na rezultate postignute drugim dostupnim tehnikama za stvaranje sustava višestrukih klasifikatora. DFSE tehnika je uspoređena s tehnikama Bagging i Boosting na hrvatskom i njemačkom skupu podataka. Radi stvaranja šire slike kvalitete korištenih tehnika u istraživanju su pored točnosti mjerene performanse prema tri dodatne mjere: greška tipa I, greška tipa II i AUC. Dodatno, u okviru istraživanja su mjerena vremena potrebna za trening i testiranje modela svih tehnika uključenih u eksperiment, u svrhu utvrđivanja njihovih odnosa. Na posljetku su rezultati ostvareni DFSE tehnikom uspoređeni s onima objavljenima u literaturi.

U zaključku doktorske disertacije su sagledani rezultati istraživanja te realizacija postavljenih ciljeva.

2

Sustavi višestrukih klasifikatora

Radi pružanja zaokružene slike trenutnog stanja znanosti, u prvom poglavlju su definirani važni pojmovi i bitna istraživanja iz područja sustava višestrukih klasifikatora.

2.1 Uvod

Klasifikacija podataka predstavlja povezivanje ulaznih primjera s unaprijed predodređenim klasama (Murphy, 2012). Obzirom da na ulazu nikada nisu dostupni svi primjeri nekog problema, treningom se nastoji što bolje aproksimirati željena funkcija. Aproksimacija ciljne funkcije predstavlja temeljni koncept na kojem se zasniva paradigma strojnog učenja. Iz razloga što trening podaci najčešće sadrže šum i što se mogu koristiti različiti pristupi klasifikaciji, krajnji rezultat izrade klasifikatora za neki problem su često različite producirane funkcije (Ranawana i Palade, 2006). Na toj različitosti počiva ideja da se umjesto odabiranja jedne, ispravnim kombiniranjem više hipoteza mogu postići rezultati koji su u najlošijem slučaju podjednako dobri kao najbolji pojedinačni član.

Intuitivno gledajući, opisani pristup se koristi u svakodnevnom životu gdje nastojimo dobiti „drugo mišljenje“. Na primjer: prije odluke odlaska na medicinski zahvat tražimo mišljenje nekoliko liječnika, prilikom kupnje nekog proizvoda skloni smo čitanju većeg broja korisničkih ocjena tog proizvoda ili prilikom zapošljavanja novog djelatnika provjerit ćemo sve njegove reference (Polikar, 2012). Osim navedenih postoje i brojni drugi primjeri, zapravo i ovu disertaciju će ocijeniti višečlana komisija za ocjenjivanje doktorskog rada. U svakom od navedenih slučajeva cilj je kombinacijom više ocjena izbjeći pogrešnu odluku, premda je nemoguće u svakoj odluci izbjeći pogrešnu, opisanim načinom svakodnevno nastojimo umanjiti vjerojatnost da se pogrešna odluka dogodi.

Iz navedenih primjera korištenje takvih sustava se čini opravdano, a ponekad i jednostavno, međutim, za konstruiranje korisnog sustava postoje faktori koji se moraju uzeti u obzir. Kombinacija više klasifikatora s istim ciljem se naziva sustav višestrukih klasifikatora, a dva najvažnija faktora prilikom konstruiranja takvih sustava su točnost i raznolikost (Kuncheva et al., 2003; Ranawana i Palade, 2006). Glavni cilj ovog uvodnog poglavlja je dati prikaz tehnika konstrukcije uspješnih sustava višestrukih klasifikatora s naglaskom na različite pristupe kojima se nastoji osigurati veća kvaliteta cjelokupnog sustava.

U literaturi postoje brojni pristupi pri konstruiranju sustava višestrukih klasifikatora; u ovom pregledu fokus je stavljen na recentnije radove (Wang et al., 2011; Finlay, 2011; Zeng et al., 2014) s različitim pristupima, a sve s ciljem stvaranja teorijskog uvoda i prikaza trenutnog stanja znanosti na području koje je predmet ove disertacije. Kratki osvrt na teoriju mudrosti gomile (Surowiecki, 2004) je dan u odjeljku dva. Sukladno ciljevima ovog rada u odjeljku broj tri će se definirati sustavi višestrukih klasifikatora i dati teorijski dokaz da

ispravna kombinacija više klasifikatora utječe na smanjenje stope greške pojedinačnih klasifikatora. U četvrtom odjeljku su definirani osnovni faktori za kreiranje uspješnog sustava višestrukih klasifikatora. Potom su u petom odjeljku istraženi pristupi konstrukcije sustava višestrukih klasifikatora te njihove primjene, da bi se u šestom odjeljku analizirala strategija za konstruiranje sustava odabranog u ovoj disertaciji. Zadnji sedmi odjeljak je zaključak poglavlja.

2.2 Mudrost gomile

Iako je korištenje više stručnjaka za rješavanje nekog problema intuitivan pristup i vrlo lako se mogu pronaći brojni primjeri u svakodnevnom životu, neki od njih su istaknuti u uvodu ovog rada, primjena istog u znanstvenom istraživanju klasifikacije podataka je relativno nova te je blisko povezana s konceptom mudrosti gomile. Ideja da velika grupa ljudi može biti pametnija od pojedinaca koji predstavljaju stručnjake u nekom području jest iznesena u knjizi autora Surowiecki (2004). Spomenuti autor nije prvi koji je iznio tu teoriju, međutim način kako je opisan koncept i elementi koji utječu na gomilu da postane mudra su vrlo zanimljivi te su prenosivi kao smjernice na konstruiranje sustava višestrukih klasifikatora. Stoga kao uvod u disertaciju slijedi kratki osvrt na teoriju mudrosti kako je predstavljena u (Surowiecki, 2004).

Na tragu koncepta mudrosti gomile jest tvrdnja „*dvojica su pametnija od jednoga*“ koja se često može čuti u raznim prilikama gdje je potrebno donijeti odluku. Analogno se može zaključiti da su troje pametniji od dvoje, i tako slijedom se može doći do gomile. Tvrdnja da grupa ljudi može biti pametnija od pojedinaca koji predstavljaju elitu u nekom području, može biti teško prihvatljiva i predstavlja značajan udar na način kako se vodi poslovanje, stvara novo znanje i kako ljudi žive svoje svakodnevne živote.

„Mudrost gomile objašnjava princip grupnog razmišljanja, i koncept koji u suštini tvrdi da su mase bolje za rješavanje problema, stvaranje predviđanja i donošenje odluka u odnosu na bilo kojeg pojedinca (Surowiecki, 2004).“

Često se ističe da je mudrost gomile otkrivena pronalaskom Francisa Galtona koji je 1906. godine uočio, na temelju igre pogađanja u kojoj je sudjelovalo otprilike 800 ljudi, da je grupa koja je sudjelovala u prosjeku vrlo precizno predvidjela traženu težinu životinje na poljoprivrednom sajmu u Plymouth-u, Velika Britanija. Tog dana je spomenuti istraživač

naišao na jednostavnu ali vrlo snažnu i korisnu zakonitost; pod pravim okolnostima, grupe mogu biti izvanredno inteligentne. Tim grupama ne moraju dominirati najpametniji ljudi da bi bile inteligentne, i čak ako većina ljudi nije najbolje informirana ili racionalna, one ipak mogu donositi inteligentne odluke. Međutim ne može se za svaku grupu ili gomilu reći da je pametnija od pojedinca, moraju postojati određena pravila unutar grupe da bi ona bila mudra. Autor Surowiecki navodi četiri elementa koja su potrebna da bi se mogla prepoznati mudrost gomile.

Generiranje raznolikih rješenja nekog problema je osnovni uvjet korištenja više jedinki za donošenje odluka. Raznolikost u konceptualnom i kognitivnom smislu, može utjecati na istraživanje različitih ideja, generirajući važne razlike između tih ideja. Što je raznolikost manja, manje su raznoliki koncepti i ideje, stoga se zanemaruje potencijalno veliki raspon mogućih rješenja. Veća raznolikost povećava vjerojatnost da će netko predložiti radikalnu ili malo vjerojatnu ideju. Autor ističe da će se problem prije uspješno riješiti tako da se slučajnim odabirom izaberu članovi grupe nego ukoliko se odabiru samo pametni ljudi. Posljedično zaključuje, da će grupa ljudi s različitim razinama inteligencije donositi bolje odluke u odnosu na pametne pojedince. Iz toga se može zaključiti da ljudi koji donose manje znanja u grupu, paradoksalno, poboljšavaju performanse grupe. S druge strane postoje istraživanja koja ukazuju da se eksperti iz nekog područja često ne slažu oko mnogo stavova, i da su skloni čvrstom ustrajanju u svojim stavovima. Iako se time nikako ne podcjenjuje snaga ekspertnog znanja i njegova vrijednost, kombiniranje stavova i u ovom slučaju ima smisla u stvaranju dobre odluke.

Drugi bitan aspekt kolektivnog donošenja odluka jest **neovisnost**, koja se može preklapati s raznolikošću u jednom poimanju tog bitnog faktora. Ostvarivanje raznolikosti je neophodno za očuvanje neovisnosti. U ovom kontekstu neovisnost znači sloboda od utjecaja drugih. Neovisnost je bitna kako bi se spriječilo da pogreške koje nastaju kod nekih članova, postanu korelirajuće. Što su ljudi unutar grupe više povezani i imaju češći kontakt to su skloniji imati ista mišljenja i činiti iste pogreške. U skladu s pretpostavkom koja kaže da je u nesigurnim uvjetima najbolje prikloniti se stavu mnoštva, članovi grupe da bi izbjegli strah od neuspjeha ili osude i izrugivanja drugih priklanjaju se mnoštvu i ne iznose svoj stav. Takvo ponašanje je u suprotnosti s neovisnosti u odlučivanju i čest uzrok neuspješne gomile.

„**Umjetnost decentralizacije**“ je slijedeći uvjet mudrosti gomile, koji ohrabruje individualne članove grupe da donose bitne odluke, i to ne samo na jednoj lokaciji s jednim specifičnim tipom informacija, već disperzirano kroz nekoliko lokacija iz kojih se crpi i razmjenjuje lokalno znanje. To podrazumijeva da će skupina samostalnih pojedinaca koji se

vode svojim znanjem raditi zajedno na decentralizirani način na istom problemu. Na taj način odluke se donose od strane pojedinaca, iz različitih lokacija na način da se uvažava njihovo specifično znanje. Takav pristup suradnji može biti bolji u odnosu na pristup gdje jedna osoba upravlja svim odlukama od vrha prema nižim razinama hijerarhije.

Posljednji element koji utječe na uspješno korištenje mudrosti gomile, iznesenog u knjizi (Surowiecki, 2004), jest **agregacija** odluka. Grupa mora imati način kako agregirati pojedinačne odluke i mišljenja članova kako bi učinkovito donosila ukupne odluke. Prilikom agregiranja svaki glas mora imati priliku stvoriti utjecaj na konačnu odluku, tj. nitko ne smije biti zanemaren.

U suprotnosti navedenim uvjetima za uspješnu kolaboraciju pojedinaca u donošenju odluka, autor navodi i najčešće razloge neuspjeha gomile koji se događa kada je grupa: previše homogena, previše centralizirana, previše podijeljena, previše imitativna i previše emocionalna. U knjizi (Surowiecki, 2004) su navedeni primjeri koji potkrepljuju ove razloge kao i njihovo objašnjenje, koje zbog sažetosti ovog kratkog pregleda neće biti opisani.

Izneseni pogledi pružaju smjernice kombiniranja pojedinačnih odluka, od kojih se mnoge mogu prenijeti na klasifikaciju podataka. Dani okvir će se u slijedećem poglavlju formalno definirati u kontekstu teorije binarne klasifikacije podataka, a kasnije će se u disertaciji pokušati pronaći način da se ispune svi potrebni uvjeti za konstruiranje sustava višestrukih klasifikatora koji djeluje kao mudra gomila tj. kao sustav s rezultatima boljima od najboljeg pojedinca.

2.3 Definicija sustava višestrukih klasifikatora

Sustav višestrukih klasifikatora (engl. *Multi-Classifier System*) (u daljnjem tekstu SVK) je skup različitih klasifikatora čije individualne odluke su kombinirane na jedan od načina (najčešće ponderirano ili običnim glasanjem) s ciljem klasifikacije novih primjera (Dietterich, 2000). Formalno SVK se može definirati prema Definiciji 2.1.

Definicija 2.1 Svaki sustav koji:

- sadrži L pojedinačnih klasifikatora, gdje svaki pojedinačni član D_i generira labelu klase $s_i \in \Omega$, i = 1,…, L gdje je Ω skup svih dostupnih klasa
- za svaki klasifikacijski slučaj $x \in \Re^n$ generira vektor $s = [s_1, \dots, s_L]^T \in \Omega^L$ i

- koristi neku od funkcija koja kombinira sve pojedinačne odluke u jedinstvenu odluku sustava,

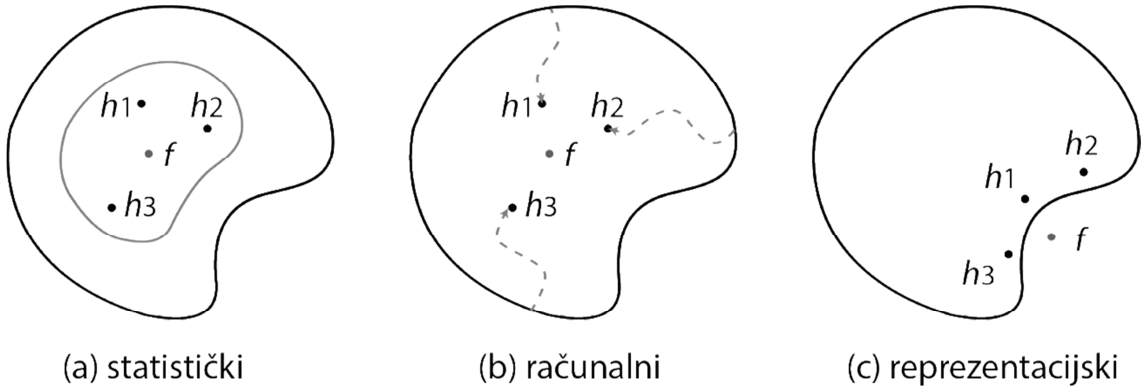
naziva se sustavom višestrukih klasifikatora.

U literaturi se osim korištenog naziva, sustavi višestrukih klasifikatora, mogu pronaći i nazivi: ansambli klasifikatora (engl. *ensembles*) i odbori (engl. *committees*) koji označavaju isti pojam.

Dietterich (2000) navodi, a citiraju ga (Zhang i Ma, 2012; Ranawana i Palade, 2006), tri osnovna razloga (problema) zbog kojih bi SVK trebao biti bolji od pojedinačnih klasifikatora.

Algoritam učenja može biti promatran kao pretraga prostora hipoteza *H* s ciljem identificiranja najbolje hipoteze u zadanom prostoru. Prvi, statistički problem nastaje kada je dostupna količina trening podataka premala u usporedbi s veličinom prostora hipoteza. Bez dovoljne količine podataka, algoritam može pronaći mnogo različitih hipoteza u prostoru hipoteza *H* koje daju istu točnost na trening podacima. Ukoliko se odabere samo jedan klasifikator kao rješenje, tada postoji rizik pogrešnog odabira (Kuncheva et al., 2003). Konstruiranjem SVK iz svih pronađenih preciznih klasifikatora, algoritam agregacije može npr. s prosjekom njihovih glasova smanjiti rizik odabira krivog klasifikatora.

Slika 2.1 u prikazu (a) oslikava navedenu situaciju. Vanjska krivulja označava prostor hipoteza *H*. Unutarnja krivulja označava skup hipoteza koje daju zadovoljavajuću točnost na podacima za trening. Točka označena s *f* je istinita hipoteza. Cilj sustava je s nekom vrstom agregacije hipoteza, najčešće glasanjem, pronaći dobru aproksimaciju točke *f* (Soni i Shavlik, 2011).



Slika 2.1 Tri prednosti SVK-a nad pojedinačnim klasifikatorima (prema (Dietterich, 2000))

Mnogi algoritmi funkcioniraju na temelju izvođenja nekog oblika lokalne pretrage i postoji mogućnost da se zaustave u lokalnom minimumu. Na primjer, neuronske mreže

koriste gradijentni spust (engl. *gradient descent*) za minimiziranje greške funkcije, a algoritmi stabla odluke koriste pohlepno pravilo dijeljenja (engl. *greedy splitting rule*) za stvaranje stabla. U slučajevima gdje postoji dovoljno trening podataka (tako da ne postoji statistički problem), može biti računalno vrlo teško pronaći najbolju hipotezu. SVK konstruiran tako da lokalna pretraga kreće iz različitih startnih točaka može pružiti bolju aproksimaciju istinite, nepoznate hipoteze koja je bolja od bilo kojeg individualnog klasifikatora kao što je prikazano na slici 2.1 prikaz (b).

Kao treći razlog navode se primjeri strojnog učenja kada istinita funkcija *f* ne može biti reprezentirana s niti jednom hipotezom unutar *H*. Korištenjem pondera rezultata (engl. *weighted sums*) hipoteza izvučenih iz *H*, moguće je proširiti prostor funkcija koje se mogu reprezentirati. Slika 2.1 prikaz (c) oslikava navedenu situaciju.

Iako su navedeni razlozi isključivo teorijska razmatranja, predstavljaju motivaciju za istraživanje sustava višestrukih klasifikatora. Osim teoretskih prednosti sustava višestrukih klasifikatora moguće je dokazati i utjecaj sustava na smanjenje stope greške klasifikacije u odnosu na pojedinačne klasifikatore. Autor (Dietterich, 2000) navodi da je najjednostavniji dokaz utjecaja SVK na smanjenje stope greške pomoću modela u kojem svi pojedinačni klasifikatori ostvaruju jednaku točnost (1-*p*) i rade nepovezane greške. Ovaj primjer podrazumijeva idealni slučaj kada je raznolikost klasifikatora savršena.

Vjerojatnost da će točno *k* klasifikatora od *N* krivo klasificirati neki slučaj je dana formulom (Dietterich, 2000)

$$\binom{N}{k} p^k (1-p)^{N-k} \tag{2.1}$$

a formula za vjerojatnost da će SVK krivo klasificirati zadani slučaj je dana formulom 2.2:

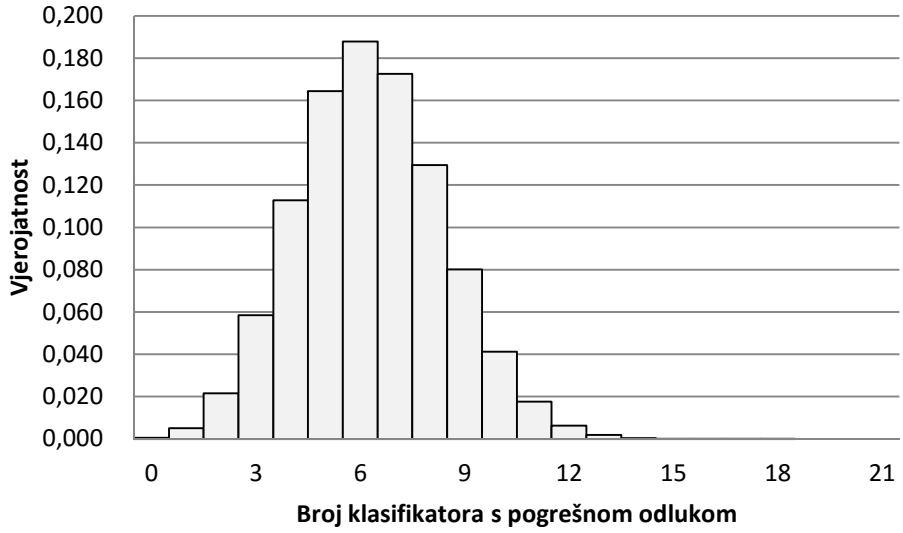
$$\sum_{k>N/2}^{N} \binom{N}{k} p^k (1-p)^{N-k} \tag{2.2}$$

Korištenjem formule 2.1 može se izračunati vjerojatnost da će SVK koji sadrži 21 hipotezu, od kojih svaka ima stopu pogreške od 0.3, pogrešno klasificirati neki primjer tj. vjerojatnost slučaja da će 11 i više hipoteza istodobno krivo klasificirati neku pojavu iznosi približno 0.026, što je značajnije manje od stope greške individualnih hipoteza (Dietterich, 2000). Navedeni slučaj je prikazan na slici 2.2.

Uvjet da bi se stopa pogreške mogle mjeriti pomoću definirane formule jest da svi pojedinačni klasifikatori unutar sustava rade u potpunosti nezavisne pogreške. Potpuna

nezavisnost između klasifikatora je teorijski koncept koji u empirijskim istraživanjima ne postoji jer je uvijek prisutna određena zavisnost između klasifikatora.

Kao što je pokazano ranije, raznolikost odluka (engl. *diversity*) klasifikatora je povezana sa stopom greške sustava na način da povećanje raznolikosti utječe na smanjenje greške, a samim time i na točnost klasifikacije.



Slika 2.2 Vjerojatnost da će točno *k* (od 21) klasifikatora napraviti grešku, pod pretpostavkom da klasifikatori imaju stope greške 0.3 i da rade greške neovisno o drugim članovima

Točnost i raznolikost, dva ključna faktora prilikom konstruiranja SVK-a, su definirani i opisani u slijedećem poglavlju.

2.4 Točnost i raznolikost

Kvaliteta SVK-a ovisi o dva faktora: točnosti i raznolikosti odluka. SVK neće funkcionirati ukoliko nisu zadovoljena dva temeljna uvjeta:

1) pojedinačni klasifikatori moraju imati veću točnost klasifikacije od 50% tj. ne smiju biti slučajni klasifikatori i
2) klasifikatori moraju generirati greške na različitim primjerima, što znači da ne smiju biti u potpunosti međusobno zavisni (Ranawana i Palade, 2006).

Točnost klasifikatora ovisi o prikladnosti algoritma i procesu prilagodbe algoritma podacima tj. treningu. Brojna istraživanja su posvećena kreiranju takvih modela klasifikatora

koji imaju cilj dati najveću točnost na određenom skupu kreditnih podataka (Oreski et al., 2012; Oreski i Oreski, 2014; Khashman, 2010).

Kada se promatra točnost u okviru SVK-a, tada se može primijetiti da iako je neophodna za funkcioniranje sustava točnost često nije najveća prepreka prilikom konstruiranja SVK-a. Najveći izazov predstavlja postizanje raznolikosti odgovora klasifikatora, članova sustava. Raznolikost odluka je česti fokus interesa mnogih istraživača tijekom proteklih godina (Brown et al., 2005). Definicija pojmova točnosti i raznolikosti klasifikatora slijedi u nastavku.

2.4.1. Točnost klasifikatora

Procjena kreditnog rizika kao i druge domene koje počivaju na binarnom problemu klasifikacije ima dvije dostupne klase. To znači da je klasifikacijski primjer (u ovom slučaju klijenta) potrebno svrstati, u jednu od dvije postojeće klase („loše“ ili „dobre“ klijente). Primjeri iz prve klase se nazivaju pozitivnim, a primjeri iz druge klase se zovu negativni primjeri (Kononenko i Kukar, 2007). Analiza točnosti klasifikatora se izvodi pomoću elemenata matrice konfuzije (engl. *confusion matrix*), prikazane u tablici 2.1. Sama matrica konfuzije je detaljnije objašnjenja za potrebe istraživanja u poglavlju 4. ove disertacije.

Tablica 2.1 Prikaz elemenata matrice konfuzije

ispravna klasa	klasificirano kao		Σ
	P	N	
P	TP	FN	POS = TP + FN
N	FP	TN	NEG = FP + TN
Σ	PP = TP + FP	PN = FN + TN	n=TP+FN+FP+TN

* *P – pozitivna klasa, POS – broj pozitivnih primjera, n – ukupan broj primjera*
N – negativna klasa, NEG – broj negativnih primjera
TN - broj ispravno predviđenih negativnih ishoda
FP - broj pogrešno predviđenih pozitivnih ishoda
FN - broj pogrešno predviđenih negativnih ishoda
TP - broj ispravno predviđenih pozitivnih ishoda

Da bi se utvrdile performanse treniranog klasifikacijskog modela potrebno je testirati njegove odluke pomoću primjera koji su izdvojeni iz skupa podataka i nisu korišteni u ciklusu treninga. Točnost klasifikacije je jednostavna mjera koja računa omjer između testnih primjera kojima je klasa točno predviđena i ukupnog broja slučajeva. Za binarnu klasifikaciju računa se pomoću elemenata matrice konfuzije prema formuli 2.3.

$$acc = \frac{TP + TN}{n} \tag{2.3}$$

Točnost pojedinih klasifikatora unutar SVK-a je bitan faktor prilikom konstrukcije sustava, te može utjecati na kvalitetu predikcije. Cilj je koristiti članove sa što većom točnosti predikcije, međutim to nije uvijek garancija staranja sustava koji će postići rezultate bolje od pojedinačnih klasifikatora. U obzir treba uzeti i raznolikost klasifikatora ili sposobnost članova sustava da neovisno donose odluke, drugi faktor koji bitno doprinosi performansama sustava.

2.4.2. Raznolikost odluka klasifikatora

Ukoliko je dostupan savršeni klasifikator koji ne generira greške, tada nam SVK nije potreban. Međutim ukoliko klasifikator radi greške, tada je potrebno pronaći drugi klasifikator (najčešće više njih) kao njegov komplement, koji će raditi pogreške na drugim primjerima. Raznolikost se odnosi na način kako klasifikatori čine greške tj. za skup klasifikatora se kaže da su raznoliki ukoliko ne rade iste greške na istim primjerima (Tang et al., 2006; Wang i Yao, 2013). Raznolikost označava „specijalizaciju“ pojedinih klasifikatora na različitim dijelovima ulaznog skupa podataka i iznimno je bitna za uspješnost sustava.

Izazov izrade SVK-a jest u treniranju što preciznijih pojedinačnih klasifikatora, koji rade greške na različitim primjerima. U teoriji postoji nekoliko mjera raznolikosti, koje se dijele na uparene i neuparene mjere (engl. *pairwise and non-pairwise measures*) (Kuncheva, 2004). Međutim u istraživanjima, daleko najčešća mjera koja se koristi za mjerenje raznolikosti odluka klasifikatora je Q statistika (engl. *the Q Statistics*) koja spada u uparene mjere. Mjera je korištena u istraživanju utjecaja klasne nejednakosti na sustav višestrukih klasifikatora (Wang i Yao, 2009). Ista je odabrana i u istraživanju (Kuncheva et al., 2003) performansi SVK-a u zavisnosti od: performansi klasifikatora članova, njihovog broja i ostvarene mjere raznolikosti. Q statistika je korištena i za uklanjanje klasifikatora slučajne šume (engl. *random forest*) koji negativno utječu na rezultate sustava (Banfield et al., 2005), gdje je autori odabiru kao najbolju mjeru zbog jednostavne primjenjivosti i razumljivosti. Prema preporukama istraživača koji su koristili spomenutu mjeru, u disertaciji će za potrebe mjerenja raznolikosti također biti korištena Q statistika. Slijedi opis odabrane mjere.

Uzmimo da je $X = \{x_1, \dots, x_N\}$ skup podataka koji opisuje neki klasifikacijski problem. Rezultat nekog klasifikatora D_i možemo prikazati kao N-dimenzionalni binarni

vektor $y_i = [y_{1,i}, \dots, y_{N,i}]^{\mathrm{T}}$, takav da je $y_{j,i} = 1$ ukoliko D_i prepozna ispravno x_j i -1 u suprotnom, uz $i = 1, \dots, L$, pri čemu je L broj klasifikatora. Tada se odnos između dva klasifikatora D_k i D_i može prikazati pomoću tablice 2.2. Četiri su moguća ishoda, ovisno o odlukama klasifikatora na temelju kojih se sastavlja tablica odnosa. Pomoću tablice u kojoj su evidentirani svi test primjeri računaju se uparene mjere raznolikosti za bilo koja dva klasifikatora, između ostalih i Q statistika.

Tablica 2.2 Prikaz elemenata Q statistike

	D_k ispravno (1)	D_k pogrešno (-1)
D_i ispravno (1)	N^{11}	N^{10}
D_i pogrešno (-1)	N^{01}	N^{00}
Ukupno	$N = N^{00} + N^{01} + N^{10} + N^{11}$	

Autori (Kuncheva, L. I., & Whitaker, 2003; Zhou, 2012) definiraju Q statistiku (negdje i Yulova Q statistika) za dva klasifikatora D_i i D_k:

$$Q_{i,k} = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}} \tag{2.4}$$

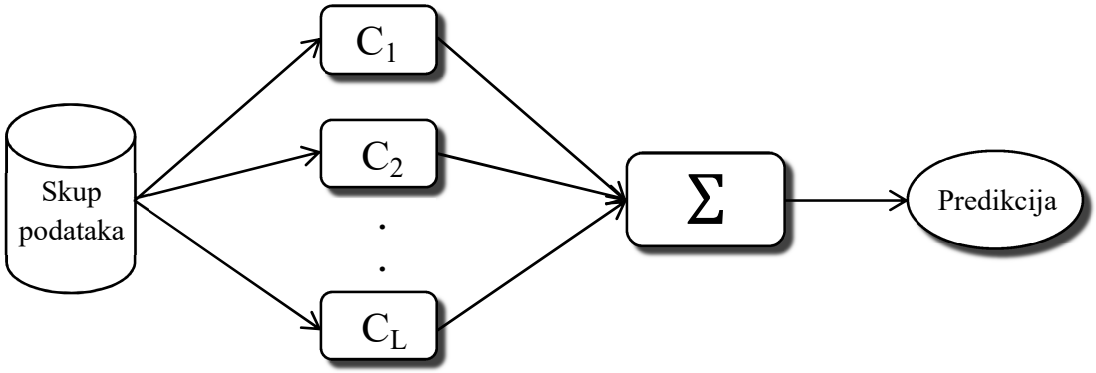
gdje je N^{ab} broj elemenata x_j od X za koje $y_{j,i} = a$ i $y_{j,k} = b$ (prikaz tablica 2.2). Za nezavisne klasifikatore, očekivana vrijednost $Q_{i,k}$ je 0. Vrijednost Q može varirati između -1 i 1. Klasifikatori koji više prepoznaju iste slučajeve jednako će imati pozitivne vrijednosti Q, a oni koji više rade greške na različitim slučajevima će imati Q vrijednosti negative. Za sustav D od L klasifikatora, prosjek Q statistike za sve parove klasifikatora je,

$$Q = \frac{2}{L(L-1)} \sum_{i=1}^{L-1} \sum_{k=i+1}^{L} Q_{i,k} \tag{2.5}$$

Osim Q statistike postoje i druge mjere koje se koriste ali u vrlo malom broju istraživanja te u ovom radu neće biti pojedinačno predstavljene (Brown et al., 2005; Wang i Yao, 2009; Yu et al., 2008). **Mjere za raznolikost služe da bi se utvrdilo koliko su klasifikatori nezavisni, a način kako se postiže nezavisnost odluka klasifikatora predstavlja strategiju o kojoj ovisi konačan rezultat kombiniranja klasifikatora.** Mogući pristupi za konstruiranje SVK-a su opisani u sljedećem poglavlju.

2.5 Konstruiranje sustava višestrukih klasifikatora

Prilikom izrade SVK-a, utjecaj na konačni rezultat je moguće ostvariti kroz različite odluke koje se tiču same konstrukcije. Osnovni elementi koji su nužni za konstruiranje SVK-a su prikazani na slici 2.3. Kao svaka tehnika strojnog učenja i SVK nužno uključuje povijesne podatke koji su nužni za trening klasifikacijskih modela. Povijesni podatci su tipični skupovi podataka za nadzirano učenje koji sadrže pojedinačne primjere promatrane pojave. Osim podataka za izradu SVK-a potrebni su klasifikatori (C_1 do C_L na slici 2.3), koji će biti trenirani na trening skupu i koji će zasebno donositi predikcijske odluke. Prilikom odabira klasifikacijskog algoritma ili više njih koji će biti korišteni, potrebno je uzeti u obzir dva bitna faktora koji su opisani u prethodnom poglavlju. Posljednja odluka tijekom konstrukcije jest kako će biti izvršena agregacija pojedinačnih odluka. Odabrana funkcija spajanja mora uzeti u obzir sve odgovore, te na temelju njih odabrati konačan odgovor koji će se pojaviti na izlazu sustava. Funkcija spajanja može samo odabrati jedan od predloženih odgovora te nikako drugačije ih mijenjati ili utjecati na iste.



Slika 2.3 Jednostavna shema sustava višestrukih klasifikatora

Možda najteža i najvažnija odluka prilikom konstrukcije SVK-a jest vezana uz raznolikost odgovora klasifikatora tj. kako je postići. Ovisno o donesenoj odluci stvara se cjelokupna strategija. Osnovne strategije postizanja raznolikosti se najčešće mogu svesti na nekoliko metoda (Ranawana i Palade, 2006): poduzorkovanje testnog skupa, manipulacija ulaznih atributa, manipulacija izlaznih klasa, uključivanje slučajnosti i manipulacija parametara algoritma. U svakom SVK-u kombinacija odluka članova je korisna isključivo ukoliko se članovi ne slažu u klasifikaciji inputa. Stoga je cilj prilikom treniranja klasifikatora postići što veću raznolikost među članovima ali ujedno i što veću točnost klasifikacije.

Različiti istraživači su koristili različite pristupe i vidljivo je da ne postoji jedna metoda koja bi bila najbolja za sve probleme klasifikacije (engl. *„no free lunch“ theorem*).

2.5.1. Poduzorkovanje testnog skupa

Poduzorkovanje testnog skupa podrazumijeva izvršenje trening algoritma nekoliko puta na različitom podskupu trening podataka s ciljem stvaranja različitih hipoteza. Takva metoda je korisna za nestabilne trening algoritme, to jest one koji generiraju velike promjene na izlazu kao odgovor na male promjene u ulaznim podacima. Najbolji primjer nestabilnih algoritama su SVM, neuronske mreže i stabla odlučivanja. Na poduzorkovanju testnog skupa zasnivaju se dvije najpopularnije tehnike: „Bagging“ i „Boosting“ (Bar et al., 2012).

Tehnika Bagging, skraćeni naziv od bootstrap aggregating, sadrži možda najjednostavniji algoritam za kreiranje SVK-a koji postiže jako kvalitetne rezultate. Kao rezultat poduzorkovanja testnog skupa, algoritam jest dobar i za izbjegavanje problema pretreniranosti (engl. *overfitting*) modela (Bühlmann, 2012). Za svaki klasifikator generira se poseban skup podataka dobiven bootstrap postupkom iz izvornog skupa podataka. Obzirom da se koristi uzorkovanje sa zamjenom neki od primjera mogu više puta ući u skup podataka za trening, dok neki primjeri neće niti jedanput biti izabrani, što osigurava različiti skup trening primjera u svakom ciklusu. Metoda je našla primjenu u brojnim istraživanjima iz različitih domena (Fraz et al., 2012; Kempa et al., 2011; Kim i Kang 2010).

Tehnikom Boosting skup podataka za trening se odabire na temelju performansi prethodnog klasifikatora u nizu. Slučajevi koji su neispravno klasificirani imaju veću šansu da uđu u skup na temelju kojeg uči slijedeći klasifikator (Brown i Kuncheva, 2010). Stoga Boosting nastoji producirati nove klasifikatore koji mogu bolje predviđati primjere za koje su prijašnji klasifikatori bili loši (Maclin i Optiz, 2011).

Zbog svoje jednostavnosti i kvalitete rezultata, Bagging i Boosting su najpopularnije tehnike korištene u dosadašnjim istraživanjima. Iste tehnike će biti korištene u posljednjem poglavlju ove disertacije u kojem će biti uspoređene s novom tehnikom koja je konstruirana. Detaljniji opis algoritama i njihov pseudo kod se nalazi u 5. poglavlju.

2.5.2. Manipulacija ulaznih atributa

Manipulacijom ulaznih atributa klasifikatori prilikom svakog treninga uče na različitim podskupovima atributa. Prema autoru (Dietterich, 2000) takva trening metodologija se pokazala uspješna samo na skupovima s visoko redundantnim atributima. Kao primjer

uspješne implementacije je istraživanje s ciljem identifikacije vulkana na planetu Veneri (Cherkauer, 1996). Autori su koristili 32 neuronske mreže u sustavu višestrukih klasifikatora s 8 različitih podskupova atributa koji su odabrani iz skupa od 119 atributa.

Istraživanje koje je dovelo do zaključka da skupovi podataka s manjom redundancijom dimenzija nisu prikladni za manipulaciju ulaznih atributa je provedeno na skupu sonarnih podataka (Tumer i Ghosh, 1996). Skup sadrži ukupno četiri klase i 25 različita atributa. SVK je kreiran od četiri klasifikatora tako da svaki sadrži 22 atributa s najvećom korelacijom prema pojedinoj klasi. Premda je istraživanje provedeno na jednom skupu podataka, zaključak je autora da se izostavljanjem nekih atributa iz malog broja atributa narušavaju performanse klasifikatora.

2.5.3. Manipulacija izlaznih klasa

Manipulacija izlaznih klasa se zasniva na mijenjanju broja izlaznih klasa skupa podataka koji se koristi prilikom treninga algoritma. Ova metoda je prikladna u slučaju da klasifikacijski problem sadrži veliki broj izlaznih klasa. Istraživači mogu podijeliti izlazne klase u manji broj podskupova klasa koji mogu biti korišteni prilikom treninga. Sve klase u podskupu se preimenuju s istom oznakom prije treninga. Stoga trenirani klasifikator može razlikovati manji broj klasa na kojim je treniran. Ponavljanjem ovog postupka moguće je trenirati više klasifikatora.

2.5.4. Uključivanje slučajnosti i manipulacija parametara algoritma

Uključivanjem slučajnosti u trening modela može se utjecati na performanse klasifikatora. Najbolji primjer za uključivanje slučajnosti je algoritam neuronske mreže. Ukoliko ne mijenjamo arhitekturu neuronskih mreža te inicijaliziramo trening s drugim ponderima mreža može generirati različite hipoteze. Mijenjanjem pondera možemo trenirati željeni broj klasifikatora. Slično se može manipulirati s parametrima algoritama. Svaki algoritam ima određen broj parametra pomoću kojih se može kontrolirati tijek treninga. Vrijednosti parametara se mogu mijenjati na istom skupu podataka s ciljem postizanja raznolikosti u rezultatima.

U slijedu istraživanja provedenih u ovoj doktorskoj disertaciji i opisanih u slijedećim poglavljima, konstruiranje SVK-a se bazira na manipulaciji ulaznih atributa. Odabir strategije se temelji na ranijim istraživanjima autora (Oreski et al., 2012; Oreski i Oreski, 2014) iz kojih

se može zaključiti da uključivanje odabira atributa u konstruiranje klasifikatora može donijeti poboljšanje performansi na kreditnom skupu podataka.

2.6 Razvoj SVK-a temeljenog na manipulaciji ulaznih atributa

Premda su neki autori (Cherkauer, 1996; Tumer i Ghosh, 1996) prilično rano istraživali SVK manipulacijom ulaznih atributa i došli do zaključka da skupovi podataka s manjom redundancijom dimenzija nisu prikladni za manipulaciju ulaznih atributa, u prvom redu zbog toga što izostavljanje nekih atributa na malom broja atributa narušava performanse klasifikatora, time nije otklonjena mogućnost postizanja dobrih rezultata navedenom metodom. U novije vrijeme postoji više razloga koji daju povod za istraživanjem ove vrste SVK. Prije svega kao razlog se može navesti činjenica da je u međuvremenu došlo do podatkovne eksplozije; sustavi autonomno generiraju sve veće količine raznovrsnih, visokodimenzionalnih podataka o različitim entitetima, koji su često dobra podloga za klasifikaciju tih entiteta. Redundancija dimenzija u podatcima, o kojoj je ranije bilo riječi, u novije vrijeme sve češće postoji ali se do njezina otkrivanja ne dolazi jednostavno, već primjenom različitih mjera.

Konkretno, ako za klasifikaciju instanci nekog skupa primjenom korelacije, kao evaluacijske funkcije (tehnike odabira atributa) za određivanje relevantnosti atributa, dolazimo do jednog podskupa atributa koji dobro opisuju instance tog skupa, a primjenom npr. informacijske dobiti kao evaluacijske funkcije dolazimo do nekog drugog podskupa atributa koji također dobro opisuju instance tog skupa, tada se može zaključiti da u navedenom skupu podataka postoji redundancija dimenzija do koje dolazimo primjenom različitih metoda. Ili možda još bolje, postoje podskupovi relevantnih atributa koji su međusobno dovoljno različiti. Tu pretpostavku možemo iskoristiti za konstrukciju efikasnih SVK.

Stoga će se predloženi pristup konstrukciji SVK temeljiti ne na slučajnom odabiru atributa nego na predefiniranom načinu odabira atributa za svaki klasifikator. Evaluacijska funkcija koja se koristi za određivanje relevantnosti atributa mjeri sposobnost atributa, ili podskupa atributa, da uzorke podataka svrsta u neku od klasa. Pri tome, relevantnost atributa ovisi o evaluacijskoj funkciji te se odabirom različitih evaluacijskih funkcija (tehnika odabira atributa) dobivaju različiti podskupovi kao optimalni. Vjerojatnost da će u podskup atributa

biti uključeni različiti atributi u značajnoj mjeri ovisi o tome što evaluacijske funkcije mjere, odnosno, na čemu se one temelje. U skladu s tom pretpostavkom može se očekivati da će tako konstruirani SVK osigurati potrebnu raznolikost rezultata klasifikatora.

S obzirom na ono što mjere, evaluacijske funkcije se dijele u pet grupa (Dash i Liu, 1997.):

1) mjere udaljenosti,
2) mjere informacija,
3) mjere ovisnosti,
4) mjere dosljednosti i
5) mjere pogreške klasifikatora.

Broj različitih tehnika koje će se odabrati, nije ograničen npr. u istraživanjima provedenim u okviru disertacije je korišteno pet različitih tehnika odabira atributa koje dobro pokrivaju različite mjere. Korištene su: korelacija, Gini indeks, omjer informacijske dobiti, reljefna metoda, mjera nesigurnosti.

Jedan od problema koji se pojavljuje u realizaciji opisane tehnike za konstrukciju sustava višestrukih klasifikatora je kako odrediti optimalni broj atributa koji će selektirati svaka od tehnika. Na taj problem nema jednoznačnog odgovora te je često potrebno konstruirati nekoliko modela s različitim brojem atributa koji će pokazati s kojim brojem atributa se postižu najbolji rezultati.

Temeljem opisanih polaznih osnova za konstrukciju SVK, jedan od većih problema novijeg doba koji se ogleda u velikim količinama podataka s mnoštvom dimenzija pretvara se u prednost prilikom klasifikacije.

2.7 Zaključci poglavlja

Baze podataka poslovnih sustava širom svijeta kumuliraju ogromne količine podataka o svojim klijentima i poslovnim partnerima. Ti podatci mogu biti značajni kapital i strateška prednost ako su iskorišteni na adekvatan način. Metode i tehnike umjetne inteligencije predstavljaju alat za iskorištavanje tih podataka. Jedan od ključnih zadataka koji se pri tome javlja i stavlja pred umjetnu inteligenciju je klasifikacija velikih količina podataka. Rješavanju tog problema, teorijski i praktično, može se pristupiti na više načina, odnosno na više područja. Odabir varijabli ključnih za konstrukciju klasifikacijskih modela jedno je od

područja. Konstrukcija naprednih klasifikacijskih algoritama je drugo područje djelovanja, a kombiniranje klasifikatora predstavlja treće važno područje za klasifikaciju podataka.

U uvodu ovog doktorskog rada je predstavljen koncept sustava višestrukih klasifikatora, potom i osnovni faktori o kojim treba voditi brigu prilikom sastavljanja sustava, te su konačno opisane strategije konstruiranja SVK koje uključuju: poduzorkovanje testnog skupa, manipulacija ulaznih atributa, manipulacija izlaznih klasa, uključivanje slučajnosti i manipulacija parametara algoritma. Teorija sustava višestrukih klasifikatora je šira od ovog uvodnog dijela, i postoje elementi teorije koji nisu ovdje uključeni. Opisano u ovom poglavlju predstavlja osnovni uvod za istraživanja koja slijede u ovoj disertaciji. Istraživanja će se fokusirati na primjenu sustava višestrukih klasifikatora temeljenih na odabiru atributa kod problema procjene kreditnog rizika.

Konstrukcija sustava višestrukih klasifikatora na skupu kreditnih podataka logični je nastavak rada autora ovog rada, nakon što su u ranijim istraživanjima predmet istraživanja bili odabir ključnih varijabli i konstrukcija naprednih algoritama za klasifikaciju. Premda je u ranijim istraživanjima nekih autora u drugim domenama bilo pokušaja da se SVK konstruira manipulacijom ulaznih atributa, promjena pristupa kao i promjena okolnosti u odnosu na vrijeme tih pokušaja opravdavaju očekivanja u uspješnost novih istraživanja. Upravo promijenjene okolnosti kao i promijenjen pristup problemu mogu biti izvor dobrih rezultata i poticaj drugim znanstvenicima za istraživanje ovog izazova na novi način.

3

Algoritam za smanjivanje sustava višestrukih klasifikatora

U ovom poglavlju je razvijen vlastiti pohlepni algoritam za smanjivanje sustava višestrukih klasifikatora koji za metriku odabira koristi novu mjeru pod nazivom koeficijent uspješnosti klasifikatora.

3.1 Uvod

Područje istraživanja vezano uz sustave višestrukih klasifikatora (SVK) je postalo vrlo popularno tijekom posljednjih godina zahvaljujući njihovom potencijalu u poboljšavanju točnosti klasifikacije podataka (Partalas et al., 2012; Lacy et al, 2015). Sve više znanstvenika iz različitih područja istraživanja kao što su: statistika, strojno učenje, prepoznavanje uzoraka u podacima (engl. *Pattern Recognition*) i otkrivanje znanja u bazama podataka (engl. *Knowledge Discovery in Databases*)(Tsoumakas et al., 2009), zaključuje da kombiniranje zasebnih klasifikacijskih modela postaje pouzdan pristup koji unapređuje performanse klasifikacijskih tehnika (Gaber i Bader-El-Den, 2012; Coletta et al., 2015). Tsoumakas i suradnici (2008) navode da uspjeh SVK proizlazi iz činjenice da nude dobra rješenja za nekoliko zanimljivih problema iz domene strojnog učenja tj. nude poboljšanje performansi predikcije za: učenje iz više fizički podijeljenih izvora podataka, skaliranje induktivnih algoritama na velike baze podataka i učenje iz podataka koji su podložni promjeni svojstava zavisne varijable (engl. *concept drift*).

