UNIVERZITET U NIŠU

ELEKTRONSKI FAKULTET

Ivana P. Marković

IZBOR ATRIBUTA INTEGRACIJOM ZNANJA O DOMENU PRIMENOM METODA ODLUČIVANJA KOD PREDIKTIVNOG MODELOVANJA VREMENSKIH SERIJA NADGLEDANIM MAŠINSKIM UČENJEM

doktorska disertacija

Niš, 2017.

UNIVERSITY OF NIŠ

FACULTY OF ELECTRONIC ENGINEERING

Ivana P. Marković

FEATURE SELECTION BY INTEGRATING DOMAIN-SPECIFIC KNOWLEDGE USING DECISION MAKING METHODS FOR PREDICTIVE MODELING OF TIME-SERIES DATA WITH SUPERVISED MACHINE LEARNING METHODS

Doctoral dissertation

Niš, 2017

Sa iskrenim zadovoljstvom želim da se zahvalim svom mentoru, dr Mileni Stanković, redovnom profesoru Elektronskog fakulteta Univerziteta u Nišu, na svim naučnim savetima i konstruktivnim diskusijama koje su mi pomogle da oblikujem svoj naučni rad. Posebno želim da izrazim svoju zahvalnost na svoj pruženoj podršci, pokazanom razumevanju i strpljenju, koji su mi bili od suštinskog značaja i konačno doveli do izrade ove doktorske disertacije.

Želim da se zahvalim i svojim bliskim saradnicima i kolegama na višegodišnjoj saradnji koja je posredno doprinela i razvoju ove doktorske disertacije.

U Nišu, novembra 2017. godine

Podaci o doktorskoj disertaciji

Mentor:	Dr Milena Stanković, redovni profesor, Univerzitet u Nišu, Elektronski fakultet
Naslov:	Izbor atributa integracijom znanja o domenu primenom metoda odlučivanja kod prediktivnog modelovanja vremenskih serija nadgledanim mašinskim učenjem
Rezime:	Cilj istraživanja predstavljenih u okviru ove doktorske disertacije jeste razvoj metodologije za izbor atributa integracijom znanja o domenu primenom matematičkih metoda odlučivanja, kako bi se poboljšala preciznost metoda nadgledanog mašinskog učenja kod prediktivnog modelovanja vremenskih serija. Za integraciju znanja o domenu korišćena je višekriterijumska metoda odlučivanja, odnosno Analitički hijerarhijski proces, koji se pokazao uspešnim u mnogim eksperimentima koji su do sada obavljani. Ovaj pristup je izabran zbog sposobnosti da vrši selekciju između skupa faktora na osnovu njihove relevantnosti čak i kod međusobno suprotstavljenih kriterijuma. Kod predviđanja kretanja vremenskih serija istražene su mogućnosti integracije faktora korisnosti atributa kod metoda podržavajućih vektora mašinskog učenja kako bi se poboljšale njihove predikcione performanse. Predložena metodologija primenjena je kao metod izbora atributa kod prediktivnog modelovanja kretanja finansijskih vremenskih serija. Za razliku od većine postojećih pristupa, kod kojih se metodi izbora atributa zasnivaju na kvantitativnoj analizi ulaznih vrednosti, predložena metodologija vrši kvalitativnu procenu atributa u odnosu na domen predviđanja i predstavlja način da se integriše *a priori* znanje o domenu predviđanja.
Naučna oblast:	Elektrotehničko i računarsko inženjerstvo (Računarstvo i informatika)
Naučna disciplina:	Data mining – Mašinsko učenje
Ključne reči:	izbor atributa, težinska kernel funkcija, prediktivno modelovanje, vremenske serije
UDK:	(004.431.2:532.5):004.414.22
CERIF klasifikacija:	T120 Sistemski inženjering, računarska tehnologija
Tip licence Kreativne zajednice:	**CC BY-NC-ND**

Data on Doctoral Dissertation

Doctoral Supervisor:	Prof. Milena Stanković, PhD, Full Professor Faculty of Electronic Engineering, University of Niš
Title:	Feature selection by integrating domain-specific knowledge using decision making methods for predictive modeling of time-series data with supervised machine learning methods
Abstract:	The aim of the research presented within this doctoral dissertation is to develop a feature selection methodology through integrating domain-specific knowledge by applying mathematical methods of decision-making, to improve the feature selection process and the precision of supervised machine learning methods for predictive modeling of time series. To integrate domain-specific knowledge, a multi-criteria decision making method is used, i.e. an analytical hierarchical process proven to be successful in numerous studies carried out to date. This approach was selected because it allows the selection of a set of factors based on their relevance, even in the case of mutually opposite criteria. In predicting the movement of time series, the possibility of integrating feature relevance into support vector machines to improve their prediction accuracy was studied. The proposed methodology was applied as a feature-selection method for the predictive modelling of movement of financial time series. Unlike existing approaches, where the feature selection method is based on a quantitative analysis of the input values, the proposed methodology carries out a qualitative evaluation of the attributes in relation to the prediction domain and represents a means of integrating a priori knowledge of the prediction domain.
Scientific Field:	Electronic and Computer Engineering (Computer science)
Scientific Discipline:	Data mining – Machine learning
Key Words:	Feature selection, Weighted kernel function, Predictive modeling, Time series
UDC:	(004.431.2:532.5):004.414.22
CERIF Classification:	T120 System engineering, computer technology
Creative Commons License Type:	**CC BY-NC-ND**

SADRŽAJ

SPISAK SLIKA

SPISAK TABELA

SKRAĆENICE

AHP - *Analytic hierarchy process*

ANNs - *Artificial neural networks*

ANP - *Analytical network process*

ARCH - *Autoregressive conditional heteroskedasticity*

ARIMA - *Autoregressive integrated moving average*

CART - *Classification and regression tree*

EMR - *Empirical risk minimization*

GARCH - *Generalized autoregressive conditional heteroskedasticity*

GAs - *Genetic algorithms*

ICA - *Independent component analysis*

ID3 - *Iterative Dichotomiser 3*

IID - *Independent and identically distributed*

KB - *Kernel based estimator*

k-NN - *k nearest neighbours*

LS-SVMs - *Least squares support vector machines*

MI - *Mutual information*

MLP - *Multilayer perceptron*

PCA - *Principal component analysis*

PDF - *Probability density function*

RBF - *Radial basis function*

SRM - *Structural risk minimization*

SVMs - *Support vector machines*

POGLAVLJE 1

UVOD

Svedoci smo vremena u kome su podaci iz različitih sfera ljudskog delovanja opšte dostupni, pri čemu je i obim podataka sve veći. Paradigma našeg vremena odslikava se u konceptima *Big data* i *Internet of Things*. Shodno tome ne iznenađuje sve veća potreba da se dostupni podaci analiziraju i pretvore u korisne informacije o procesima koji ih generišu. Mašinsko učenje je aktuelna oblast računarstva koja se vrlo intezivno razvoja upravo usled rastućih potreba da se iz podataka dobiju korisne informacije.

Prema definiciji, nadgledano mašinsko učenje predstavlja sposobnost algoritma da vrši generalizaciju na osnovu prethodno naučenih veza između podataka. Prediktivno modelovanje sa druge strane „podrazumeva korišćenje statističkih, data mining i tehnika mašinskog učenja u cilju prepoznavanja struktura i šablona u podacima kako bi se predvideo trend promene podataka u budućnosti“ [81]. Osnovno pitanje u oblasti prediktivnog modelovanja je kako poboljšati kvalitet predikcija.

Oblasti primene metoda mašinskog učenja su raznovrsne: u medicini, bioinformatici, zarad unapređenja poslovnih performansi preduzeća. Modelovanje vremenskih serija je značajno u neuroinformatici kod analize biosignala [164], kao i u meteorologiji na primer, gde se za hidrološke prognoze koriste različite metode prediktivnog modeliranja hidroloških veličina [29]. Primenom koncepta vremenskih serija može se predviđati i koliko će puta naučni rad u budućnosti biti citiran [38] ili modelovati kolika će biti posećenost nekog sajta [89].

Predviđanje kao tehnika ima posebnu prednost kada se dobijene informacije primenjuju tokom donošenja poslovnih odluka na svim nivoima upravljanja. Na primer: na nivou pružanja marketinških usluga gde se vrši segmentacija kupaca prema preferencijama ka određenim proizvodima ili u bankama gde se vrši klasifikacija kreditnih zahteva klijenata.

Glavni motiv za primenu predikcionih metoda u poslovanju su svakako smanjenje rizika poslovanja i ostvarivanje profita.

Predikcioni modeli sastoje se od dva dela: izbora atributa koji će se koristiti za obuku modela i izbora predikcionog metoda i njegovog algoritma za obuku. Poznato je da kreiranje kompleksnijih algoritama ne mora nužno dovesti do poboljšanja performansi prediktora, dok bolja reprezentacija problema predviđanja uglavnom vodi do značajnih poboljšanja [81].

Atribut (engl. *feature*) – je informacija koja je potencijalno korisna za predikciju. Sa druge strane prema definiciji „Inženjering atributa je proces transformacije neobrađenih podataka u atribute koji bolje predstavljaju problem iz domena predviđanja prediktivnim modelima i vode do unapređenja kvaliteta predikcije“ [81].

Određivanje dovoljnog i neophodnog skupa atributa je od suštinske važnosti za obučavanje dobrog predikcionog modela tako da je u dosadašnjim radovima većina rešenja upravo bila orijentisana ka razvoju algoritama i kriterijuma za izbor atributa. Ako je broj atributa nedovoljan, preciznost prediktora može biti nezadovoljavajuća i suprotno tome ako ima previše atributa, može se pojaviti redundantnost u podacima, a modeli mogu imati loše sposobnosti generalizacije. U slučaju velikog broja ulaznih atributa d, određivanje optimalnog podskupa atributa direktnom evaluacijom svih 2^d podskupova za dati skup atributa je kao i mnogi problemi u vezi sa izborom atributa problem iz klase *NP – hard* [3] i [63].

U većini radova u kojima se obrađuje problem izbora atributa fokus je na analizi numeričkih vrednosti atributa i njihovih međusobnih relacija. Tek nedavno počelo se sa istraživanjima u pravcu kvalitativne analize atributa specifičnom za domen predviđanja. Aktuelan problem kod izbora atributa je i činjenica da većina algoritama mašinskog učenja pretpostavlja da svi ulazni atributi imaju istu relevantnost. Međutim, zajednička osobina podataka iz realnog okruženja je da su određeni atributi u većoj ili manjoj meri relevantni za posmatrani domen.

U cilju prevazilaženja ovih problema postaje značajno dodeljivanje težina atributima kao i adekvatna formalizacija znanja o specifičnostima oblasti primene metoda mašinskog učenja.

Maksimalne performanse se dobijaju unapređenjem oba dela predikcionog modela kako procesa izbora atributa tako i prilagođavanjem same metode učenja. Takvi pristupi su u praksi i najređi zato što inženjering atributa zahteva aktivnosti i znanja usko vezane za domen predikcije, dok su algoritmi mašinskog učenja uglavnom opšte namene [42].

Istraživanja koja su predstavljena u ovoj doktorskoj disertaciji primarno su fokusirana na optimizaciju prediktivnog modelovanja kretanja finansijskih vremenskih serija. Kretanja na finansijskim tržištima su oduvek bila u fokusu stručne i akademske javnosti, zbog mogućnosti ostvarivanja profita investiranjem u finansijske instrumente na tržištu kapitala, pri čemu uspešno predviđanje cena i kretanja finansijskih instrumenata postaje sve značajnije.

Novije studije pokazuju da strategije trgovanja vođene prognozama o pravcu promene cena mogu biti efikasnije i generisati veći prinos u odnosu na tradicionalna predviđanja nivoa cena finansijskih instrumenata.

U ovoj doktorskoj disertaciji razmatrana su oba aspekta predikcionog modela i izbor atributa i prilagođavanje metoda mašinskog učenja sa ciljem povećanja preciznosti.

Predložena metodologija formalizuje znanja specifična za finansijska tržišta na način da ih kroz matematički metod odlučivanja konceptualizuje i integriše u postupak izbora atributa za obuku predikcionog modela. Kao metoda izbora atributa iz ulaznog skupa atributa biće korišćena višekriterijumska analiza metodom korisnosti poznatija kao Analitički hijerarhijski proces (engl. *Analytic hierarchy process* – AHP) [132]. Metod se pokazao uspešnim kod višekriterijumskog odlučivanja zbog svoje sposobnosti da vrši evaluaciju skupa faktora na osnovu njihove relevantnosti čak i kod međusobno suprotstavljenih kriterijuma, a bez prethodnog znanja o strukturi njihovih odnosa. U okviru doktorske disertacije predložena je metodologija za izbor atributa, zasnovana na analizi korisnosti inicijalnog skupa atributa, kao i sami kriterijumi poređenja u skladu sa kojima se vrši rangiranje i selekcija atributa.

Kako bi se postiglo maksimalno poboljšanje na posmatranom domenu, u ovom radu fokus je i na poboljšanju predikcionih sposobnosti predikcionih metoda sa integracijom težina u kernel. Koriste se metoda podržavajućih vektora (engl. *support vector machines* - SVMs) i metod najmanjih kvadrata podržavajućih vektora (engl. *least squares support vector machines* - LS-SVMs), koji su zbog osobine da uvek pronalaze optimalno rešenje, malog broja potrebnih trening parametara i jako dobre sposobnosti generalizacije već našli široku primenu u oblasti predikcije finansijskih vremenskih serija. Za prediktivno modelovanje kretanja finansijskih vremenskih serija, u kombinaciji sa izborom atributa implementirana je binarna klasifikacija.

Evaluacija predložene metodologije obavljena je na modelovanju problema predviđanja kretanja finansijskih vremenskih serija sa tržišta kapitala.

Na kraju uvodnog poglavlja potrebno je istaći i napomene o stilu pisanju i usvojenim nomenklaturama u okviru ove doktorske disertacije. Prema saznanjima autora za pojedine termine koji se koriste u oblastima koje su predmet istraživanja ne postoji koncenzus stručne i

naučne javnosti o adekvatnim ekvivalentima na srpskom jeziku niti prihvaćena standardizacija terminologije. Iz tog razloga prihvaćeni stil pisanja u ovoj doktorskoj disertaciji podrazumeva da su strani uskostručni i univerzalno prepoznatljivi termini pisani u tekstu doktorske disertacije u svojim originalnim stranim nazivima, uz navođenje srpskih ekvivalenata u zagradama na način kako se najčešće mogu naći u domaćoj literaturi. Svrha usvojenih prilagođavanja je da se preduprede sve potencijalne nedoumice koje nameću jezička ograničenja i obezbedi uniformnost u imenovanju.

Disertacija se sastoji od devet struktuiranih celina. Nakon uvodnog poglavlja, u drugom poglavlju detaljno je izložena oblast istraživanja, analiziraju se oblasti primene vremenskih serija kao i prediktivnog modelovanja, dat je prikaz predikcionih modela i izloženi su osnovni pojmovi potrebni za razumevanje svojstva algoritama mašinskog učenja. Prikazana su i ograničenja i problemi u dosadašnjim pristupima kreiranju predikcionih modela.

U trećem poglavlju izložen je princip kernel funkcija i date su teorijske osnove SVM i LS-SVM metoda koji su korišćeni u simulacijama za formiranje predikcionih modela. U ovom poglavlju su posebno analizirane teorijske osnove potrebne za razumevanje i definisanje težinske kernel funkcije.

U četvrtom poglavlju detaljno su analizirani problem i oblast izbora atributa u nadgledanom mašinskom učenju, a zatim je prema najšire prihvaćenoj klasifikaciji izvršena komparativna analiza prikazanih strategija.

U petom poglavlju razmatran je problem reprezentacije znanja kod metoda mašinskog učenja, posebno je razmatran pojam inženjeringa atributa i mogućnosti označavanja vektora i dodeljivanje težina atributima.

U šestom poglavlju dat je prikaz sinergije metoda mašinskog učenja i optimizacionih metoda, zajedno sa pregledom literature koja upućuje na integraciju metoda odlučivanja i algoritama mašinskog učenja. U poslednjem delu istog poglavlja prikazane su teorijske osnove potrebne za razumevanje jednog od najkorišćenijeg metoda višekriterijumskog odlučivanja, Analitičkog hijerarhijskog procesa.

U sedmom poglavlju, kao doprinos, predstavljena je predložena metodologija za izbor podskupa atributa zasnovana na određivanju težina primenom Analitičkog hijerarhijskog procesa.

U osmom poglavlju izvršena je evaluacija rezultata primene predložene metodologije u kombinaciji sa integracijom težina u kernel kod SVM i LS-SVM metoda na različitim skupovima podataka. U okviru osmog poglavlja prikazani su i rezultati komparacije

predložene metodologije sa drugim algoritmima izbora atributa kao i rezultati komparacije predloženog predikcionog modela sa drugim modelima nadgledanog mašinskog učenja. Na kraju osmog poglavlja, razmatrani su uslovi primene predložene metodologije za izbor atributa i diskutovana su uopštenja.

U devetom poglavlju su uz kratak rezime i zaključak izneti i pravci budućih istraživanja.

POGLAVLJE 2

PREDMET ISTRAŽIVANJA

U poglavlju koje sledi izloženi su oblast i predmet istraživanja kroz prikaz aktuelnih trendova u oblasti i uočenih mogućnosti optimizacije pristupa.

2.1. PREDIKTIVNO MODELOVANJE VREMENSKIH SERIJA

Kako je prethodno definisano „Prediktivno modelovanje podrazumeva korišćenje statističkih, data mining i tehnika mašinskog učenja u cilju prepoznavanja struktura i šablona u podacima kako bi se poboljšao kvalitet predikcije“ [81].

U osnovi statistika se koristi za razumevanje procesa koji generišu podatke, sa primarnim ciljem da se izvrši testiranje različitih hipoteza o samim procesima. Data mining tehnike nastoje da pronađu ranije nepoznate i korisne veze među podacima. Nasuprot njima, fokus mašinskog učenje je na tačnom i efikasnom predviđanju nepoznatih podataka odnosno generalizaciji.

Vremenska serija predstavlja sekvencu vrednosti dobijenih merenjima koja se obično izvode u uzastopnim ekvidistantnim vremenskim trenucima. U opštem slučaju, vremenska serija može ispoljavati nelinearnost, nestacionarnost, periodičnost, prisustvo haotične komponente i prisustvo šuma.

Poseban izazov kod modelovanje vremenskih serija predstavljaju vremenske serije nelinearnih dinamičkih sistema, sa značajnom primenom u neuroinformatici u analizi biosignala, kao i u meterologiji na primer, gde se za hidrološke prognoze koriste različite metode prediktivnog modelovanja hidroloških veličina ili modelovanje finansijkih vremenkih serija kod projekcija tokove gotovine.

Dodatni izazov u oblasti predviđanja vremenskih serija predstavljaju predviđanja kretanja vremenskih serija. Dok se kod regresionih predikcija vrši predviđanje buduće

vrednosti posmatrane pojave na primer nivoa finansijskih instrumenata, kod prediktivnog modelovanja koje se predstavlja problemom binarne klasifikacije vrši se predviđanje promena trenda kretanja vrednosti vremenskih serija odnosno da li će u narednom vremenskom trenutku doći do rasta ili do opadanja posmatrane vrednosti.

Problemi koji se javljaju kod predviđanja pravca kretanja ogledaju se u dinamičkoj promeni posmatrane karakteristike koja po pravilu zavisi od više faktora. U oblasti finansija je na primer poznato da se dobijanjem novih informacija ponašanje tržišnih učesnika stalno menja, dok je donošenje odluka o delovanju usko povezano sa iskustvom i intuicijom investitora.

2.2. PREDIKCIONI MODELI

U [7] i [176] istaknuto je da se najšire rasprostranjeni predikcioni modeli sastoje iz dva dela. Prvi deo predstavlja korak izbora atributa za obuku modela, dok drugi predstavlja izbor predikcionog metoda i njegovog algoritma za obuku.

Poznato je i da kreiranje kompleksnijih algoritama ne mora nužno dovesti do poboljšanja performansi prediktora, dok bolja reprezentacija problema predviđanja uglavnom vodi do značajnih poboljšanja [81].

Kako je navedeno u [8], najznačajniji korak u kreiranju pouzdanog predikcionog modela predstavlja selekcija ulaznih atributa za obučavanje modela, pri čemu je od suštinskog značaja izbor metoda za određivanje podskupa skupa ulaznih atributa.

Praktično uspešnost predikcionog modela je u osnovi uslovljena odnosom između korišćene metode predikcije i samog algoritma za izbor atributa.

2.2.1. ATRIBUTI

Atributima se predstavljaju različite osobine koje se prepoznaju u posmatranom domenu, kao i odnosi među tim svojstvima. Atribut prema definiciji iz [56] predstavlja merljivo svojstvo procesa koji se posmatra. Za adekvatnu reprezentaciju modela izazov predstavlja izbor optimalnog skupa atributa.

Posmatrano od strane predikcionog metoda atribut predstavlja informaciju koja je potencijalno korisna za predikciju. Pri čemu po definiciji iz [81] „Inženjering atributa predstavlja proces transformacije neobrađenih podataka u atribute koji bolje predstvaljaju problem iz domena predviđanja prediktivnim modelima i vode do unapređenja kvaliteta predikcije“.

Kod izbora atributa dominantna su dva pristupa: filter metode (engl. *filter*), kod kojih je selekcija atributa nezavisna od korišćenog algoritma učenja, i *wrapper* metode (metode probnog ili prethodnog učenja), kod kojih je proces selekcije atributa zasnovan na evaluaciji doprinosa posmatranog atributa kvalitetu predviđanja. Prema nekim klasifikacijama, treći pravac predstavljaju *embeded* (ugrađeni) metodi koji za redukciju skupa atributa koriste sam algoritam učenja.

U poslednje dve decenije značajna grupa radova razmatrala je problem odabira odgovarajućeg algoritma izbora atributa za određene domene [12], [40], [66], [91] i [96]. Međutim, bez obzira na širok spektar raspoloživih algoritama, izbor uslovno optimalnog algoritma za konkretni problem se pokazao kao izuzetno zahtevan zadatak.

U većini radova u kojima se obrađuje problem izbora atributa fokus je na analizi numeričkih vrednosti atributa i njihovih međusobnih relacija. Tek nedavno počelo se sa istraživanjima u pravcu kvalitativne analize atributa specifičnom za domen predviđanja. Otvoren problem kod izbora atributa je i činjenica da većina algoritama mašinskog učenja predpostavlja da svi ulazni atributi imaju istu relevantnost. Međutim, zajednička osobina podataka iz realnog okruženja je da su određeni aitributi u većoj ili manjoj meri relevantni za posmatrani domen.

2.2.2. NADGLEDANO MAŠINSKO UČENJE

Oblast mašinskog učenja odnosi se na proučavanje algoritama koji su sposobni da u skladu sa „iskustvom“, odnosno na osnovu podataka, a bez eksplicitnog programiranja automatski poboljšaju svoje performanse [135].

Formalna definicija upućuje na sledeća svojstva:

“Za računarski program se može reći da uči iz iskustva E u odnosu na neke vrste zadataka T i merilo performansi P, ako se njegove performanse na zadacima iz T, merene sa P, unapređuju sa iskustvom E” [106].

Metode mašinskog učenja se najčešće dele na metode nadgledanog mašinskog učenja i na metode koje pripadaju grupi metoda nenadgledanog mašinskog učenja. Dodatno se još može govoriti o polunadgledanom učenju (engl. *semi-supervised*) i učenju uslovljavanjem, odnosno učenju uz podsticaje (engl. *reinforcement learning*). U osnovi, većina problema za koje se mašinsko učenje koristi pripada grupi nadgledanog mašinskog učenja [42].

Prema definiciji, nadgledano mašinsko učenje predstavlja sposobnost algoritma da vrši generalizaciju na osnovu prethodno naučenih veza između atributa.

Definišimo trening skup na osnovu parova {ulazna vrednost, ciljna vrednost} u oznaci, $\{x_k, y_k\}, k = 1,...N$, gde N predstavlja broj trening primera odnosno ulazno/ciljnih parova, i označimo sa X prostor ulaznih vrednosti, $x_k \in R^n$, a sa Y prostor ciljnih vrednosti $y_k \in R$. Algoritmi nadgledanog mašinskog učenja imaju za cilj da na osnovu zadatog trening skupa pronađu uslovno optimalnu funkciju predviđanja *h* (engl. *hypothesized function*), $h: X \to Y$, koja za svaku pojedinačnu ulaznu vrednost daje dovoljno dobre aproksimacije ciljne vrednosti. U slučaju klasifikacije prediktivna metoda koristi selektovane atribute i oznake klasa kako bi naučila funkciju predviđanja *h* kojom se atributi preslikavaju u izlazne promenljive [153]. Generalizacija je sposobnost algoritma mašinskog učenja da primenom funkcije predviđanja za ulazne vrednosti kojima je ciljna vrednost nepoznata izvrši tačnu procenu ciljne vrednosti.

Već je u uvodnom poglavlju navedeno da je osnovno pitanje u oblasti mašinskog učenja kako poboljšati kvalitet predikcija. Može se očekivati da se maksimalne performanse predikcionog modela dobijaju kako unapređenjem procesa izbora atributa tako i prilagođavanjem same metode učenja. Kako je prethodno istaknuto, takvi pristupi su u praksi retki uzevši u obzir da proces inženjeringa atributa zahteva aktivnosti i znanje usko vezano za domen predikcije i da su algoritmi mašinskog učenja uglavnom opšte namene [42].

2.3. OBLAST ISTRAŽIVANJA

Predviđanje kao tehnika ima posebnu prednost kada se dobijene informacije primenjuju tokom donošenja poslovnih odluka. Glavni motiv za primenu predikcionih metoda su svakako smanjenjenje rizika poslovanja i ostvarivanje profita.

Finansijski modeli se zasnivaju na određenim pretpostavkama o finansijskom tržištu i ponašanju tržišnih učesnika. Iako suštinski značajno različiti, prema dobijenim empirijskim rezultatima, finansijski instrumenti pokazuju određene zajedničke karakteristike. Posmatrano sa statističkog stanovišta, naizgled slučajne promene vrednosti finansijskih instrumenata dele izvesne zajedničke osobine [146].

Finansijske vremenske serije cena i prinosa na finansijsku aktivu karakteriše niz specifičnosti, koje nameću zahtev za njihovom detaljnom analizom u cilju adekvatnog postavljanja finansijskih modela [146].

Finansijsko modeliranje se zasniva kako na karakteristikama serije podataka, tako i na određenim pretpostavkama o finansijskom tržištu i ponašanju tržišnih učesnika. Iako se

standardni i opšteprihvaćeni modeli baziraju na teoriji efikasnog tržišta, finansijsko tržište je kompleksan, evolutivni i dinamičan sistem, koji se ponaša izrazito nelinerano [68].

Ukoliko se pretpostavi da su finansijska tržišta efikasna, ne bi bilo moguće predložiti model koji bi obezbedio dodatne prinose investitorima. Umesto toga, slaba forma hipoteze efikasnog tržišta (engl. *the efficient market hypothesis* - EMH) [47] pretpostavlja da se sve prošle promene cene finansijskih instrumenata reflektuju u današnjoj, tako da bi svaki pokušaj modeliranja cene finansijskih instrumenata negirao najznačajniju hipotezu na osnovu koje se objašnjava funkcionisanje finansijskih tržišta u nauci. U realnosti, međutim, primena različitih metoda predviđanja promene vrednosti finansijskih instrumenata, posebno tehnička analiza, omogućavaju investitorima ostvarenje određenih prinosa.

Predikciju u oblasti finansija uslovljavaju velika raznovrsnost, ali i nestacionarnost i nestruktuiranost podataka s visokim stepenom nestabilnosti i izraženim skrivenim vezama [73]. Poznato je da se sa dobijanjem novih informacija ponašanje tržišnih učesnika stalno menja i da je donošenje odluka o delovanju usko povezano sa iskustvom i intuicijom investitora.

Interesovanje stručne i akademske javnosti za kretanja na finansijskim tržištima uzrokovalo je brojna istraživanja na tom polju.

Ostvarivanje profita investiranjem u finansijske instrumente na tržištu kapitala bazira se na mogućnosti uspešnog predviđanja cena finansijskih instrumenata u budućnosti [68] i [73]. Poznato je da su precizna predviđanja kretanja indeksa cena akcija veoma važna za razvoj efikasne strategije trgovanja na tržištu [73]. Većina trgovinskih praksi usvojenih od strane finansijskih analitičara se oslanja na tačna predviđanja nivoa cena finansijskih instrumenata. Međutim, novije studije su pokazale da strategije trgovanja vođene prognozama o pravcu promene cena mogu biti efikasnije i generisati veći prinos u odnosu na tradicionalna predviđanja nivoa cena finansijskih instrumenata. Stoga se investiciona strategija može smatrati efektivnom samo ukoliko se zasniva na preciznom predviđanju trenda promene vrednosti konkretnog tržišnog indeksa [73] i [147].

Investitori uobičajeno koriste fundamentalnu i/ili tehničku analizu u analizi cene finansijske aktive i odlučivanju. Fundamentalna analiza proučava faktore koji utiču na razvoj privrede i privrednih društava u cilju određenja unutrašnje vrednosti finansijske aktive. Na makroekonomskom nivou ova analiza se fokusira na ekonomske podatke kao što su inflacija, nezaposlenost i nivo kamatne stope, kako bi procenila trenutnu i predvidela buduću stopu privrednog rasta. Na nivou privrednih društava, fundamentalna analiza se zasniva na finansijskoj racio analizi, ali može da uključuje i analizu konkurencije, menadžmenta i

poslovnih koncepata, dok se faktori koji utiču na ponudu i tražnju specifičnih proizvoda razmatraju na nivou privredne grane.

Za razliku od fundamentalne analize tržišta kapitala, tehnička analiza se zasniva na pretpostavci da kretanja na tržištu kapitala pružaju dovoljno informacija za predviđanje budućih vrednosti. Tehnička analiza se oslanja na brojne kvalitativne i kvantitativne metode u cilju predviđanja trenda promene cene finansijske aktive. Najjednostavniji kvalitativni metodi, koji se koriste u okviru ove analize, se baziraju na grafičkom prikazivanju cena finansijske aktive i obima trgovanja. Ovi metodi pomažu u prepoznavanju obrazaca promene koji mogu biti korišćeni u svrhu ostvarivanja profita u trgovanju hartijama od vrednosti. Prema [80] tehnička analiza predstavlja jednu od opšte prihvaćenih ekonomskih metoda za predviđanje trenda, koja se primenjuje na svetskim tržištima kapitala.

U odnosu na konvencionalne metode predviđanja finansijskih vremenskih serija, razvijenih tokom 70-ih i 80-ih godina, od kojih su najpopularnije ARCH model [45], GARCH [13] i Box-Jenkins ARIMA [14], u mnogim studijama algoritmi mašinskog učenja pokazali su se veoma efikasnim. Najčešće korišćeni algoritmi mašinskog učenja za predikciju na finansijskim tržištima jesu veštačke neuronske mreže (engl. *artificial neural networks – ANNs, NN*) [7] i [73], metodi podržavajućih vektora (engl. *support vector machines - SVMs*) [21], [68], [85] i [111] i jedna od reformulacija SVM metode, metod najmanjih kvadrata podržavajućih vektora (engl. *least squares support vector machines - LS-SVMs*) [23], [100] i [175]. U [121] je pokazano da metoda podržavajućih vektora i metod najmanjih kvadrata podržavajućih vektora, postižu bolje predikcione rezultate u ovoj oblasti u odnosu na ostale algoritme mašinskog učenja. Upravo se iz tog razloga metode podržavajućih vektora koriste za formiranje predikcionih modela u ovoj doktorskoj disertaciji.

POGLAVLJE 3

KERNEL FUNKCIJE I METODE PODRŽAVAJUĆIH VEKTORA

U okviru ovog poglavlja dat je pregled teorijskih osnova neophodnih za razumevanje oblasti mašinskog učenja, kernel funkcija u mašinskom učenju i metoda podržavajućih vektora i najmanjih kvadrata podržavajućih vektora kao najpoznatijih predstavnika kernel metoda mašinskog učenja. U okviru poglavlja koje sledi prikazana su i teorijska razmatranja neophodna za definisanje težinske kernel funkcije, koja će biti korišćena u eksperimentalnom delu ove doktorske disertacije.

3.1. KERNEL FUNKCIJE

Kod nadgledanog mašinskog učenja problemi se analiziraju kroz definisanje disjunktnih skupova podataka koji se koriste za obuku i procenu performansi predikcionog metoda. Najpre se definiše trening skup odnosno skup koji se koristi za obučavanje predikcionog metoda, $S = \{(x_k^{(j)}, y_k)\}$, $k = 1,\ldots,N$ i $j = 1,\ldots,d$ gde N predstavlja broj trening primera (engl. *training example, instance*), a svaki trening primer $\left(x_k^{(1)}, x_k^{(2)}, \ldots, x_k^{(d)}, y_k\right)$ je sastavljen od trening vektora $x_k^{(j)}$ koji se sastoji od d atributa (engl. *inputs, features*) i pridružene ciljne promenljive y_i (engl. *output, target, labels*). Višedimenzionalni prostor ulaznih vrednosti (engl. *input space, instance set*) označava se sa X, $x^{(j)} \in X$. Prostor ciljnih vrednosti (engl. *finate set, label set*) označen je sa Y, $y \in \mathrm{Y}$. Funkcija koja vrši preslikavanje $f : X \to Y$, je nepoznata i naziva se ciljna funkcija f (engl. *target functio*n).

Zadatak algoritama nadgledanog mašinskog učenja je da na osnovu zadatog trening skupa pronađe funkciju predviđanja h (engl. *hypothesized function*) iz skupa raspoloživih hipoteza H, $h \in H$, $H = \{h \mid h : X \to Y\}$ koja najbliže aproksimira ciljnu funkciju f [148].

Test skup se koristi za procenu performansi predikcionog modela. Osnovna pretpostavka o podacima je da elemente trening i test skupa čine međusobno nezavisne slučajne promenljive koje pripadaju istoj fiksnoj i nepoznatoj raspodeli *p*, i da su i trening i test skupovi nezavisno i identično distribuirani (engl. *independent and identically distributed, i.i.d*) u odnosu na distribuciju *p* [36].

Slično se može posmatrati predikcija iz ugla atributa, prediktivna metoda zapravo koristi selektovane atribute i u slučaju klasifikacije oznake (labele) klasa kako bi naučila funkciju mapiranja h kojom se atributi preslikavaju u izlazne promenljive [153].

Kako je navedeno u [67] algoritmi mašinskog učenja po pravilu daju dobre rezultate kada se koriste za rešavanje linearnih problema, ali su primene u realnim problemima i zavisnosti među podacima po pravilu nelinearne.

Cilj algoritma mašinskog učenja je da na osnovu trening skupa nauči zavisnost koja postoji među podacima, a da zatim predvidi izlazne vrednosti za prethodno nepoznate podatke. Kako je navedeno u [138] učenje je moguće pod pretpostavkom da postoji određena mera „sličnosti“ između test instance i trening skupa, kernel se upravo može intuitivno shvatiti kao mera sličnosti između tačaka skupa podataka. Iako bi se moglo pretpostaviti da je izbor mere sličnosti za izlazne veličine jednostavan kod primera binarne klasifikacije gde su predikcije ili tačne ili netačne, takav izbor je zapravo prema navodima iz [138] suštinsko pitanje u oblasti mašinskog učenja.

Osnovna ideja na kojoj se zasnivaju kernel metodi učenja je da se izvrši projekcija ulaznih podataka u takozvani vektorski prostor atributa.

Prednosti takve transformacije podatka ogledaju se u tome da je na taj način moguće transformisati nelinearne relacije između podataka u primarnom prostoru u linearne relacije u dualnom prostoru atributa, pri čemu se umesto eksplicitnog preslikavanja (projekcije) podataka u dualnom prostoru koristi skalarni proizvod između svih parova vektora u dualnom prostoru, a tako dobijene informacije su nezavisne od dimenzionalnosti preslikanog prostora.

Najpre će na karakterističnom primeru iz [138] biti predstavljena suština kernel transformacija. Primeri su prikazani u kontekstu prethodno definisanih skupova koji se koriste kod problema nadgledanog mašinskog učenja.

Razmotrimo mapiranje iz dvodimenzionalnog u trodimenzionalni prostor atributa. U slučaju dvodimenzionalnog ulaznog prostora, $X \in R^2$, svaki trening vektor je sastavljen od dva atributa, prema prethodno definisanoj notaciji: $x_k = (x_k^{(1)}, x_k^{(2)})$, $k = 1,...,N$, gde N predstavlja broj trening primera. Nadalje će se radi pojednostavljenja zapisa u konkretnom

primeru koristiti forma $x = (x_1, x_2)$. Za konkretni vektor, moguće je odrediti funkciju ϕ kojom se vrši preslikavanje iz dvodimenzionalnog ulaznog prostora u trodimenzionalni prostor atributa u oznaci $\phi : R^2 \to R^3$, sa sledećim svojstvima:

$$(x_1, x_2) \to (\nu_1, \nu_2, \nu_3) := (x_1^2, \sqrt{2}x_1x_2, x_2^2) \tag{3.1}$$

Nakon preslikavanja dobijaju se koordinate u višedimenzionalnom prostoru, označene sa (ν_1, ν_2, ν_3). Dakle $\phi(x) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$.

Uzmimo sada primer dve instance iz ulaznog dvodimenzionalnog prostora, $x_k = (x_k^{(1)}, x_k^{(2)})$, za $k = 1,2$, $x_1 = (x_1^{(1)}, x_1^{(2)})$ i $x_2 = (x_2^{(1)}, x_2^{(2)})$ na dalje u tekstu će se radi pojednostavljenja zapisivati: $x = x_1 = (x_1^{(1)}, x_1^{(2)}) = (x_1, x_2)$ i $z = x_2 = (x_2^{(1)}, x_2^{(2)}) = (z_1, z_2)$. Skalarni proizvod između ovih vektora predstavljen je sa $\langle x, z \rangle = x^T z$. Kvadrat skalarnog proizvoda se može definisati kao:

$$\begin{aligned} \langle x, y \rangle^2 &= (x_1z_1 + x_2z_2)^2 \\ &= x_1^2z_1^2 + 2x_1z_1x_2z_2 + x_2^2z_2^2 \\ &= \langle (x_1^2, \sqrt{2}x_1x_2, x_2^2), (z_1^2, \sqrt{2}z_1z_2, z_2^2) \rangle \\ &= \langle \phi(x), \phi(z) \rangle \end{aligned} \tag{3.2}$$

Sada možemo zameniti kvadrat skalarnog proizvoda funkcijom *k* na način $\langle x, z \rangle^2 = (x^T z)^2 = k(\phi(x), \phi(z))$, gde *k* predstavlja skalarni proizvod vektora u višedimenzionalnom prostoru i naziva se kernel funkcija:

Odnosno u suprotnom redosledu izvođenja:

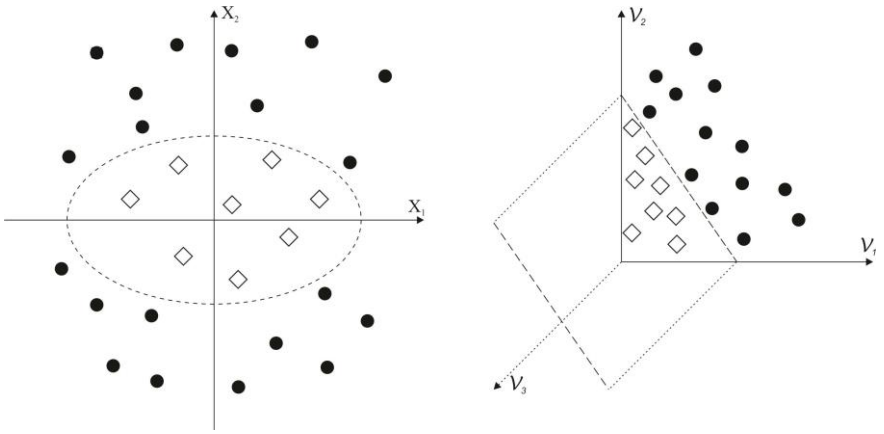
$$\begin{aligned} \langle \phi(x), \phi(z) \rangle &= (x_1^2, \sqrt{2}x_1x_2, x_2^2)^T (z_1^2, \sqrt{2}z_1z_2, z_2^2) \\ &= ((x_1, x_2)^T (z_1, z_2))^2 \\ &= \langle x, z \rangle^2 \end{aligned} \tag{3.3}$$

Dakle:

$$\langle \phi(x), \phi(z) \rangle_{R^3} = \langle x, z \rangle_{R^2}^2 \tag{3.4}$$

Pri čemu $k(x, z)$ u konkretnom primeru predstavlja polinomalnu kernel funkciju drugog stepena.