Generalno govoreći, izrada SVK se sastoji od dva koraka: prvo se trenira više zasebnih modela s jednim ciljem, a potom se ti modeli kombiniraju na neki način, najčešće glasanjem (engl. *voting*), s ciljem ostvarivanja boljih performansi (Li et al., 2012). Povrh navedenih, moguće je razmatrati i korak kroz koji se radi redukcija pojedinačnih modela prije njihovog kombiniranja u SVK. Spomenuti korak u literaturi se naziva smanjivanje sustava višestrukih klasifikatora (engl. *ensemble pruning*) i jedan je od predmeta istraživanja u ovoj disertaciji. Smanjivanje SVK je važno iz dva razloga: *efikasnosti* i *performansi predviđanja* (Tsoumakas et al., 2009). Bez obzira na dobre rezultate klasifikacije, nedostatak SVK jest da je ponekad potrebno kombinirati veliki broj klasifikatora da bi se osigurala konvergencija greške do asimptotske vrijednosti. To zahtjeva velike memorijske i procesorske kapacitete računala i usporava samu klasifikaciju (Prodromidis i Stolfo, 2001). Rješenje za opisani problem jest odabir samo dijela tj. podskupa dostupnih modela što rezultira smanjenjem kompleksnosti konačnog sustava. Stvaranjem SVK s manjim brojem članova postiže se *efikasniji* sustav koji zahtjeva manje računalnih resursa i potrebnog vremena za klasifikaciju.

Podjednako važan je i drugi razlog, *performanse predviđanja*. Kada SVK sadrži veliki broj modela (članova, hipoteza) tada postoji mogućnost da uz modele koji imaju dobre performanse predikcije, sustav sadrži i one druge koji to nemaju. Potonji ne doprinose uspješnosti i mogu negativno utjecati na ukupne performanse sustava. Štoviše, SVK može

sadržavati više modela koji imaju slične odluke što značajno smanjuje raznolikost i sposobnost korekcije greške (Tsoumakas et al., 2009). Cilj smanjivanja SVK je izbaciti modele koji negativno utječu na kvalitetu odluke sustava i prepustiti utjecaj samo onim modelima koji su međusobno raznoliki i pridonose performansama predikcije. Postojeća istraživanja (Martinez-Muoz et al., 2009; Li et al., 2014; Ma et al. 2015) potvrđuju da smanjeni SVK (engl. *pruned ensemble*) može poboljšati generalizacijske performanse i pozitivno utjecati na točnost klasifikacije.

Problem smanjivanja SVK se može definirati kao optimizacijski problem; odnosno zadatak pronalaska jednog podskupa dostupnih modela, članova sustava, koji optimizira mjeru koja indicira performanse generalizacije (npr. točnost na odvojenom validacijskom skupu podataka). Obzirom da je broj mogućih podskupova, sustava koji ima T članova, jednak 2^T-1 (prazan skup nije uključen), iscrpna pretraga postaje računski skupa za sustav umjerene veličine (Partalas et al.,2010). Učinkoviti pristup za smanjivanje SVK jest korištenje heurističke pretrage prostora potencijalnih rješenja, koja ovisno o paradigmi pretraživanja može biti pohlepna ili stohastička. Neovisno o načinu pretraživanja, pretraga mora biti vođena nekom metrikom za evaluaciju svakog potencijalnog kandidata, tj. podskupa (Tsoumakas et al., 2008). Pohlepna pretraga je jedna od najpopularnijih metoda u znanstvenoj literaturi i predmet je istraživanja u ovom poglavlju, dok stohastička pretraga uključuje slučajnost prilikom izbora slijedećeg podskupa kandidata te se na taj način nastoji izbjeći „zapinjanje“ u lokalnom optimumu.

U ovom poglavlju, autor istražuje novi, ovdje predloženi, pohlepni algoritam za smanjivanje sustava višestrukih klasifikatora pod nazivom COP (engl. *Classifier Odds based Pruning*). Algoritam se zasniva na koeficijentu uspješnosti klasifikatora (engl. *Classifier Odds, CO*) kao novoj metrici koja se koristi prilikom odabira novih klasifikatora u podskup. Cilj novog algoritma je na temelju nove mjere CO, stvoriti brzi odabir podskupa klasifikatora uz povećanje efikasnosti i poboljšanje performansi sustava. Istraživanje se fokusira na popularnu postavku SVK gdje se članovi kombiniraju glasanjem (engl. *voting*), a dodavanje novih klasifikatora trenutnom sustavu (trenutno odabranom podskupu) zasniva se na usporedbi kvalitete novih SVK-a nakon dodavanja kandidata, tj. nakon dodavanja novih klasifikatora. Svaki novi SVK se promatra kao jedinstveni klasifikator koji kombinira odluke svojih članova. Istraživanje predstavlja teorijsku analizu utjecaja novog COP algoritma na efikasnost (broja članova) i performanse (točnost klasifikacije) odabranog podskupa u odnosu na cijeli SVK na generiranom skupu podataka i daje obećavajuće rezultate po pitanju oba kriterija na testiranim primjerima.

Preostali dijelovi ovog poglavlja su organizirani na sljedeći način. Odjeljak 2 opisuje problem smanjivanja SVK koji se proučava u ovom poglavlju i analizira literaturu povezanu s ovom temom. Pohlepni algoritmi za smanjivanje SVK su objašnjeni u trećem poglavlju. Odjeljak 4 opisuje dizajn eksperimenta u pogledu: generiranja skupa podataka, definiranja COP algoritma i CO mjere, klasifikacije, ocjenjivanja i usporedbe rezultata. Odjeljak 5 analizira i diskutira rezultate eksperimenta s fokusom na efikasnost i performanse dobivenih rezultata u odnosu na cijeli SVK. Zaključci i smjernice za budući rad su izneseni u odjeljku 6.

3.2 Opis problema i pregled literature

Smanjivanje sustava višestrukih klasifikatora, u suštini, jest pokušaj da se uklone oni klasifikatori koji ne pridonose poboljšanju performansi sustava. Uz dani skup treniranih klasifikatora, odabir podskupa SVK s najboljim generalizacijskim performansama je izazov iz dva razloga: (1) nije jednostavno procijeniti generalizacijsku sposobnost podskupa SVK i (2) pronalazak optimalnog podskupa je problem kombinatorne pretrage s eksponencijalnom računalnom složenošću, stoga je računanje egzaktnog rješenja iscrpnom pretragom često neprovedivo te je potrebna heuristička pretraga (Li et al., 2012; Oreski, 2014). Literatura nudi različite pristupe rješavanju navedenog problema koji ugrubo mogu biti klasificirani u dvije grupe: (1) globalna pretraga i (2) lokalna pretraga. Pohlepni algoritmi spadaju u drugu grupu koja koristi heuristiku za sužavanje prostora pretrage; to je pristup koji nastoji iskoristiti prednosti tehnike koja efikasno pretražuje prostor da bi se računski jeftino pronašao podskup klasifikatora koji po svojim performansama nadmašuje SVK temeljen na cijelom skupu treniranih klasifikatora.

Cilj ovog istraživanja je razviti vlastiti algoritam za smanjivanje SVK-a, pod nazivom COP, koji bi za metriku odabira koristio novu mjeru pod nazivom koeficijent uspješnosti klasifikatora (CO). Novi algoritam treba biti brz i računalno ne zahtjevan te treba pružiti dobre rezultate u vidu efikasnosti i performansi predviđanja. Razlika u odnosu na prethodne algoritme dostupne u literaturi jest nova mjera koja efikasno bira smjer odabira rješenja, tj. novi klasifikator koji će dodati u SVK.

Prije iznošenja metodologije istraživanja, a u svrhu analize trenutnog stanja znanosti na spomenutom području u nastavku slijedi pregled literature. Smanjivanje SVK, kao što je već spomenuto, je vrlo atraktivna tema i postoje brojni radovi koji su dali doprinos tom

području istraživanja. Dobro polazište za istraživanje smanjivanja SVK-a jest rad (Tsoumakas et al., 2009) koji predlaže taksonomiju postojećih metoda. Osim pregleda postojećih metoda rad vrlo jasno iznosi njihove prednosti i nedostatke ali i definicije nekih ključnih pojmova iz tog područja. U širem kontekstu, istraživanje smanjivanja SVK-a se može podijeliti na dva pristupa. **Prvi pristup** je korištenjem globalne pretrage prostora s ciljem pronalaska optimalno ili približno optimalnog rješenja.

Pomoću genetskih algoritama u istraživanju, autori Zhou et al. (2002) su smanjivali homogeni sustav neuronskih mreža tako što su evoluirali težinske faktore koji označavaju prikladnost (engl. fitness) uključivanja pojedinih modela u sustav. Nova metoda GASEN je u istraživanju testirana na 20 skupova podatka i uspoređena je s metodama Bagging i Boosting. Rezultati su pokazali da je nova metoda bolja po pitanju smanjivanja veličine i generalizacijskih sposobnosti sustava u odnosu na uspoređene metode.

Autori Lazarevic i Obradovic (2001) koriste pristup klasteriranja za grupiranje sličnih klasifikatora te potom izbacuju redundantne klasifikatore. Klasteriranje se izvodi pomoću algoritma k-srednjih vrijednosti (k-means based clustering) na tri realna skupa podataka. Istraživanje pokazuje da se upotrebom predloženog pristupa mogu ostvariti isti ili bolji rezultati u odnosu na cijeli sustav uz smanjenje od 50%-70% članova. Iste zaključke iznose i drugi autori koji su koristili sličan pristup (Lin et al., 2014).

U objavljenim istraživanjima su korištene i druge metode globalne pretrage, pa se tako problem odabira podskupa SVK-a može definirati i kao problem kvadratnog cjelobrojnog programiranja (engl. *quadratic integer programming problem*) koji se učinkovito rješava primjenom semidefinitnog programiranja (engl. *semi-definite programming, SDP*) (Zhang et al, 2006; Xu & Gray, 2013). Metoda optimizacije rojem čestica je također uspješno korištena (Zhang & Chau, 2009) za smanjivanje višeslojnog SVK-a.

Zajednička karakteristika svih spomenutih metoda jest da su vremenski i računalno vrlo zahtjevne. Iako često postižu dobre rezultate, njihova negativna strana u vidu računalne zahtjevnosti može biti poprilično značajna pogotovo na većim sustavima (Li et al., 2012). U nastojanju da bi se uklonio spomenuti nedostatak istraživači koriste **drugi pristup** za smanjivanje SVK-a koji koristi lokalnu pretragu prostora. Lokalna pretraga se izvodi pomoću pohlepnih algoritama koji nastoje pronaći najbolje globalno rješenje pohlepnim odabirom slijedećeg koraka prilikom promjene podskupa (Partalas et al. 2012). Bitan element odabira slijedećeg koraka (člana skupa) jest evaluacijska mjera koju algoritmi koriste.

Ranija istraživanja su koristila mjere zasnovane na performansama klasifikatora ili na njihovoj raznolikosti. Smanjivanja koje su se bazirala na performansama pojedinih

klasifikatora, uključivala su mjeru točnosti klasifikacije u odabir novih članova (Fan et al., 2002). Osim točnosti korištene su i druge mjere: korijen sredine kvadrata odstupanja (engl. *root mean squared error*), F-Score mjera, preciznost/odziv, ROC mjera (Caruana et al., 2004). Druga istraživanja (Tang et al.,2006; Banfield et al., 2005) su analizirala mjere raznolikosti za odabir podskupova klasifikatora, kao što je Q statistika. Sinteza zaključaka istraživanja koja su se temeljila na smanjivanju SVK-a na temelju raznolikosti jest da ne postoji suglasnost o efektu utjecaja raznolikosti na odabir članova sustava. Neka istraživanja tvrde da je nemoguće napraviti dobar podskup sustava na temelju raznolikosti (Tang et al., 2006) dok postoje i druga istraživanja koja ostvaruju dobre rezultate (Martınez-Munoz i Suárez, 2004; Banfield et al., 2005). Međutim, brojna novija istraživanja kombiniraju ova dva opisana pristupa i stvaraju vlastite algoritme za odabir najboljeg podskupa klasifikatora.

Kombinacija točnosti i raznolikosti korištena je u novoj mjeri WAD (engl. *Weighted Accuracy and Diversity*) u istraživanju (Zeng et al., 2014). Cilj istraživanja je bio pronaći ravnotežu između točnosti i raznolikosti s ciljem ostvarivanja boljih rezultata klasifikacije. Predložena je jedinstvena mjera koja kombinira oba elementa uz pomoć dva težinska faktora kojim se određuje bitnost pojedinog elementa. Istraživanja provedena na različitim skupovima podataka pokazuju da WAD mjera može producirati sustav prihvatljive veličine i robusnosti i da WAD ima bolje performanse od tri osnovna slučaja, tj. originalnog sustava, sustava smanjenog pomoću samo preciznosti ili raznolikosti. Sličan pristup kombiniranja performansi i raznolikosti odluka može se naći i u istraživanju autora Fu et al. (2013).

Istraživanje (Yang et al., 2013) predlaže odabir članova SVK-a, iz inicijalnog skupa klasifikatora koji koriste algoritam neuronskih mreža, na temelju nove mjere bazirane na osjetljivosti klasifikatora. Ideja istraživanja je da veća osjetljivost rezultata na ulazne vrijednosti uzrokuje i veću raznolikost produciranih odgovora. Autori treniranju 20 različitih klasifikatora te ih potom grupiraju na temelju nove mjere. Potom odabiru iz svake grupe po jedan član i pridjeljuju im težinske faktore.

Algoritam koji dodaje slučajni faktor u pohlepnu pretragu s ciljem izbjegavanja lokanog optimuma jest također zanimljiv pristup odabira podskupa SVK-a (Liu et al., 2014). Predloženi algoritam GraspEnS se bazira na GRASP (engl. *Greedy Randomized Adaptive Search Procedure*) proceduri pretrage. Razlika u odnosu na klasične pohlepne algoritme je u tome da se u prvom koraku koristi generirani slučajni faktor koji može usmjeriti odabir rješenja van lokalnog optimuma u kojem bi klasični pohlepni algoritam završio. Istraživanje je pokazalo da uključivanje slučajnosti može rezultirati s efikasnim podskupovima članova koji poboljšavaju performanse ukupnog sustava.

U istraživanju smanjivanja SVK-a autor Dai (2013.) predlaže novu mjeru za odabir članova sustava CEPCV (engl. *Competitive measure for Ensemble Pruning based on Cross-Validation*) koja se bazira na udaljenosti koju ostvaruju klasifikatori u unakrsnoj validaciji (engl. *Cross Validation*) testnog skupa. Autor za trening klasifikatora koristi *n*BBC-COP-ES (engl. *n-Bits Binary Coding ICBP Ensemble System*) sustav, a potom reducira članove pomoću algoritma koji koristi novu mjeru. Razlika novog algoritma koji koristi CEPCV tehniku u odnosu na druge jest da se izvodi u fazi izvođenja klasifikacije na testnim podacima stoga autor navodi da je to on-line smanjivanje SVK-a koje adresira problem promjene svojstava zavisne varijable (engl. *concept drift*). Novi pristup je testiran na šest skupova podataka i zaključak je da novi algoritam može značajno pozitivno utjecati na učinkovitost SVK-a. Problem dinamičkog smanjivanja SVK-a istražuju i autori (Markatopoulou et al., 2015) koji također predlažu vlastitu metodu te donose slični zaključak.

Sinteza dosadašnjih istraživanja na području smanjivanja SVK-a ukazuje na vidljivi pomak od jednostavnijih metoda za odabir podskupa, koje su se većinom bazirale na točnosti ili raznolikosti, prema metodama koje su složenije. Pristupi temeljeni na globalnoj i lokalnoj pretrazi su i danas zastupljeni u novijoj literaturi s tim da istraživači pokušavaju prevladati njihove nedostatke. Bez obzira na pristup, metodologija izrade i testiranja novih metoda je ista i čine je slijedeći koraci: (1) konstrukcija temeljnih klasifikatora na način da postižu što točnije rezultate i da im se odluke što više razlikuju, (2) razvoj evaluacijske mjere za ocjenjivanje pojedinih modela i (3) konstruiranje algoritma za odabir članova na temelju evaluacijske mjere.

3.3 Metodologija

3.3.1. Algoritam za generiranje odluka klasifikatora fiksne točnosti i raznolikosti

Generiranje većeg broja nezavisnih klasifikatora zadane točnosti $\boldsymbol{p}$ nije složen zadatak. Međutim kada je potrebno modelirati određenu raznolikost između istih, problem nije trivijalan. Bez obzira što je krajnji cilj disertacije kreiranje SVK temeljenog na upravljanom odabiru ulaznih atributa prilikom procjene kreditnog rizika građana, razvoj modela će biti proveden na teorijskom skupu podataka. Razlog zašto je istraživanje u dijelu razvoja modela provedeno na generiranom skupu podataka je proučavanje ponašanja algoritma ukoliko se neka od karakteristika mijenja npr. kao što je raznolikost. Provođenje takvih simulacija na

stvarnim podacima je teško ili gotovo nemoguće, zbog ograničenog skupa podataka kao i njihovih karakteristika koje ne zadovoljavaju neke teorijske pretpostavke, stoga je korištena simulacijska rutina (algoritam) koja kroz parametre prima karakteristike klasifikatora, a na izlazu daje matricu odluka. Na taj način se karakteristike skupova mogu kontrolirano mijenjati i proučavati.

Autori Kuncheva i Kountchev (2002) su u svom radu ponudili algoritam za generiranje odluka klasifikatora s mogućnošću specificiranja individualne točnosti (***p***), uparenih zavisnosti (***Q***) i broja primjera u skupu podataka (***N***). Predloženi algoritam pruža dobru aproksimaciju veza između klasifikatora s aspekta točnosti i raznolikosti za dovoljno veliki ***N*** i bit će korišten za generiranje odgovora klasifikatora u ovom istraživanju. Raznolikost se definira pomoću Q statistike; uparene mjere za zavisnosti klasifikatora u SVK objašnjene u poglavlju 3.4. Rezultat algoritma su binarni vektori (točna/pogrešna klasifikacija) za hipotetski skup podataka.

Uzmimo u obzir dva klasifikatora D_i i D_k, i njihove respektivne vektore odgovora y_i i y_k. Neka klasifikator D_i ima točnost A, tako da otprilike $N \times A$ elemenata vektora y_i su 1, a $N \times (1\text{-}A)$ elemenata su 0. Pretpostavimo da se svaki element 1 iz vektora y_i invertira s vjerojatnošću P_1 i svaki element 0 s vjerojatnošću P_2. Novi rezultat će biti vektor odgovora, recimo y_k, čija točnost A_{new} može biti izračunata iz A, P_1, P_2 i N. Broj jedinica y_k će biti zbroj onih koje se nisu promijenile, $N \times A \times (1\text{-} P_1)$ otprilike, i onih koje su se promijenile iz nula, $N \times (1\text{-}A) \times P_2$ (Kuncheva i Kountchev, 2002). Stoga će nova točnost biti:

$$A_{new} = \frac{1}{N}(NA(1-P_1) + N(1-A)P_2) \tag{3.1}$$

$$= (A(1-P_1) + (1-A)P_2) \tag{3.2}$$

Potrebne su vrijednosti P_1 i P_2 tako da se D_k izvede putem opisane procedure, uz uvjet da su željene točnosti $p_i = A$ i $p_k = A_{new}$ te $Q_{i,k}$ postavljeno na željenu vrijednost. Vidi se da su varijable iz tablice veza dva klasifikatora (tablica 2.2) jednake:

$$N^{11} = NA(1-P_1) \tag{3.3}$$

$$N^{10} = NAP_1 \tag{3.4}$$

$$N^{01} = N(1-A)P_2 \tag{3.5}$$

$$N^{00} = N(1-A)(1-P_2) \tag{3.6}$$

i ako se formule primjene u (formula 2.4), kroz algebarsko sređivanje dobije se:

$$Q_{i,k} = \frac{(1-P_1)(1-P_2)-P_1P_2}{(1-P_1)(1-P_2)+P_1P_2} \tag{3.7}$$

$$= \frac{1-P_1-P_2}{1-P_1-P_2+2P_1P_2}. \tag{3.8}$$

Zamijeni li se p_i=A i p_k=A_{new} u formuli 3.2 i istovremeno riješi formula 3.2 i 3.8 za P_1 i P_2, dobije se:

$$P_1 = 1 - P_2 - \frac{p_k}{p_i} - \frac{P_2}{p_i} \tag{3.9}$$

$$P_2 = \frac{-(1-Q_{i,k}+2Q_{i,k}(p_i-p_k))\pm\sqrt{Discr}}{4Q_{i,k}(1-p_i)} \tag{3.10}$$

gdje je:

$$Discr = (1 - Q_{i,k} + 2Q_{i,k}(p_i - p_k))^2 - 8Q_{i,k}(1 - p_i)p_k\big(Q_{i,k} - 1\big). \tag{3.11}$$

Za $Q_{i,k}$ = 0 tj. nezavisne klasifikatore D_i i D_k, P_1 = 1 – A_{new} i P_2 = A_{new}. Opisane formule se mogu jednostavno primijeniti za slučaj dva klasifikatora, međutim kada je potrebno modelirati zavisnosti više klasifikatora tada postupak nije tako jednostavan. Mijenjanje vrijednosti odluka jednog klasifikatora (zvanog *osnovni* klasifikator) da bi se dobile odluke drugog klasifikatora (zvanog *zavisni* klasifikator) se moraju „dijeliti“ između svih L klasifikatora. Ukoliko se prvo generira jedan osnovni klasifikator i na temelju njega zavisni klasifikator, te se tada koristi jedan od klasifikatora za generiranje novog zavisnog klasifikatora tada nema garancije da će odnos između ostalih klasifikatora koji su izuzeti biti jednaka definiranoj vrijednosti Q.

1. Ulazni parametri: $\boldsymbol{p}$ (veličine L), matrica $\boldsymbol{Q}$ (veličine LxL) i N (broj primjera u skupu).
2. Računa se $P_1(i,k)$ i $P_2(i,k)$ za sve $i,k=1,\ldots,L$, $i<>k$ prema formulama (3.9) i (3.10).
3. Za svaki j=1:N,
   a) Odaberi slučajnu permutaciju $\{i_1,\ldots,i_L\}$ od (1,2,…,L).
   b) Postavi j-tu odluku klasifikatora D_{i_1}, y_{j,i_1} na 1 s vjerojatnošću p_{i_1}, ili u suprotnom na 0.
   c) Za svaki t=2:L,
      i. Koristeći $D_{i_{t-1}}$ kao osnovni klasifikator, postavi j-tu odluku za D_{i_t}, y_{j,i_t} u skladu s vjerojatnostima $P_1(i_{t-1},i_t)$ i $P_2(i_{t-1},i_t)$
      ii. Kraj t.
   d) Kraj j.
4. Vrati $y_1,\ldots,y_L$

Slika 3.1 Pseudo kod za generiranje matrice s odgovorima klasifikatora iz vektora točnosti p i matrice raznolikosti Q prema (Kuncheva i Kountchev, 2002, slika 1.).

Da bi se izbjegao navedeni problem za svaki element skupa se generira slučajna permutacija brojeva od 1 do L, koja se koristi prilikom odabira osnovnih i zavisnih klasifikatora. Algoritam za generiranje L klasifikatora s fiksnom točnošću i definiranom raznolikošću između članova je prikazan na slici 3.1. U prilogu B ove disertacije dana je autorova implementacija navedenog algoritma u programu MatLab. Funkcija prima opisane ulazne parametre i vraća matricu odgovora svih klasifikatora.

3.3.2. Pohlepni algoritmi za smanjivanje sustava višestrukih klasifikatora

Pohlepni algoritmi nastoje pronaći najbolje globalno rješenje pohlepnim odabirom slijedećeg klasifikatora prilikom proširivanja podskupa (Partalas et al. 2012). Pohlepni odabir podrazumijeva odabir jednog smjera bez razmatranja alternativa koje se istovremeno odbacuju. Pohlepni algoritmi tipično sadrže pet komponenti (Miglani & Rana, 2011):

1. skup kandidata, iz kojih se rješenje stvara,
2. funkciju odabira, koja odabire najboljeg kandidata u rješenje,
3. funkciju izvodljivosti, koja procjenjuje da li kandidat može doprinijeti rješenju (evaluacijska tehnika),
4. funkciju objektivnosti, koja dodjeljuje vrijednost rješenju ili parcijalnom rješenju i
5. funkciju rješenja, koja indicira da je algoritam pronašao konačno rješenje.

Pohlepni algoritmi produciraju dobra rješenja na nekim problemima, ali ne na svima. Većina problema na kojima postižu uspješne rezultate posjeduju dva svojstva: (1) svojstvo pohlepnog odabira i (2) optimalnu sub-strukturu (Cormen, 2009). Prvo ključno svojstvo jest svojstvo pohlepnog odabira tj. slučaj kada se globalno optimalno rješenje može postići odabirom lokalno optimalnih rješenja. Drugim riječima, kada se razmatra koji odabir napraviti, odabire se uvijek onaj koji izgleda najbolji za trenutni problem, ne razmatrajući rezultate pod-problema. Stoga odluka pohlepnog algoritma može zavisiti o odlukama napravljenima u povijesti ali nikako o onim odlukama u budućnosti. Drugo svojstvo jest optimalna sub-struktura problema. Neki problem ima optimalnu sub-strukturu ukoliko optimalno rješenje unutar sebe sadrži i optimalna rješenja pod-problema.

Iako pohlepni algoritmi iscrpno ne pretražuju područje pretrage i moguće je da ih neki raniji odabir usmjeri dalje od optimalnog rješenja, u slučajevima gdje je moguće dokazati da ostvaruju optimalna rješenja često su prvi izbor zbog svoje brzine.

Pohlepni algoritmi se često koriste za smanjivanje SVK-a što je vidljivo iz poglavlja 3.2 tj. pregleda literature. U nastavku će biti opisani neki od elemenata algoritma bitni za istraživanje: smjer pretrage algoritma, korištena evaluacijska tehnika i veličina konačnog SVK-a.

3.3.3. Smjer pretrage

Pohlepni algoritmi se na temelju smjera pretrage mogu podijeliti u dvije grupe: algoritmi s odabirom unaprijed i s eliminacijom unazad (Liu et al., 2014).

Algoritmi s odabirom unaprijed započinju stvaranje SVK-a tako što skup klasifikatora S inicijaliziraju kao prazan skup te inkrementalno u svakom koraku dodaju novi član. Iterativno se odabire jedan klasifikator h_t koji optimizira evaluacijsku funkciju i dodaje se skupu $S_{NEW} = S \cup h_t$. Na primjer, evaluacijska funkcija može mjeriti točnost skupa $S \cup h_t$, gdje bi se odabrao onaj h_t koji maksimizira spomenutu mjeru. Algoritmi s odabirom unazad funkcioniraju obrnuto tj. skup S se inicijalizira sa svim članovima sustava SVK-a te se u svakom koraku odbacuje jedan član. Slično kao i u prethodnoj grupi iterativno se odabire jedan klasifikator h_t koji optimizira evaluacijsku funkciju i odbacuje se iz skupa $S_{NEW}=S\setminus\{ h_t \}$.

U ranijim istraživanjima smanjivanja SVK-a (Li et al., 2012) je zaključeno da algoritmi s odabirom unaprijed ostvaruju bolje rezultate u odnosu na druge. Iako se rezultati po pitanju performansi ne razlikuju ovisno o smjeru pretrage, algoritmi koji koriste odabir unaprijed često stvaraju manje skupove klasifikatora.

3.3.4. Evaluacijska tehnika

Evaluacijska tehnika je glavna komponenta pohlepnog algoritma jer ona određuje smjer pretrage. Ovisno o rezultatima dobivenim iz evaluacijske tehnike algoritam odlučuje koji će biti slijedeći odabir iz skupa svih dostupnih alternativa. Evaluacijska tehnika je u suštini funkcija koja ocjenjuje različite smjerove tijekom pretrage prostora rješenja. Ako promatramo SVK prezentiran skupom S i potencijalnim kandidatom h_t za ulazak u skup, tada funkcija ocjenjuje korisnost uključivanja (ili izbacivanja) kandidata h_t u skup S, na temelju dostupnog skupa podataka (Partalas et al. 2012). Podjela evaluacijskih tehnika kao i prednost i nedostatci pojedinih pristupa su opisani ranije u pregledu literature.

3.3.5. Veličina konačnog SVK

Prilikom odabira klasifikatora u sustav klasifikatora pojavljuje se problem odabira najbolje veličine novog SVK-a. Algoritmi inkrementalno dodaju (ili izbacuju) nove klasifikatore stoga je pitanje kada je najbolje zaustaviti proces pretrage. Odgovor na to pitanje ovisi da li se od novog sustava više očekuje poboljšanje efikasnosti ili performansi predviđanja. U prvom slučaju, ako je glavni cilj efikasnost novog SVK-a tada se broj članova unaprijed zadaje kao postotak originalnog sustava. Takav pristup je korišten u radovima (Martınez-Munoz i Suárez, 2004; Banfield et al., 2005). U suprotnom, ako je cilj poboljšati performanse sustava tada se algoritam izvršava do kraja (dok nisu svi članovi iz originalnog sustava uključeni) te se bira onaj podskup koji maksimizira zadanu mjeru performansi npr. točnost. U radu (Caruana et al., 2004) je primijenjen takav pristup.

3.4 Dizajn eksperimenta

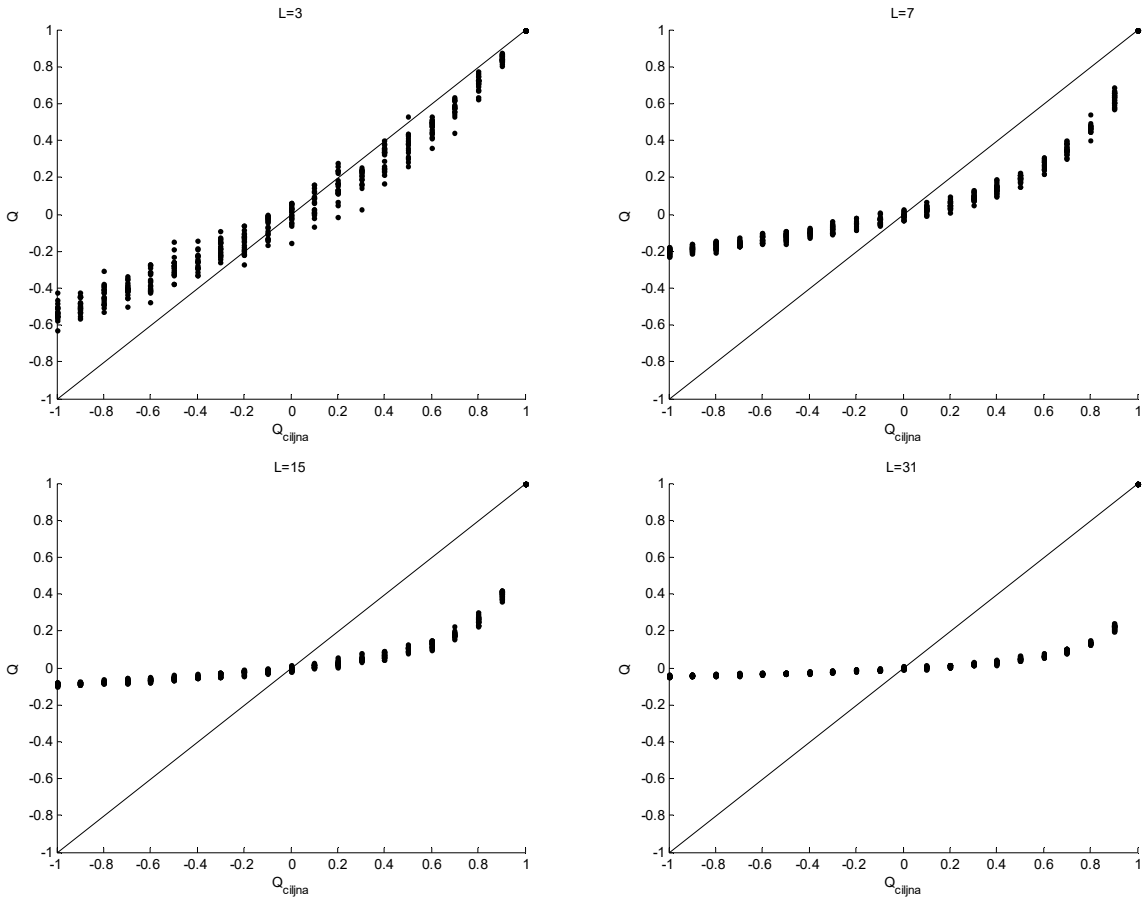
Prikaz eksperimenta je podijeljen u dva koraka; u prvom koraku se opisuje dizajn i analiziraju postavke modificiranog algoritama za generiranje odluka klasifikatora fiksne točnosti i raznolikosti kojim se kreira skup podataka na kojemu će se testirati novi algoritam, a u drugom koraku se opisuje konstrukcija novog algoritma COP za smanjivanje SVK i mjera za odabir članova u podskup rješenja, koeficijent uspješnosti klasifikatora (CO).

3.4.1. Generiranje skupova podataka

3.4.1.1. Modificirani algoritam za generiranje odluka klasifikatora fiksne točnosti i raznolikosti

Performanse algoritma za generiranje odluka klasifikatora, u originalnom radu (Kuncheva i Kountchev, 2002) su prezentirane na malom broju klasifikatora. Najbolje rezultate, u smislu ostvarene točnosti i raznolikosti u odnosu na zadane, algoritam postiže ukoliko se ciljani skup sastoji od dva klasifikatora (L=2). Međutim, uspješnost algoritma pada s porastom L, što se očituje u ostvarenoj vrijednosti Q koja s rastom broja klasifikatora konvergira nuli tj. neovisnosti. Grafički prikaz kretanja Q mjere za skupove veličine 3, 7, 15 i 31 klasifikatora je dan na slici 3.2. Skupovi prikazani na slici su generirani 20 puta s veličinom skupa N=1000,

točnost klasifikatora je p=0.9, za svaki od mjera raznolikosti od -1 do 1 s korakom promjene 0.1.



Slika 3.2 Odstupanje Q od Q_{ciljne} vrijednosti za skupove veličine 3, 7, 15 i 31, korištenjem algoritma prikazanog na slici 3.3.

Slika 3.2 prikazuje da s povećanjem broja klasifikatora raste i odstupanje od idealnog rezultata, prezentiranog dijagonalnom linijom na grafu. Vidljivo je da vrijednosti Q s porastom L, osim što se približavaju nuli, imaju i manju disperziju rezultata. Autori, Kuncheva i Kountchev, (2002) su prepoznali nedostatak predloženog algoritma koji se očituje na sustavima s većim bojem klasifikatora i predložili strategiju odabira samo onih skupova koji su ostvarili dobre rezultate dok bi se ostali skupovi odbacili. Međutim takva strategija obzirom na odstupanja dobivenih Q vrijednosti u odnosu na ciljne vrijednosti za $L > 3$ ne daje zadovoljavajuće rezultate.

Da bi se otklonio opisani nedostatak u nastavku je predložen poboljšan algoritam koji pomoću korekcijskog faktora ispravlja konvergenciju rezultata prema 0 za veće vrijednosti L. Korekcijski faktor se primjenjuje na vrijednosti P_2 u zavisnosti od broja klasifikatora (L) i ciljne mjere raznolikosti (Q).

1. Ulazni parametri: $\boldsymbol{p}$ (veličine L), matrica $\boldsymbol{Q}$ (veličine LxL) i N (broj primjera u skupu).
2. Za svaki $i,k=1,\ldots,L$, $i<>k$ računa se $P_1(i,k)$ i $P_2(i,k)$ prema formulama (3.9) i (**3.13**).
3. Za svaki j=1:N,
   a) Odaberi slučajnu permutaciju $\{i_1,\ldots,i_L\}$ od $(1,2,\ldots,L)$.
   b) Postavi j-tu odluku klasifikatora D_{i_1}, y_{j,i_1} na 1 s vjerojatnošću p_{i_1}, ili u suprotnom na 0.
   c) Za svaki t=2:L,
      i. Koristeći $D_{i_{t-1}}$ kao osnovni klasifikator, postavi j-tu odluku za D_{i_t}, y_{j,i_t} u skladu s vjerojatnostima $P_1(i_{t-1},i_t)$ i $P_2(i_{t-1},i_t)$
      ii. Kraj t.
   d) Kraj j.
4. Vrati $y_1,\ldots,y_L$

Slika 3.3 Pseudo kod poboljšanog algoritma za generiranje matrice s odgovorima klasifikatora iz vektora točnosti p i matrice raznolikosti Q

Modificirani (poboljšani) algoritam se koristi za skupove gdje je $L > 2$. Korekcijski faktor $C(L,Q)$ se računa pomoću formule:

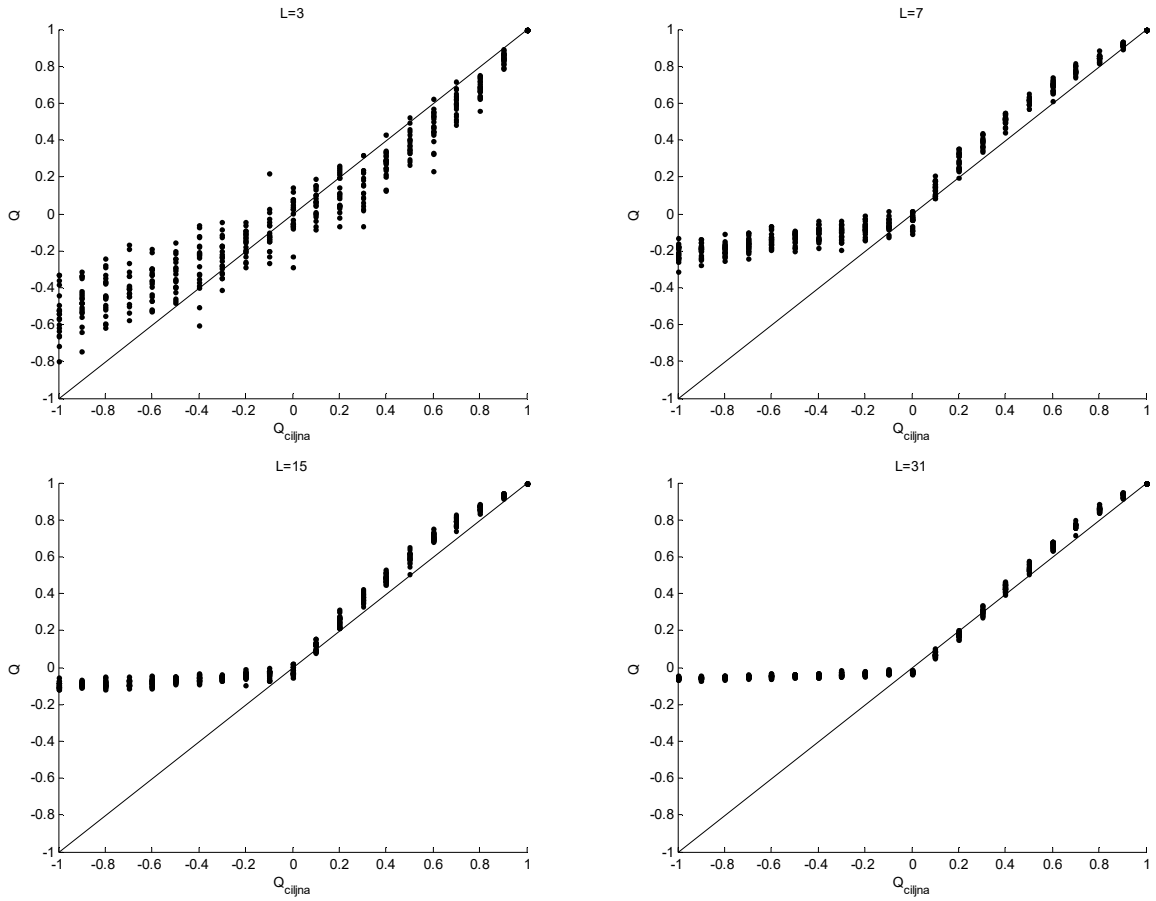
$$C(L,Q) = \left(\frac{3}{L}\right)^{\frac{|Q|+Q}{2}} \tag{3.12}$$

Korekcijski faktor se primjenjuje samo na P_2 tako da ga umanjuje ovisno o L i Q; što ciljni skup ima više klasifikatora i utjecaj faktora će biti značajniji. Rezultati za skup podataka od L=3 su isti u odnosu na originalni algoritam. Osim na P_2 faktor posredno utječe i na vrijednost P_1 koja također pada s porastom varijable L. Smanjenje navedenih vrijednosti dovodi do bolje kontrole raznolikosti između klasifikatora što je i bio cilj modificiranja algoritma. Eksponent korekcijskog faktora poništava utjecaj na Q vrijednosti koje su negativne, stoga one ostaju iste u odnosu na originalni algoritam. Modificiranim algoritmom se ne mogu precizno modelirati Q vrijednosti manje od 0. Da bi se utjecalo na skupove za koje vrijedi $Q < 0$ potrebno je dodati korekcijski faktor i na P_1 koji bi korigirao vrijednosti samo za Q manji od nula. Međutim, vrijednosti $Q < 0$ nisu predmet ovog istraživanja jer na stvarnim primjerima na kojima se primjenjuje ovaj algoritam ne postoje klasifikatori koji imaju negativni Q.

Prema opisanom postupku uzme li se formula (3.10) za P_2 i formula (3.12) za izračun korekcijskog faktora, nova formula za P_2 je sljedeća:

$$P_2 = \frac{-(1-Q_{i,k}+2Q_{i,k}(p_i-p_k))\pm\sqrt{Discr}}{4Q_{i,k}(1-p_i)}\ C\left(L,Q_{i,k}\right) \tag{3.13}$$

Na slici 3.4 je prikazano odstupanje Q od Q_{ciljne} vrijednosti za skupove veličine 3, 7, 15 i 31 klasifikatora, korištenjem novog modificiranog algoritma. Parametri generiranih skupova su ostali isti kao i kod slike 3.2, tj. skupovi prikazani na slici su generirani 20 puta s veličinom skupa $N = 1000$. Parametar točnost je za sve pojedinačne klasifikatore isti; $p = 0.9$, za svaki od mjera raznolikosti od -1 do 1 s korakom promjene 0.1.



Slika 3.4 Odstupanje Q od Q_{ciljne} vrijednosti za skupove veličine 3, 7, 15 i 31 klasifikatora, korištenjem algoritma prikazanog na slici 3.3

Iz slike 3.4 je vidljivo da generirani skupovi na temelju modificiranog algoritma daju značajno bolje rezultate za Q vrijednost te da su isti bliže optimalnoj liniji za $Q_{ciljna} \geq 0$. Ostali rezultati gdje je $Q_{ciljna} < 0$ su očekivano isti. U biti, rezultati prikazani na slici 3.2, a dobiveni algoritmom prikazanim na slici 3.3 (Kuncheva i Kountchev, 2002) nisu bili iskoristivi za skupove veće od 3 člana. Ovako generirane matrice s odgovorima klasifikatora (slika 3.4) mogu značajno unaprijediti dalja teorijska i praktična istraživanja na području kreiranja i smanjivanja većih SVK-a.

U slijedećem odjeljku će biti analizirani parametri za generiranje teorijskih skupova podataka na kojima će se testirati novi algoritam za smanjivanje SVK-a. Obzirom da će biti

promatrane samo vrijednosti Q koje se nalaze u intervalu [0.7, 0.9] za skup od 31 klasifikatora, modifikacija predloženog algoritma samo za $Q_{ciljna} \geq 0$ jest opravdana. Generirani skupovi će biti analizirani u odjeljku 3.5.1.

3.4.1.2. Parametri za generiranje skupa podataka

Autori Nanni i Lumini (2009) su objavili istraživanje u kojem su istražili performanse nekoliko modela baziranih na SVK za predviđanje bankrota i procjenu kreditnog rizika. U istraživanju su koristili tri skupa podataka među kojima je i njemački skup podataka koji će biti korišten u ovoj disertaciji. Između nekoliko pristupa istražen je i pristup slučajnog odabira ulaznih atributa koji je po autorima dao najbolje rezultate klasifikacije. Osim točnosti i AUC (engl. *the area under the Receiver Operating Characteristic curve*) mjere, u radu su dali i mjeru Q statistike za raznolikost članova za sve SVK, po svim skupovima. Prosječna Q vrijednost za SVK konstruirane temeljem slučajnog odabira ulaznih atributa na njemačkom skupu kreditnih podataka je $Q = 0.795 \pm 0.098$. Točnost klasifikacije na istim primjerima je $p = 0.719 \pm 0.014$.

Q vrijednost i točnost klasifikacije na njemačkom skupu podataka mjerili su i autori Banfield et al. (2003) prilikom konstruiranja nove mjere PCDM (engl. *Percentage Correct Diversity Measure*) za raznolikost klasifikatora SVK. Izmjerena točnost SVK nakon smanjivanja jest $p = 0.7102$, a mjera raznolikosti $Q = 0.85$.

Iako se spomenuta istraživanja razlikuju od postavljenog u ovoj disertaciji, karakteristike odluka klasifikatora unutar SVK na kreditnom skupu podataka su korisne prilikom generiranja odluka teoretskih klasifikatora. Iznesena točnost i raznolikost će poslužiti za postavljanje ulaznih parametara funkcije za generiranje odluka klasifikatora. Da bi se modelirala osjetljivost algoritma na raznolikost klasifikatora, generirane odluke klasifikatora će imati tri vrijednosti Q mjere za raznolikost odluka; $Q_{i,k} = q,\ q \in \{0.7, 0.8, 0.9\}$ i četiri različite vrijednosti točnosti $p \in \{0.65, 0.70, 0.75, 0.8\}$. Veličina skupa je N=1000 primjera, a eksperiment će se ponoviti 100 puta s istim postavkama koje su sistematizirane u tablici 3.1.

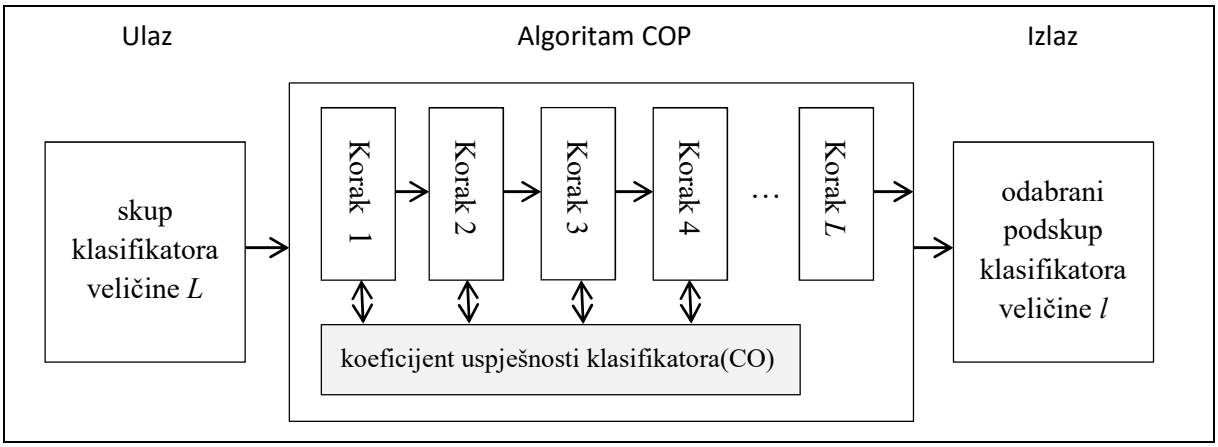
Tablica 3.1 Postavke eksperimenta

RB	Postavka	Oznaka	Vrijednost
1	broj testova	T	100
2	broj primjera	N	1000
3	točnost klasifikatora	p	{0.65, 0.70, 0.75, 0.8}
4	raznolikost klasifikatora	Q	{0.7, 0.8, 0.9}
5	broj klasifikatora	L	31

Analiza generiranih rezultata je obavljena u odjeljku 3.5.1.