Dakle, kernel odgovara skalarnom proizvodu u višedimenzionalnom prostoru atributa, u kome su metodi predviđanja linearni. Na slici 3.1, predstavljen je efekat prethodno opisanog preslikavanja.



SLIKA 3.1 Binarna klasifikacija, transformacija funkcije razdvajanja 2D $\rightarrow$ 3D

Na osnovu slike 3.1, preuzete iz [138], može se uočiti da je primena polinomalne funkcije preslikavanja omogućila da se nelinerana elipsoidna granica razdvajanja u dvodimenzionalnom prostoru prevede u linearno razdvojivu hiper-ravan u trodimenzionalnom prostoru atributa.

Dalje je moguće izvršiti uopštavanja. Uzmimo za primer kernel funkciju $k(x,z)=(1+x^Tz)^2$, prateći prethodne oznake moguće je zapisati:

$$\begin{aligned} k(x,y) &= (1+x_1z_1+x_2z_2)^2 \\ &= (1+2x_1z_1+2x_2z_2+x_1^2z_1^2+2x_1z_1x_2z_2+x_2^2z_2^2) \\ &= (1,\sqrt{2}x_1,\sqrt{2}x_2,x_1^2,\sqrt{2}x_1x_2,x_2^2)^T(1,\sqrt{2}z_1,\sqrt{2}z_2,z_1^2,\sqrt{2}z_1z_2,z_2^2) \\ &= \phi(x)^T\phi(z) \end{aligned} \tag{3.5}$$

U prethodnom primeru vrši se dakle preslikavanje dvodimenzionalnog ulaznog prostora atributa u šestodimenzionalni prostor $\phi: R^2 \rightarrow R^6$.

Kernel trik se oslanja na dobro poznate matematičke formalizacije. Na osnovu detaljnih razmatranja predstavljenih u [152], može se istaći da za svaku simetričnu kontinualnu funkciju *K(x, z)* koja zadovoljava *Mercerove* uslove (*Mercer's conditions)*, postoji Hilbertov prostor *H*, preslikavanje $\phi(\cdot): R^n \rightarrow H$ i pozitivni brojevi λ_i, takvi da važi:

$$K(x,z)=\sum_{i=1}^{n_H}\lambda_i\phi_i(x)\phi_i(z) \tag{3.6}$$

gde $x, z \in R^n$ a n_H predstavlja dimenziju *H*. *Mercerovi* uslovi zahtevaju da za svaku kvadratnu integrabilnu funkciju *g(x)* bude ispunjena relacija:

$$\int K(x,z)g(x)g(z)dxdz \geq 0 \tag{3.7}$$

Izraz (3.7) se može napisati kao $K(x,z)=\sum_{i=1}^{n_H}\sqrt{\lambda_i}\phi_i(x)\sqrt{\lambda_i}\phi_i(z)$, gde su $\varphi_i(x)=\sqrt{\lambda_i}\phi(x)$ i $\varphi_i(z)=\sqrt{\lambda_i}\phi(z)$. Nakon zamene kernel funkcija se preko skalarnog proizvoda predstavlja na sledeći način:

$$K(x,z)=\varphi(x)^T\varphi(z) \tag{3.8}$$

Kako bi važila prethodna relacija kernel funkcija mora biti simetrično pozitivno definisana, što uslovljava da je i sam kernel *K* separabilan.

Na osnovu pregleda literature može se zaključiti da su najčešće korišćene kernel funkcije linearna, polinomna, gausova kernel funkcija (engl. *radial basis function* - RBF) i sigmoidna funkcija odnosno MLP kernel (engl. *multi layer perceptron*) [138] i [141]. Pri čemu se RBF kernel najčešće koristi kod problema nelinearne klasifikacije [55] i [152]. Jednačine 3.9 - 3.12 predstavljaju matematičke formulacije najznačajnijih kernela, $k(\cdot,\cdot)$:

Linearan: $$k(x,x_k)=x_k^T x \tag{3.9}$$

Polinomialan: $$k(x,x_k)=(\tau+x_k^T x)^d \tag{3.10}$$

Prilikom izbora kernel funkcije potrebno je napomenuti da su *Mercerovi* uslovi ispunjeni za sve pozitivne vrednosti parametra τ kod polinomalnog kernela.

MLP kernel: $$k(x,x_k)=(\tau+x_k^T x)^d \tag{3.11}$$

Takođe, prilikom izbora kernel funkcije treba prema [152] biti svestan ograničenja, da prethodno istaknuti *Mercerovi* uslovi nisu ispunjeni za određene kombinacije *k1* i *k2* u slučaju MLP kernela.

Gausov RBF kernel: $$k(x,x_k)=e^{\frac{-\|(x-x_k)\|^2}{\sigma^2}} \tag{3.12}$$

U slučaju RBF kernela *Mercerovi* uslovi su ispunjeni za sve vrednosti parametara σ. Takođe, uz određena uprošćenja i primenom Tejlorovog razvoja, na osnovu analogije sa polinomalnim kernelom, može se zaključiti da RBF kernel vrši mapiranje ulaznog prostora u potencijalno beskonačno dimenzionalni prostor.

Prethodno smo definisali kernel funkcije koje se najčešće koriste u praktičnim primenama, međutim važno je istaći da su kernel funkcije zatvorene za linearne kombinacije i da je na osnovu svojstava kernel funkcije moguće izdvojiti i niz dozvoljenih operacije nad kernelima pri čemu su i novo kreirani kerneli pozitivno definisane funkcije. U [141] i [152] imogu se naći kompleni dokazi i bliža objašnjenja za sledeće operacije nad kernelima na osnovu kojih je moguće definisati nove kernele iz već postojećih:

$$k(x,z) = ak(x,z), a > 0 \tag{3.13}$$

$$k(x,z) = k_1(x,z) + b, b > 0 \tag{3.14}$$

$$k(x,z) = x^T Pz \, (P = P^T > 0) \tag{3.15}$$

$$k(x,z) = k_1(x,z) + k_2(x,z) \tag{3.16}$$

$$k(x,z) = k_1(x,z)k_2(x,z) \tag{3.17}$$

$$k(x,z) = f(x)f(z) \tag{3.18}$$

$$k(x,z) = k_3(\phi(x), \phi(z)) \tag{3.19}$$

$$k(x,z) = p(k_4(x,z)) \tag{3.20}$$

$$k(x,z) = \exp(k_4(x,z)) \tag{3.21}$$

Pri čemu $a, b \in R^+$, dok su k_1, k_2, k_3 i k_4 simetrične pozitivno definisane kernel funkcije, $f(\cdot): R^n \to R$, $\phi(\cdot): R^n \to R^{n_h}$, a $p(x)$ je polinom sa pozitivnim koeficijentima. P je simetrična pozitivno semidefinitna matrica veličine $n \times n$. Takođe, u [152] i [141], mogu se naći i dodatna razmatranja o načinima konstruisanja, problemima izbora i tipovima kernel funkcija.

Pored navedenih tipova kernel funkcija postoji grupa kernela koja je definisana nad grafovima [77], biosekvencama [9], slikama [65] kao i stringovima i tekstualnim podacima [97], koji uslovno vrše projekciju opštih skupova podataka u euklidski prostor gde se može izvršiti obučavanje algoritma [52]. Pri čemu upravo sposobnost da se obrađuju opšti tipovi podataka jeste jedna od glavnih inovacija uvedena kernel pristupom [33].

Kernel funkcija dozvoljava da se podaci koriste kao da su projektovane vrednosti u više dimenzionalnom prostoru, ali da se svi proračuni zapravo izvršavaju u originalnom prostoru atributa. Prethodno opisan efekat, u literaturi se navodi kao kernel trik. Na osnovu predstavljenih teorijskih postulata upotrebom kernel trika konstruisane su mnoge klase algoritma uključujući metode podržavajućih vektora i *Kernel Principal Components Analysis - KPCA* [139].

Na osnovu analiza iz studije [110] mogu se izdvojiti sledeće osobine kernel metoda: 1) Kernel metodi koji obimno koriste optimizacione metode, eksplicitno zasnovani na teorijskom modelu učenja i kernel funcije kao nelinearne mere sličnosti uspešno prevazilaze ograničenja prethodnih modela učenja koji su se zasnovali na nekim heuristikama ili analogijama sa prirodnim sistemima učenja, kao što su neuronske mreže - ANN; 2) Implicitna preslikavanja su i ranije postojala kod nekih algoritama mašinskog učenja u skrivenim slojevima ANN mreža na primer, ali kernel metodi učenja nisu pogođeni

problemima lokalnog minimuma pošto u fazi obuke podrazumevaju konveksnu optimizaciju [33]. Posebno treba istaći da se kod učenja ne uzima u obzir dimenzionalnost problema već kompleksnost funkcije razdvajanja u višedimenzionalnom prostoru. Pored toga da bi se iskoristile prednosti upotrebe kernela, ne mora se vršiti projekcija ili preporcesiranja ulaznog prostora podataka u višedimenzionalni prostor, a broj atributa ne utiče na parametre koji se optimizuju.

Metoda podržavajućih vektora - SVM i metod najmanjih kvadrata podržavajućih vektora - LS-SVM, predstavljaju najpoznatije algortme učenja koji se baziraju na kernel funkcijama.

U [152] se navodi da je najznačajniji napredak u teoriji SVM [155] načinjen upravo kada je linearni SVM proširen kernel funkcijama i postao primenljiv na klasu nelinearnih problema. Metodi podržavajućih vektora se uspešno primenjuju na mnogim realnim problemima i uobičajeno se koriste za rešavanje nelinearnih klasifikacionih problema [152]. Oba metoda imaju i široku primenu u oblasti predikcije finansijskih vremenskih serija [23], [68] i [111].

3.2. METOD NAJMANJIH KVADRATA PODRŽAVAJUĆIH VEKTORA

Prethodno definisani pristup kernel transformacija je u okviru metoda podržavajućih vektora uobličen kroz notaciju mapiranja ulaznih atributa u prostor atributa korišćenjem nelinearne funkcije projekcije $\varphi : R^n \to F$, gde F predstavlja prostor skalarnih proizvoda odnosno prostor atributa, gde se zapravo vrši pronalaženje razdvajajuće hiper-ravni.

U okviru ovog poglavlja iznosi se teorijska osnova LS-SVM metoda, kao novijeg te i uslovno manje poznatog metoda, bez oslanjanja na SVM teoriju najpre jer su osnove SVM metoda dobro poznate [155]. Takođe, najznačajnija poboljšanja u izvedenim simulacijama u eksperimentalnom delu ove doktorske disertacije dobijena su upravo primenom LS-SVM metoda te je i to jedan od razloga.

Ovde je važno istaći da su sva prethodna razmatranja o kernel funkcijama podjednako primenjiva i na SVM i na LS-SVM model, što je kasnije i pokazano u eksperimentalnom delu ove doktorske disertacije.

Osnovna terminologija u vezi sa zadacima algoritma mašinskog učenja definisana je u uvodnom poglavlju. Sada ćemo na osnovu [152] definisati notaciju potrebnu za razumevanje klasifikacije primenom metoda najmanjih kvadrata podržavajućih vektora.

Trening skup se definiše vrednostima $\{x_k, y_k\}, k = 1,...N$, gde N predstavlja ukupan broj trening primera sa ulaznim vrednostima iz skupa $x_k \in R^n$, dok ciljne vrednosti pripadaju skupu $y_k \in \{-1,1\}$. Predikcioni model u primarnom prostoru ulaznih atributa može se formirati korišćenjem nelinearnog mapiranja $\varphi(\cdot): R^n \to R^{n_h}$ kojim se vrši preslikavanje ulaznog prostora atributa u višedimenzionalni prostor i definiše se na sledeći način:

$$y(x) = sign\left[\omega^T \varphi(x) + b\right] \tag{3.22}$$

gde ω predstavlja težinski vektor, a *b* označava *bias term* (pomeraj).

Optimizacioni problem formulisan je u primarnom prostoru jednačinom:

$$\min_{\omega,b,e} J_p(\omega, e) = \frac{1}{2}\omega^T\omega + \frac{1}{2}\gamma\sum_{k=1}^{N} e_k^2 \tag{3.23}$$

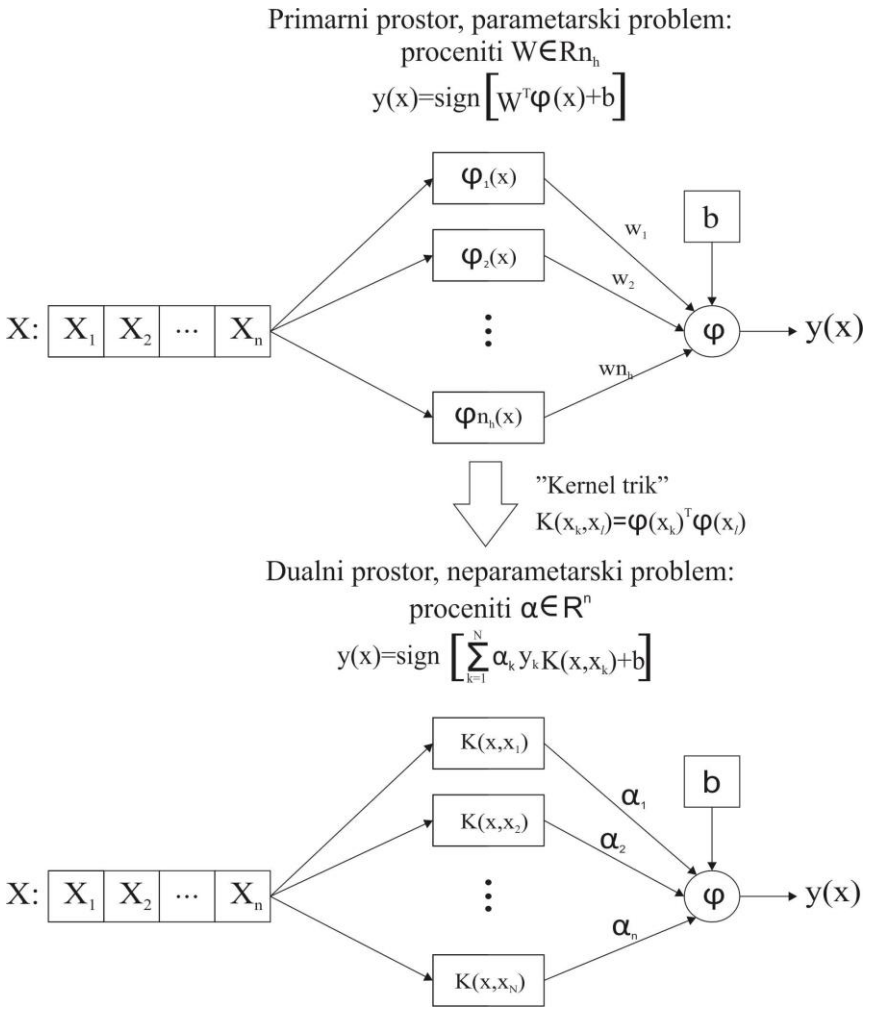
sa sledećim ograničenjima:

$$y_k\left[\omega^T\varphi(x_k) + b\right] = 1 - e_k,\ k = 1,...,N \tag{3.24}$$

gde vrednosti e_k predstavljaju dozvoljene greške prilikom kreiranja predikcionog modela, a γ je parametar koji im dodeljuje relativne težine.

U slučaju LS-SVM metoda (ista postavka važi i za SVM) početna formulacija problema je u primarnom težinskom prostoru sa višedimenzionalnim prostorom atributa koji je dobijen primenom preslikavanja $\varphi(x)$. U primarnom prostoru problem je određen parametrima ω. U primarnom prostoru problem je dakle parametarski sa fiksnom dimenzijom vektora $\omega \in R^{n_h}$, gde n_h predstavlja dimenziju prostora preslikavanja. Optimizacioni problem se rešava kroz konstruisanje Lagranžeove dualne forme problema. Problem se dalje rešava u dualnom prostoru Lagranžeovih multiplikatora nakon primene kernel trika, kojim se omogućava rad u višedimenzionalnom prostoru atributa bez eksplicitnog obavljanja izračunavanja u njemu. Na taj način, optimizacioni problem postaje neparametarski u dualnom prostoru. U dualnom prostoru određuje se vektor α i složenost rešenja raste isključivo sa porastom broja trening primera *N*. Dakle, sva izračunavanja se obavljaju u polaznom nepreslikanom prostoru sa vektorima dimenzije *n*.

Primarno dualna interpretacija LS-SVM metoda može se videti na slici 3.2, adaptacija iz [148].



SLIKA 3.2 Primarno-dualna interpretacija LS-SVM metoda

Nakon rešavanja optimizacionog problema klasifikacioni model u dualnom prostoru može se predstaviti sa:

$$y(x) = sign\left[\sum_{k=1}^{N} \alpha_k y_k K(x, x_k) + b\right] \tag{3.25}$$

Skalarni proizvod:

$$K(x, x_k) = \varphi(x)^T \varphi(x_k) \tag{3.26}$$

predstavlja kernel funkciju, a α_k su Lagranžeovi multiplikatori. Kompletno izvođenje i rešavanje optimizacionog problema može se naći u [152].

Prilikom korišćenja RBF funkcije definisane sa:

$$K(x, x_k) = e^{\frac{-\|(x-x_k)\|^2}{\sigma^2}} \tag{3.27}$$

kernel parametar σ nije unapred poznat. Optimalna kombinacija paremetara (γ, σ) se određuje primenom *grid-search* algoritma u kombinaciji sa *k-fold cross-validation* pristupom

kao najčešće korišćenim metodom [5].

Kako je prethodno istaknuto, kernel funkcije ne uzimaju u obzir razlike koje postoje u relevantnosti atributa.

Prema navodima u radovima [60], [167] i [172] različita raspodela informacija među atributima se ističe korišćenjem težinskih kernela. Kao rezultat dodeljivanja težina zapravo se uključuje relativni značaj svakog od atributa, a promena u vrednosti kernela koja je posledica dodeljivanja težine potencijalno unapređuje sposobnost generalizacije i smanjuje greške prilikom predikcije.

U narednoj sekciji predstavljene su osnove težinske kernel funkcije u kontekstu LS-SVM metoda nadgledanog mašinskog učenja.

3.3. TEŽINSKI KERNEL

Težinska kernel funkcija definisana je kao $K(\theta x, \theta x_k)$ gde θ predstavlja vektor sa težinama izabranog skupa atributa. Kompletne matematičke derivacije za težinsku kernel funkciju mogu se naći u [167] za slučaj SVM metode i mogu se prilagoditi za LS-SVM metodu. Klasifikacioni problem u dualnom prostoru atributa sa dodeljenim težinama atributima formulisan je u jednačini (3.28) sa naznakom da se težine atributa takođe uzimaju u obzir prilikom računanja parametara α_k i b.

$$y(x) = \text{sign}\left[\sum_{k=1}^{N} \alpha_k y_k K(\theta x, \theta x_k) + b\right] \tag{3.28}$$

Iz jednačine (3.28) i prema [172] može se zaključiti da definisana težinska funkcija jezgra ne zavisi od tipa kernel funkcije.

Na osnovu [60] i [167] vektor težina treba da zadovoljava sledeće uslove:

$$\begin{aligned} &0 \le \theta_i \le 1 \quad i = 1,...,d \\ &i\ da\ je \\ &\sum_{i=1}^{d} \theta_i = 1 \end{aligned} \tag{3.29}$$

Težinska RBF kernel funkcija sada može biti zapisana na sledeći način:

$$K(x, x_k) = e^{\frac{-\|\Theta(x - x_k)\|^2}{\sigma^2}} \tag{3.30}$$

Pri čemu je $\Theta = \text{diag}[\theta_1, \theta_2, \cdots, \theta_n]$.

U narednom koraku kao i kod RBF kernel funkcije, treba ustanoviti optimalne vrednosti kombinacije parametara (γ, σ).

Izbor kernel funkcije je jedan od osnovnih izazova prilikom kreiranja predikcionih modela. Odluka o tipu kernel funkcije može se doneti i uz pomoć eksperta iz domena primene, pošto je mera sličnosti između podataka ekspertima prepoznatljiva na osnovu iskustva. Kerneli koji su specifični za domen predviđenja se intezivno koriste u bioinformatici ili kod problema prepoznavanja teksta [141].

Dodeljivanje težina atributima kroz kernel funkciju smatra se jednim od potencijalnih načina da se razmatra značaj različitih atributa [172].

POGLAVLJE 4

IZBOR ATRIBUTA U MAŠINSKOM UČENJU

Problem izbora skupa atributa kojim se prikazuje posmatrani domen je veoma bitan ali i zahtevan korak u procesu analize podataka, u prilog tome govore brojna istraživanja [40], [108] i [133]. Činjenica je i da je do sada razvijen i predložen veliki broj metoda izbora atributa, ali i da se novi metodi i dalje aktivno razvijaju [41] i [95].

U poglavlju koje sledi izložene su teorijske osnove neophodne za razumevanje problema izbora atributa, a zatim su prikazani različiti metodi izbora atributa, zajedno sa komparativnom analizom njihovih svojstava.

4.1. OSNOVNE KATEGORIZACIJE METODA IZBORA ATRIBUTA

U poslednje dve decenije značajna grupa radova razmatrala je problem odabira odgovarajućeg algoritma izbora atributa za određene domene [40] i [96]. Međutim, bez obzira na širok spektar raspoloživih algoritama, izbor uslovno optimalnog algoritma za konkretni problem se pokazao kao izuzetno zahtevan zadatak.

Prema [56] izbor atributa sa aspekta eliminacije atributa smanjuje vreme izračunavanja, smanjuje efekte dimenzionalnosti podataka, poboljšava predikcione sposobnosti i pomaže u boljem razumevanju podataka. Prema [108] izbor atributa kao ključan korak u analizi podatka obuhvata analizu originalnog skupa atributa i izbor optimalnog podskupa na osnovu nekog od kriterijuma evaluacije.

Cilj algoritama za izbor atributa je da se pronađe podskup ulaznog skupa atributa koji adekvatno reprezentuje ulazne podatke, doprinosi smanjenju šuma, vrši redukciju irelevantinh podataka i pri tome omogućava postizanje dovoljno dobrih predikcionih rezultata [81].

Sa praktičnog stanovišta posmatrano algoritmi mašinskog učenja uče rešenje problema na osnovu uzoraka podataka, a izabrani atributi matematički definisano predstavljaju minimalni podskup nezavisnih promenljivih koji objašnjavaju šablone (engl. *patterns*) koji postoje u podacima [43].

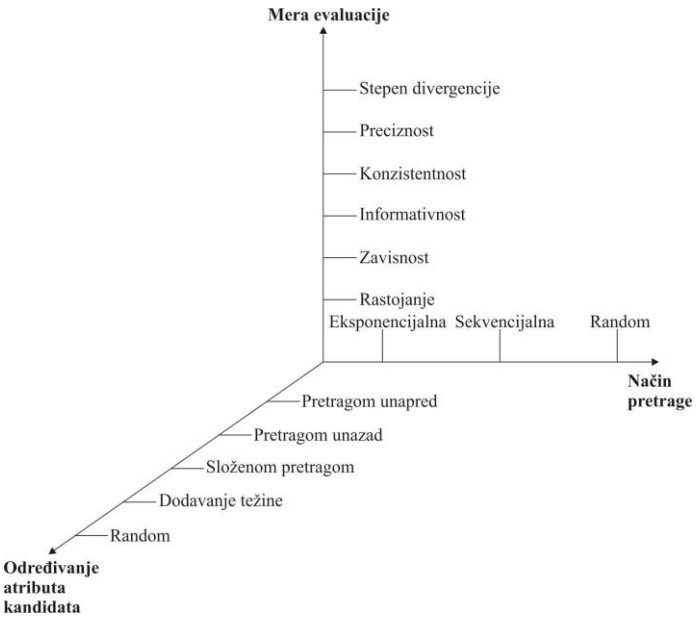
Formalno definisano prema [108] izbor atributa svodi se na određivanje podskupa skupa atributa koji dovodi do zadovoljenja jednog od narednih uslova:

- Pronaći podskup skupa atributa, često predefinisane veličine, koji daje optimalnu (najveću/najmanju) vrednost optimizacionog kriterijuma (najmanja u slučaju da se optimizuje funkcija greške)
- Pronaći podskup skupa atributa koji daje vrednost optimizacione funkcije veću ili jednaku najmanjoj vrednosti koja je postavljena kao spoljni kriterijum optimizacije, to jest koja zadovaljava ograničenja optimizacionog kriterijuma
- Pronaći kompromisni rezultat između veličine skupa podskupa atributa i maksimalne vrednosti optimizacione funkcije.

U teorijskoj perspektivi prema [12], izdvajaju se tri osnovne karakteristike kojima se definiše svaki algoritam izbora atributa:

1. Organizacija pretraživanja prostora atributa;
2. Način na koji se određuje naredni atribut za procenu i
3. Mera evaluacije, odnosno funkcija po kojoj se procenjuje značaj atributa. Pojam relevantnosti atributa, sam po sebi je takođe deo širih teorijskih razmatranja koja se mogu naći u [35], [63] i [108]. Prihvaćeno stanovište relevatnosti u okviru ove doktorske disertacije, je po definiciji iz [108], „relevatnost u odnosu na cilj“.

Na slici 4.1, preuzetoj iz [108], prikazan je trodimenzionalni okvir za praćenje karakteristika algoritma za izbor atributa Strategija pretrage i metrika predstavljaju dva osnovna faktora kod svih algoritama izbora atributa. Iz tog razloga su kao prve dve dimenzije upravo postavljeni ovi kriterijumi. Treća dimenzija predstavlja načine određivanja narednog atributa kandidata. Što se tiče strategija pretrage prostora atributa, mogu se istaknuti strategije kompletne pretrage (iscrpljivanje, engl. *complete*), sekvencijalna pretraga (engl. s*equential*) i strategija nasumične odnosno slučajne pretrage (engl. *random*).



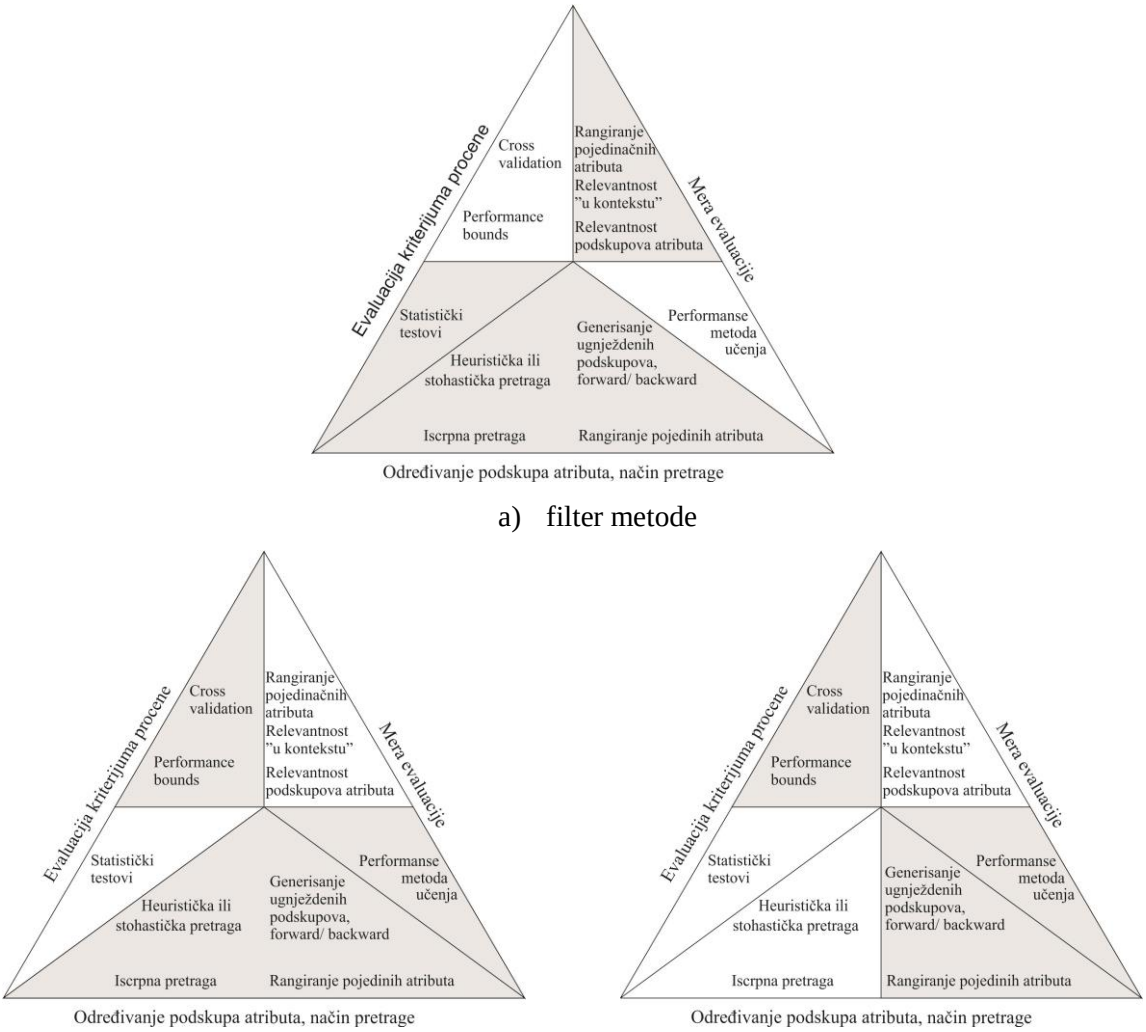
SLIKA 4.1 Karakteristike metoda izbora atributa

Na slici 4.1, preuzetoj iz [108] se mogu uočiti pet osnovnih načina generisanja narednog člana skupa: pretragom unapred, pretragom unazad, složenom pretragom, *random* pretragom i pretragom na osnovu težina. Kod mera evaluacije izdvajaju se: informativnost, preciznost, stepen divergencije. Kod strategija pretrage definišu se eksponencijalna, sekvencijalna i *random*. Šira teorijska osnova za razmatranje problema izbora atributa može se naći, na primer, u [35], [63] i [108].

U okviru istraživanja [153] se navodi da se prema tipu zadatka klasifikaciji ili klasterovanju, odnosno na osnovu toga da li su podaci označeni ili nisu, algoritmi izbora atributa mogu podeliti u nadgledane, nenadgledane i polunadgledane (engl. *semi-supervised*). U vezi sa time potrebno je istaći da će predmet razmatranja u okviru ove doktorske disertacije biti samo nadgledani metodi izbora atributa dok se detaljnije informacije za ostale dve kategorije mogu naći u [70], [88], [129] i [181].

Razmatranjem interakcije šeme izbora atributa sa predikcionim metodom koji će se koristiti kod kreiranja predikcinog modela, odnosno na osnovu odnosa između metoda izbora atributa i algoritma učenja dolazi se do najopštije podele nadgledanih metoda izbora atributa koja uključuje postojanje dve kategorije: filter metode i *wrapper* metode [63]. *Wrapper* metode se u literaturi mogu naći i pod nazivom metode prethodnog (probnog) učenja [112]. U [56] i [96] podjednako se navodi i podela na tri opšte kategorije odnosno metode izbora atributa, koja je zbog širine pristupa prihvaćena kao metodološki okvir u ovoj doktorskoj disertaciji. Prvi pristup svakako predstavljaju filter metode koje ispituju opšte karakteristike podataka nezavisno od korišćenog metoda predikcije. Drugu grupu metoda čine *wrapper*

metode, kod kojih je potraga za optimalnim podskupom skupa atributa zavisna od korišćenog predikcionog metoda. Poslednju grupu čine *embedded* metodi kod kojih je izbor atributa korak u okviru samog procesa učenja odnosno treniranja predikcionog metoda [56].



a) filter metode

b) *wrapper* metode

c) ugrađene metode

SLIKA 4.2 Odnos filter, *wrapper* i ugrađenih metoda izbora atributa

U odnosu na opštu kategorizaciju metoda izbora atributa prikazanu na slici 4.1, na slici 4.2, preuzetoj iz [62], može se videti odnos filter, *wrapper* i *embedded* metoda i komponente koje oni koriste. U narednim poglavljima je na osnovu predstavljene podele dat pregled svojstava svake od metoda izbora atributa.

4.2. FILTER METODE

Kao što je prethodno navedeno kod filter metoda funkcija izbora atributa je nezavisna od samog predikcionog metoda.

Prema [63], [112] i [133] moguće je izdvojiti podelu filter metoda na dve osnovne kategorije, metoda koje vrše vrednovanje pojedinačnih atributa i druga podgrupa metoda koja vrši vrednovanje na nivou podskupa skupa atributa.

Obe grupe filter metode se često prema [62] postovećuju samo sa metodama rangiranja koje koriste različite tehnike rangiranja atributa kao kriterijume za izbor atributa i u osnovi predstavljaju korak preprocesiranja podataka. Značaj svakog pojedinačnog atributa se određuje na osnovu različitih metrika poređenja, ali nezavisno od uticaja izabranog podskupa atributa na performanse samog algoritma učenja.

Osnovna stvar kod filter metoda rangiranja predstavlja određivanje značajnosti određenog atributa u odnosu na ostale atribute ili željene izlazne vrednosti. Prema [56] atribut se može smatrati irelevantnim ako je uslovno nezavistan od posmatrane izlazne vrednosti. Odnosno, atribut koji se smatra značajnim, može biti nezavistan od ostalih ulaznih atributa, ali mora biti zavistan od izlazne veličine.

Karakteristike metoda koji pripadaju prvoj grupi su da se vrši vrednovanje atributa, a zatim se podskup atributa određuje ili na osnovu unapred definisanog praga ili na osnovu već zadate veličine podskupa atributa. Ujedno potreba da se spolja odrede kriterijumi za odabir predstavlja i najveći nedostatak ovakvog pristupa. Drugi nedostatak ogleda se u tome što se vrednovanjem pojedinačnih atributa ne može uočiti njihova međusobna korelisanost, pa može doći do pojave redundatnih atributa u izabranom podskupu.

U slučaju da je potrebno izvršiti vrednovanje na nivou podskupova skupa atributa potrebno je koristiti i neku od strategije pretrage ulaznog skupa atributa, dok se kao funkcije vrednovanja koriste određene multivarijantne statističke analize.

Pošto su strategije pretrage u osnovi iste kod filter i *wrapper* metoda, slika 4.2, u ovoj doktorskoj disertaciji strategije pretrage prostora atributa razmatrane su samo u okviru *wrapper* metoda, gde su korišćene i u svojstvu klasifikacije *wrapper* metoda izbora atributa.

U nastavku ovog poglavlja smatraće se da se ulazni skup podataka sastoji od N primera, $\{x_k^{(j)}, y_k\}$, $k = 1,\ldots,N$ od kojih svaki sadrži d atributa $j = 1,\ldots,d$. Gde je x_k k-ti primer dok je y_k u opštem slučaju izlazna vrednost, diskretna ili kontinualna veličina. Sa x_i označićemo vektor sa vrednostima svih trening primera posmatranog atributa iz skupa ulaznih atributa, u oznaci $x_i = (x_k^{(j)})$, $i = 1,\ldots,N$. Za svaku konkretnu instancu (x_i, y_k) smatra se da predstavlja realizaciju slučajnih promenljivih (X, Y). U [56], [12] i [63] predstavljeni su različiti okviri i metrike za ispitivanje relevantnosti atributa. U nastavku rada biće

predstavljeni neki od najznačajnijih pristupa.

4.2.1. KORELACIONI KRITERIJUMI

Kod ovog tipa metoda vrši se procena korelisanosti između dva skupa podataka, odnosno izračunavanje korelacionih koeficijenata koji predstavljaju mere povezanosti između skupova podataka. Dakle, vrši se provera relevantnosti svakog od atributa ili podskupa skupa atributa tako što se ocenjuje koliko se varijacije u jednom skupu podataka reflektuju na drugi skup podatka.

Pristup korelacionih kriterijuma odražava stav da podskup skupa atributa treba da se sastoji od atributa koji su u visokoj korelaciji sa izlaznim podacima, dok su međusobno atributi slabo korelisani [64].

Jedan od najjednostavnijih kriterijuma koji se koristi za određivanje koeficijenta korelacije među podacima jeste Pirsonov koeficijent korelacije, koji predstavlja kovarijansu izraženu u jedinicama standarnih devijacija varijabli koje se posmatraju, predstavljen je u jednačini 4.1:

$$R(i) = \frac{\mathrm{cov}(x_i, y)}{\sqrt{\mathrm{var}(x_i)\,\mathrm{var}(y)}} \tag{4.1}$$

gde x_i predstavlja sve vrednosti posmatranog atributa iz prostora X, a Y je slučajna promenljiva koja predstavlja realizaciju izlaza *y*. U prethodnoj formuli $\mathrm{cov}(x_i, y) = \sum_{i=1}^{N}(x_i - \bar{x}_i)(y_i - \bar{y})$, N je broj trening primera u skupu. Dok su $\mathrm{var}(x_i) = \sum_{i=1}^{N}(x_i - \bar{x}_i)^2$ i $\mathrm{var}(y) = \sum_{i=1}^{N}(y_i - y)^2$, gde $\bar{x}_i$ predstavlja aritmetičku sredinu uzorka za izabrani atribut, a $\bar{y}$ predstavlja aritmetičku sredinu uzorka izlaznih vrednosti. Ograničenje predstavlja da je moguće utvrditi samo linernu zavisnost između posmatanog atributa i ciljne vrednosti i to u slučaju kontinulanih vrednosti, ali se u [63] navode i načini da se ovo ograničenje prevaziđe.

Primenom CFS metoda (engl. *correlation-based feture selection*) [64] određuje se relevantnost podskupa atributa a ne individualnih atributa i može se definisati kao:

$$CFS_s = \frac{kr_{cf}}{\sqrt{k + k(k-1)\bar{r}_{ff}}} \tag{4.2}$$

pri čemu *k* predstavlja broj atributa u podskupu atributa obeleženim sa *S*, $\bar{r}_{ff}$ predstavlja prosek vrednosti atributa u odnosu na korelaciju sa drugim atributima, dok r_{cf} predstavlja prosek vrednosti korelacionog koeficijenta atributa u odnosu na klasnu korelaciju [64]. U

[64] je dato i proširenje za diskretne vrednosti podataka, te je metod moguće primeniti i za određivanje korelacije kod diskretnih vrednosti atributa.

T-test statistika je neredni često korišćeni pristup koji u praksi daje bolje rezultate kod primene u problemima klasifikacije.

4.2.2. PROCENE ZAJEDNIČKIH INFORMACIJA

Metode koji pripadaju ovoj grupi koriste za izbor atributa kriterijume zasnovane na teoriji informacije odnosno entropiji koja se uobičajeno koristi u proceni [63], [79] i [158]. Definicija zajedničkih (uzajamnih) informacija (engl. *mutual information – MI*) je izvedena u [30]. MI određuje količina zajedničkih informacija koju promenljive dele, odnosno koliko poznavanje jedne od promenljivih smanjuje neodređenost one druge. MI se u formulama označava kao $I(X,Y)$.

U slučaju kontinulanih slučajnih promenljivih X i Y, količina zajedničke informacije $I(X,Y)$ se izračunava prema jednačini:

$$I(X,Y) = \int_{x_k}\int_{y} p(x_k,y)\log\frac{p(x_i,y)}{p(x_i)p(y)}dxdy \tag{4.3}$$

U jednačini 4.3 $p(x_i,y)$ predstavlja zajedničku gustinu verovatnoće (engl. *joint probability density function - PDF*) kontinulanih slučajnih prmenljivih X i Y dok su $p(x_i)$ i $p(y)$ marginalne gustine. MI predstavlja zapravo razliku koja se dobija kada se od vrednosti entropije za promenljivu *Y* oduzme vrednost entropije za *Y* kada je promenljiva *X* poznata. Pošto zajedničke gustine verovatnoća nisu unapred poznate one se u praktičnim problemima ne obračunavaju prema prethodnoj jednačini već se vrši procena na osnovu skupa podataka.

MI između dve diskretne slučajne promenljive X, Y izračunava se prema jednačini:

$$I(X,Y) = \sum_{x_i}\sum_{y} P(X=x_i,Y=y)\log\frac{P(X=x_i,Y=y)}{P(X=x_i)P(Y=y)} \tag{4.4}$$

Pri čemu kod diskretnih slučajnih promenljivih, $P(X=x_i)$ i $P(Y=y)$ predstavljaju verovatnoću da posmatrana slučajna promenljiva ima neku od mogućih vrednosti observacije. Odnosno $P(X=x_i,Y=y)$ predstavlja verovatnoću za skup zajedničkih opservacija.

MI je uvek veće od nule ili jednako nuli. Odnosno, ako su promenljive X i Y nezavisne, onda je MI jednako nuli, u suprotnom je MI veće od nule. Bliža određenja i definicije mogu se naći u [63], [79] i [158].

Određivanjem MI za razliku od korelacionih koeficijenata mogu se otkriti i nelinerane

zavisnosti između slučajnih promenljivih, a pri tome MI može biti definisan i za podskupove atributa.

Na kraju slično kao i kod korelacionih kriterijuma, ukoliko postoji mnogo zajedničkih informacija između atributa i izlazne klase, onda je u teoriji takav atribut značajan i koristan za treniranje predikcionog metoda.

4.2.3. RELIEF ALGORITMI

RELIEF grupa algoritama predstavlja jednostavan i efikasan način za određivanje težina atributima pri čemu se vrši odabir statistički značajnih atributa.

Prva verzija RELIEF algoritma rešavala je problem izbora optimalnog skupa atributa kod problema binarne klasifikacije [75]. Kod RELIEF algoritama težine atributa se određuju kroz iterativni postupak. Inicijalno su težine svih atributa postavljene na vrednost *0*. Zatim se vrši nasumični odabir jednog primera iz skupa podataka za učenje i određuju se njegovi najbliži susedi koji pripadaju istoj klasi odnosno suprotnoj klasi. U narednom koraku se vrši upoređivanje vrednosti svakog atributa u izabranom primeru sa određnim susedima. Ako promena vrednosti atributa vodi do promene klase, onda se tom atributu povećava težina, i suprotno inicijalna težina se smanjuje ako promena vrednosti atributa ne utiče na promenu klase. Procedura ažuriranja težina atributa se sprovodi ili nad slučajno izabranim uzorkom podataka ili nad celokupnim skupom podataka. Na kraju se vrši normalizacija dobijenih težina, te su konačno dobijene težine u opsegu vrednosti $[-1, 1]$, a treba ih razumeti kao vrednost verovatnoće da je posmatrani atribut značajan. Izbor podskupa skupa atributa za obuku predikcionih metoda na kraju se svodi na odabir onih atributa čija težina prelazi unapred definisani prag vrednosti.

Algoritam RELIEF-F je nadogradnja RELIEF algoritma predstavljena u [130], kojom se prevazilaze problemi šumova u podacima, dozvoljava da se radi sa nekompletnim podacima i može se koristiti i kod problema klasifikacije u više klasa. Prema navodima iz [130] RELIEF-F algoritmi uspešno identifikuju uslovnu zavisnost među podacima i daju sveobuhvatan pogled na atribute kako za probleme klasifikacije tako i za regresione probleme.

Složenost metoda je u linearnoj zavisnosti sa brojem atributa i veličinom trening skupa, tako da se pristup veoma efikasno može koristiti kod velikih skupova podataka kao i kod skupova atributa kod kojih postoji velika međuzavisnost [130]. Metod nije ni vremenski zahtevan, uzevši u obzir činjenicu da se broj provera može zadati kao konstantan argument

funkcije i ne moraju se nužno svi podaci uzeti za obradu.

Osnovni nedostatak ovog metoda leži u potrebi da se odredi prag vrednosti (engl. *threshold*) kao i u nasumičnom odabiru primera od koga se kreće sa postupkom dodeljivanja težina atributima.

4.3. *WRAPPER* METODE

Metode izbora atributa koje pripadaju ovoj grupi određuju podskup skupa atributa na osnovu evaluacije poboljšanja predikcije samog prediktora korišćenjem posmatranog podskupa atributa. U suštini za evaluaciju optimalnog skupa atributa koristi se sam predikcini metod. U osnovi kod ovog načina selekcije atributa, odbacuju se iz skupa atributa atributi koji ne doprinose poboljšanju kvaliteta predikcije. Prema [63] i [153] kada se *wrapper* metode primenjuju potrebno je definisati određene kriterijume. Prvi se odnosi na način pretrage svih podskupova skupa atributa. Zatim je potrebno odrediti i izvršiti evaluaciju podskupova atributa u odnosu na postignute predikcione performanse koje se sa svakim podskupom iznova procenjuju, što uključuje i kriterijume zaustavljanja. Konačno, prva dva koraka se ponavljaju, ali je potrebno odrediti i koji predikcioni model će biti korišćen.

Prema nekim od okvira klasifikacije [56] *wrapper* metode najšire se mogu klasifikovati prema načinu pretrage skupa atributa, na deterministčke i *randomized wrapper* metode. Deterministčke *wrapper* metode, koriste potpunu strategiju pretrage prostora atributa i sekvencijalne strategije odonosno heuristčke pretrage (engl. *sequential selection algorithms, heuristic search algorithms*). *Randomized* metode oslanjaju se na stohastčke pristupe pretrage.

4.3.1. ALGORITMI SA DETERMINISTIČKOM PRETRAGOM

Osnovna karakteristika algoritama koji pripadaju ovoj grupi je da se izbor atributa obavlja kroz višestruke iteracije osnovnog procesa. Po definiciji za dati skup podatka u istim eksperimentalnim uslovima, odnosno za identčnu inicijalizaciju algoritma, deterministčke strategije pretrage će uvek generisati isti podskup skupa atributa. Algoritmi iz ove grupe u zavisnosti od pristupa počinju sa praznim ili punim skupom atributa i proces dodavanja ili odbacivanja atributa se vrši dok se ne dostigne maksimalna vrednost funkcije optimizacije sa minimalnim brojem atributa.

4.3.1.1 Potpuna pretraga

Iscrpno pretraživanje podskupova atributa (engl. *complete, exhaustive search method, enumerative search method*), garantuje odabir optimalnog podskupa skupa atributa za korišćeni algoritam učenja, ali je i računski najzahtevniji. Primena ovog metoda pretrage je opravdana kada je inicijalni skup atributa $d<10$ [123], ali sa porastom broja atributa vreme izbora raste eksponencionalno, složenost algoritma je $O(2^d)$, s obziom na činjenicu da je broj kombinacija za pretragu jednak 2^d. Sve prethodno navedeno čini ovakav pristup robusnim i kod višedimezionalnih skupova podataka onemogućava njegovu primenu u praksi.

4.3.1.2 Sekvencijalne tehnike pretraživanja

Prethodno ukazane slabosti iscrpnog pretraživanja, vodile su ka definisanju različitih modela pretraživanja ulaznog skupa atributa, pri čemu su se kao veoma uspešne kod različitih tipova problema izdvojile takozvane „pohlepne“ tehnike pretraživanja poznate kao *greedy* strategije. *Greedy* tehnike pripadaju grupu algoritama sa sekvencijalnom odnosno heurističkom strategijom pretrage. Osnovni algoritmi u ovoj kategoriji su *forward selection* (pretraga unapred), *backword selection* (pretraga unazad/unatrag, eliminacija unazad) i kombinovana *forward-backward selection* (bidirekciona) tehnika pretraživanja. *Greedy* tehnike pretraživanja ne vrše proveru svih podskupova skupa atributa već odlučuju na osnovu kriterijuma izbora da li se posmatrani atribut uključuje ili isključuje iz skupa. Osnovni nedostatak *greedy* tehnika je da njihovom upotrebom može biti izabran neoptimalan podskup atributa usled zaustavljanja algoritma na lokalnom minimumu.

Tehnika *forward selection* podrazumeva da algoritam započinje pretragu praznim skupom ulaznih atributa. U svakom koraku algoritma vrši se dodavanje jednog novog atributa i proračun kriterijuma izbora. Ako posmatrani atribut poboljšava karakteristike predikcionog metoda (povećava vrednost optimizacione funkcije ili smanjuje vrednost funkcije greške) on se pridodaje inicijalnom skupu atributa i postupak se nastavlja sa svaki atribut iz ulaznog skupa atributa. Ovakvim pristupom redukuje se broj ispitivanih kombinacija na $d(d+1)/2$ u odnosu na 2^d koliko se formira korišćenjem strategije potpune pretrage, pri čemu postaje i praktično moguće razmatranje većih skupova ulaznih atributa.

Primenom ove tehnike ne mora se nužno dobiti optimalan podskup skupa atributa, pošto se algoritam može zaustaviti u nekom od lokalnih minimuma, ali sigurno se biraju relevantni atributi. S obzirom na činjenicu da su ulazni atributi retko međusobno potpuno

nezavisni, korišćenjem ove tehnike uzimaju se u razmatranje i efekti kombinovanja više atributa.

Tehnikom *backward selection* počinje se pretraga od punog skupa ulaznih atributa i vrši se uklanjanje po jednog atributa u svakom koraku i to onog čije uklanjanje iz skupa daje najmanje smanjenje performansi predikcije, odnosno maksimizuje kriterijum izbora između ostalih atributa i ciljnih vrednosti. *Backward* algoritmom izbora se takođe može generisati redundatni podskup skupova podataka. Složenost algoritma je ista kao i kod prethodnog $O(d^2)$, pošto se ponovo ispituje $d(d+1)/2$ kombinacija podskupova ulaznih atributa.

Treći algoritam pretrage *forward-backward* strategija nastala je kao pokušaj da se prevaziđe jednosmernost u pretrazi u prethodno opisanim pristupima. U tom smislu ovaj algoritam uzima u ponovno razmatranje prethodno odbačene atribute, ali i ostavlja mogućnost da budu izbačeni iz konačnog skupa atributa koji su mu već pridodati. U svakom koraku se vrši izračunavanje vrednosti kriterijuma izbora. Ukoliko je stara vrednost koju vraća kriterijum izbora veća od nove, prelazi se na razmatranje narednog atributa. U suprotnom skup ulaznih atributa se ažurira dodavanjem ili odbacivanjem tekućeg atributa i postupak se ponavlja dok se ne dođe do ažuriranja skupa atributa koji dalje ne povećavaju vrednost kiterijuma izbora. Ulazni skup pretrage može biti prazan, pun ili slučajno inicijalizovan. *Forward-backward* strategija pretrage smanjuje mogućnost izbora neoptimalnog podskupa skupa atributa, ali početna inicijalizacija početnog ulaznog skupa ne pruža garancije ni da je selektovani podskup atributa optimalan.

U prethodnim sekcijama predstavljeni su osnovni algoritmi sekvencijalnih pretraga. Kako u načelu takav vid pretrage u mnogim oblastima primene daje dobre rezulatate, i uspešno održava balans između računske zahtevnosti i kvaliteta predikcije, u literaturi se mogu naći primeri raznih modifikacija osnovnih algoritama.

U modifikacije bazičnih metoda spadaju pristup *floating search* – FS i *adaptive floating search* – AFS u obe varijante selekcije *sequential backward* i *sequential forward* prikazani u [145]. Kod *floating search* - FS algoritama uvodi se dodatni *backtracking* korak u kome se u svakom koraku zajedno sa dodavanjem atributa sa maksimalnom vrednošću kriterijuma izbora vrši i određivanje vrednosti kriterijuma izbora nakon oduzimanja iz skupa najranije unetog atributa u skup. Kod AFS pristupa se u „pokretu“ određuje broj atributa za uključivanje ili isključivanje iz skupa, na osnovu koga se dobija fleksibilnija varijanta. Odluka o uključivanju ili izuzimanju atributa u selektovani podskup zavisi od doprinosa atributa kvalitetu predikcije. Postupak se nastavlja dok se ne iscrpu svi atributi iz skupa

ulaznih atributa. U [56] je dat opširiniji prikaz radova koji predstavljaju najznačajnija unapređenja.

Sva prethodna unapređenja ipak ne garantuju mogućnost da se izbegne selekcija podskupa sa redundantnim podacima pošto se i dalje ne vrši provera korelisanosti selektovanih atributa.

4.3.2. STOHASTIČKI ALGORITMI PRETRAGE

Stohastički algoritmi pretrage vrše procene različitih podskupova atributa kako bi optimizovali vrednost optimizacione funkcije odnosno poboljšale performanse prediktora. Stohastičke strategije pretrage imaju dovoljnu konzistentnost rezultata, a pri tome su usled brže konvergencije zbog sopstvene prirode manje osetljive na *overfitting* (preprilagođavanje). Podskupovi se mogu generisati na različite načine, nasumičnom pretragom u prostoru ulaznih atributa ili kroz direktno generisanje rešenja za optimizacioni problem. Tehnike *random* pretraživanja mogu dati dovoljno dobre rezultate uz uštedu vremena procesiranja u odnosu na iscrpno pretraživanje. Najpoznatiji predstavnik ove grupe algoritama su svakako genetski algoritami, te će u narednoj sekciji biti prikazana njihova osnovna struktura.

4.3.2.1 Genetski algoritmi

Pripadaju široj klasi takozvanih populacionih metoda odnosno evolucionih algoritama koji koriste stohastičku optimizaciju. Genetski algoritmi jedino početnu populaciju biraju *random* (nasumično), u kasnijim koracima postupak odabira populacije je strogo definisan. Koraci genetskog algoritma se iterativno ponavljaju dok se ne dostigne željena ciljna vrednost, odnosno kriterijum zaustavljanja algoritma. Najpre se vrši nasumični odabir početne odnosno inicijalne populacije. U narednom koraku se vrši evaluacija *fitness* funkcije, i zatim sve dok ne dođe do zadovoljenja kriterijuma zaustavljanja vrši se odabir podskupa atributa za dalju evoluciju i manipulacije, rekombinacije i mutacije, pri čemu se svaka naredna populacija bira tako da vrednosti odabira konvergiraju željenoj ciljnoj vrednosti.

Prednosti upotrebe genetskih algoritama ogledaju se u sposobnosti da se njima uspešno obrađuju veliki skupovi atributa (primena u bioinformatici, dimenzionalnost skupa atributa $d>1000$), pri čemu sama upotreba ne zahteva nikakva *a priori* znanja o procesima koji generišu podatke, kao i samih odnosa između atributa.

Na kraju treba dodati da se upotrebom genetskih algoritama konkretno kod *wrapper* metoda ne prevazilazi njihov problem računarske zahtevnosti. Sami proračuni potrebni za

genetske algoritme su kompleksni, a zbog prirode samog algoritma može biti produženo i vreme konvergencije. Detaljni pregled istraživanje u oblasti evolucionih algoritama može se naći u radu [168].

4.4. UGRAĐENE METODE

Kako samo ime ove grupe metoda ukazuje, proces izbora atributa je kod ovih metoda ugrađen u algoritam učenja odnosno deo je samog algoritma učenja koji se primenjuje. Izbor atributa se vrši u fazi učenja modela i specifičan je u odnosu na algoritam učenja.

Izborom atributa u trening fazi, primenom metoda iz ove klase se bolje sagledavaju svojstva atributa pošto se ne vrši podela podataka na skup za validaciju. Takođe, metodi iz ove grupe su računski manje kompleksni jer ne zahtevaju ni ponovno retreniranje modela. Karakteristični predstavnici ove grupe metoda su: stabla odlučivanja CART [17], *Random forest* [16] i različite tehnike regularizacije, kao što je na primer *lasso* (*least absolute shrinkage and selection operator*) tehnika [154].

4.4.1. STABLA ODLUČIVANJA

Stabla odlučivanja (engl. *decision tree*) predstavljaju grupu prediktivnih modela koji se kreiraju iterativnim postupkom u toku koga se vrši razdvajanje skupa podataka u zavisnosti od vrednosti posmatranog atributa, odnosno u zavisnosti od značaja posmatranog atributa za sam problem klasifikacije.

Postoje mnoge varijacije osnovnog algoritma u zavisnosti od vremenske efikasnosti, prediktivne preciznosti i sposobnosti generalizacije. Najznačajniji primeri su *ID3* [125], CART i *C4.5* [124].

Kao kriterijumi za određivanje relevatnosti atributa koriste se varijante koncepta zajedničkih informacija (*ID3, CART*) ili podela atributa i procena može biti izvršena i sa ciljem kreiranja stabla sa najmanjom empirijskom greškom generalizacije. Kao rezultat procesa podele skupa atributa u mnogim slučajevima stabla odlučivanja na kraju uključuju samo određeni podskup skupa ulaznih atributa.

4.4.2. RANDOM FOREST

Random forest (slučajne šume) predstavljaju varijantu stabala odlučivanja kod kojih se vrši višestruko generisanje stabla odlučivanja bez ikakvog odsecanja i sa fiksnim brojem slučajno odabranih atributa u svakom čvoru granjanja. Konačna odluka predstavlja proizvod

glasanja svakog od stabala u algoritmu.

Prema algoritmu generisanje podataka vrši se ponovnim uzorkovanjem *bootstrapping* metodom te se na taj način poboljšava iskorišćenost podataka [16].

Random forest se može istovremeno koristiti i za ocenu važnosti atributa. Prema [16], selekcija atributa se kod slučajnih šuma vrši ili obračunom ukupne značajnosti atributa ili računanjem *Gini importance* skora.

Značajnost varijable se određuje oduzimanjem broja ispravno klasifikovanih podataka na takozvanom *out-of-bag* skupu podatka koji odgovara 1/3 veličine resemplovanog skupa sa permutovanim vrednostima atributa na broj ispravno klasifikovanih podataka sa originalnim vrednostima atributa. Prosek tako dobijenih brojeva nad svim stablima u skupu određuje *raw importance* vrednost za posmatrani atribut. Dodatnu varijantu za određivanje značajnosti podataka predstavlja i određivanje *z-score* pod pretpostavkom da ne postoji korelacija među stablima.

Gini importance skor se računa određivanjem smanjenja *Gini impurity* (Gini indeks) mere, kojom se određuje koliko dobro potencijalni atribut deli u posmatranom čvoru uzorke klase, u direktnim potomcima u odnosnu na posmatrani čvor. U odnosu na mere permutacije *Gini importance* skor se brže izračunava, a prema navodima iz literature daje slične rezultate selekcije atributa.

Algoritmi učenja stabala minimizuju funkciju gubitka dodavanjem u konačni skup samo atribute čije ispitivanje dovoljno smanjuje grešku na obučavajućem skupu.

4.5. NAPREDNE METODE IZBORA ATRIBUTA

Na kraju ovog poglavlja treba istaći da se lista metoda može proširiti metodama koje su izvan predstavljene klasifikacije a predstavljaju pravce aktuelnih istraživanja u oblasti izbora atributa. Šire sagledavanje limita metoda izbora atributa predstavljeno je u [91]. Ističu se metode koje u različitim fazama selekcije kombinuju filter i *wrapper* metode, kao u radovima [19], [34], [119] i [140]. Dalje tendencije upućuju na mogućnost kompozitnog (engl. *ensamble*) kombinovanja rezultata različitih pristupa izbora atributa [120]. Takođe, pojavljuju se i tendencije kombinovanja metoda izbora i ekstrakcije atributa kao u [10].

U praksi proces izbora atributa podrazumeva sledeće korake: generisanje podskupova skupa atributa, evaluaciju relevantnosti atributa, kriterijum zaustavljanja i validaciju rezultata [35] i [96]. U fazi generisanja podskupova oslanjamo se na strategiju pretrage kojom se dolazi do kandidata za dalju evaluaciju. Evaluacija svakog podskupa ili atributa kandidata se

vrši u odnosu na postavljeni kriterijum evaluacije, pri čemu se zadržava skup sa boljim karakteristikama. Proces pretrage i evaluacije se završava dostizanjem željenog kriterijuma zaustavljanja. U poslednjem koraku vrši se validacija selektovanog podskupa atributa na test skupu podataka.

4.6. KOMPARATIVNA ANALIZA METODA IZBORA ATRIBUTA

Postavka u okviru ovog poglavlja zahteva da se izvrši komparativna analiza metoda izbora atributa na nivou kategorija, a zatim bi se podrazumevalo i da se izvrši komparacija pripadajućih algoritama. Međutim, odmah je potrebno istaći da uprkos postojanju velikog broja radova koji se bave poređenjem i komparativnom analizom različitih metoda selekcije atributa ne postoji već dokazano univerzalno najbolji metod selekcije atributa već je uspešnost svakog pojedinačnog metoda uslovljena samim domenom predviđanja [40]. Ova uslovljenost dodatno usložnjava proces komparacije metoda uzevši u obzir da se stalno pojavljuju novi metodi izbora atributa koji upravo predstavljaju odgovore na uočene slabosti i specifične zahteve određenih domena predikcije. Takva komparacija je izrazito zahtevna i odvojena od teme ove doktorske disertacije.

U nastavku će biti pobrojane osnovne smernice klasifikacije i najčešće isticane prednosti i mane svake od prethodno tri posmatrane kategorije metoda izbora atributa kao i svojstva koja mogu biti korisna prilikom inicijalnog odlučivanja o odabiru algoritma izbora atributa.

U tabeli 4.1, preuzetoj iz [133], na osnovu prethodno prikazanih analiza predstavljene su zajedničke karakteristike i sličnosti između metoda ali i specifičnosti svakog od metoda izbora atributa.

Osnovna prednost svih filter metoda leži u njihovoj jednostavnosti. Osnovni nedostatak se ogleda u činjenici da izabrani podskup skupa atributa ne mora biti optimalan i da se može javiti redundantnost među podacima, s obzirom na činjenicu da neki od metoda ne vrše procenu međusobne korelisanosti između atributa. Upravo se u [12] i [63] obrazlažu problemi značaja koji atributi mogu imati u kontekstu kombinacije sa drugim atributima odnosno relevantnosti i redundanse među podacima. Evidentno je i da nije definisan idealan način za određivanje veličine podskupa skupa atributa.

TABELA 4.1 Prikaz svojstava tehnika izbora atributa

	PREDNOSTI	NEDOSTACI
FILTER (UNIVARIJANTNI)	• Brzi • Skalabilni • Ne zavise od klasifikatora	• Ignorišu se međuzavisnosti atributa • Nezavisnost od klaisifikatora
FILTER (VIŠEVARIJANTNI)	• Modeluju zavisnosti između atributa • Ne zavise od klasifikatora • Jednostavnija izračunavanja u odnosu na *wrapper* metode	• Sporije od univarijantnih tehnika • Manje skalabilne u odnosu na univarijantne pristupe • Nezavisnost od klaisifikatora
WRAPPER (DETERMINISTIČKA PRETRAGA)	• Jedostavni • Interaguju sa klasifikatorim • Modeluju zavisnosti između atributa	• Neotporni na preprilagođavanje • Mogućnost lokalnog optimuma • Izbor zavisi od klasifikatora
WRAPPER (*STOHASTIČKA* PRETRAGA)	• Otporni na lokalne optimume • Iteraguju sa klasifikatorim • Modeluju zavisnosti između atributa	• Računarski zahtevni • Izbor zavisi od klasifikatora • Visok rizik od preprilagođavanja
UGRAĐENE METODE	• Jednostavnija izračunavanja u odnosu na metode probnog učenja • Modeluju zavisnosti između atributa, • Interaguju sa klasifikatorima	• Izbor zavisi od klasifikatora

Uzevši u obzir zavisnost od samog predikcionog metoda, u mnogim slučajevima primene realno je očekivati bolje performanse *wrapper* metoda u odosu na filter metode, pri čemu treba imati u vidu da su *wrapper* metode vremenski zahtevnije. U praktičnoj primeni *embedded* metode su te koje uspostavljaju balans između vremena izvršavanja i kvaliteta predikcije.

Na kraju ovog dela doktorske disertacije treba istaći i neke od mogućnosti daljih istraživanja. Pre svega prema navodima iz opsežnih studija [88] i [91] i nameću se izazovi u pogledu vrste podataka, najpre problemi izbora atributa kod generičkih podataka, heterogenih podataka iz više izvora ili generisanih tokom više procesa, kao i kod dinamičkih (engl. *streaming*) podataka. Posebno se u pravcima budućih razvoja izdvaja problem izbora atributa u analizi sa izrazito velikom dimenzionalnošću *d~10000* [61], [91], [177] i [183].

Prema navodima iz opsežnih studija [40] i [96] u okviru istraživanja koja se bave komparativnom analizom nedovoljno je istraženo i pitanje teorijske i eksperimentalne provere stepena konzistenstnosti rezultata različitih metoda izbora atributa. Odnosno potrebno je potencirati zaključak iz studija [40] i [91] u kojima se ističe da „različiti metodi izbori atributa sa sličnim svojstvima u smislu preciznosti ili stabilnosti ne moraju nužno birati ni isti ni sličan skup atributa, dok sa druge strane podskupovi skupova atributa koji imaju i mnogo zajedničkih atributa ne moraju nužno dati iste predikcione rezulatate“.

4.7. PREGLED RADOVA IZ OBLASTI ISTRAŽIVANJA

U brojnim radovima koji obrađuju problem izbora atributa u oblasti predikcije trenda kretanja finansijskih vremenskih serija sa tržišta kapitala, odnosno berzanskih indeksa, ulazni atributi se selektuju na osnovu analize numeričkih vrednosti finansijskih instrumenata, uključujući vrednosti indeksa, obim trgovanja, finansijske racio brojeve i tehničke indikatore. U odnosu na prethodno prikazane metode izbora atributa mogu se naći primeri i radovi sa upotrebom gotovo svakog pristupa.

U [85], se kombinuju F-score i *Supported Sequential Forward Search* (F_SSFS), kako bi se iskoristile prednosti i filter i *wrapper* pristupa u procesu selekcije atributa radi dobijanja optimalnog podskupa atributa iz inicijalnog skupa atributa.

U [111], kao metoda izbora atributa koristi se fraktalna analiza, a zatim se odabrani atributi integrišu sa SVM metodom kako bi se poboljšala efikasnost algoritma u predviđanju smera kretanja berzanskih indeksa.

Autori Yu, Wang, i Lai u [179] koriste hibridni data mining pristup, SVM i genetski algoritam - GA, kao metodu izbora atributa.

U [66] su vršena istraživanja sa različitim metodama izbora i ekstrakcije atributa, PCA (engl. p*rincipal component analysis*), GA i *sequential forward* tehnika pretrage.

Integrisani pristup sa korišćenjem ICA (*engl. independent component analysis*) i PCA predstavljen je u [94], gde je predložena metodologija testirana predviđanju tržišnih kretanja na kineskoj berzi kapitala.

U [69] predstavljen je primer upotrebe *wrapper* pristupa i kompozicije klasifikatora u predviđanju na tržištima kapitala.

Obimna i sveobuhvatna studija i pregled literature o tehnikama predviđanja zajedno sa prikazom metoda izbora atributa na tržištima kapitala može se naći u [7].

Tek nedavno se krenulo u pravcu kvalitativne analize podataka kako bi se unapredio kvalitet predikcije. Pristupi koji se mogu naći u radovima istraživanja [51], [107], [174] i [180] koriste znanje o događajima koji utiču na procese koji generišu podatke i o samim vrednostima vremenskih serija u cilju povećanja preciznosti. U radovima [51], [107] i [180], predviđanje trenda u kretanju vrednosti berzanskih indeksa vrši se pomoću *text mining* tehnika kroz aktivni pregled novinskih izveštaja.

Takođe, novija istraživanja u oblasti predviđanja kretanja vrednosti na tržištima kapitala predstavljena u [117] upućuju na korišćenje signala trgovanja generisanih strategijama trgovanja kao atributa za obuku modela umesto do sada korišćenih pristupa koji

se zasnivaju na učenju na osnovu vrednosti tehničkih indikatora.

Prema [7] i [104] precizno predviđanje trenda na finansijskim tržištima trebalo bi da inkorporira način na koji berzanski eksperti uče i obrađuju informacije. U istim radovima navodi da se trgovanje na berzi najpreciznije definiše kao proces donošenja odluka koji je pod uticajem dinamičkih tržišnih uslova i potencijalnih rizika trgovanja. Upravo je takav pristup služio kao osnov za istraživanja predstavljena u ovoj doktorskoj disertaciji, gde se predstavlja pristup inkorporacije *a priori* znanja o domenu sa finansijskih tržišta u proces izbora atributa i samu metodu učenja.

POGLAVLJE 5

REPREZENTACIJA ZNANJA I INŽENJERING ATRIBUTA

„Korišćenje atributa u zadatoj formi nije ujedno i najbolji način korišćenja atributa“ [163]. Transformacija atributa može se postići na više načina uključujući i korišćenje znanja o domenu. Odgovarajuća reprezentacija atributa utiče na smanjenje vremena potrebnog za obuku algoritma učenja, utiče na povećanje preciznosti predikcionog modela kao i na smanjenje korelisanosti među atributima [163].

U poglavlju koje sledi biće prikazani načini koji omogućavaju bolju reprezentaciju problema algoritmima mašinskog učenja i utiču na poboljšanje kvaliteta predikcije. Po pravilu inicijalno izabrani skupovi atributa predstavljaju upravo formalizaciju prethodnog znanja o problemu predviđanja.

5.1. ZNANJE O DOMENU I MAŠINSKO UČENJE

Proces mašinskog učenja u praksi podrazumeva najpre poznavanje domena iz oblasti predikcije, dovoljno prethodnog znanja i postavljanje ciljeva, u kojima je osnovni korak razgovor sa ekspertom. Zatim idu koraci integracije podataka, izbora atributa i preprocesiranja, koji po pravilu odnose najviše vremena u celom procesu. Naredni korak je korak izbora i obuke modela, zatim sledi interpretacija rezultata i primena otkrivenih znanja. Proces je iterativan u smislu da se nastavlja sve dok se ne dobiju rezultati koji se mogu primenjivati u praksi. U praksi najveći deo uspeha algoritma mašinskog učenja zapravo predstavlja uspeh u inženjeringu atributa tako da algoritam učenja može bolje da „razume“ sam proces koji generiše atribute. Takođe, performanse algoritama mašinskog učenja zavise od načina na koji je predstavljen problem (engl. *problem representation*).

Relevantnost atributa može se razlikovati u odnosu na opseg podataka čak i u slučaju kada atribute generiše isti proces. Atributi mogu biti različite relevatnosti u odnosu na model

koji se koristi za predikciju, ali njihov značaj može da zavisi i od korišćene metrike za predstavljanje podataka. Nameće se zaključak da striktno algoritamski pristup procesu izbora atributa često nije dovoljan za razrešavanje svih skrivenih problema.

U mnogim oblastima primene data mining tehnika postoje eksperti iz oblasti koji poseduju potrebna znanja izgrađena na osnovu iskustva za rešavanje problema iz posmatranog domena. Međutim, ekspertsko znanje je u prirodi heurističko, što znači da po pravilu eksperti nisu u stanju da na adekvatan način formalizuju pravila koja koriste prilikom rešavanja problema. Iz tog razloga je prikupljanje ekspertize zahtevan i izazovan posao [144]. Sa druge strane, poznato je da heurističko znanje može pomoći pri rešavanju samo određenih problema, i da za razliku od algoritamskog rešavanja problema nije moguće garantovati da će heuristički pristup uvek dati željene rezultate.

Po navodima iz [182], znanje o domenu predviđanja se može najpre iskoristiti kako bi se kreirali atributi višeg nivoa apstrakcije koji bi potencijalno mogli unaprediti predikciju.

Upravo u prilog tome govore i rezultati eksperimenata iz [49]. U radu se ističe da je najbolji način izbora atributa zapravo korišćenje znanja o domenu i jasno razumevanje šta podaci zaista znače u realnom problemu. Dalje se navodi da je jedan od najznačajnijih pristupa izbora atributa i određivanje njihove relevatnosti upravo na osnovu ekspertskog znanja o domenu bez obzira na veliki broj istraživanja posvećenih metodama izbora atributa i brojnih algoritama i alata koji automatizuju proces selekcije. Očigledno je da adekvatna ekspertiza omogućava uvid u strukturu problema koju u realnom vremenu teško može dostići bilo koji algoritam mašinskog učenja.

U [49] istaknut je značaj koje znanje o domenu ima u procesu konstruisanja efikasnog skupa ulaznih atributa višeg nivoa koje su specifični za domen predviđanja u odnosu na neobrađene podatke iz sistema. U radu se ističe da u praksi uspešni data mining projekti prave fuziju ekspertskog znanja u vezi sa samim podacima i problemom koji je predmet posmatranja i primenjenih algoritama, pri čemu znanje o procesu koji generiše podatke upravo pomaže u pravilnoj upotrebi data mining alata.

U [182] se navodi na osnovu teorijskog pregleda radova drugih autora da generalno uspeh klasifikatora više zavisi od reprezentacije podataka nego od odabira algoritma učenja i strukture modela. U praksi je upravo reprezentacija ulaznih atributa najzahtevnije deo procesa a ujedno i zavisi od oblasti primene [182]. Suština svih transformacija je da se nađe reprezentacija i atributi koji optimalno opisuju ciljni koncept.

Nedovoljno istraživanja u pravcu kombinovanja data mining tehnika sa znanjem o domenu uslovilo je da to postane važan pravac razvoja ove oblasti, fuzija data mininga i

ekspertskih sistema zasnovanih na znanju. Najpre ćemo odrediti okvire za proučavanje problema iz ove oblasti, s obzirom na različitu terminologiju kojom se definišu slični pojmovi, navešćemo više sličnih ali u osnovi različitih tumačenja.

Pođimo dakle od najopštijih definicija. U [84] i [138] može se naći definicija po kojoj „prethodno znanje upućuje na sve informacije o problemu koje su dostupne pored trening skupa“.

U [49] navode se različite kategorije znanja o domenu i ističe se da ono uključuje i znanje o uzročno posledičnim vezama koje postoje u domenu primene. Neke tehnike mašinskog učenja kao što je *Bayesian Networks* već i same omogućavaju inkorporiranje *a priori* znanja. U istom radu se dalje navode ograničenja algoritama učenja. Većina algoritama mašinskog učenja podrazumeva da su svi podaci koji se odnose na jednu opservaciju entiteta koji je predmet prediktivne analize zapamćeni u okviru jednog zapisa, pa se analiza oslanja na tabelarni pristup podacima. Takav pristup sam po sebi uslovljava određena pojednostavljenja u karakterizaciji domena predviđanja.

U [20] se već ističu određeni nivoi inteligencije i ekspertize koji su potrebni za razumevanje problema i definiše se pojam *domain intelligence*, kao resursi iz domena predviđanja koji pomažu u razumevanju i pri rešavanju problema. Takozvana *domain intelligence* se po navodima autora sastoji od dva tipa znanja, kvalitativne i kvantitatvne inteligencije koje upućuju na aspekte kao što su poznavanje domena, osnovne informacije, ograničenja, organizacioni faktori i poslovni procesi, kao i inteligencija u okruženju.

Prema [84] i [138], bez korišćenja prethodnog znanja verovatno će postavka problema biti neakdekvatna i problem neće moći da se okarakteriše jedinstvenim modelom. Naročito ako se ima u vidu svojstvo većine klasifikatora koji se baziraju na pretpostavci sličnosti, slični trening i test podatak će uglavnom biti svrstani u istu klasu. Značaj prethodnog znanja je istaknut u [84] kroz podsećanje na „*no free lunch”* teoriju kojom se iskazuje stav da većina algoritama ima iste prosečne rezultate na problemima iz različitih oblasti i da zapravo poboljšanje performansi proističe tek iz korišćenja specijalizovanih algoritama i uključivanjem na naki način prethodnog znanja o problemu.

Jedan od tekućih pravaca razvoja algoritama mašinskog učenja upravo se odnosi na integraciju prethodnog znanja o domenu u proces učenja. Imajući u vidu da je za algoritam učenja neohodno da ima i adekvatne podatke, adekvatno znanje o domenu, kako bi mogao na ispravan način da izvede zaključke i pronađe veze o posmatranim konceptima i procesima.