3.4.2. Konstrukcija COP algoritma za smanjivanje SVK-a

Cilj istraživanja prezentiranog u ovom poglavlju je razviti vlastiti algoritam za smanjivanje SVK koji će u funkciji odabira koristiti novu metriku pod nazivom koeficijent uspješnosti klasifikatora (CO). U uvodu je već naglašeno da se problem smanjivanja SVK može definirati kao optimizacijski problem, odnosno problem pronalaska jednog podskupa dostupnih modela, članova sustava, koji optimizira mjeru koja indicira performanse generalizacije. Odabir podskupa u algoritmu COP se izvodi pomoću CO metrike. COP je u suštini funkcija koja na ulazu prima rezultate klasifikacije svih temeljnih klasifikatora, a na izlazu daje odabrani podskup kao smanjeni SVK. Postupak odabira je iterativan, svaka iteracija prima trenutno odabrani podskup SVK-a (kojeg promatra kao jedan klasifikator) i potencijalne članove koji još nisu uključeni u sustav. Na izlazu svaki potencijalni član dobiva mjeru performansi temeljem koje se odabire sljedeći član koji ulazi u podskup. Na slici 3.5 dan je pojednostavljen prikaz algoritma COP na kojem je vidljivo u kojem su odnosu algoritam i CO metrika za odabir novih članova sustava.



Slika 3.5 Pojednostavljen prikaz COP algoritma

Opis dizajna novog algoritma, radi jednostavnosti je podijeljen u dva dijela: (1) dizajn mjere CO za izračun koeficijenta uspješnosti klasifikatora temeljem koje se vrši odabir novih članova u sustav i (2) dizajn algoritma COP temeljenog na CO za smanjivanje SVK-a. U nastavku će se koristiti slijedeća notacija: originalni SVK će biti notiran kao $H = \{h_t, t = 1,2,\dots,L\}$, gdje je L neparni broj. Trenutni SVK je označen sa *S,* za koji vrijedi $S \subseteq H$, a $S' = H \setminus S$. S se u procesu odabira optimalnog podskupa klasifikatora inkrementalno povećava za novog člana i nastaje novi SVK oznake S_{NEW}, veličine l, $l = |S_{NEW}|$ za koji vrijedi $S_{NEW} =$

$S \cup h_t$, $h_t \in H \setminus S$. Skup podataka na kojem se provodi klasifikacija je notiran s $D = \{(x_i, y_i), i = 1,2, \dots, N\}$ gdje je x_i vektor s vrijednostima atributa, a y_i ispravna vrijednost klase za dani vektor x_i, tj. $y_i = f(x_i) \wedge y_i \in \{-1,1\}$. Uz navedenu notaciju vrijedi ukoliko je $h_t(x_i) = y_i$ tada je klasifikator t ulazne atribute x_i klasificirao u ispravnu klasu. Skup podataka za klasifikaciju D, bez atributa koji označava klasu, notiran je sa $X = \{x_i, i = 1,2, \dots, N\}$, gdje je svaki x_i, kao što je ranije navedeno, vektor s vrijednostima prediktorskih atributa.

3.4.3. Koeficijent uspješnosti klasifikatora (engl. *Classifier Odds – CO*)

3.4.3.1. Definicija mjere koeficijent uspješnosti klasifikatora

U osnovi, koeficijent uspješnosti klasifikatora (CO) je mjera koja izražava dobrotu nekog SVK u odnosu na točnost klasifikacije svih primjera iz skupa podataka za klasifikaciju. To je mjera za kvantifikaciju korisnosti uključivanja novih klasifikatora u SVK temeljena na vrednovanju težine primjera. CO se izračunava kao suma pojedinačnih koeficijenata izraženih za svaki primjer iz skupa podataka za klasifikaciju. Ako je većina članova SVK-a ispravno klasificirala neki primjer x_i tada je mjera CO za taj primjer pozitivna i obrnuto. Za neki primjer x_i, $CO(x_i)$ može ostvariti slijedeće vrijednosti:

$$\mathrm{CO}(x_i) = \begin{cases} < 0, & \textit{većina članova je pogrešno klasificirala } x_i \\ 0, & \textit{pola članova je ispravno klasificialo } x_i \\ > 0, & \textit{većina članova je ispravno klasificirala } x_i \end{cases}$$

$CO(x_i)$ se može formalno definirati na slijedeći način: Ako sa C označimo odluku sustava višestrukih klasifikatora koji odlučuje jednostavnim glasanjem članova, tada se odluka SVK-a za slučaj x_i može prikazati kao:

$$C(x_i) = sign\left(\sum_{t=1}^{L} h_t(x_i)\right) \tag{3.14}$$

gdje je $C(x_i) \in \{-1,1\}$, jer je, kao što je ranije navedeno, L neparni broj. Kada se odluka SVK-a za slučaj x_i ponderira s ispravnom vrijednosti klase, tj. sa y_i, za dani slučaj x_i, dobije se odgovor kako je sustav tj. većina članova SVK-a klasificirala slučaj x_i:

$$O(x_i) = C(x_i) y_i, \tag{3.14a}$$

gdje je $O(x_i) = 1$ ako je sustav točno klasificirao instancu x_i i $O(x_i) = -1$ ako je sustav netočno klasificirao instancu x_i.

S obzirom da izračun ukupne mjere CO kreće od pojedinačnih slučajeva za klasifikaciju x_i te od temeljnih klasifikatora iz SVK, potrebna nam je formula koja izražava odnos između temeljnog klasifikatora h_t i ispravne klase za slučaj x_i:

$$O_t(x_i) = h_t(x_i)y_i, \tag{3.15}$$

gdje je $O_t(x_i) = 1$ ako je klasifikator h_t točno klasificirao instancu x_i i $O_t(x_i) = -1$ ako je klasifikator h_t netočno klasificirao instancu x_i.

Sada se za jedan primjer iz skupa *X*, tj. za x_i, može definirati koeficijent uspješnosti za skup S_{NEW}, za koji vrijedi $S_{NEW} = S \cup h_t,\ h_t \in H \setminus S$, pomoću formule:

$$CO_{S_{NEW}}(x_i) = \sum_{t=1}^{l} h_t(x_i)\, y_i Coef(x_i)$$

ili

$$CO_{S_{NEW}}(x_i) = \sum_{t=1}^{l} O_t(x_i)\, Coef(x_i) \tag{3.16}$$

gdje $O_t(x_i)$ izražava odnos klasifikatora h_t spram ispravnosti klasifikacije instance x_i, a $Coef(x_i)$ određuje koliko klasifikator h_t doprinosi postojećem SVK u povećanju ili smanjenju CO vrijednosti za primjer x_i, ukoliko predmetni klasifikator h_t proširi *S* kao njegov slijedeći član. Neka je $P(x_i)$ broj klasifikatora iz skupa *S* koji su ispravno klasificirali primjer x_i i neka je $N(x_i)$ broj onih koji su negativno klasificirali x_i tada se *Coef* računa:

$$Coef(x_i) = \frac{1}{|P(x_i) - N(x_i)| + \left(\frac{(1 + (\, h_t(x_i)\, y_i\, sign(P(x_i) - N(x_i)\,))}{2^{sign(|P(x_i) - N(x_i)|)}}\right)}. \tag{3.17}$$

Nakon što se izračuna CO za sve primjere x_1, x_2, x_3, … , x_N iz skupa *X*, ukupni CO nekog klasifikatora, npr. S_{NEW}, se računa kao suma svih $CO(x_i)$ predmetnog klasifikatora S_{NEW}:

$$CO_{S_{NEW}} = \sum_{i=1}^{N}\sum_{t=1}^{l} h_t(x_i)\, y_i Coef(x_i)$$

ili

$$CO_{S_{NEW}} = \sum_{i=1}^{N}\sum_{t=1}^{l} O_t(x_i)\, Coef(x_i) \tag{3.18}$$

Tako izračunat CO predstavlja jednu sveobuhvatnu mjeru koja se koristi za izražavanje opravdanosti uključivanja klasifikatora h_t u skup *S*.

Na temelju prezentiranih formula konstruirana je funkcija koja vrši izračun CO-a. Funkcija na ulazu prima dva parametra: (1) odluke pojedinačnih klasifikatora iz S za sve primjere iz *D* (2) i odluke potencijalnih kandidata h_t za koje vrijedi $h_t \in H \setminus S$ ili $h_t \in S'$ za sve primjere iz *D*. Ukoliko je $S = \emptyset$ tada se kao prvi parametar prosljeđuje vektor veličine N s vrijednostima 0 na svim pozicijama. Rezultat funkcije su vektori veličine N koji sadrže pojedinačne koeficijente uspješnosti klasifikatora $CO(x_i)$ za sve primjere iz N i sve moguće podskupove klasifikatora $S \cup h_t$, $\forall\, h_t \in S'$. Stoga je konačni rezultat matrica veličine $|S'| \times N$.

Iz ulaznih parametara je vidljivo da se uvijek na ulazu funkcije dostavlja skup *S*, time se koeficijent uspješnosti klasifikatora inkrementalno gradi kako se klasifikatori dodaju u skup. Algoritam za smanjivanje SVK-a koji primi odgovor iz funkcije za izračun CO može odabrati bilo koji stupac (klasifikator) kao slijedeći član, ovisno o svojoj strategiji. Nakon toga on pohranjuje svoj odabir i prosljeđuje odabrani S_{NEW} kao novi skup S u slijedećem pozivu funkcije CO. Na taj način se izračunava CO u zavisnosti kako novi članovi ulaze u SVK.

Drugi parametar na ulazu funkcije je potreban da bi svi potencijalni kandidati (skup *S′*) bili dostupni funkciji za izračun CO-a. Funkcija računa CO za svaki S_{NEW}, koji se sastoji od starog skupa S uvećanog za jedan klasifikator iz *S′*. Ukoliko algoritam odabire samo jedan član na temelju CO-a, skup *S* se povećava, a skup S' smanjuje u svakoj iteraciji za jedan član (stupac).

Iako se odabrani članovi sustava u konačnici kombiniraju jednostavnim glasanjem (engl. *voting*) gdje svi imaju jednaku važnost, prilikom računanja CO-a njihove odluke imaju različitu težinu, što je vidljivo iz formule (3.18). Koliko će koji klasifikator povećati ili smanjiti CO određuje koeficijent *Coef* iz formule (3.17). Detaljniji opis načina na koji se vrednuju odluke klasifikatora dan je u sljedećem odjeljku.

Pseudo kod za programsku implementaciju matrice CO-a prikazan je na slici 3.6. Unutar dvije petlje računaju se CO vrijednosti prema ulaznim parametrima. Rezultat je matrica čiji stupci predstavljaju klasifikatore koji još nisu uključeni u smanjeni sustav višestrukih klasifikatora, a redovi sve primjere iz skupa *D*. Na temelju te matrice COP algoritam odlučuje o odabiru novog člana SVK-a, odnosno, o sastavu skupa *S* za idući krug.

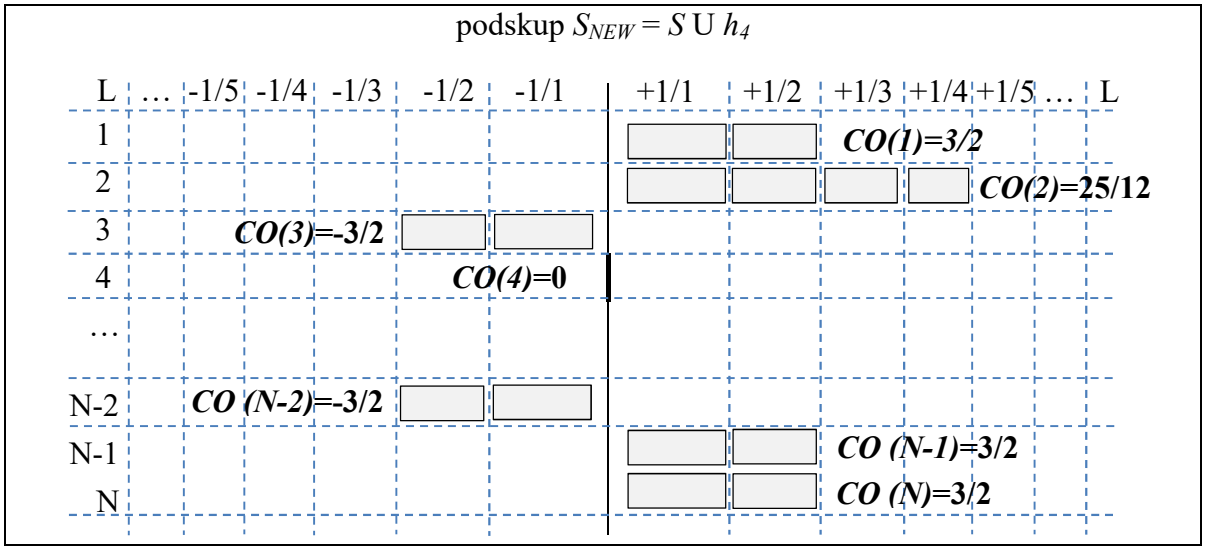
```
1. Ulaz: matrica M (veličine N x |S|), matrica M' (veličine N x |S'|)
   Ako je |S|=0 ⇒ M veličine N x 1, ∀i=1:N: M_{1,i} = 0
2. za svaki a = 1:|S'|
        a. za svaki i=1:N
                računaj T_{i,a} = CO_{S_NEW}(x_i) prema formuli(3.16)
        b. kraj i
3. kraj a
4. Izlaz: T (veličine N x |S'|)
```

Slika 3.6 Pseudo kod za računanje matrice pojedinačnih CO-a za sve potencijalne klasifikatore

3.4.3.2. Način vrednovanja odluka klasifikatora

Da bi se objasnio način na koji se vrednuju odluke klasifikatora potrebno je definirati pojam koeficijenta uspješnosti. Koeficijent uspješnosti klasifikatora CO odnosi se na izglede klasifikatora da klasificira neki primjer u ispravnu klasu. Primjeri koji su klasificirani od većeg broja klasifikatora ispravno imaju pozitivan utjecaj i obrnuto, primjeri pogrešno klasificirani od više klasifikatora imaju negativan utjecaj na ukupnu uspješnost klasifikatora. Pojednostavljeno, proces izračuna $CO(x_i)$ se može promatrati na slijedeći način; ako svaka ispravna klasifikacija dodaje jedan, a neispravna oduzima vrijednost jedan od koeficijenta, tada se $CO(x_i)$ za skup klasifikatora veličine 5, tj. $|S|$=5, za neki primjer x_i, može kretati u intervalu [-5, 5]. Ukoliko su svi klasifikatori klasificirali x_i ispravno tada je $CO(x_i)$ = 5, a obratno $CO(x_i)$ = -5. Kako se novi klasifikator uključuje u skup S tako se $CO(x_i)$ mijenja ovisno o odlukama novog klasifikatora, te ga on može smanjiti ukoliko vrijedi $h_t(x_i) \neq y_i$ ili povećati u slučaju $h_t(x_i) = y_i$.

Postoji nekoliko razlika između prethodnog, pojednostavljenog primjera u odnosu na stvarni izračun CO-a u kojemu svi klasifikatori ne donose jednaku promjenu koeficijenta. U svrhu objašnjavanja izračuna CO-a razmotrimo slijedeći primjer: 15 različitih klasifikatora čine originalni SVK koji je predmet smanjivanja; primjenom COP algoritma, koji koristi CO, odabrano je pet članova S={h_1, h_2, h_3, h_9, h_{15}} koji ulaze u smanjeni SVK; računa se CO za preostale klasifikatore iz S' s ciljem odabira šestog člana. Na slici 3.7 slikovito je prikazano računanje CO-a za podskup S_{NEW} u koji je dodan klasifikator h_4 kao šesti član SVK-a.



Slika 3.7 Grafički prikaz računanja CO-a

Najveća vrijednost promjene CO-a koju jedan potencijalni član h_t može ostvariti na jednom slučaju x_i za klasifikaciju je $\Delta CO(x_i) = \pm 1$. Iz slike je vidljivo da se $max\ \Delta CO(x_i)$ može ostvariti samo ako je $CO_s(x_i)$=0. U tom slučaju pola klasifikatora u skupu S ispravno klasificira slučaj x_i i svaki potencijalni član koji ispravno klasificira takav primjer je maksimalno nagrađen jer direktno utječe na ispravnu klasifikaciju. Istom logikom, ako je $h_t(x_i) \neq y_i$ tada je klasifikator h_t maksimalno „kažnjen“ za -1.

Manju vrijednost $\Delta CO(x_i)$ ostvaruje klasifikator na primjerima koji imaju vrijednost $CO_s(x_i) \neq 0$. Koliku promjenu vrijednosti će ostvariti klasifikator h_t ovisi o vrijednosti $CO_s(x_i)$, tj. o vrijednosti CO(x_i) za ulazni SVK. Što je veća vrijednost $|CO_s(x_i)|$, to je manji značaj odluke novog klasifikatora koji ulazi u skup; npr. ako su svi članovi iz skupa S ispravno klasificirali promatrani primjer x_i, tada šesti član koji se dodaje, ako također ispravno odredi klasu donosi 1/6 vrijednosti. Međutim, ukoliko šesti član neispravno procjeni klasu tada se od zbroja oduzima 1/5 ili iznos koji je donio zadnji klasifikator iz skupa S. Na isti način klasifikator koji je negativno klasificirao primjer uvijek oduzima istu vrijednost koju je zadnji ispravni klasifikator donio, pod pretpostavkom da je zadnji ispravni klasifikator neposredno prethodio. Slikovito rečeno, briše jedan pravokutnik iz reda za slučaj x_i (slika 3.7). Isto pravilo vrijedi obrnuto, ukoliko je $CO_s(x_i) < 0$. Pri tome, ako u skupu S ima članova, 0 označava da je jedak broj klasifikatora ispravno i neispravno klasificirao primjer, što je vidljivo na primjeru 4 slika 3.7.

U suštini, temeljni klasifikator koji ispravno klasificira primjere koje postojeći članovi SVK-a ne klasificiraju jednoznačno postiže veću vrijednost CO-a u odnosu na one klasifikatore koji ispravno klasificiraju isti broj primjera koji su već jednoznačno klasificirani,

bilo pozitivno ili negativno. Na taj način se jače vrednuju klasifikatori koji su uspješniji na težim primjerima.

Primjer 1. Razmotrimo binarne vektore odluka klasifikatora z_1=[-1 1 1 1 1 1]^T, z_2=[1 1 -1 1 1 1]^T, z_3=[1 -1 1 -1 1 -1]^T veličine 6, kao odluke klasifikatora h_1, h_2 *i* h_3 respektivno. Neka vrijedi da je $z_t(x_i)$ = 1 ako je $h_t(x_i)=y_i$, gdje je y_i ispravna vrijednost klase za instancu x_i, i obrnuto, $z_t(x_i)$ = -1 ako je $h_t(x_i)<>y_i$. Uzmimo da je skup $S = \emptyset$ stoga je CO_S=[0 0 0 0 0 0]^T . U nastavku je dan izračun CO-a za ulazne vektore *z* prema pseudo kodu na slici 3.6 s ciljem pronalaska prvog člana skupa *S*.

Tablica 3.2 Matrica S' - ulaz u funkciju za izračun CO-a

X	z_1	z_2	z_3
1	-1	1	1
2	1	1	-1
3	1	-1	1
4	1	1	-1
5	1	1	1
6	1	1	-1
∑	*4*	*4*	*0*

Tablica 3.3 Matrica - rezultat funkcije za izračun CO-a

X	CO(h_1)	CO(h_2)	CO(h_3)
1	-1,00	1,00	1,00
2	1,00	1,00	-1,00
3	1,00	-1,00	1,00
4	1,00	1,00	-1,00
5	1,00	1,00	1,00
6	1,00	1,00	-1,00
∑	*4*	*4*	*0*

Funkcija za CO kao odgovor vraća matricu iz tablice 3.3, na temelju koje COP algoritam sukladno svojoj strategiji odabire jedan klasifikator kao novi član sustava.

Iz primjera 1. (tablica 3.3) vidljiv je rezultat izračuna CO-a za tri klasifikatora kada nije izabran još niti jedan član u sustav. Može se uočiti da je, u uvjetima kada još nema niti jednog člana u S, |S| = 0, svaka promjena vrijednosti CO(x_i) maksimalna, bilo pozitivna, bilo negativna. Navedeno rezultira time da je ukupni CO na razini klasifikatora jednak razlici točnih i netočnih odluka klasifikatora. Ovakav rezultat CO algoritma svojstven je samo odabiru prvog člana u sustav S. Sljedeći primjer to jasno demonstrira.

Primjer 2. Uzmimo druge binarne vektore z_3=[1 1 -1 1 1 -1]^T, z_4=[1 1 1 -1 -1 1]^T, z_5=[-1 1 1 -1 1 -1]^T veličine 6, kao odluke klasifikatora h_3, h_4 *i* h_5 respektivno. Neka vrijedi da je $z_t(x_i)$ = 1 ako je $h_t(x_i)=y_i$, gdje je y_i ispravna vrijednost klase za instancu x_i, i obrnuto $z_t(x_i)$ = -1. Definirajmo da je skup $S = \{h_1, h_2\}$ te CO_S=[0 1.5 0 1.5 1.5 1.5]^T i |S|=2. U tablicama 3.4 i 3.5 je prikazan izračun matrice CO-a za ulazne vektore *z* prema pseudo kodu na slici 3.6 s ciljem pronalaska trećeg člana skupa *S*.

Tablica 3.4 Matrica S' - ulaz u funkciju za izračun CO-a

X	z_3	z_4	z_5
1	1	1	-1
2	1	1	1
3	-1	1	1
4	1	-1	-1
5	1	-1	1
6	-1	1	-1
∑	2	2	0

Tablica 3.5 Matrica - rezultat funkcije za izračun CO-a

X	$CO_{S \cup h3}$	$CO_{S \cup h4}$	$CO_{S \cup h5}$
1	1,00	1,00	-1,00
2	1,83	1,83	1,83
3	-1,00	1,00	1,0
4	1,83	1,00	1,00
5	1,83	1,00	1,83
6	1,00	183	1,00
∑	6,49	7,66	5,66

Primjer 2. pokazuje izračun matrice CO-a za novi SVK u koji je uključen po jedan od tri (h_3, h_4 *i* h_5) potencijalna klasifikatora, u slučaju kada SVK već sadrži neke članove. Izračun se razlikuje u odnosu na prošli primjer kao što je ranije u ovom poglavlju objašnjeno. Svaka promjena vrijednosti CO(x_i) više nije maksimalna jer po svakoj instanci x_i ovisi o rezultatima ranije uključenih klasifikatora. Navedeno rezultira time da ukupni CO na razini klasifikatora nije jednak razlici točnih i netočnih odluka klasifikatora, niti se njegov doprinos novom sustavu može izračunati jednostavnim pribrajanjem njegovih rezultata klasifikacije. Stoga se ukupni CO nekog SVK-a može izračunati samo izračunom za svaku instancu x_i posebno, uvažavajući sve članove SVK-a. Odluka koji će klasifikator biti odabran u skup S_{NEW}, kao što je već spomenuto, je prepuštena algoritmu. U slijedećem poglavlju detaljno je opisana konstrukcija COP algoritma.

3.4.4. COP algoritam za smanjivanje SVK-a

Algoritam COP odabire nove članove u SVK na temelju koeficijenta uspješnosti klasifikatora - CO. Dodavanje novih članova se može promatrati kao pohlepan izbor slijedećeg stanja iz susjedstva trenutnog stanja. Spomenuta stanja, u ovom slučaju su različiti podskupovi klasifikatora; trenutno stanje je skup S, a susjedstvo skupa S se sastoji od onih podskupova koji mogu biti konstruirani tako da se doda jedan h_t iz skupa S' u skup *S*. Algoritam se prema smjeru pretrage može svrstati u grupu s odabirom unaprijed, što znači da počinje s praznim skupom i u svakom koraku dodaje jedan novi član sustava (slika 3.8). Već odabranim klasifikatorima u skup *S* dodaje se u svakoj iteraciji po jedan klasifikator h_t iz skupa *S',* dotle dok se ne zadovolji uvjet da je skup $S' = \emptyset$, a skup S = H. Ključni element koji donosi COP algoritam jest upravo strategija odabira novih članova, koja zavisi da li se radi o odabiru parnog ili nepranog člana. Cilj strategije jest kombinirati dva konfliktna cilja prilikom dizajniranja algoritma za pretragu; diverzifikaciju i intenziviranje. Diverzifikacija se odnosi

na sposobnost pretraživanja mnogo različitih područja u prostoru rješenja, dok intenziviranje podrazumijeva sposobnost pronalaska kvalitetnih rješenja unutar tih područja. Dobar algoritam za pretragu treba pronaći ravnotežu između navedena dva cilja (Lozano i García-Martínez, 2010; Oreski, 2014). Pseudo kod za COP je prikazan na slici 3.8.

```
1. Ulaz: matrica rezultata Z (veličine N x L)
2. Inicijalno S_NEW = ∅, A = ∅, OPT = 0, S_OPT = ∅
3. Za svaki i=1:L
        a. S = S_NEW
        b. Poziv funkcije za izračun matrice CO-a slika(3.6)
        c. Odaberi novi član h_next ako je i neparan
             ii. na temelju Definicije 1
        d. U suprotnom
             ii. na temelju Definicije 2
        e. Kraj Odaberi
        f. Dodaj odabrani član h_next skupu S (S_NEW = S U h_next)
        g. Ažuriraj vektor A sa S_NEW
4. Kraj i
5. Za svaki i=1:L, i = i + 2
        a. Ako je p(A_i) > p(S_OPT)
           S_OPT = A_i, OPT = i
        b. Kraj Ako
6. Kraj i
7. Vrati: smanjeni SVK S_OPT
```

Slika 3.8 Pseudo kod za COP

Obzirom da COP koristi glasovanje kao funkciju kombinacije, broj klasifikatora u sustavu mora biti neparan. Stoga algoritam ima neparan broj koraka L, jer se sukcesivno u sustav dodaje samo po jedan novi član. Algoritam se može promatrati kao unija parnih $\{K_2, K_4, \ldots, K_{L-1}\}$ i neparnih koraka $\{K_1, K_3, \ldots, K_L\}$ koji imaju suštinski različite strategije odabira, čime se u algoritmu postiže dinamična ravnoteža diverzifikacije i intenziviranja. U svakom nepranom koraku COP algoritam intenzivira pretragu tako da odabire onaj klasifikator S_{NEW} koji ima ukupno najveći CO do kojeg se dolazi zbrajanjem svih $CO(x_i)$. Vrijednost $CO(x_i)$ ovisi o svakom članu skupa S_{NEW} (formula 3.18). Onaj temeljni klasifikator h_t koji se u nepranom koraku dodaje postojećem skupu S, za koji S_{NEW} ostvari najveći CO, nazivamo najboljim sljedećim neparnim članom i obilježavamo sa h_{next}. Za njega vrijedi da je element skupa S' te da se njegovim uključivanjem u postojeći S ostvaruje maksimalnu vrijednost CO.

Definicija 1

Najbolji sljedeći neparni član koji se dodaje skupu S je član h_{next} tako da $h_{next} \in S'$ i $\forall h_t \in S' \setminus h_{next}$: $CO_{S \cup h_{next}} \geq CO_{S \cup h_t}$.

U prvom koraku se na opisani način uvijek odabire onaj klasifikator koji ima najveću točnost, jer isti ima i najveći zbroj *CO*-a, što je prikazano u primjeru 1. Potom se u svakom slijedećem neparnom koraku odabire klasifikator koji zajedno s ranije odabranim skupom S ima najveću vrijednost *CO*-a (formula 3.19). Prema načinu izračuna CO koeficijenta više se vrednuju oni primjeri koji su bliže nuli tj. neizvjesnosti, to znači da klasifikator koji bolje klasificira „neizvjesne" slučajeve ima veće šanse da uđe u skup *S* kao sljedeći član. Traženje maksimalne vrijednosti CO intenzivira pretragu prostora na one h_t koji prema definiciji CO koeficijenta najviše pridonosi uspješnosti skupa S.

Da bi se pretražilo široko područje pretrage i izbjeglo zapinjanje u lokalnom minimumu u parnim koracima se radi diverzifikacija. Diverzifikacija se obično provodi kroz neku vrstu slučajnosti, kao primjer može se navesti mutacija u genetskom algoritmu (Mitchell,1996; Michalewicz, 1998).

U COP algoritmu diverzifikacija, u pranom koraku, usmjerava pretragu na one klasifikatore koji otvaraju prostor za poboljšanja performansi u slijedećoj iteraciji. Najbolji sljedeći član koji se dodaje skupu *S* u parnom koraku može se definirati na sljedeći način:

Definicija 2

Najbolji sljedeći parni član koji se dodaje skupu S je član h_{next} tako da $h_{next} \in S'$ i $\forall h_t \in S' \backslash h_{next}$: count(sign($CO_{S \cup h_{next}}(x_i)$)=0) + konst) / ($N$ − count(sign($CO_{S \cup h_{next}}(x_i)$) = 1) + ($konst/2$)) ≥ count(sign($CO_{S \cup h_t}(x_i)$)=0) + konst) / (N − count(sign($CO_{S \cup h_t}(x_i)$) = 1) + ($konst/2$))

ili

Najbolji sljedeći parni član koji se dodaje skupu S je član h_{next} tako da $h_{next} \in S'$ i $\forall h_t \in S' \backslash h_{next}$: (brojNeutralnih($S \cup h_{next}$) + konst) / ($N - brojTočnoKlasificiranih(S \cup h_{next})$ + ($konst/2$)) ≥ (brojNeutralnih($S \cup h_t$) + konst) / ($N - brojTočnoKlasificiranih(S \cup h_t)$ + ($konst/2$)).

pri čemu je *konst* dovoljno mali broj npr. 0,01 koji značajno ne utječe na vrijednost izraza, a neophodna je u slučajevima kada nema neutralno klasificiranih slučajeva ili kada su svi slučajevi točno klasificirani. U svakom koraku se odabire onaj klasifikator koji ima najveću vrijednost. Postoji mogućnost da dva klasifikatora imaju jednak rezultat, u tom slučaju se slučajnim odabirom bira najbolji sljedeći član.

3.4.5. Dizajn eksperimenta

Na temelju generiranih skupova podatka u istraživanju će se testirati performanse COP algoritma. Za svaku kombinaciju točnosti i raznolikosti generirat će se 100 različitih skupova podataka koji će biti učitani u COP algoritam. Nakon obrade podataka uspoređivat će se dvije ostvarene vrijednosti: (1) broj članova u skupu S i (2) točnost klasifikacije odabranih članova u odnosu na originalni (cijeli) SVK prije smanjivanja.

Broj članova u konačnom skupu *S* odredit će se nakon što: (1) algoritam prođe kroz sve korake i sukcesivno odabere sve dostupne klasifikatore i doda ih u *S*, (2) izračuna se mjera točnosti za sve neparne podskupove od *S* i (3) odabere onaj skup koji ima najveću točnost. Što je konačni broj članova manji to je sustav efikasniji. Osim efikasnosti COP ima za cilj i poboljšati performanse u odnosu na originalni SVK. Točnost klasifikacije odabranog podskupa od *H* će se usporediti s točnosti svih klasifikatora u skupu *H*, tj. s na originalnim SVK-om.

3.5 Empirijska analiza

Empirijska analiza počinje generiranjem skupova podataka određenih karakteristika za provođenje testa te njihovom analizom. Nakon toga se testiraju efikasnost i performanse COP algoritma za smanjivanje SVK-a.

3.5.1. Analiza generiranog skupa podataka

Algoritam predložen od autora Kuncheva i Kountchev (2002), u ovom radu modificiran korekcijskim faktorom opisanim u poglavlju 3.4.1, izveden je 100 puta za svaki od 31 klasifikatora za 3 raznolikosti i 4 točnosti, s ulaznim parametrima kako slijedi: $Q_{i,k} \in \mathrm{b}\ \forall i,k = 1,2,\dots,31, i \neq k,$ i $Q_{i,i} = 1$ gdje je $\mathrm{b} = \{0.7, 0.8, 0.9\}$ i $p \in \{0.65, 0.7, 0.75, 0.8\}$. U eksperimentu je konstruirano 31 X 3 X 4 X 100 tablica, (ukupno 37200 različitih tablica klasifikacije) koje sadržavaju po 1000 primjera. Korekcijski faktori *C(L,Q)* za *L*=31 i $Q \in b$ su računati prema formuli (3.12).

Srednje vrijednosti i standardne devijacije za ciljne točnosti *p*, dobivene generiranjem odgovora klasifikatora, bez obzira na mjeru Q, prikazane su u tablici 3.6.

Tablica 3.6 Srednje vrijednosti i standardne devijacije za ciljne točnosti p

Ciljni p	Dobivene prosječne točnosti
0.65	0.6505 (±0.0145)
0.70	0.7000 (±0.0145)
0.75	0.7502 (±0.0135)
0.80	0.8003 (±0.0124)

Srednje vrijednosti i standardne devijacije za mjeru raznolikosti Q dobivene generiranjem odgovora klasifikatora prikazane su u tablici 3.7.

Tablica 3.7 Aritmetičke sredine i standardne devijacije za mjeru Q u ovisnosti o ciljnim točnostima i raznolikostima klasifikatora

Ciljni Q	Ciljni p			
	0.65	0.70	0.75	0.80
0.7	0.6604 (±0.0150)	0.6747 (±0.0160)	0.6820 (±0.0169)	0.7032 (±0.0169)
0.8	0.7969 (±0.0127)	0.8032 (±0.0118)	0.8107 (±0.0119)	0.8216 (±0.0126)
0.9	0.9101 (±0.0078)	0.9119 (±0.0077)	0.9161 (±0.0074)	0.9184 (±0.0070)

Radi testiranja performansi COP algoritma za smanjivanje SVK-a korišten je modificirani algoritam za generiranje odgovora klasifikatora fiksne točnosti i raznolikosti, čiji pseudo kod je prikazan na slici 3.3. Modificirani algoritam aproksimira odgovore klasifikatora prema ciljnim parametrima koji se zadaju na ulazu. Iz generiranih skupova podataka, koji predstavljaju odgovore klasifikatora, izračunate su srednje vrijednosti za p i Q, prikazane u tablicama 3.6 i 3.7, respektivno. Iz tablica je vidljivo da su ostvarene vrijednosti u skladu s zadanim parametrima. Jedino mjera raznolikosti Q za neke parametre, npr. $Q = 0.7$ i $p = 0.65$, ima nešto veće odstupanje od ciljnih vrijednosti ali je disperzija rezultata vrlo mala što predstavlja dobar input za testiranje performansi COP algoritma.

3.5.2. Primjena COP algoritma

Temeljem provedenih testova analiziraju se efikasnost (broj članova) i performanse (točnost klasifikacije) smanjenog SVK-a, kreiranog pomoću COP algoritma. U tablici 3.8 dan je, za svaku kombinaciju, prosječan broj članova sustava nakon smanjivanja, dobiven temeljem 100 zasebnih testova koji su provedeni za svaku kombinaciju točnosti i raznolikosti na skupu od 31 člana.

Odabrani podskupovi su značajno manji od originalnog SVK-a, tj. algoritam je sa zadanim parametrima u prosjeku smanjivao broj članova od 50% do 80%. Ako se rezultati promatraju iz aspekta točnosti, tada je vidljivo da su sustavi čiji članovi imaju manju točnost više smanjeni u odnosu na one koji imaju veću točnost. Algoritam stvara efikasnije sustave iz

članova koji imaju manju točnost, međutim, takvi sustavi imaju značajno slabije rezultate točnosti klasifikacije u odnosu na sustave s članovima većih točnosti.

Tablica 3.8 Aritmetička sredina i standardna devijacija broja članova sustava nakon smanjivanja u ovisnosti o zadanim točnostima i raznolikostima klasifikatora

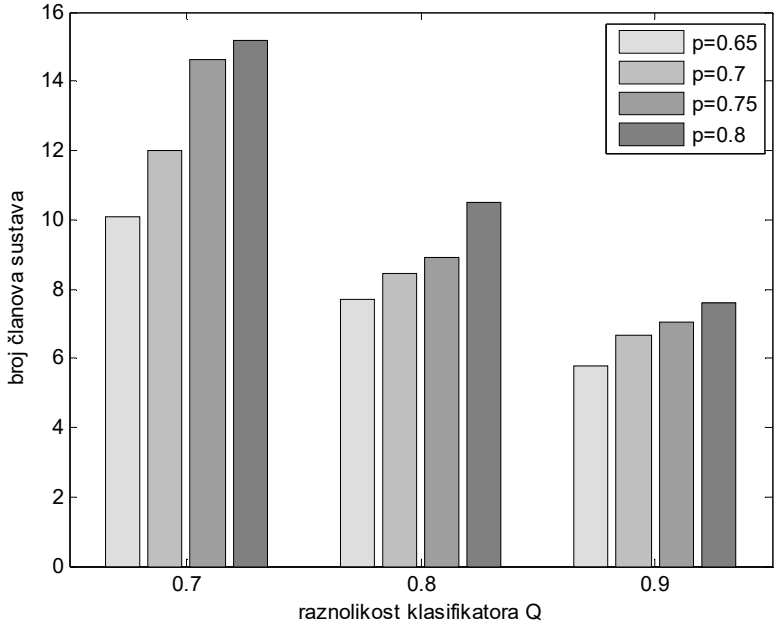
Ciljni Q	Ciljni p			
	0.65	0.70	0.75	0.80
0.7	10.08 (±6.0464)	12.0 (±7.2055)	14.64 (±7.3272)	15.16 (±6.8442)
0.8	7.70 (±4.3519)	8.46 (±4.1082)	8.90 (±5.0642)	10.50 (±5.6381)
0.9	5.76 (±3.2879)	6.68 (±2.7957)	7.04 (±3.2409)	7.58 (±3.0918)

Intuitivno, takve rezultate potvrđuje svakodnevno iskustvo. Naime, ako izbornik odabire neki tim za natjecanje i pri tome veliki broj kandidata ostvaruje odlične rezultate, na natjecanje će povesti tim s većim brojem članova jer si time značajno diže vjerojatnost uspjeha tima na natjecanju. Nasuprot tome, ako su rezultati kandidata relativno loši, na natjecanje će povesti tim s manjim brojem članova. Dok je logika za prvi slučaj, izbor većeg broja kandidata, ako neki od članova slučajno ostvari slabiji rezultat, veći broj drugih članova će taj podbačaj „popraviti“ svojim rezultatima. Kod drugog slučaja logika je obrnuta, ako neki od kandidata ostvari dobar rezultat veći broj ostalih kandidata bi taj rezultat značajno pokvario, stoga formira manji tim i nada se. Ovdje prikazani rezultati eksperimenta provedenog na vrlo velikom skupu slučajeva potvrđuju racionalno ponašanje izbornika pri odabiru članova tima.

Općenito, ako se analizira točnost članova u odnosu na veličinu SVK-a, tada se jednostavno može zaključiti da stvaranje sustava nema smisla ukoliko je točnost klasifikatora 0.5 ili 1. U oba slučaja veličina sustava je jednaka 1 jer, bez obzira radi li se o slučajnom odabiru klase ili o savršenom klasifikatoru, kombiniranje ne može donijeti poboljšanje po pitanju performansi. Klasifikatori s točnosti za koje vrijedi $0.5 < p < 1$ kombiniranjem mogu donijeti poboljšanje performansi. Unutar određenih granica, algoritam COP jasno pokazuje korelaciju između veličine skupa i točnosti tj. što je veća točnost klasifikatora to je i broj članova veći. Opisane rezultate prikazuje slika 3.9.

Analizom efikasnosti iz aspekta mjere raznolikosti se također dolazi do snažne korelacije između Q vrijednosti i broja članova sustava. Originalni sustavi koji imaju manju mjeru raznolikosti (članovi su nezavisniji) nakon korištenja COP algoritma imaju veći broj članova u odnosu na sustave s većim Q vrijednostima. Takvi rezultati su u skladu s definicijom raznolikosti između klasifikatora, koja kaže da klasifikatori s raznolikijim odlukama postižu bolje performanse kada se kombiniraju u sustav.

U idealnom slučaju kada je $Q = 0$ (najveća raznolikost sustava) tada svi temeljni klasifikatori doprinose konačnoj odluci, pa smanjivanje SVK-a nema smisla. Obrnuto u slučaju kada svi klasifikatori generiraju iste odluke ili kada je $Q = 1$ tada je dovoljan samo jedan od članova jer je konačan rezultat isti. Stoga od raznolikosti $Q = 0$ gdje je broj članova sustava jednak L, do $Q = 1$ gdje je broj članova 1, prosječna veličina sustava pada. Navedeno svojstvo je potvrđeno ostvarenim rezultatima (slika 3.9).



Slika 3.9 Prikaz odnosa broja članova sustava i raznolikosti klasifikatora

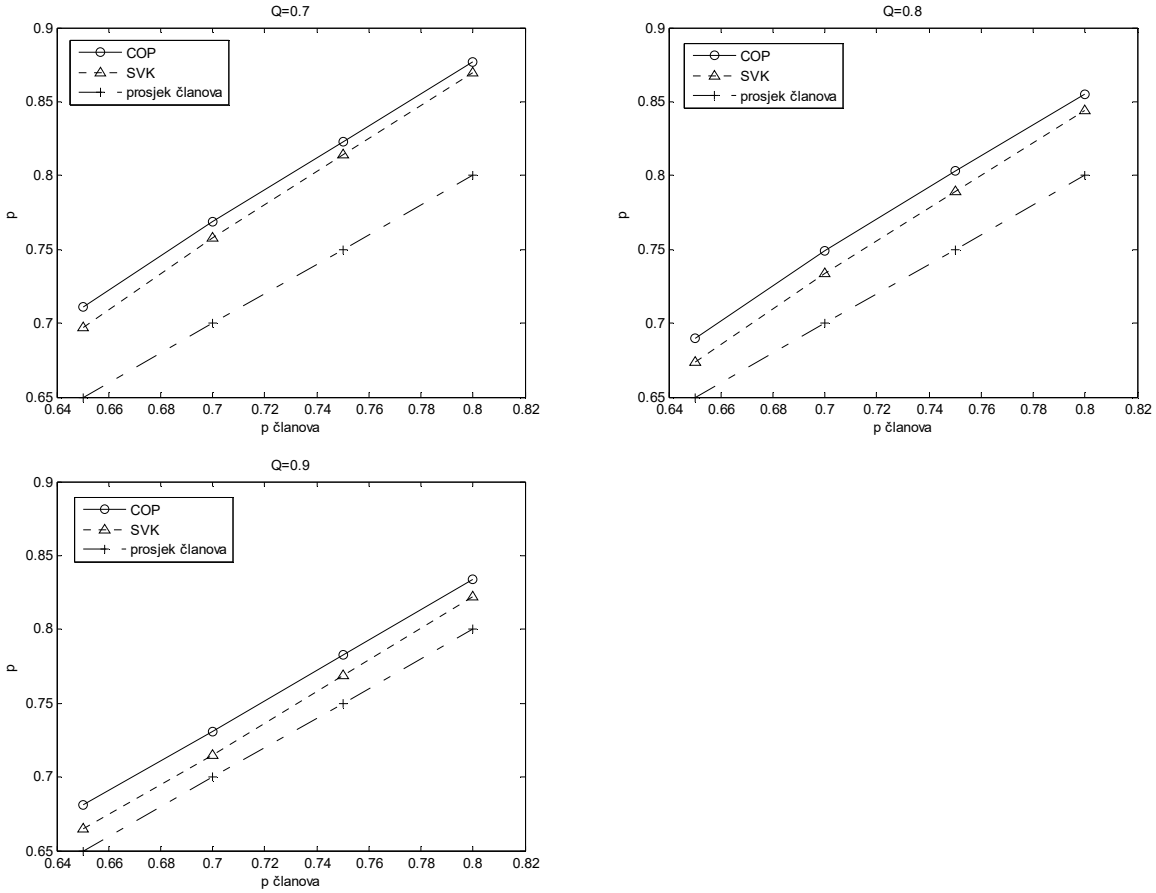
Iz provedenog istraživanja može se zaključiti da je COP pozitivno utjecao na efikasnost originalnog sustava jer je smanjenje SVK-a, unutar zadanih parametara točnosti i raznolikosti, bilo od 50 – 80%.

Tablica 3.9 Aritmetička sredina i standardna devijacija točnosti originalnih i smanjenih SVK-a u ovisnosti o zadanim točnostima i raznolikostima klasifikatora

Ciljni	Originalni SVK za ciljni Q			COP SVK za ciljni Q		
p	0.7	0.8	0.9	0.7	0.8	0.9
0.65	0,697 ±0,016	0,674 ±0,015	0,665 ±0,012	0,711 ±0,014	0,690 ±0,015	0,681 ±0,011
0.70	0,758 ±0,014	0,734 ±0,015	0,715 ±0,014	0,769 ±0,011	0,749 ±0,013	0,731 ±0,013
0.75	0,814 ±0,012	0,789 ±0,011	0,769 ±0,014	0,823 ±0,011	0,803 ±0,011	0,783 ±0,013
0.80	0,870 ±0,012	0,844 ±0,012	0,822 ±0,013	0,877 ±0,012	0,855 ±0,010	0,834 ±0,011

Pored analize efikasnosti COP algoritma, provedena je i analiza točnosti smanjenog SVK-a. Rezultati analize točnosti u obliku prosječne ostvarene točnosti originalnih SVK-a i smanjenih pomoću COP algoritma nalaze se u tablici 3.9.

Rezultati su u skladu s postavkama sustava višestrukih klasifikatora kako su definirane u ranijim poglavljima disertacije. Što sustav ima manju vrijednost Q i veću vrijednost p to su njegove performanse u vidu točnosti klasifikacije bolje. Na slici 3.10 prikazana su 3 linijska grafa s tri različite linije koje uspoređuju točnost: prosjeka članova sustava, originalnog SVK-a i smanjenog SVK-a pomoću COP algoritma, u zavisnosti od vrijednosti Q.



Slika 3.10 Usporedba točnosti članova sustava, originalnog SVK-a i smanjenog SVK-a COP algoritmom u zavisnosti od vrijednosti Q

Iz slike se vidi koliko donosi spajanje klasifikatora u sustav, ali i koliko doprinosi smanjivanje SVK-a u odnosu na originalni sustav. Vidljivo je, što je Q vrijednost veća to smanjivanje značajnije doprinosi točnosti u sustavu. Navedeno pravilo se najznačajnije vidi na trećem grafu $Q = 0.9$ gdje su doprinos spajanja u SVK i doprinos smanjivanja SVK-a gotovo izjednačeni. Stoga se može zaključiti da je korištenje COP algoritma opravdano s naglaskom da svoj najveći doprinos daje na klasifikatorima čije odluke međusobno više zavise što je ujedno najčešći slučaj u praksi.