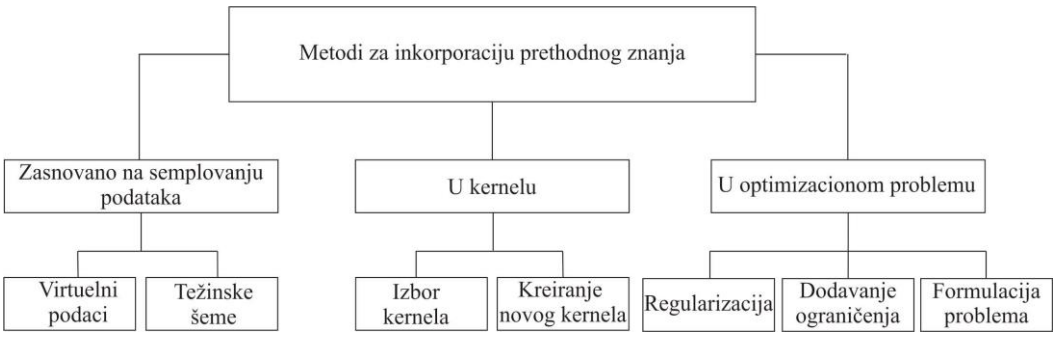
U dosadašnjim radovima takvi pristupi su ređi upravo zbog kompleksnosti postupka integracije takve vrste specifičnih znanja. Znanje koje poseduju eksperti je previše opšte i ne

može se lako iskoristiti u algoritmima učenja bez adekvatne konceptualizacije. Inkorporacija i konceptualizacija ekspertskog znanja sa druge strane bi zahtevala i da sam ekspert do određene mere razume osnovne postulate algoritama mašinskog učenja što opet može predstavljati dodatno ograničenje. U [138] se navodi da se znanje odnosi i na relacije ili regulative koje uopšte nisu prikazani u trening skupu, a mogu same po sebi uticati na pouzdanije predikcije. Bez obzira na činjenicu da ne postoje načini da se takvo znanje direktno inkorporira u algoritam učenja ono može pomoći da poveća *inductive bias* i da se uvede *domain-specific bias* u algoritam učenja [138].

Znanje o domenu u osnovi upućuje na poznavanje svojstava procesa koji generiše podatke, odnosno načina kako funkcionišu veze među podacima i predstavlja šire znanje u odnosu na sam klasifikacioni problem.

Najveći nedostatak svih pristupa inkorporiranja znanja u algoritme mašinskog učenja prema [138] ogleda se upravo u činjenici da je znanje o domenu lokalnog karaktera tako da ne postoji način za širu primenu algoritama koji bi imali za cilj da koriste ili prikupljaju određena znanja. Dodatno, znanje o domenu može biti jedino integrisano kroz određene aproksimacije pa otuda i stav prema uticaju na *bias* algoritma učenja.

U [84] koji konkretno obrađuje načine inkorporacije prethodnog znanja u metode podržavajućih vektora prikazana je sistematizacija pristupa u integraciji prethodnih znanja. Na slici 5.1 prikazana je osnovna podela pristupa inkorporacije prethodnog znanja na osnovu klasifikacije predstavljene u [84] i [138].



SLIKA 5.1 Metodi inkorporacije prethodnog znanja o domenu

Na osnovu slike 5.1, preuzete iz [84], može se videti da su metodi klasifikovani prema prikazanom pristupu u tri osnovne kategorije: metode semplovanja, kernel metode i optimizacione metode.

Prvoj kategoriji pripadaju metodi semplovanja (engl. *sample methods*). U ovakvom pristupu prethodno znanje se integriše tako što se vrši generisanje novih podataka, virtuelnih

podataka ili kroz dodavanje težina instancama, odnosno kroz modifikaciju načina na koji se podaci koriste za učenje.

Drugu grupu metoda čine kernel metode koje integrišu prethodno znanje na dva načina, kroz izbor odgovarajućeg kernela upotrebom znanja o domenu i poznavanjem transformacija koje vode do željene ciljne vrednosti ili kroz kreiranje novog kernela koji odgovara domenu predviđanja. Popularnost pristupa integracije prethodnog znanja preko kernel funkcija leži u tome što se tako kreirani kerneli mogu koristiti i za učenje kod različitih kernelizovanih metoda, SVM ili KPCA.

Treću kategoriju čini grupa optimizacionih metoda kod kojih se prethodno znanje inkorporira na tri moguća načina. Najpre uvođenjem određenih regularizacionih parametara, zatim kroz drugačiji pristup u definisanju problema posmatranja i treći pristup uključuje dodavanje novih ograničenja za posmatrani domen predikcije. Primeri [86] i [160] su konkretne implementacije optimizacionog pristupa. U [86] se inkorporacija prethodnog znanja postiže kroz modifikaciju prostora hipoteza a ne samog optimizacionog problema. Na taj način se smanjuje veličina trening skupa, dok se ujedno i poboljšavaju performanse prediktora. U [160] je obrađena problematika inkorporacije prethodnog znanja u problemima obrade slike (engl. *image retrieval*), na način da je konstruisana nova ciljna funkcija optimizacije koja se zasniva na upotrebi prethodnog znanja. Ekseperimentalni rezultati u radu ukazuju da je takvim pristupom moguće poboljšati učenje naročito kod problema kada postoji nedovoljan broj označenih podataka.

Razlika između grupa uslovljena je time koje alate koriste algoritmi kako bi integrisali prethodno znanje i u odnosu na načine na koje dolazi do izmena osobina procesa koji se posmatra.

Prema navodima iz [84] postoje dakle i druge forme prilagođavanja koje nisu obuhvaćene ovom klasifikacijom zato što su specifične i zavise od konkretne primene za konkretni domen, te su to metodi koji su proizašli iz specifičnosti aplikacija. Cilj prikazanog rada je i bio po pisanju autora da predstavi generalne metode koje se mogu koristiti kod različitih aplikacija a ne da se pruža lista specifičnih znanja sa primerima praktične inkorporacije u sistem.

U [22] prikazan je specifičan pristup za inkorporaciju znanja kod primena algoritama mašinskog učenja u poslovanju. Opisano je da u poslovanju znanje o domenu pored različitih oblika preciznih znanja, obuhvata i koncepte, verovanja, odnose i neformalne prednosti i lične preferencije ili predrasude eksperta koje se ogledaju u odabranim strategijama delovanja. U oblasti marketinga različiti marketinški pristupi upravo oslikavaju lične

preferencije eksperta. Prema navodima iz istog rada znanje o domenu može značajno doprineti uspehu data mining algoritama.

U praksi i problemima iz realnog okruženja uvek unapred postoji neka količina informacija o problemu. U [84] i [138] se navode tri karakteristična slučaja koja se po pravilu javljaju u fazi prikupljanja podataka. Prvo, mogu postojati neobeležene instance za koje se može pretpostaviti pripadnost klasi; zatim može se javiti neizbalansiranost u trening skupu podataka koja je posledica velike dominacije podataka iz jedne klase u odnosu na drugu klasu. I kao treće, postoje varijacije u kvalitetu podataka u zavisnosti od uzorka koji se uzme u razmatranje. U svakom od slučaja prethodno znanje pomaže u obradi specifičnosti i nedostataka u podacima kako bi se poboljšale performanse klasifikatora.

Na kraju ovog dela potrebno je uvesti i najširi kontekst razmatranja reprezentacije problema koji onda upućuje na problematiku iz oblasti adaptacije na domen, karakteristične kod mnogih realnih problema mašinskog učenja.

U trećem poglavlju ove doktorske disertacije već je istaknuto da se sposobost generalizacije kod većine metoda mašinskog učenja zasniva na pretpostavci da su elementi trening i test skupa međusobno nezavisne slučajne promenljive koje pripadaju istoj raspodeli, IID. Međutim, poznato je i da u podacima iz realnog okruženja IID često nije slučaj. Upravo se u kontekstu razlike u distribuciji podataka u test skupu u odnosu na distribuciju podataka u trening skupu prepoznaje problem prilagođavanja odnosno adaptacija na domen.

Karakterističan primer predstavljaju potrebe da se izvrše personalizacije spam filtera kako bi se postigle bolje performanse metoda mašinskog učenja, koji se po pravilu treniraju na nekom skupu javno dostupnih podataka odnosno spam mailova. Slični problemi postoje i kod obrade prirodnog jezika i govora (engl. *natural language procesing, domain adaptation, transfer learning*).

Iako problem kontekstualne adaptacije predstavlja osnovni problem u oblasti mašinskog učenja, on dobija veću pažnju naučne javnosti tek u poslednjoj deceniji. U [36] i [71] dat je prikaz radova iz oblasti. Po navodima iz literature upravo su određeni specijalni slučajevi iz oblasti adaptacija domena poznati od ranije, pod nazivom neizbalansiranost klasa (engl. *class imbalance*) i *covariate shift*.

Ukratko koncept *data shift* (najčešći tipovi, *covariate shift* i *concept drift*), upućuje na različite raspodele podataka ili na iste raspodele ali sa različitim parametrima u trening i test skupu. Formalno definisano, *data shift* se odnosi na tehnike, algoritme i aplikacije mašinskog učenja u okolnostima da postoji nestacionarnost u distribiciji test skupa podataka, odnosno da je distribucija trening i test skupa različita, ali da je koncept učenja stacionaran [126]. U

takvim okolnostima standarne tehnike za odabir modela više ne daju optimalne rezultate. Kod metoda koje obrađuju problem *covariate shift* karakteristično se vrše davanje težina instancama trening skupa tako da nakon ponderisanja distribucija podataka adekvatnije reprezentuje distribuciju vrednosti u test skupu. Težine se dodeljuju u fazi kreiranja modela na osnovu metrike sličnosti koja se obračunava između trening i test instance. Konstruisani su algoritmi koji obrađuju navedeni problem kao što je metoda iz [150], za koji su autori pokazali da i u uslovima *covariate shift* ne podleže povećanju *bias-a.*

Detaljnije o pristupima rešavanja ove grupe problema može se naći u [36], [71], [116] i [126].

Imajući u vidu da se u okviru ove doktorske disertacije razmatra upotreba domenskog znanja u procesu izbora atributa, dva slučaja reprezentacije prethodnog znanja zavređuju posebnu pažnju i biće izložena u narednim sekcijama. Prvi je problem dodele težina atributima kako bi im se pridružila relativna važnost kod reprezentacije podataka i drugi je problem određivanja relevantnosti trening vektora koji će se koristiti u postupku učenja, kroz označavanje vektora.

5.2. OZNAČAVANJE VEKTORA

Jedan od osnovnih pristupa za adresiranje problema reprezentacije znanja je da se pridruže težine trening vektorima kako bi se minimizovali očekivani gubici usled distribucije podataka [71].

Davanje težina instancama pomaže da se integrišu u algoritam učenja specifične forme znanja o podacima koje se odnose na neizbalansiranost klasa, relativan kvalitet trening instanci ili ako postoji znanje o neobeleženim instancama.

Najjednostavniji način da se prevaziđe problem neizbalansiranosti klasa je resemplovanje podataka. Problem pretpostavlja da uobičajeni pristup semplovanja ne doprinosi dobijanju dodatnih informacija iz podataka, a može se i izgubiti nešto od korisnih informacije ako se vrši *undersampling* klase iz koje preovladavaju podaci. Zato postoji i drugi pristup rešavanja problema neizbalansiranosti klasa. Kod takvog pristupa vrši se dodeljivanje težina vektorima u skladu sa njihovim značajem u okviru problema koji se posmatra. Veća težina upućuje na veći značaj.

Među načinima inkorporacija znanja o domenu izdvaja se integracija znanja o *class-invariance,* koja karakteriše nepromenjivost klase predikcije na transformacije ulaznih podataka, označena sa *transformation-invariance* [138]. *Class invariance* se odnosi na

nepromenljivost podatka u odnosu na transformacije i permutacije u ulaznom prostoru podataka. Prvi pristup se odnosi na generisanje virtuelnih uzoraka i fokusira se na inkorporaciju invarijanse u odnosu na transformaciju, dok se drugi odnosi na određivanja relevatnosti različitih uzoraka koji se koristi kako bi se integrisalo znanje o podacima. Kao primer može da posluži prepoznavanje brojeva, gde treba uzeti u obzir da i rotacije i translacije karaktera i dalje predstavljaju isti karakter, pa bi prilikom kreiranja predikcionog modela trebalo postići da klasifikator bude invarijantan na takve promene. Virtuelne instance, proizašle su iz stava da veći skup ulaznih podataka pomaže pri postizanju boljih performasi bez obzira na kvalitet samog skupa.

Drugi pristup integracije znanja o podacima se odnosi na integraciju znanja o neobeleženim podacima, kvalitetu podataka ili neizbalansiranosti klasa u trening skupu podataka.

Za postojeći skup podataka vrši se inicijalizacija težina za svaki vektor iz skupa trening vektora, na osnovu najveće posterior verovatnoće za posmatrani trening vektor [48]. U ovakvom pristupu prednjače stabla odlučivanja, kod kojih je moguće izvršiti dve vrste dodeljivanja težina. U osnovi se nalazi ideja da se svim vektorima odredi značaj u zavisnosti od njihovog uticaja kod obučavanja modela.

Označavanje vektora se vrši na način da se težina dodeljuje svakom pojedinačnom zapisu u skupu trening vektora i da se veća vrednost dodeljuje vektorima koji su značajniji u odnosu na ostatak skupa.

Što se tiče samih metoda podržavajućih vektora koji se koriste u eksperimentalnom delu ove doktorske disertacije već je odmah nakon promovisanja algoritama prema navodima u [84] koji se oslanjaju na zapažanja iz rada autora LeCun, 1995. godine, istaknuto da metode imaju dobre generalizacione osobine bez obzira na činjenicu da ne vrše inkorporaciju prethodnog znanja. Odnosno ukazano je na svojstvo algoritama da bi proizveo zadovoljavajuće rezultate na primeru problema prepoznavanja slika i kada bi došlo do permutacije piksela po nekoj predefinisanoj šemi. Upravo se uvođenjem kernel funkcije koja se iz ranijih razmatranja može smatrati merom sličnosti pokušava da izvrši inkorporacija prethodnog znanja, koristeći *a priori* znanje prilikom definisanja same kernel funkcije. U prilog tome u [137] je istaknut način inkorporacije prethodnog znanja kroz kreiranje adekvatnog kernela u problemima prepoznavanja slike što dovodi do značajnih poboljšanja u pogledu performansi algoritma i ukazuje na prepoznavanje problema inkorporacije znanja. Isto stanovište se prepoznaje i u [50] u kome je istaknuto da se upravo nedovoljan broj radova bavi mogućnošću inkorporacije prethodnog znanja u SVM za razliku od neuronskih mreža.

U [166] prikazan je pristup koji omogućava inkorporaciju znanja kod problema prepoznavanja teksta kod SVM metoda učenja tako da su u stanju da sa određenom verovatnoćom odrede pripadnost klasi za neoznačene trening instance. Algoritam je u stanju da vrši predviđanja i kod potpuno neoznačenih klasa, s obzirom na osobinu da se svakoj trening instanci pridružuje nivo pouzdanosti čija je vrednost između *0.0* i *1.0* i koji meri kolika je verovatoća da trening instanca pripada baš određenoj klasi. Prethodno znanje se uključuje kroz generisanje pseudo trening skupa čije su klase određene prema verovatnoći stepena pripadnosti nekoj od klasa.

U grupu algoritama koji koriste pristup označavanja vektora spadaju i *Weighted-SVM* (WSVM) [171] kao i *Weighted-LS-SVM* (WLS-SVM) [151] koji predlažu različite težine koje se pridružuju svakom od uzoraka u odnosu na relevatnost koja se određuje ili na osnovu prethodnog znanja ili kroz algoritamske procedure za označavanje podataka.

Prema [84] upravo je integracija prethodnog znanja u SVM metod prednost koja omogućava algoritmu da postigne poboljšanja u mnogim oblastima primene.

U pogledu novijih istraživanja, u [59], se ispituju uticaj inkorporacije prethodnog znanja u srednji sloj *deep supervised neural networks*, sa takođe uočenim pozitivnim efektima. U [32] znanje o domenu je korišćeno kako bi se dobile relevatnije informacije iz biosignala i ispostavilo se da su takve informacije od suštinske važnosti za postizanje boljih rezultata u kliničkim predviđanjima. U radu se generišu i vrši se simulacija kompleksnih mreža koje su analogne sa biološkim sistemom čoveka a zatim se podaci iz simulacija koriste kako bi se obučavali algoritmi mašinskog učenja koji bi predviđali ponašanje mreže. Poređene su performanse algoritma nakon što su obučavane korišćenjem znanja o procesima koji objašnjavaju funkcionisanje kompleksnih mreža i dokumentovana su poboljšanja.

Izbor vektora je specijalan slučaj dodeljivanja težina vektorima. U [76] je istaknuto da *instance-based* metodi postižu značajne rezultate u mnogim praktičnim primerima pristupa učenja koji je zavistan od trening vektora. Osnovna ideja u ovakvom pristupu je da se za obučavanje predikcionog modela koriste vektori koji su prema odabranoj metrici najbliži svojstvima zadatog test vektora. Dakle, prema [2] *instance-based* algoritmi učenja zasnivaju se na pamćenju svih dostupnih trening podataka ili neke od selekcija i ne vrši se konstrukcija globalne reprezentacije ciljne funkcije, već se na osnovu test instance koju je potrebno klasifikovati lokalno kreira predikcioni model za koji se trening primeri biraju tako da pripadaju okruženju vrednosti test primera. Jedan takav pristup zasnovan na metodi zajedničkih informacija može se naći u [148] i [149]. U [25] predstavljen je algoritam kod koga je izbor vektora za treniranje predikcionog modela zasnovan na konceptu *k-NN* i

zajedničkih informacija. Poboljšanje predikcionih performansi postiže se kroz modifikaciju LS-SVM metoda na osnovu trening instanci.

5.3. DODELJIVANJE TEŽINA ATRIBUTIMA

Fokus četvrtog poglavlja disertacije bio je na analizi algoritama koji se koriste kako bi se eksplicitno odredili najznačajniji atributi odnosno podskup atributa. Takođe, u svim prethodnim radovima rešenja i optimizacije se postižu na osnovu informacija koje se dobijaju iz procesa koji generišu same vektore. U [159] prikazana je podela na prostor vektora i prostor atributa i upravo nastojanja da se optimizacije postižu zahvaljujući informacijama iz prostora obuhvaćenog atributima. U ovom pogavlju razmotrićemo dva značajna koncepta iz postora atributa, određivanje težina atributima i inženjering atributa.

Inženjering atributa je oblast reprezentacije problema čiji je cilj da se nađe najbolja moguća reprezentacija podataka kako bi se što uspešnije našlo rešenje problema. Pojam inženjeringa atributa još uvek nije potpuno formalno definisan termin, ali jeste termin koji se često sreće kod praktične upotrebe algoritama mašinskog učenja i odnosi se na „proces transformacije neobrađenih podataka u atribute koji bolje predstavaljaju problem iz domena predviđanja prediktivnim modelima i vode do unapređenja kvaliteta predikcije“ [81]. Inženjerig atributa u osnovi predstavlja proces korišćenja znanja o domenu kako bi se konstruisali atributi koji će omogućiti kvalitetniju obuku predikcionog metoda. Kako bi se efikasno izvršio inženjering atributa potrebno je najpre razumeti svojstva problema koji se posmatra, a zatim i način na koji ti atributi mogu interagovati sa modelima predikcije. Sama reprezentacija problema čini značajan deo procesa i fazu učenja o strukturi problema koji je potrebno modelovati. Izbor i inženjering aributa se međusobno dopunjuju i nisu uzajamno isključivi [18] i [42].

Neformalno definisano inženjering atributa je oblast koja se odnosi na dobijanje više informacija iz postojećih skupova podataka, ne radi se o dodavanju novih podataka već o načinu da se postojeći podaci učine smislenijim za algoritam mašinskog učenja [42]. Proces se može izvoditi u dva osnovna koraka. Prvi korak odnosi se na transformaciju atributa, dok se drugi korak odnosi na kreiranje atributa koji mogu doprineti unapređenju performansi predikcionog modela.

U kontekstu mašinskog učenja inženjering atributa odnosi se i na „ručno“ konstruisanje ulaznih atributa. Konstrukcija ili inženjering atributa predstavlja proces „ručnog“ kreiranja novih promenljivih i podrazumeva korišćenje znanja o domenu kako bi se

kreirali novi atributi koji su sposobni da unaprede predikcione performanse modela. Konstrukcija atributa je u osnovi proces transformacije podataka.

Pojam koji se izdvaja u oblasti inženjeringa podataka je i *handcrafted features*, odnosno atributi koji su rezultat formalizovanog znanja do kojeg se došlo tokom postupka definisanja atributa. Takvo znanje se može prikupljati na osnovu više skupova podatka koji predstavljaju isti problem mašinskog učenja [184].

Inženjering atributa je proces koji se u praktičnoj primeni algoritama mašinskog učenja i dalje oslanja na eksperte u celom procesu mašinskog učenja tako da ujedno predstavlja i izazov za dalja istraživanja u smislu kreiranja automatizovanog sistema kako bi se dizajnirao potpuno autonoman računarski sistem [72].

U [182] se obrađuje problem iz domena predviđanja bankrota banaka u kojima su atributi učenja konstruisani iz različitih bankovnih varijabli u obliku finansijskih racio brojeva (engl. *financial ratio*). Empirijski su proverene performanse klasifikacije kada se vrši učenje algoritama na osnovu neobrađenih ulaznih podataka i nakon upotrebe ekspertskog znanja u konstruisanju atributa. Dobijeni rezultati ukazuju na značajno poboljšanje performansi koje se postiže naknadnim konstruisanjem atributa korišćenjem znanja o domenu. Evaluacija je izvršena za četiri data mining algoritma: *logistic regression*, *decision tree*, *ANN*, i *k-NN*. U zavisnosti od korišćenog metoda varira stepen poboljšanja, ali je određeno poboljšanje prisutno kod svakog od metoda.

Sve češće se mogu naći i istraživanja koje upućuju na kombinaciju ekspertskog znanja i automatozovanih pristupa kao u [109] čiji eksperimenti pokazuju da ekspertsko znanje drastično unapređuje data mining algoritme za izbor atributa. Ekspertsko znanje koristi se dvostruko ili za konstruisanje novih atributa ili za dobijanje novih korisnih infomacija koje adekvatnije reprezentuju problem posmatranja.

U odnosu na izbor i inženjering atributa, u smislu inkorporacije znanja iz prostora atributa ističe se i pristup dodeljivanja težina atributima prema njihovoj relevatnosti.

Kako upućuju prethodna razmatranja zajednička osobina dostupnih podataka iz realnog okruženja je da su njihove različite karakteristike u manjoj ili većoj meri značajne za razumevanje posmatranog problema. Bez obzira na tu činjenicu, većina metoda mašinskog učenja podrazumeva da svi ulazni atributi za obuku modela imaju istu relevantnost.

U nedavno izvedenim studijama, davanje težina atributima postaje važan pristup, pre svega kod algoritama klasterovanja. Takođe, dodeljivanje težina atributima karakteristično je i za ugrađene metode kod kojih se atributima pridružuje težinska funkcija kojom se određuje težina atributa.

Postupak dodeljivanja težina atributima suštinski vrši preslikavanje uticaja samog atributa na njegovu vrednost sa idejom je da se svakom atributu prema njegovom značaju dodeli i odgovarajuća težina.

Težine atributa ne moraju se dodeljivati globalno i biti konstantne u celom skupu podataka već mogu da variraju lokalno [163]. Lokalne težine mogu se dodeljivati atributima na osnovu dva principa: prvi je da težine variraju kao funkcija vrednosti atributa, a drugi je da težine variraju na nivou trening vektora kao funkcija distribucije vrednosti atributa. Kod prethodno navedenih pristupa postoje dve uočene slabosti: lokalne težine su osetljive na šum u trening podacima i različite lokalne funkcije mogu izgubiti korisne infomacije iz atributa. Treća kategorija dodavanja težina atributima je zapravo korišćenje znanja o domenu u cilju izbora i određivanja značaja atributa.

Zaključci iz [76] upućuju da se dodeljivanjem težina sigurno poboljšavaju performanse *nearest neighbor* prediktora, za dodeljivanje težina iz skupa vrednosti $[0, 1]$, dok kod većeg skupa vrednosti težina treba uzeti u obzir da se može pozitivno uticati na performanse algoritma učenja odnosno smanjiti *bias* kod algoritma učenja, ali sa druge strane može se povećati *variance* i preprilagođavanje. Zbog toga je preporuka istraživača da se ne dodeljuju težine svim atributima već samo onima koji su informativni u smislu različitih karakteristika atributa. U konkretnoj oblasti atributi se sagledavaju u smeru opšte značajnosti atributa, ali i u smislu osobenosti karakteristika koje predstavljaju. Stoga je preporuka da treba istražiti pod kojim uslovima su koji atributi značajni.

Davanje težina atributim je tehnika preprocesiranja atributa i kod klasifikacije teksta koja bitno utiče i na poboljšanje indeksiranja i na bolju klasifikaciju. U oblasti prepoznavanja teksta karakteristično je da je za klasifikaciju dokumenta potrebno odabrati mali skup karakterističnih atributa (reči), inače može doći i do urušavanja perfomansi prediktora. Preporuka je u ovoj oblasti da se nauče težine pri čemu veća težina upućuje na veći značaj atributa. Ovakvi algoritmi se u praksi mogu naći kao algoritmi podešavanja težina odnosno tehnike prilagođavanja težina (engl. *feature weight adjustment, weight adjustment techniques*).

U svakoj šemi dodeljivanja težina potrebno je odrediti se prema sledećim stavkama, [163]: najpre potrebno je težinu dodeliti u skladu sa sposobnošću atributa da vrši razdvajanje između klasa i drugo, težine atributa moraju biti tako podešene da preslikaju značaj koji atribut ima u razdvajanju klasa. U slučaju izbora atributa nakon primene određene šeme za određivanje težine atributima potrebno je odrediti određeni prag odsecanja.

Mnogi pristupi dodeljivanja težina, uključujući i *perceptron updating rule*, oslanjaju se na metod gradijentnog spusta (engl.*gradient descen*t), te se dodeljivanje težina svodi na više uzastopnih prolaza kroz trening skup pri čemu dolazi do promena težina svih atributa u svakoj iteraciji. Takođe, algoritam RELIEF-F, prethodno opisan u četvrtom poglavlju ove doktorske disertacije, predstavlja jedan od načina da se dodeljuju težine atributima.

Za razliku od selekcije atributa korišćenje težinskih šema omogućava klasifikatoru da parcijalno razmotri atribute dajući im određeni stepen važnosti u klasifikacionom zadatku. U suštini davanje težina atributima može se shvatiti i kao binarna varijanta metoda izbora atributa kod kojih je se atributima dodeljuju težine iz skupa $[0, 1]$. Međutim, potrebno je istaći različitu motivaciju u ovim pristupima. Prema [12] selekcija atributa se vrši kada je namera da se dobijeni rezultati koriste od strane drugog algoritma ili postoji potreba da budu razumljivi ljudima. Sa druge strane težinske šeme su u osnovi motivisane povećanjem efikasnosti i lakše se implementiraju u *on-line* sistemima. Takođe, dodeljivanje težina atributima se preporučuje kada je potrebno da se uzmu u razmatranje korisni atributi koji su manjeg značaja umesto da se kao kod pristupa izbora atributa takav atribut potpuno odbaci [134].

Težinske šeme su već duže vreme u fokusu istraživanja kod *k-NN* klasifikatora. U konkretnom metodu cilj dodeljivanja težina atributima je da se smanji osetljivost *k-NN* na redundantne, irelevantne atribute i šum, što se postiže modifikacijom *similarity function* uključivanjem težina atributa [134]. Veća težina upućuje na veći značaj koji se dodeljuje konkretnom atributu prilikom klasifikacije određenog primera Kod *k-NN* klasifikatora pravilni odabir težinske šeme atributa može imati presudni doprinos u unapređenju performansi [37] i [134].

Sve prethodno navedeno su upravo razlozi koji ukazuju na prednosti korišćenja težinskih šema kako bi se istakli najznačajniji atributi za posmatrani domen problema sa ciljem da se kroz odabir odgovarajućih težinskih šema poboljša preciznost prediktora. Kada se vrši poređenje sa pristupom dodeljivanja jednakih težina svim atributima, nedostaci težinskih šema se ogledaju jedino u činjenici da sa porastom atributa koje treba ponderisati raste i kompleksnost proračuna, koji sa razvojem računarske tehnologije sve više gubi na značaju.

U [159] se nasuprot pristupu dobijanja informacija iz prostora trening instanci upravo obrađuju načini da se inkorporira konkretno u SVM klasifikator znanje o informacijama sadrženim u samim atributima, odnosno iz prostora atributa. U predloženom algoritmu težine

atributima se određuju na osnovu metrike koja je izabrana da odredi diskriminativnu vrednost svakog od atributa u odnosu na posmatranu klasu. Rad je značajan upravo po tome što se dodatno znanje dobija iz samih atributa a ne kroz unapređivanje načina da se prethodno poznate informacije inkorporiraju u algoritam učenja.

U [80] predložen je način za inkorporaciju prethodnog znanja tako što bi se kreirali meta-atributi pored atributa koji predstavljaju podatke iz procesa koji se obrađuje. Primer obrađuje problemu prepoznavanja slika, gde se kao meta-atributi mogu koristiti pozicije piksela, dok su sami pikseli atributi. U metodološkom pristupu koji se predlaže, težine se dodeljuju svakom atributu na osnovu vrednosti meta-atributa. Eksperimentalni rezultati upućuju na to da se na taj način poboljšavaju generelazicione sposobnosti algoritma.

Kako bi se povećao efekat značajnijih atributa kod algoritama učenja u [57] uveden je kriterijum procene zajedničkih informacija na osnovu koga se dodeljuju težine atributima kako bi se odredila njihova relevantnost za specifičan zadatak.

Autori u [60] predlažu spektralnu težinsku fukciju jezgra, koja se preslikava na prostor atributa kroz dodeljivanja težina, kao način da se inkorporira teoretsko znanje o neuniformnoj distribuciji vrednosti u metode mašinskog učenja.

U [27] predstavljen je novi medicinski sistem za dijagnostiku koji je zanovan na *fuzzy* logici za dodeljivanje težina i preprocesiranju atributa u kombinaciji sa LS-SVM metodom. Dobijeni rezultati na standardnom skupu podataka za oboljenje jetre iz *UCI Machine Learning Repository* (*BUPA liver disorders dataset*) ukazuju da se na ovaj način dobija najveći stepen predikcije i da se sitem može koristiti u svrhu medicinske dijagnostike.

Za dalji pregled ove oblasti može se pogledati prikaz značajnih radova iz [159].

Na kraju ovog poglavlja treba istaći dodatnu aktualizaciju problema inkorporacije i reprezentacije znanja u okviru stručne javnosti kroz radove [156] i [157], autora Vapnik *et al*., (2009) i Vapnik *et al.*, (2015), na uspostavljanju napredne paradigme učenja kroz takozvane privilegovane informacije, dostupne u fazi obučavanja modela ali ne i u fazi testiranja i ulozi takozvanog učitelja koji je dobio adekvatne formalizacije u kontekstu mašinskog učenja.

Na osnovu svega izloženog može se zaključiti da izbor predikcionog modela i određivanje njegovih parametara u mnogome zavise od poznavanja osnovnih svojstava samog problema predikcije. Na kraju važi i obrnuto, proces transformacije i izbora atributa može na osnovu tipa i vrednosti selektovanih atributa pružiti korisne informacije o samoj strukturi problema koji se posmatra.

POGLAVLJE 6

INTEGRACIJA MAŠINSKOG UČENJA I METODA ODLUČIVANJA

Aktuelna istraživanja u oblasti mašinskog učenja upućuju na korišćenje optimizacionih metoda kako bi se došlo do rešenja za data mining probleme ili na korišćenja data mining alata kako bi se rešili određeni optimizacioni problemi. U narednom poglavlju upravo će biti prikazan presek ove dve oblasti i kako jedna utiče na drugu, sa posebnim osvrtom na integraciju algoritama mašinskog učenja i metoda odlučivanja.

6.1. PREGLED LITERATURE

U poslednje vreme sve je veći broj radova u kojima se pažnja naučne i stručne javnosti posvećuje načinima sinergije između oblasti operacionih istraživanja u čijoj osnovi su optimizacione tehnike i oblasti data mininga [28], [105] i [113]. U [105] se ističe da je proces obostran odnosno da svaka oblast može doprineti drugoj. Prema navodima iz istog rada optimizacione tehnike mogu doprineti u rešavanju sledećih data mining problema: najpre mogu uticati na povećanje efikasnosti data mining tehnika, zatim upotrebom optimizacionih metoda mogu se poboljšati rešenja iz mnogih oblasti data mining problema koji zahtevaju veću fleksibilnost i manje rigorozna ograničenja, i na kraju dobra sinergija upućuje na dalja komplementarna istraživanja i upotrebe tehnika.

Upravo sam SVM i njegova formulacija kao optimizacionog problema u [113] predstavlja najznačajniji spoj optimizacionih tehnika (matematičko programiranje) i data miniga. Ostale konkretne oblasti interakcije uključuju upotrebe različitih metaheuristika kod rešavanja kombinatornih problema i upotreba optimizacionih tehnika kod klasifikacije. Optimizacione tehnike su značajne i kod vizualizacije podataka.

U okviru ove doktorske disertacije razmatra se pitanje izbora atributa, koje se iz ugla optimizacionih tehnika može posmatrati kao kombinatorni optimizacioni problem čiji je cilj

izbor najboljeg mogućeg podskupa skupa atributa, dok se neizabrani atributi odbacuju. Predstavimo skup raspoloživih atributa sa $\mathrm{x} = [\mathrm{x}_1, \mathrm{x}_2, ..., \mathrm{x}_\mathrm{d}]$, gde *d* prema notacija iz drugog poglavlja doktorske disertacije, predstavlja ukupan broj atributa dostupan u procesu selekcije. Selektovani skup može se predstaviti kao skup binarnih vrednosti, gde svaki atribut može imati vrednost *1* u slučaju da je izabran ili 0 ako nije izabran u procesu selekcije, što se prema [113] može predstaviti i na sledeći način:

$$x_i = \begin{cases} 1 & ako\ je\ i-ti\ atribut\ izabran \\ 0 & ako\ i-ti\ atribut\ nije\ izabran \end{cases} \tag{6.1}$$

za *i=1,2,...d*. Optimizacioni problem predstavlja pronalaženje minimalne vrednosti optimizacione funkcije, $\min f(x)$ tako da budu ispunjeni sledeći zahtevi:

$$K_{\min} \le \sum_{i=1}^{m} x_i \le K_{\max} \tag{6.2}$$

gde $x_i \in \{0,1\}$ i K_{min} predstavlja najmanji broj atributa koji može biti izabran u procesu izbora atributa, dok K_{max} predstavlja najveći broj atributa koji mogu biti selektovani u procesu izbora atributa.

U procesu optimizacije glavni cilj predstavlja selekcija optimizacione funkcije $f(x)$, čiji je izbor zavistan i od cilja predviđanja i od samog domena predviđanja.

Kako je već navedeno u trećem poglavlju ove doktorske disertacije ne postoji jedinstven metod evaluacije atributa koji garantovano može dati optimalna rešenja u svakom posmatranom domenu.

U [113] i [169] se navodi da se u procesu izbora atributa upotrebom optimizacionih tehnika često mogu dobiti rezultati koji su lakši za interpretaciju i dalje korišćenje. Primer se može naći u [178] gde se poznate metaheuristike *tabu* pretraga, *simulated annealing* i genetski algoritmi koriste u procesu selekcije atributa. U [170] autori unapređuju svoj pristup izbora atributa predstavljanjem metaheuristke adaptivnog *nested partitions* methoda za rešavanje oba prethodno navedena optimizaciona problema i na taj način postižu zapažena unapređenja na posmatranom test skupu podataka.

Primeri integracija metoda odlučivanja i tehnika mašinskog učenja sa ciljem unapređenja performansi se mogu naći i u [1] gde je predstavljena tehnika koja kombinuje TOPSIS (engl. *technique for order performance by similarity to ideal solution metod*) i *F-score* metode kako bi se izabrao podskup relevatnih gena u problemima klasifikacije kancera. Rezultati kombinovane tehnike su iskorišćeni za učenje različitih hibridnih predikcionih modela, *k-NN*, *decision tree*, SVM i *Naive Bayes*. Eksperimentalni rezulatati upućuju na

poboljšanje preciznosti klasifikatora sa ovako odabranim atributima.

U [143] TOPSIS je korišćen za rangiranje tehnika izbora atributa, kako bi omogućio izbor najadekvatnijeg podskupa atributa koji su dobijeni upotrebom razlučitih metoda izbora atributa i kako bi se upravo rešio problem odabira optimalne metode izbora za dati domen, u konkretnom radu kod *intrusion detection systems* - IDS. Evaluacija rezultata je izvršena na skupu deset različitih tehnika izbora prilikom analize KDD-CUP 99 skupa podataka (*UCI knowledge discovery in databases archive*).

Na kraju može se istaći i pristup iz [115], gde je već sam TOPSIS metod mimo svojih osnovnih karakteristika u rangiranju alternativa, iskorišćen kao klasifikator u cilju predviđanja bankrota banaka. Empirijski rezultati ukazuju na jako dobre performanse i na *in-sample* i na *out-of-sample* skupu podataka, sa značajnim uticajem metoda u modelovanju i analizi rizika. Korišćenje TOPSIS metoda kao neparametarskog klasifikatora se preporučuje kao konkuretni predikcioni pristup u bankarstvu i investicijama. Performanse su testirane na *UK dataset of bankrupt and non-bankrupt firms* koje se se nalazili na listi *London Stock Exchange* (LSE) tokom 2010–2014. godine.

U [99] dat je prikaz istraživanja iz oblasti višekriterijumskog odlučivanja u periodu od 2010-2013. godine, sa više od 390 razmatranih radova, razvrstanih po različitim kategorijama. U ovom prikazu se kao individualni model na prvom mestu po upotrebi nalazi Analitički hijerarhijski proces, (engl. *Analytic hierarchy process* - AHP) dok su hibridni MCDM (engl. *multiple criteria decision-making*) modeli na drugom mestu po upotrebi kod integrisanih sistema. Najznačnije oblasti primene metoda višekriterijumskog odlučivanja su po navodima iz rada, inženjerstvo, zaštita životne sredine i održivi razvoj. AHP metod se koristi i u softverskom inženjerstvu kod odabira softverskih paketa ili kod izbora komponenata i u prioritizaciji zahteva kupaca. Dakle, Analitički hijerarhijski proces koga je predložio Saaty 1980. godine [132] ima višestruku primenu u višekriterijumskom odlučivanju gde je neophodno adekvatno izvršiti evaluaciju i rangiranje alternativa.

Do sada je AHP korišćen u više istraživanja u kombinaciji sa algoritmima mašinskog učenja pre svega kod algoritama klasterovanja.

U [162] predstavljena je nova hibridna šema za segmentaciju kupaca prema preferencijama ka određenim brendovima, kako bi se odredile pravilno investicione šeme i ostvarila prednost u odnosu na konkurenciju. AHP se koristi kako bi se izvršila prioritizacija i određivanje kriterijuma prilikom investiranja u promotivne delatnosti i kako bi se ustanovio novi indeks evaluacije parametara. Zatim se koristi *K-means* algoritam klasterovanja kako bi se na osnovu rezultata AHP analize izvršila podela po brendovima. Dobijeni rezultati se

mogu koristiti prilikom kreiranja marketinških strategija.

U [128] AHP i *k-means* algoritam klasterovanja su korišćeni kako bi se izvršilo klasterovanje i rangiranje glavnih univerzitetskih predmeta. U prikazanom pristupu najpre je izvršeno klasterovanje univerziteta prema sličnostima i razlikama, a zatim je izvršeno rangiranje posmatranih univerziteta na osnovu devet različitih kriterijuma.

U [92] predstavljen je novi pristup za preporuku proizvoda na osnovu integracije grupnog odlučivanja kroz AHP metodu i tehnike klasterovanja. U sistemima preporuke, određivanje vrednosti potrošača (engl. *customer lifetime value* - CLV) se vrši korišćenjem RFM metoda (*recency, frequency, monetary* - RFM). AHP metod se koristi kako bi se predstavile relativne težine RFM varijabli, a zatim je izvedeno klasterovanje kako bi se prema RFM kriterijumima izvršila segmentacija potrošača. Na kraju su korišćena asocijativna pravila koja obezbeđuju preporuku proizvoda za svaku grupu potrošača. Eksperimentalni rezultati potvrđuju da ovakav pristup daje bolje performanse u odnosu na kolaborativno filtriranje ili RFM sa podjednakom težinom varijabli.

U [11] *k-NN* algoritam integrisan je sa AHP metodom zasnovanoj na *Granger causality* vrednostima kako bi se poboljšao kvalitet predikcije, sa naročitim uspehom kod problema prepoznavanja rukopisa i lica. Test *Granger causality* se sprovodi u cilju određivanja preferencija svakog od kriterijuma AHP evaluacije. Zatim se AHP metodom vrši određivanje težina za različite atribute u odnosu na dva posmatrana kriterijuma i njihove prethodno obračunate relativne odnose. U poslednjem koraku dobijene težine se koriste kako bi se konstruisala *weighted distance function* za *k-NN* klasifikator. Eksperimenti izvedeni nad 15 skupova podataka iz *UCI Machine Learning Repository* dokazuju uspešnost ovakvog pristupa.

U [54] predstavljen je hibridni sistem zasnovan na AHP i neuronskim mrežama za predviđanja prinosa na kratkoračna ulaganja u hartije od vrednosti koje čine berzanski indeks sa grčkog tržišta kapitala. U predloženom sistemu najpre je obučavana neuronska mreža na osnovu stohastičkih varijabli i pokretnih proseka koji su računati na osnovu dostupnih podataka o vrednosti indeksa iz prethodnog vremenskog perioda. Rezultati neuronske mreže predstavljaju predviđene visine prinosa u ciljanom intervalu od pet nedelja. U narednom koraku, kvalitativni faktori koji mogu uticati na formiranje cena zajedno sa dobijenim predikcijama integrisani su u AHP model. Konačni rezultat obrade sistema je najverovatniji interval u kome se može očekivati kretanje prinosa u posmatranom periodu.

U [44] predstavljen je pristup zasnovan na neuronskim mrežama i AHP metodi, koji unapređuje detekciju različitih tipova ruda. Ekspertsko znanje je korišćeno radi unapređenja

kvaliteta atributa i kod određivanja težina.

U [82] predstavljen je sistem u kome se najpre primenom AHP vrši upoređivanje kriterijuma za procenu performasi različitih vendora, a zatim se rezultati poređenja i težine prosleđuju *NN* algoritmu za inicijalizaciju težina u skrivenim slojevima mreže kako bi se unapredila selekcija.

U [103] predložen je model neuronske mreže koji bi omogućio izvođenje proces odlučivanja u skladu sa AHP metodologijom. Analizom rezultata uočeno je da model zanovan na neuronskim mrežama uspeva da održi validnim proces odlučivanja i u okolnostima kada nisu dostupne sve informacije potrebne za donošenje odluke. Validnost modela je dalje potvrđena kroz različite simulacije.

U nekim primerima ukazuje se na integraciju u smislu konstriusanja kompozitnih modela kako bi se unapredile performanse u posmatranim domenima. Takav primer imamo u [90] gde je prikazan pristup integracije *fuzzy* AHP i SVM metoda u slučaju izbora 3PL logističkih provajdera, (engl. *third-party logistics provider* - 3PL) koji predstavlja komplesan i nelinearan problem odlučivanja. U konkretnom radu SVM se koristi kod izbora i evaluacije logistike a zatim *fuzzy* AHP kod odlučivanja na osnovu dobijenih informacija.

U [93] predložen je hibridni algoritam za određivanje značajnosti atributa u *intrusion detection* sistemima. Najpre je SVM metod korišćen kako bi se odredio kvalitativan značaj svakog od posmatranih atributa u smislu kvaliteta predikcije sa njegovim postojanjem u skupu atributa i nakon njegovog izostavljanja. Nakon tog koraka AHP metod je korišćen na osnovu prethodno dobijenog odnosa vrednosti za određivanje relevatnosti svakog od atributa u skupu i za konačnu selekciju atributa.

U [142] AHP se kombinuje sa geografskim informacionim sistemima - GIS za analizu rizika poplava u određenoj regiji. *Fuzzy* AHP pristupom se kombinuju ekspertska znanja, geografski, istorijski i statistički podaci. Nakon obrade se dobijaju relativne težine identifikovanih faktora rizika koje se integrišu u *Quantum GIS software* kako bi se kreirale mape sa zonama procenjenog rizika plavljenja. Takve mape značajno poboljšavaju mogućnost upravljanja rizikom plavljenja.

U [161] je predstavljena integracija između AHP metoda i genetskog algoritma kao hibridnog modela AHP-GA u cilju rešavanja problema održavanja avionskih motora. Problem je definisan kroz više optimizacionih faktora i sa nametnutnim spoljnim ograničenjima, u smislu što manje cene održavanja i što većeg broja upotrebe motora nakon održavanja. U radu se najpre koristi AHP kako bi se definisali svi potrebni zahtevi i preferencija donosioca odlika, a zatim genetski algoritam kao celobrojna reprezentacija

problema kojom se u analiziranom periodu pronalazi najbolji odnos cene i performansi.

Dalja upotreba AHP modela odnosno ANP, (engl. *analytical network process*) varijante može se naći u [173]. U radu se ANP prvi put predlaže kao nelinearna metoda za fuziju informacija iz različitih modela (agregacione šeme). Kod takvih problema najznačajnije šeme fuzije su linerana težinska šema (engl. *linear weighted fusion*, *linear combination*) ili većinsko glasanje (engl. *majority voting*). Pristup se oslanja na dve osnovne ideje, da postoji zavisnost klase od atributa i da postoji međuzavisnost u klasama korišćenih modela.

Interaktivni pristup za otkrivanje lažnih odštetnih zahteva u zdravstvenom osiguranju baziran na mašinskom učenju i sa primenom AHP može se naći u [78].

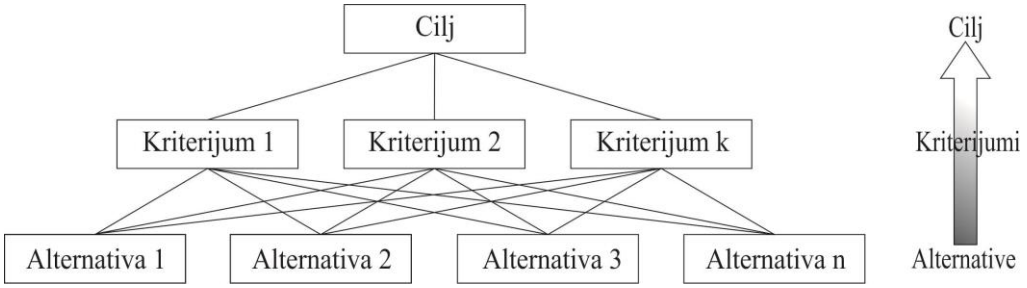
Uspešna primena AHP metoda u različitim emprijskim analizama podataka, koja je posledica jasnosti upotrebljenih matematičkih principa kao i sposobnost da se izvrši evaluacija konzistenstnosti procesa donošenja odluke, uslovila je da se AHP koristi i u eksperimentalnom delu ove doktorske disertacije.

U narednoj sekciji biće prikazane teorijske osnove potrebne za razumevanje AHP metoda.

6.2. OSNOVE ANALITIČKOG HIJERARHIJSKOG PROCESA

Analitički hijerarhijski proces predstavlja struktuiranu formu odlučivanja koja nam daje mogućnost izbora između alternativa u odnosu na njihovu relevatnost. Metod se pokazao uspešnim kod višekriterijumskog odlučivanja zbog svoje sposobnosti da vrši evaluaciju skupa faktora čak i kod međusobno suprostavljenih kriterijuma, a bez prethodnog znanja o strukturi njihovih odnosa.

AHP metod [131] i [132] podrazumeva najpre kreiranje hijerarhije, zatim formiranje matrica parnog upoređivanja, postavljanje prioriteta kroz određivanje odnosno ocenjivanje relativnih važnosti lokalno na novou kriterijuma i globalno na nivou problema odlučivanja i na kraju proveru konzistentnosti u procesu odlučivanja kroz sprovođenje analize osetljivosti. Proces rešavanja složenih problema odlučivanja primenom AHP metode podrazumeva dekompoziciju problema na cilj, kriterijume i alternative, kroz hijerarhijsku strukturu. Upravo se na slici 6.1 može videti hijerarhijska dekompozicija problema sa k kriterijuma i n alternativa.



SLIKA 6.1 Struktura AHP hijerarhije

Na slici 6.1, adaptacija na osnovu [136], grafički je prikazana struktura AHP hijerarhije i koraci koji vode ka dostizanju cilja. AHP proračuni se mogu predstaviti sledećim koracima.

Prema matematičkom modelu u AHP analizi se kreiraju matrice parnog upoređivanja, kroz definisanje relativne važnosti faktora na istom hijerarhijskom nivou u odnosu na elemente na prvom višem nivou, na osnovu kojih se obračunavaju vrednosti sopstvenog vektora (engl. *eigenvector*) odnosno dobijaju relativne važnosti atributa prema posmatranim kriterijumima.

Prilikom kreiranja matrice parnih upoređenjivanja koristi skala relativnih prioriteta sa vrednostima od 1 do 9 prema [132].

TABELA 6.1 Skala relativnih prioriteta

INTEZITET VAŽNOSTI	ZNAČAJ
1	JEDNAK
3	UMEREN
5	JAKA DOMINANTNOST
7	VEOMA JAK
9	OGROMAN
2,4,6,8	MEĐUVREDNOSTI

Na osnovu tabele 6.1 [132] može se zaključiti da se faktori mogu klasifikovati kao istog značaja, označeno na skali sa rangom 1, ako su dva elementa istog značaja u odnosu na cilj. Slaba odnosno umerena dominantnost jednog u odnosu na drugog označena je na skali rangom 3, što znači da iskustvo neznatno favorizuje jedan element u odnosu na drugi. Suštinska ili jaka dominantnost označava se rangom 5, upućuje da iskustvo ili rasuđivanje u velikoj meri favorizuju jedan element u odnosu na drugi. Rang 7 upućuje na deomonstriranu veoma jaku dominantnost, što znači da je preferentnost nekog od elemenatna potvrđena u praksi. Ogromna, odnosna ekstremna dominantnost se označava rangom 9. Preostali rangovi 2,4,6 i 8 predstavljaju međuvrednosti, koje ukazuju da je kod rasuđivanja potreban kompromis ili dalja podela. Takođe, prilikom kreiranja matrice parnog upoređivanja treba

imati u vidu, princip reciprociteta, odnosno ako se odnos aktivnosti *i* prema aktivnosti *j* opisuje nekom od vrednosti iz tabela 6.1, onda se vrednost aktivnosti *j* prema aktivnosti *i* opisuje recipročnom vrednošću.

Matrice parnog upoređivanja su oblika:

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \tag{6.3}$$

U matrici A, *n* predstavlja broj elemenata koji se porede, kriterijumi ili alternative, a a_{ij} stepen dominacije elementa *i* prema elementu *j* određenog prema skali iz tabele 6.1. Pri čemu je $a_{ij} = 1/a_{ji}$ za svako $i \neq j$ i $a_{ii} = 1$.

U okviru ove doktorske disertacije usvojena je notacija prema [31], za obeležavanje matrica parnog upoređivanja na različitim nivoima AHP hijerarhije.

Prema strukturi procesa predstavljenoj u [31], u praksi prvi korak podrazumeva kreiranje matrice parnog upoređivanja počevši od nivoa kriterijuma, kako bi se dobio sopstveni vektor označen sa RVV (engl. *relative value vector*).

Za određivanje težina i dobijanje vektora prioriteta $W = (w_1, w_2, \dots, w_n)^T$ u RVV matrici, ali i u svim ostalim matricama parnog upoređivanja, prema [136], najčešće se koriste metodi geometrijske i aritmetičke sredine, predstavljeni u relacijama (6.4) i (6.5), respektivno:

$$w_i = \frac{1}{n} \sum_{i=1}^{n} \frac{a_{ij}}{\sum_{k=1}^{n} a_{kj}}, i = 1,2,...,n \tag{6.4}$$

$$w_i = \frac{\left(\prod_{j=1}^{n} a_{ij} \right)^{1/n}}{\sum_{k=1}^{n} \left(\prod_{j=1}^{n} a_{kj} \right)^{1/n}}, i = 1,2,...,n \tag{6.5}$$

Zatim se za svaki posmatrani kriterijum kreira matrica parnog upoređivanja pod nazivom PCM (engl. *pairwise comparison matrix*) kojom se karakteriše relevatnost alternative u odnosu na svaki pojedinačni kriterijum.

Nakon tog koraka kreira se evaluaciona OPM matrica (engl. *option performance matrix*) kojom se predstavljaju lokalne težine alternativa u odnosu na odabrane kriterijume.

U poslednjem koraku je potrebno odrediti globalne težine alternativa u odnosu na

posmatrani cilj. Prema matematičkom modelu vrši se množenje RVVxOPM, kako bi se izvršila sinteza svih težina i odredili ukupni rangovi, odnosno težinski koeficijenti svih alternativa.

Kod precizno određenih prioriteta, matrica A_{ij} je tranzitivna i sopstveni vektor (engl. *eigenvector*) W reda n može se izračunati tako da je $AW=\lambda W$, gde λ predstavlja sopstvene vrednosti. S obzirom na inkonzistentnosti u procesu donošenja odluka u opštem slučaju težinski vektor W generalno zadovoljava jednačinu $AW=\lambda maxW$, pri čemu λmax predstavlja najveću sopstvenu vrednost matrice poređenja A, i računa se prema formuli:

$$\lambda_{\max} = \frac{1}{n}\sum_{i=1}^{n}\frac{(AW)_i}{\omega_i} \tag{6.6}$$

pri čemu za dobijene vrednosti važi da je $\lambda max \geq n$. Ukoliko postoji nekonzistentnost u procesu donošenja odluka, odnos između λmax i n određuje stepen nekonzistentnosti odluke, pri čemu jednakost između dve vrednosti upućuje na konzistentnost. Upravo se zbog mogućnosti da identifukuje nekonzistentnost u procesu donošenja odluka ili tokom dodeljivanja težina AHP i ubraja u popularne višekriterijumske metode odlučivanja.

Indeks konzistentnosti CI (engl. *consistency index*) računa se po formuli:

$$CI=(\lambda max-n)/(n-1) \tag{6.7}$$

Konačni stepen konzistentnosti (engl. *consistency ratio* - *CR*) određuje se prema jednačini:

$$CR = CI/RI \tag{6.8}$$

gde *RI* (engl. *random consistency index*) predstavlja odgovarajuću vrednost slučajnog indeksa konzistentnosti RI koja se očitava iz tabele 6.2, preuzete iz [131]. U tabeli 6.2 predstavljene su vrednosti slučajnih indeksa za različite redove matrice.

TABELA 6.2 Slučajni indeksi RI

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0.0	0.0	0.58	0.9	1.12	1.24	1.32	1.41	1.45	1.49	1.51	1.48	1.56	1.57	1.59

U slučaju i da je broj alternativa za razmatranje veći od 15 onda se RI može i dalje odrediti indirektnim obračunom [93]. Najpre se generiše 500 matrica parnih upoređivanja sa slučajnom raspodelom međusobih značajnosti kriterijuma, te se na osnovu njih odredi srednja najveća vrednost sopstvenog vektora za svih 500 posmatranih matrica u oznaci $\lambda'max$. RI se zatim obračunava kao:

$$\text{RI}= \lambda'max-n/n-1 \tag{6.9}$$

Prema [131] ustanovljeno je da vrednost CR koja ne prelazi vrednost 0.1, $CR \leq 0.1$, upućuje na konzistentne odluke, a da vrednost CR=0 upućuje na perfektno konzistentne odluke.

Primer koji sledi predstavlja opisane AHP proračune, modifikacija na osnovu [136]: Određena kompanije želi da nabavi novi deo opreme i uzima u razmatranje sledeće aspekte prilikom odlučivanja: troškove nabavke (K1), pouzdanost opreme (K2) i šira upotrebna vrednost, odnosno korisnost (K3). Na raspolaganju su tri dobavljača, označeni dalje u primeru sa A1, A2 i A3, u smislu alternativa izbora. Donosilac odluke favorizuje pouzdanost opreme u odnosu na cenu nabavke i umereno je favorizuje u odnosu na korisnost za druge namene.

Matrica evaluacija kriterijuma, prema definisanim kriterijumima predstavljena je u tabeli 6.3:

TABELA 6.3 Matrica parnog upoređivanja kriterijuma

	K1	K2	K3
K1	1	1/3	1/5
K2	3	1	1/3
K3	5	3	1

Za prethodnu matricu vrši se obračun težina u cilju izračunavanja RVV vrednosti primenom metode aritmetičke sredine, relacija 6.4, na sledeći način: najpre se dele elemenati svake kolone sumom vrednosti te kolone, izračunava se suma elemenata po svakoj vrsti i određuju srednje vrednosti svake vrste, prikazano u tabeli 6.4.

TABELA 6.4 Matrica obračuna težina kriterijuma, RVV

	K1	K2	K3	Σ_v	$W = \Sigma / n$
K1	$1/\Sigma_{K1}$	$1/3/\Sigma_{K2}$	$1/5/\Sigma_{K3}$	0.318	0.106
K2	$3/\Sigma_{K1}$	$1/\Sigma_{K2}$	$1/3/\Sigma_{K3}$	0,781	0.206
K3	$5/\Sigma_{K1}$	$3/\Sigma_{K2}$	$1/\Sigma_{K3}$	1.900	0.633
Σ_k	Σ_{K1} 9	Σ_{K2} 4.33	Σ_{K3} 1,53		
$\lambda_{max} = 3.039$					
CR = 0.0375					

Na osnovu prethodne tabele mogu se očitati sledeće vrednosti za vektor težina kriterijuma, $RVV = (0.106, 0.206, 0.633)^T$, prema dobijenim vrednostima K1 ima najveću težinu, nakon njega sledi K2, pa kriterijum K3. Da bi se odredio stepen konzistentnosti odluke potrebno je

odrediti λ_{max} i na osnovu njega CI i CR prema relacijama (6.7) i (6.8). Vrednost λ_{max} se određuje nakon matričnih izračunavanja množenjem polazne matrice parnog upoređivanja kriterijuma matricom težina kriterijuma, a zatim deljenjem elemenata dobijene matrice sa odgovarajućim vrednostima težine kriterijuma, sumiranjem dobijenih vrednosti i deljenjem sa ukupnim brojem kriterijuma što je u posmatranom slučaju tri. Obračunate vrednosti prikazane su u tabeli 6.4.

Nakon određivanja RVV vrednosti, formiraju se pojedinačne PCM matrice tako što se isti postupak primeni za određivanje posmatranih alternativa u odnosu na svaki pojedinačni kriterijum. Za posmatrani primer konkretne vrednosti alternativa u odnosu na kriterijume postavljene su u skladu sa inicijalnom matricom odlučivanja, čije su vrednosti predstavljene u tabeli 6.5:

TABELA 6.5 Vrednosti matrice odlučivanja

	K1	K2	K3
A1	VRLO DOBAR	DOBAR	VRLO DOBAR
A2	VRLO DOBAR	PROSEČAN	ODLIČAN
A3	PROSEČAN	DOBAR	PROSEČAN

Vrednosti u matrici parnog upoređivanja alternativa u odnosu na kriterijume određene su na osnovu vrednosti iz tabele 6.6

TABELA 6.6 Poređenje alternativa po K1, PCM_{k1} matrica

K1	A1	A2	A3	*W*
A1	1	1	3	0.429
A2	1	1	3	0.429
A3	1/3	1/3	1	0.143
CR=0				

Na osnovu dobijenih CR vrednosti može se zaključiti da su odluke konzistentne.

TABELA 6.7 Poređenje alternativa po K2, PCM_{k2} matrica

K2	A1	A2	A3	*W*
A1	1	1/3	1	0.200
A2	3	1	3	0.600
A3	1	1/3	1	0.200
CR=0				

Na osnovu dobijenih vrednosti može se zaključiti da su i u odnosu na kriterijum K2 odluke konzistentne.

TABELA 6.8 Poređenje alternativa po K3, PCM_{k3} matrica

K3	A1	A2	A3	*W*
A1	1	1/3	3	0.260
A2	3	1	5	0.633
A3	1/3	1/5	1	0.106
CR=0,0375				

Na osnovu dobijenih vrednosti može se zaključiti da su odluke i za kriterijum K3 konzistentne.

Nakon izračunavanja lokalnih težina kriterijuma, na osnovu pojedinačnih PCM matrica kreira se OPM matrica lokalnih težina. U poslednjem koraku vrši se množenje matrice OPM lokalnih težine alternativa matricom težina kriterijuma $OPM \circ RVV$ kako bi se dobile gobalne težine alternativa.

TABELA 6.9 Matrica globalnih težina

$[OPM]$				$[RVV]$		$[W]$
0.429	0.200	0.260		0.106		0.262
0.429	0.600	0.633	.	0.206	=	0.602
0.143	0.200	0.106		0.633		0.134

Na osnovu dobijenih vrednosti može se zaključiti da je najbolja alternativa A3, odnosno da je ponuda trećeg dobavljača najbolje rangirana, zatim sledi ponuda drugog dobavljača i na kraju ponuda prvog dobavljača.

Imajući u vidu značaj koji proces izbora atributa sam po sebi ima u kreiranju dobrog predikcionog modela, može se zaključiti da optimalni izbor atributa utiče i na kvalitet predikcionih modela i na smanjenje vremena proračuna [113]. Takođe prema [113] se ističe da je čest slučaj da jednostavnija rešenja sa manjim brojim izabranih atributa u mnogim radovima pružaju bolje prediktivne performanse. Sve prethodno navedeno imalo se u vidu prilikom razvijanja metodologije za izbor atributa predloženoj u narednom poglavlju.

POGLAVLJE 7

RAZVOJ METODOLOGIJE IZBORA ATRIBUTA

U poglavlju koje sledi biće predstavljen razvoj metodologija za izbor atributa i algoritam za određivanje težina atributa zasnovan na Analitičkom hijerarhijskom procesu.

Prezentovana istraživanja motivisana su istraživanjima prikazanim u radovima [6], [60], [93], [114], [176] i predstavljaju nastavak razvoja prethodnih modela [100], [101] i generalizaciju pristupa iz radova [102] i [147].

Predloženom metodologijom se formalizuju znanja specifična za finansijska tržišta. Kroz matematički metod odlučivanja vrši se konceptualizacija i integracija prethodnog znanja u postupak izbora atributa za obuku predikcionog modela. Selekcija atributa vrši se primenom Analitičkog hijerarhijskog procesa. Metodološkim okvirom predlažu se i kriterijumi poređenja u skladu sa kojima se vrši rangiranje i selekcija atributa. Kako bi se postiglo maksimalno poboljšanje predikcije na posmatranom domenu, predlaže se integracija težina atributa u kernel kod SVM metoda i metoda najmanjih kvadrata podržavajućih vektora, LS-SVM.

7.1. DEFINISANJE KRITERIJUMA EVALUACIJE

Kako bi se izvršila procena relevantnosti posmatranog skupa ulaznih atributa neophodno je uvesti kriterijume za AHP evaluaciju, čime se u osnovu vrši adaptacija predikcionog modela pomoću prethodnog znanja o posmatranom finansijskom tržištu.

U ovoj disertaciji predlaže se kreiranje tehničkih strategija trgovanja kojima se može predstaviti mera uspeha svakog od tehničkih indikatora na koji se strategija oslanja. Tehničke strategije trgovanja sastoje se od skupa pravila trgovanja koji se koriste kako bi se definisao signal trgovanja. U najvećem broju slučajeva strategije trgovanja se oslanjaju na korišćenje jednog do dva tehnička indikatora kojima se definišu signali trgovanja [74] i [118].

Kriterijumi potrebni za AHP evaluaciju mogu se posmatrati na dva načina. U prvu grupu mogu se ubrajati kriterijumi koji se koriste kako bi se procenila ekonomska relevatnost posmatranih atributa: gde se mogu ubrojati kumulativni bruto prinos, kao mera profitabilnosti finansijskog tržišta i sistemski rizik kao mera tržišne volatilnosti. Treći kriterijum predstavlja poređenje signala koji su generisani na osnovu strategija trgovanja i aktuelne promene vrednosti tržišnog indeksa, koji se porede na osnovu postignute preciznosti.

Bruto prinos

Prinos na investicije u slučaju posmatranog tržišnog indeksa se računa kao razlika između dnevnih vrednosti indeksa u nacionalnoj valuti pomnoženih sa generisanim signalom trgovanja za konkretni dan. Bruto prinos se definiše kao kumulativni kapitalni prinos za određeni vremenski period na sledeći način:

$$R = \sum_{t=1}^{n} S_t * \left(CP_t - CP_{t-1}\right) \tag{7.1}$$

gde S_t predstavlja signal za trgovanje koji je generisan u okviru određene strategije trgovanja. Ovako obračunat prinos na investicije u finansijske instrumente omogućava poređenje uspešnosti trgovanja, koje se bazira na izabranom setu tehničkih indikatora. Kako bi se izvršila evaluacija kriterijuma prema vrednostima ordinalne težinske skale, tabela 6.1 , primenjena je proporcionalna funkcija u odnosu na *min-max* odnos rezultujućih vrednosti. Po istoj funkciji je izvršeno i skaliranje vrednosti za preostala dva kriterijuma.

Sistemski rizik

U ovom istraživanju, pored prinosa, uvodi se i rizik u model predikcije kao jedan od kriterijuma za evaluaciju u AHP analizi, s obzirom na činjenicu da je nivo prinosa u trgovanju finansijskim instrumentima uslovljen rizikom [8] i [127]. Sistemski rizik se može odrediti na sledeći način**:**

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{t=1}^{n}(R_t - \overline{R})^2} \tag{7.2}$$

gde $\overline{R}$ predstavlja srednju vrednost bruto prinosa R u određenom vremenskom periodu *t.*

Stopa predviđanja

Kao opšta mera procene uspešnosti predviđanja u disertaciji se koristi stopa pogodka - HR (engl. *hit ratio*). Stopa pogodka se izračunava na osnovu broja pravilno generisanih

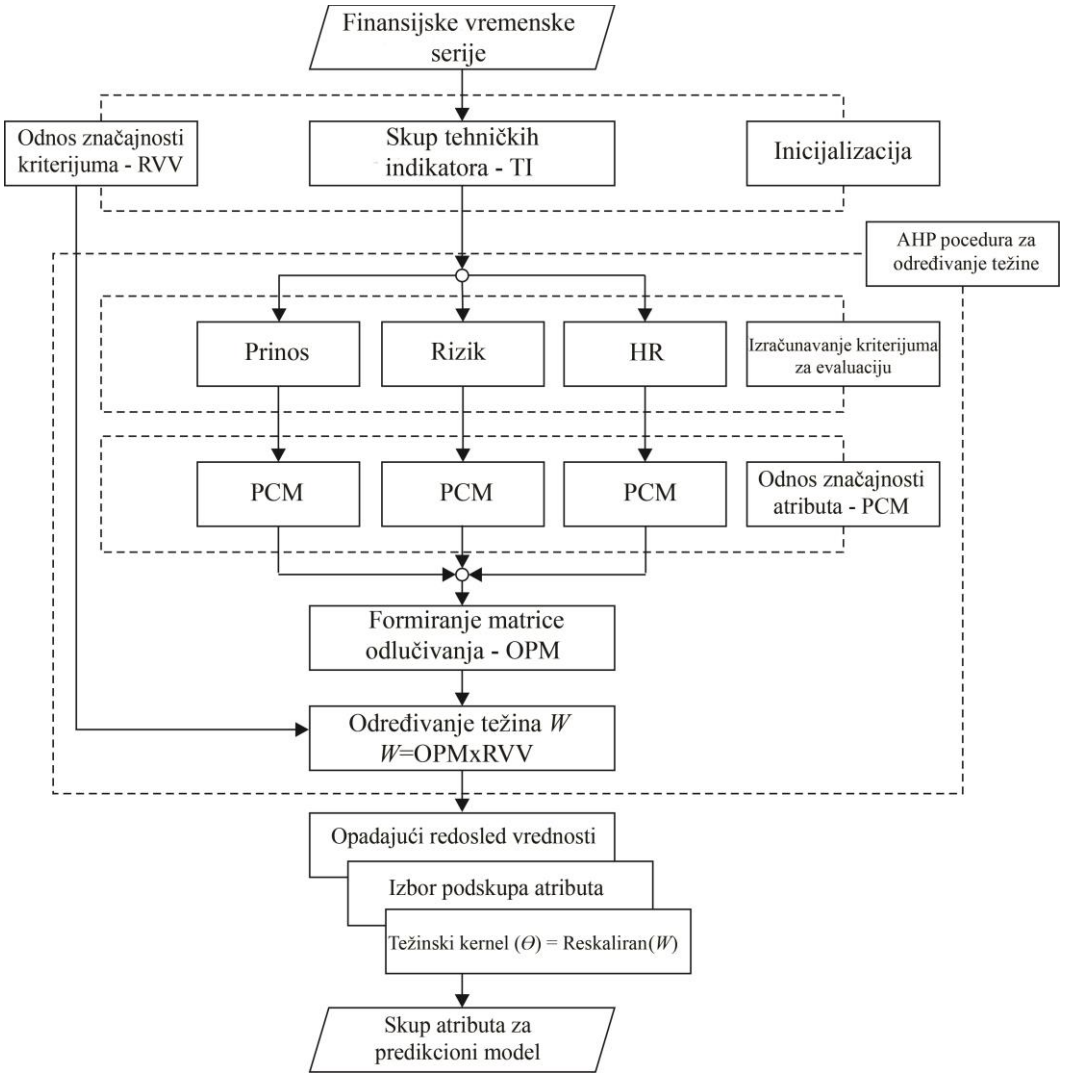
signala trgovanja u okviru posmatrane grupe:

$$HR = \frac{1}{m}\sum_{i=1}^{m} PO_i \tag{7.3}$$

U jednačini PO_i predstavlja predikcioni rezultat *i-tog* dana trgovanja, odnosno S_t za posmatranu strategiju trgovanja. PO_i je jednako *1* ako je predikcioni rezultat jednak aktuelnoj vrednosti u posmatranom danu trgovanja, u suprotnom PO_i je jednako 0, dok broj *m* predstavlja broj podataka u korišćenom skupu podataka.

7.2. ALGORITAM IZBORA ATRIBUTA

Predložena metodologija za izbor atributa koja se zasniva na AHP evaluaciji indikatora predstavljena je na slici 7.1:



SLIKA 7.1 Algoritam za izbor atributa

Nakon kreiranja početnog skupa tehničkih indikatora, prvi korak u algoritmu predstavljenom na slici 7.1 predstavlja obračunavanje kriterijuma za AHP evaluaciju. Odnosno da se za definisane strategije trgovanja nad tehničkim indikatorima izračunaju vrednosti: bruto prinosa, sistemskog rizika i stope predviđanja. Vektor odnosa značajnosti kriterijuma - RVV se obračunava na način predstavljen u poglavlju 6.2. Nakon toga kreiraju se tri nezavisne matrice za predstavljanje odnosa značajnosti atributa - PCM. Po AHP porceduri težine u matricama predstavljaju rangiranje indikatora u odnosu na posmatrani kriterijum. Sledeći korak predstavlja kreiranje matrice odlučivanja - OPM matrice kako bi se dobile globalne vrednosti težina za svaki od atributa iz posmatranog skupa atributa, odnosno vektor težina $W = (w_1, w_2, \ldots, w_d)^T, i = 1,...,d$, a d predstavlja broj posmatranih atributa. Svaka pojedinačna vrednost iz vektora W određuje relativni značaj (rang) svakog pojedinačnog atributa u odnosu na posmatrane kriterijume. U narednom koraku, vrši se sortiranje skupa atributa u opadajućem redosledu prema vrednostima težina *w*. Cilj ovog koraka je da se pronađe podskup skupa atributa koji će se koristiti za obučavanje predikcionog modela. Konkretno, podrazumeva se da atribut koji ima najveću vrednost težinskog faktora, može najviše doprineti učenju predikcionog modela. Ako bi se iscrtao grafik na osnovu dobijenih vrednosti težina, na grafikonu bi se moglo uočiti opadanje relevatnosti atributa, što bi uzrokovalo pojavljivanje prepoznatljivog efekta „lakta“ na grafikonu. Nakon izbora podskupa atributa, procenjene težine atributa treba reskalirati kako bi zadovoljile relaciju 3.29, čime se dobijaju vrednosti za vektor $\Theta = \mathrm{diag}[\theta_1, \theta_2, \cdots, \theta_m]$, gde *m* predstavlja broj izabranih atributa. U poslednjem koraku predstavljenom na slici 7.1 težinski kernel se izračunava množenjem vrednosti atributa sa reskaliranim vrednostima težina u primarnom prostoru atributa.

Procedura izbora atributa prikazana pseudokodom u algoritmu **1**, predstavlja generalizaciju i opšti koncept prethodno opisane metodologije, kako bi ona bila primenljiva i na druge skupove podataka.

Algoritam 1. Selekcija i izbor atributa primenom AHP metode

*Ulaz:*Polazni skup atributa $\left(x^{(1)}, x^{(2)}, ..., x^{(d)}\right)$

Rezultat: Podskup skupa atributa $\left(x^{(1)}, x^{(2)}, ..., x^{(j)}\right), K\min \le j \le K\max$

1. Inicijalizacija

   Za izabrani domen predikcije izvršiti formalizaciju i konceptualizaciju znanja o domenu

2. Definisanje alternativa i kriterijuma

   Za skup kandidata atributa definisati hijerarhiju i obračunati vrednosti kriterijuma evaluacije

3. Formiranje rezultata

   Na osnovu AHP proračuna odrediti težine atributima kandidatima

   $W = (w_1, w_2, \ldots, w_d)^T$

4. Uređenje rezultata

   Izvršiti sortitanje dobijenih vrednosti u opadajućem redosledu

   $Sort\ W = (w_1, w_2, \ldots, w_d)^T$

5. Izbor atributa

   Vizualizacijom dobijenih vrednosti težina ili na osnovu unapred definisanog praga vrednosti, izvršiti izbor podskupa skupa atributa:

   $W_s \subset W$

   $W_s = (w_1, w_2, \ldots, w_s)^T$, s predstavlja broj selektovanh atributa

6. Reskaliranje

   $W_s \overset{(3.29)}{\longrightarrow} \Theta$

7. Težine izabranog podskupa atributa

   $\Theta = diag[\theta_1, \theta_2, \cdots, \theta_m]$

8. Inkorporacija težina u kernel:

   $K(x, x_k) = K(\theta x, \theta x_k)$

9. Obuka predikcionog modela

Prema klasifikaciji metoda izbora atributa u nadgledanom mašinskom učenju predstavljenoj u trećem poglavlju ove doktorske disertacije, predložena metoda može se svrstati u filter metode rangiranja, s obzirom na njenu nezavisnot od korišćenog metoda predikcije.

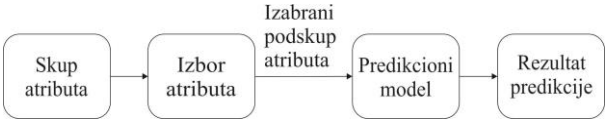
POGLAVLJE 8

ANALIZA REZULTATA

U okviru ovog poglavlja predstavljeni su eksperimentalni rezultati primene metodologije izbora atributa predložene u ovoj doktorskoj disertaciji. Izvršena je komparativna analiza predložene metodologije za rangiranje i izbor atributa u kombinaciji sa metodom najmanjih kvadrata podržavajućih vektora - LS-SVM sa drugim metodama izbora atributa kao i sa drugim metodama mašinskog učenja. Poglavlje počinje opisom skupa podataka koji su korišćeni u eksperimentu, zatim je predstavljen eksperimentalni okvir i na kraju poglavlja data je diskusija rezultata.

8.1. EKSPERIMENTALNI OKVIR

Iako mnoga istraživanja ukazuju da promene cena akcija nisu potpuno nasumične, posmatrano u dužem vremenskom intervalu promena cena se aproksimira slučajnim procesom (engl. *random walk*). Stoga se stepen preciznosti od oko 60% koji se dobija korišćenjem metoda mašinskog učenja često opisuje kao zadovoljvajući za predviđanja na tržištima kapitala [146] i [83]. Postojeći sistemi za predviđanje tržišnih kretanja se po pravilu fokusiraju na sledeće karakteristike: izbor atributa, izbor predikcionog modela i evaluacija rezultata.

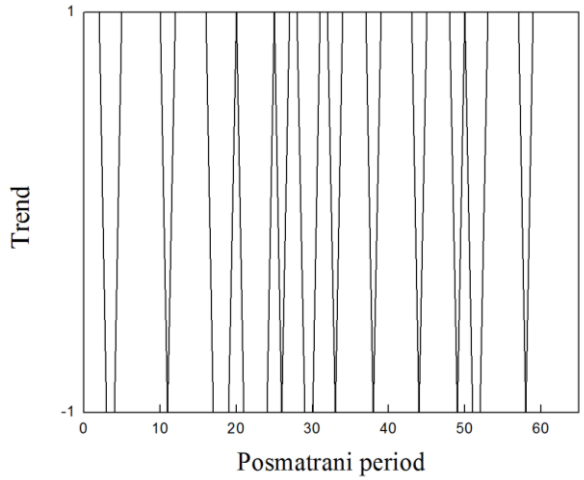


SLIKA 8.1 Struktura predikcionog procesa

Najšire rasprostranjeni predikcioni modeli sastoje se kao što je prikazano na slici 8.1 prema [6] i [176] iz dva dela. Prvi deo predstavlja korak izbora atributa za obuku modela, dok drugi predstavlja izbor predikcionog metoda i njegovog algoritma učenja. U ovoj doktorskoj

disertaciji kako je već istaknuto razmatrana su oba aspekta predikcionog modela i izbor atributa i prilagođavanje metoda mašinskog učenja sa ciljem povećanja preciznosti.

Problem modelovanja promene trenda kretanja tržišnih indeksa se uobičajeno modeluje kao problem binarne klasifikacije, pri čemu su oznake klase kategorije *-1* i *1*. Klasa *-1* indikuje da je cena na zatvaranju posmatranog dana veća od cene na zatvaranju narednog dana, dok klasa *1* indikuje suprotno odnosno da je cena na zatvaranju posmatranog dana manja od cene na zatvaranju narednog dana. Na slici 8.2 prikazan je uzorak koji predstavlja karakteristične promene kretanja vrednosti tržišnih indeksa.



SLIKA 8.2 Fluktuacije trenda

Na osnovu slike 8.2 može se uočiti da su promene pravce kretanja vrednosti tržišnih indeksa intezivne, odnosno da je ponašanje finansijskih tržišta dinamično što predstavlja poseban predikcioni izazov.