Također možemo zaključiti da SVK nema smisla u slučajevima kada pojedinačni temeljni klasifikator daje bolje performanse klasifikacije od SVK-a, ali takvi slučajevi su

mogući. U takvim slučajevima pruning ima ključni doprinos u sprečavanju nelogične situacije jer smanjivanjem, SVK se izjednačava, tj. postaje pojedinačni klasifikator.

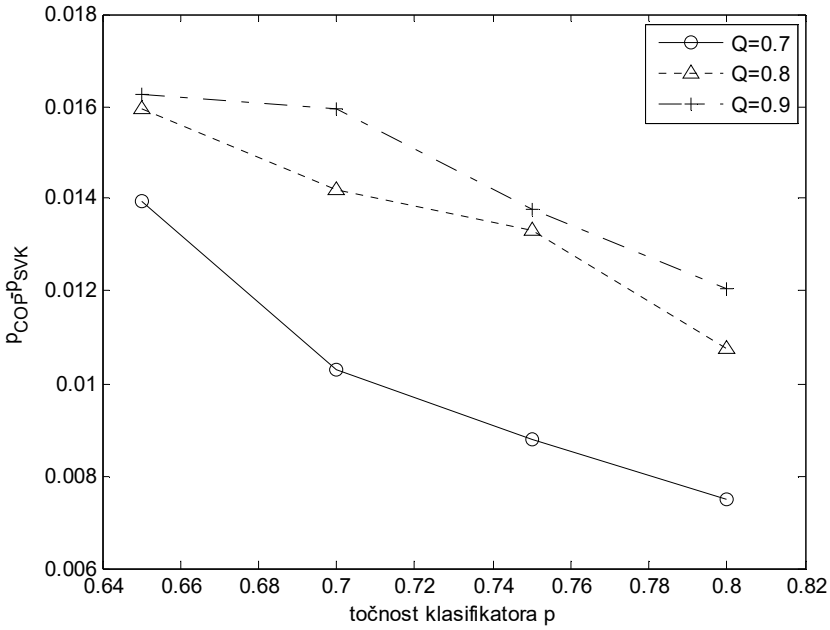
Analiza performansi SVK-a, s aspekta točnosti pojedinačnih članova, pokazuje da temeljni klasifikatori koji imaju veću točnost postižu bolje poboljšanje performansi u SVK u odnosu na one temeljne klasifikatore koji imaju manju točnost (tablica 3.9), što je u negativnoj korelaciji s poboljšanjem performansi koje dobijemo s osnova smanjivanja SVK-a COP algoritmom, jer to poboljšanje opada s porastom točnosti (tablica 3.10).

Tablica 3.10 Poboljšanje performansi primjenom COP algoritma

Ciljni Q	Ciljni p			
	0.65	0.70	0.75	0.80
0.7	0,0139 (±0,0072)	0,0103 (±0,0064)	0,0088 (±0,0051)	0,0075 (±0,0045)
0.8	0,0159 (±0,0071)	0,0142 (±0,0062)	0,0133 (±0,0058)	0,0108 (±0,0053)
0.9	0,0163 (±0,0052)	0,0160 (±0,0062)	0,0138 (±0,0050)	0,0120 (±0,0050)

Razlika između točnosti sustava prije smanjivanja i poslije je veća na klasifikatorima s manjom točnosti. Navedeno je u skladu s ranije prikazanim rezultatima (Tablica 3.7) iz kojih je vidljivo da je broj članova sustava nakon smanjivanja manji kod SVK čiji članovi imaju manju točnost. Temeljem navedenog proizlazi da SVK-i s manjom točnosti članova, u promatranom intervalu točnosti od 0,65 do 0,80 i raznolikosti od 0,7 do 0,9 , postižu bolju točnost uz manji broj članova te se i najveći doprinos smanjivanjem postiže kod tih klasifikatora. Na takav zaključak navode rezultati provedenog istraživanja s generiranim skupom podataka.

Sličan zaključak se može izvesti promatrajući rezultate točnosti u odnosu na raznolikost. Originalni sustavi koji su ostvarili veće poboljšanje točnosti spajanjem članova u sustav (oni s većom raznolikosti, manji Q) imaju manju korist od smanjivanja nego sustavi s manjom raznolikosti.



Slika 3.11 Poboljšanje točnosti smanjenog SVK-a, promatrano prema točnosti i raznolikosti temeljnih klasifikatora

Algoritam COP, kao što rezultati istraživanja pokazuju, je na svim skupovima podataka donio poboljšanje. Njegovo korištenje je opravdano po pitanju efikasnosti i performansi smanjenog sustava. U nastavku slijede statistički testovi kojima se testira statistička značajnost rezultata.

3.5.3. Statistički testovi

Statistička komparaciju rezultata izvedena je pomoću *t*-testa za zavisne skupove i neparametarskog Wilcoxonovog testa za zavisne skupove (engl. *Wilcoxon matched-pairs signed rank test*). Tim testovima utvrđuje se da li se razlike procijenjenih srednjih vrijednosti točnosti mogu smatrati značajnim.

Ukupan broj provedenih testova kao i veličina svakog pojedinog skupa sa 1000 opažanja, dobre su pretpostavke za provođenje uparenog *t*-testa (Japkowicz i Shah, 2011). Nulta hipoteza je da ne postoji razlika između prosječne točnosti dviju tehnika; originalnog i smanjenog SVK-a. Na temelju dvostranog, uparenog *t*-testa, prikazanog na slici 3.12, odbacili smo nultu hipotezu zato što je izračunata p-vrijednost < 0,0001 i signifikantna je u odnosu na odabranu razinu signifikantnosti ($\alpha = 0{,}05$).

Osim s uparenim *t*-testom, testirali smo performanse originalnog i smanjenog SVK koristeći Wilcoxonov test uparenih parova. Taj test pruža dobru alternativu uparenom *t*-testu

kad je populacija razlika rezultata simetrično raspoređena. Taj test je malo manje snažan od *t*-testa kada su podaci raspoređeni po normalnoj distribuciji, a može biti znatno bolji kada su razlike rezultata simetrično (ali ne nužno i normalno) distribuirane s jakim repovima (Myers i Well, 2003).

Wilcoxon Signed Rank test uparenih parova testira nultu hipotezu da je medijan razlike jednak 0.0 nasuprot alternativne hipoteze da je medijan razlike različit od 0.0, tj. nulta hipoteza je da su oba SVK-a jednako dobra. Iz razloga što je ostvarena p-vrijednost < 0,0001 i signifikantna, možemo odbaciti nultu hipotezu s 95.0% pouzdanosti i odbaciti ideju da je razlika slučajna, stoga zaključujemo da populacije imaju različite medijane u korist novo predloženog smanjenog SVK-a.

Kao što je prikazano, ostvarena prosječna točnost klasifikacije je uspoređena pomoću parametarskog udvojenog *t*-testa i neparametarskog Wilcoxon Signed Rank testa uparenih parova. Rezultati oba statistička testa pokazuju da podaci potvrđuju pretpostavku da rezultati sustava smanjenog COP algoritmom postižu bolje rezultate od originalnog sustava sa svim članovima, na razini signifikantnosti $\alpha = 0{,}05$.

```
> t.test(dokStatTestPog3$Orig.SVK,dokStatTestPog3$COP.SVK,paired=TRUE)

                              Paired t-test

data:  dokStatTestPog3$Orig.SVK and dokStatTestPog3$COP.SVK
t = -68.749, df = 1199, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.01308815 -0.01236185
sample estimates:
mean of the differences
              -0.012725

> wilcox.test(dokStatTestPog3$Orig.SVK,dokStatTestPog3$COP.SVK,paired=TR
UE)

          Wilcoxon signed rank test with continuity correction

data:  dokStatTestPog3$Orig.SVK and dokStatTestPog3$COP.SVK
V = 0, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Slika 3.12 Rezultati statističkih testova na ostvrenim rezultatima istraživanja

3.6 Zaključci poglavlja

U ovom poglavlju je razvijen algoritam za smanjivanje SVK koji za metriku odabira klasifikatora koristi novu mjeru pod nazivom koeficijent uspješnosti klasifikatora.

Istraživanje je provedeno na teorijskim skupovima podataka generiranim pomoću modificiranog algoritma autora Kuncheva i Kountchev (2002). Ukupno je korišteno 1200 skupova podataka po 1000 primjera svaki. Analizirane su 3 različite raznolikosti klasifikatora od 0.7 – 0.9 te 4 točnosti od 0.65 – 0.80. Navedeni spektar karakteristika uzet je zbog sličnih karakteristika stvarnih kreditnih skupova podataka (Njemačkog i Hrvatskog) i pretpostavke da, ako na generiranim podatcima navedenih karakteristika novi algoritam za smanjivanje SVK-a pokaže dobre rezultate, tada će slične rezultate ostvariti i na stvarnim skupovima podataka.

Na temelju dobivenih rezultata vidljivo je da su smanjeni sustavi manji od 50%-80% što predstavlja značajno smanjenje. Točnost je u prosjeku poboljšana za 1.27 postotna boda u odnosu na originalne sustave.

Statistički je potvrđeno da novi algoritam daje bolje rezultate u vidu efikasnosti i performansi predviđanja u odnosu na originalni SVK bez smanjivanja. Također, kao dodatni doprinos istraživanja prezentiranog u ovom poglavlju je značajno unaprjeđen generator teorijskih skupova podataka određenih točnosti i raznolikosti. Ostvareni rezultati su obećavajući i daju dobar temelj za primjenu tehnike na stvarnim skupovima podataka.

Istraživanje će se nastaviti primjenom ovdje kreiranog COP algoritma na stvarnim skupovima kreditnih podataka. Osim performansi testirat će se i brzina novog algoritma.

4

Odabir atributa kao podloga za kombiniranje klasifikatora na kreditnim podacima

U poglavlju je opisana nova tehnika za kombiniranje klasifikatora te je testirana opravdanost takvog kombiniranja kroz ostvareno poboljšanje performansi u odnosu na pojedinačne članove sustava.

4.1 Uvod

Problem procjene kreditnog rizika se jednostavno može definirati kao problem odlučivanja da li će tražitelj kredita biti sposoban otplatiti kredit u skladu s ugovorenim uvjetima ili će isti imati problema prilikom otplate. Ovisno o odluci, kreditna institucija će odobriti ili odbaciti zahtjev tražitelja kredita. Važnost procjene kreditnog rizika je često naglašavana u kontekstu velike financijske krize koje je započela 2007. godine i koja je otkrila nedostatke financijskog sustava po pitanju spomenutog problema. Pogrešne odluke, kao što se pokazalo, mogu utjecati ne samo na stabilnost banaka, već zbog njihove uloge financijskog krvotoka i na cijelo gospodarstvo. Iz opisanog odnosa proizlazi važnost i opravdanost istraživanja promatranog problema iz različitih područja kako financija i računovodstva tako i područja umjetne inteligencije koja sve više ima utjecaja na donošenje kreditnih odluka.

Razlog uključivanja teorije umjetne inteligencije u problem procjene kreditnog rizika proizlazi iz potrebe za stvaranjem sve sofisticiranijih modela u kreditnim institucijama. S naglim rastom kreditne industrije i potrebe za upravljanjem sve većih portfelja kredita, evaluacija kreditnih prijava se sve više oslanja na računalne kreditne modele. Modeli za ocjenu kreditnog rizika klasificiraju kredite, sukladno ranijoj definiciji problema, u dvije grupe: (1) dobra kreditna grupa (prihvaćeni zahtjevi) i (2) loša kreditna grupa (odbijeni zahtjevi). Klasifikacija se izvodi na temelju niza karakteristikama kao što su visina prihoda, rashodi, povijest otplate i slično (Chen i Huang, 2003). Prednosti razvoja i korištenja pouzdanog kreditnog modela su (Tsai i Wu, 2008):

a) reduciranje troška analize kreditnih zahtjeva,
b) omogućavanje bržeg donošenja odluka i
c) osiguravanje naplate kredita i umanjivanje mogućeg rizika.

Svako, čak i malo poboljšanje točnosti klasifikacije može smanjiti kreditni rizik i pozitivno utjecati na buduće financijske rezultate (Tsai i Wu, 2008). Jedan od načina poboljšanja klasifikacije jest kroz kombiniranje klasifikatora, već afirmirano područje istraživanja. Ideja kombiniranja nije nova stoga već postoje brojna istraživanja vezana uz kreditni rizik (Twala, 2010; Finlay, 2011; Wang i Ma, 2012) s pozitivnim rezultatima. Dostupni radovi kao ključ uspjeha ističu generiranje klasifikatora koji rade pogreške na različitim primjerima (Zhang i Zhou, 2013; Didaci et al., 2013). Iako se problem generiranja klasifikatora koji na različitim primjerima rade pogreške čini jednostavan, često je upravo suprotno. Pristupi stvaranja raznolikosti u odlukama klasifikatora su brojni, u ovom

istraživanju raznolikost će se nastojati postići kroz upravljani odabir atributa iz skupa dostupnih atributa. Ideja je, koristiti više tehnika koje rangiraju atribute prema njihovoj povezanosti s zavisnom varijablom. Potom se prilikom treniranja klasifikatora koriste različiti podskupovi najbolje rangiranih atributa svake tehnike. Pristup konstrukcije sustava višestrukih klasifikatora s više tehnika za odabir atributa dosada nije istraživan na kreditnim podacima, a kao dodatni korak predložene tehnike će se koristiti novi COP algoritam za optimalni odabir članova. Cilj istraživanja je poboljšati rezultate klasifikacije izradom tehnike koja nije računski zahtjevna, a dovoljno je robusna za primjenu na različitim problemima iz različitih domena.

Preostali dijelovi ovog poglavlja su organizirani na sljedeći način. U sljedećem odjeljku je definiran problem te je dan pregled literature vezano uz načine postizanja različitosti prilikom konstruiranja sustava višestrukih klasifikatora. Kratki opis tehnika i koncepata koji će se koristiti u istraživanju prezentiranom u ovom poglavlju je dan u odjeljku 3. Odjeljak 4 opisuje eksperimentalni dizajn tehnike za konstruiranje sustava višestrukih klasifikatora na temelju odabira atributa i tehnike COP. Empirijska analiza u kojoj se rezultati sustava uspoređuju u odnosu na pojedinačne članove sustava je dana u odjeljku 5. U zadnjem 6. odjeljku izneseni su zaključci i smjernice za buduća istraživanja.

4.2 Definiranje problema i pregled literature

Problem odabira optimalnog podskupa atributa je važan problem strojnog učenja jer rezultati tog odabira imaju značajan utjecaj na rezultate učenja klasifikatora. On se sastoji od prepoznavanja i uklanjanja nebitnih i redundantnih atributa iz skupa podataka za učenje, tako da se algoritam za učenje fokusira samo na one podatke koji su korisni za analizu i predikciju (Guyon et al., 2008). Težina problema je u potencijalno velikom prostoru pretrage i činjenici da ukoliko se koriste: (1) iscrpne tehnike pretrage prostora, pronalazak rješenja može biti vremenski vrlo zahtjevan ili (2) pohlepne pretrage prostora, tada rješenja mogu biti lošija zbog velikog broja lokalnih optimuma u prostoru pretrage. Odabir atributa je važan iz slijedećih razloga (Cunningham i Carney, 2000):

a) treniranje klasifikatora postiže bolje performanse nakon uklanjanja nebitnih atributa iz skupa podataka,
b) izrada efikasnijih modela koji su što je moguće više jezgrovitiji i

c) otkrivanje znanja koji su atributi, u kojoj domeni, utjecajni na rezultat te onih koji to nisu.

Uz navedene razloge postoji i jedan dodatni koji se pojavio u istraživanjima vezanima za sustave višestrukih klasifikatora. Istraživači koriste odabir atributa za postizanje raznolikosti u odlukama klasifikatora. Raznolikost u ovom kontekstu se odnosi na način kako klasifikatori čine greške tj. za klasifikatore se kaže da su raznoliki ukoliko ne rade iste greške na istim primjerima (Tang et al., 2006; Wang i Yao, 2013). Raznolikost označava „specijalizaciju“ pojedinih klasifikatora na različitim dijelovima ulaznog skupa podataka.

Na primjer, ukoliko jedan član SVK-a, h_i, klasificira ulazni primjer x netočno i njegova odluka je nepovezana s druga dva člana iz SVK-a, h_j i h_k, koji ispravno klasificiraju isti primjer, tada će glasanje većine od sva tri člana osigurati da je primjer x ispravno klasificiran zbog utjecaja h_j i h_k. U suprotnom, bez raznolikosti, postoji rizik od neispravne klasifikacije istih primjera. U gornjem primjeru, ukoliko bi članovi h_j i h_k imali snažno korelirajuće odgovore s h_i, svi članovi bi neispravno klasificirati x, čime bi i odgovor SVK-a bio netočan; a time ne bi postojala korist kombiniranja klasifikatora u sustav (Bhowan et al., 2014).

Glavni cilj istraživanja prikazanog u ovom poglavlju jest saznati da li je pristup temeljen na filtarskim tehnikama odabira atributa prikladan za stvaranje raznolikosti klasifikatora što bi prilikom kombinacije u sustav rezultiralo s poboljšanjem performansi klasifikacije na kreditnim skupovima podataka. Istraživanje teži jednostavnosti, stoga nije cilj koristiti zahtjevne tehnike odabira atributa koje zahtijevaju računalno velike resurse u pronalasku optimalnog podskupa. Ideja je kombiniranjem brzih tehnika i algoritama za klasifikaciju bez ili sa što manje prilagođavanja parametara iskoristiti prednosti kombiniranja zasebnih modela. Takva robusna tehnika bi bila široko primjenjiva i prikladna za korištenje kod širokog kruga istraživača na području strojnog učenja. U skladu s time je postavljena glavna hipoteza istraživanja:

H1: *Sustav višestrukih klasifikatora koji je temeljen na odabiru različitih podskupova atributa pomoću filtarskih tehnika te konstruiran na temelju u ovom radu predloženog algoritma za smanjivanje sustava će postizati statistički značajno veću točnost klasifikacije od pojedinačnih klasifikatora uključenih u sustav na razini statističke značajnosti* $p \leq 0,05$.

Svako i najmanje poboljšanje u odnosu na performanse pojedinačnih algoritama, koje se može statistički dokazati, smatrat će se zadovoljavajućim obzirom da je i najmanji postotak poboljšanja efikasnosti od velikog značaja kreditnim institucijama.

Istraživanje definirano H1 hipotezom uključuje tri izazova: (1) smanjivanje dimenzionalnosti inicijalnog skupa podatka ili odabir atributa, (2) kombiniranje klasifikatora u sustav i (3) smanjivanje SVK-a na optimalni broj članova. Literatura vezana uz kombiniranje klasifikatora je detaljno opisana u poglavlju 2, a literatura za smanjivanje SVK-a u poglavlju 3, stoga u nastavku slijedi pregled literature vezan za odabir atributa s naglaskom na kombinaciju tehnika.

Odabir atributa (engl. *feature selection*) je bitna aktivnost u pretprocesiranju podataka koja je uspješno primijenjena na mnogo različitih klasifikacijskih problema kao npr.: klasifikacija teksta (Gomez et al., 2102; Basu i Murthy, 2012), slika (Dhasal et al., 2012), kreditnog rizika (Oreski i Oreski, 2014), neovlaštenog upada (Mukherjee i Sharma, 2012), mikro-polja gena u DNK (engl. *gene expression microarray*) (Lazar et al., 2012). Većina istraživača se slaže da ne postoji najbolja metoda i njihova nastojanja su usmjerena na pronalazak „*dobre*“ metode za specifični problem. Stoga se često pojavljuju nove metode za odabir atributa koje koriste različite strategije (Bolón-Canedo et al., 2012):

a) reinterpretacija postojećih algoritama (Sun & Li, 2006), da bi se adaptirali nekom problemu (Sun et al., 2008);
b) konstruiranje novih metoda za primjenu na još neriješenim problemima (Chidlovskii & Lecerf, 2008);
c) korištenje nekoliko metoda istog pristupa za odabir atributa, kao npr. metoda filtarskog pristupa (Zhang et al., 2008), ili kombiniranjem različitih pristupa, tipično filtarskog i tehnika omotača (El Akadi et al., 2011);
d) kombiniranjem odabira atributa s drugim tehnikama, kao što je ekstrakcija atributa (Vainer et al., 2011); i
e) korištenje sustava višestrukih tehnika (ansambla) za odabir atributa (Bolón-Canedo et al., 2012).

Strategije uglavnom počivaju na dvije vrste tehnika koje se dijele prema evaluacijskom kriteriju na zavisne i nezavisne, ovisno o korištenju algoritma za učenje. Nezavisne ili filtarske tehnike ne uključuju algoritam za učenje, već iskorištavaju karakteristike samih podataka za ocjenu kvalitete podskupa atributa (Kumar i Minz, 2014). Često su odabrane zbog jednostavnosti primjene i brzine obrade podataka (Bolón-Canedo et al., 2013).

Zavisne tehnike podrazumijevaju korištenje algoritma za učenje. Performanse algoritma se koriste za evaluaciju dobrote odabranog podskupa atributa. Kod toga odabrani podskup atributa najbolje odgovara specifičnom algoritmu. Stoga su rezultati ovog pristupa uobičajeno bolji ali uz cijenu veće računalne složenosti jer se za svaki podskup mjere performanse. Zavisne tehnike se nazivaju tehnike omotača i također su zastupljene u literaturi (Bolón-Canedo et al., 2013).

Međutim primjena tehnika za odabir atributa nije ograničena samo na izbor filtarske ili tehnike omotača već se mogu koristiti i njihove kombinacije (Bolón-Canedo et al., 2014). Primjer takve strategije jest serijska integracija dvije filtarske tehnike: ReliefF i MRMR (engl. *Minimal-Redundancy-Maximal-Relevance*) za odabir atributa predstavljene u radu (Zhang et al., 2008). Tehnike se izvode u dvije faze: prvo se originalni skup filtrira pomoću ReliefF tehnike, a potom se izvodi MRMR tehnika. U usporedbi s korištenim tehnikama izvođenim zasebno i nekim drugim tehnikama za odabir atributa novo predstavljena kombinacija postiže bolje rezultate.

Osim serijskog, gdje jedna tehnika na ulazu prima odabrane atribute druge, tehnike se mogu kombinirati paralelno. Nekoliko različitih varijanti paralelnog kombiniranja je konstruirano u istraživanju (Shen et al., 2012). Autori su implementirali tri nova koncepta koja se razlikuju ovisno o kombiniraju tehnika. Ovisno o konceptu korišten je ili genetski algoritam koji se bazira na stohastičkoj pretrazi ili kombinacija više različitih tehnika koje deterministički biraju najbolje atribute. Odluke tehnika se u konačnici kombiniraju u jedan podskup atributa koji koristi algoritam za klasifikaciju.

Korištenje genetskog algoritam za stvaranje raznolikosti klasifikatora temeljenih na neuronskim mrežama je ranije istraživao Opitz (1999) pod nazivom GEFS (Genetic Ensemble Feature Selection) algoritam. Predloženi algoritam u kombinaciji s klasifikacijskim algoritmom je uspoređen s Bagging i Boosting tehnikama, a rezultati pokazuju poboljšanje performansi u odnosu na spomenute tehnike.

Upotreba više tehnika za postizanje stabilnosti i robusnosti odabira atributa je predložena u istraživanju (Saeys et al., 2008). Robusnost autori definiraju kao varijaciju u rezultatima odabira atributa kao rezultat malih promjena u skupu podataka. Četiri različite tehnike se uz pomoć Bagging algoritma, gdje je za svaku novu iteraciju algoritma korištena jedna od četiri tehnike, kombiniraju u jedan sustav. Istraživanje je provedeno na 6 skupova podataka, a rezultati pokazuju da se korištenjem više tehnika postiže robusniji odabir nego njihovim pojedinačnim korištenjem.

U ovom istraživanju koncept kombiniranja je drugačiji u odnosu na opisane, koristit će se više filtarskih tehnika ali se njihovi rezultati neće kombinirati, već će svaka tehnika biti dodijeljena zasebnom klasifikatoru. U slijedećem poglavlju slijedi metodologija istraživanja u kojem su opisane korištene tehnike.

4.3 Metodologija

Cilj istraživanja prikazanog u ovom poglavlju je konstruirati sustav višestrukih klasifikatora na temelju upravljanog odabira atributa i primijene novog COP algoritam za smanjivanje sustava. Istraživanje u fokus stavlja sposobnost postizanja raznolikosti odgovora klasifikatora temeljem različitih podskupova atributa. Stoga je glavni naglasak na odabiru različitih podskupova atributa pomoću filtarskih tehnika. Između velikog broja tehnika dostupnih u literaturi odabrano je pet različitih filtarskih tehnika: omjer informacijske dobiti, Relief, Gini indeks, mjera nesigurnosti i korelacija. Odabrane tehnike su baziraju na nekoliko različitih mjera za vrednovanje kvalitete atributa, što bi trebalo osigurati raznolikost u odabiru, odnosno široki raspon odabranih atributa. Osim toga, u predloženim modelima, za hrvatski i njemački skup kreditnih podataka koriste se dvije vrste sustava višestrukih klasifikatora: homogeni i heterogeni. Heterogeni sustav koristi više različitih algoritama za učenje s ciljem postizanja veće raznolikosti odluka klasifikatora. U ovom odjeljku ćemo ukratko opisati navedene tehnike i koncepte koji se koriste u istraživanju prikazanom u ovom poglavlju.

4.3.1. Koeficijent korelacije

Korelacija izražava statistički odnos između dviju slučajnih varijabli ili dva skupa podataka. Takav odnos između podataka često se kvantificira pomoću koeficijenta korelacije (Vořechovský, 2012). Za mjerenje stupanja korelacije postoji nekoliko koeficijenata korelacije, često se označavaju sa ρ ili r. Najčešći od njih je Pearsonov koeficijent korelacije, koji je osjetljiv samo na linearni odnos između dvije varijable. Pri tome je poznato da je Pearsonov koeficijent korelacije jako pod utjecajem rubnih vrijednosti, tj. stršećih ili ekstremnih podataka (engl. *outlier data*). Ostali koeficijenti korelacije su razvijeni kako bi bili više robusni na stršeće podatke od Pearsonova koeficijenta korelacije i osjetljiviji na nelinearne odnose (Oreški, 2014).

Pearsonov produkt moment koeficijent korelacije (engl. *product moment correlation coefficient*), također poznat kao koeficijent linearne korelacije *r, R*, ili Pearsonov *r*, mjeri jačinu i smjer linearnog odnosa između dviju varijabli i definiran je pomoću kovarijance varijabli (uzorka) podijeljenih s njihovim standardnim devijacijama. Matematička formula za računanje je (Niven i Deutsch, 2012):

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{4.1}$$

To također može biti napisano kao:

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2}} \tag{4.2}$$

gdje su $\overline{x}$ i $\overline{y}$ sredine uzoraka od X i Y, a *n* je broj parova podataka.

Vrijednost koeficijenta *r* je takva da je $-1 \leq r \leq +1$. Predznak koeficijenta korelacije je pozitivan ako su varijable izravno povezane i negativan ako su obrnuto povezane. Bliskošću s 1 ili -1 mjeri se jakost linearnog odnosa. Savršena korelacija ± 1 javlja se samo kad sve točke leže točno na ravnoj liniji. Ako je predznak koeficijenta korelacije pozitivan, nagib te linije je pozitivan i obrnuto. U nekim slučajevima svega nekoliko ekstremnih vrijednosti smanjuje inače visoku korelaciju. U ovom radu tehnika korištena za odabir atributa koristi Pearsonov koeficijent korelacije. Ovaj koeficijent korelacije nije mjerilo ukupnih odnosa jer ne daje informaciju o tome postoji li ili ne postoji nelinearni odnos između dviju varijabli. Stoga, korelacija 0 ne mora nužno značiti da su varijable nezavisne (Oreški, 2014).

4.3.2. Relief

Tehnike koje odabir baziraju na udaljenosti traže takav podskup atributa gdje su primjeri iz različitih klasa odvojeni velikom udaljenosti, dok su oni iz iste klase blizu jedni drugima. Često korištena tehnika koja se bazira na udaljenosti je Relief (Freeman et al., 2015).

Glavna ideja Relief algoritma jest procijeniti kvalitetu atributa prema tome koliko dobro se njihove vrijednosti razlikuju između instanci (primjera) koje su blizu jedna drugoj (Robnik-Šikonja i Kononenko, 2003). Proces vrednovanja atributa tehnikom Relief izgleda ovako: algoritam prvo odabire jednu instancu slučajnim odabirom potom traži dva najbliža susjeda; jedan iz iste klase, zvan najbliži pogodak, a drugi iz alternativne klase, zvan najbliži promašaj. Nakon odabira sva tri primjera (slučajni primjer, najbliži pogodak i najbliži

promašaj) obnavlja se vektor koji sadrži vrijednosti kvalitete predviđanja za sve atribute iz odabranog primjera. Vrijednosti vektora se obnavljaju na slijedeći način:

a) ukoliko odabrani slučaj i najbolji pogodak imaju različite vrijednosti promatranog atributa tada atribut razdvaja dva primjera s istim vrijednostima klase što nije poželjno stoga se vrijednost kvalitete predviđanja za taj atribut smanjuje i
b) s druge strane, ukoliko odabrani slučaj i najbolji promašaj imaju različite vrijednosti promatranog atributa tada atribut razdvaja dva primjera s različitim vrijednostima klase što je poželjno stoga se vrijednost kvalitete predviđanja za taj atribut povećava.

Cijeli proces se ponavlja onoliko puta koliko se zada na ulaznom parametru algoritma, koji je definiran od strane korisnika.

Pretpostavimo da primjeri I_1,I_2,…,I_n čine prostor instanci i da su opisani s vektorima atributa A_i, i = 1,…,a, gdje je a broj pripadnih atributa svakog primjera, te da su označeni s labelom s vrijednosti τ_j. Tada instance predstavljaju točke u a-dimenzionalnom prostoru (Robnik-Šikonja i Kononenko, 2003). Pseudo kod algoritma Relief je prikazan na slici 4.1.

```
Ulaz: za svaki primjer iz trening skupa vektor atributa i vrijednost klase
Izlaz: vektor W procjena kvalitete atributa

1.  postavi sve težine W[A]=0;
2.  za svaki i=1 do m radi slijedeće
          a.  odaberi primjer R_i slučajnim odabirom;
          b.  pronađi najbliži pogodak H i najbliži promašaj M;
          c.  za svaki A=1 do a radi slijedeće
                  i.  W[A]=W[A]-diff(A, R_i, H)/m + diff(A, R_i, M)/m;
3.  kraj;
```

Slika 4.1 Pseudo kod algoritma Relief (Robnik-Šikonja i Kononenko, 2003)

Funkcija *diff*(A, I_1, I_2) računa razliku između vrijednosti atributa A i dvije instance I_1 i I_2. Za nominalne atribute definiran je kao (Robnik-Šikonja i Kononenko, 2003):

$$diff(A,\ I_1, I_2) = \begin{cases} 0; & value(A, I_1) = value(A, I_2) \\ 1; & suprotno \end{cases} \tag{4.3}$$

a za numeričke atribute kao:

$$diff(A,\ I_1, I_2) = \frac{|value(A, I_1) - value(A, I_2)|}{\max(A) - \min(A)} \tag{4.4}$$

Funkcija *diff* se također koristi za kalkulaciju razlike između instanci za pronalazak najbližih susjeda. Ukupna udaljenost se jednostavno računa kao suma udaljenosti svih atributa.

Jedan nedostatak primjene Relief algoritma za odabir atributa jest prisutan na redundantnim atributima. Ako je većina atributa relevantna za koncept, odabrat će ih sve, iako je možda samo nekolicina dovoljna za opis koncepta (Domazet-Lošo, 2006). Algoritam se može primijeniti samo na problemima binarne klasifikacije što predstavlja još jedno ograničenje.

4.3.3. Simetrična nesigurnost

Drugi pristup baziran na informacijsko-teorijskom konceptu entropije, je mjera nesigurnosti slučajne varijable. Entropija varijable A je definirana kao:

$$H(A) = -\sum_{i} P(a_i)\, log_2(P(a_i)) \tag{4.5}$$

a entropija za A nakon prikupljenih vrijednosti za drugu varijablu C je definirana kao:

$$H(A|C) = -\sum_{j} P(c_j) \sum_{i} P(a_i|c_j)\, log_2\left(P(a_i|c_j)\right) \tag{4.6}$$

gdje su $P(a_i)$ a priori vjerojatnosti za sve vrijednosti varijable *A*, a $P(a_i|c_j)$ su posteriori vjerojatnosti za *A* uz dane vrijednosti *C*. Iznos za koji se smanjuje entropija za varijablu *A* dodaje dodatnu informaciju o *A* koju pruža *C* i zove se informacijska dobit (Quinlan, 2014) a računa se :

$$IG(A|C) = H(A) - H(A|C) \tag{4.7}$$

Prema mjeri, klasa C se smatra da više korelira s atributom *A* nego atributom *B*, ako vrijedi $IG(A|C) > IG(B|C)$.

Poznato je da je informacijska dobit simetrična za dvije varijable (Yu i Liu, 2003), što je poželjno za mjerenje korelacije između atributa. Međutim, informacijska dobit je pristrana u korist atributa koji imaju veći broj vrijednosti. Nadalje vrijednosti moraju biti normalizirane da bi se mogle usporediti i da imaju isti utjecaj. Stoga, se koristi simetrična nesigurnost (Yu i Liu, 2003), definirana kao:

$$SU(A|C) = 2\frac{IG(A|C)}{H(A) + H(C)} \tag{4.8}$$

Mjera kompenzira pristranost informacijske dobiti prema atributima s više vrijednosti i normalizira vrijednosti u raspon [0,1] gdje vrijednost 1 indicira da znanje vrijednosti bilo koje

od dvije varijable predviđa vrijednost druge i u suprotnom vrijednost 0 indicira da su A i C nezavisni. Simetrična nesigurnost također tretira dvije varijable simetrično (Yu i Liu, 2003). Mjere koje se baziraju na entropiji koriste nominalne atribute, ali se također mogu primijeniti i za mjerenje korelacije između kontinuiranih atributa, ako su vrijednosti unaprijed diskretizirane (Liu et al., 2002). U ovom istraživanju za jednu od tehnika za odabir atributa odabrana je simetrična nesigurnost.

4.3.4. Gini indeks

Gini indeks mjeri nečistoću podataka tj. atributa. Pretpostavimo da je d n-torka primjera iz skupa podataka D, i da labela klase ima m različitih vrijednosti, koje definiraju različite klase C_i, (i=1,2,…,m). Prema vrijendostima klasa, D može biti podijeljen u m podskupa (D_i, i=1,2,…,m). Ako je D_i podksup primjera koji pripadaju klasi C_i, i d_i broj primjera u podskupu D_i, tada se Gini indeks skupa D računa (Zhu i Lin, 2013):

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2, \tag{4.9}$$

gdje je p_i vjerojatnost da n-torka iz D pripada klasi C_i, a procjenjuje se pomoću $|C_{i,D}|$ / $|D|$. Zbroj se izračunava nad m klasa. Ginijev indeks za binarni atribut A koji dijeli D na particije D_1 i D_2 je

$$Gini_A(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2). \tag{4.10}$$

Smanjenje nečistoće nastalo po atributu A je

$$\Delta Gini(A) = Gini(D) - Gini_A(D). \tag{4.11}$$

Atribut koji maksimira smanjenje nečistoće (ili, ekvivalentno, ima minimalni Ginijev indeks) je najbolje rangiran po zadanim trening n-torkama.

Ginijev indeks je tehnika odabira atributa temeljena na stupnju nečistoće u rezultirajućim particijama. Za utvrđivanje ranga pojedinog atributa potrebno je usporediti stupanj nečistoće skupa (prije podjele) sa stupnjem nečistoće rezultirajućih particija (nakon podjele). Pošto je nečistoća skupa prije podjele jednaka za sve atribute, maksimiziranje dobiti

ekvivalentno je minimiziranju srednje vrijednosti mjera nečistoća rezultirajućih particija. Što je njihova razlika veća to je odabrani atribut bolji (Oreški, 2014).

4.3.5. Omjer informacijske dobiti

Već je spomenuto da tehnika informacijske dobiti preferira atribute s većim brojem vrijednosti. Ako uzmemo primjer atributa *ID*, podjela po njemu dovest će do onolikog broja podskupova koliko imamo primjera u skupu primjera za učenje, a svaki od podskupova imat će samo jedan element. Kako svaki element u takvim podskupovima pripada samo jednoj klasi, informacija koju možemo dobiti jednaka je nuli, tj. $H(D) = 0$. Dobit je u slučaju podjele po tom atributu maksimalna, taj atribut dobiva najviši rang, a zapravo takva podjela nedonosi nikakvu korist (Oreški, 2014).

Slično kao simetrična nesigurnost, nova tehnika koja pokušava prevladati pristranost iz navedenog slučaja naziva se omjer dobiti (engl. *gain ratio*). Ona to postiže uz svojevrsnu normalizaciju informacijske dobiti pomoću novodefinirane vrijednosti "*SplitInfo*" čija vrijednost je definirana analogno s $H(D)$ formula 4.5 :

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times log_2\left(\frac{|D_j|}{|D|}\right). \tag{4.12}$$

Ova vrijednost uzima u obzir broj n-torki koje imaju određenu vrijednost atributa u odnosu na ukupan broj n-torki u D. Omjer dobiti se definira kao

$$GainRatio(A) = \frac{H(A)}{SplitInfo(A)}. \tag{4.13}$$

Ako se sad vratimo primjeru sa ID-ovim sa početka, možemo vidjeti da će u tom slučaju *GainRatio* biti mali. Ako imamo n primjera za učenje, koji spadaju u m klasa, uz uvjet da imamo puno više primjera za učenje nego klasa, lako možemo vidjeti da će *GainRatio* biti malen, jer će informacijska dobit iznositi maksimalno $\log_2(m)$ dok će SplitInfo imati vrijednost $\log_2(n)$ koja je puno veća od $\log_2(m)$. Atribut s višim omjerom je bolje rangiran po danim trening n-torkama (Oreški, 2014).

4.3.6. Homogeni i heterogeni sustavi višestrukih klasifikatora

U istraživanju, će se primijeniti dva različita načina konstruiranja SVK-a. Na jednom skupu podataka koristit će se homogeni model, koji podrazumijeva izvođenje istog algoritma za učenje na različitim podacima (Tsoumakas et al., 2009). Drugi, heterogeni, prema nekim istraživanjima superiorni način kombiniranja modela (Gashler et al., 2008), koji koristi više različitih algoritama za učenje na istom ili različitom skupu podataka, će biti primijenjen na drugom skupu kreditnih podataka. Klasifikacijski algoritmi koji će se koristiti su: stabla odluke, tehnika potpornih vektora, neuronske mreže, logistička regresija i Ripper algoritam. Algoritmi neće biti detaljno opisani, za pronalazak više infromacija o istima se preporučuje knjiga, „*Data classification: algorithms and applications*“ (Aggarwal, 2014) i Weka dokumentacija[1] u kojima su opisani.

4.4 Razvoj tehnike

Predložena tehnika DFSE (engl. *Directed Feature Selection Ensemble*) je razvijena u svrhu provjere hipoteze istraživanja, odnosno s ciljem poboljšanja performansi klasifikacije pojedinačnih modela na temelju razlike proizašle iz odabira atributa i kombiniranja najboljih pojedinačnih modela. DFSE je robusna tehnika koja je široko primjenjiva, a njezina implementacija ne zahtjeva temeljito poznavanje metoda strojnog učenja. Iako je poželjno koristiti prethodno prikupljeno znanje za određeni skup podataka, tehnika ne podrazumijeva interveniranje u parametre algoritama već poboljšanje rezultata temelji na raznolikosti odabranih algoritama za odabir atributa i kombiniranju najboljih modela.

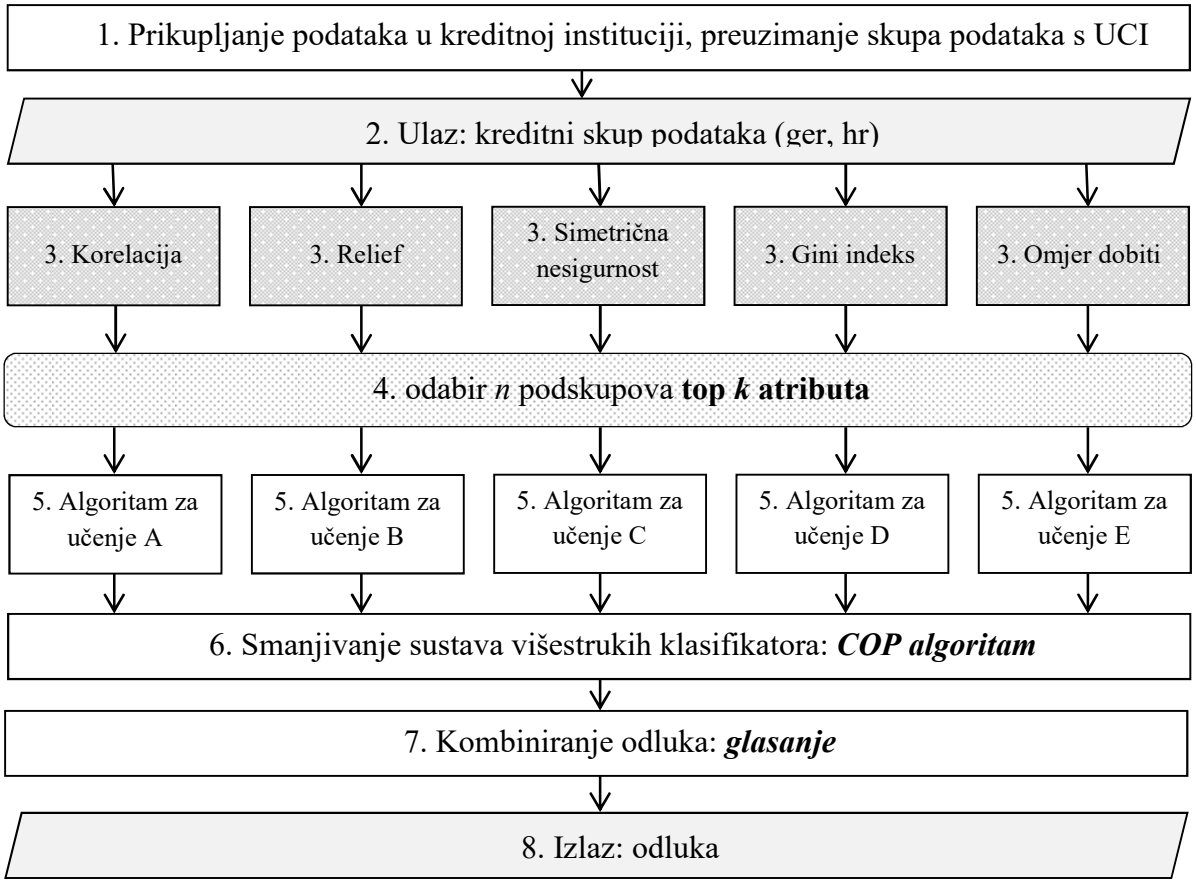
Tehnika je grafički prikazana na slici 4.2. Za razliku od konstruiranja SVK-a koji se zasniva na manipulaciji primjera iz ulaznog skupa podataka, inicijalni skup podataka se prosljeđuje u originalnom obliku sa svim primjerima.

Korištenje više različitih filtarskih tehnika znači da svi atributi ne moraju nužno biti izabrani u procesu učenja. Prema tome, ne postoji iscrpno korištenje svih atributa na način da se cijeli skup podijeli u manje podskupove koji se dostavljaju klasifikatorima, već se u potpunosti odbacuju oni atributi koji nisu prepoznati kao bitni od niti jedne tehnike, ukoliko

[1] Weka dokumentacija je dostupna na stranicama http://www.cs.waikato.ac.nz/ml/weka/documentation.html (datum posljednjeg otvaranja 7.10.2015)

takvi postoje. U ovom koraku nagalasak je na korištenju različith mjera za vrednovanje kvalitete atributa. Na primjer, mjere koje su korištene u istraživanju su:

a) bazirane na udaljenosti: odabiru se oni atributi koji podupiru instance iste klase da ostanu unutar iste udaljenosti. Instance iste klase trebaju biti bliže u smislu udaljenosti od instanci drugih klasa. Tehnika Relief koristi mjeru udaljenosti za ocjenu kvalitete atributa (García et al., 2015).
b) bazirane na informacijama, a uključuju mjere: entropiju, informacijsku dobit i zajedničku informaciju. Zajedničko tim mjerama jest da se baziraju na entropiji tj. mjeri nesigurnosti slučajne varijable. Od odabranih tehnika tu spadaju omjer informacijske dobiti, Gini indeks i simetrična nesigurnost (Srivastava, 2013).
c) bazirane na zavisnosti podataka: mjera zavisnosti istražuje korelaciju između atributa i klase, i koliko je atribut povezan s rezultatom u smislu vrijednosti klase. Tehnika korelacije koja se koristi u ovom istraživanju bazira se na mjeri zavisnosti podataka (Srivastava, 2013).



Slika 4.2 Grafički prikaz tehnike DFSE po fazama

Iako je u ovom istraživanju korišteno pet različitih tehnika, taj broj se može jednostavno povećati, a prednost je korištenje što različitijih tehnika i mjera za vrednovanje atributa tako da korisnik ne mora brinuti da li je odabrao adekvatnu tehniku jer one pokrivaju široki spektar mjera i mogu biti primijenjene na podacima s različitim karakteristikama.

Slijedeći korak je odabir veličine skupa atributa. Odabir najbolje rangiranih atributa, temeljem različitih kriterija, osigurava klasifikatorima dobru osnovu za izradu prediktivnih modela. Postizanje raznolikosti odluka je glavni cilj korištenja različitih tehnika, međutim i točnost klasifikacije je bitan faktor, a povećanje raznolikosti na štetu točnosti može negativno utjecati na performanse. U tom kontekstu moguće je definirati dva tipa problema:

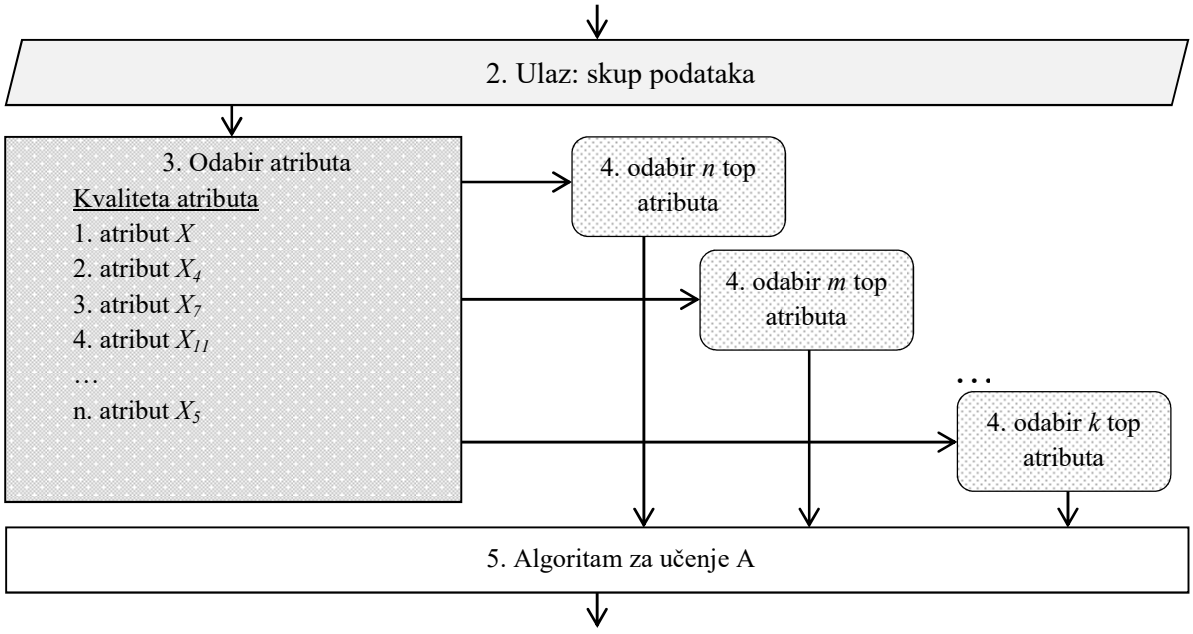
a) problem tipa A se događa kada trenirani modeli međusobno ostvaruju dobru raznolikost ali pojedinačno imaju lošu točnost klasifikacije i
b) problem tima B se događa kada klasifikatori ostvaruju veliku točnost ali međusobno imaju malu raznolikost.

Iako je u teoriji moguće izraziti savršenu raznolikost, u stvarnosti ona je isključivo dostupna u slučaju da su klasifikacijski modeli slučajni tj. da im je točnost predviđanja 50%. To je ekstreman slučaj koji dokazuje da raznolikost nije jedini bitni faktor u kombiniranju klasifikatora. S druge strane kombiniranje dobrih klasifikatora koji ne prave razlike u greškama ne može poboljšati ukupne rezultate.

Pitanje koje se nameće jest koliko najboljih atributa je potrebno odabrati od svake tehnike da se postigne zadovoljavajuća raznolikost odgovora i točnost klasifikacije? Jedan način je pojedinačno, za svaki skup podataka, empirijski ispitati koliko atributa iz pojedine tehnike je potrebno odabrati da bi se postigli najbolji rezultati klasifikacije. Međutim to je vremenski zahtjevno i ne doprinosi jednostavnosti modela. DFSE tehnika primjenjuje jednostavniji i vremenski promatrano brži pristup tako da odabire više podskupova atributa za svaku tehniku odjednom.

Broj podskupova koji će biti odabrani kao i broj atributa koji čine podskupove ipak ovise o skupu podataka i procjenjuju se temeljem iskustva istraživača te značajno ovise o domeni istraživanja. Odabir većeg broja podskupova će rezultirati s više pojedinačnih modela što utječe na veću raznolikost klasifikatora. Ako postoji dilema da li izabrati npr. 4 ili 5 podskupova atributa, onda treba istaknuti da tehnika DFSE u kasnijoj fazi odbacuje one kombinacije atributa koje ne daju pozitivan doprinos konačnom rješenju. Stoga sam odabir broja podskupova i broja atributa u tim podskupovima kod DFSE tehnike nema presudni utjecaj na konačni rezultat klasifikacije. Isječak slike 4.2 koji obuhvaća jednu tehniku za odabir atributa, a prikazuje odabir više podskupova atributa je dan slikom 4.3.

Nakon odabira tehnika te broja i veličine podskupova slijedi klasifikacija podataka. Jednako kao što se odabir atributa može zasnivati na različitim mjerama tako se mogu kombinirati i klasifikacijski algoritmi. Algoritmi funkcioniraju na temelju izvođenja nekog oblika pretrage i postoji mogućnost da se zaustave u lokalnom minimumu. Na primjer, neuronske mreže s propagacijom greške unazad koriste gradijentni spust (engl. *gradient descent*) za minimiziranje greške funkcije, a algoritmi stabla odluke koriste pohlepno pravilo dijeljenja (engl. *greedy splitting rule*) za stvaranje stabla. U slučajevima gdje postoji veliki broj lokalnih optimuma u trening podacima, može biti računalno vrlo zahtjenvno pronaći najbolju hipotezu. SVK konstruiran tako da pretraga kreće iz različitih startnih točaka može pružiti bolju aproksimaciju istinite, nepoznate hipoteze koja je bolja od bilo kojeg individualnog klasifikatora.



Slika 4.3 Isječak tehnike DFSE koji detaljnije prikazuje korak odabira top atributa

Bitan korak nove tehnike jest uklanjanje modela koji ne pridonose performansama sustava. Generiranje velikog broja modela koji su prethodno trenirani na temelju skupova podataka dobivenih pomoću različitih tehnika za odabir atributa, različitih veličina podskupova atributa i različitih algoritama za učenje može rezultirati s pojedinačnim modelima lošijih performansi. Primjena algoritma COP za smanjivanje SVK-a je korekcijski korak izrade modela, koji ima za cilj isključivanje loših klasifikatora nastalih uslijed lošeg odabira ili ne kompatibilnosti tehnika za odabir atributa i algoritama za učenje.