8.1.1. TEHNIČKE STRATEGIJE TRGOVANJA

Tehnički indikatori imaju značajnu ulogu u predviđanju trenda promene vrednosti finansijskih instrumenata. Mogu se podeliti u sledeće grupe: indikatori trenda, indikatori obima trgovanja i oscilatori.

Indikatori trenda identifikuju i prate trend vrednosti finansijskih instrumenata, dok se indikatori obima zasnivaju na promeni u obimu trgovanja finansijskih instrumenata, čime upotpunjuju informacije koje pružaju indikatori trenda u formiranju trgovinskih strategija. Oscilatori su vodeći indikatori koji generišu rane signale upozorenja na promene u trendu vrednosti finansijskih instrumenata i određuju jačinu postojećeg trenda, kao i trenutak kada se promena trenda dešava.

Empirijski podaci koji se koriste za analizu finansijskih tržišta se uobičajeno generišu na dnevnom nivou. Količina podataka, koji se koriste u analizi, se ubrzano povećava, što dovodi do povećane potrebe za razvojem odgovarajućih matematičkih alata za praktične primene u oblasti klasifikacije i razumevanja funkcionisanja sistema finansijskog tržišta. Ako je predviđanje trendova na finansijskom tržištu preciznije, investitori mogu ostvariti veće prinose.

Dva najviše razmatrana pristupa u razumevanju procesa na finansijskom tržištu jesu teorija slučajnog hoda i sa njom povezana hipoteza efikasnog tržišta [47], i tehnička analiza koja se zasniva na prepoznavanju trendova.

Da bi se mogli pratiti kvantitativni efekti, u ovoj doktorskoj disertaciji predloženo je formiranje tehničkih strategija trgovanja, čijim bi se rezultatima mogao meriti i uspeh svakog tehničkog indikatora. Prilikom kreiranja startegija trgovanja predložen je pristup u kome se ostavlja mogućnost investitorima da zadrže prethodno stanje. Na taj način izvršena je realna simulacija procesa na tržištima kapitala u cilju da se kroz adekvatniju reprezentaciju ponašanja tržišnih učesnika dobiju što realniji pokazatelji o kvalitetu strategija trgovaja.

Sada ćemo razmotriti skup potencijalnih ulaznih atributa. S obzirom na to da je cilj ovog rada predviđanje tendencija u promenama vrednosti indeksa, u disertaciji su korišćeni isključivo tehnički indikatori za formiranje predikcionog modela. Najčešće korišćeni tehnički indikatori u radovima su:

- EMA (engl. *exponential moving average*) – eksponencijalni pokretni prosek vrednosti finansijskog instrumenta na zatvaranju,
- RSI (engl. *relative strength index*) – indeks koji meri brzinu i promenu kretanja vrednosti finansijskog instrumenta,
- MACD (engl. *moving average convergence-divergence*) – indikator koji meri jačinu i pravac trenda finansijskih instrumenata,
- SO (engl. *stochastic oscilator*) – indikator koji pokazuje održivost trenda i signalizira promenu vrednosti finansijskog instrumenta,
- ROC (engl. *rate of change*) – indikator koji pokazuje procentualnu promenu vrednosti finansijskog instrumenta na zatvaranju,
- CCI (engl. *commodity channel index*) – indikator koji se koristi za otkrivanje cikličnih promena vrednosti finansijskog instrumenta merenjem odstupanja vrednosti od njene statističke sredine,
- SAR (engl. *parabolic stop and reverse*) – indikator koji otkriva pravac trenda

promene vrednosti finansijskog instrumenta i koristi se za određivanje momenata za trgovanje.

Detaljna procedura za izračunavanje tehničkih indikatora i pravila za generisanje signala trgovanja predstavljeni su u tabeli 8.1.

TABELA 8.1 Tehnički inidkatori i strategije trgovanja

TEHNIČKI INDIKATORI	FORMULA	TEHNIČKE STRATEGIJE TRGOVANJA - S_T
CP_t - Closing price, t= 1,2, ... n, LP_N / HP_N – Lowest/ Highest price in the past N days		
EMA	$\mathrm{EMA}_t = \mathrm{CP}_t * k + \mathrm{EMA}_{t-1} * (1-k), k = 2/(N+1)$	$\begin{cases} 1 \text{ if } \mathrm{EMA}_{1,t} > EMA_{10,t} \\ -1 \text{ if } \mathrm{EMA}_{1,t} < EMA_{10,t} \end{cases}$
MACD	$MACD_t = EMA_{12,t} - EMA_{26,t}$	$\begin{cases} 1 \text{ if } \mathrm{MACD}_t > EMA_{9,t} \\ -1 \text{ if } \mathrm{MACD}_t < \mathrm{EMA}_{9,t} \end{cases}$
RSI	$RSI_t = 100 - \frac{100}{1+RS_t}\ RS_t = \frac{\sum_{n=t}^{t-d} \max(0, CP_n - CP_{n-1})}{\sum_{n=t}^{t-d} \lvert\min(0, CP_n - CP_{n-1})\rvert}$	$\begin{cases} 1 \text{ if } \mathrm{RSI}_{t-1} \geq 30 \text{ and } \mathrm{RSI}_t \geq 30 \\ -1 \text{ if } \mathrm{RSI}_{t-1} \leq 70 \text{ and } \mathrm{RSI}_t \leq 70 \end{cases}$
CCI	$CCI_t = (M_t - SM_t)/0.0015D_t$ $M_t = HP_t + CP_t + LP_t\ SM_t = \sum_{i=t-m+i}^{t} M_i / m$ $D_t = \sum_{i=t-m+i}^{t} \lvert M_i - SM_t \rvert / m$	$\begin{cases} 1 \text{ if } \mathrm{CCI}_t > 100 \text{ or } \mathrm{CCI}_t > -100 \\ -1 \text{ if } \mathrm{CCI}_t < 100 \text{ or } \mathrm{CCI}_t < -100 \end{cases}$
SO	$\%K_t = 100((CP_t - LP_{14})/(HP_{14} - LP_{14}))$ $\%D = \frac{1}{t}\sum_{t=1}^{3} \%K_t$	$\begin{cases} 1 \text{ if } \%\mathrm{D} < 0.2 \text{ and } \%\mathrm{K}_t > \%\mathrm{D} \\ -1 \text{ if } \%\mathrm{D} > 0.8 \text{ and } \%\mathrm{K}_t < \%\mathrm{D} \end{cases}$
SAR	$SAR_{t+1} = SAR_t + \alpha(EP - SAR_t)$ EP -the extreme point α – the acceleration factor	$\begin{cases} 1 \text{ if } \mathrm{CP}_t > \mathrm{SAR}_t \\ -1 \text{ if } \mathrm{CP}_t < \mathrm{SAR}_t \end{cases}$
ROC	$ROC_t = 100\ ((CP_t - CP_{t-n})/CP_{t-n})$	$\begin{cases} 1 \text{ if } \mathrm{ROC}_t > 0 \\ -1 \text{ if } \mathrm{ROC}_t < 0 \end{cases}$

Kako je navedeno u poglavlju Poglavlje 7, na osnovu tehničkih indikatora kreirane su strategije trgovanja. Tehničke strategije trgovanja sastoje se od skupa pravila trgovanja koja se koriste kako bi se definisao signal trgovanja. U najvećem broju slučajeva strategije trgovanja se oslanjaju na korišćenje jednog do dva tehnička indikatora kojima se definišu signali trgovanja [74] i [118]. Promenom kombinacije tehničkih indikatora može se definisati mnoštvo različitih pravila trgovanja.

Eksponencijalni pokretni prosek – **EMA** predstavlja prosek vrednosti finansijskog instrumenta u određenom vremenskom periodu. Ovo je često korišćeni indikator, čiji je cilj da, eliminišući tekuće oscilacije, izravna trend vrednosti finansijskog instrumenta i istakne njegov osnovni tok.

S obzirom na to da je trend tržišnih vrednosti vrlo volatilan, nije praktično primenjivati pravila trgovanja zasnovana samo na trendu, jer se na taj način može generisati previše signala, uključujući i lažne signale. Stoga se u definisanju pravila trgovanja najčešće koristi neki tip pokretnog proseka (engl. *moving average* – MA) – aritmetički ili prosti MA

(engl. *simple* MA), ponderisani MA (engl. *weighted* MA) i eksponencijalni MA (engl. *exponential* MA). U ovoj disertaciji koristiće se eksponencijalni MA (EMA), koji u obračunu proseka vrednosti za određeni period *t* veći značaj pridaje vrednostima ostvarenim u bližim vremenskim intervalima. Može se obračunati primenom sledeće formule:

$$EMA_t = CP_t * k + EMA_{t-1} * (1 - k), k = 2/(N + 1) \quad (8.1)$$

gde upotrebljeni simboli imaju sledeće značenje: CP_t – vrednost finansijskog instrumenta u periodu *t*, k – faktor za izravnanje trenda, N – broj perioda za koji se EMA obračunava.

Strategije trgovanja koje se baziraju na EMA, signale za kupovinu i prodaju generišu u preseku vrednosti finansijskog instrumenta i EMA ili u preseku dva ili više EMA. U ovoj doktorskoj disertaciji primenjene su strategije koje se baziraju na preseku dva EMA, tzv. sistem dvostrukih preseka pokretnih proseka (engl. *dual moving average crossover system*). Ovakava pravila trgovanja uključuju dve EMA vrednosti koje se obračunavaju za različite vremenske intervale. Ako se vrednost EMA obračunava za period od 20 dana ili manje ona se naziva kratkoročna (engl. *short-term*) EMA. Ako je period obračuna vrednosti od 20 do 50 dana, govori se o srednjoročnoj (engl. *medium-term*) EMA i za period veći od 50 dana, računa se dugoročna (engl. *long-term*) EMA [26]. Uzevši u obzir karakteristike posmatranih tržišta kapitala, kratkoročna EMA je obračunata za period od jednog dana, a srednjoročna za period od 10 dana.

U slučaju ovog tipa sistema trgovanja, signali prodaje i kupovine, koji se generišu u trenutku *t*, nastaju upoređivanjem dve EMA vrednosti prema sledećim pravilima:

$$S_t = \begin{cases} 1 \; ako \; EMA_{s,t} > EMA_{(m,l),t} \\ -1 \; ako \; EMA_{s,t} < EMA_{(m,l),t} \end{cases} \quad (8.2)$$

Prema ovako definisanim pravilima trgovanja, signal za kupovinu u trenutku *t* $(S_t = 1)$ se generiše kada je vrednost kratkoročne EMA ($EMA_{s,t}$) veća od vrednosti srednjoročne EMA ($EMA_{m,t}$) ili dugoročne EMA ($EMA_{l,t}$). Ovakva situacija ukazuje na porast vrednosti finansijskog instrumenta i rastući trend, kada je poželjno biti na tržištu. Signal za prodaju u trenutku *t* $(S_t = -1)$ se dobija u obrnutoj situaciji – kada je vrednost kratkoročne EMA ($EMA_{s,t}$) manja od vrednosti srednjoročne EMA ($EMA_{m,t}$) ili dugoročne EMA ($EMA_{l,t}$). U ovakvoj situaciji investitori se povlače sa tržišta, jer ovakav signal ukazuje na opadajući trend vrednosti finansijskog instrumenta. S obzirom na to da se u ovoj doktorskoj disertaciji za generisanje signala trgovanja koriste pokretni proseci obračunati u različitim vremenskim intervalima bez unapred definisanih granica za trgovanje, signal za kupovinu ili prodaju se generiše svakog dana trgovanja.

Indeks konvergencije ili divergencije pokretnih proseka (engl. *moving average convergence-divergence* – MACD) predstavlja široko korišćeni indikator za prepoznavanje i praćenje trenda, ali i promene trenda. Računski se ovaj indikator u trenutku *t* može izraziti kao razlika između vrednosti kratkoročne EMA i vrednosti srednjoročne ili dugoročne EMA finansijskog instrumenta, odnosno:

$$MACD_t = EMA_{s,t} - EMA_{(m,l),t} \tag{8.3}$$

Uobičajeno korišćena kombinacija EMA jeste kratkoročna EMA obračunata za period od 12 dana i srednjoročna EMA obračunata za period od 26 dana. Međutim, mogu se koristiti i druge kombinacije u zavisnosti od uslova na finansijskom tržištu i cilja investitora [4] i [46].

Trend povećanja i smanjenja vrednosti finansijskog instrumenta mogu se prepoznati prema vrednosti MACD: u slučaju kada je trend rastući, vrednost ovog indikatora će biti pozitivna, dok negativna vrednost MACD signalizira opadajući trend. Signali za kupovinu i prodaju finansijskog instrumenta definišu se na bazi sistema preseka indikatora, gde signalnu liniju predstavlja EMA indikatora MACD obračunata za period od 9 dana i to na sledeći način:

$$S_t = \begin{cases} 1 \; ako \; MACD_t > \mathrm{EMA}_{9,t} \\ -1 \; ako \; MACD_t < EMA_{9,t} \end{cases} \tag{8.4}$$

Signal za kupovinu u trenutku t $(S_t = 1)$ generiše se ukoliko je vrednost $MACD_t$ veća od vrednosti EMA indikatora MACD obračunate za period od 9 dana ($\mathrm{EMA}_{9,t}$), dok je signal za prodaju finansijskog instrumenta u trenutku t $(S_t = -1)$ pad vrednosti $MACD_t$ ispod vrednosti $EMA_{9,t}$.

Indeks relativne snage (engl. *relative strength index – RSI*) jeste tehnički indikator koji meri opseg oscilacija vrednosti finansijskog instrumenta, odnosno brzinu i promenu kretanja vrednosti finansijskog instrumenta. Vrednost ovog oscilatora se dobija primenom sledeće formule:

$$\mathrm{RSI}_t = 100 - \frac{100}{1 + \mathrm{RS}_t}; \mathrm{RS}_t = \frac{\sum_{n=t}^{t-d} \max(0, \mathrm{CP}_n - \mathrm{CP}_{n-1})}{\sum_{n=t}^{t-d} \left|\min(0, \mathrm{CP}_n - \mathrm{CP}_{n-1})\right|} \tag{8.5}$$

Vrednost ovog racija oscilira između 0 i 100. Ukoliko je vrednost bliža gornjoj granici, može se zaključiti da je porast vrednosti finansijskog instrumenta veći i učestaliji u odnosu na njihov pad, što takođe sugeriše da bi konkretan finansijski instrument trebalo prodati, jer se u takvoj situaciji očekuje povratak cene na prosečan nivo, odnosno očekuje se okretanje trenda (engl. *trend reversal*). Ukoliko se, pak, vrednost racija približava donjoj

vrednosti, to je signal da je pad vrednosti finansijskog instrumenta mnogo učestaliji, ali sledeći logiku okretanja trenda, konkretni finansijski instrument bi trebalo kupiti, jer će se u skorijoj budućnosti gubici neutralizovati zbog promene trenda.

Signali za trgovanje se mogu generisati na različite načine u zavisnosti od toga kako su postavljene granice [165]. U eksperimentalnom delu ove doktorske disertacije gornja i donja granica su postavljene na nivou od 70 i 30. Signal za kupovinu u trenutku t $(S_t = 1)$ se generiše u trenutku kada vrednost RSI pređe donju granicu od 30 odozdo, dok se signal za prodaju u trenutku t $(S_t = -1)$ generiše kada vrednost RSI pređe gornju granicu od 70 odozgo:

$$S_t = \begin{cases} 1 \ \ ako\ RSI_{t-1} \geq 30\ i\ RSI_t \geq 30 \\ -1\ ako\ RSI_{t-1} \leq 70 \ \ i \ \ RSI_t \leq 70 \end{cases} \tag{8.6}$$

U slučaju kada se vrednost RSI nalazi između postavljenih graničnih vrednosti, investitor samo zadržava prethodnu poziciju.

Robni indeks kanala (engl. *commodity channel index* – CCI) pripada grupi oscilatora i prvobitno se koristio za otkrivanje cikličnih promena vrednosti robe, ali je njegova primena proširena na tržište kapitala i tržište novca. CCI indeks meri razliku između uobičajene vrednosti finansijskog instrumenta u trenutku t (M_t) i prosečne vrednosti (SM_t). Ukoliko se pretpostavi da m iznosi 20 dana, vrednost CCI indeksa može se utvrditi primenom sledeće formule:

$$CCI_t = (M_t - SM_t) / 0.0015\ D_t \tag{8.7}$$

pri čemu D_t predstavlja vrednost standardne devijacije uobičajene vrednosti finansijskog instrumenta, dok se konstanta od 0,0015 koristi u svrhe skaliranja dobijenih rezultata.

Uobičajena vrednost finansijskog instrumenta u posmatranom trenutku t može se izračunati na sledeći način:

$$M_t = HP_t + CP_t + LP_t \tag{8.8}$$

pri čemu upotrebljeni simboli predstavljaju: HP_t - maksimalna vrednost, LP_t - minimalna vrednost i CP_t - vrednost finansijskog instrumenta na zatvaranju trgovanja u posmatranom trenutku t.

Prosečna vrednost finansijskog instrumenta (SM_t) obračunava se kao jednostavni pokretni prosek, dok se vrednost standardne devijacije uobičajene vrednosti finansijskog instrumenta izračunava na sledeći način:

$$D_t = \sum_{i=t-m+i}^{t} |M_i - SM_t| / m \tag{8.9}$$

Vrednost CCI indeksa oscilira oko nule, pri čemu su gornja i donja granica određene na nivou od +100 i -100, redom. U ovim granicama se javljaju normalne oscilacije vrednosti indeksa. Kada vrednost CCI indeksa pređe granicu od +100, smatra se da je trend promene vrednosti finansijskog instrumenta rastući, što predstavlja signal za kupovinu. Ovu poziciju u trgovanju treba zatvoriti u trenutku kada vrednost CCI indeksa padne ispod +100. Nasuprot tome, ukoliko vrednost CCI indeksa opadne ispod donje granice, trend promene vrednosti finansijskog instrumenta je opadajući, što predstavlja signal za prodaju. Porast vrednosti CCI indeksa iznad -100 predstavlja signal za kupovinu. Stoga se pravila trgovanja pomoću ovog indeksa mogu definisati na sledeći način:

$$S_t = \begin{cases} 1 \text{ako } \mathrm{CCI_t} > 100 \text{ ili } \mathrm{CCI_t} > \text{-}100 \\ \text{-}1 \text{ako } \mathrm{CCI_t} < 100 \text{ ili } \mathrm{CCI_t} < \text{-}100 \end{cases} \tag{8.10}$$

Indikator stope promene (engl. *rate of change – ROC*), koji pripada grupi oscilatora, meri brzinu kojom se trend promene vrednosti finansijskog instrumenta menja. Izračunava se kao procentualna promena vrednosti finansijskog instrumenta u konkretnom periodu u odnosu na izabrani prethodni period primenom sledeće formule:

$$ROC_t = 100\,((CP_t - CP_{t-n}) / CP_{t-n}) \tag{8.11}$$

S obzirom na to da se ovim indikatorom upoređuje tekuća vrednost finansijskog instrumenta (CP_t) u odnosu na vrednost iz *n*-tog perioda (CP_{t-n}), može se zaključiti da vrednost ovog indikatora fluktuira od pozitivne ka negativnoj. Pozitivna vrednost ROC indikatora ukazuje na povećanje vrednosti finansijskog instrumenta, dok negativna vrednost označava smanjenje vrednosti finansijskog instrumenta. Gornja vrednost indikatora nije definisana, što znači da investitori na tržištu mogu ostvariti neograničene prinose. Ubrzani rast vrednosti finansijskog instrumenta odražava se kroz povećanje vrednosti ovog indikatora. Donja granica se, pak, može odrediti na nivou vrednosti od -100% što znači da se vrednost finansijskog instrumenta može smanjiti do nule. Usporavanje rasta ili opadanje vrednosti finansijskog instrumenta uzrokuje smanjenje, odnosno negativnu vrednost ROC indikatora. Granice u kojima se kreće vrednost ovog indikatora omogućavaju identifikovanje uslova na tržištu u kojima je vrednost finansijskog instrumenta precenjena (engl. *overbought*) i potcenjena (engl. *oversold*). Za definisanje strategije trgovanja potrebno je odrediti referentnu vrednost u odnosu na koju će se meriti promena trenda. U ovoj doktorskoj disertaciji referentna vrednost iznosi 0, pa su pravila trgovanja postavljena na sledeći način:

$$S_t = \begin{cases} 1 \text{ ako } \mathrm{ROC_t} > 0 \\ -1 \text{ ako } \mathrm{ROC_t} < 0 \end{cases} \tag{8.12}$$

Prema ovom pravilu trgovanja, signal za kupovinu u trenutku t $(S_t = 1)$ je generisan kada vrednost ROC indikatora od negativne pređe u pozitivnu, odnosno bude veća od *0*. Ovakva situacija ukazuje na činjenicu da je vrednost finansijskog instrumenta bila potcenjena i da se u budućnosti može očekivati povećanje vrednosti. Signal za prodaju finansijskog instrumenta u trenutku t $(S_t = -1)$ dobijen je u suprotnoj situaciji – kada vrednost ROC indikatora od pozitivne pređe u negativnu, što predstavlja znak da je vrednost finansijskog instrumenta bila precenjena i da će u budućnosti njena vrednost opadati.

Parabolični SAR (engl. *stop and reverse* – SAR) je indikator koji investitorima omogućava da utvrde pravac i intenzitet promene vrednosti finansijskog instrumenta, kao i momenat kada će verovatnoća promene pravca biti najveća. Vrednost ovog indikatora može se izračunati primenom sledeće formule:

$$\mathrm{SAR}_{t+1} = \mathrm{SAR}_t + \alpha(\mathrm{EP} - \mathrm{SAR}_t) \tag{8.13}$$

pri čemu upotrebljeni simboli imaju sledeće značenje: *EP* – ekstremna tačka koja označava najveću vrednost finansijskog instrumenta u obračunskom periodu, ukoliko je trend rastući, odnosno najmanju vrednost finansijskog instrumenta u obračunskom periodu, ukoliko je trend opadajući; α – faktor akceleracije čija je inicijalna vrednost 0,02, ali se povećava svaki put kada vrednost finansijskog instrumenta u obračunskom periodu dostigne novi ekstremni nivo.

U slučaju kada dolazi do povećanja vrednosti finansijskog instrumenta, predviđena vrednost ovog indikatora je manja od vrednosti finansijskog instrumenta i konvergira ka njoj. U obrnutom slučaju, kada je trend vrednosti finansijskog instrumenta opadajući, predviđena vrednost ovog indikatora je veća od vrednosti finansijskog instrumenta i konvergira ka njoj. Sistem pravila trgovanja, koja se baziraju na ovom indikatoru, jeste funkcija pravca trenda vrednosti finansijskog instrumenta i vremena u okviru kojeg se promena dešava, a signali za trgovanje finansijskim instrumentom mogu se generisati na sledeći način:

$$S_t = \begin{cases} 1 \ \textit{ako} \ CP_t > SAR_t \\ -1 \ \textit{ako} \ CP_t < SAR_t \end{cases} \tag{8.14}$$

Ukoliko je vrednost SAR manja od vrednosti finansijskog instrumenta u trenutku *t*, trend promene vrednosti finansijskog instrumenta je rastući, što predstavlja signal za kupovinu $(S_t = 1)$. U suprotnom slučaju, vrednost SAR veća od vrednosti finansijskog instrumenta u trenutku *t* ukazuje na opadajući trend promene vrednosti finansijskog instrumenta i signalizira prodaju konkretnog instrumenta $(S_t = -1)$.

Stohastički oscilator (engl. *stochastic oscilator* – *S*O) je indikator pravca i intenziteta promene trenda vrednosti finansijskog instrumenta, koji je određen sa dva pokazatelja: $\%K_t$ – nivo potpore (engl. *support level*) i *%D* – nivo otpora (engl. *resistance level*). Nivo potpore predstavlja promenu vrednosti ispod koje se ne očekuje dalji pad i može se izračunati na sledeći način:

$$\%K_t = 100((CP_t - LP_n)/(HP_n - LP_n)) \tag{8.15}$$

U praksi se ovaj pokazatelj najčešće obračunava za period od 5, 9 i 14 dana, a u ovoj doktorskoj disertaciji period za koji se obračunava *n* iznosi 14 dana.

Nivo otpora, nasuprot prethodnom, označava trenutak kada bi ponuda trebalo da prevaziđe tražnju, a vrednost finansijskog instrumenta počne da opada, i može se izračunati na sledeći način:

$$\%\mathrm{D} = \frac{1}{\mathrm{t}}\sum_{\mathrm{t}=1}^{3}\%\mathrm{K}_{\mathrm{t}} \tag{8.16}$$

Osnovna pretpostavka od koje se polazi prilikom generisanja signala trgovanja jeste da će vrednost finansijskog instrumenta oscilirati blizu utvrđenih nivoa pre nego što promeni trend. U ovoj doktorskoj disertaciji donji granični nivoi je određen na 20, a gornji na 80, pa su u skladu sa tim definisana i pravila trgovanja na sledeći način:

$$S_t = \begin{cases} 1 \ \ ako \%D < 0.2 \ i \ \%K_t > \%D \\ -1 \ \ ako \%D > 0.8 \ \ i \ \ \%K_t < \%D \end{cases} \tag{8.17}$$

U slučaju kada je nivo otpora manji od donjeg limita i nivo potpore u trenutku *t* veći od nivoa otpora, može se zaključiti da je vrednost finansijskog instrumenta potcenjena i da će se u budućnosti povećavati, zbog čega se signalizira kupovina u datom trenutku $t\,(S_t = 1)$. U suprotnom slučaju, ako je u trenutku *t* nivo otpora veći od gornjeg limita i nivo potpore manji od nivoa otpora, generiše se signal za prodaju finansijskog instrumenta $(S_t = \text{-}1)$, jer ovakva situacija ukazuje na to da je vrednost finansijskog instrumenta precenjena i da će u budućem periodu doći do njenog smanjenja.

8.1.2. OSNOVNE POSTAVKE SIMULACIJE

Prvi korak prema algoritmu 7.1 je da se obračunaju vrednosti za vektor odnosa značajnosti kriterijuma - RVV. Kriterijumi za procenu značajnosti rizika i bruto prinosa su zasnovani na pretpostavkama iz teorije ekonomske nauke odnosno ponašanje učesnika na finansijskim tržištima da su investitori uglavnom osetljivi na rizik [87] i [98]. Treći kriterijum, stopa predviđanja, procenjen je kao najznačajniji, uzevši u obzir činjenicu da se

izbor atributa vrši sa ciljem povećanja preciznosti predikcionog modela.

Inicijalne težine mogu biti dodeljene ili na osnovu mišljenja donosioca odluke ili mogu naći utemeljenje u ekonomskoj teoriji. Prilikom odabira opsega podataka koji će učestvovati u AHP proračunima razmatrani su dugoročni obrasci promene vrednosti finansijskih instrumenata koji se mogu prepoznati u cikličnim promenama vrednosti tržišnih indeksa, a koji takođe imaju svoje utemeljenje u ekonomskoj teoriji [74] i [122].

Za potrebe AHP proračuna posmatran je četvorogodišnji period počevši od početka 2009. godine do kraja 2012. godine, čime su u obzir uzeti dugoročni obrasci promene vrednosti finansijskih instrumenata koji se mogu uočiti na osnovu teorijskih pretpostavki o ciklusnom kretanju na tržištima kapitala [74].

TABELA 8.2 Matrica poređenja značajnosti kriterijuma

	RETURN	RISK	HR	RVV
RETURN	1	1/4	1/6	0.082
RISK	4	1	1/4	0.236
HR	6	4	1	0.682
$\lambda_{max} = 3.1078$				
$CR = 0.09297$				

Sopstveni vektor odnosa značajnosti kriterijuma RVV, izračunat je prema prethodno opisanom postupku u poglavlju 6.2 prema relaciji 6.5 i dobijene su vrednosti $\text{RVV} = (0.082, 0.236, 0.682)^{\text{T}}$. Dobijeni brojevi ukazuju redom na relativnu značajnost svakog od kriterijuma posebno, bruto prinosa, sistemskog rizika i stope predviđanja. Dobijena vrednost od 0.682 upućuje na to da postavljeni model najviše vrednuje kriterijum stope predviđanja. Vrednost 0.236 ukazuje da se sistemski rizik vrednuje manje i na kraju vrednost 0.082 upućuje da model najmanje vrednuje značaj bruto prinosa. Odnos konzistenstnosti CR je 0,09297, što je manje od vrednosti kritičnog limita 0.1, te se može zaključiti da je model konzistetntan u izborima.

Prethodno smo spominjali termine tehnički indikatori i strategije trgovanja. Kako bi se dalje pojednostavila notacija ova dva termina će se smatrati sinonimima iako model vrši izbor tehničkih indikatora.

Simulacije su izvođene nad podacima o vrednosti tržišnih indeksa sa Beogradske berze (BELEX15), tržišta kapitala SAD (S&P 500) i Londonske berze (FTSE 100). Vrednost indeksa određuju cene najlikvidnijih akcija kojima se trguje na regulisanom tržištu posmatranih berzi. Empirijski podaci koji se koriste za analizu finansijskih tržišta evidentiraju se u fiksnim vremenskim intervalima, uobičajeno na dnevnom nivou. Podaci su preuzeti sa

https://finance.yahoo.com/ servisa.

8.2. INDEKS BELEX15

BELEX15 je vodeći indeks Beogradske berze, čiju vrednost određuju cene najlikvidnijih akcija, kojima se trguje na regulisanom tržištu Beogradske berze. Indeks je ponderisan tržišnom kapitalizacijom i obračunava se u realnom vremenu. Finansijska tržišta u razvoju, gde se prema stepenu razvijenosti može svrstati i finansijsko tržište Srbije, smatraju se perspektivnim tržištima, s obzirom na to da ova tržišta predstavljaju značajan alternativni izvor investicionih mogućnosti za strane investitore.

Za potrebe simulacija podaci su podeljeni u dve grupe. Prva grupa sastoji se od zapisa koji su korišćeni za treniranje modela, i obuhvata period od 26. oktobra 2005. godine do 31. decebra 2012. godine. Trening skup za BELEX15 indeks obuhvata 1793 trgovinskih dana. Za potrebe testiranja modela uzeto je 252 dana odnosno cela godina trgovanja, u priodu od 3. januara 2013. godine do 31. decembra iste godine. Rezultati predstavljaju predikciju tipa jedan korak unapred na produženom vremenskom horizontu.

Radi dobijanja kompletne i jasne slike o statističkim osobinama posmatranog berzanskog indeksa, potrebno je obračunati i analizirati osnovne statističke osobine posmatranog skupa atributa.

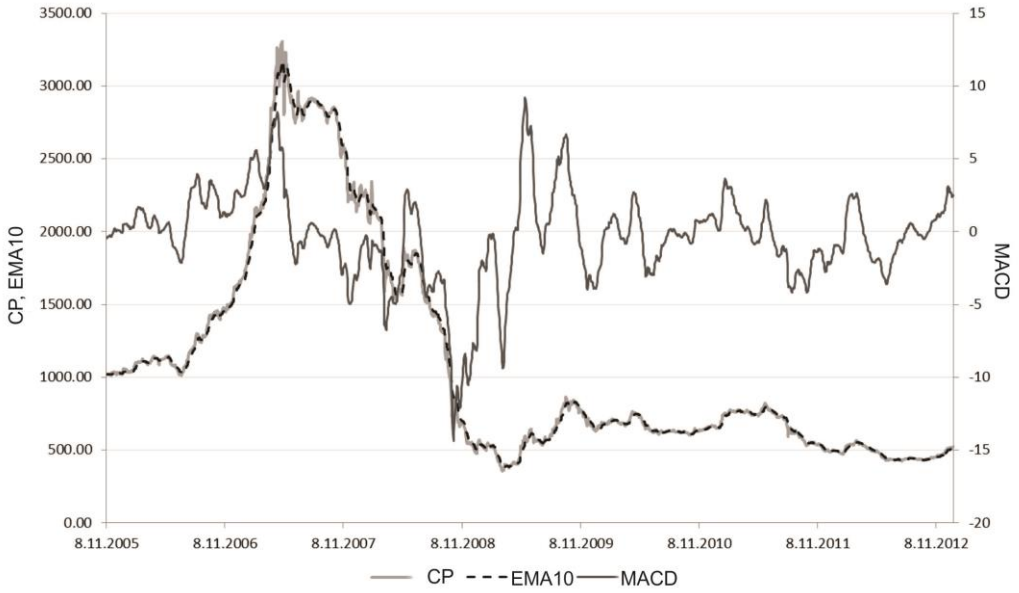
U tabeli 8.3 predstavljena je osnovna deskriptivna statistika BELEX15 indeksa. U tabeli 8.3 su predstavljene najmanje i najveće verdnosti za svaki posmatrani atribut. Zatim aritmetička sredina serija, kao izabrana mera centralne tendencije i prikazana je standardna devijacija kao jedna od mera disperzije u cilju prikazivanja varijabiliteta uzorka.

TABELA 8.3 BELEX15 deskriptivna statistika selektovanih atributa

TEHNIČKI INDIKATORI	BELEX15			
	MIN	MAX	MEAN	STDEV
ROC	-37.41	38.93	-0.10	7.16
CCI	-400.8	454.5	0.79	118.3
RSI	0	100	49.46	23.95
%K	0	100	48.97	33.32
%D	1,33	99.04	48.98	31.52
EMA1	354.39	3304.6	1028.56	722.9
EMA10	378.61	3173.1	1029.60	721.3
MACD	-147.2	230.14	-1.66	39.15
SAR	347.46	3335.2	1033.14	727.3

Na osnovu podataka iz prethodne tabele mogu se učiti opsezi kretanja vrednosti posmatranih atributa.

Na slici 8.3 može se videti modelovanje posmatrane vremenske serije BELEX15 indeksa pomoću EMA i MACD indikatora.



SLIKA 8.3 Odnos vrednosti indeksa na zatvaranju i korišćenih tehničkih indikatora

Na osnovu slike 8.3, može se uočiti da prethodno navedene transformacije doprinose stacionarnosti serije podataka. Na taj način dodatno se utiče na poboljšanje efikasnosti korišćenog algoritma mašinskog učenja i povećanje brzine izračunavanja numeričkih kalkulacija.

8.2.1. IZBOR ATRIBUTA

Prema AHP proračunima predstavljenim u pogavlju 6.2 ove doktorske disertacije, na osnovu pojedinačnih matrica odnosa značajnosti atributa, vrši se poređenje ulaznih atributa u odnosu na definisane kriterijume evaluacije: bruto prinos, sistemski rizik i stopu predviđanja.

TABELA 8.4 Matrica parnog upoređivanja po prinosu

PRINOS		EMA	MACD	RSI	CCI	SO	ROC	SAR
917.85	EMA	1	1/4	1	1/3	5	1/2	1/5
1083.85	MACD	4	1	5	2	9	3	1
436.84	RSI	1	1/5	1	1/4	4	1/3	1/6
843.05	CCI	3	1/2	4	1	7	2	1/3
-70.73	SO	1/5	1/9	1/4	1/7	1	1/6	1/9
710.05	ROC	2	1/3	3	1/2	6	1	1/4
1108.59	SAR	5	1	6	3	9	4	1
$\lambda_{max} = 7.3808$								
$CR = 0.0481$								

Sopstveni vektor matrice je $(0.1593, 0.2386, 0.0564, 0.1336, 0.0204, 0.0932, 0.2985)^T$, i vrednost CR je 0.0481, tako da se proces odlučivanja može smatrati konzistentnim.

TABELA 8.5 Matrica parnog upoređivanja u odnosu na rizik

RIZIK		EMA	MACD	RSI	CCI	SO	ROC	SAR
18.391	EMA	1	1/6	1/9	1/6	1/9	1/6	1/6
8.092	MACD	6	1	1/4	1	1/3	1	1
2.704	RSI	9	4	1	4	1	4	4
8.120	CCI	6	1	1/4	1	1/3	1	1
3.686	SO	9	3	1	3	1	3	3
8.132	ROC	6	1	1/4	1	1/3	1	1
8.088	SAR	6	1	1/4	1	1/3	1	1
$\lambda_{max} = 7.2555$								
$CR = 0.0322$								

Sopstveni vektor matrice je $(0.0212, 0.1000, 0.2738, 0.1219, 0.2831, 0.1000, 0.1000)^T$, i vrednost CR je 0.0322. Kao rezultat, proces odlučivanja može se smatrati konzistentnim.

TABELA 8.6 Matrica parnog upoređivanja po kriterijumu preciznosti

HR		EMA	MACD	RSI	CCI	SO	ROC	SAR
0.580	EMA	1	1	9	1	9	1	1
0.567	MACD	1	1	9	1	8	1	1
0.050	RSI	1/9	1/9	1	1/9	1	1/9	1/9
0.553	CCI	1	1	9	1	8	1	1
0.107	SO	1/9	1/8	1	1/8	1	1/8	1/8
0.544	ROC	1	1	9	1	8	1	1
0.574	SAR	1	1	9	1	8	1	1
$\lambda_{max} = 7.5604$								
$CR = 0.0707$								

Sopstveni vektor matrice je $(0.2024, 0.1991, 0.0308, 0.1454, 0.0241, 0.1991, 0.1991)^T$, dok je CR vrednost 0.0707. Kao rezultat, proces odlučivanja može se smatrati konzistentnim.

Prema postupku definisnom u poglavlju 6.2 naredni korak predstavlja kreiranje OPM matrice sa lokalnim težinama za svaki pojedinačni atribut. Poslednji korak predstavlja množenje matrica OPM i RVV kako bi se izračunale globalne težine. Znači, u poslednjem koraku se kreira matrica odlučivanja i uz pomoć nje obračunavaju težine za posmatrane ulazne atribute.

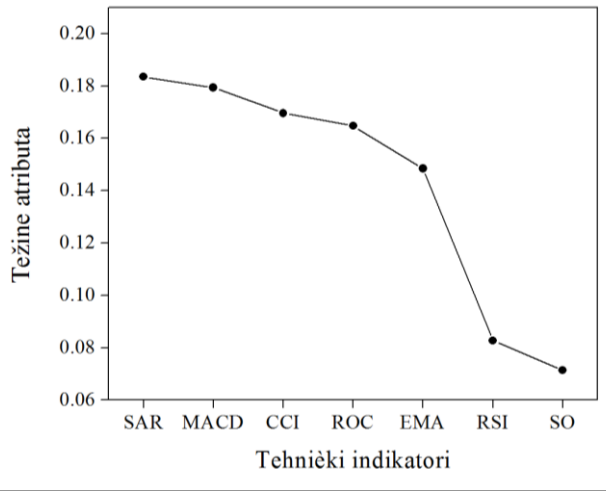
U tabeli 8.7 predstavljene su konačno dobijene vrednosti težina prema opisanom postupku.

TABELA 8.7 BELEX15 matrica globalnih težina

	TEŽINE ZA PRINOS (0.082)	TEŽINE ZA RIZIK (0.236)	TEŽINE ZA HR (0.682)	TEŽINE ATRIBUTA
EMA	0.1593	0.0212	0.2024	0.1561
MACD	0.2386	0.1000	0.1991	0.1789
RSI	0.0564	0.2738	0.0308	0.0903
CCI	0.1336	0.1219	0.1454	0.1389
SO	0.0204	0.2831	0.0241	0.0850
ROC	0.0932	0.1000	0.1991	0.1670
SAR	0.2985	0.1000	0.1991	0.1838

Prema kalkulacijama, dobijene su sledeće vrednosti za relevatnost svakog od atributa: SAR-0.1838, MACD - 0.1789, ROC - 0.1670, EMA - 0.1561, CCI - 0.1389, RSI - 0.0903, i SO - 0.0850.