```
F – vektor odabranih filtarskih tehnika, M – vektor veličina podskupova
 a) za svaki f od 1 do |F|
     i. za svaki k od 1 do |M|
              i. odaberi M(k) top atributa A koristeći filtar F(f)
             ii. napravi klasifikator C_{f,k} s odabranim atributima A
            iii. dohvati predikciju P_{f,k} od klasifikatora C_{f,k}
    ii. kraj k
 b) kraj f
 c) odaberi klasifikatore temeljem algoritma COP na predikcijama P_{1,1}-P_{f,k}
 d) primjeni funkciju glasanja na odabranim predikcijama P_{1,1}-P_{f,k}
 e) dohvati predikciju P
```

Slika 4.4 Pseudo kod tehnike DFSE

Posljedni korak je kombinacija klasifikatora glasanjem. Odabrani modeli se kombiniraju pomoću glasanja većine gdje se odabire ona klasa za koju je glasala većina članova. DFSE tehnika se može opisati pomoći presudo koda prikazanog na slici 4.4.

Kompleksnost SVK-a predloženog u istraživanju zavisi o korištenim algoritmima za strojno učenje. Uzmimo da su K i J kompleksnosti algoritama za odabir atributa i algoritama za učenje, respektivno, a F broj filtera korištenih u SVK. Tada bi kompleksnost bila F*max(K,J). Kod procjene kompleksnosti algoritma zanemarujemo konstante (Oreski, 2014). Tako je svejedno ima li neki algoritam $5n^2$, $7n^2$, ili $1000n^2$ koraka, važno je samo da se kompleksnost ponaša kao kvadrat od n. Kažemo da predmetni algoritam zahtijeva vrijeme $O(n^2)$ gdje je veliko O Landauov simbol.

Stoga, promatrajući kompleksnost algoritma SVK može se reći da je kompleksnost takvog SVK određena onom metodom koja ima veću kompleksnost (K ili J) te da nije računalno kompleksniji od samostalno korištenih klasifikatora.

4.5 Postavke eksperimenta

4.5.1. Skupovi podataka

DFSE tehnika će se testirati na hrvatskom i njemačkom skupu kreditnih podatka. Osnovne karakteristike oba skupa su prikazane u tablici 4.1.

Tablica 4.1 Osnovne karakteristike korištenih skupova podataka

Naziv skupa	Broj primjera	Broj atributa	Broj klasa	Omjer klasa
Hrvatski kreditni skup	1000	35	2	75:25
Njemački kreditni skup	1000	24	2	70:30

Istraživanje se provodi na dva različita skupa kreditnih podataka. Prvi je u tu svrhu prikupljen unutar jedne hrvatske kreditne institucije, a drugi, njemački kreditni skup je javno objavljeni u online repozitoriju za strojno učenje. Skupovi podataka će biti ukratko opisani u nastavku poglavlja.

Prvi skup podataka iz primarnog izvora odnosi se na kreditne podatke hrvatske kreditne institucije. Unutar dostupne baze podataka promatrat će se jedinke tj. odobreni krediti fizičkim osobama čija je vrijednost uzetog kredita manja ili jednaka 100.000,00 HRK i koji imaju tekući račun u banci najmanje 15 mjeseci prije dana odobravanja kredita. Ukupnu populaciju čine sve jedinke koje su uzele kredit u razdoblju od rujna 2004. do rujna 2006. godine, a čije je vraćanje kredita praćeno do rujna 2011. godine.

Unutar populacije postoje dvije homogene podskupine: klijenti kojima je odobren kredit i koji su ga uspješno vratili (grupa 1) i klijenti kojima je odobren kredit ali su imali problema prilikom vraćanja kredita (grupa 2). Smatra se da je klijent imao problema prilikom vraćanja kredita ukoliko nije uspješno izvršio svoje obveze vraćanja rate kredita u roku od 90 dana od dogovorenog roka, tada pripada grupi 2. Grupa 1 predstavlja pozitivne primjere u skupu podataka a grupa 2 negativne primjere. Navedeno pravilo je u skladu s Baselskim sporazumom o kapitalu koji definira loše klijente kao one koji imaju prekoračenje vraćanja obveza preko 90 dana, iznimno u nekim državama 180 dana od dogovorenog roka (Basel, 2014).

Svaka jedinka je inicijalno opisana s ukupno 37 varijabli, nakon obrade podataka i odstranjivanja varijabli zbog činjenice da imaju identičnu vrijednost ili imaju ekstremno visoku korelaciju, preostale su 33 regularne varijable te dvije posebne varijable (identifikator i labela). Sve varijable su podijeljene u ukupno pet glavnih grupa: (1) osnovne karakteristike (2) povijest plaćanja (mjesečni prosjek) (3) financijski pokazatelji (4) povijest neplaćanja (5) iskustva s prijašnjim kreditima (Oreški et al, 2012). Po tako utvrđenim pravilima u prethodnom istraživanju (Oreški et al, 2012) kreiran je probabilistički uzorak koji će se koristiti u istraživanju. Radi se o stratificiranom uzorku u omjeru 75% grupe 1 i 25% grupe 2.

Obzirom da se kreditni modeli podataka temelje na podacima već odobrenih kredita u bankama, u inicijalnom skupu podataka omjer između grupe 1 i grupe 2 u normalnim ekonomskim uvjetima je otprilike 96% prema 4%, respektivno. S padom ekonomske moći građana uslijed financijske krize, posljednjih godina taj omjer iznosi otprilike 90% prema 10% u korist grupe 1, s daljom tendencijom povećanja udjela grupe 2. Međutim u inicijalni skup podataka ne ulaze kreditni zahtjevi koje su klasificirani kao loši od strane kreditnih referenata,

stoga navedeni omjeri podataka nisu realni u odnosu na stvarni omjer s kojim se banke susreću. Postotak odbijenih kredita ovisi o pojedinačnoj banci i njezinoj politici te ne postoji jedinstveni postotak odbijenih kredita. Procjenjuje se da se taj postotak kreće između 15% i 35%. Takav omjer je u skladu s praksom i preporukama istraživača koji su koristili iste ili jako slične omjere (Abdou et al., 2008; Šušteršič et al., 2009). Uzorak na kraju sačinjavaju 1000 jedinki od kojih 750 jedinki iz grupe 1 i 250 kredite iz grupe 2. Tablica s deskriptivnom statistikom hrvatskog skupa podataka je dana u prilogu A ove disertacije.

Drugi skup podataka, iz sekundarnog izvora, na kojem će se paralelno provesti istraživanje je njemački skup kreditnih podataka. Originalni skup podataka je objavio prof. Hoffman i danas je najkorišteniji u procjeni kreditnog rizika građana (Zhang et al., 2007; Zhou et al., 2011; Han et al., 2012; Marques et al., 2012; Zhu et al., 2013). Skup podataka sadrži 1000 slučajeva od kojih 700 kredita klasificiranih kao dobri i 300 kredita klasificiranih kao loši. Svaki slučaj je opisan s 24 različite varijable. Skup podataka je preuzet iz repozitorija Sveučilišta iz Aucklanda (https://www.stat.auckland.ac.nz/~reilly/, 28.09.2015)

4.5.2. Klasifikacija i evaluacija

Obrada je provedena pomoću programa Waikato Environment for Knowledge Analysis (*WEKA)* verzije 3.6.10, dostupnog na službenim web stranicama[2]. Svi testovi su izvršeni na računalu konfiguracije Intel Core i3 CPU 2.13 GHz, 4GB RAM, Win 7 64bit.

Testiranja su provedena na dva prikupljena skupa podataka opisana u prethodnom odjeljku. Postavke i parametri korišteni za izradu SVK-a nisu identični na oba skupa podataka, već su testirane dvije moguće izvedbe predloženog modela. Tehnike za odabir atributa su iste u oba slučaja a radi se o implementacijama iz „*attributeSelection*“ paketa. Odabrane su: korelacija, Relief, simetrična nesigurnost, Gini indeks i omjer informacijske dobiti. Parametri tehnika nisu mijenjani tj. korišteni su inicijalno zadani parametri programa.

Odabir atributa je dinamička faza kreiranja sustava te ovisi o skupu podataka tj. korisniku. Moguće je odabrati više podskupova podataka s različitim brojem atributa (slika 4.3). Način odabira atributa za predmetno istraživanje je prikazan u tablici 4.2.

Omjer između odabranih i ukupnog broja atributa se procjenjuje prilikom konstruiranja sustava, ukoliko je omjer pogrešno odabran i ne doprinosi kvaliteti isti će biti izbačen pomoću

[2] Program Weka se može preuzeti na stranici http://www.cs.waikato.ac.nz/~ml/weka/index.html (poveznica otvorena 19.09.2015); korisnički priručnik za upotrebu alata dostupan je na web stranici http://www.cs.waikato.ac.nz/~ml/weka/documentation.html (poveznica otvorena 07.10.2015)

COP algoritma. Stoga korisnik može istraživati različite mogućnosti jer model kasnije odbacuje „pogrešne“ procjene nastale tijekom izrade modela.

Tablica 4.2 Pregled veličine odabranih podskupova atributa i omjer u odnosu na inicijalni skup

RB	Hrvatski skup podataka		Njemački skup podataka	
	Broj atributa	Omjer u %	Broj atributa	Omjer u %
1.	12	34	9	37
2.	16	45	12	50
3.	20	57	15	62
4.	24	68	18	75
5.	28	80	23	95

Klasifikacija podataka za hrvatski skup podataka će se izvoditi pomoći algoritma neuronskih mreža implementiranog u alatu Weka; modul pod nazivom „*MultilayerPerceptron*“. Parametri za trening pojedinačnih 25 modela (5 tehnika odabira atributa * 5 različitih veličina podskupa najboljih atributa) su isti i nisu mijenjani. Vrijednosti parametara neuronskih mreža su dani u tablici 4.3.

Tablica 4.3 Parametri za neuronske mreže za hrvatski skup podataka

Naziv	Vrijednost
trening ciklusi	500
stopa učenja	0.6
momentum	0.3
slučajan odabir	da
normalizacija	da

Parametri *stopa učenja* i *momentum* su izmijenjeni u odnosu na inicijalne vrijednosti i to na način da je njihov omjer postavljen sukladno ranijim istraživanjima. Njihove vrijednosti nisu rezultat pretrage već su aproksimativno postavljene u omjeru 2:1, koji se ranije pokazao dobrim. U konstruiranju homogenog klasifikatora vrlo je bitno koristiti ispravan klasifikator prilagođen podacima jer izostankom dobrog klasifikatora točnost klasifikacije padne na razinu da niti veća raznolikost ne može unaprijediti performanse na željenu razinu; problem tipa A.

Na njemačkom skupu podataka će biti primijenjen drugačiji pristup koji koristi više klasifikacijskih algoritama za trening modela. Algoritmi koji će se koristiti su: stabla odluke, tehnika potpornih vektora, neuronske mreže, logistička regresija i Ripper algoritam. Korištene implementacije (*LMT, SMO, MultilayerPerceptron, Logistic i JRip*) odabranih algoritama se nalaze u „*classifier*“ paketu Weka programa. Parametri algoritama nisu mijenjani u odnosu na njihove inicijalne postavke.

Generiranje sustava za oba skupa podataka je ponovljeno 30 puta sa slučajnim odabirom seed-a validacije. U izradi modela korištena je unakrsna validacija s 10 preklapanja (engl. *10 fold cross validation*).

4.6 Komparacija rezultata

Za ocjenu učinkovitosti modela koristi se mjera točnost klasifikacije. Do mjere se dolazi tako da se računa koliko je primjera iz testnog skupa podataka klasifikator točno klasificirao. Te vrijednosti se uvrštavaju u tablicu pod nazivom matrica konfuzije (engl. *confusion matrix*). Tablica 4.4 prikazuje matricu konfuzije za binarni klasifikacijski problem. Svaka pozicija N_{ij} u tablici označava broj primjera iz klase *i* koji su klasificirani kao klasa *j*. Na primjer, N_{10} je broj primjera iz klase 1 netočno klasificiranih kao klasa 0. Prema unosima u matricu konfuzije računa se: ukupan broj točnih predikcija promatranog modela (N_{11}+N_{00}) i ukupan broj netočnih predikcija (N_{10}+N_{01}) (Pang-Ning et al., 2006).

Tablica 4.4 Binarna matrica konfuzije

		Rezultat klasifikacije	
		1	0
Istinita klasa	1	N_{11}	N_{10}
	0	N_{01}	N_{00}

Iako matrica konfuzije pruža informaciju potrebnu za određivanje kvalitete klasifikacije podataka nekog modela, prikazivanje tih podataka kao jedinstvenog broja je mnogo prikladnije za usporedbu performansi različitih modela (Pang-Ning et al., 2006). Jedna od metrika performansi jest točnost, a definira se na slijedeći način:

$$acc = \frac{N_{11} + N_{00}}{N_{11} + N_{10} + N_{01} + N_{00}}. \tag{4.14}$$

U komparaciju učinkovitosti promatranih modela uključeni su: najtočniji klasifikator iz SVK-a, svi klasifikatori iz SVK-a, tj. ukupni SVK te smanjeni SVK nakon primjene algoritma COP. Potrebno je provjeriti hipotezu istraživanja da SVK nakon primjene algoritma COP postiže statistički značajno veću točnost klasifikacije od klasifikatora uključenih u sustav na razini statističke značajnosti $p \leq 0{,}05$.

Odluka o postojanju značajnih razlika u učinkovitosti promatranih modela ne smije se temeljiti samo na analizi dobivenih rezultata temeljem prethodno stečenog iskustva i znanja

istraživača. Potrebno je na znanstveni način pokazati da su, primjerice, promatrane razlike između izmjerenih srednjih vrijednosti za različite modele tijekom pokusa na nekom skupu podataka statistički značajne. Postaje jasno da su za ispitivanje statističke značajnosti potrebni statistički testovi kako bi se mogli donijeti zaključci da je nešto postiglo ili nije postiglo statistički značajnu razliku (Marusteri i Bacarea, 2010).

Ispitivanje statističke značajnosti i donošenje statističkih odluka temeljem eksperimentalnih podataka gotovo uvijek se izvodi pomoću takozvanih testova za ispitivanje nulte hipoteze. Tekst ispitivanja nulte hipoteze za ispitivanje razlike obično ima oblik: „Ne postoji (statistički značajna) razlika između skupina“ (Marusteri i Bacarea, 2010). Alternativna se hipoteza ne može potvrditi. Možemo samo odbaciti nultu hipotezu (u tom slučaju prihvaćamo alternativnu hipotezu) ili prihvatiti nultu hipotezu (Marusteri i Bacarea, 2010).

Mora se primijetiti da istraživač ne može biti 100% siguran u rezultat testa o promatranoj razlici, čak i kada je ta razlika statistički značajna. Radi kontrole nesigurnosti uveden je pojam razina značajnosti (engl. *significance level*, α ili alpha). Pojednostavljeno, razina značajnosti može se definirati kao vjerojatnost odluke o odbijanju nulte hipoteze kada je nulta hipoteza zapravo istinita. U statistici takva odluka je poznata kao pogreška tipa I ili lažno pozitivna odluka. Najčešće korištene razine značajnosti su 5%, 1% i 0,1%, što empirijski odgovara razini pouzdanosti od 95%, 99% i 99,9% (Marusteri i Bacarea, 2010).

Dva glavna statistička testa smatraju se najprikladnijim za ispitivanje postojanja ili ne postojanja statistički značajne razlike u rezultatima većeg broja klasifikatora nad više međusobno neovisnih uzoraka. Prvi od njih je vrlo dobro poznati parametarski test za testiranje većeg broja hipoteza: analiza varijance (ANOVA) za ponovljena mjerenja. Drugi je rjeđe korištena neparametarska alternativa navedenom testu - Friedmanov test (Japkowicz i Shah, 2011). ANOVA test za ponovljena mjerenja je generalizacija uparenog *t* testa (*t* testa za ponovljena mjerenja) koja može usporediti performanse različitih klasifikatora kroz različite skupove podataka s ciljem utvrđivanja da li su promatrane razlike statistički značajne. *Null* hipoteza je da razlike u performansama klasifikatora kroz različite skupove nisu statistički značajne. Ako odbijemo *null* hipotezu tada možemo zaključiti da najmanje jedan klasifikator ima različite performanse od ostalih.

Međutim, Demšar (2006) navodi da takva usporedba može biti konceptualno neprimjerena i statistički nesigurna jer se parametarski testovi temelje na različitim pretpostavkama (normalnosti, homogenosti varijanci) koje često nisu zadovoljene zbog prirode problema.

Friedmanov test je neparametarska verzija ANOVA testa za ponovljena mjerenja. Dok parametarski test ANOVA pretpostavlja normalnu distribuciju i homogene varijance, Friedmanov test nema tih ograničenja (Demšar, 2006). Cijena te slobode je plaćena manjom snagom Friedmanova testa u usporedbi s parametarskim ANOVA testom.

Statistika Friedmanova testa se temelji na prosječnim rangovima (R) performansi klasifikacijskih algoritama po svim skupovima podataka, a računa se na sljedeći način:

$$\chi_F^2 = \frac{12N}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right] \quad \text{,gdje je} \quad R_j = \frac{1}{N}\sum_i r_i^j \tag{4.15}$$

U formuli (4.15) N označava broj različitih skupova podataka (različitih testova) za koje su izvršena mjerenja, k je broj kompariranih klasifikatora, a r_i^j je rang j-tog klasifikatora na i-tom skupu podataka. Statistika za Friedman test je Hi-kvadrat s [k-1] stupnjeva slobode (Demšar, 2006).

Pretpostavke potrebne za Friedmanov test su kako slijedi:

1) Podaci se sastoje od N međusobno neovisnih uzoraka nad kojima je konstruirano k modela. Neovisni uzorci ili blokovi su kod nas različiti skupovi podataka za koje su izvršena mjerenja, a k je broj kompariranih klasifikatora.
2) Varijabilne od interesa su kontinuirane ili barem ordinarne.
3) Nema interakcija između blokova i tretmana.
4) Promatranja u svakom bloku mogu biti rangirana prema veličini.

Sve pretpostavke za Friedmanov test su u našem slučaju ispunjene.

Hipoteze kod Friedmanova testa su:

H_0: Nema razlike u distribucijama rangova kod k modela.

$$H_0: M_1 = M_2 = \cdots = M_k$$

H_A: Distribucije rangova su različite, tj. najmanje jedna od jednakosti nije zadovoljena.

4.6.1. Post hoc analiza Friedmanova testa

Ako Friedmanov test da signifikantnu p-vrijednost, to znači da neki od k modela u podacima ima različitu distribuciju od nekog drugog modela, ali iz tog se pokazatelja još ne zna koji se model statistički značajno razlikuje od kojeg. Stoga je cilj sljedećeg koraka u analizi, pronaći koji parovi modela su statistički značajno međusobno različiti. Provjera svih parova zahtjeva višestruke komparacije koje se izvode nekim od post hoc testova koje možemo shvatiti kao

Wilcoxon testove za uparene uzorke s korekcijom za višestruka izvođenja. U empirijskoj analizi izvest ćemo Nemenyi post hoc test.

4.7 Empirijska analiza

U poglavlju empirijska analiza su izneseni rezultati istraživanja dva različita sustava višestrukih klasifikatora na dva skupa podataka. Ukratko, homogeni sustav višestrukih klasifikatora koristi neuronske mreže (NN) za algoritam učenja i sastoji se od 25 zasebnih modela kreirana i testirana na hrvatskom skupu podatka. Drugi, heterogeni sustav koristi pet različitih klasifikacijskih algoritama za trening: stabla odluke (SO), tehniku potpornih vektora (PV), neuronske mreže, logističku regresiju (LR) i Ripper (RI). Također se sastoji od 25 modela kreiranih i testiranih na njemačkom skupu podataka. Oba sustava sadrže pet različitih tehnika za odabir atributa: omjer informacijske dobiti (OD), RELIEF (RE), Gini indeks (GI), mjera nesigurnosti (MN) i korelacija (KO).

Radi lakšeg referenciranja modela u analizi i tablicama će se koristiti skraćene oznake na slijedeći način: prvi dio oznake predstavlja skraćenicu korištene tehnike za odabir atributa, drugi dio je veličina odabranog podskupa najboljih atributa, a treći algoritam klasifikacije. Npr. oznaka „*OD-12-NN*“ označava model treniran na 12 najznačajnijih atributa prema odabiru tehnike omjer informacijske odbiti s klasifikatorom neuronske mreže.

4.7.1. Rezultati na njemačkom skupu

Tablica 4.5 prikazuje rezultate svih 30 testiranja na njemačkom skupu podataka. Promatrane vrijednosti se odnose na: odabrani podskup klasifikatora odabran COP algoritmom uključenim u DFSE tehniku, sustav višestrukih klasifikatora sa svim članovima, kombiniran glasanjem (SVK), pojedinačni model koji ostvaruje najtočnije rezultate, prosječnu točnost svih modela i mjeru raznolikosti Q za ukupni sustav. Kolona pod nazivom „razlika“, prikazuje razliku točnosti DFSE sustava i promatrane točnosti. Na podnožju tablice je izražen prosjek svih mjerenja.

Prosječna veličina sustava dobivenog DFSE tehnikom je 6,7 što je približno 7 članova. U odnosu na 25 tj. ukupan broj klasifikatora koji je treniran u svakom eksperimentu, prosječna veličina čini 28%. Izmjerene vrijednosti se kreću između minimalne vrijednosti 1 i maksimalne 13 i ne koreliraju s Q mjerom za raznolikost. U primjerima gdje je odabrana

veličina 1 točnost sustava DFSE je manja od prosječne vrijednosti, dok već kombinacijom od 3 klasifikatora ta vrijednost raste.

Tablica 4.5 Rezultati testiranja DFSE tehnike na njemačkom skupu podataka

RB	DFSE		SVK		Najtočniji član		Prosjek SVK		SVK Q
	Veličina	Točnost	Točnost	Razlika	Točnost	Razlika	Točnost	Razlika	
t1	5	0,777	0,766	-0,011	0,775	-0,002	0,747	-0,030	0,9018
t2	3	0,78	0,761	-0,019	0,777	-0,003	0,747	-0,033	0,8999
t3	3	0,786	0,774	-0,012	0,78	-0,006	0,751	-0,035	0,9077
t4	7	0,78	0,764	-0,016	0,778	-0,002	0,748	-0,032	0,908
t5	7	0,783	0,772	-0,011	0,775	-0,008	0,749	-0,034	0,8895
t6	9	0,779	0,767	-0,012	0,774	-0,005	0,748	-0,031	0,9021
t7	13	0,778	0,767	-0,011	0,772	-0,006	0,747	-0,031	0,902
t8	3	0,779	0,764	-0,015	0,776	-0,003	0,747	-0,032	0,9087
t9	3	0,782	0,766	-0,016	0,776	-0,006	0,745	-0,037	0,8983
t10	3	0,778	0,769	-0,009	0,772	-0,006	0,747	-0,031	0,8993
t11	3	0,781	0,77	-0,011	0,773	-0,008	0,747	-0,034	0,9012
t12	13	0,78	0,773	-0,007	0,777	-0,003	0,753	-0,027	0,9046
t13	5	0,782	0,773	-0,009	0,776	-0,006	0,751	-0,031	0,9101
t14	13	0,782	0,775	-0,007	0,777	-0,005	0,753	-0,029	0,9014
t15	5	0,778	0,764	-0,014	0,77	-0,008	0,750	-0,028	0,8978
t16	1	0,777	0,763	-0,014	0,777	0	0,746	-0,031	0,9077
t17	9	0,788	0,771	-0,017	0,776	-0,012	0,750	-0,038	0,9024
t18	7	0,781	0,77	-0,011	0,774	-0,007	0,749	-0,032	0,9035
t19	7	0,783	0,766	-0,017	0,776	-0,007	0,747	-0,036	0,8989
t20	5	0,78	0,765	-0,015	0,773	-0,007	0,747	-0,033	0,9047
t21	3	0,781	0,766	-0,015	0,779	-0,002	0,749	-0,032	0,8993
t22	11	0,781	0,769	-0,012	0,775	-0,006	0,752	-0,029	0,9071
t23	11	0,782	0,763	-0,019	0,776	-0,006	0,751	-0,031	0,9051
t24	5	0,78	0,769	-0,011	0,777	-0,003	0,748	-0,032	0,9018
t25	11	0,778	0,768	-0,01	0,773	-0,005	0,748	-0,030	0,9005
t26	13	0,785	0,772	-0,013	0,774	-0,011	0,749	-0,036	0,8993
t27	7	0,781	0,767	-0,014	0,779	-0,002	0,751	-0,030	0,9002
t28	1	0,779	0,768	-0,011	0,779	0	0,750	-0,029	0,9059
t29	11	0,779	0,768	-0,011	0,776	-0,003	0,750	-0,029	0,9005
t30	5	0,779	0,766	-0,013	0,774	-0,005	0,748	-0,031	0,9094
Avg:	6,733	0,781	0,768	-0,013	0,776	-0,005	0,749	-0,032	0,903

Prosječna točnost smanjenog sustava iznosi 78,1% što je najveća izmjerena preciznost u usporedbi sa sličnim sustavima iz literature koja je provedena u poglavlju 5. Iako su u teoriji točnost i raznolikost povezane na način da povećanje raznolikosti dovodi do povećanja točnosti, iz tablice 4.5 se to ne može zaključiti. To pripisujemo činjenici da su razlike raznolikosti razmjerno male, a Q statistika kao mjera za raznolikost nije toliko osjetljiva da odražava ovako male promjene u točnosti.

DFSE sustav u odnosu na sustav sa svim članovima je u prosjeku točniji 1.3 postotna boda. Razlog značajnog poboljšanja se općenito može pronaći u jednom od dva razloga: (1) tenirani klasifikatori ostvaruju veliku raznolikost odluka te se time povećava točnost smanjenog sustava ili (2) pojedini klasifikatori nisu pogodni za kombiniranje i smanjuju

točnost. Visoke izmjerene vrijednosti Q mjere i značajno smanjenje veličine DFSE sustava u odnosu na originalni pokazuju da u inicijalnom skupu postoji mnogo klasifikatora koji negativno utječu na performanse sustava, što čini ranije predloženi razlog (2) vjerojatnijim.

Najtočniji član u ovom kontekstu se odnosi na klasifikator koji je ostvario najveću točnost u danom mjerenju. U višestrukim mjerenjima postoji mogućnost pojavljivanja različitih klasifikatora kao najboljih članova. Prosječno, najtočniji član iz skupa od 25 je 0.5 postotna boda manji od točnosti smanjenog sustava i 0.8 postotna boda veći od sustava koji uključuje sve članove. Odnos između najtočnijeg člana i ukupnog sustava pokazuje da su prilikom konstruiranja korišteni modeli koji nisu doprinosili raznolikosti. Iz prezentiranih rezultata je razvidno da algoritam COP uspješno odabire klasifikatore koji poboljšavaju točnost u odnosu na najbolji član, odnosno, da isključuje modele koji ne doprinose točnosti. Pojedinačno najtočniji klasifikator je „*RE-23-LR*“ s 77.3% točnosti.

Prosječna točnost svih klasifikatora je srednja vrijednost 25 klasifikatora. Poboljšanje DFSE sustava u odnosu na prosječnu točnost je 3.2 postotna boda. Iz prosječne točnosti koja ima standardnu devijaciju 0,021 može se zaključiti da postoji razlike u generiranim modelima. Razlika od 3.2 postotna boda dokazuje da je u spomenutim razlikama moguće pronaći raznolike klasifikatore.

Između 30 provedenih testova slučajnim odabirom je izabran jedan koji je predstavljen u tablici 4.6. Tablica prikazuje uključivanje klasifikatora u sustav prema ciklusima COP algoritma te samim time demonstrira način rada algoritma. U svakom ciklusu uključena su dva nova člana koja su dodana postojećim, već odabranim klasifikatorima. Algoritam obuhvaća 13 ciklusa s napomenom da prvi, odabir najtočnijeg klasifikatora, nije prikazan u tablici. Vrijednost unesena na presjeku naziva modela i broja članova (klasifikatora) označava postignutu točnost i ciklus u kojem je dotični klasifikator uključen. Osim uključivanja modela u podnožju tablice su dane i vrijednosti: točnost sustava ovisno o odabranim članovima, mjera raznolikosti sustava, najveća točnost člana i prosječna točnost SVK-a. Istaknuti stupac predstavlja odabir DFSE tehnike.

Promatrani sustav višestrukih klasifikatora ostvaruje najveću točnost 78,1% kada uključuje tri klasifikatora: „*GI-18-SO*“, „*MN-23-NN*“ i „*RE-12-LR*“. U odnosu na rezultat najtočnijeg člana to predstavlja poboljšanje od 0.8 postotnih poena, i 1.1 postotni poen u odnosu na sustav koji uključuje sve klasifikatore. Ostvareno poboljšanje je rezultat raznolikosti odgovora, koji nastaje uslijed pogrešne klasifikacije različitih primjera. Rezultati pokazuju da COP algoritam iskorištava sva 3 elementa implementirana u sustav s ciljem povećavanja raznolikosti. Odabrani klasifikatori u sustavu veličine 3 (DFSE-3) koriste: tri

različite tehnike (GI, MN, RE), tri različite veličine podskupova atributa (18, 23, 12) i tri različita algoritma (SO, NN, LR). Performanse pojedinačnih klasifikatora DFSE-3 u prosjeku iznose 75.4% što je za 0.7 postotna boda više od prosjeka DFSE-25. Poboljšanje rezultata sustava u odnosu na prosječne pojedinačne rezultate sustava DFSE-3 s 2.7 postotnih poena u odnosu na 2.3 sustava DFSE-25 potvrđuje ostvarenu raznolikost sustava s manjim brojem članova.

Tablica 4.6 Detaljan prikaz faza COP algoritma za odabrani test na njemačkom skupu podataka

Naziv modela	Broj članova DFSE tehnike											
	3	5	7	9	11	13	15	17	19	21	23	25
GI-09-SO												0,751
GI-12-SO							0,755	0,755	0,755	0,755	0,755	0,755
GI-15-SO										0,767	0,767	0,767
GI-18-SO	0,773	0,773	0,773	0,773	0,773	0,773	0,773	0,773	0,773	0,773	0,773	0,773
GI-23-SO					0,772	0,772	0,772	0,772	0,772	0,772	0,772	0,772
KO-09-PV											0,757	0,757
KO-12-PV											0,757	0,757
KO-15-PV									0,765	0,765	0,765	0,765
KO-18-PV		0,772	0,772	0,772	0,772	0,772	0,772	0,772	0,772	0,772	0,772	0,772
KO-23-PV			0,769	0,769	0,769	0,769	0,769	0,769	0,769	0,769	0,769	0,769
MN-09-NN								0,731	0,731	0,731	0,731	0,731
MN-12-NN							0,697	0,697	0,697	0,697	0,697	0,697
MN-15-NN		0,703	0,703	0,703	0,703	0,703	0,703	0,703	0,703	0,703	0,703	0,703
MN-18-NN					0,705	0,705	0,705	0,705	0,705	0,705	0,705	0,705
MN-23-NN	0,722	0,722	0,722	0,722	0,722	0,722	0,722	0,722	0,722	0,722	0,722	0,722
RE-09-LR								0,744	0,744	0,744	0,744	0,744
RE-12-LR	0,768	0,768	0,768	0,768	0,768	0,768	0,768	0,768	0,768	0,768	0,768	0,768
RE-15-LR						0,763	0,763	0,763	0,763	0,763	0,763	0,763
RE-18-LR						0,767	0,767	0,767	0,767	0,767	0,767	0,767
RE-23-LR				0,771	0,771	0,771	0,771	0,771	0,771	0,771	0,771	0,771
OD-09-RI										0,729	0,729	0,729
OD-12-RI									0,732	0,732	0,732	0,732
OD-15-RI			0,746	0,746	0,746	0,746	0,746	0,746	0,746	0,746	0,746	0,746
OD-18-RI												0,739
OD-23-RI				0,72	0,72	0,72	0,72	0,72	0,72	0,72	0,72	0,72
Q	0.851	0.877	0.875	0.879	0.876	0.891	0.891	0.889	0.892	0.894	0.901	0.901
Točnost sustava	0,781	0,775	0,774	0,775	0,771	0,772	0,776	0,773	0,775	0,771	0,77	0,77
Najtočniji član	0,773	0,773	0,773	0,773	0,773	0,773	0,773	0,773	0,773	0,773	0,773	0,773
Prosječna točnost	0,754	0,747	0,750	0,749	0,747	0,750	0,746	0,748	0,746	0,746	0,747	0,747

Ako se analizira Q vrijednost za odabrani primjer prema veličini sustava, tada je vidljivo da ona pada s povećanjem broja klasifikatora u sustavu. Promatramo li kao startnu točku DFSE-3 gdje Q postiže najmanju vrijednost 0,851 tada se može uočiti da točnost sustava s porastom Q generalno pada. Porast Q vrijednosti za 0,5 od DFSE-3 do DFSE-25 pratio je pad točnosti za 1.1 postotni poen. Kao što je već ranije utvrđeno, iako zavisnost između raznolikosti i točnosti postoji, ne postoji idealna korelacija koja bi bila evidentna na svakoj i najmanjoj promjeni veličina.

Primjer pokazuje da se Q vrijednost povećava, a raznolikost pada s povećanjem broja klasifikatora u sustavu. Iz takvih rezultata se nameće zaključak da odabir više podskupova najboljih atributa ne donosi značajniju raznolikost u treniranim modelima. Kada algoritam iskoristi sve pojedinačne tehnike za odabir atributa i klasifikacijske algoritme dolazi do značajnijeg povećanja Q vrijednosti. Najveći porast vrijednosti je vidljiv između DFSE-11 i DFSE-13 kada su uključena dva klasifikatora temeljena na istim tehnikama. Ovakvi rezultati se mogu povezati s početnom dvojbom tipa: koliko različitih podskupova atributa generirati? S obzirom na činjenicu da veći broj podskupova najboljih atributa ne donosi značajniju raznolikost u treniranim modelima, odgovor na prethodno pitanje bi bio da velik broj podskupova neće značajno pridonijeti performansama SVK-a te, također, da broj atributa u tim podskupovima ne treba biti vrlo mali.

Vidljivo je da su modeli koji se baziraju na podskupu od 9 najboljih atributa posljednji uključivani u sustav te da nisu doprinijeli raznolikosti niti točnosti. Može se zaključiti da u heterogenim sustavima (više različitih klasifikatora) modeli temeljeni na malom broju atributa nemaju izraženu specijalizaciju.

4.7.2. Rezultati na hrvatskom skupu

Rezultati primjene DFSE tehnike za konstruiranje homogenog sustava višestrukih klasifikatora na hrvatskom skupu podataka se nalaze u tablici 4.7. Kao i na njemačkom skupu podataka istraživanje je ponovljeno 30 puta sa slučajnim odabirom seed-a za validaciju. Mjereni su isti elementi kao i u prijašnjem primjeru a obuhvaćaju: DFSE sustav, SVK sustav, najtočniji član i prosječna točnost članova SVK. Kolona pod nazivom „razlika“, kao i ranije, prikazuje razliku točnosti DFSE sustava i promatrane točnosti.

Prosječna veličina smanjenog sustava je 8,67, približno 9 što je 36% originalnog sustava. Broj odabranih klasifikatora se kreće u rasponu od 3 do 17.

Točnost klasifikatora kombiniranih u DFSE sustav postiže prosječnu točnost od 83.4%. Ostvarena točnost je poboljšanje u odnosu na rezultat ostvaren u prijašnjem istraživanju primjene HGA-NN tehnike na istom skupu podatka. Najbolji rezultat HGA-NN je postigao koristeći 50 generacija genetskog algoritama, koji je u prosjeku 11 mjerenja iznosio 82.88% (Oreski i Oreski, 2014).

Rezultati točnosti svih članova uključenih u sustav i najtočnijeg člana su gotovo podjednaki, razlika iznosi 0.1 postotna boda. Pri tome valja napomenuti da se u višestrukim mjerenjima pojavljuju različiti klasifikatori kao najbolji. Stoga se „najtočniji član“ može

promatrati kao statistička kategorija, a nikako kao moguće rješenje. Iako nije ostvareno poboljšanje kombiniranjem svih klasifikatora, svako poboljšanje sustava pomoću koraka koji uključuje COP algoritam će predstavljati poboljšanje u odnosu na najtočniji član. Razlika od 0.1 potvrđuje raniju tvrdnju o pouzdanosti Q statistike kao mjere raznolikosti. U ovom slučaju prosječna Q mjera je veća nego na njemačkom skupu ali je SVK ostvario bolje rezultate u odnosu na najtočniji član. Klasifikator „GI-28-NN“ je u prosjeku 30 mjerenja ostvario najbolju točnost 82,08%.

Tablica 4.7 Rezultati testiranja DFSE tehnike na hrvatskom skupu podataka

RB	DFSE		SVK		Najtočniji član		Prosjek SVK		SVK Q
	Veličina	Točnost	Točnost	Razlika	Točnost	Razlika	Točnost	Razlika	
t1	13	0,828	0,82	-0,008	0,824	-0,004	0,799	-0,029	0,909
t2	5	0,829	0,826	-0,003	0,825	-0,004	0,801	-0,028	0,910
t3	3	0,837	0,821	-0,016	0,829	-0,008	0,799	-0,038	0,907
t4	5	0,842	0,828	-0,014	0,83	-0,012	0,802	-0,040	0,918
t5	11	0,83	0,82	-0,01	0,822	-0,008	0,800	-0,030	0,915
t6	9	0,835	0,82	-0,015	0,824	-0,011	0,800	-0,035	0,919
t7	5	0,831	0,813	-0,018	0,818	-0,013	0,798	-0,033	0,913
t8	11	0,835	0,825	-0,01	0,825	-0,010	0,798	-0,037	0,914
t9	7	0,837	0,826	-0,011	0,83	-0,007	0,800	-0,037	0,901
t10	9	0,83	0,82	-0,01	0,818	-0,012	0,800	-0,030	0,911
t11	11	0,831	0,823	-0,008	0,819	-0,012	0,801	-0,030	0,924
t12	5	0,834	0,824	-0,01	0,821	-0,013	0,800	-0,034	0,911
t13	5	0,835	0,823	-0,012	0,831	-0,004	0,799	-0,036	0,904
t14	9	0,833	0,823	-0,01	0,826	-0,007	0,798	-0,035	0,907
t15	5	0,838	0,829	-0,009	0,825	-0,013	0,799	-0,039	0,898
t16	7	0,835	0,824	-0,011	0,827	-0,008	0,802	-0,033	0,913
t17	7	0,831	0,822	-0,009	0,826	-0,005	0,801	-0,030	0,918
t18	5	0,832	0,823	-0,009	0,823	-0,009	0,799	-0,033	0,909
t19	15	0,834	0,825	-0,009	0,822	-0,012	0,800	-0,034	0,907
t20	7	0,83	0,819	-0,011	0,816	-0,014	0,796	-0,034	0,915
t21	7	0,837	0,825	-0,012	0,821	-0,016	0,797	-0,040	0,905
t22	9	0,835	0,82	-0,015	0,822	-0,013	0,800	-0,035	0,901
t23	7	0,833	0,819	-0,014	0,822	-0,011	0,799	-0,034	0,913
t24	15	0,836	0,828	-0,008	0,827	-0,009	0,803	-0,033	0,917
t25	11	0,831	0,821	-0,01	0,825	-0,006	0,801	-0,030	0,917
t26	17	0,83	0,828	-0,002	0,822	-0,008	0,798	-0,032	0,911
t27	9	0,835	0,816	-0,019	0,823	-0,012	0,800	-0,035	0,915
t28	5	0,836	0,821	-0,015	0,824	-0,012	0,800	-0,036	0,914
t29	15	0,835	0,819	-0,016	0,824	-0,011	0,800	-0,035	0,918
t30	11	0,839	0,829	-0,01	0,825	-0,014	0,803	-0,036	0,909
Avg:	8,667	0,834	0,823	-0,011	0,824	-0,010	0,800	-0,034	0,911

U nastavku je analiziran jedan test koji je slučajnim odabirom izdvojen iz primjera od 30 testova. Odabrani primjer je prikazan u tablici 4.8. Tablica prikazuje koji članovi su odabrani u pojedinim iteracijama algoritma COP. Podskup klasifikatora koji daje najbolji rezultat istaknut je u donosu na ostale. Brojevi unutar slobodnih polja tablice predstavljaju točnost pojedinih modela.

U odabiru klasifikatora, COP algoritam prepoznaje različite tehnike i raznolikost koju one donose. U prvim iteracijama postupno se odabiru klasifikatori koji koriste različite tehnike. Sustav s 5 članova (DFSE-5) koristi svih pet tehnika i postiže najveću raznolikost 0.831. U ovom primjeru još jednom se potvrđuje da Q statistika iako najčešće korištena mjera za raznolikost nije toliko precizna da bi se njome mogla odrediti točnost nekog sustava. Sustav s najraznolikijim odlukama nije najtočniji već je treći po točnosti. Najtočniji sustav ima 7 članova i to podijeljenih 1-2-1-2-1 prema odabranim tehnikama OD-RE-GI-MN-KO, respektivno. U odabranom sustavu DFSE-7 modeli koji koriste mjeru nesigurnosti imaju najslabije, pete rezultate po pitanju točnosti, a modeli koji koriste Relief imaju treće rezultate, međutim obje su jedine odabrane dva puta. Iz takvog odabira vidljivo je da algoritam pronalazi veću raznolikost u klasifikatorima sa slabijom točnosti. Kao potvrda toga, modeli koji koriste tehniku korelacije u prosjeku imaju najbolje performanse točnosti, 4 od 5 modela je uključeno u sustav u posljednjim fazama izvođenja algoritma. Bez obzira na ostvarenu točnost u tim modelima njihova raznolikost u odnosu na ostale tehnike je premala da bi poboljšala performanse sustava.

Tablica 4.8 Detaljan prikaz faza COP algoritma za odabrani test na hrvatskom skupu podataka

Naziv modela	Broj članova DFSE tehnike											
	3	5	7	9	11	13	15	17	19	21	23	25
OD-12-NN							0,773	0,773	0,773	0,773	0,773	0,773
OD-16-NN		0,784	0,784	0,784	0,784	0,784	0,784	0,784	0,784	0,784	0,784	0,784
OD-20-NN					0,787	0,787	0,787	0,787	0,787	0,787	0,787	0,787
OD-24-NN						0,794	0,794	0,794	0,794	0,794	0,794	0,794
OD-28-NN										0,782	0,782	0,782
RE-12-NN								0,793	0,793	0,793	0,793	0,793
RE-16-NN							0,801	0,801	0,801	0,801	0,801	0,801
RE-20-NN			0,811	0,811	0,811	0,811	0,811	0,811	0,811	0,811	0,811	0,811
RE-24-NN					0,815	0,815	0,815	0,815	0,815	0,815	0,815	0,815
RE-28-NN	0,821	0,821	0,821	0,821	0,821	0,821	0,821	0,821	0,821	0,821	0,821	0,821
GI-12-NN											0,804	0,804
GI-16-NN						0,796	0,796	0,796	0,796	0,796	0,796	0,796
GI-20-NN		0,817	0,817	0,817	0,817	0,817	0,817	0,817	0,817	0,817	0,817	0,817
GI-24-NN									0,818	0,818	0,818	0,818
GI-28-NN				0,818	0,818	0,818	0,818	0,818	0,818	0,818	0,818	0,818
MN-12-NN			0,784	0,784	0,784	0,784	0,784	0,784	0,784	0,784	0,784	0,784
MN-16-NN	0,771	0,771	0,771	0,771	0,771	0,771	0,771	0,771	0,771	0,771	0,771	0,771
MN-20-NN				0,766	0,766	0,766	0,766	0,766	0,766	0,766	0,766	0,766
MN-24-NN								0,765	0,765	0,765	0,765	0,765
MN-28-NN										0,775	0,775	0,775
KO-12-NN												0,804
KO-16-NN												0,805
KO-20-NN	0,818	0,818	0,818	0,818	0,818	0,818	0,818	0,818	0,818	0,818	0,818	0,818
KO-24-NN											0,814	0,814
KO-28-NN									0,805	0,805	0,805	0,805
Q	0.887	0.831	0.897	0.901	0.906	0.900	0.900	0.896	0.904	0.894	0.900	0.905
Točnost sustava	0.828	0.83	0.837	0.831	0.83	0.829	0.828	0.825	0.824	0.823	0.823	0.825
Najtočniji član	0,821	0,821	0,821	0,821	0,821	0,821	0,821	0,821	0,821	0,821	0,821	0,821
Prosječna točnost	0,803	0,802	0,801	0,799	0,799	0,799	0,797	0,795	0,797	0,795	0,796	0,797

Najbolji pojedinačni klasifikator ima točnost 82,1%, a ispravnom kombinacijom više klasifikatora koji koriste različite tehnike za odabir atributa postignuta točnost je 83,7 što je 1,6 postotna boda više. Značajno poboljšanje dokazuje postignutu raznolikost čija je direktna posljedica. Raznolikost sustava mjerena Q statistikom, u ovom primjeru kao i u prethodnom, pokazuje porast vrijednosti s povećanjem broja članova.

4.7.3. Statistička komparacija rezultata

Kao što je u postavkama eksperimenta navedeno, za ocjenu performansi modela koristi se mjera točnost klasifikacije. Odluku o postojanju značajnih razlika u performansama promatranih modela donijet će se na temelju statističkih testova. Testovi će pokazati da li su razlike između izmjerenih srednjih vrijednosti za različite modele tijekom pokusa na eksperimentalnim skupovima podataka statistički značajne. Dva glavna statistička testa smatraju se najprikladnijim za ispitivanje postojanja ili ne postojanja statistički značajne razlike u rezultatima većeg broja klasifikatora nad više međusobno neovisnih uzoraka. Prvi od njih je parametarski test za testiranje većeg broja hipoteza: analiza varijance (ANOVA) za ponovljena mjerenja. *Null* hipoteza je da razlike u performansama klasifikatora kroz različite skupove nisu statistički značajne. Ako odbijemo *null* hipotezu tada možemo zaključiti da najmanje jedan klasifikator ima različite performanse od ostalih.

Na slici 4.5 prikazani su rezultati ANOVA testa za rezultate istraživanja na hrvatskom skupu podataka. Obzirom da je p-vrijednost ANOVA testa manja od *alpha*=0.05 može se zaključiti da za barem jedan par ostvarenih rezultata postoji statistički značajna razlika u točnosti klasifikacije. Za utvrđivanje između kojih rezultata postoji razlika korišten je Tukey test. Post-hoc test pokazuje da su rezultati temeljeni na novoj tehnici DFSE statistički značajno različiti (bolji) u donosu na ostale rezultate.