Na osnovu definisanih koraka u poglavlju 7.2 ove doktorske disertacije, vrši se uređenje dobijenih vrednosti težina atributa, kako bi se dobila konačna procena značajnosti svakog od ulaznih atributa Na slici 8.4 grafički su prikazane težine i relevatnost svakog od atributa.



SLIKA 8.4 BELEX15 opadajući redosled dobijenih težina atributa

Nakon izračunavanja vrednosti težina za posmatrani skup atributa, analizirane su dobijene vrednosti, slika 8.4, i izvršen je izbor podskupa skupa atributa za obuku predikcionog modela. Na osnovu predstavljenih grafikona može se zaključiti da težine indikatora značajno opadaju nakon drugo rangiranog atributa za BELEX15 indeks.

Kao rezultat analize, selektovana su prva dva indikatora kao ulazni atribitu za kreiranje predikcionog modela.

8.2.2. KOMPARATIVNA ANALIZA DOBIJENIH REZULTATA

Predložena strategija za izbor atributa je dodatno poređena sa različitim metodama izbora atribura: primenom metoda zajedničkih informacija (engl. *mutual information* - MI) sa sekvencijalnom *forward-backward* selekcijom atributa [58], *random forest* (RF) za izbor atributa [53] i linerni klasifikator sa *sequential forward* metodom izbora atributa (engl. *linear discriminant classifier* - LDC) [66]. U tabeli 8.8 predstavljeni su podsupovi skupova atributa dobijeni primenom različitih metoda za izbor atributa u slučaju BELEX15 indeksa.

TABELA 8.8 BELEX15 komparativna analiza metoda izbora atributa

METODA IZBORA	IZABRANI ATRIBUTI	BROJ ATRIBUTA
AHP	SAR, MACD	2
MI	%K, %D	2
RF	CCI, RSI,%K,%D	4
LDC	CCI,RSI,%K,%D, MACD	5

Na osnovu podataka u tabeli 8.8 može se zaključiti da u zavisnosti od izabranog metoda za izbor atributa, varira broj i sam podskup selektovanih atributa. Primenom predloženog metoda bio bi izabran podskup od dva ulazna atributa isto kao i primenom MI metoda izbora, dok bi RF i LDC selektovali podskup od 4 i 5 atributa respektivno.

U cilju testiranja rezultata predložene metodologije izbora atributa kreirano je više različitih predikcionih modela. Najpre je izvršena komparativna analiza predikcionih rezultat dobijenih predloženim pristupom izbora atributa sa inkorporacijom težina u kernel kod metoda podržavajućih vektora - SVM i metoda najmanjih kvadrata podržavajućih vektora - LS-SVM u odnose na druge metode izbora atributa u kombinaciji sa istim predikcionim modelima.

Kako je navedeno u [55] na osnovu dvadeset različitih skupova podataka najbolju opštu stopu predviđanja dali su LS-SVM klasifikatori sa RBF kernelom. Pored toga u slučajevima kada je broj primera za klasifikaciju mnogo veći od broja atributa (dimenzija vektora - prostora) takođe se preporučuje korišćenje RBF kernela. Za formiranje LS-SVM predikcionog modela korišćena je biblioteka LS-SVMlab [15]. Pri čemu je radi analize šire primenjivosti predloženog metoda izbora atributa, korišćen linearni kernel kod SVM metoda, kod koga je parametar C fiksiran na C=1. Za formiranje SVM predikcionog modela korišćena je biblioteka LibSVM [24].

Posledično su kreirani sledeći modeli označeni skrećenicom navedenih pristupa selekcije atributa u tabeli i korišćenog predikcionog metoda.

- MI-LS-SVM – model obučavan sa podskupom atributa selektovanim na osnovu pristupa zajedničkih informacija
- RF-LS-SVM - LS-SVM model obučavan nad poskupom atributa izabranog pomoću RF algoritma
- LDC-LS-SVM - LS-SVM treniran na osnovu atributa dobijenih primenom *forward selection* metode i LDC (engl. *linear discriminant classifier*) klasifikatora
- AHP-WK-LS-SVM – implementira metod za izbor atributa predložen u ovoj doktorskoj disertaciji zajedno sa težinskom kernel funkcijom LS-SVM modela
- AHP-WK-SVM – model koristi podskup atributa selektovan primenom AHP metode i inkorporira dobijene težine u linerani SVM model.

Modeli su kreirani korišćenjem MATLAB alata uz korišćenje dodatnih biblioteka gde je potrebno LS-SVMlab [15], LibSVM [24] i MILCA – MI [79].

Rezultati predstavljaju najbolje dobijene vrednosti predikcije tipa jedan korak unapred na produženom vremenskom horizontu. Ista komparativna osnova korišćena je za procenu poboljšanja predloženog modela u svim nadalje prikazanim komparacijama u okviru ovog poglavlja.

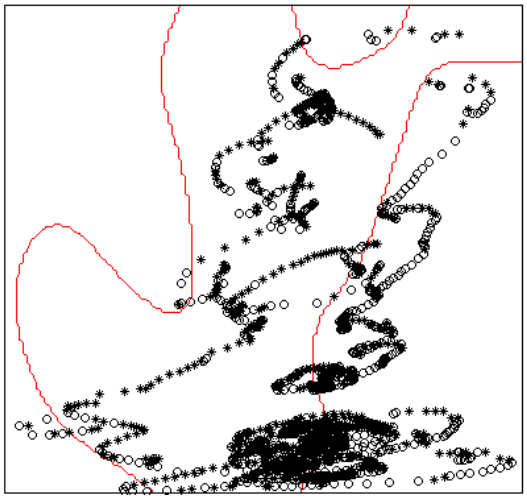
TABELA 8.9 BELEX15 predikcija sa različitim metodama izbora atributa

PREDIKCIONI MODEL	HR
LS-SVM	53.57
MI-LS-SVM	52.38
RF-LS-SVM	52.38
LDC-LS-SVM	53.17
AHP-WK-LS-SVM	61.11
SVM	53.98
AHP-WK-SVM	57.54

Na osnovu rezultata iz tabele 8.9 može se zaključiti da se primenom predložene metodologije dobijaju značajna poboljšanja na posmatranom BELEX15 skupu podataka. U odnosu na LS-SVM metod bez izbora atributa razlika u stepenu predviđenja iznosi 7.54%. Poboljšanja su redom uočljiva i kod ostalih *benchmark* modela i kreću se od 8.73% za MI-LS-SVM i RF-LS-SVM model do 7.94% za LDC-LS-SVM model.

Pored primene dobijenih težina i podskupa skupa atributa u predloženom AHP-WK-LS-SVM predikcionom modelu, na osnovu tabele 8.9 može se uvideti i da inkorporacija dobijenih AHP težina u SVM linerana kernel vodi do poboljšanja i kod AHP-WK-SVM predikcionog modela, u vrednosti od 3.56%.

S obzirom na to da je dva atributa odabrano primenom metodologije predložene u okviru ove doktorske disertacije za dalju obuku predikcionog metoda moguće je izvršiti i vizualizaciju granice razdvajanja, te se na slici 8.5 može uočiti izrazita adaptibilnost predloženog modela.



SLIKA 8.5 Mapiranje atributa i granica razdvajanja AHP-WK-LS-SVM metoda

Na slici 8.5 predstavljena je granica razdvajanja dobijena primenom AHP-WK-LS-SVM modela sa parametrima σ^2=1.369 and γ=30.9793 i jasno se može videti kako izgleda adaptilnost modela.

Konačno kako bi se procenilo poboljšanje predloženog modela, izvršena je komparativna analiza dobijenih rezultata sa rezultatima dobijenim drugim klasifikacionim algoritmima.

Poređene su preciznosti sa Random Forest (RF) [16], Linear SVM [24] i veštačke neuronske mreže, ANN.

Za kreiranje stabla odlučivanja korišćeno je 1000 stabala, i broj atributa za svaku podelu je postavljen na vrednost kvadratnog korena dimenzionalnosti atributa. Prilikom simulacija korišćena je ANN mreža sa dva skrivena sloja od po 100 neurona. Poslednji model predstavlja takozvani RW – *Random Walk* model koji koristi trenutnu vrednost da predvidi buduću, pretpostavljajući da će vrednost u narednom periodu (y_{t+1}) biti jednaka tekućoj vrednosti (y_t).

Rezultati stope predviđanja dobijenih na posmatranom skupu vrednosti na osnovu inicijalne podele podataka na 90 procenata veličine skupa za obuku modela i 10 procenata veličine dostupnog skupa za testiranje modela predstavljeni su u tabeli 8.20. Svi modeli su postavljeni na osnovu jedinstvene eksperimentalne postavke u okviru korišćenog skupa

podataka.

Modeli su kreirani korišćenjem MATLAB alata uz korišćenje dodatnih biblioteka gde je potrebno LS-SVMlab [15], LibSVM [24].

Konačno izvršena je i komparativna analiza predloženih predikcionih modela sa drugim modelima nadgledanog mašinskog učenja.

Kako bi se procenilo poboljšanje predloženog modela, izvršena je komparativna analiza dobijenih rezultata sa rezultatima dobijenim drugim klasifikacionim algoritmima. Poređene su preciznosti sa *Random Forest* (RF) [16], *Linear SVM* [24] i veštačke neuronske mreže, *ANN*.

Za kreiranje stabla odlučivanja korišćeno je 1000 stabala, i broj atributa za svaku podelu je postavljen na vrednost kvadratnog korena dimenzionalnosti atributa. Za linearni SVM parametar C je fiksiran na C=1, dok je za ANN korišćena mreža sa dva skrivena sloja od po 100 neurona. Model RW – *Random Walk* model koristi trenutnu vrednost da predvidi buduću, pretpostavljajući da će vrednost u narednom periodu (y_{t+1}) biti jednaka tekućoj vrednosti (y_t).

Rezultati stope predviđanja dobijenih na posmatranom skupu vrednosti na osnovu inicijalne podele podataka na 90 procenata veličine skupa za obuku modela i 10 procenata veličine dostupnog skupa za testiranje modela predstavljeni su u tabeli 8.10. Svi modeli su postavljeni na osnovu jedinstvene eksperimentalne postavke u okviru korišćenog skupa podataka.

Modeli su kreirani korišćenjem MATLAB alata uz korišćenje dodatnih biblioteka gde je potrebno LS-SVMlab [15] i LibSVM [24].

TABELA 8.10 BELEX15 komparativna analiza predikcionih modela

PREDIKCIONI MODEL	HR
AHP-WK-LS-SVM	61.11
SVM	53.98
AHP-WK-SVM	57.54
ANN	53.97
RF	50.00
RW*	50.00

Na osnovu podataka predstavljenih u tabeli 8.10 može se zaključiti da za BELEX15 skup podataka postoje određena poboljšanja koja se postižu primenom predloženog AHP-WK-LS-SVM predikcionog modela. U odnosu na RW i RF model pobljšanja su značajna od razlika dobijenih vrednosti je 11.11%, razlike su manje izražene u odnosu na ANN i SVM modela kod kojih su dobijene vrednosti poboljšanja od 7.13%.

8.3. INDEKS S&P 500

S&P 500 je berzanski indeks čiju vrednost određuje 500 najlikvidnijih akcija Njujorške berze (NASDAQ). Ovaj indeks odražava performanse – rizik i prinos, hartija od vrednosti koje emituju kompanije velike tržišne kapitalizacije a koje posluju na teritoriji SAD, zbog čega se smatra najboljim reprezentom tržišta kapitala SAD. Izbor komponenti ovog indeksa određen je pravilima, koja podrazumevaju da kompanije ispune kriterijume u pogledu tržišne kapitalizacije, likvidnosti, sedišta, sektorske klasifikacije, dužine trgovanja akcijama na berzi, broja akcija kojima se trguje i finansijske održivosti. Prilikom utvrđivanja vrednosti indeksa, učešće pojedinih hartija od vrednosti određeno je njihovom tržišnom kapitalizacijom, koja se obračunava isključivo na osnovu broja akcija kojima se javno trguje na berzi. Vrednost indeksa se obračunava u realnom vremenu i tokom trajanja sesija trgovanja obelodanjuje na svakih 15 sekundi. Za indeks S&P 500, trening skup se sastoji od 1775 zapisa.

Radi dobijanja kompletne i jasne slike o statističkim osobinama izabranog berzanskog indeksa, ponovo su obračunate i analizirane osnovne statističke osobine posmatranog skupa atributa. U tabeli 8.11 predstavljena je osnovna deskriptivna statistika.

TABELA 8.11 S&P 500 deskriptivna statistika selektovanih atributa

TEHNIČKI INDIKATORI	S&P500			
	MIN	MAX	MEAN	STDEV
ROC	-25.19	20.57	0.29	3.91
CCI	-396.8	286.42	26.32	108.4
RSI	9.34	99.3	54.80	15.59
%K	0	100	61.29	31.54
%D	1.57	99.15	61.27	28.18
EMA_1	676.53	1842	1297.58	218.9
EMA_{10}	713.78	.1824	1296.32	216.4
MACD	-77.2	25.81	1.94	14.98
SAR	666.79	1813.6	1290.96	219

Na osnovu podataka iz prethodne tabele mogu se uočiti opsezi kretanja vrednosti posmatranih atributa.

8.3.1. IZBOR ATRIBUTA

U cilju izbora atributa, na osnovu pojedinačnih matrica odnosa značajnosti atributa, vrši se poređenje ulaznih atributa u odnosu na definisane kriterijume evaluacije: bruto prinos, sistemski rizik i stopu predviđanja. U poslednjem koraku kreira se matrica odlučivanja i uz

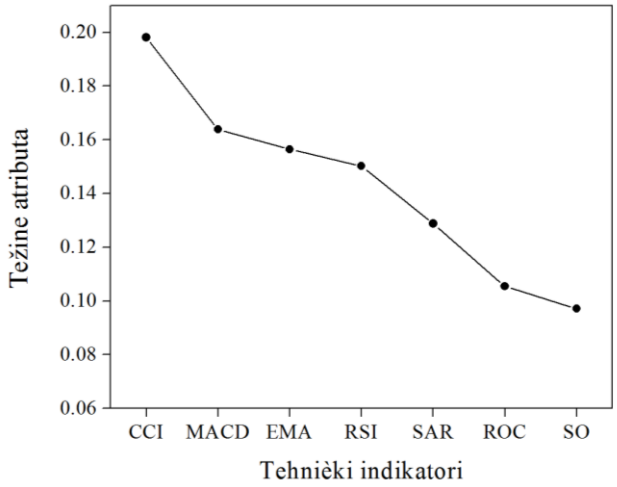
pomoć nje obračunavaju globalne težine za posmatrane ulazne atribute. Tabela 8.12 predstavlja konačno dobijene vrednosti težina prema opisanom postupak.

TABELA 8.12 S&P 500 matrica globalnih težina

	PRINOS (0.082)	RIZIK (0.236)	HR (0.682)	TEŽINE ATRIBUTA
EMA	0,039	0,047	0,209	0,157
MACD	0,132	0,046	0,209	0,164
RSI	0,350	0,451	0,022	0,150
CCI	0,334	0,047	0,234	0,198
SO	0,083	0,315	0,023	0,097
ROC	0,025	0,047	0,135	0,105
SAR	0,037	0,168	0,168	0,129

Na osnovu finalnih kalkulacija, vrši se uređenje dobijenih vrednosti težina atributa, kako bi se dobila konačna procena značajnosti svakog od ulaznih atributa. Na slici 8.6 može se videti uređenje dobijenih vrednosti.

Nakon izračunavanja vrednosti težina za posmatrani skup atributa, na osnovu algoritma predstavljenog u poglavlju 7.2 analiziran je dobijeni grafikon predstavljen na slici 8.6 i izvršen je izbor podskupa skupa atributa za obuku predikcionog modela.



SLIKA 8.6 S&P 500 opadajući redosled dobijenih težina atributa

Na osnovu predstavljenih grafikona može se zaključiti da težine indikatora značajno opadaju nakon drugo rangiranog atributa za S&P 500 indeks. Kao rezultat analize, selektovana su prva dva indikatora kao ulazni atribiti za kreiranje predikcionog modela za S&P 500.

8.3.2. KOMPARATIVNA ANALIZA DOBIJENIH REZULTATA

Predložena strategija za izbor atributa je kao i u slučaju BELEX15 indeksa dodatno poređena sa različitim metodama izbora atributa, podsetimo: primenom metoda zajedničkih inofrmacija (engl. *mutual information* - MI) sa sekvencijalnom *forward-backward* selekcijom atributa [58], *random forest* (RF) za izbor atributa [53] i linerni klasifikator sa *sequential forward selection* (engl. *linear discriminant classifier - LDC*) [66]. U tabeli 8.13 predstavljeni su podskupovi skupova atributa dobijenih primenom različitih metoda za izbor atributa.

TABELA 8.13 S&P 500 komparativna analiza metoda izbora atributa

METOD IZBORA	IZABRANI ATRIBUTI	BROJ ATRIBUTA
AHP	CCI, MACD	2
MI	%K,%D,EMA_1,EMA_{10}, SAR	5
RF	ROC, CCI, RSI,%K,%D	5
LDC	%K, %D,MACD, SAR	4

Na osnovu podataka iz tabele 8.13 može se zaključiti da u zavisnosti od izabranog metoda za izbor atributa, varira broj kao i sam podskup selektovanih atributa. Konkretno, primenom predloženog AHP pristupa izbora atributa selektovana su dva atributa iz polaznog skupa, CCI i MACD, dok je primenom svih ostalih metoda selektovan veći broj atributa za obuku predikcionog modela. Primenom MI i RF selektovano je pet atributa, dok je primenom LDC selektovano četiri atributa.

Dalje su u cilju komparativne analize dobijenih rezultata kreirani isti predikcioni modeli kao i uslučaju BELEX15 indeksa i dobijene vrednosti predstavljene su u tabelama 8.13 i 8.15.

TABELA 8.14 S&P 500 predikcija sa različitim metodama izbora atributa

PREDIKCIONI MODEL	S&P500
LS-SVM	53.97
MI-LS-SVM	53.18
RF-LS-SVM	51.19
LDC-LS-SVM	51.98
AHP-WK-LS-SVM	57.14
SVM	53.57
AHP-WK-SVM	54.76

Na osnovu rezultat iz tabele 8.14 može se zaključiti da se primenom predložene metodologije dobijaju značajna poboljšanja na posmatranom S&P 500 skupu podataka. U odnosu na LS-SVM metod bez izbora atributa razlika u dobijenim stopama predviđanja

iznosi 3,17%. Poboljšanja su redom uočljiva i kod ostalih *benchmark* modela i kreću se od 3.96% za MI-LS-SVM, preko 5.16% za LDC-LS-SVM model do 5.95% za RF-LS-SVM model do.

Pored primene dobijenih težina i podskupa skupa atributa u predloženom AHP-WK-LS-SVM predikcionom modelu, na osnovu tabele 8.14 može se uvideti i da inkorporacija dobijenih AHP težina u SVM sa lineranim kernelom vodi do poboljšanja i kod AHP-WK-SVM predikcionog modela, u vrednosti od 1.19%. u odnosu na SVM model bez izbora atributa.

TABELA 8.15 S&P 500 komparativna analiza predikcionih modela

PREDIKCIONI MODELI	S&P 500
AHP-WK-LS-SVM	57.14
SVM	53.57
AHP-WK-SVM	54.76
ANN	58.33
RF	51.19
RW	48.41

Na osnovu podataka predstavljenih u tabeli 8.15 može se zaključiti da za S&P 500 skup podataka postoje određena poboljšanja koja se postižu primenom predloženog AHP-WK-LS-SVM predikcionog modela. U odnosu na RW i RF model razlike su značajne u iznosu od 8.73% i 5.95% respektivno. Kod S&P 500 skupa podataka, u poređenju sa veštačkim neuronskim mrežama AHP-WK-LS-S daje nešto slabije rezultate, za 1.19% je niža stopa predviđanja.

Prema podacima iz tabele 8.15 može se videti i da inkorporacija dobijenih AHP težina vodi do poboljšanji i kod AHP-WK-SVM predikcionog modela za jedan procenat kod berzanskog indeksa S&P500.

8.4. INDEKS FTSE 100

FTSE 100 je indeks Londonske berze i jedan od vodećih svetskih indeksa za praćenje vrednosti hartija od vrednosti. Vrednost ovog indeksa određuje 100 najlikvidnijih akcija, koje u ukupnoj tržišnoj kapitalizaciji Londonske berze učestvuju sa oko 80%. Akcije velikih kompanija imaju veće učešće u vrednosti indeksa, jer je indeks ponderisan tržišnom kapitalizacijom. Obračunava se u realnom vremenu, a u periodu kada je berza aktivna podaci o ovom indeksu se obelodanjuju na svakih 15 sekundi. Iako se na ovaj indeks gleda kao na opšti pokazatelj stanja privrede i kompanija UK, deo kompanija, čije su akcije uključene u

korpu ovog indeksa, posluje u državama širom sveta, zbog čega je globalno prihvaćen. Trening skup za FTSE 100 indeks obuhvata 1843 trgovinskih dana.

Radi dobijanja kompletne i jasne slike o statističkim osobinama izabranog berzanskog indeksa, kao i u prethdonim slučajevima, obračunate su i analizirane osnovne statističke osobine posmatranog skupa atributa.

TABELA 8.16 FTSE 100 deskriptivna statistika selektovanih atributa

TEHNIČKI INDIKATORI	FTSE 100			
	MIN	MAX	MEAN	STDEV
ROC	-22.83	15.88	0.17	3.83
CCI	-343.94	272.93	17.36	108.6
RSI	9.06	98.52	53.35	15.92
%K	0	100	58.19	31.1
%D	2.61	99.69	58.18	27.94
EMA_1	3512.1	6840.3	5701.21	674.1
EMA_{10}	3670.5	6725.8	5698.79	665.9
MACD	-318.28	123.77	3.80	62.97
SAR	3460.7	6875.6	5676.33	689.3

U tabeli 8.16 predstavljena je osnovna deskriptivna statistika. Na osnovu podataka mogu se učiti opsezi kretanja vrednosti posmatranih atributa.

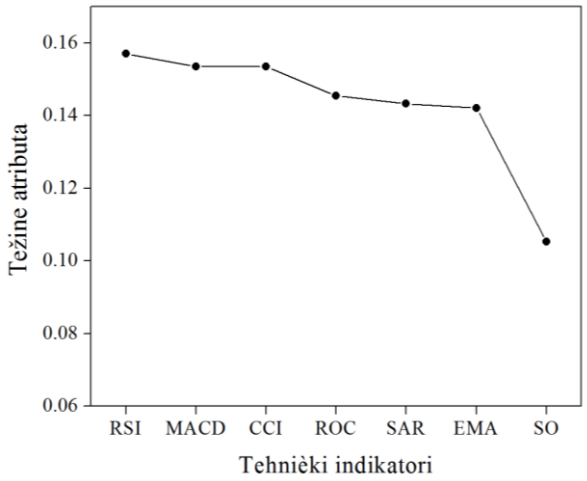
8.4.1. IZBOR ATRIBUTA

Na osnovu predložene metodologije izbor ataributa, na osnovu pojedinačnih matrica odnosa značajnosti atributa, vrši se poređenje ulaznih atributa u odnosu na definisane kriterijume evaluacije: bruto prinos, sistemski rizik i stopu predviđanja. U poslednjem koraku se kreira se matrica odlučivanja i uz pomoć nje obračunavaju težine za posmatrane ulazne atribute. Tabela 8.17 predstavlja konačno dobijene globalne vrednosti težina za opisani postupak.

TABELA 8.17 FTSE 100 matrica globalnih vrednosti težina

	PRINOS (0.082)	RIZIK (0.236)	HR (0.682)	TEŽINE ATRIBUTA
EMA	0,028	0,045	0,189	0,142
MACD	0,140	0,045	0,193	0,153
RSI	0,466	0,441	0,021	0,157
CCI	0,140	0,045	0,193	0,153
SO	0,140	0,333	0,022	0,105
ROC	0,043	0,045	0,193	0,146
SAR	0,043	0,045	0,189	0,143

Na osnovu finalnih kalkulacija, vrši se uređenje dobijenih vrednosti težina atributa, kako bi se dobila konačna procena značajnosti svakog od ulaznih atributa. Na slici 8.7 može se videti uređenje dobijenih vrednosti.



SLIKA 8.7 FTSE 100 opadajući redosled dobijenih težina atributa

Nakon izračunavanja globalnih vrednosti težina za posmatrani skup atributa, na osnovu algoritma predstavljenog u poglavlju 7.2 analiziran je dobijen grafikon predstavljen na slici 8.7 i izvršen je izbor podskupa skupa atributa za obuku predikcionog modela. Na osnovu predstavljenog grafikona može se zaključiti da težine indikatora značajno opadaju nakon trećeg atributa. Kao rezultat analize, selektovana su prva tri indikatora kao ulazni atribiti za kreiranje predikcionog modela.

8.4.2. KOMPARATIVNA ANALIZA DOBIJENIH REZULTATA

Predložena strategija za izbor atributa je kao i u prethodna dva slučaja dodatno poređena sa različitim metodama izbora atributa: primenom metoda zajedničkih inofrmacija (engl. *mutual information* - MI) sa sekvencijalnom *forward-backward* selekcijom atributa [58], *random forest* (RF) za izbor atributa [53] i linerni klasifikator sa *sequential forward selection* (engl. *linear discriminant classifier, LDC*) [66]. U tabeli 8.18 predstavljeni su podskupovi skupova atributa dobijenih primenom različitih metoda za izbor atributa.

TABELA 8.18 FTSE 100 komparativna analiza metoda izbora atributa

METOD IZBORA	IZABRANI ATRIBUTI	BROJ ATRIBUTA
AHP	RSI, MACD CCI	3
MI	CCI, % K, %D	3
RF	ROC, CCI, RSI, %K, %D, MACD	6
LDC	CCI, %K, %D	3

Na osnovu podataka u tabeli 8.18 može se zaključiti da u zavisnosti od izabranog metoda za izbor atributa, varira broj kao i sam podskup selektovanih atributa. Pretpostavka kod izbora atributa je ta da će se treningom modela sa onim podskupom atributa koji deli maksimalnu količinu informacija sa ciljnim vrednostima trening skupa postići bolja predviđanja modela. Kao što rezultati pokazuju, ovo ne mora biti uvek slučaj, čak i kada je prema MI kriterijumu izabran optimalan podskup atributa.

I u slučaju FTSE 100 berzanskog indeksa izvršene su komparatvne analize dobijenih rezultata, konstruisanjem istih predikcionih modela kao i u slučaju Belex15 i S&P 500 indeksa. Dobijeni rezultati predstavljeni su u tabelama 8.19 i 8.20.

TABELA 8.19 FTSE 100 predikcija sa različitim metodama izbora atributa

PREDIKCIONI MODELI	FTSE 100
LS-SVM	51.58
MI-LS-SVM	50.79
RF-LS-SVM	51.19
LDC-LS-SVM	51.19
AHP-WK-LS-SVM	57.54
SVM	52.78
AHP-WK-SVM	55.56

Na osnovu rezultat iz tabele 8.9 može se zaključiti da se primenom predložene metodologije dobijaju značajna poboljšanja na posmatranom FTSE 100 skupu podataka. U odnosu na LS-SVM metod bez izbora atributa razlika u dobijenim stopama pogodka iznosi 5.96%. Poboljšanja su uočljiva i kod ostalih benchmark modela i kreću se od 6.75% za MI-LS-SVM, do 6.35% za RF-LS-SVM i LDC-LS-SVM model.

Pored primene dobijenih težina i podskupa skupa atributa u predloženom AHP-WK-LS-SVM predikcionom modelu, na osnovu tabele 8.9 može se uvideti i da inkorporacija dobijenih AHP težina u SVM sa lineranim kernelom doprinosi poboljšanju kvaliteta predikcije i kod AHP-WK-SVM predikcionog modela u iznosu od 2.78%.

TABELA 8.20 FTSE 100 komparativna analiza predikcionih modela

PREDIKCIONI MODELI	FTSE 100
AHP-WK-LS-SVM	57.54
SVM	52.78
AHP-WK-SVM	55.56
ANN	54.37
RF	53.57
RW	50.00

Na osnovu podataka predstavljenih u tabeli 8.20 može se zaključiti da za FTSE 100 skup podataka postoje određena poboljšanja koja se postižu primenom predloženog AHP-WK-LS-SVM predikcionog modela. U odnosu na RF model pobljšanje iznosi 3.97%, dok se u odnosu na RW model mogu uočiti značajna poboljšanja od 7.54%. Razlike su manje izražene u odnosu na ANN model za koji je razlika u dobijenim vrednostima 3.17%.

8.5. VALIDACIJA METODOLOGIJE

U okviru ove sekcije u cilju dalje validacije predložen emetodologije izvršene su dodatne analize dobijenih rezultata.

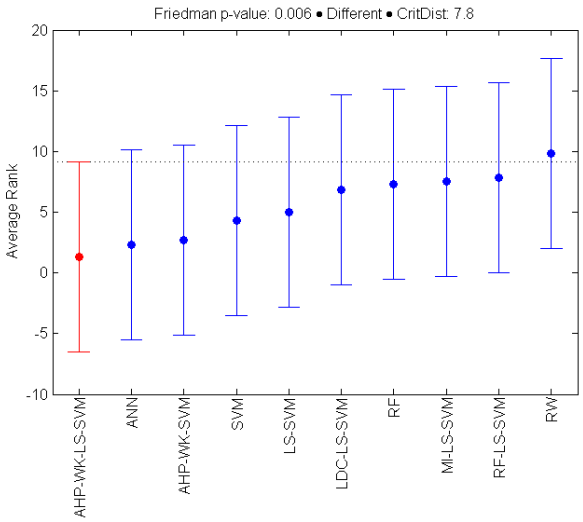
Kako bi se odredila statistička značajnost dobijenih rezultata, pri višestrukim poređenjima, odnosno kod poređenja više predikcionih modela na više skupova podataka, preporučuje se primena dvostepene procedure [39].

Najpre se primenjuje Fridmanov test (eng. *Friedman's test*) kako bi se odredilo da li modeli koji se upoređuju imaju generalno značajne razlike u performansama, ako je nulta hipoteza odbačena onda se u drugom koraku primenjuje određeni post-hoc test. Fridmanov test je neparametarski test koji je kreiran da utvrdi razlike između dve ili više grupa. U tabeli 8.21 grupisani su prethodno dobijeni rezultati nad kojima je izvršen Fridmanov test.

TABELA 8.21 Komparativna analiza predikcionih modela

PREDIKCIONI MODELI	BELEX15	S&P500	FTSE100
LS-SVM	53.57	53.97	51.58
MI-LS-SVM	52.38	53.18	50,79
RF-LS-SVM	52.38	51.19	51.19
LDC-LS-SVM	53.17	51.98	51.19
AHP-WK-LS-SVM*	61.11	57.14	57.54
SVM	53.98	53.57	52.78
AHP-WK-SVM	57.54	54.76	55.56
ANN	53.97	58.33	54.37
RF	50.00	51.19	53.57
RW*	50.00	48.41	50.00
*Nemenyi post-hoc $p \leq 0.05$			

Primenom Fridmanovog testa, dobijena je *p-valu*e od 0.0057, pri kojoj se odbacuje nulta hipoteza na 5% nivou značajnosti. Odbacivanje hipoteze upućuje na postojanje značajne statističke razlike kod poređenja upoređivanih predikcionih modela. Na slici 8.8 predstavljen je grafik dobijenih rezultata.



SLIKA 8.8 Grafički prikaz Fridman testa

Kao *post-hoc* test korišćen je *Nemenyi* test koji indikuje da ne postoje značajne statističke razlike na 0.05 nivou značajnosti između posmatranih predikcionih modela osim značajne razlike koja je utvrđena između AHP-WK-LS-SVM i RW predikcionog modela indikovano u tabeli 8.21.

Radi dalje evaluacije predložene metodologije, u okviru ove sekcije izvršeno je i poređenje preciznosti predloženog predikcionog modela sa prethodno definisanim predikcionim modelima u zavisnosti od veličine raspoloživog trening skupa, i nezavisno od interpretacije izvršenih analiza i rezultata u kontekstu procesa koji generiše podatke. Dobijeni rezultati stope predviđanja u zavisnosti od veličine trening skupa predstavljeni su u tabelama 8.22 - 8.25.

TABELA 8.22 Komparacija rezultata na osnovu 30% veličine trening skupa

PREDIKCIONI MODEL	BELEX15	S&P 500	FTSE 100
LS-SVM	57.62	50.92	48.27
MI-LS-SVM	57.20	50.63	49.21
RF-LS-SVM	56.78	50.63	49.69
LDC-LS-SVM	57.62	51.99	48.87
AHP-WK-LS-SVM	58.81	54.79	51.05
SVM	58.11	50.07	47.65
AHP-WK-SVM	58.18	52.68	50.64
ANN	52.72	55.08	50.24
RF	57.13	51.48	49.72

Na osnovu rezultata iz tabele 8.22 može se zaključiti da kod svih posmatranih serija podataka, sa smanjenjem veličine trening skupa opada i preciznost predikcionih modela. U slučaju učenja sa manjim skupom podataka može se uočiti da za FTSE 100 indeks i S&P 500

indeks ANN daju više vrednosti stope predviđanja u poređenju sa LS-SVM modelom obučavanom sa svim dostupnim atributima. Kod predstavljene podele skupa podataka, AHP-WK-LS-SVM model vodi do određenih poboljšanja najpre u odnosu na LS-SVM model a zatim i u odnosu na ostale predikcione modele. Poboljšanja su kreću prema rezultatima prikazanim u tabeli u iznosu od 1.19% do 3.87% u odnosu na LS-SVM modele koji implementiraju druge pristupe izbora atributa. U odnosu na ANN model kod BELEX15 indeksa poboljšanje iznosi 6.1%, dok u odnosu na RF model iznosi 1.68%. Određena poboljšanja mogu se uočiti i na primeru FTSE 100 indeksa. Kod S&P 500 indeksa najbolje vrednosti dobijene su primenom neuronskih mreža.

TABELA 8.23 Komparacija rezultata na osnovu 40% veličine trening skupa

PREDIKCIONI MODEL	BELEX15	S&P 500	FTSE 100
LS-SVM	56.38	52.10	48.29
MI-LS-SVM	55.66	51.03	48.92
RF-LS-SVM	55.75	50.69	48.85
LDC-LS-SVM	55.91	51.76	48.92
AHP-WK-LS-SVM	57.62	55.39	50.83
SVM	57.21	50.62	47.19
AHP-WK-SVM	57.37	50.95	47.42
ANN	50.20	56.14	51.38
RF	54.21	52.36	49.76

Na osnovu rezultata iz tabele 8.23 može se zaključiti da za podelu trening i test skupa u odnosu 40:60 kod posmatranih skupova podataka metod za rangiranje i selekciju atributa zasnovan na metodologiji predloženoj u ovoj doktorskoj disertaciji daje poboljšanja u primeni sa LS-SVM i SVM predikcionim modelima.

Kod BELEX15 berzanskog indeksa, predloženi predikcioni model u predstavljenoj podeli trening i test skupa podataka ima najveću stopu predviđanju u odnosnu na ostale predikcione modele obuhvaćene komparativnom analizom. Kod S&P 500 indeksa, najveću stopu preciznosti ima model veštačkih neuronskih mreža dok je predloženi predikcioni model na drugoj poziciji po uspešnosti. Takođe, kod predikcija FTSE 100 berzanskog indeksa na osnovu obučavanja predikcionog modela sa 40% veličine trening skupa bolje rezultate u odnosu na predloženi predikcioni model postigao je ANN model.

U tabeli 8.24 predstavljena je komparacija dobijenih stopa predviđanja na osnovu obučavanja predikcionih modela sa 50% veličine trening skupa.

TABELA 8.24 Komparacija rezultata na osnovu 50% veličine trening skupa

PREDIKCIONI MODEL	BELEX15	S&P 500	FTSE 100
LS-SVM	55.47	52.31	49.28
MI-LS-SVM	53.72	51.92	49.47
RF-LS-SVM	54.79	51.82	49.01
LDC-LS-SVM	54.59	51.92	49.47
AHP-WK-LS-SVM	57.63	55.07	51.67
SVM	57.24	51.92	48.05
AHP-WK-SVM	57.24	53.89	49.38
ANN	50.69	55.86	50.62
RF	54.31	53.84	49.71

Na osnovu rezultata iz tabele 8.24 može se zaključiti da kod svih posmatranih serija podataka, sa povećanjem veličine trening skupa raste i preciznost predikcionih modela. U slučaju učenja sa predstavljenom veličinom skupa podataka može se uočiti da za FTSE 100 indeks i S&P 500 indeks predikcioni ANN model daje više vrednosti stope predviđanja u poređenju sa LS-SVM modelom obučavanom sa svim dostupnim atributima. Na osnovu prikazanih rezultata može se zaključiti da i kod ovakve podele skupa podataka, AHP-WK-LS-SVM model vodi do određenih poboljšanja najpre u odnosu na LS-SVM model a zatim i u odnosu na ostale predikcione modele u slučaju BELEX15 i FTSE 100 indeksa. Kod S&P 500 indeksa najbolje vrednosti dobijene su primenom neuronskih mreža.

TABELA 8.25 Komparacija rezultata na osnovu 75% veličine trening skupa

PREDIKCIONI MODEL	BELEX15	S&P 500	FTSE 100
LS-SVM	54.71	52.77	49.33
MI-LS-SVM	55.49	52.96	49.33
RF-LS-SVM	55.49	53.16	49.14
LDC-LS-SVM	55.88	51.78	49.53
AHP-WK-LS-SVM	59.61	55.34	53.33
SVM	56.08	51.38	48.38
AHP-WK-SVM	58.04	52.77	53.33
ANN	50.39	55.54	52.00
RF	50.00	51.19	53.57

Na osnovu rezultata iz tabele 8.25 može se zaključiti da za predstavljenu podelu trening i test skupa kod posmatranih skupova podataka metod za rangiranje i selekciju atributa zasnovan na metodologiji predloženoj u ovoj doktorskoj disertaciji daje poboljšanja u primeni sa LS-SVM i SVM predikcionim modelima. Kod FTSE 100 berzanskog indeksa, jedino kod 75% podele trening skupa RF algoritam daje bolje predikcione rezultate u odnosu predloženi predikcioni model. Kod S&P 500 indeksa, najveću stopu preciznosti ima i kod

predstavljene podele model veštačkih neuronskih mreža dok je predloženi predikcioni model na drugoj poziciji po uspešnosti.