```
> aov.ex1 = aov(vrijednosti~grupa,data=doktStatTestHr)
> summary(aov.ex1)
                    One Way Analysis of Variance

Df   Sum Sq  Mean Sq F value Pr(>F)
grupa         3 0.018657 0.006219     603 <2e-16 ***
Residuals   116 0.001196 0.000010
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> TukeyHSD(aov.ex1)
                  Tukey multiple comparisons of means
                    95% family-wise confidence level

Fit: aov(formula = vrijednosti ~ grupa, data = doktStatTestHr)

$grupa
                            diff          lwr           upr     p adj
Najtocniji-DFSE    -0.009933333 -0.012094711 -0.0077719558 0.0000000
Prosjek-DFSE       -0.034033333 -0.036194711 -0.0318719558 0.0000000
SVK-DFSE           -0.011133333 -0.013294711 -0.0089719558 0.0000000
Prosjek-Najtocniji -0.024100000 -0.026261378 -0.0219386224 0.0000000
SVK-Najtocniji     -0.001200000 -0.003361378  0.0009613776 0.4727649
SVK-Prosjek         0.022900000  0.020738622  0.0250613776 0.0000000
```

Slika 4.5 ANOVA test za rezultate na hrvatskom skupu podataka

Drugi statistički test korišten za utvrđivanje razlika u rezultatima ostvarenih tehnika jest neparametarski Friedmanov test. Rezultati testa za hrvatski skup podataka su vidljivi na slici 4.6. Friedmanov test potvrđuje ranije provedeni ANOVA test; rezultati temeljeni na novoj tehnici DFSE su statistički značajno bolji u donosu na ostale rezultate.

```
> friedman.test(doktStatTestHr$vrijednosti, doktStatTestHr$grupa, doktSta
tTestHr$mjerenje)
                        Friedman rank sum test

data:  doktStatTestHr$vrijednosti, doktStatTestHr$grupa and doktStatTestH
r$mjerenje
Friedman chi-squared = 81.906, df = 3, p-value < 2.2e-16

> posthoc.friedman.nemenyi.test(doktStatTestHr$vrijednosti, doktStatTestH
r$grupa, doktStatTestHr$mjerenje)

        Pairwise comparisons using Nemenyi multiple comparison test
             with q approximation for unreplicated blocked data

data:  doktStatTestHr$vrijednosti , doktStatTestHr$grupa and doktStatTest
Hr$mjerenje

           DFSE    Najtocniji Prosjek
Najtocniji 0.00016 -          -
Prosjek    2.7e-14 9.4e-06    -
SVK        9.4e-06 0.93208    0.00016
```

Slika 4.6 Friedmanov test za rezultate istraživanja na hrvatskom skupu podataka

Postupak je ponovljen za istraživanja na njemačkom skupu podataka. Temeljem ANOVA i post-hoc Tukey testa, može se zaključiti da su rezultati ostvareni pomoću nove tehnike DFSE statistički značajno bolji u donosu na ostale rezultate uz dani *alpha*=0.05.

```
> aov.ex1 = aov(vrijednosti~grupe,data=doktStatTestGer)
> summary(aov.ex1)
                  One Way Analysis of Variance

             Df   Sum Sq  Mean Sq F value Pr(>F)
grupe         3 0.017506 0.005835   807.7 <2e-16 ***
Residuals   116 0.000838 0.000007
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> TukeyHSD(aov.ex1)
               Tukey multiple comparisons of means
                 95% family-wise confidence level

Fit: aov(formula = vrijednosti ~ grupe, data = doktStatTestGer)

$grupe
                            diff          lwr          upr p adj
Najtocniji-DFSE    -0.005100000 -0.006909047 -0.003290953     0
Prosjek-DFSE       -0.031800000 -0.033609047 -0.029990953     0
SVK-DFSE           -0.012766667 -0.014575714 -0.010957620     0
Prosjek-Najtocniji -0.026700000 -0.028509047 -0.024890953     0
SVK-Najtocniji     -0.007666667 -0.009475714 -0.005857620     0
SVK-Prosjek         0.019033333  0.017224286  0.020842380     0
```

Slika 4.7 ANOVA test za rezultate istraživanja na njemačkom skupu podataka

Isti zaključak se može izvesti iz neparametarskog testa, gdje su sve vrijednosti *p* koje se odnose na DFSE tehniku manje od alpha=0.05.

```
> friedman.test(doktStatTestGer$vrijednosti, doktStatTestGer$grupe, doktS
tatTestGer$mjerenja)
                        Friedman rank sum test

data:  doktStatTestGer$vrijednosti, doktStatTestGer$grupe and doktStatTes
tGer$mjerenja
Friedman chi-squared = 89.436, df = 3, p-value < 2.2e-16

> posthoc.friedman.nemenyi.test(doktStatTestGer$vrijednosti, doktStatTest
Ger$grupe, doktStatTestGer$mjerenja)

       Pairwise comparisons using Nemenyi multiple comparison test
            with q approximation for unreplicated blocked data

data:  doktStatTestGer$vrijednosti , doktStatTestGer$grupe and doktStatTe
stGer$mjerenja

           DFSE    Najtocniji Prosjek
Najtocniji 0.026   -          -
Prosjek    2.9e-14 6.4e-09    -
SVK        2.2e-08 0.010      0.014
```

Slika 4.8 Friedmanov test za rezultate istraživanja na njemačkom skupu podataka

4.8 Zaključci poglavlja

U ovom poglavlju provedeno je istraživanje u kojem je konstruirana nova tehnika za izradu sustava višestrukih klasifikatora. Tehnika se zasniva na korištenju više klasifikatora koji na ulazu primaju različite podatke za trening. Različitost podataka se očituje u različitim atributima koji su rezultat primjene više tehnika za njihov odabir. U kombiniranje modela je dodatno uključen korak smanjivanja sustava pomoću COP algoritma, kojim se izostavljaju klasifikatori koji negativno utječu na sustav. Istraživanje je provedeno na njemačkom i hrvatskom skupu kreditnih podataka s ciljem provjere temeljne hipoteze; da li sustav koji je temeljen na različitim tehnikama za odabir atributa može postići dovoljnu raznolikost da bi imalo smisla kombinirati klasifikatore.

Rezultati istraživanja potvrđuju da je kombiniranje modela treniranih na različitim atributima opravdano i da se rezultati u odnosu na pojedinačne klasifikatore mogu poboljšati uz uvjet da se zadrže samo oni koji pozitivno utječu na rezultat. U oba primjera samo kombiniranje klasifikatora nije donijelo značajno poboljšanje performansi zbog nedostatka faza u pretprocesiranju kao što su: prilagođavanja parametara, pronalaženja optimalnog broja atributa i analize kompatibilnosti algoritama. Proizvoljan odabir parametara, odnosno ne provođenje pretprocesiranja, prilikom konstruiranja tehnike povlači nedostatak garancije kvalitete pojedinačnih modela. Međutim upravo takav odabir elemenata je prednost tehnike, u tome se ogleda njezina robusnost i brzina jer smanjuje vrijeme potrebno za konstrukciju modela, a problem loših modela se rješava smanjivanjem sustava gdje se isti odbacuju.

Rezultati DFSE tehnike su pokazali značajno poboljšanje performansi klasifikacije i efikasnosti sustava. Veličina konačnog sustava ovisi o broju odabranih tehnika; najbolji rezultati se pojavljuju u prvim fazama izvođenja algoritma za smanjivanje. Jednom kada se iskoriste sve tehnike, raznolikost ovisi samo o veličini odabranih podskupova atributa, tada rezultati počinju opadati. Prednost nove tehnike je što se najveća raznolikost ostvaruje u ranijim koracima što znači efikasnije sustave.

Jedno od bitnih obilježja je jednostavnost i brzina konstruiranja. U slijedećem poglavlju će se usporediti performanse nove tehnike s najčešće korištenim tehnikama Bagging i Boosting. Osim kvalitete klasifikacije istražit će se i vrijeme izvođenja.

5

Eksperimentalna usporedba DFSE tehnike s tehnikama Bagging i Boosting

U svrhu vrednovanja performansi novo predložene tehnike DFSE, provedeno je istraživanje u kojem se rezultati tehnike uspoređuju s tehnikama Bagging i Boosting.

5.1 Uvod

Odobravanje kredita je složeni proces koji je originalno izvođen isključivo subjektivnim procjenama kreditnog referenta. Iako se i danas u nekim bankama izvođenje tog procesa zasniva na iskustvu djelatnika, korištenje kreditnih modela kao pomoć ili zamjena dijela procesa jest sve češća. Vrsta kreditnih modela i način na koji će sudjelovati u procesu ovisi o njihovoj pouzdanosti i korisnosti te odluci kreditne institucije.

Inicijalni način odobravanja, zvan procjena eksperta (engl. *expert-judgment*) ili subjektivna procjena (engl. *subjective-judgment*) se u potpunosti zasniva na znanju i pristupu djelatnika. Odluka se često donosi na temelju adaptiranih pravila ili principa koji su donijeli iskusni kreditni referenti (Ibtissem i Bouri, 2013), a cilj takvih pravila je smanjivanje prevelike subjektivnosti u odobravanju kredita. Na primjer, prema evaluacijskom kriteriju 5C procjena se bazira na: općim podacima, kapitalu, sredstvima osiguranja, zaduženosti klijenta i ekonomskim uvjetima (engl. *character, capital, collateral, capacity i economic conditions*) (Yap et al., 2011). S porastom kreditnih zahtjeva javila se potreba za bržom analizom podataka koja ne podrazumijeva samo rad djelatnika (Wang et al., 2011).

Analiza podataka o podnositelju zahtjeva i donošenje odluke jest najiscrpniji i najvažniji dio procesa, stoga je trud istraživača, u svrhu poboljšanja učinkovitosti i olakšavanja rada referenta, usmjeren na automatizaciju tih podprocesa. U širem kontekstu, automatsko odobravanje kredita se može klasificirati u dvije grupe ovisno o korištenim tehnikama na: statističke tehnike i tehnike umjetne inteligencije (Wang i Ma, 2012).

U ranijim istraživanjima statističke tehnike su češće korištene za izradu kreditnih modela, u tu skupinu spadaju: linearna diskriminantna analiza (Baesens et al., 2003), logistička regresija (Tomas, 2000; West, 2000) i multivarijantna adaptivna regresija – MARS (engl. *Multivariate Adaptive Regression Splines*) (Friedman, 1991). Međutim, problem primjene statističkih tehnika na problem procjene kreditnog rizika leži u pretpostavkama, kao što je multivarijatna pretpostavka normalnosti za nezavisne varijable, koje često ne vrijede u stvarnim skupovima podataka, što takve tehnike čine teoretski nevaljanima (Huang et al., 2004). Da bi se otklonio opisani nedostatak novija istraživanja koriste tehnike koje ne podrazumijevaju ranije navedene pretpostavke: umjetne neuronske mreže (Yu et al., 2008; Oreski i Oreski, 2014), stabla odlučivanja (Yap et al., 2011), tehniku potpornih vektora (Bellotti i Crook, 2009; Harris, 2013) i druge. Za razliku od statističkih, tehnike umjetne inteligencije ne podrazumijevaju određene distribucije podataka već automatski izvlače

znanje iz primjera obuhvaćenih skupovima za trening (Wang et al., 2011). Takve tehnike su u dosadašnjim istraživanjima ostvarivale superiorne rezultate u odnosu na statističke tehnike, pogotovo na nelinearnim klasifikacijskim uzorcima (Huang et al., 2004). Isti autori navode da je neuronska mreža s povratnom propagacijom načešće korištena tehnika umjetne inteligencije.

Međutim, ne postoji jedna najbolja tehnika koja bi odgovarala svim skupovima podataka već prikladnost ovisi o karakteristikama podataka, korištenim postavkama i cilju klasifikacije (Yu et al., 2008). Noviji smjer razvoja klasifikacijskih modela na području kreditnog rizika podrazumijeva kombiniranje pojedinačnih klasifikatora s ciljem poboljšanja kvalitete predikcije. Postoji mnogo različitih načina konstruiranja sustava višestrukih klasifikatora koji su velikim dijelom već objašnjeni u ovoj disertaciji. Od svih spomenutih pristupa dvije metode se smatraju najpopularnijim za konstruiranje sustava višestrukih klasifikatora: Bagging i Boosting (Paleologo et al., 2010; Wang et al., 2011). Razlog njihove popularnosti leži u: jednostavnosti primjene, širokom prostoru primjene i kvaliteti ostvarenih rezultata.

Na temelju iznesenih razmatranja, u ovom poglavlju će biti provedena eksperimentalna usporedba dvije najpopularnije tehnike Bagging i Boosting s novo predloženom tehnikom DFSE. Istražit će se performanse odabranih tehnika po pitanju: točnosti klasifikacije i vremena potrebnog za trening modela ali i greške tipa I i II te AUC mjere. Poseban naglas je stavljen na vrijeme potrebno za trening jer odabrane tehnike Bagging i Boosting su vremenski zahtjevne, što bi alternativnu tehniku s podjednako dobrim rezultatima ali kraćim vremenom treninga učinilo atraktivnom dopunom postojećim tehnikama. Na kraju poglavlja rezultati tehnike DFSE će biti uspoređeni s rezultatima ostalih istraživanja iz literature na istim skupovima podataka.

Ostatak poglavlja je organiziran na slijedeći način. Odjeljak 5.2 opisuje problem i daje kratak pregled literature vezan uz korištenje Bagging i Boosting tehnika na kreditnim skupovima podataka. Opis metodologije i koncepata korištenih u istraživanju je dan u odjeljku 5.3. U odjeljku 5.4 je predstavljen eksperimentalni dizajn provedenog istraživanja. Ostvareni rezultati i njihova analiza su dani u odjeljku 5.5, a usporedba s dostupnim istraživanjima iz literature u odjeljku 5.6. U posljednjem odjeljku su izneseni zaključci i smjer budućih istraživanja.

5.2 Opis problema i pregled literature

U poglavlju 4 je uz konstruiranje tehnike DFSE provedeno istraživanje opravdanosti kombiniranja klasifikatora koristeći predloženu tehniku. Rezultati pokazuju da je u odnosu na pojedinačne klasifikatore uključene u sustav ostvareno poboljšanje performansi koje se može statistički dokazati. Iz zaključka istraživanja prezentiranog u poglavlju 4 slijedi da trening klasifikacijskih modela na različitim atributima uz primjenu algoritma za smanjenje sustava generira dovoljnu raznolikost koja čini kombiniranje klasifikatora u sustav opravdanim. Iako je dokazano poboljšanje performansi u odnosu na pojedinačne klasifikatore, prethodno istraživanje nije dalo odgovor koliko su ostvareni rezultati kvalitetni. Istraživanje opisano u ovom poglavlju predstavlja logički nastavak kojim će se usporediti DFSE tehnika s postojećim tehnikama koje se najčešće koriste kao mjerilo u sličnim istraživanjima s ciljem vrednovanja kvalitete rezultata. U skladu s ciljem dana je hipoteza istraživanja:

H2: *Sustav višestrukih klasifikatora koji je temeljen na odabiru različitih podskupova atributa pomoću filtarskih tehnika te konstruiran na temelju u ovom radu predloženog algoritma za smanjivanje sustava će postizati statistički jednake ili bolje rezultate u odnosu na najpopularnije tehnike, Bagging i Boosting primijenjene na originalnim skupovima podataka (njemačkom i hrvatskom) sa svim karakteristikama.*

Mjera kojom će se vrednovati rezultati tehnika je točnost klasifikacije, koja je standardna i najčešće korištena mjera za evaluaciju klasifikacijskih modela u literaturi. Iako je kod vrednovanja hipoteze naglasak stavljen na točnost klasifikacije, osim točnosti u ovom istraživanju radi stvaranja šire slike o kvaliteti nove tehnike koristit će se i druge mjere.

Jedna od njih je vrijeme potrebno za generiranja modela. Kompleksnost DFSE tehnike se ne može jednoznačno odrediti jer se konstruira dinamički i zavisi o korištenim tehnikama, ali tehnika nije kompleksnija od pojedinačnih korištenih elemenata. U okviru ovog istraživanja mjerit će se vrijeme potrebno za trening modela i usporediti s vremenima tehnika Bagging i Boosting. Postojeća istraživanja (Fersini et al., 2014, Xia et al., 2014) kao glavi nedostatak odabranih tehnika za usporedbu navode vremensku složenost prilikom treninga ukoliko se ne koriste slabi klasifikatori (engl. *weak classifiers*). Ukoliko DFSE tehnika postigne jednake rezultate uz kraće vrijeme generiranja modela, tada predstavlja dobru alternativu tehnikama za konstruiranje SVK-a.

Dodatno, uz usporedbu performansi i vremena postignutih različitim tehnikama u okviru ovog istraživanja, rezultati će se usporediti s rezultatima iz dostupnih istraživanja iz literature koja su provedena na istom skupu podataka s istim postavkama validacije. Odabrana istraživanja su prikazana u pregledu literature u nastavku poglavlja.

Autori West el al. (2005) su istraživali nekoliko strategija SVK-a između kojih i Bagging i Boosting tehnike na kreditnim skupovima podataka gdje su neuronske mreže odabrane kao osnovni klasifikacijski algoritam. Istraživanje je povedeno na tri različita skupa podataka: australskom, njemačkom i vlastitom skupu podataka o bankrotu. Iz rezultata je vidljivo da sve tehnike SVK-a koje su korištene postižu bolje rezultate od najbolje pojedinačne neuronske mreže. Autori ističu da smanjenje greške, koje se kreće između 3-5%, iako se čini skromno potencijalno može uštedjeti industriji 1.3 milijarde dolara. Analizom rezultata vidljivo je da za postizanje najboljih rezultata Bagging tehnika koristi manji broj pojedinačnih klasifikatora od algoritma Boosting. U istraživanju se zaključuje da je zbog boljih generalizacijskih sposobnosti SVK superiorniji u odnosu na pojedinačne klasifikatore.

Postizanje raznolikosti je ključan faktor prilikom konstruiranja SVK-a, a tehnike Bagging i Boosting raznolikost postižu različitim tehnikama uzorkovanja testnog skupa podataka. Istraživanje (Marqués et al., 2012) ispituje kombinaciju dva pristupa: (1) uzorkovanja primjera (Bagging i Boosting) i (2) odabira podskupa atributa (slučajni odabir, rotacija šume (engl. *rotation forest*)). Algoritmi uključeni u istraživanje se kombiniraju u paru serijski tako da se iskoriste sve moguće njihove kombinacije. Istraživanje je provedeno na više skupova kreditnih podataka koji uključuju i njemački skup. Eksperimentalni rezultati i statistički testovi pokazuju da konstruirani sustavi u dvije faze postižu bolje rezultate od pojedinačnih klasifikatora i sustava koji koriste samo jednu fazu. Ukoliko se usporede rezultati promatranih tehnika Bagging i Boosting kada se samostalno primjene, tada Bagging ostvaruje malo bolje rezultate. Istraživanje (Marqués et al., 2012) je primjer usporedbe novog pristupa konstruiranja SVK-a s već dokazanim tehnikama Bagging i Boosting, a zaključak je da integracija slučajnog odabira atributa i tehnike Boosting dovodi do poboljšanja klasifikacije kreditnih podataka. Slični zaključak donosi i rad (Wang i Ma, 2011) proveden na drugim skupovima podatka.

Istraživanje (Ghodselahi, 2011) ispituje kvalitetu novo predloženog hibridnog pristupa s tehnikama Bagging i Boosting koje koriste različite algoritme za učenje. U istraživanju su korišteni algoritmi: neuronske mreže, tehnika potpornih vektora i stabla odlučivanja, a eksperiment je proveden na njemačkom skupu podataka. Nova hibridna tehnika se zasniva na tehnici potpornih vektora, točnije na korištenju različitih jezgri na temelju kojih se stvara

razlika u treniranim modelima. Rezultati pojedinačnih klasifikacijskih algoritama su lošiji u usporedbi s Bagging i Boosting tehnikama, a nova predložena tehnika je ostvarila najbolje performanse. Ovisno o korištenom algoritmu može se promatrati odnos tehnika Bagging i Boosting, u slučaju neuronskih mreža i stabla odlučivanja Bagging tehnika je ostvarila bolje rezultate dok je u slučaju tehnike potpornih vektora obrnuto.

Autori istraživanja (Marqués et al., 2012a) vrednuju performanse sedam različitih algoritama za učenje kao članova različitih tehnika SVK-a. Tehnike Bagging i Boosting su uključene u istraživanje, a rezultati potvrđuju ranije pronalaske gdje je tehnika Bagging ostvarila bolje rezultate. Tehnika Bagging ostvaruje najbolje rezultate po pitanju točnosti klasifikacije kada koristi neuronske mreže, a tehnika Boosting logističku regresiju na njemačkom skupu podataka. Osim njemačkog skupa podataka u istraživanje je uključeno pet dodatnih skupova. U širem kontekstu kada se promatraju svi rezultati autori predlažu tehniku stabla odlučivanja kao najbolji izbor prilikom konstruiranja sustava. Istraživanje je pokazalo da algoritam koji daje najbolji pojedinačni rezultat ne mora biti dati najbolje rezultate kada je odabran za konstrukciju sustava.

Dvije nove tehnike temeljene na SVK-a koje koriste neuronske mreže za klasifikaciju kreditnih podataka su predložene u radu (Alaraj, 2014). Tehnike predstavljaju kombinaciju unakrsne validacije (CV) i Bagging algoritma, a testirane su na njemačkom i australskom skupu podataka. Bagging algoritam je izabran na temelju ranijih istraživanja (Wang et al., 2011) koja su pokazala da Bagging postiže bolje rezultate od tehnike Boosting. Predložene tehnike su uspoređene s individualnim neuronskim mrežama i tehnikom Bagging zasebno. Rezultati pokazuju da model CV-Bagging daje najbolje rezultate u odnosu na ostale tehnike.

Iscrpna usporedba 41 različita klasifikatora prema 6 mjera performansi na 8 različitih skupova kreditnih podataka je provedena u istraživanju (Lessmann et al., 2015). Klasifikatori su podijeljeni u tri različite grupe: individualni klasifikatori, homogeni sustavi i heterogeni sustavi višestrukih klasifikatora. Istraživanje pokazuje da neki klasifikatori ostvaruju značajno bolje performanse u odnosu na druge. Autori preporučuju korištenje neuronskih mreža i logističke regresije kao standarda prilikom usporedbe s novim tehnikama, jer pružaju najbolje rezultate u kategoriji individualnih klasifikatora i dostupne su u većini alata. Međutim, najbolji rezultat na testu je postigla integracija brze tehnike za smanjivanje sustava klasifikatora i tehnike Bagging. Ukoliko se usporede prosjeci po grupama najbolje rezultate ostvaruju heterogeni sustavi višestrukih klasifikatora.

Pregled literature pokazuje veliko zanimanje znanstvenika za problem kreditnog rizika, s naglaskom na povećanje broja istraživanja u posljednjim godinama (Marqués et al.,

2012a; Alaraj, 2014; Oreski i Oreski, 2014; Lessmann et al., 2015). Velika većina radova objavljena na tom području koristi njemački skup podataka koji je svojevrsno mjerilo (engl. *benchmark*) za usporedbu postignutih rezultata. Osim njemačkog skupa podataka istraživači koriste i australski ali i vlastite skupove podataka prikupljene u kreditnim institucijama.

Novija istraživanja su usmjerena prema kombiniranju klasifikatora s ciljem poboljšanja performansi. Zajednički je stav istraživača da su tehnike Bagging i Boosting mjerilo kvalitete novih sustava, stoga se koriste za usporedbu ostvarenih performansi. Različiti su pristupi stvaranju novih tehnika između kojih vrijedi istaknuti: kombiniranje Bagging i Boosting tehnika s drugim elementima kao što je korištenje algoritma za smanjivanje sustava ili redukciju dimenzionalnosti. U većini provedenih istraživanja (Wang et al., 2011; Marqués et al., 2012a; Marqués et al., 2012) tehnika Bagging je ostvarila nešto bolje rezultate od tehnike Boosting.

Mjere koje se koriste za evaluaciju performansi klasifikatora variraju ovisno o istraživanju. Mjere uključuju: točnost klasifikacije, grešku tipa I i grešku tipa II. Osim navedenih neki radovi koriste i AUC mjeru površine ispod krivulje, koja je posebno prikladna u uvjetima nebalansiranih uzoraka podataka.

5.3 Metodologija

U ovom istraživanju su korištene Bagging i Boosting, najpopularnije tehnike za konstruiranje sustava višestrukih klasifikatora. Razlog njihove velike prisutnosti u istraživanjima leži u njihovoj jednostavnosti i dostupnosti u većini alata za konstruiranje modela strojnog učenja. Obje tehnike se zasnivaju na uzorkovanju primjera iz skupa podataka, a njihov detaljan opis slijedi u nastavku ovog poglavlja.

5.3.1. Bagging

Bagging (skraćeno za bootstrap aggregating) je jedan on najranijih algoritama za konstrukciju sustava višestrukih klasifikatora. Odlikuju ga karakteristike najintuitivnijeg i najjednostavnijeg algoritma za implementaciju s iznenađujuće dobrim performansama. Raznolikost u Bagging algoritmu se postiže korištenjem uzorkovanja s ponavljanjem (engl. *bootstrap*) na inicijalnom skupu podataka, čime se postižu replike trening podataka (Wang et al., 2011). Uzorkovanje s ponavljanjem jest slučajan odabir primjera iz skupa, gdje se

odabrani primjer ne isključuje iz inicijalnog skupa već može biti odabran i u slijedećim iteracijama. Veličina uzorkovanog podskupa je najčešće veličine inicijalnog skupa u kojem se neki primjeri mogu pojaviti više puta dok drugi ne moraju biti odabrani niti jedanput. Vjerojatnost pojavljivanja barem jednom jest 63.2%, što znači da otprilike 36.8% primjera pojedinačni klasifikator ne koristi prilikom izrade modela. Dobiveni podskupovi trening podatka se koriste za treniranje različitih klasifikacijskih modela pomoću algoritma istog tipa.

Kombinacija odluka svih klasifikatora generiranih prilikom treninga se postiže glasanjem. Glasanje je jednostavna funkcija kombinacije glasova kojom se odabire ona klasa koja se pojavila na izlazu većine klasifikatora. Broj klasifikatora koji se kombiniraju glasanjem mora biti neparan, radi izbjegavanja neriješenih ishoda glasanja. Pseudo kod za Bagging tehniku je prikazan na slici 5.1.

Da bi se dodatno povećala raznolikost odluka generiranih modela, za trening modela se koristite slabi (engl. *weak learners*) i nestabilni klasifikatori (engl. *unstable learners*). Klasifikator je slab ukoliko su ostvarene performanse na klasifikaciji testnih primjera malo bolje od performansi slučajnog klasifikatora. Najčešći primjer slabog algoritma su jednostavna stabla odluke dubine 1 (engl. *decision stumps*). Međutim vrijedno je napomenuti, da iako se većina teorijskih analiza radi na slabim, klasifikatori koji se koriste u istraživanjima ne moraju to nužno biti, štoviše kombiniranje jakih klasifikatora često rezultira boljim performansama (Zhou, 2012).

Klasifikator je nestabilan kada male promjene ulaznih podataka generiraju velike promjene unutar treniranih modela, a posljedično i promjene u odlukama. Nestabilnost klasifikatora je poželjna karakteristika kod svih tehnika kombiniranja, pa tako i Bagginga. U nestabilne algoritme se najčešće svrstavaju: neuronske mreže i tehnika potpornih vektora.

```
Ulaz:  Skup podataka D={(x1,y1),(x2,y2),..,(xm,ym)};
       Osnovni algoritam za učenje L;
       Broj iteracija učenja T;

Proces:
       Za svaki t=1,2,…,T
              Dt = Bootstrap(D);   %gen. podskup primjera uzorkovanjem s
ponavljanjem
              ht = L(Dt)           %koristi alg.L za trening podataka odabranog
podskupa
       kraj.

Izlaz: H(x) = argmax_{y∈Y} Σ_{t=1}^{T} 1(y = ht(x))      %vrijednost 1(α) je 1 ako je α istina i
0 obrnuto
```

Slika 5.1 Pseudo kod Bagging tehnike

5.3.2. Boosting

Boosting algoritmi obuhvaćaju skup tehnika za konstruiranje sustava višestrukih klasifikatora, koje karakterizira korištenje više slabih klasifikatora s ciljem stvaranja jakog klasifikatora. U odnosu na Bagging tehniku dodane su dvije modifikacije:

1. umjesto slučajnog odabira primjera iz skupa podataka, koristi se odabir s težinskim faktorima tako da se učenje usmjerava prema težim primjerima i
2. umjesto kombiniranja jednostavnim glasanjem gdje svaki član ima jednak glas, koristi se glasanje s težinskim faktorima (engl. *weighted vote*).

Ideja Boosting tehnike je serijska primjena istog klasifikatora na modificiranim verzijama testnih podataka, s ciljem produciranja različitih modela koji se u konačnici kombiniraju. Modifikacija podataka se ostvaruje uzorkovanjem trening skupa prema promjenjivoj distribuciji koja se ponovno definira u svakom ciklusu izvođenja tehnike. Distribucija se sastoji od težine primjera i mijenja se zavisno o ispravnosti klasificiranja pojedinih primjera.

U prvom koraku svi primjeri iz originalnog skupa podataka su inicijalizirani s istim težinama što čini odabir primjera slučajnim. Nakon inicijalizacije, u svakoj Boosting iteraciji se trenira model podataka odabranim klasifikacijskim algoritmom na odabranom trening skupu podataka. U praksi klasifikator mora koristiti algoritam koji zna koristiti težine primjera iz trening skupa. Alternativno, kada to nije moguće, podskup novih primjera se može uzorkovati poštujući težine i tako odabrani (bez težina) primjeri se mogu koristiti za trening klasifikatora (Freund, 1999). Nakon treninga, računa se greška klasifikacije te se smanjuje težina ispravno klasificiranim primjerima a povećava onima koji su pogrešno klasificirani. Opisani postupak se ponavlja preddefinirani broj puta.

Konačna odluka Boosting modela je linearna kombinacija modela iz svih koraka s uključenom težinom koja predstavlja vlastitu mjeru važnosti. Iako postoji nekoliko verzija Boosting algoritma najčešće se koristi AdaBoost (engl. *Adaptive Boosting*) algoritam koji su predložili Freund i Schapire (1996). Stoga će u ovom radu biti korišten AdaBoost algoritam. Pseudo kod AdaBoost algoritma je dan na slici 5.2.

Ulaz: Skup podataka $D=\{(x_1,y_1),(x_2,y_2),..,(x_m,y_m)\}$;
Osnovni algoritam za učenje L;
Broj iteracija učenja T;

Proces:
$D_1(i)=1/m$ %inicijalizacija distribucije težina
Za svaki $t=1,2,\ldots,T$:
$h_t = L(D,D_t)$; %trening klasifikatora h_t na D koristeći distribuciju Dt
$\varepsilon_t = \Pr_{i\sim Di}[h_t(x_i)\neq y_i]$; %mjerenje pogreške klasifikatora h_t
$\alpha_t = \frac{1}{2} ln \frac{1-\varepsilon_t}{\varepsilon_t}$; %određivanje težine od h_t
$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} \exp(-\alpha_t) \; if \; h_t(x_i) = y_i \\ \exp(\alpha_t) \; if \; h_t(x_i) \neq y_i \end{cases}$ %obnavljanje distribucije, gdje je Z,
$= \frac{D_t(i)\exp\,(-\alpha_t y_i h_t(x_i))}{Z_t}$ %norm. faktor za distribuciju D_{t+1}
Kraj.

Izlaz: $H(x) = sign(f(x)) = sign \sum_{t=1}^{T} \alpha_t \, h_t(x)$

Slika 5.2 Pseudo kod AdaBoost tehnike (Freund, 1999)

5.4 Dizajn eksperimenta

5.4.1. Skupovi podataka

Istraživanje je provedeno na stvarnim skupovima podataka korištenim i u prethodnim istraživanjima ove doktorske disertacije. Odabrani skupovi uključuju hrvatski i njemački skup kreditnih podataka. Oba skupa podataka sadrže 1000 primjera stvarno odobrenih kredita. Njemački skup podataka je javno dostupan i preuzet iz repozitorija Sveučilišta iz Aucklanda, a hrvatski skup podataka je prikupljen za potrebe ranijih istraživanja u jednoj hrvatskoj kreditnoj instituciji. Prikupljeni podaci se mogu svrstati u pet glavnih skupina : (i) osnovne karakteristike , (ii) povijest plaćanja (mjesečni prosjeci) , (iii) financijski uvjeti , (iv) delikvencijska povijest, i (v) prošla kreditna iskustava (Oreski et al., 2012).

Prikupljanje i obrada podataka je detaljno opisana u poglavlju 4.5, a popis varijabli hrvatskog skupa s njihovim obrazloženjem i deskriptivnom statistikom za podatke iz uzorka dan je u dodatku A.

5.4.2. Evaluacijski kriterij

Evaluacijski kriteriji za vrednovanje kvalitete rezultata su preuzeti iz recentnih istraživanja kao utvrđene standardne mjere korištene prilikom testiranja novih tehnika na kreditnim podacima. Korištene mjere uključuju: točnost klasifikacije, grešku tipa I, grešku tipa II i AUC (površinu ispod krivulje ROC)(engl. *Area Under the ROC Curve*). Svaka mjera ima svoje

prednosti i ograničenja, a njihovim kombiniranjem postiže se šira slika kvalitete uspoređenih tehnika, stoga su u istraživanju umjesto jedne korištene četiri različite mjere.

Točnost klasifikacije je najčešća mjera koja se u mnogim istraživanjima koristi samostalno a mjeri odnos ispravno klasificiranih primjera i svih primjera testnog skupa podataka. Mjera je opisana u poglavlju 4.5.2.

Greška tipa I mjeri broj primjera koji pripadaju u klasu loših kredita, a neispravno su klasificirani u klasu dobrih kredita tj. slučaj odobravanja kredita lošem klijentu. Mjera se formalno može definirati pomoću tablice 4.4. na slijedeći način (Tsai i Wu, 2008):

$$Greška\ tipa\ I = \frac{f_{01}}{f_{11} + f_{01}} \tag{5.1}$$

Greška tipa II mjeri broj primjera koji pripadaju u klasu dobrih kredita, a neispravno su klasificirani u klasu loših tj. slučaj neodobravanja kredita dobrom klijentu. Definicija mjera pomoću tablice 4.4. glasi (Tsai i Wu, 2008):

$$Greška\ tipa\ II = \frac{f_{10}}{f_{00} + f_{10}} \tag{5.2}$$

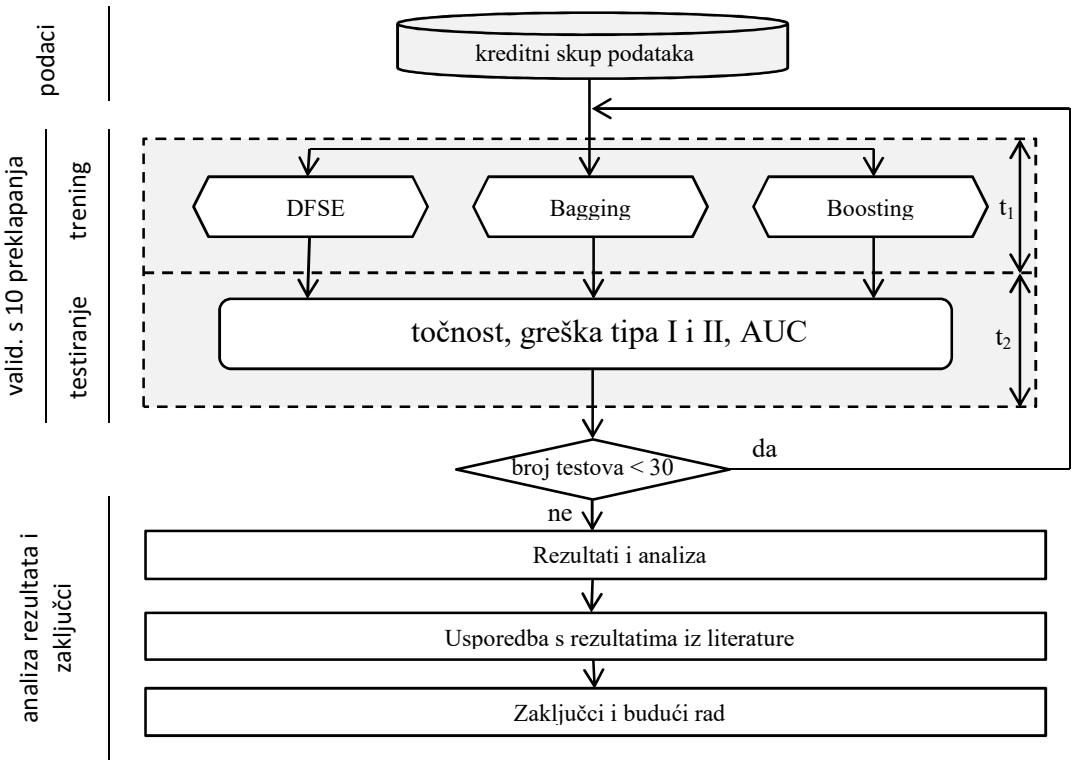
ROC krivulja (engl. *Receiver Operating Characteristic curve*) je standardna tehnika za prikaz performansi klasifikatora koja daje uravnoteženu sliku odnosa točno klasificiranih primjera pozitivne klase u odnosu na ukupan broj pozitivnih primjera (engl. *true positive rate,* TPR) i netočno klasificiranih primjera negativne klase u odnosu na ukupan broj negativnih primjera (engl. *false positive rate,* FPR). Površina ispod ROC krivulje je tradicionalno prihvaćena kao mjera performansi za neuravnotežene skupove (Mazurowski et al, 2008).

5.4.3. Eksperimentalna procedura

Uspoređene su tri različite tehnike za konstruiranje sustava višestrukih klasifikatora. Novo predložena tehnika DFSE je uspoređena s tehnikama Bagging i Boosting na kreditnim skupovima podataka. U suštini, razlika u odabranim tehnikama je u načinu postizanja raznolikosti, gdje se uspoređuje uzorkovanje podataka s odabirom atributa. Dodatno uz odabir atributa, tehnika DFSE koristi i algoritam za smanjivanje sustava kao korekcijski korak kojim se odbacuju klasifikatori koji ne pridonose poboljšanju performansi. Da bi se postigla usporedivost, osnovni klasifikacijski algoritam ne smije donositi razliku u produciranim modelima tj. algoritmi moraju biti isti. U skladu s tim, odabrani algoritmi koji su korišteni u konstruiranju DFSE tehnike u poglavlju 4, su izabrani za ovo istraživanje. Jedina razlika jest u odabiru klasifikacijskog algoritma za njemački skup podataka, gdje su umjesto stabla

odlučivanja s logističkom regresijom, odabrana stabla odlučivanja s Gini indeksom. Mjerenja su provedena za sve neprane veličine sustava od 3 od 25 članova, ukupno 12 različitih veličina.

U istraživanju se koristi validacija s 10 preklapanja. Tehnika validacije s *k*-preklapanja je bolja od jednostavne tehnike validacije, jer jednostavna tehnika validacije dijeli skup podataka na skup s uzorcima za učenje i na skup s uzorcima za testiranje (engl. *holdout sample*) s kojima testira učinkovitost modela. S obzirom da se za najbolji model odabire onaj model koji najbolje klasificira jedan podskup, holdout, jednostavna tehnika validacije često procjenjuje pravu stopu pogreške preoptimistično (Malhotra i Malhotra, 2003). U postupku validacije s *k*-preklapanja, kreditni skup se dijeli na *k* nezavisnih skupova. Model se trenira korištenjem prvih *k*-1 skupova uzoraka, a trenirani model se testira na *k*-tom skupu. Ovaj postupak se ponavlja sve dok svaki od skupova ne bude korišten jednom kao skup za testiranje. Ukupna točnost skoring modela je prosječna točnost ostvarena kroz svih *k* skupova. Značajka validacije s *k*-preklapanja je da je model kreditnog skoringa razvijen na temelju velikog dijela svih raspoloživih podataka te da su svi podaci korišteni za testiranje konačnog modela (Oreski, 2014).



Slika 5.3 Dijagram eksperimentalnog procesa

Slučajnim odabirom je određeno 30 različitih seed-ova validacije, čime je postignuto 30 različitih uzoraka za validaciju podataka. U svakoj validaciji u fazi testiranja modela

prikupljaju se četiri mjere opisane u poglavlju „*Evaluacijski kriteriji*“. Postupak se ponavlja dok se ne prikupe rezultati svih 30 iteracija.

Povrh odabranih mjera za vrednovanje performansi tehnika, u istraživanju se mjeri vrijeme potrebno za trening i testiranje modela na zadanim skupovima podataka. Na slici 5.3 prikazana su vremena t_1 za trening i t_2 vrijeme testiranja, koja su uspoređena u analizi istraživanja.

Nakon analize rezultata provest će se i komparacija s rezultatima ostvarenim u literaturi, gdje će biti odabrana ona istraživanja koja koriste istu validacijsku tehniku.

5.5 Rezultati istraživanja i analiza

Unutar poglavlja rezultati su podijeljeni u tri cjeline, zasebno su predstavljeni za hrvatski skup podataka, potom za njemački skup te na kraju su iznesena vremena potrebna za trening i test klasifikatora. Odabrane mjere su prikazane u zasebnim tablicama na način da svaki skup podataka ima četiri tablice s rezultatima. Analiza podataka i grafički prikaz izmjerenih vrijednosti su dani uz rezultate.

Tablice u kojima su prikazani rezultati istraživanja sadrže i rezultate provedenih statističkih testova. Testovi su provedeni zasebno za svako mjerenje (ponovljeno 30 puta) prema veličini sustava koji su korišteni. Statistički testovi uključuju parametarski test ANOVA i neparametrijski Friedmanov test.

Preduvjeti za korištenje ANOVA testa su provjereni Shapiro-Wilk testom normalnosti i Levene-ovim testom za homogenost standardnih devijacija. Zavisne varijable su prosječne mjere korištene u istraživanju i spadaju u kvantitativne kvocijentne varijable. Analiza rezultata dobivenih ANOVA testom je napravljena Tukey post-hoc testom.

Naparametrijski Friedmanov test je korišten s post-hoc Friedman-Nemenyi testom. U svim provedenim statističkim testovima zadana je statistička značajnost na razini *alpha*=0,05.

Prikazani rezultati u tablicama se odnose na rezultate post-hoc testova i tumače se na slijedeći način:

- svi rezultati su prikazani s tri simbola koji predstavljaju odnose rezultata tehnika prema slijedećem redoslijedu: (1) DFSE-Bagging, (2) DFSE-Boosting i (3) Bagging-Boosting

- simbol „*x*“ označava da je *p* vrijednost post-hoc testa manja od 0,05 što znači da postoji signifikantna razlika u rezultatima tehnika, u suprotnom slučaju korišten je simbol „-“
- simbol „*?*“ označava da jedan ili više uvjeta za provođenje ANOVA testa nisu zadovoljeni .

5.5.1. Hrvatski skup podataka

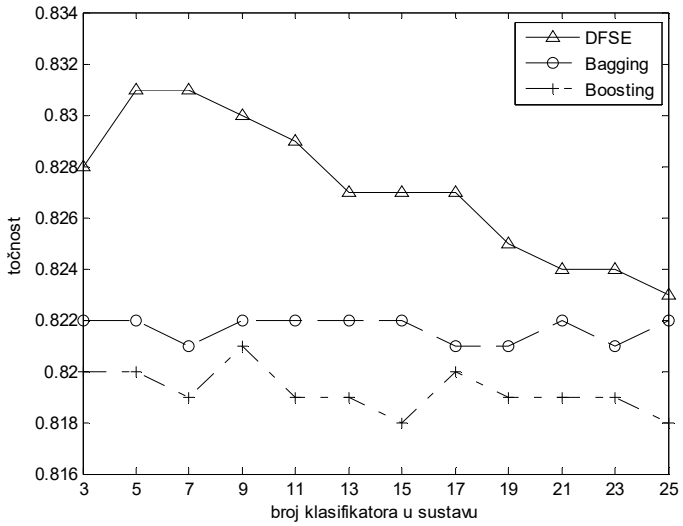
Ostvarena točnost klasifikacije odabranih tehnika na hrvatskom skupu podataka je dana u tablici 5.1. Najveću točnost ostvarila je DFSE tehnika u svim promatranim veličinama sustava. Na slici 5.4 na kojoj su prikazani rezultati, vidljivo je da je razlika najznačajnija na sustavima veličine 5 i 7 članova, dok se povećanjem broja članova razlika između DFSE tehnike i ostalih tehnika smanjuje. Sukladno zaključcima istraživanja opisanim u poglavlju 4 ove disertacije, performanse DFSE tehnike su povezane s brojem tehnika za odabir atributa. Prema ranijem istraživanju raznolikost sustava je najveća kada odabrani klasifikatori koriste različite tehnike i niti jedna tehnika se ne koristi više puta. Uključivanjem istih tehnika više puta smanjuje se raznolikost što utječe na točnost klasifikacije SVK-a.

Tablica 5.1 Ostvarena točnost i standardna devijacija tehnika na hrvatskom skupu podataka

Veličina sustava	DFSE		Bagging		Boosting		Stat. testovi	
	Točnost	Std	Točnost	Std	Točnost	Std	Anova	Fried.
3	0,828	0,004	0,822	0,005	0,820	0,005	xx-	xx-
5	0,831	0,005	0,822	0,005	0,820	0,005	xx-	xx-
7	0,831	0,004	0,821	0,005	0,819	0,005	xx-	xx-
9	0,830	0,004	0,822	0,004	0,821	0,005	xx-	xx-
11	0,829	0,005	0,822	0,005	0,819	0,005	xx-	xx-
13	0,827	0,004	0,822	0,004	0,819	0,005	xx-	xx-
15	0,827	0,005	0,822	0,005	0,818	0,005	xxx	-xx
17	0,827	0,004	0,821	0,004	0,820	0,004	xx-	xx-
19	0,825	0,004	0,821	0,005	0,819	0,005	xx-	xx-
21	0,824	0,004	0,822	0,005	0,819	0,006	-x-	-x-
23	0,824	0,004	0,821	0,004	0,819	0,005	-x-	-xx
25	0,823	0,004	0,822	0,004	0,818	0,005	-x-	-xx
Prosjek	0,827	-	0,822	-	0,819	-	-	-

U kontekstu zaključaka iz poglavlja 4 mogu se tumačiti ostvareni rezultati DFSE tehnike, gdje se infleksija događa na sustavima veličine 5 i 7 članova. Broj različitih korištenih tehnika za odabir atributa je 5 što se poklapa s veličinom sustava s najvećom preciznosti, a potvrđuje rezultate ranijih istraživanja. Redundantnost dobivena ponavljanjem tehnika nije utjecala na rezultat u prvom koraku s 7 članova, dok je u svakom slijedećem

vidljiv negativan utjecaj. Iz statističkih testova se može zaključiti da je na sustavima koji imaju do 21-og člana tehnika DFSE ostvarila bolje rezultate od obje tehnike korištene u usporedbi. Na sustavima s brojem članova koji prelaze 21, razlika između ostvarenih rezultata DFSE i Bagging tehnike se, uz zadani *alpha,* ne može statistički potvrditi.



Slika 5.4 Odnos točnosti i broja klasifikatora prema izvedenim mjerenjima na hrvatskom skupu podataka

Povećanje broja klasifikatora u sustavu nije povezano s velikim promjenama u performansama tehnika Bagging i Boosting. Točnost klasifikacije tih tehnika je gotovo konstantna vrijednost, koja indicira da poduzorkovanje skupa podataka na kreditnim podacima ne utječe na raznolikost koja bi donijela poboljšanje točnosti klasifikacije. Rezultati Bagging tehnike, iako za vrlo mali postotak, su bolji u donosu na Boosting tehniku, što je u skladu s dosadašnjim rezultatima iz literature.

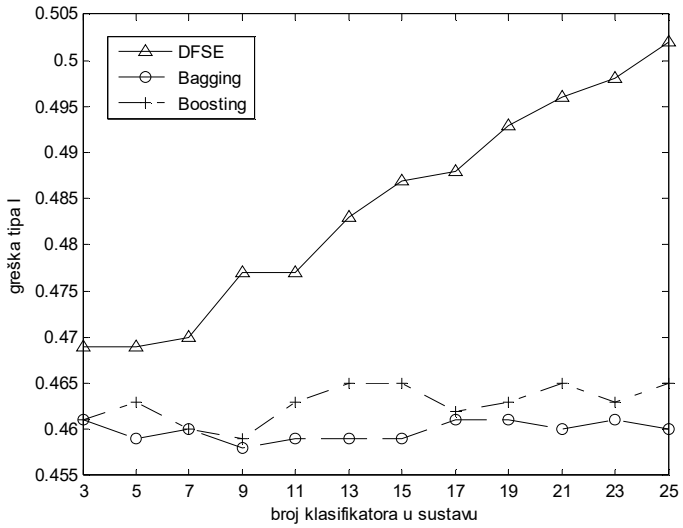
Greška tipa I je negativan ishod klasifikacije kada se loš klijent pogrešno klasificira kao dobar. Cilj klasifikacije je minimiziranje greške, što u ovom slučaju, za razliku od mjere točnosti, znači da klasifikator koji ostvari manju vrijednost ima bolje performanse. Vrijednosti prosječne greške tipa I za odabrane tehnike su prikazane u tablici 5.2.

Osim što tehnika DFSE ima najveću prosječnu grešku, s porastom broja klasifikatora u sustavu raste i vrijednost greške. Promatraju li se rezultati prema veličini sustava u rasponu u kojem je DFSE tehnika ostvarila najbolju točnost, vidljivo je da su u tim slučajevima ostvarene razlike greške tipa I najmanje. Štoviše, statistički testovi za sustave veličine 3, 5 i 7 pokazuju da između ostvarenih rezultata tehnike DFSE i Boosting ne postoji razlika, dok je ona između DFSE i Bagging tehnike granična.

Tablica 5.2 Ostvarena greška tipa I i standardna devijacija tehnika na hrvatskom skupu podataka

Veličina sustava	DFSE		Bagging		Boosting		Stat. testovi	
	Gr. tipa I	Std	Gr. tipa I	Std	Gr. tipa I	Std	Anova	Fried.
3	0,469	0,015	0,461	0,012	0,461	0,012	x--	x--
5	0,469	0,014	0,459	0,013	0,463	0,015	x--	x--
7	0,470	0,015	0,460	0,011	0,460	0,013	xx-	x--
9	0,477	0,014	0,458	0,012	0,459	0,013	xx-	xx-
11	0,477	0,017	0,459	0,011	0,463	0,013	xx-	xx-
13	0,483	0,015	0,459	0,011	0,465	0,012	?	xxx
15	0,487	0,013	0,459	0,011	0,465	0,011	xx-	xx-
17	0,488	0,012	0,461	0,012	0,462	0,011	xx-	xx-
19	0,493	0,012	0,461	0,012	0,463	0,013	xx-	xx-
21	0,496	0,011	0,460	0,013	0,465	0,016	?	xx-
23	0,498	0,009	0,461	0,012	0,463	0,013	xx-	xx-
25	0,502	0,008	0,460	0,011	0,465	0,013	xx-	xx-
Prosjek	0,484	-	0,460	-	0,463	-	-	-

Za razliku od DFSE tehnike, rezultati Bagging i Boosting tehnika promatrani mjerom greške tipa I su gotovo konstantni tj. izmjerene vrijednosti ne odstupaju mnogo od srednjih vrijednosti. Obje tehnike postižu manju grešku od DFSE, gdje je razlika veća u slučaju kada sustav ima više članova što je grafički prikazano na slici 5.5.



Slika 5.5 Odnos greške tipa I i broja klasifikatora prema izvedenim mjerenjima na hrvatskom skupu podataka

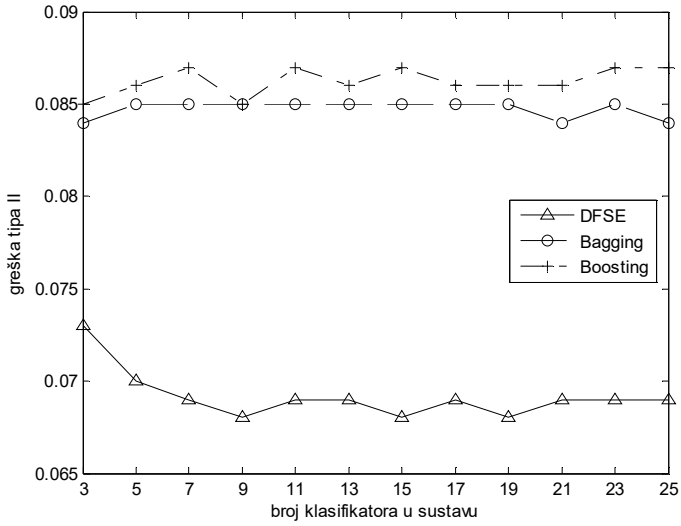
Greška tipa II predstavlja suprotan slučaj kada banka klasificira dobrog klijenta pogrešno i odbije njegov zahtjev za odobravanje kredita. U tablici 5.3 se nalaze rezultati istraživanja koji se odnose na prosječnu mjeru greške tipa II na hrvatskom skupu podataka. Odabrane tehnike su ostvarile gotovo konstante vrijednosti bez obzira na veličinu sustava, tj. postignuti rezultati se ne mijenjaju značajno s dodavanjem novih klasifikatora u sustav. Ako se na slici 5.6 promatra samo DFSE tehnika vidljivo je da se izmjerene vrijednosti značajnije razlikuju samo na sustavu koji ima 3 klasifikatora, gdje je greška veća za 0.4 postotna boda u

odnosu na prosjek. Razlika rezultata u korist DFSE u odnosu na druge tehnike u prosjeku iznosi 1.6 i 1.7 postotna boda za Bagging i Boosting, respektivno. Rezultati tehnika Bagging i Boosting su vrlo slični stoga je i njihov prosjek gotovo isti, tehnika Bagging ima manji prosjek za 0.1 postotna boda u odnosu na tehniku Boosting.

Tablica 5.3 Ostvarena greška tipa II i standardna devijacija tehnika na hrvatskom skupu podataka

Veličina sustava	DFSE		Bagging		Boosting		Stat. testovi	
	Gr. tipa II	Std	Gr. tipa II	Std	Gr. tipa II	Std	Anova	Fried.
3	0,073	0,005	0,084	0,004	0,085	0,004	xx-	xx-
5	0,070	0,005	0,085	0,004	0,086	0,003	xx-	xx-
7	0,069	0,004	0,085	0,004	0,087	0,004	xx-	xx-
9	0,068	0,005	0,085	0,003	0,085	0,004	xx-	xx-
11	0,069	0,005	0,085	0,004	0,087	0,005	xx-	xx-
13	0,069	0,004	0,085	0,004	0,086	0,005	xx-	xx-
15	0,068	0,004	0,085	0,004	0,087	0,004	xx-	xx-
17	0,069	0,004	0,085	0,003	0,086	0,003	xx-	xx-
19	0,068	0,005	0,085	0,004	0,086	0,005	xx-	xx-
21	0,069	0,004	0,084	0,004	0,086	0,004	xx-	xx-
23	0,069	0,004	0,085	0,003	0,087	0,004	xx-	xx-
25	0,069	0,004	0,084	0,004	0,087	0,004	xx-	xx-
Prosjek	0,069	-	0,085	-	0,086	-	-	-

Statistički testovi jednoznačno pokazuju da u svim mjerenjima postoji razlika između rezultata DFSE tehnike te tehnika Bagging i Boosting, dok u rezultatima te dvije tehnike razlika međusobno ne postoji.



Slika 5.6 Odnos greške tipa II i broja klasifikatora prema izvedenim mjerenjima na hrvatskom skupu podataka

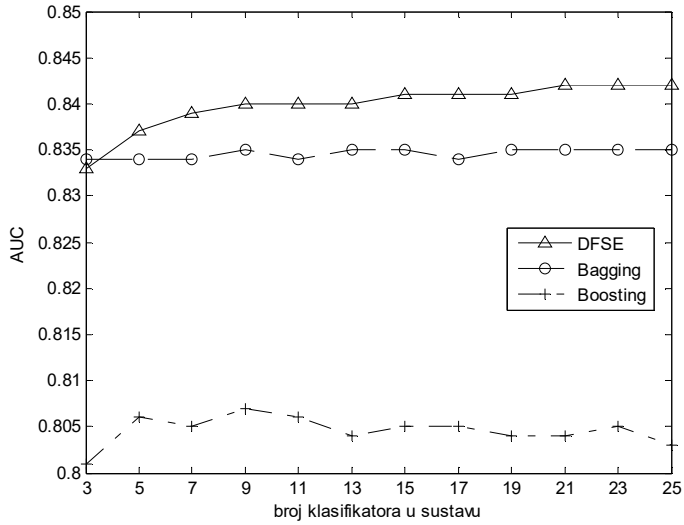
Četvrta mjera koja je korištena za analizu performansi klasifikatora je AUC, a rezultati su prikazani u tablici 5.4. Dok se točnost klasifikacije zasniva na jednoj graničnoj vrijednosti (engl. *cut-off value*), ROC krivulja koju koristi AUC mjera iscrtava mnogo graničnih

vrijednosti koje predstavljaju odnos između TPR (broj ispravnih klasifikacija u pozitivnoj klasi u odnosu na ukupan broj pozitivnih primjera) i FPR (broj krivih klasifikacija u pozitivnoj klasi u odnosu na ukupan broj negativnih primjera). Optimizacija kreditnih modela maksimizacijom AUC mjere se često koristi kada: trošak pogrešne klasifikacije nije jednak za obje klase ili zastupljenost pojedinih klasa po broju primjera nije jednaka. U domeni procjene kreditnog rizika, AUC mjera se smatra relevantnijom od mjere točnosti klasifikacije.

Tablica 5.4 Ostvarena AUC mjera i standardna devijacija tehnika na hrvatskom skupu podataka

Veličina sustava	DFSE		Bagging		Boosting		Stat. testovi	
	AUC	Std	AUC	Std	AUC	Std	Anova	Fried.
3	0,833	0,006	0,834	0,004	0,801	0,009	-xx	-xx
5	0,837	0,006	0,834	0,004	0,806	0,007	-xx	-xx
7	0,839	0,005	0,834	0,004	0,805	0,007	xxx	-xx
9	0,840	0,005	0,835	0,004	0,807	0,007	xxx	-xx
11	0,840	0,006	0,834	0,004	0,806	0,008	xxx	-xx
13	0,840	0,005	0,835	0,004	0,804	0,010	xxx	xxx
15	0,841	0,005	0,835	0,004	0,805	0,007	xxx	xxx
17	0,841	0,005	0,834	0,004	0,805	0,006	xxx	xxx
19	0,841	0,004	0,835	0,004	0,804	0,007	xxx	xxx
21	0,842	0,004	0,835	0,004	0,804	0,009	xxx	xxx
23	0,842	0,004	0,835	0,004	0,805	0,006	xxx	xxx
25	0,842	0,004	0,835	0,004	0,803	0,009	xxx	xxx
Prosjek	0,840	-	0,834	-	0,805	-	-	-

Rezultati pokazuju da je tehnika DFSE postigla najbolje rezultate u svim mjerenjima osim prvom koji uključuje sustave veličine 3, gdje je tehnika Bagging ostvarila neznatno veći AUC. Izmjereni AUC tehnike DFSE polagano raste s povećanjem broja klasifikatora u sustavu, iako je porast vrijednosti konstantan povećanje nije veliko i iznosi 0,05 između sustava od 5 i 25 članova.



Slika 5.7 Odnos AUC mjere i broja klasifikatora prema izvedenim mjerenjima na hrvatskom skupu podataka

Razlika u performansama između Boosting i ostalih tehnika je značajna, što potvrđuju statistički testovi, i iznosi u prosjeku 2,9 za Bagging i 3,5 za DFSE. Na mjerenjima koja su uključivala sustave s manjima brojem članova, razlika između tehnika DFSE i Bagging se statistički ne može dokazati prema odabranoj vrijednosti *p*. Rezultati za mjeru AUC su prikazani na grafu 5.7.

5.5.2. Njemački skup podataka

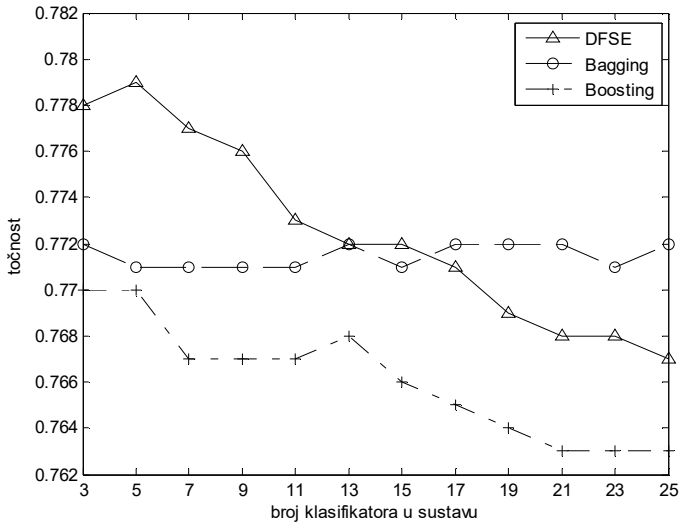
Eksperiment je proveden 30 puta na njemačkom skupu podatka za svaku veličinu sustava od 3 do 25. Točnost klasifikacije odabranih metoda se nalazi u tablici 5.5.

Tablica 5.5 Ostvarena točnost i standardna devijacija tehnika na njemačkom skupu podataka

Veličina sustava	DFSE		Bagging		Boosting		Stat. testovi	
	Točnost	Std	Točnost	Std	Točnost	Std	Anova	Fried.
3	0,778	0,004	0,772	0,004	0,770	0,005	xx-	xx-
5	0,779	0,004	0,771	0,004	0,770	0,005	xx-	xx-
7	0,777	0,005	0,771	0,004	0,767	0,006	xxx	xxx
9	0,776	0,006	0,771	0,004	0,767	0,005	xxx	xxx
11	0,773	0,008	0,771	0,004	0,767	0,006	-x-	-x-
13	0,772	0,007	0,772	0,004	0,768	0,006	---	---
15	0,772	0,006	0,771	0,004	0,766	0,007	-xx	-xx
17	0,771	0,006	0,772	0,004	0,765	0,007	-xx	-xx
19	0,769	0,006	0,772	0,004	0,764	0,009	?	-xx
21	0,768	0,006	0,772	0,004	0,763	0,010	-xx	x-x
23	0,768	0,005	0,771	0,004	0,763	0,007	xxx	x-x
25	0,767	0,006	0,772	0,004	0,763	0,007	xxx	x-x
Prosjek	0,772		0,771		0,766		-	-

Tehnika DFSE je ostvarila najbolje rezultate za sustave koji imaju manji broj članova, dok je za sustave od veličine 17 do 25 Bagging tehnika ostvarila bolje rezultate. Rezultati potvrđuju ranije pronalaske, da se ponovnom upotrebom istih tehnika za odabir atributa treniraju klasifikatori koji ne stvaraju raznolike odgovore što utječe na performanse točnosti.

Opadanjem kvalitete predviđanja tehnike DFSE, točnost ostvarena korištenjem Bagging tehnike, koja je gotovo konstanta kroz sva mjerenja, postaje viša. Od mjerenja kada je sustav veličinom prešao broj klasifikatora, do tog broja da je svaka tehnika za odabir atributa mogla biti korištena tri puta, točnost sustava je pala ispod rezultata Bagging tehnike. Bitno je naglasiti, a što je jasno vidljivo na slici 5.8 koja prikazuje rezultate, da su performanse ostvarene Bagging tehnikom na većim sustavima, gdje je ona najbolja, značajno slabije od rezultata ostvarenih DFSE tehnikom na manjim. Stoga je korisnost rezultata Bagging tehnike vrlo mala. Post-hoc testovi pokazuju da između tehnika DFSE i Bagging postoji statistički značajna razlika u točnosti klasifikacije na sustavima veličine do 11 članova i sustavima većim od 23 člana. Razlika na manjim sustavima jest u korist DFSE tehnike jer ostvaruje bolje rezultate dok je na većim sustavima odnos tehnika obrnut. Kao što je naglašeno najbolji rezultati se postižu na manjim sustavima, a postiže ih DFSE tehnika.



Slika 5.8 Odnos točnosti i broja klasifikatora prema izvedenim mjerenjima na njemačkom skupu podataka

Tehnika Boosting je ostvarila najlošije rezultate u provedenim mjerenjima, s time da ostvarena točnost pada s porastom veličine SVK-a.

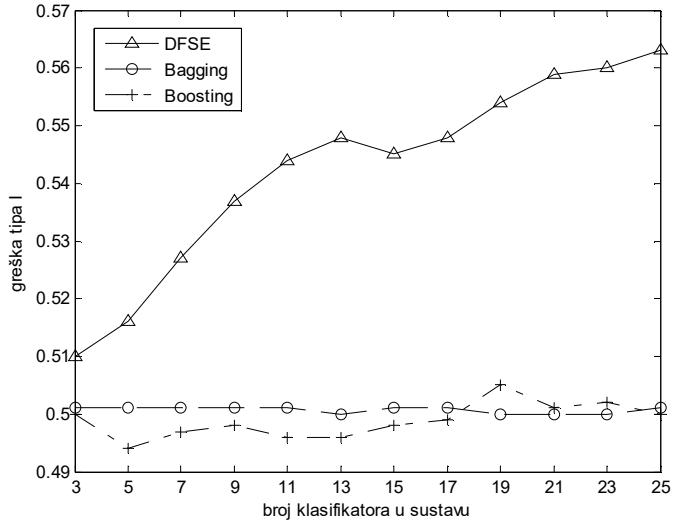
Performanse tehnika promatrane pomoću mjere greške tipa I su drugačije u odnosu na ranije promatranu točnost (tablica 5.6). DFSE tehnika je ostvarila najlošije rezultate jer u svim mjerenjima podijeljenim prema veličini sustava ima najveću vrijednost prosječne greške, što

se odražava i na najvećoj vrijednosti prosjeka svih mjerenja. Kada se promatra ukupan prosjek svih mjerenja (zadnji red tablice) tada je vidljivo da je tehnika Boosting ostvarila nešto bolji rezultat od tehnike Bagging.

Tablica 5.6 Ostvarena greška tipa I i standardna devijacija tehnika na njemačkom skupu podataka

Veličina sustava	DFSE		Bagging		Boosting		Stat. testovi	
	Gr. tipa I	Std	Gr. tipa I	Std	Gr. tipa I	Std	Anova	Fried.
3	0,510	0,014	0,501	0,008	0,500	0,010	xx-	xx-
5	0,516	0,017	0,501	0,009	0,494	0,015	xx-	xx-
7	0,527	0,026	0,501	0,009	0,497	0,018	?	xx-
9	0,537	0,047	0,501	0,009	0,498	0,014	?	xx-
11	0,544	0,064	0,501	0,009	0,496	0,014	?	xx-
13	0,548	0,058	0,500	0,009	0,496	0,014	?	xx-
15	0,545	0,051	0,501	0,009	0,498	0,015	?	xx-
17	0,548	0,047	0,501	0,008	0,499	0,019	?	xx-
19	0,554	0,043	0,500	0,008	0,505	0,019	?	xx-
21	0,559	0,040	0,500	0,008	0,501	0,021	?	xx-
23	0,560	0,037	0,500	0,008	0,502	0,018	?	xx-
25	0,563	0,038	0,501	0,008	0,500	0,014	?	xx-
Prosjek	0,542		0,501		0,499		-	-

Najniža vrijednost greške iznosi 0,494 i rezultat je tehnike Boosting provedene na sustavu od 5 klasifikatora. Razlika prosjeka između Boosting tehnike i ostalih tehnika je: 0,02 u odnosu na Bagging i 4,3 u odnosu na DFSE tehniku. Rezultati pokazuju (slika 5.9) da DFSE tehnika, jednako kao i na hrvatskom skupu podataka, ostvaruje porast greške s povećanjem broja klasifikatora u sustavu stoga je razlika između tehnike Boosting i DFSE na manjim sustavima značajnije manja.



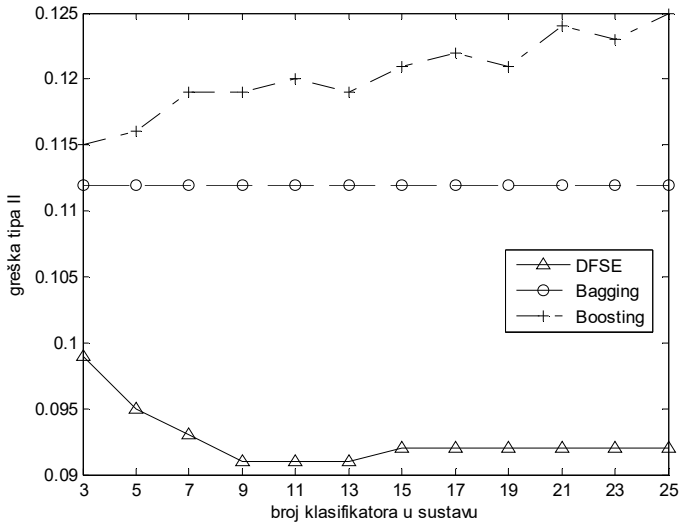
Slika 5.9 Odnos greške tipa I i broja klasifikatora prema izvedenim mjerenjima na njemačkom skupu podataka

Rezultati greške tipa II prikazani u tablici 5.7 su također usporedivi s onima postignutim na hrvatskom skupu podataka. DFSE tehnika je u oba slučaja postigla znatno bolje rezultate od tehnika Bagging i Boosting i to u prosjeku za 1.9 i 2.7 postotna boda respektivno, što predstavlja značajno poboljšanje.

Tablica 5.7 Ostvarena greška tipa II i standardna devijacija tehnika na njemačkom skupu podataka

Veličina sustava	DFSE		Bagging		Boosting		Stat. testovi	
	Gr. tipa II	Std	Gr. tipa II	Std	Gr. tipa II	Std	Anova	Fried.
3	0,099	0,005	0,112	0,005	0,115	0,008	xx-	xx-
5	0,095	0,006	0,112	0,005	0,116	0,007	xxx	xxx
7	0,093	0,007	0,112	0,005	0,119	0,009	xxx	xxx
9	0,091	0,015	0,112	0,005	0,119	0,008	xxx	xxx
11	0,091	0,022	0,112	0,005	0,120	0,008	?	xxx
13	0,091	0,020	0,112	0,005	0,119	0,006	?	xxx
15	0,092	0,019	0,112	0,005	0,121	0,008	?	xxx
17	0,092	0,016	0,112	0,005	0,122	0,008	?	xxx
19	0,092	0,014	0,112	0,005	0,121	0,011	?	xxx
21	0,092	0,012	0,112	0,005	0,124	0,010	xxx	xxx
23	0,092	0,011	0,112	0,005	0,123	0,006	xxx	xxx
25	0,092	0,011	0,112	0,005	0,125	0,009	xxx	xxx
Prosjek	0,093		0,112		0,120		-	-

Na njemačkom skupu podataka jednako kao i na hrvatskom DFSE tehnika je napravila nešto veću grešku na sustavu s tri člana, koja je smanjena nakon povećanja veličine sustava što se može vidjeti na slici 5.10 Najniža vrijednost prosječne greške tipa II je 0,91 a ostvarena je DFSE tehnikom na sustavima veličine 9, 11 i 13.



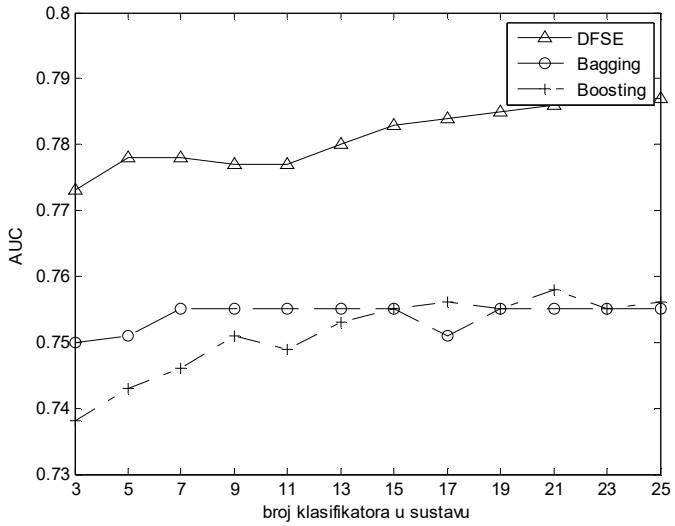
Slika 5.10 Odnos greške tipa II i broja klasifikatora prema izvedenim mjerenjima na njemačkom skupu podataka

AUC vrijednosti eksperimenta na njemačkom skupu podataka se nalaze u tablici 5.8. Najviše vrijednosti po svim mjerenjima kao i ukupno najvišu vrijednost je ostvarila DFSE tehnika. Najviše vrijednosti su ostvarene na sustavima koji uključuju najveći broj klasifikatora.

Tablica 5.8 Ostvarena AUC mjera i standardna devijacija tehnika na njemačkom skupu podataka

Veličina sustava	DFSE		Bagging		Boosting		Stat. testovi	
	AUC	Std	AUC	Std	AUC	Std	Anova	Fried.
3	0,773	0,021	0,750	0,008	0,738	0,010	?	xxx
5	0,778	0,012	0,751	0,008	0,743	0,009	xxx	xx-
7	0,778	0,011	0,755	0,008	0,746	0,010	xxx	xxx
9	0,777	0,010	0,755	0,008	0,751	0,012	xx-	xx-
11	0,777	0,008	0,755	0,007	0,749	0,010	xxx	xx-
13	0,780	0,007	0,755	0,008	0,753	0,009	xx-	xx-
15	0,783	0,007	0,755	0,008	0,755	0,008	xx-	xx-
17	0,784	0,006	0,751	0,008	0,756	0,009	xx-	xx-
19	0,785	0,005	0,755	0,008	0,755	0,011	xx-	xx-
21	0,786	0,005	0,755	0,007	0,758	0,011	xx-	xx-
23	0,787	0,005	0,755	0,008	0,755	0,010	xx-	xx-
25	0,787	0,004	0,755	0,008	0,756	0,012	xx-	xx-
Prosjek	0,781		0,754		0,751		-	-

Odnos ostvarenih AUC vrijednosti svih korištenih tehnika jest vidljiv na slici 5.11 koja prikazuje rezultate iz tablice 5.8.



Slika 5.11 Odnos AUC mjere i broja klasifikatora prema izvedenim mjerenjima na njemačkom skupu podataka

5.5.3. Vremena treninga i testa podataka

Istraživanje je provedeno na osobnom računalu konfiguracije: Intel Core i3 CPU M330 2.13GHz, 4 GB RAM, Windows 7 64bit. Tijekom provođenja mjerenja bilježeno je vrijeme potrebno za treniranje modela podataka t_1 i vrijeme potrebno za testiranje t_2 u milisekundama (ms) kako je prikazano na slici 5.3.

U slučaju DFSE tehnike, vrijeme treninga uključuje vrijeme izvođenja: tehnike za odabir atributa, operatora za odabir top k atributa i vrijeme izrade modela klasifikacijskog algoritma. Kako tehnike Bagging i Boosting ne koriste tehnike za odabir atributa njihovo trajanje treninga je određeno samo vremenom izvođenja klasifikacijskog algoritma. Testno vrijeme je za sve tehnike isto i odnosi se na vrijeme potrebno za testiranje 1000 primjera tj. 10 preklapanja po 100 primjera iz validacije.

Tablica 5.9 Vrijeme potrebno za trening i test modela prikupljeno prilikom provođenja istraživanja

Tehnika	Hrvatski skup		Njemački skup	
	Trening (ms)	Test (ms)	Trening (ms)	Test (ms)
Bagging	791.761	11	677.234	17
Boosting	849.672	15	719.697	19
DFSE	184.050	15	84.820	16

Vremena prikazana u tablici 5.9 su prosjek 30 testova koji su provedeni na veličini sustava 25. Vremena trajanja pojedinačnih testova su gotovo identična stoga nije navedena standardna devijacija skupa jer je zanemariva.

Iz prikupljenih podataka može se zaključiti da je DFSE tehnici potrebno najmanje vremena za trening modela konstruiranih u ovom istraživanju. Poznato je da su Bagging i Boosting tehnike vremenski zahtjevne kada se koriste klasifikacijski algoritmi koji nisu slabi. U slučaju hrvatskog skupa podataka DFSE tehnika je završila trening modela oko 4.3 puta prije nego Bagging i 4.6 puta prije tehnike Boosting. Razlika na njemačkom skupu podatka je još izraženija jer je vrijeme DFSE tehnike za trening modela otprilike 8 puta kraće od tehnike Bagging i 8.5 od tehnike Boosting.

Vrijeme potrebno za testiranje modela je vremenski zanemarivo kratko i traje podjednako bez obzira na korištenu tehniku.

5.6 Pregled rezultata iz literature

U nastavku slijedi vrednovanje ostvarenih rezultata DFSE tehnike pomoću rezultata autoru dostupnih istraživanja u posljednjih 5 godina. Uvjet da bi rezultati bili u potpunosti usporedivi, jest da istraživanja moraju koristiti iste: skupove podataka, validacije i mjere evaluacije. Tablica 5.10 sadrži popis dostupnih istraživanja i njihov opis kroz osnovne karakteristike potrebne za usporedbu. Iako neka istraživanja u tablici ne odgovaraju u potpunosti zadanim uvjetima, ista su dodana radi stvaranja šire slike postignuća na području klasifikacije kreditnog rizika.

Istraživanja razmatrana u ovom poglavlju unutar samog rada sadrže nekoliko, često i mnogo klasifikacijskih modela čije performanse se uspoređuju. Radi ograničenog prostora i sažetosti usporedbe iz svakog istraživanja odabran je onaj model koji je ostvario najbolje rezultate. Odabir je izvršen pomoću slijedećeg pravila; model koji je ostvario najbolje rezultate točnosti klasifikacije je proglašen najboljim i ostale mjere evaluacije ukoliko su dostupne u istraživanju odnose se na odabrani model. Na primjeru istraživanja provedenog u poglavlju 5. ove disertacije, za njemački skup kreditnih podataka, temeljem opisanih pravila izabrana je tehnika DFSE s 5 članova i najvišom točnosti od 0,779. Ostale mjere su prepisane za odabrani klasifikator. Točnost klasifikacije kao glavi faktor odabira modela jest izabran na temelju učestalosti korištenja, jer gotovo sva istraživanja koriste tu mjeru, a često i isključivo samo nju.

Osim mjera za evaluaciju i tipa validacije u tablici je dostupan i podatak o klasifikacijskom algoritmu na kojem počiva najbolji klasifikacijski model pojedinog istraživanja. Navedeni podatak označava samo klasifikacijski algoritam za trening i ne sadrži podatke o cijelom korištenom modelu koji dodatno može uključivati: algoritme za smanjenje dimezionalnosti, algoritam za odabir optimalnih parametara ili bilo koju drugu strategiju pretprocesiranja podataka.

Odabrane vrijednosti DFSE modela predstavljaju pesimističnu evaluaciju performansi modela. Tehnika DFSE odabire optimalan broj članova pomoću smanjivanja sustava na onaj broj koji daje najbolje rezultate. U ovom slučaju broj članova je apriori odabran i ne predstavlja samo ona mjerenja gdje je skup od 5 klasifikatora najbolje rješenje već uključuje sva mjerenja i njihove rezultate u kojima je broj članova jednak 5. Usporedbe radi, u istraživanju provedenom u poglavlju 4. gdje je DFSE tehnika korištena sa smanjivanjem na

najbolji podskup klasifikatora, prosječna točnost na njemačkom skupu podataka je iznosila 0.781 a na hrvatskom 0,834.

Tablica 5.10 Rezultati istraživanja klasifikacije podataka na njemačkom i hrvatskom skupu kreditnih podataka iz literature

Istraživanje	Klasifikacijski algoritam*	Skup podataka	Validacija **	Evaluacija			
				Točnost	Greška tipa I	Greška tipa II	AUC
DFSE* **	**SVK**	**Njemački**	**10F**	**0,779**	**0,516**	**0,095**	**0,778**
Nanni i Lumini, 2009.	SVK	Njemački	70/30	0,739	0,662	0,12	0,78
Peng et al., 2011	SVM	Njemački	10F	0,774	-	-	0,693
Ping & Yongheng, 2011	HIB	Njemački	10F	0,766	-	-	-
Wang et al., 2011	SVK	Njemački	80/20	0,763	0,543	0,105	-
Hens i Tiwari, 2012	HIB	Njemački	10F	0,777	-	-	-
Marqués et al., 2012a	SVK	Njemački	5F	0,774	-	0,15	-
Marqués et al., 2012b	SVK	Njemački	5F	-	0,55	0,09	0,80
Salama i Freitas, 2012	NB	Njemački	10F	0,756	-	-	-
Zhu et al., 2013	LogR	Njemački	80/20	0,766	0,506	0,116	-
Zurada, 2013	kNN	Njemački	10F	0,76	0,569	0,099	0,738
Alaraj et al., 2014	SVK (GS)	Njemački	80/20	0,783	-	-	-
Oreški i Oreški, 2014	HIB (GS)	Njemački	10F	0,789	-	-	-
Bouaguel i Li., 2015	NN	Njemački	10F	0,769	-	-	0,645
DFSE* **	**SVK**	**Hrvatski**	**10F**	**0,831**	**0,47**	**0,069**	**0,839**
Oreški & Oreški, 2012	HIB (GS)	Hrvatski	10F	0,824	-	-	-
Oreški & Oreški, 2014	HIB (GS)	Hrvatski	10F	0,828	-	-	-

** Stupac „Klasifikacijski algoritam" sadrži podatak na kojem se klasifikacijskom algoritmu temelji najbolji klasifikator istraživanja. Korištene skraćenice su: SVK – sustav višestrukih klasifikatora, HIB – hibridna tehnika, SVM – tehnika potpornih vektora, LogR – logistička regresija, NB – naivni Bayes, NN – neuronska mreža, kNN – algoritam k najbližih susjeda. Skraćenica GS označava da su rezultati ostvareni pomoću algoritma koji koristi globalnu pretragu prostora rješenja.*

*** Stupac „Validacija" prikazuje koja vrstu validacije korištenu u istraživanju. Korištene skraćenice su: xF – x unakrsnih validacija s preklapanjem, y/z – hold-out validacija gdje je y postotak korištenih primjera za trening a z preostali postotak za test*

**** U usporedbi rezultata je korištena pesimistična evaluacija performansi DFSE tehnike.*

Točnost klasifikacije je zbog jednostavnosti izračuna i dostupnosti u alatima za klasifikaciju najčešće korištena (gotovo benchmark) mjera za evaluaciju performansi modela. Prilikom analize performansi potrebno je uzeti u obzir tip klasifikacijskog algoritma, te razdvojiti algoritme koji koriste globalnu pretragu od onih koji ju ne koriste. Iako globalna pretraga često rezultira s boljim rezultatima (vidljivo i u ovoj usporedbi) cijena jest značajno povećanje potrebnih procesorskih (vremenskih) i memorijskih računalnih resursa. U svojoj osnovi tehnika DFSE je konstruirana kao brza tehnika s razumno jednostavnom implementacijom, koja ima za cilj omogućiti „dobre" rezultate istraživačima koji nisu detaljno upoznati s teorijom strojnog učenja. U tom kontekstu na njemačkom skupu podataka DFSE tehnika je ostvarila lošije rezultate od klasifikatora koji su koristili globalnu pretragu za 0,4 (Alaraj et al., 2014) i 1 (Oreški i Oreški, 2014) postotni bod. Pri tome je bitno napomenuti da je u istraživanju (Alaraj et al., 2014) korištena hold-out validacija koja može rezultirati s

iskrivljenim procjenama performansi u odnosu na unakrsnu validaciju. DFSE tehnika u odnosu na ostala istraživanja koja ne koriste globalnu pretragu postiže bolje rezultate po pitanju točnosti.

Pogrešna klasifikacija loših klijenata kao dobrih čini grešku tipa I, stoga se minimizacijom ove greške smanjuje broj loših klijenata čiji zahtjevi su prihvaćeni jer su prepoznati kao dobri. Kod svih istraživanja greška ovog tipa je značajno veća od greške tipa II zbog neravnoteže u broju loših klijenata u odnosu na dobre klijente u trening skupu podataka, gdje su loši klijenti manjinska klasa (engl. *minority class*). Odnos greške tipa I i greške tipa II je poseban izazov prilikom konstruiranja klasifikatora koji se povezuje s troškovima klasifikacije te nije obuhvaćen ovom disertacijom. Idealno, trenirani klasifikatori bi trebali smanjivati obje greške. Istraživanje (Zhu et al., 2013) je postiglo najnižu grešku tipa I koja iznosi 0,506 i ako je usporedimo s DFES tehnikom ona je manja za 1 postotni bod (ukoliko se zanemari različita validacija korištena u istraživanjima). Visina greške tipa I u ostalim istraživanjima je viša u odnosu na rezultate DFSE tehnike.

Greška tipa II mjeri pogrešnu klasifikaciju dobrih primjera kao loših, stoga se minimizacijom ove greške smanjuje broj dobrih klijenata čiji zahtjevi su odbijeni jer su prepoznati kao loši. Od dostupnih istraživanja koja su koristila tu mjeru, jedno istraživanje (Marqués et al., 2012b) ima gotovo isti rezultat kao DFSE tehnika, razlika se ne može utvrditi jer navedeno istraživanje iskazuje rezultat na dvije decimale. Osim spomenutog i istraživanje (Zurada, 2013) ima manju vrijednost od 0.10, dok sva ostala istraživanja imaju značajnije veću grešku tipa II. Analiza pokazuje da tehnika DFSE postiže dobre rezultate u usporedbi s ostalim tehnikama iz literature po pitanju greške tipa II.

Rezultati prikazani AUC mjerom pokazuju da je klasifikator korišten u istraživanju (Marqués et al., 2012b) postigao 0.8 tj. najbolji rezultat među prezentiranim radovima. Taj rezultat je za 2.2 postotna boda veći od AUC mjere postignute DFSE tehnikom i 2 postotna poena veći od istraživanja (Nanni i Lumini, 2009.).

Istraživanja prikupljena za hrvatski skup podataka predstavljaju radove u kojima je autor disertacije sudjelovao. U objavljenim radovima klasifikacija se bazirala na hibridnoj tehnici koja je kombinirala genetski algoritam i neuronske mreže. Kreirane tehnike GA-NN i HGA-NN su vremenski i memorijski mnogo zahtjevnije u odnosu na DFSE tehniku jer koriste genetski algoritam koji je znatno računalno zahtjevniji kod izvođenja. Unatoč tome, tehnika DFSE jest ostvarila bolje rezultate, za 0.03 postotna boda. Za razliku od njemačkog skupa podataka gdje je korištena strategija heterogenih klasifikatora na hrvatskom skupu su

korištene isključivo neuronske mreže koje su rezultirale s poboljšanjem performansi u odnosu na prijašnje rezultate.

5.7 Zaključci poglavlja

U ovom poglavlju jest provedeno šire vrednovanje kvalitete klasifikacije DFSE tehnike kroz četiri različite mjere. Ocjena kvalitete rezultata se temelji na usporedbi s trenutno dvije najpopularnije tehnike kombiniranja klasifikatora: Bagging i Boosting.

Eksperiment je proveden 30 puta za svaku veličinu SVK-a na njemačkom i hrvatskom kreditnom skupu podataka, gdje je u odnosu na prethodno istraživanje heterogeni sustav korišten na njemačkom skupu minimalno izmijenjen, dok je homogeni ostao identičan. Rezultati slijede zaključke istraživanja iz poglavlja 4 tj. DFSE tehnika je postigla najbolje rezultate kada veličina sustava odgovara broju različitih tehnika korištenih za odabir atributa. Iako se korištenjem smanjivanja sustava unutar DFSE tehnike broj klasifikatora sam optimizira, najbolji rezultati ovog istraživanja pokazuju veličinu sustava koja će u većini slučaja biti odabrana. Sposobnost postizanja najboljih rezultata na manjim sustavima znači stvaranje efikasnih sustava, što predstavlja jedan od ciljeva disertacije, a ujedno i prednost ove tehnike nad tehnikama Bagging i Boosting.

Osim efikasnosti rezultati pokazuju i bolje performanse na tri od četiri odabrane mjere u odnosu na Bagging i Boosting. Najbolji rezultati prema mjerama: točnosti, greške tipa II i AUC su ostvareni DFSE tehnikom. Točnost klasifikacije ovisi o raznolikosti sustava te je povezana s brojem klasifikatora u sustavu. Na manjim veličinama sustava, gdje je superiorna, tehnika DFSE ostvaruje rezultate koji su značajno bolji od ostalih tehnika. Rezultati greške tipa II i AUC mjere ostvareni korištenjem DFSE tehnike su u svim provedenim mjerenjima značajno bolji od onih tehnika Bagging i Boosting. Četvrta mjera je za razliku od prve tri minimalnu vrijednost u eksperimentu postigla tehnikom Boosting. Razlika između Boosting i DFSE rezultata prema greški tipa I se povećava s porastom veličine sustava, čime oni manji sustavi odabrani od DFSE tehnike imaju rezultate koji nisu mnogo lošiji od ostalih tehnika. Bitno je napomenuti da iako je istraživanje provedeno pomoću četiri različite mjere, glavni fokus je zbog načina na koji je definirana istraživačka hipoteza stavljen na točnost klasifikacije. Stoga je i prilikom konstruiranja DFSE tehnike korišten COP algoritam koji

jednako vrednuje primjere bez obzira na klasu, što se jednostavno u slučaju drugačijeg istraživačkog pristupa može promijeniti tako da se uzme u obzir i važnost klasa.

Postignuti rezultati su uspoređeni s ostalim istraživanjima iz literature. Rezultati DFSE tehnike promatrani kroz sve četiri mjere se mogu ocijeniti kao vrlo obećavajući u odnosu na postavljenje ciljeve. Na njemačkom skupu podataka, po pitanju točnosti i greške tipa II (ukoliko izuzmemo algoritme s globalnom pretragom), DFSE tehnika je ostvarila najbolje rezultate od uspoređenih istraživanja. U preostalim mjerama su rezultati unutar 3 najbolja od svih korištenih koji su iskazali potrebne mjere evaluacije. Potencijal nove tehnike je također vidljiv na rezultatima ostvarenim na hrvatskom skupu podataka koji su bolji u odnosu na tehnike koje koriste kompleksnije algoritme.

Nakon provedenog istraživanja, na temelju ostvarenih rezultata i statističke potvrde istih, promatrajući sustave veličine koje odabire DFSE tehnika pomoću smanjivanja tj. sustave manjeg broja članova, prihvaća se istraživačka hipoteza H2. U obzir treba uzeti i način prikazivanja (izvođenja) rezultata DFSE tehnike u ovom istraživanju, uspoređeni rezultati se mogu smatrati pesimističnom procjenom jer nije korišten odabir najboljeg modela. U konačnici može se zaključiti da tehnika DFSE koja se bazira na kombinaciji različitih tehnika za odabir atributa i bez manipulacije parametara može svojim rezultatima predstavljati dodatak drugim, već dostupnim tehnikama.

U budućim istraživanjima bilo bi zanimljivo primijeniti DFSE tehniku na druge domene kako bi se istražila njezina primjenjivost na probleme koji nisu procjena kreditnog rizika.

6

Zaključak

Doktorska disertacija se zaključuje analizom realizacije znanstvenih ciljeva istraživanja i očekivanog znanstvenog doprinosa rada.

Svrha doktorske disertacije je bila istražiti primjenjivost sustava višestrukih klasifikatora temeljnog na upravljanom odabiru atributa na problem procjene kreditnog rizika građana.

U skladu s definiranom svrhom postavljen je i glavni cilj istraživanja: razviti brzu, robusnu tehniku za kombiniranje klasifikatora koja će na temelju upravljanog odabira atributa stvarati efikasne i kvalitetne sustave za ocjenu sposobnosti tražitelja kredita da vrati kredit. Povrh toga, nova tehnika mora biti dovoljno jednostavna za laku implementaciju i široku primjenu u istraživačkoj zajednici.

Realizacija glavnog i popratnih ciljeva je provedena kroz detaljnije ciljeve istraživanja:

- Razviti mjeru za kvantifikaciju korisnosti uključivanja novih klasifikatora u sustav višestrukih klasifikatora, temeljenu na vrednovanju težine primjera.
- Kreirati robusnu tehniku za konstruiranje sustava višestrukih klasifikatora temeljenu na upravljanom odabiru atributa, koja daje:
    a. bolje rezultate od pojedinačnih klasifikatora uključenih u sustav
    b. jednako dobre ili bolje rezultate klasifikacije kod procjene kreditnog rizika nego najpopularnije tehnike Bagging i Boosting.
- Utvrditi zavisnost između mjere raznolikosti iskazane Q statistikom i mjere točnosti.

Pojedinačni ciljevi su realizirani provođenjem više međusobno povezanih istraživanja, koja su opisana u zasebnim poglavljima:

- U prvom dijelu je dan uvod u sustave višestrukih klasifikatora. Cilj je bio dati opći uvod i definirati osnovne pojmove koji se koriste u istraživanjima predstavljenim u disertaciji. Opisane su strategije konstruiranja sustava višestrukih klasifikatora s naglaskom na odabranu; strategiju temeljenu na manipulaciji ulaznih atributa.
- U slijedećem poglavlju, kao preduvjet za istraživanje postavljenih hipoteza, dano je istraživanje u kojem je razvijen novi algoritam za smanjivanje sustava višestrukih klasifikatora COP. Pohlepni algoritam COP odabire klasifikatore u sustav pomoću nove mjere CO, koja se temelji na novom pogledu kombiniranja glasova u kojem se vrednuje težina pojedinih primjera. Ideja na kojem počiva novi pogled se može definirati; što je klasifikacija primjera, glasanjem uključenih članova, neizvjesnija to je odluka slijedećeg klasifikatora koji se uključuje u sustav značajnija. Novi algoritam je temeljito testiran na vrlo velikom broju generiranih skupova podataka.
- Na temelju COP algoritma i odabira atributa opisana je konstrukcija nove DFSE tehnike. Performanse DFSE tehnike su testirane na dva realna skupa podataka te su

uspoređene s pojedinačnim klasifikatorima unutar sustava. Istražena su dva različita pristupa odabiru klasifikacijskih algoritama, tako da je na jednom skupu korišten homogeni, a na drugom heterogeni sustav klasifikatora. U sklopu istraživanja napravljena je i analiza odnosa Q statistike kao mjere raznolikosti i točnosti klasifikacije.

- Nakon utvrđivanja opravdanosti kombiniranja klasifikatora novom tehnikom, vrednovani su ostvareni rezultati u odnosu na rezultate postignute drugim dostupnim tehnikama za konstrukciju sustava višestrukih klasifikatora. DFSE tehnika je uspoređena s tehnikama Bagging i Boosting. Radi stvaranja šire slike kvalitete korištenih tehnika, u istraživanju su mjerene performanse prema četiri različite mjere: točnost, greška tipa I, greška tipa II i AUC. Dodatno, u okviru istraživanja su mjerena vremena potrebna za trening i testiranje modela svih tehnika uključenih u eksperiment, u svrhu utvrđivanja njihovih odnosa. Na posljetku su rezultati ostvareni DFSE tehnikom uspoređeni s onima objavljenima u literaturi.

Pregledom provedenih istraživanja može se zaključiti da je u radu predstavljen novi pogled na kombiniranje odluka klasifikatora. Osnovna komponenta novog pristupa je upravljani odabir atributa kojim se stvaraju klasifikatori s raznolikim odlukama. Dodatno uz odabir atributa korišteno je i smanjivanje sustava kao korekcijski korak koji dopušta istraživanje različitih postavki prilikom treninga klasifikatora. Obje karakteristike su uključene u novu tehniku DFSE, koja radi željene jednostavnosti ne zahtjeva mijenjanje početnih parametara korištenih tehnika.

Realizacija znanstvenih ciljeva istraživanja i očekivanog znanstvenog doprinosa rada analizira se prihvaćanjem ili odbacivanjem hipoteza istraživanja.

Hipoteze istraživanja

H1: *Sustav višestrukih klasifikatora koji je temeljen na odabiru različitih podskupova atributa pomoću filtarskih tehnika te konstruiran na temelju u ovom radu predloženog algoritma za smanjivanje sustava će postizati statistički značajno veću točnost klasifikacije od pojedinačnih klasifikatora uključenih u sustav na razini statističke značajnosti* $p \leq 0,05$.

Na oba skupa korištena u istraživanju DFSE tehnika je ostvarila statistički značajno veću točnost klasifikacije od najboljeg pojedinačnog klasifikatora uključenog u sustav. Na

njemačkom skupu podataka razlika između rezultata DFSE tehnike i najboljeg člana iznosi 0,013 postotna boda, a na hrvatskom skupu podataka razlika iznosi 0,011 postotna boda. Dobiveni rezultati podupiru hipotezu H1. Stoga se može zaključiti da se

prihvaća prva hipoteza istraživanja.

H2: *Sustav višestrukih klasifikatora koji je temeljen na odabiru različitih podskupova atributa pomoću filtarskih tehnika te konstruiran na temelju u ovom radu predloženog algoritma za smanjivanje sustava će postizati statistički jednake ili bolje rezultate u odnosu na najpopularnije tehnike, Bagging i Boosting primijenjene na originalnim skupovima podataka (njemačkom i hrvatskom) sa svim karakteristikama.*

Točnost klasifikacije koja se u većini istraživanja uzima kao mjera kvalitete klasifikatora je i u okviru provedenog istraživanja uzeta kao mjera performansi. Izmjereni podaci pokazuju da najbolje rezultate ostvaruje DFSE tehnika na manjim veličinama sustava. Iako je istraživanje provedeno na različitim veličinama sustava radi stvaranja šire slike, veličina sustava koju odabire DFSE tehnika postiže statistički značajnije veću točnost klasifikacije u odnosu na Bagging i Boosting tehnike. Ostvareno poboljšanje je dovoljno veliko da bi bilo znanstveno i praktično interesantno. Stoga se

prihvaća druga hipoteza istraživanja.

Prihvaćanjem hipoteza istraživanja H1 i H2 potvrđuje se uspješna realizacije postavljenih znanstvenih ciljeva.

Osim znanstvenog doprinosa postignuti rezultati mogu dati i značajan društveni doprinos u vidu:

- skraćivanja vremena potrebnog za trening sustava višestrukih klasifikatora,
- primjene istraživanja u poslovanju banaka što bi unaprijedilo korištenje tehnika dubinske analize podataka u bankarstvu,
- stvaranja koraka prema ostvarivanju automatizacije procesa odobravanja kredita i
- olakšavanje pristupa kreditima klijentima banke.

Ostvareni rezultati potvrđuju očekivani potencijal uključivanja teorije zasnovane na mudrosti gomile u konstrukciju klasifikacijskih modela. Pri tome valja naglasiti da razvoj

modela povećane prediktivne točnosti predstavlja trajni izazov za strojno učenje, ali primjena tih modela u području otkrivanja znanja u podacima nije nimalo manji izazov. Kombiniranje postojećeg ljudskog znanja i ovakvih modela znatno može pomoći u boljoj alokaciji kapitala u bankarstvu, odnosno za poboljšanje stabilnosti i uspješnosti banaka. Možda sustavi višestrukih klasifikatora nisu zlatni gral u strojnom učenju, ali su sigurno vrijedni daljnjeg istraživanja i primjene. Štoviše bit će zanimljivo pratiti budući razvoj tog područja.

LITERATURA

[1] Abdou, H., Pointon, J., & El-Masry, A. (2008). Neural nets versus conventional techniques in credit scoring in Egyptian banking. Expert Systems with Applications, 35(3), 1275-1292.

[2] Aggarwal, C. C. (Ed.). (2014). Data classification: algorithms and applications. CRC Press.

[3] Alaraj, M., Abbod, M., & Hunaitii, Z. (2014). Evaluating Consumer Loans Using Neural Networks Ensembles. In Int. Conf. on Machine Learning, Electrical and Mechanical Engineering (ICMLEME'2014) (pp. 55-61).

[4] Ansari, D., Nilsson, J., Andersson, R., Regnér, S., Tingstedt, B., & Andersson, B. (2013). Artificial neural networks predict survival from pancreatic cancer after radical surgery. The American Journal of Surgery, 205(1), 1-7.

[5] Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. Journal of the Operational Research Society, 54(6), 627-635.

[6] Banfield, R. E., Hall, L. O., Bowyer, K. W., & Kegelmeyer, W. P. (2005). Ensemble diversity measures and their application to thinning. Information Fusion, 6(1), 49-62.

[7] Banfield, R. E., Hall, L. O., Bowyer, K. W., & Kegelmeyer, W. P. (2003). A new ensemble diversity measure applied to thinning ensembles. In Multiple Classifier Systems (pp. 306-316). Springer Berlin Heidelberg.

[8] Bar, A., Rokach, L., Shani, G., Shapira, B., & Schclar, A. (2012). Boosting Simple Collaborative Filtering Models Using Ensemble Methods. arXiv preprint arXiv:1211.2891.

[9] Basel, International regulatory framework for banks (Basel III) http://www.bis.org/bcbs/basel3.htm. zadnji pristup: 22.07.2015.

[10] Basu, T., & Murthy, C. A. (2012, December). Effective text classification by a supervised feature selection approach. In Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on (pp. 918-925). IEEE.

[11] Belciug, S., & Gorunescu, F. (2013). A hybrid neural network/genetic algorithm applied to breast cancer detection and recurrence. Expert Systems, 30(3), 243-254.

[12] Bellotti, T., & Crook, J. (2009). Support vector machines for credit scoring and discovery of significant features. Expert Systems with Applications, 36(2), 3302-3308.

[13] Bhowan, U., Johnston, M., Zhang, M., & Yao, X. (2014). Reusing genetic programming for ensemble selection in classification of unbalanced data. Evolutionary Computation, IEEE Transactions on, 18(6), 893-908.

[14] Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2012). An ensemble of filters and classifiers for microarray data classification. Pattern Recognition, 45(1), 531-539.

[15] Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2013). A review of feature selection methods on synthetic data. Knowledge and information systems, 34(3), 483-519.

[16] Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A., Benítez, J. M., & Herrera, F. (2014). A review of microarray datasets and applied feature selection methods. Information Sciences, 282, 111-135.

[17] Bouaguel, W., & Limam, M. (2015, January). An Ensemble Wrapper Feature Selection for Credit Scoring. In Proceedings of Fourth International Conference on Soft Computing for Problem Solving (pp. 619-631). Springer India.

[18] Brown, G., & Kuncheva, L. I. (2010). "Good" and "bad" diversity in majority vote ensembles. In Multiple Classifier Systems (pp. 124-133). Springer Berlin Heidelberg.

[19] Brown, G., Wyatt, J., Harris, R., & Yao, X. (2005). Diversity creation methods: a survey and categorisation. Information Fusion, 6(1), 5-20.

[20] Bühlmann, P. (2012). Bagging, boosting and ensemble methods. In Handbook of Computational Statistics (pp. 985-1022). Springer Berlin Heidelberg.

[21] Caruana, R., Niculescu-Mizil, A., Crew, G., & Ksikes, A. (2004, July). Ensemble selection from libraries of models. In Proceedings of the twenty-first international conference on Machine learning (p. 18). ACM.

[22] Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. Computers & Electrical Engineering, 40(1), 16-28.

[23] Chen, M. C., & Huang, S. H. (2003). Credit scoring and rejected instances reassigning through evolutionary computation techniques. Expert Systems with Applications, 24(4), 433-441.

[24] Cherkauer, K. J. (1996, March). Human expert-level performance on a scientific image analysis task by a system using combined artificial neural networks. In Working notes of the AAAI workshop on integrating multiple learned models (pp. 15-21).

[25] Chidlovskii, B., & Lecerf, L. (2008). Scalable feature selection for multi-class problems. In Machine learning and knowledge discovery in databases (pp. 227-240). Springer Berlin Heidelberg.

[26] Coletta, L. F., Hruschka, E. R., Acharya, A., & Ghosh, J. (2015). A differential evolution algorithm to optimise the combination of classifier and cluster ensembles. International Journal of Bio-Inspired Computation, 7(2), 111-124.

[27] Cormen, T. H. (2009). Introduction to algorithms. MIT press.

[28] Cunningham, P., & Carney, J. (2000). Diversity versus quality in classification ensembles based on feature selection. In Machine Learning: ECML 2000 (pp. 109-116). Springer Berlin Heidelberg.

[29] Dai, Q. (2013). A competitive ensemble pruning approach based on cross-validation technique. Knowledge-Based Systems, 37, 394-414.

[30] Dhasal, P., Shrivastava, S. S., Gupta, H., & Kumar, P. (2012). An Optimized Feature Selection for Image Classification Based on SVM-ACO. International Journal of Advanced Computer Research, sept.

[31] Didaci, L., Fumera, G., & Roli, F. (2013). Diversity in classifier ensembles: Fertile concept or dead end?. In Multiple Classifier Systems (pp. 37-48). Springer Berlin Heidelberg.

[32] Dietterich, T. G. (2000). Ensemble methods in machine learning. In Multiple classifier systems (pp. 1-15). Springer Berlin Heidelberg.

[33] Domazet-Lošo, M. (2006). Usporedba postupaka dubinske analize primijenjenih nad biološkim podacima (Doctoral dissertation, M. Sc. Thesis, University of Zagrebu).

[34] El Akadi, A., Amine, A., El Ouardighi, A., & Aboutajdine, D. (2011). A two-stage gene selection scheme utilizing MRMR filter and GA wrapper. Knowledge and Information Systems, 26(3), 487-500.

[35] Fan, W., Chu, F., Wang, H., & Yu, P. S. (2002, July). Pruning and dynamic scheduling of cost-sensitive ensembles. In Proceedings of the National Conference on Artificial Intelligence (pp. 146-151). Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

[36] Fersini, E., Messina, E., & Pozzi, F. A. (2014). Sentiment analysis: Bayesian ensemble learning. Decision support systems, 68, 26-38.

[37] Finlay, S. (2011). Multiple classifier architectures and their application to credit risk assessment. European Journal of Operational Research, 210(2), 368-378.

[38] Fraz, M. M., Remagnino, P., Hoppe, A., Uyyanonvara, B., Rudnicka, A. R., Owen, C. G., & Barman, S. A. (2012). An ensemble classification-based approach applied to retinal blood vessel segmentation. Biomedical Engineering, IEEE Transactions on, 59(9), 2538-2548.

[39] Freeman, C., Kulić, D., & Basir, O. (2015). An evaluation of classifier-specific filter measure performance for feature selection. Pattern Recognition, 48(5), 1812-1826.

[40] Freund, Y., Schapire, R., & Abe, N. (1999). A short introduction to boosting. Journal-Japanese Society For Artificial Intelligence, 14(771-780), 1612.

[41] Friedman, J. H. (1991). Multivariate adaptive regression splines. The annals of statistics, 1-67.

[42] Fu, B., Wang, Z., Pan, R., Xu, G., & Dolog, P. (2013, January). An integrated pruning criterion for ensemble learning based on classification accuracy and diversity. In 7th International Conference on Knowledge Management in Organizations: Service and Cloud Computing (pp. 47-58). Springer Berlin Heidelberg.

[43] Gaber, M. M., & Bader-El-Den, M. B. (2012). Optimisation of Ensemble Classifiers using Genetic Algorithm. In KES (pp. 39-48).

[44] Gamberger, D. (2011) Otkrivanje znanja dubinskom analizom podataka. Institut Ruđer Bošković, dostupno na: http://lis.irb.hr/Prirucnik/prirucnik-otkrivanje-znanja.pdf

[45] García, S., Luengo, J., & Herrera, F. (2015). Data preprocessing in data mining. New York: Springer.

[46] Gashler, M., Giraud-Carrier, C., & Martinez, T. (2008, December). Decision tree ensemble: Small heterogeneous is better than large homogeneous. In Machine Learning and Applications, 2008. ICMLA'08. Seventh International Conference on (pp. 900-905). IEEE.

[47] Ghodselahi, A. (2011). A hybrid support vector machine ensemble model for credit scoring. International Journal of Computer Applications, 17(5), 1-5.

[48] Gomez, J. C., Boiy, E., & Moens, M. F. (2012). Highly discriminative statistical features for email classification. Knowledge and information systems, 31(1), 23-53.

[49] Guyon, I., Gunn, S., Nikravesh, M., & Zadeh, L. A. (Eds.). (2008). Feature extraction: foundations and applications (Vol. 207). Springer.

[50] Han, L., Han, L., & Zhao, H. (2013). Orthogonal support vector machine for credit scoring. Engineering Applications of Artificial Intelligence, 26(2), 848-862.

[51] Harris, T. (2013). Quantitative credit risk assessment using support vector machines: Broad versus Narrow default definitions. Expert Systems with Applications, 40(11), 4404-4413.

[52] Hens, A. B., & Tiwari, M. K. (2012). Computational time reduction for credit scoring: An integrated approach based on support vector machine and stratified sampling method. Expert Systems with Applications, 39(8), 6774-6781.

[53] Huang, Z., Chen, H., Hsu, C. J., Chen, W. H., & Wu, S. (2004). Credit rating analysis with support vector machines and neural networks: a market comparative study. Decision support systems, 37(4), 543-558.

[54] Ibtissem, B., & Bouri, A. (2013). Credit risk management in microfinance: The conceptual framework. ACRN Journal of Finance and Risk Perspectives, 2(1), 9-24.

[55] Kempa, O., Lasota, T., Telec, Z., & Trawiński, B. (2011). Investigation of bagging ensembles of genetic neural networks and fuzzy systems for real estate appraisal. In Intelligent Information and Database Systems (pp. 323-332). Springer Berlin Heidelberg.

[56] Khashman, A. (2010). Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes. Expert Systems with Applications, 37(9), 6233-6239.

[57] Kim, M. J., & Kang, D. K. (2010). Ensemble with neural networks for bankruptcy prediction. Expert Systems with Applications, 37(4), 3373-3379.

[58] Kononenko, I., & Kukar, M. (2007). Machine learning and data mining. Elsevier.

[59] Kumar, V., & Minz, S. (2014). Feature Selection. SmartCR, 4(3), 211-229.

[60] Kuncheva, L. I. (2004). Combining pattern classifiers: methods and algorithms. John Wiley & Sons.

[61] Kuncheva, L. I., & Kountchev, R. K. (2002). Generating classifier outputs of fixed accuracy and diversity. Pattern recognition letters, 23(5), 593-600.

[62] Kuncheva, L. I., & Whitaker, C. J. (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Machine learning, 51(2), 181-207.

[63] Kuncheva, L. I., Whitaker, C. J., Shipp, C. A., & Duin, R. P. (2003). Limits on the majority vote accuracy in classifier fusion. Pattern Analysis & Applications, 6(1), 22-31.

[64] Lacy, S. E., Lones, M. A., & Smith, S. L. (2015). Forming Classifier Ensembles with Multimodal Evolutionary Algorithms. In Proc. of the 2015 IEEE Congress on Evolutionary Computation. IEEE.

[65] Lazar, C., Taminau, J., Meganck, S., Steenhoff, D., Coletta, A., Molter, C., ... & Nowe, A. (2012). A survey on filter techniques for feature selection in gene expression microarray analysis. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 9(4), 1106-1119.

[66] Lazarevic, A., & Obradovic, Z. (2001). Effective pruning of neural network classifier ensembles. In Neural Networks, 2001. Proceedings. IJCNN'01. International Joint Conference on (Vol. 2, pp. 796-801). IEEE.

[67] Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: A ten-year update. European Journal of Operational Research.

[68] Li, L., Hu, Q., Wu, X., & Yu, D. (2014). Exploration of classification confidence in ensemble learning. Pattern Recognition, 47(9), 3120-3131.

[69] Li, N., Yu, Y., & Zhou, Z. H. (2012). Diversity regularized ensemble pruning. In Machine Learning and Knowledge Discovery in Databases (pp. 330-345). Springer Berlin Heidelberg.

[70] Lin, C., Chen, W., Qiu, C., Wu, Y., Krishnan, S., & Zou, Q. (2014). LibD3C: ensemble classifiers with a clustering and dynamic selection strategy. Neurocomputing, 123, 424-435.

[71] Liu, H., & Motoda, H. (2012). Feature selection for knowledge discovery and data mining (Vol. 454). Springer Science & Business Media.

[72] Liu, H., Hussain, F., Tan, C. L., & Dash, M. (2002). Discretization: An enabling technique. Data mining and knowledge discovery, 6(4), 393-423.

[73] Liu, Z., Dai, Q., & Liu, N. (2014). Ensemble selection by GRASP. Applied intelligence, 41(1), 128-144.

[74] Lozano, M., & García-Martínez, C. (2010). Hybrid metaheuristics with evolutionary algorithms specializing in intensification and diversification: Overview and progress report. Computers & Operations Research, 37(3), 481-497.

[75] Ma, Z., Dai, Q., & Liu, N. (2015). Several novel evaluation measures for rank-based ensemble pruning with applications to time series prediction. Expert Systems with Applications, 42(1), 280-292.

[76] Maclin, R., & Opitz, D. (2011). Popular ensemble methods: An empirical study. arXiv preprint arXiv:1106.0257.

[77] Malhotra, R., & Malhotra, D. K. (2003). Evaluating consumer loans using neural networks. Omega, 31(2), 83–96.

[78] Markatopoulou, F., Tsoumakas, G., & Vlahavas, I. (2015). Dynamic ensemble pruning based on multi-label classification. Neurocomputing, 150, 501-512.

[79] Marqués, A. I., García, V., & Sánchez, J. S. (2012). Two-level classifier ensembles for credit risk assessment. Expert Systems with Applications, 39(12), 10916-10922.

[80] Marqués, A. I., García, V., & Sánchez, J. S. (2012a). Exploring the behaviour of base classifiers in credit scoring ensembles. Expert Systems with Applications, 39(11), 10244-10250.

[81] Martınez-Munoz, G., & Suárez, A. (2004). Aggregation ordering in bagging. In Proc. of the IASTED International Conference on Artificial Intelligence and Applications (pp. 258-263).

[82] Martinez-Muoz, G., Hernández-Lobato, D., & Suarez, A. (2009). An analysis of ensemble pruning techniques based on ordered aggregation. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 31(2), 245-259.

[83] Mazurowski, M. A., Habas, P. A., Zurada, J. M., Lo, J. Y., Baker, J. A., & Tourassi, G. D. (2008). Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. Neural networks, 21(2), 427-436.

[84] Michalewicz, Z. (1998). Genetic algorithms + data structures = evolution programs. Springer.

[85] Miglani, N., & Rana, P. (2011). Ranking of Software Reliability Growth Models using Greedy Approach. Global Journal of Business Management and Information Technology, 1(11).

[86] Mitchell, M. (1996). An introduction to genetic algorithms. Cambridge, MA: MIT Press/Addison-Wesley.

[87] Mukherjee, S., & Sharma, N. (2012). Intrusion detection using naive Bayes classifier with feature reduction. Procedia Technology, 4, 119-128.

[88] Murphy, K. P. (2012). Machine learning: a probabilistic perspective. MIT press.

[89] Niven E.B., Deutsch C.V. (2012). Calculating a robust correlation coefficient and quantifying its uncertainty. Computers & Geosciences 40, 1–9.

[90] Opitz, D. W. (1999, July). Feature selection for ensembles. In AAAI/IAAI (pp. 379-384).

[91] Oreski, S. (2014). Hybrid Techniques of Combinatorial Optimization with Application to Retail Credit Risk Assessment. ARTIFICIAL INTELLIGENCE, 1(1).

[92] Oreski, S. (2014). Hybrid techniques of combinatorial optimization based on genetic algorithms with application to feature selection in retail credit risk assessment (Doctoral dissertation, Fakultet organizacije i informatike, Sveučilište u Zagrebu).

[93] Oreski, S., & Oreski, G. (2014). Genetic algorithm-based heuristic for feature selection in credit risk assessment. Expert systems with applications, 41(4), 2052-2064.

[94] Oreski, S., Oreski, D., & Oreski, G. (2012). Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment. Expert systems with applications, 39(16), 12605-12617.

[95] Paleologo, G., Elisseeff, A., & Antonini, G. (2010). Subagging for credit scoring models. European Journal of Operational Research, 201(2), 490-499.

[96] Pang-Ning, T., Steinbach, M., & Kumar, V. (2006). Introduction to data mining. In Library of Congress (p. 74).

[97] Partalas, I., Tsoumakas, G., & Vlahavas, I. (2010). An ensemble uncertainty aware measure for directed hill climbing ensemble pruning. Machine Learning, 81(3), 257-282.

[98] Partalas, I., Tsoumakas, G., & Vlahavas, I. (2012). A study on greedy algorithms for ensemble pruning. Technical Report TR-LPIS-360-12, Department of Informatics, Aristotle University of Thessaloniki, Greece.

[99] Peng, Y., Wang, G., Kou, G., & Shi, Y. (2011). An empirical study of classification algorithm evaluation for financial risk prediction. Applied Soft Computing, 11(2), 2906-2915.

[100] Ping, Y., & Yongheng, L. (2011). Neighborhood rough set and SVM based hybrid credit scoring classifier. Expert Systems with Applications, 38(9), 11300-11304.

[101] Polikar, R. (2012). Ensemble learning. In Ensemble Machine Learning (pp. 1-34). Springer US.

[102] Prodromidis, A. L., & Stolfo, S. J. (2001). Cost complexity-based pruning of ensemble classifiers. Knowledge and Information Systems, 3(4), 449-469.

[103] Quinlan, J. R. (2014). C4. 5: programs for machine learning. Elsevier.

[104] Ranawana, R., & Palade, V. (2006). Multi-classifier systems: Review and a roadmap for developers. International Journal of Hybrid Intelligent Systems, 3(1), 35-61.

[105] Robnik-Šikonja, M., & Kononenko, I. (2003). Theoretical and empirical analysis of ReliefF and RReliefF. Machine learning, 53(1-2), 23-69.

[106] Saeys, Y., Abeel, T., & Van de Peer, Y. (2008). Robust feature selection using ensemble feature selection techniques. In Machine learning and knowledge discovery in databases (pp. 313-325). Springer Berlin Heidelberg.

[107] Salama, K. M., & Freitas, A. A. (2012). ABC-Miner: an ant-based Bayesian classification algorithm. In Swarm Intelligence (pp. 13-24). Springer Berlin Heidelberg.

[108] Shen, Q., Diao, R., & Su, P. (2012). Feature Selection Ensemble. Turing-100, 10, 289-306.

[109] Soni, A., & Shavlik, J. (2011, August). Probabilistic ensembles for improved inference in protein-structure determination. In Proceedings of the 2nd ACM Conference on Bioinformatics, Computational Biology and Biomedicine (pp. 264-273). ACM.

[110] Srivastava, R. (Ed.). (2013). Research Developments in Computer Vision and Image Processing: Methodologies and Applications: Methodologies and Applications. IGI Global.

[111] Sun, Y., & Li, J. (2006, June). Iterative RELIEF for feature weighting. In Proceedings of the 23rd international conference on Machine learning (pp. 913-920). ACM.

[112] Sun, Y., Todorovic, S., & Goodison, S. (2008, April). A Feature Selection Algorithm Capable of Handling Extremely Large Data Dimensionality. In SDM (pp. 530-540).

[113] Surowiecki, J. (2004). The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business. Economies, Societies and Nations.

[114] Šušteršič, M., Mramor, D., & Zupan, J. (2009). Consumer credit scoring models with limited data. Expert Systems with Applications, 36(3), 4736-4744.

[115] Tang, E. K., Suganthan, P. N., & Yao, X. (2006). An analysis of diversity measures. Machine Learning, 65(1), 247-271.

[116] Thomas, L. C. (2000). A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. International journal of forecasting, 16(2), 149-172.

[117] Tsai, C. F., & Wu, J. W. (2008). Using neural network ensembles for bankruptcy prediction and credit scoring. Expert Systems with Applications, 34(4), 2639-2649.

[118] Tsoumakas, G., Partalas, I., & Vlahavas, I. (2008, July). A taxonomy and short review of ensemble selection. In ECAI 2008, workshop on supervised and unsupervised ensemble methods and their applications (pp. 41-46).

[119] Tsoumakas, G., Partalas, I., & Vlahavas, I. (2009). An ensemble pruning primer. In Applications of supervised and unsupervised ensemble methods (pp. 1-13). Springer Berlin Heidelberg.

[120] Tumer, K., & Ghosh, J. (1996). Error correlation and error reduction in ensemble classifiers. Connection science, 8(3-4), 385-404.

[121] Twala, B. (2010). Multiple classifier application to credit risk assessment. Expert Systems with Applications, 37(4), 3326-3336.

[122] Vainer, I., Kraus, S., Kaminka, G. A., & Slovin, H. (2011). Obtaining scalable and accurate classification in large-scale spatio-temporal domains. Knowledge and information systems, 29(3), 527-564.

[123] Vořechovský M. (2012). Correlation control in small sample Monte Carlo type simulations II: Analysis of estimation formulas, random correlation and perfect uncorrelatedness. Probabilistic Engineering Mechanics 29, 105–120.

[124] Wang, G., & Ma, J. (2011). Study of corporate credit risk prediction based on integrating boosting and random subspace. Expert Systems with Applications, 38(11), 13871-13878.

[125] Wang, G., & Ma, J. (2012). A hybrid ensemble approach for enterprise credit risk assessment based on Support Vector Machine. Expert Systems with Applications, 39(5), 5325-5331.

[126] Wang, G., Hao, J., Ma, J., & Jiang, H. (2011). A comparative assessment of ensemble learning for credit scoring. Expert systems with applications, 38(1), 223-230.

[127] Wang, S., & Yao, X. (2009, March). Diversity analysis on imbalanced data sets by using ensemble models. In Computational Intelligence and Data Mining, 2009. CIDM'09. IEEE Symposium on (pp. 324-331). IEEE.

[128] Wang, S., & Yao, X. (2013). Relationships between diversity of classification ensembles and single-class performance measures. Knowledge and Data Engineering, IEEE Transactions on, 25(1), 206-219.

[129] West, D. (2000). Neural network credit scoring models. Computers & Operations Research, 27(11), 1131-1152.

[130] West, D., Dellana, S., & Qian, J. (2005). Neural network ensemble strategies for financial decision applications. Computers & operations research, 32(10), 2543-2559.

[131] Xia, J., Du, P., He, X., & Chanussot, J. (2014). Hyperspectral remote sensing image classification based on rotation forest. Geoscience and Remote Sensing Letters, IEEE, 11(1), 239-243.

[132] Xin-She Yang (2011) Metaheuristic Optimization. Scholarpedia, 6(8):11472.

[133] Xu, J., & Gray, J. B. (2013). Quadratic programming algorithms for ensemble models. Wiley Interdisciplinary Reviews: Computational Statistics, 5(1), 41-47.

[134] Yang, J., Zeng, X., Zhong, S., & Wu, S. (2013). Effective neural network ensemble approach for improving generalization performance. Neural Networks and Learning Systems, IEEE Transactions on, 24(6), 878-887.

[135] Yap, B. W., Ong, S. H., & Husain, N. H. M. (2011). Using data mining to improve assessment of credit worthiness via credit scoring models. Expert Systems with Applications, 38(10), 13274-13283.

[136] Yu, L., & Liu, H. (2003, August). Efficiently handling feature redundancy in high-dimensional data. In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 685-690). ACM.

[137] Yu, L., & Liu, H. (2004, August). Redundancy based feature selection for microarray data. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 737-742). ACM.

[138] Yu, L., Wang, S., & Lai, K. K. (2008). Credit risk assessment with a multistage neural network ensemble learning approach. Expert Systems with Applications, 34(2), 1434-1444.

[139] Zeng, X., Wong, D. F., & Chao, L. S. (2014). Constructing Better Classifier Ensemble Based on Weighted Accuracy and Diversity Measure. The Scientific World Journal, 2014.

[140] Zhang, C., & Ma, Y. (2012). Ensemble Machine Learning. Springer.

[141] Zhang, D., Huang, H., Chen, Q., & Jiang, Y. (2007). A comparison study of credit scoring models.

[142] Zhang, J., & Chau, K. W. (2009). Multilayer Ensemble Pruning via Novel Multi-sub-swarm Particle Swarm Optimization. J. UCS, 15(4), 840-858.

[143] Zhang, M. L., & Zhou, Z. H. (2013). Exploiting unlabeled data to enhance ensemble diversity. Data Mining and Knowledge Discovery, 26(1), 98-129.

[144] Zhang, Y., Burer, S., & Street, W. N. (2006). Ensemble pruning via semi-definite programming. The Journal of Machine Learning Research, 7, 1315-1338.

[145] Zhang, Y., Ding, C., & Li, T. (2008). Gene selection algorithm by combining reliefF and mRMR. BMC genomics, 9(Suppl 2), S27.

[146] Zhou, X., Jiang, W., Shi, Y., & Tian, Y. (2011). Credit risk evaluation with kernel-based affine subspace nearest points learning method. Expert Systems with Applications, 38(4), 4272-4279.

[147] Zhou, Z. H. (2012). Ensemble methods: foundations and algorithms. CRC Press.

[148] Zhou, Z. H., Wu, J., & Tang, W. (2002). Ensembling neural networks: many could be better than all. Artificial intelligence, 137(1), 239-263.

[149] Zhu X., Li J., Wu D., Wang H., & Liang C. (2013). "Balancing accuracy, complexity and interpretability in consumer credit decision making: A c-topsis classification approach," Knowledge-Based Systems, vol. 52, pp. 258–267.

[150] Zhu, W., & Lin, Y. (2013). Using gini-index for feature weighting in text categorization. Journal of Computational Information Systems, 9(14), 5819-5826.

[151] Zhu, X., Li, J., Wu, D., Wang, H., & Liang, C. (2013). Balancing accuracy, complexity and interpretability in consumer credit decision making: A C-TOPSIS classification approach. Knowledge-Based Systems, 52, 258-267.

[152] Zurada, J. (2013). An Investigation Of The Effect Of Variable Reduction On Classification Accuracy Rates Of Consumer Loans. Review of Business Information Systems (RBIS), 17(4), 135-140.

PRILOZI

Dodatak A

Tablica A.1 Popis atributa s deskriptivnom statistikom za hrvatski skup podataka

Attributi	Tip	Opis	Statistika	Raspon
att1	integer	*Id - redni broj zapisa*		[1; 1000]
att2	integer	*Starost*	avg = 46.198 +/- 14.097	[20; 80]
att3	integer	*Spol*	avg = 0.506 +/- 0.500	[0; 1]
att4	integer	*Pošta (selo,grad)*	avg = 0.581 +/- 0.494	[0; 1]
att5	integer	*Telefon (0-nema, 1-ima)*	avg = 0.906 +/- 0.292	[0; 1]
att6	integer	*Vrijeme na sadašnjoj adresi (izraženo u godinama)*	avg = 2.722 +/- 1.594	[0; 6]
att7	integer	*Klijent banke (izraženo u godinama)*	avg = 14.039 +/- 8.347	[1; 50]
att8	integer	*Starost tekućeg računa(izraženo u godinama)*	avg = 9.825 +/- 7.566	[1; 34]
att9	integer	*Mjesec izdavanja kredita (u toku godine)*	avg = 6.400 +/- 3.265	[1; 12]
att10	integer	*Rok povrata kredita (u mjesecima)*	avg = 39.378 +/- 19.562	[11; 61]
att11	integer	*Namjena kredita (svaka vrsta kredita ima svoj broj)*	avg = 0.874 +/- 0.508	[0; 5]
att12	integer	*Iznos kredita (u HRK)*	avg = 25057.85 +/- 18075.76	[2000; 100000]
att13	integer	*Ukupne uplate na tekući račun*	avg = 4373.09 +/- 3351.98	[1; 32522]
att14	real	*Gotovinske uplate / Ukupne uplate*	avg = 0.071 +/- 0.158	[0; 1]
att15	real	*Redovne uplate (plaće, mirovine) Ukupne uplate*	avg = 0.884 +/- 0.206	[0; 1]
att16	real	*Dozvoljen minus / Ukupne uplate*	avg = 1.754 +/- 1.272	[0; 5]
att17	integer	*Ukupne isplate sa tekućeg računa*	avg = 4665.159 +/- 4967.562	[1; 117664]
att18	real	*Ukupne uplate / Ukupne isplate*	avg = 0.991 +/- 0.238	[0.010 ; 3]
att19	real	*EFTPOS isplate / Ukupne isplate*	avg = 0.164 +/- 0.170	[0; 0.950]
att20	real	*ATM isplate / Ukupne isplate*	avg = 0.398 +/- 0.303	[0; 1]
att21	real	*Samoposlužne isplate / Ukupne isplate*	avg = 0.563 +/- 0.322	[0; 1]
att22	real	*Redovne isplate s računa (rate,trajni nalozi) / Redovne uplate (plaće, mirovine)*	avg = 1.680 +/- 1.235	[0; 5]
att23	real	*Odnos isplata u tromjesečju odobravanja kredita / isto tromjesečje godinu ranije*	avg = 0.296 +/- 0.323	[0.030 ; 2]
att24	real	*Odnos uplata u tromjesečju odobravanja kredita / isto*	avg = 1.388 +/- 1.000	[0.010 ; 5]

		tromjesečje godinu ranije		
att25	integer	*Dozvoljen minus*	avg = 7921 +/- 7283.530	[0; 29000]
att26	integer	*Oročeni depoziti (saldo na dan odobravanja kredita)*	avg = 1248.602 +/- 7892.941	[0; 107590]
att27	integer	*Saldo svih računa na dan odobravanja kredita (Potr.-Dug.)*	avg = -5353.451 +/- 7085.791	[-30632; 31338]
att28	real	*Saldo tekuće računa / Dozvoljen minus*	avg = -0.448 +/- 1.473	[-9; 9]
att29	integer	*Koliko puta je klijent imao negativnu kamatu*	avg = 2.683 +/- 3.237	[0; 12]
att30	real	*Pozitivna kamata/ Negativna kamata*	avg = 2.157 +/- 3.092	[0; 10]
att31	integer	*Iznos negativne kamate*	avg = 7.514 +/- 25.154	[0; 369]
att32	integer	*Odobrenih kredita većeg iznosa od trenutnog*	avg = 0.212 +/- 0.491	[0; 3]
att33	integer	*Odobrenih kredita manjeg ili jednako iznosa od trenutnog*	avg = 0.504 +/- 0.750	[0; 4]
att34	integer	*U koliko je postojećih kredita klijent bio u zakašnjenju većem od 90 dana*	avg = 0.031 +/- 0.173	[0; 1]
att35	binominal	*Urednost klijenta (0-neuredan, 1–uredan)*	mode = 1 (750), least = 0 (250)	0 (250), 1 (750)

Dodatak B

Slika B.1 Implementacija algoritma za generiranje skupa podataka u programu MatLab

```
function [ odlukeKlasifikatora ] =
generirajOdlukeKlasifikatora(brojPrimjera, brojKlasifikatora)
%funkcija generira odluka klasifikatora na temelju točnosti i
raznolikosti
%qVrijednost definira mjeru raznolikosti
qVrijednost=0.7;
%incijalizacija vektora
p = [];
p1 = [];
p2 = [];
matrix=[];
% stvara se vektor p vrijednosti pojedinih klasifikatora
% točnost svih klasifikatora postavljena na istu vrijednost
for u=1:brojKlasifikatora
     p=horzcat(p,0.65);
end
%izračun matrica p1 i p2
for k=1:brojKlasifikatora
      for z=1:brojKlasifikatora
          if k ~= z
                  pi=p(k);
                  pk=p(z);
              if qVrijednost~=0
                  Discr = (1-qVrijednost+2*qVrijednost*(pi-pk))^2-
8*qVrijednost*(1-pi)*pk*(qVrijednost-1);
                  %vrijednosti p2 modificirane
                  p2(k,z)=(-(1-qVrijednost+2*qVrijednost*(pi-
```

```
pk))+Discr^0.5)/(4*qVrijednost*(1-
pi))*(3/brojKlasifikatora)^((abs(qVrijednost)+qVrijednost)/2);
                    p1(k,z)= (1-p2(k,z)-pk/pi+p2(k,z)/pi);
                else
                    p2(k,z)=pi;
                    p1(k,z)=1-pi;
                end
            else
                p2(k,z)=1;
                p1(k,z)=1;
            end
        end
end
%generiranje primjera prema vrijednosti iz matrica
for j=1:brojPrimjera
      randOdabir=randperm(brojKlasifikatora);
      operation = horzcat(zeros(1,100-round(p(randOdabir(1))*100)),
ones(1,round(p(randOdabir(1))*100)));
      operation = operation(randperm(100));
      matrix(j,randOdabir(1))= operation(1);
      for t=2:brojKlasifikatora
          randomBroj=rand(1);
          if matrix(j,randOdabir(t-1)) == 1
              if p1(randOdabir(t-1),randOdabir(t))<randomBroj
                  matrix(j,randOdabir(t))=1;
              else
                  matrix(j,randOdabir(t))=0;
              end
          else
              if p2(randOdabir(t-1),randOdabir(t))<randomBroj
                  matrix(j,randOdabir(t))=0;
              else
                  matrix(j,randOdabir(t))=1;
              end
          end
      end
end
%odluke
odlukeKlasifikatora=matrix;
```

Životopis

Goran Oreški je rođen 19. listopada 1987. godine u Bosanskoj Dubici, Republika Bosna i Hercegovina. Završio je osnovnu i srednju školu u Karlovcu.

Upisuje Fakultet organizacije i informatike 2005. godine. Tijekom studija je studirao na Karl-Franzes Sveučilištu u Grazu, Austrija u okviru međunarodne razmjene studenata. Diplomirao je 2010. godine na temu „*Skladište podataka kao podloga poslovnom odlučivanju u bankarstvu*".

Po završetku fakulteta 2010. godine otvara vlastitu tvrtku GO Studio d.o.o. za računalno programiranje, te se iste godine zapošljava u Karlovačkoj banci d.d. kao programer. U okviru poslova u banci sudjelovao je u: projektiranju prvog skladišta podataka banke, izradi nadzornih izvještaja za izvješćivanje HNB-a te u izradi poslovnih aplikacija za potrebe Banke.

Godine 2012. prelazi u informatičku tvrtku LC d.o.o. Zagreb, na radno mjesto projektanta informacijski sustava gdje je bio zadužen za projektiranje aplikacija unutar SPI sustava, integraciju novih funkcionalnosti u SPI sustav, razvoj informacijskog sustava (aplikacije i modela podataka) i optimizaciju postojećih modela podataka.

Krajem 2013. godine zapošljava se u vlastitoj tvrtci GO Studio d.o.o. gdje do danas radi.

Već tijekom rada u banci je uključen u istraživački rad vezano za predviđanje kreditnog rizika u bankama uz pomoć umjetne inteligencije, što je rezultiralo interesom za daljnjim usavršavanjem na području znanstveno istraživačkog rada. S istim ciljem upisuje poslijediplomski studij „Informacijske znanosti" na Fakultetu organizacije i informatike krajem 2013. godine. Kao rezultat istraživanja objavio je u koautorstvu nekoliko znanstvenih radova, od kojih dva u *Expert systems with applications*, prestižnom svjetskom časopisu za umjetnu inteligenciju. Godine 2014. je dobio nagradu za najbolji rad na jubilarnoj 25. CECIIS znanstvenoj konferenciji.

Popis radova

Izvorni znanstveni radovi u CC časopisima

1. Oreski, S., Oreski, D., & Oreski, G. (2012). Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment. Expert systems with applications, 39(16), 12605-12617.

2. Oreski, S., & Oreski, G. (2014). Genetic algorithm-based heuristic for feature selection in credit risk assessment. Expert systems with applications, 41(4), 2052-2064.

Znanstveni radovi u drugim časopisima

1. Oreški, G., & Oreški, S. (2015). Two Stage Comparison of Classifier Performances for Highly Imbalanced Datasets. *Journal of Information and Organizational Sciences*, *39*(2), 209-222.

Znanstveni radovi u zbornicima skupova s međunar.rec.

1. Oreški, G., Oreški, S., „An experimental comparison of classification algorithm performances for highly imbalanced datasets“, 25rd Central European Conference on Information and Intelligent Systems, Varaždin, 2014