8.6. USLOVI PRIMENE METODOLOGIJE ZA IZBOR ATRIBUTA

U prethodno opisanim simulacijama, najbolje rezultate pokazali su predikcioni modeli formirani sa LS-SVM algoritmom za binarnu klasifikaciju kroz integraciju težina u kernel. U odnosu na klasifikaciju metoda izbora atributa predstavljenoj u okviru četvrtog poglavlja ove doktorske disertacije, predstavljena metodologija izbora atributa pripada grupi filter metoda što znači da način primene predložene metodologije izbora atributa ne zavisi od tipa algoritma nadgledanog mašinskog učenja sa kojim se formira predikcioni model. Ipak, potrebno je posebno razmotriti *k-NN* algoritam s obzirom na činjenicu da već postoje radovi koji upućuju na značaj davanja težina atributima kod *k-NN* algoritma. Sa tim u vezi realno je očekivati da bi predložena metodologija dala poboljšanje performansi i sa *k-NN* algoritmom.

Pretpostavka od koje se polazi je da su u dostupnoj vremenskoj seriji vektorima pristupa po hronološkom redosledu bez primena metoda semplovanja ili *bootstrapping* metoda.

Iako se na osnovu prikazanih rezultata može zaključiti da se primenom predložene metodologije izbora atributa kod prediktivnog modelovanje vremenskih serija nadgledanim mašinskim učenjem postižu zapaženi rezultati, treba istaći mogućnosti za dalji razvoj. Poboljšanja u kvalitetu predikcije korišćenjem predloženog predikcionog modela zavise od selekcije i inicijalnog opsega podataka nad kojima se izvode AHP proračuni, ali i od procene inicijalne skale težina predstavljene u [131] za izabrane kriterijume. Iako je na osnovu prikazanih rezultata evidentno da određena unapređenja u kvalitetu predickije postoje nezavisno od inicijalnih odabira, u planu je da se u istraživanjima razmotre algoritmi za automatski izbor parametara i njihovu optimizaciju.

Što se tiče uopštenja, predložena metodologija izbora atributa može se koristiti i kada se u ulaznom skupu vektora nalaze i kvalitativni atributi.

Takođe, skup težina atributa se nezavisno može primeniti i kod drugih kernel zasnovanih metoda mašinskog učenja.

S obzirom na karakteristike AHP metoda za očekivati je i da se ova metodologija može primenjivati na sve skupove podataka kod kojih je moguće izvršiti hijerarhijsku dekompoziciju cilja kroz definisanje kriterijuma koji adekvatno konceptualizuju znanje o domenu.

POGLAVLJE 9

ZAKLJUČAK

Istraživanja prikazana u ovoj doktorskoj disertaciji primarno su fokusirana na izbor skupa atributa za obuku predikcionog modela i zatim na integraciju metode izbora atributa sa metodama nadgledanog mašinskog učenja. Cilj istraživanja bio je i da se poveća preciznost predikcije vremenskih serija.

U procesu izbora atributa predložena je metodologija zasnovana na Analitičkom hijerarhijskom procesu, koji ima mogućnost da vrši evaluaciju atributa i kod međusobno suprostavljenih kriterijuma.

Za kreiranje predikcionih modela korišćeni su SVM i LS-SVM metodi za klasifikaciju, sa linearnim i RBF kernelom. Unapređenje korišćenih metoda postignuto je kroz inkorporaciju težina u kernel.

U nastavku poglavlja najpre je dat rezime istraživanja prezentovanih u okviru ove doktorske disertacije, zatim su istaknuti doprinosi rezultata istraživanja i na kraju su izneti pravci budućih istraživanja.

U prvom i drugom poglavlju analizirane se oblasti primene vremenskih serija kao i prediktivnog modelovanja, dat je prikaz predikcionih modela i izloženi su osnovni pojmovi potrebni za razumevanje svojstva algoritama mašinskog učenja. Prikazana su i ograničenja i problemi u dosadašnjim pristupima kreiranja predikcionih modela.

U trećem poglavlju izložen je princip učenja kernelom i teorijske osnove SVM i LS-SVM metoda koji su korišćeni u simulacijama za formiranje predikcionih modela. U ovom poglavlju su posebno analizirane kernel funkcije i prikazane su terijske osnove potrebne za razumevanje i definisanje težinske kernel funkcije.

U četvrtom poglavlju detaljno su analizirani problem i oblast izbora atributa u nadgledanom mašinskom učenju, a zatim je prema najšire prihvaćenoj klasifikaciji izvršena

komparativna analiza prikazanih strategija.

U petom poglavlju razmatrani su problemi reprezentacije znanja kod metoda mašinskog učenja, pojmovi inženjering atributa i mogućnosti označavanja vektora i dodeljivanje težina atributima.

U šestom poglavlju dat je prikaz sinergije metoda mašinskog učenja i optimizacionih metoda, zajedno sa pregledom literature koja upućuje na integraciju metoda odlučivanja i algoritama mašinskog učenja. U poslednjem delu istog poglavlja prikazane su teorijske osnove potrebne za razumevanje AHP metoda korišćenog u eksperimentalnom delu doktorske disertacije.

U sedmom poglavlju, kao doprinos, predložena je metodologija za izbor podskupa atributa zasnovana na inkorporaciji znanja o domenu primenom metoda odlučivanja odnosno primenom Analitičkog hijerarhijskog procesa.

U osmom poglavlju prikazana je i analizirana je primena predložene metodologije na različitim skupovima podataka u kombinaciji sa integracijom težina u kernel kod SVM i LS-SVM metoda. U okviru osmog poglavlja prikazani su i rezultati komparacije predložene metodologije sa drugim algoritmima izbora atributa kao i rezultati komparacije predloženog predikcionog modela sa drugim modelima nadgledanog mašinskog učenja. Na kraju osmog poglavlja, razmatrani su uslovi primene predložene metodologije za izbor atributa i diskutovana su uopštenja.

Najznačajniji doprinos ove doktorske disertacije je predložena metodologija za izbor atributa zasnovana na određivanju težina atributima korišćenjem Analitičkog hijerarhijskog procesa. Na osnovu određenih težina, bira se podskup atributa inicijalnog skupa atributa za obuku predikcionog modela sa ciljem poboljšanja kvaliteta predviđanja. Metodologija predložena u ovoj doktorskoj disertaciji zasnovana je na konceptu primene AHP metoda za rangiranje i selekciju atributa. Pored toga, kako bi se poboljšale sposobnosti generalizacije LS-SVM modela u disertaciji je korišćena težinska kernel fukcija, pri čemu su težine određene na osnovu relevantnosti atributa određene primenom AHP postupka. Uticaj težinskog kernela i izbora atributa vodi do određenih poboljšanja u preciznosti predloženog modela.

Metodologija je testirana na finansijskim vremenskim serijama sa tržišta kapitala razvijenih i zemalja u razvoju. Primenom predloženog modela dobijaju se konkurentni rezultati predviđanja pravca kretanja vrednosti berzanskih indeksa. Naročito je značajno uzeti u obzir da se u okviru ove doktorske disertacije posebno ispituje stepen predvidljivosti kretanja indeksa cena akcija na tržištu u razvoju, kao što je tržište kapitala Republike Srbije,

nasuprot većini radova iz ove oblasti, koji se bave predviđanjem indeksa cena akcija na razvijenim tržištima.

U disertaciji je takođe izvršena i komparativna analiza predložene metodologije sa drugim metodama izbora atributa kao i predloženog predikcionog modela sa drugim metodama mašinskog učenja. Analizirano je i kako promena veličine trening skupa utiče na performanse predikciong modela i sam izbor atributa.

Za razliku od većine prethodnih istraživanja iz ove oblasti, kod kojih se izbor atributa oslanja na evaluaciju kvantitativnih pokazatelja, predložen pristup vrši procenu atributa uvodeći u razmatranje *a priori* znanje i koristeći kvalitativna svojstva atributa. Kao što rezultati pokazuju, primenom predložene metodologije je pored poboljšanja kvaliteta predviđanja moguće izvršiti i procenu kvaliteta samog skupa atributa. Odnosno, analiza atributa i procena njihove relevatnosti stoji u korespondenciji sa interpretacijom procesa odlučivanja na tržištima kapitala.

Dobijeni rezultati predstavljaju predviđanja za jedan korak unapred na produženom vremenskom horizontu, jedna godina trgovanja, i tako nadmašuju mnoge vremenske horizonte predstavljene ranije u literaturu [68], [176] i [111], ali i dalje daju komparativne rezultate u srednjem opsegu predstavljenih vrednosti. Dobijeni rezultati su pouzdani uzevši u obzir da su testitirani na celokupnom skupu dostupnih podataka koji predstavljaju sve forme ponašanja modela.

Rezultati simulacija ukazuju da se prezentovana metodologija zasnovana na AHP analizi može koristiti kao metod preprocesiranja kako bi se procenila relevatnost atributa kod vremenskih serija podataka, ali i kod svih skupova podataka kod kojih je moguće izvršiti hijerarhijsku dekompoziciju problema.

Iz prezentovanih rezultata može se zaključiti da SVM i LS-SVM modeli koji implementiraju predloženu metodologiju izbora atributa, bez obzira na korišćenu kernel funkciju, postižu bolje rezultate predviđanja u odnosu na SVM i LS-SVM modele trenirane sa svim dostupnim atributima. Odnosno, može se zaključiti da je kvalitet skupa atributa značajniji od njegove veličine, pošto predikcioni modeli trenirani sa izabranim podskupovima atributa postižu bolje rezultate predikcije.

Najveći doprinos u poboljšanju kvaliteta predviđanja postignut je u kombinaciji predložene metodologije izbora atributa i LS-SVM metoda.

Iako su prethodno u sekciji 8.6 detaljno diskutovani načini primene predložene metodologije i mogućnosti uopštenja, potrebno je rezimirati osnovno. Pretpostavka od koje se polazi je da se u dostupnoj vremenskoj seriji vektorima pristupa po hronološkom redosledu

bez primena metoda semplovanja ili *bootstrapping* metoda. S obzirom na to da predstavljena metodologija po klasifikaciji pripada grupi filter metoda, način primene metodologije izbora algoritma ne zavisi od tipa algoritma nadgledanog mašinskog učenja sa kojim se formira predickioni model. Predložena metodologija izbora atributa može se koristiti i kada se u ulaznom skupu atributa nalaze i kvalitativni atributi. Takođe, skup težina atributa se nezavisno može primeniti i kod drugih kernel zasnovanih metoda mašinskog učenja.

Dobijena poboljšanja u stopi predviđanja mogu se smatrati značajnim, s obzirom na činjenicu da se predviđanje pravca promene kretanja tržišnih indeksa vrši sa ciljem optimizacije strategije trgovanja na finansijskim tržištima. Pri čemu, povećanje stope preciznosti dobijeno primenom predložene metode može voditi do povećanja profita, imajući u vidu da vodi ka povećanju prinosa i smanjenju rizika trgovanja.

Dalja istraživanja biće usmerena pre svega na proveru performansi predloženih predikcionih modela u realnom okruženju i to na osnovu obračuna upravo prinosa koji se može ostvariti trgovanjem pomoću strategija zasnovanih na prediktivnom modelovanju, na osnovu [147] i [6] i postavke značaja takve vrste evaluacije. U tom smislu u planu je i razmatranje kreiranja kompozitne strategije trgovanja koja bi se oslanjala na AHP proračune.

U skladu sa prethodnim, dalja pobošanja će se kretati u dva smera. Jedan deo biće okrenut ka ispitivanju drugih kriterijuma relevantnih za investitore sa različitim preferencijama prema riziku i ka načinima da se dodatno dobiju informacije iz atributa i inkorporira znanje o domenu u predikcioni model.

Drugi deo budućih istraživanja odnosi će se na unapređenje predikcionih modela i same prikazane metodologije.

Kako je navedeno u okviru poglavlja Reprezentacija znanja i inženjering atributa, označavanje vektora predstavlja bitan način za unapređenje kvaliteta predikcije. U vezi sa time ispitivanje metoda izbora vektora predstavlja jedan od značajnih daljih pravaca istraživanja. S obzirom na mogućnost da sam izbor vektora može dodatno pozitivno uticati na performanse prediktora, odnosno da izbor vektora može doprineti kvalitetu predikcije.

Posebnu pažnju potrebno je posvetiti grupnom AHP metodu i *fuzzy* AHP metodi, kako bi se mogli obraditi veći skupovi atributa ili radi poboljšanja inicijalne evaluacije atributa na način da se prevaziđu neodređenosti.

Zatim istraživanjima treba obuhvatiti kreiranje i ispitivanje uticaja kreiranja ansambla (engl. *ensamble*) modela na stopu predviđanja, pri čemu bi se predikcije više različitih modela kombinovale prema nekoj od šema agregacije zasnovanoj na modelima višekriterijumskog odlučivanja u rezultujući model.

Cilj budućeg rada je i da se potvrdi da li se predložena metodologija može primenjivati na druge veće skupove podataka, kao i da se detaljno testira predložena metodologije izbora atributa sa ostalim metodama nadgledanog mašinskog učenja. U planu je da se usredsredi pažnja i da se istraživanja prošire i sa drugim merama procena karakteristika, kao i da se uključe određene optimizacione mere u razmatranje.

LITERATURA

[1] I. M. Abd-el Fattah, W. I. Khedr, and K. M. Sallam, "A TOPSIS based method for gene selection for cancer classification," *International Journal of Computer Applications*, vol. 67, no. 17, 2013.

[2] D. W. Aha, D. Kibler, and C. M. Albert, "Instance-Based Learning Algorithms," *Machine Learning*, vol. 6, pp. 37-66, 1991.

[3] E. Amaldi and V. Kann, "On the approximation of minimizing non zero variables or unsatisfied relations in linear systems," *Theoretical Computer Science*, vol. 209, pp. 237–260, 1998.

[4] G. Appel, *Technical analysis: power tools for active investors*.: FT Press, 2005.

[5] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Statistic Surveys*, vol. 4, pp. 40-79, 2010.

[6] S. G. Atsalakis and P. K. Valavanis, "Forecasting stock market short-term trends using a neuro-fuzzy based methodology," *Expert Systems with Applications*, vol. 36, no. 7, pp. 10696-1070, 2009.

[7] S. G. Atsalakis and P. K. Valavanis, "Surveying stock market forecasting techniques – Part II: Soft computing methods," *Expert Systems with Applications*, vol. 36, no. 3, pp. 5932–5941, 2009b.

[8] S. Barak and M. Modarres, "Developing an approach to evaluate stocks by forecasting effective features with data mining methods," *Expert Systems with Applications*, vol. 42, no. 3, pp. 1325–1339, 2015.

[9] A. Ben-Hur, CS Ong, S. Sonnenburg, B. Schölkopf, and G. Rätsch, "Support Vector Machines and Kernels for Computational Biology," *PLoS Comput Biol*, vol. 4, no. 10, p. e1000173, 2008.

[10] K. K. Bharti and P. K. Sing, "Hybrid dimension reduction by integrating feature selection with feature extraction method for text clustering," *Expert Systems with Applications*, vol. 42, no. 6, pp. 3105-3114, 2015.

[11] G. Bhattacharya, K. Ghosh, and S. A. Chowdhury, "Granger Causality Driven AHP for Feature Weighted kNN," *Pattern Recognition*, vol. 66, pp. 425–436, 2017.

[12] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine

learning," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 245-271, 1997.

[13] T. Bollerslev, "Generalized autoregressive conditional heteroscedasticity," *Journal of Econometrics*, vol. 31, pp. 307-327, 1986.

[14] G. E. P. Box and G. Jenkins, *Time Series Analysis, Forecasting and Control*. San Francisco: Holden- Day, 1976.

[15] K. De Brabanter et al., "LS-SVMlab Toolbox User's Guide version 1.8.," 2011. [Online]. http://www.esat.kuleuven.be/sista/lssvmlab/

[16] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[17] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*. Monterey, Calif., U.S.A: Wadsworth, Inc., 1984.

[18] J. Brownlee. (2014) Discover Feature Engineering, How to Engineer Features and How to Get Good at It. [Online]. http://machinelearningmastery.com/discover-feature-engineering-how-to-engineer-features-and-how-to-get-good-at-it/

[19] J. M. Cadenas, M. C. Garrido, and R. MartíNez, "Feature subset selection Filter–Wrapper based on low quality data," *Expert Systems with Applications*, vol. 40, no. 16, pp. 6241-6252, 2013.

[20] L. Cao, "Introduction to Domain Driven Data Mining," in *Data Mining for Business Applications*. Boston, MA: Springer, 2009.

[21] L. J. Cao and F. E. H. Tay, "Support Vector Machine With Adaptive Parameters in Financial Time Series Forecasting," *IEEE TRANSACTIONS ON NEURAL NETWORKS*, vol. 14, no. 6, pp. 1506-1518, 2003.

[22] L. Cao and C. Zhang, "The Evolution of KDD: Towards Domain-Driven Data Mining," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 21, no. 04, pp. 677-692, 2007.

[23] J. Chai, J. Du, K. K. Lai, and Lee Y. P., "A Hybrid Least Square Support Vector Machine Model with Parameters Optimization for Stock Forecasting," *Mathematical Problems in Engineering*, 2015.

[24] C.-C Chang and C.-J. Lin, "LIBSVM: A Library for Support Vector Machines," vol. 2, no. 3, pp. 27:1–27:27, 2011. [Online]. http://www.csie.ntu.edu.tw/~cjlin/libsvmtools

[25] T.T. Chen and S.J. Lee, "A weighted LS-SVM based learning system for time series forecasting," *Information Sciences*, vol. 299, pp. 99-116, 2015.

[26] W. Cheung, K. Lam, and H. Yeung, "Intertemporal profitability and the stability of technical analysis: evidences from the Hong Kong stock exchange," *Applied economics*, vol. 43, no. 15, pp. 1945-1963, 2009.

[27] E. Çomak, K. Polat, S. Güneş, and A. Arslan, "A new medical decision making system: Least square support vector machine (LSSVM) with Fuzzy Weighting Pre-processing," *Expert Systems with Applications*, vol. 32, pp. 409-414, 2007.

[28] D. Corne, C. Dhaenens, and L. Jourdan, "Synergies between operations research and data mining: The emerging use of multi-objective approaches," *European Journal of Operational Research*, vol. 221, no. 3, pp. 469-479, 2012.

[29] P. Coulibaly and C. K. Baldwin, "Nonstationary hydrological time series forecasting using nonlinear dynamic methods," *Journal of Hydrology*, vol. 307, no. 1–4, pp. 164-174, 2005.

[30] T. Cover and J. Thomas, *Elements of information theory*.: John Wiley & Sons, 1991.

[31] G. Coyle, *Practical Strategy: Structured tools and techniques*.: Pearson Education Limited, 2004.

[32] D. Craft, D. Ferranti, and D. Krane, "The value of prior knowledge in machine learning of complex network systems," *Bioinformatics*, 2017.

[33] N. Cristianini and B. Schölkopf, "Support Vector Machines and Kernel Methods The New Generation of Learning Machines," *AI Magazine*, vol. 23, no. 3, 2002.

[34] S. F. Crone and N. Kourentzes, "Feature selection for time series prediction – A combined filter and wrapper approach for neural networks," *Neurocomputing*, vol. 73, no. 10-12, pp. 1923–1936, 2010.

[35] M. Dash and H. Liu, "Feature Selection for Classification," *Intelligent Data Analysis*, vol. 1, no. 1-4, pp. 131-156, 1997.

[36] H. Daume III and D. Marcu, "Domain Adaptation for Statistical Classifiers," *Journal of Artificial Intelligence Research*, vol. 26, pp. 101-126, 2006.

[37] J. Derrac, I. Triguero, S. Garcia, and F. Herrera, "Integrating Instance Selection, Instance Weighting, and Feature Weighting for Nearest Neighbor Classifiers by Coevolutionary Algorithms," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 5, pp. 1383 - 1397, 2012.

[38] P. Desikan and J. Srivastava, "Time series analysis and forecasting methods for

temporal mining of interlinked documents. Department of Computer Science, University of Minnesota,Time series analysis and forecasting methods for temporal mining of interlinked documents," *Department of Computer Science, University of Minnesota, www-users.cs.umn.edu/~desikan/publications/TimeSeries.doc*, 2014.

[39] J. Dešmar, "Statistical Comparisons of Classifiers over Multiple Data Sets," *Journal of Machine Learning Research*, vol. 7, pp. 1-30, 2006.

[40] N. Dessì and B. Pes, "Similarity of feature selection methods: An empirical study across data intensive classification tasks," *Expert Systems with Applications*, vol. 42, no. 10, pp. 4632–4642, 2015.

[41] M. Dialameh and M. Z. Jahromi, "A general feature-weighting function for classification problems," *Expert Systems With Applications*, vol. 72, pp. 177-188, 2017.

[42] P. Domingos, "A Few Useful Things to Know about Machine Learning," *Communications of the ACM*, vol. 55, no. 10, pp. 78-87, 2012.

[43] R. O. Duda, P. E. Hart, and D. G. S Stork, *Pattern Recognition*. New York: Wiley, 2000.

[44] M. Ebrahimi, M. Abdolshah, and S. Abdolshah, "Developing a computer vision method based on AHP and feature ranking for ores type detection," *Applied Soft Computing*, vol. 49, pp. 179-188, 2016.

[45] R. F. Engle, "Autoregressive conditional heteroscedasticity with estimator of the variance of United Kingdom inflation," *Econometrica*, vol. 50, no. 4, pp. 987-1008, 1982.

[46] D. Erić, G. Andjelić, and S. Redzepagić, "Application of MACD and RVI Indicators as Functions of Investment Strategy Optimization on the Financial Market," *In Working Papers of the Faculty of Economics in Rijeka*, vol. 27, no. 1, pp. 171-196, 2009.

[47] E. Fama, "Efficient Capital Markets: A Review of Theory and Empirical Work," *Journal of Finance*, vol. 25, pp. 383-417, 1970.

[48] D. M. Farid and C. M. Rahman, "Assigning Weights to Training Instances Increases Classification Accuracy," *International Journal of Data Mining & Knowledge Management Process (IJDKP)* , vol. 3, no. 1, 2013.

[49] A. Feelders, D. Hennie, and M. Holsheimer, "Methodological and practical aspects of data mining," *Information and Management*, vol. 37, no. 5, pp. 271-281, 2000.

[50] G. M. Fung, O. L. Mangasarian, and J. W. Shavlik, "Knowledge-Based Support Vector Machine Classifiers," in *Proceeding NIPS'02 Proceedings of the 15th International Conference on Neural Information Processing Systems*, 2002, pp. 537-544.

[51] G. P. C. Fung, J. X. Yu, and W. Lam, "News sensitive stock trend prediction," *Advances in knowledge discovery and data mining*, pp. 481-493, 2002.

[52] T. Gärtner, "A survey of kernels for structured data," *ACM SIGKDD Explorations Newsletter*, vol. 5, no. 1, pp. 49-58 , 2003.

[53] R. Genuer, J-M. Poggi, and C. Tuleau-Malot, "Variable selection using Random Forests," *Pattern Recognition Letters*, vol. 31, no. 14, pp. 2225-2236, 2010.

[54] A. C. Georgiou and G. Bonias, "An AHP and neural network approach in predicting short-term returns: a case of Greek stock market," *Journal of Statistics and Management Systems*, vol. 10, no. 6, pp. 905-928, 2007.

[55] T.V. Gestel et al., "Benchmarking Least Squares Support Vector Machine Classifiers," *Machine Learning*, vol. 54, no. 1, pp. 5-32, 2004.

[56] S. Girish and F. Chandrashekar, "A survey on feature selection methods," *Computers and Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014.

[57] D. Giveki, H. Salimi, G. Bahmanyar, and Y. Khademian. (2012) Automatic Detection of Diabetes Diagnosis using Feature Weighted Support Vector Machines based on Mutual Information and Modified Cuckoo Search. [Online]. http://arxiv.org/ftp/arxiv/papers/

[58] V. Gómez-Verdejo, M. Verleysen, and J. Fleury, "Information-theoretic feature selection for functional data classification," *Neurocomputing*, vol. 72, no. 16–18, pp. 3580–3589, 2009.

[59] Ç. Gülçehre and Y. Bengio, "Knowledge matters: Importance of prior information for optimization," *Journal of Machine Learning Research*, vol. 17, no. 8, pp. 1-32, 2016.

[60] B. Guo, S. R. Gunn, and Damper R. I., "Customizing Kernel Functions for SVM-Based Hyperspectral Image Classification.," *IEEE TRANSACTIONS ON IMAGE PROCESSING*, vol. 17, no. 4, pp. 622-629, 2008.

[61] M. Gutlein, E. Frank, and M., Karwath, A. Hall, "Large-scale attribute selection using wrappers," in *In Computational Intelligence and Data Mining, 2009. CIDM'09. IEEE Symposium on (pp. 332-339). IEEE.*, 2009.

[62] I. Guyon and A. Elisseeff, "An introduction to feature extraction," in *Feature Extraction: Foundations and Applications*., 2006, pp. 1-26.

[63] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157-1182, 2003.

[64] M. A. Hall, "Correlation-based feature selection for discrete and numeric class machine learning," in *in Proceedings of the 17th International Conference on Machine Learning (ICML '00)*, San Francisco, Calif, USA, 2000, pp. 359–366.

[65] Z. Harchaoui and F. Bach, "Image Classification with Segmentation Graph Kernels," in *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, Minneapolis, MN, USA , 2007.

[66] Y. He, K. Fataliyev, and L. Wang, "Feature Selection for Stock Market Analysis," in *ICONIP 2013, Part II, LNCS 8227*.: Berlin Heidelberg Springer-Verlag., 2013, pp. 737–744.

[67] T. Hofmann, B. Scholkopf, and A. J. Smola, "Kernel methods in machine learning," *The Annals of Statistics*, vol. 36, no. 3, pp. 1171-1220, 2008.

[68] W. Huang, Y. Nakamori, and S-Y. Wang, "Forecasting stock market movement direction with support vector machine," *Computers & Operations Research*, vol. 32, no. 10, pp. 2513–2522, 2005.

[69] C.J. Huang, D.X. Yang, and Y.T. Chuang, "Application of wrapper approach and composite classifier to the stock trend prediction," *Expert Systems with Applications*, vol. 34, no. 4, pp. 2870-2878, 2008.

[70] G. D. Jennifer and E. B. Carla, "Feature Selection for Unsupervised Learning," *Journal of Machine Learning Research*, vol. 5, pp. 845-889, 2004.

[71] J. Jiang. A Literature Survey on Domain Adaptation of Statistical Classifiers. [Online]. http://sifaka.cs.uiuc.edu/jiang4/domain_adaptation/survey/

[72] J. M. Kanter and K. Veeramachaneni, "Deep feature synthesis: Towards automating data science endeavors," in *Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on*, Paris, France, 2015.

[73] Y. Kara, M.A. Boyacioglu, and Ö. K. Baykan, "Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange," *Expert Systems with Applications*, vol. 38, no. 5, pp. 5311–5319, 2011.

[74] P. J. Kaufman, *A Short Course in Technical Trading*.: John Wiley & Sons., 2003.

[75] K. Kira and L. A. Rendell, "The feature selection problem: traditional methods and a new algorithm," in *In: Proceedings of tenth national conference on artificial*, 1992.

[76] R. Kohavi, P. Langley, and Y. Yun, "The Utility of Feature Weighting in Nearest-Neighbor Algorithms," in *Proceedings of the Ninth European Conference on Machine Learning*, 1997.

[77] R. I. Kondor and J. D. Lafferty, "Diffusion Kernels on Graphs and Other Discrete Input Spaces," in *ICML '02 Proceedings of the Nineteenth International Conference on Machine Learning* , 2002.

[78] I. Kose, M. Gokturka, and K. Kilic, "An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance," *Applied Soft Computing*, vol. 36, pp. 283-299, 2015.

[79] A. Kraskov, Stogbauer H., and P. Grassberger, "Estimating mutual information," vol. 69, no. 6, p. 066138, 2004.

[80] E. Krupka and N. Tishby, "Incorporating prior knowledge on features into learning," *Artificial Intelligence and Statistics*, pp. 227-234, 2007.

[81] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. New York: Springer Science+Business Media, 2013.

[82] J. Kumar and N. Roy, "A Hybrid Method for Vendor Selection using Neural Network," *International Journal of Computer Applications*, vol. 11, no. 12, pp. 35-40, 2010.

[83] S. Lahmiri, "A Comparison of PNN and SVM for Stock Market Trend Prediction using Economic and Technical Information," *International Journal of Computer Applications*, vol. 29, pp. 24-30, 2011.

[84] F. Lauer and G. Bloch, "Incorporating Prior Knowledge in Support Vector Machines for Classification: a Review," *Neurocomputing*, vol. 71, no. 7-9, pp. 1578 - 1594, 2008.

[85] M.-C. Lee, "Using support vector machine with a hybrid feature selection method to the stock trend prediction," *Expert Systems with Applications*, vol. 36, pp. 10896–10904, 2009.

[86] Q.V. Le, A.J. Smola, and T. Gärtner, "Simpler knowledge-based support vector machines," in *Proceeding ICML '06 Proceedings of the 23rd international conference on Machine learning*, Pittsburgh, Pennsylvania, 2006, pp. 521-528.

[87] H. Levy, *Stochastic dominance Investment Decision Making under Uncertainty Second Edition.*.: Springer Science+Business Media, Inc., 2006.

[88] J. Li et al. (2016) Feature Selection: A Data Perspective. [Online]. arXiv:1601.07996, 2016

[89] J. Li and A. W. Moore, "Forecasting Web Page Views: Methods and Observations," *Journal of Machine Learning Research*, vol. 9, pp. 2217-2250, 2008.

[90] G. Liu, J. Chen, and J. Zhong, "An Integrated SVM and Fuzzy AHP Approach for Selecting Third Party Logistics Providers," *PRZEGLĄD ELEKTROTECHNICZNY (Electrical Review), ISSN 0033-2097, R. 88 NR 9b/2012*, vol. 88, pp. 5-8, 2012.

[91] H. Liu, H. Motoda, R. Setiono, and Z. Zhao, "Feature Selection: An Ever Evolving Frontier in Data Mining," in *In Proceedings of the fourth international workshop on feature selection in data mining*, 2010, pp. 4-13.

[92] D-R. Liu and Y-Y. Shih, "Integrating AHP and data mining for product recommendation based on customer lifetime value," *Information & Management*, vol. 42, no. 3, pp. 387–400, 2005.

[93] D. Liu, Z. Tian, B. Luo, and J. Xia, "Feature Ranking in Intrusion Detection by Hybrid Algorithm with Support Vector Machine and Analytic Hierarchy Process," *International Journal of Digital Content Technology and its Applications*, vol. 7, no. 7, 2013.

[94] H. Liu and J. Wang, "Integrating Independent Component Analysis and Principal Component Analysis with Neural Network to Predict Chinese Stock Market," *Mathematical Problems in Engineering*, 2011.

[95] C. Liu, W. Wang, Q. Zhao, X. Shen, and M. Konan, "A new feature selection method based on a validity index of feature subset," *Pattern Recognition Letters*, vol. 92, no. 1, pp. 1-8, 2017.

[96] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 4, pp. 491-502, 2005.

[97] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins, "Text Classification using String Kernels," *Journal of Machine Learning Research*, vol. 2, pp. 419-444, 2002.

[98] A. Lo, H. Mamaysky, and J. Wang, "Foundations of Technical Analysis:

Computational Algorithms, Statistical Inference, and Empirical Implementation," *The Journal of Finance*, vol. 55, no. 4, pp. 1705-1765, 2000.

[99] A. Mardani et al., "Multiple criteria decision-making techniques and their applications–a review of the literature from 2000 to 2014," *Economic Research-Ekonomska Istraživanja*, vol. 28, no. 1, pp. 516-571, 2015.

[100] I. Marković, M. Stojanović, M. Božić, and J. Stanković, "Stock Market Trend Prediction Based on the LS-SVM Model Update Algoritham," in *ICT Innovations 2014, Advances in Intelligent Systems and Computing*., 2014, p. 105—114.

[101] I. P. Marković, M. B Stojanović, J. Z. Stanković, and M. M. Božić, "Stock market trend prediction using support vector machines," *Facta Universitatis, Series: Automatic Control and Robotics*, vol. 13, no. 3, pp. 147-158, 2014.

[102] I. Marković, M. Stojanović, J. Stanković, and M. Stanković, "Stock market trend prediction using AHP and weighted kernel LS-SVM," *Soft Computing*, vol. 21, no. 18, pp. 5387–5398, 2017.

[103] S. Matsuda, "A neural network model for the decision-making process based on AHP," in *Neural Networks, 2005. IJCNN '05. Proceedings. 2005 IEEE International Joint Conference on*, Montreal, Que., Canada, 2005.

[104] P. D. McNelis, *Neural Networks in Finance: Gaining Predictive Edge in the Market*.: Elsevier Inc, 2005.

[105] S. Meisel and D. Mattfeld, "Synergies of Operations Research and Data Mining," *European Journal of Operational Research*, vol. 206, no. 1, pp. 1-0, 2010.

[106] T. M. Mitchell, *Machine Learning*.: McGraw Hill, 1997.

[107] M. A. Mittermayer, "Forecasting intraday stock price trends with text mining techniques," in *In the Proceedings of the Hawaii International Conference on System Sciences*, 2004.

[108] L. C. Molina, L. Belanche, and A. Nebot, "Feature Selection Algorithms: A Survey and Experimental Evaluation," in *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*, Maebashi City, Japan, 2002, pp. 306-313.

[109] S. Moro, P. Cortez, and P. Rita, "A data-driven approach to predict the success of bank telemarketing," *Decision Support Systems*, vol. 62, pp. 22–31, 2014.

[110] K-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf, "An Introduction to

Kernel-Based Learning Algorithms," *IEEE TRANSACTIONS ON NEURAL NETWORKS*, vol. 12, no. 2, pp. 181-201, 2001.

[111] L-P. Ni, Z-W. Ni, and Y-Z. Gao, "Stock trend prediction based on fractal feature selection and support vector machine," *Expert Systems with Applications*, vol. 38, pp. 5569–5576, 2011.

[112] J. Đ. Novaković, *Rešavanje klasifikacionih problema mašinskog učenja*. Štamparija SaTCIP, Vrnjačka Banja : Fakultet tehničkih nauka u Čačku , 2013.

[113] S. Olafsson, X. Li, and S. Wu, "Operations research and data mining," *European Journal of Operational Research*, vol. 187, no. 3, pp. 1429-48, 2008.

[114] E. C. Omak, K. Polat, S. Gunes, and A. Arslan, "A new medical decision making system: Least square support vector machine (LSSVM) with Fuzzy Weighting Pre-processing," *Expert Systems with Applications*, vol. 32, no. 2, pp. 409–414, 2007.

[115] J. Ouenniche, B. Pérez-Gladish, and K. Bouslah, "An out-of-sample framework for TOPSIS-based classifiers with application in bankruptcy prediction," *Technological Forecasting and Social Change*, 2017.

[116] S. J. Pan and Q Yang, "A Survey on Transfer Learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345-1359, 2010.

[117] J. Patel, S. Shah, P. Thakkar, and K Kotecha, "Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques," *Expert Systems with Applications*, vol. 42, no. 1, pp. 259-268, 2015.

[118] S. Pauwels, K. Inghelbrecht, P. Heyman, and D. Marius, "Technical Trading Rules in Emerging Stock Markets," *World Academy of Science, Engineering and Technology*, vol. 5, pp. 11-20, 2011.

[119] Y. Peng, Z. Wu, and J. Jiang, "A novel feature selection approach for classification of biomedical data," *Journal of Biomedical Informatics*, vol. 43, pp. 15-23, 2010.

[120] B. Pes, N. Dessì, and M. Angioni, "Exploiting the ensemble paradigm for stable feature selection: A case study on high-dimensional genomic data," *Information Fusion*, vol. 35, pp. 132-147, 2017.

[121] O. Phichhang and H. Wang, "Prediction of Stock Market Index Movement by Ten Data Mining Techniques," *Modern Applied Science*, vol. 3, no. 12, pp. 28-42, 2009.

[122] M. J. Pring, *The All-Season Investor*.: John Wiley & Sons. , 1992.

[123] P. Pudil and P. Somol, "Current Feature Selection Techniques in Statistical Pattern Recognition," in *Computer Recognition Systems. Advances in Soft Computing*.: Springer Berlin/Heidelberg, 2005, vol. 30, pp. 53-68.

[124] J.R. Quinlan, *C4.5: Programs for Machine Learning*.: Morgan Kaufmann, 1993.

[125] J.R. Quinlan, "Induction of Decision Trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.

[126] J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset Shift in Machine Learning*.: The MIT Press, 2009.

[127] M. Rabin, "Risk Aversion and Expected-Utility Theory: A Calibration Theorem," *Econometrica*, vol. 68, pp. 1281–1292, 2000.

[128] A. Rad, B. Naderia, and M. Soltani, "Clustering and ranking university majors using data mining and AHP algorithms: A case study in Iran ," *Expert Systems with Applications*, vol. 38, no. 1, pp. 755-763 , 2011.

[129] J. Ren, Z. Qiu, W. Fan, H. Cheng, and P.S. Yu, "Forward Semi-supervised Feature Selection," in *Advances in Knowledge Discovery and Data Mining. PAKDD 2008. Lecture Notes in Computer Science, vol 5012*.: Springer, Berlin, Heidelberg, 2008.

[130] M. Robnik-Sikonja and I. Kononenko, "Theoretical and Empirical Analysis of ReliefF and RReliefF," *Machine Learning Journal*, vol. 53, no. 1-2, pp. 23-69, 2003.

[131] T. L. Saaty, "Basic theory of the analytic hierarchy process: how to make a decision," *Rev.R.Acad.Cienc.Exact.Fis.Nat*, pp. 395-423, 1999.

[132] T. L. Saaty, *The Analytic Hierarchy Process*. New York: McGrawHill, 1980.

[133] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics," *BIOINFORMATICS*, vol. 23, no. 19, pp. 2507-2517, 2007.

[134] J. A. Sáez, J. Derrac, J. Luengo, and F. Herrera, "Statistical computation of feature weighting schemes through data estimation for nearest neighbor classifiers," *Pattern Recognition*, vol. 47, pp. 3941–3948, 2014.

[135] A. L. Samuel, "Some studies in machine learning using the game of checkers," *IBM Journal of Research and Development*, vol. 3, no. 3, pp. 210-229, 1959.

[136] S. Savić and M. Stanković, *Teorija sistema i rizika*. Beograd: Rubek, 2017.

[137] B. Schölkopf, P. Simard, A. J. Smola, and V. Vapnik, "Prior knowledge in support vector kernels," *In Advances in neural information processing systems*, pp. 640-646,

1998.

[138] B. Scholkopf and A. J. Smola, *Learning with Kernels Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, Massachusetts, London, England: The MIT Press, 2002.

[139] B. Schölkopf, A. Smola, and KR. Müller, "Kernel principal component analysis," in *In: Gerstner W., Germond A., Hasler M., Nicoud JD. (eds) Artificial Neural Networks — ICANN'97. ICANN 1997. Lecture Notes in Computer Science, vol 1327*, Springer, Berlin, Heidelberg, 1997.

[140] M. Sebban and R. Nock, "A hybrid filter/wrapper approach of feature selection using information theory," *Pattern Recognition*, vol. 35, no. 4 , pp. 835-846, 2002.

[141] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*.: Cambridge University Press, 2004.

[142] G. P. Siddayao, S. E. Valdez, and P. L. Fernandez, "Analytic Hierarchy Process (AHP) in Spatial Modeling for Floodplain Risk Assessment," *International Journal of Machine Learning and Computing*, vol. 4, no. 5, pp. 450-457, 2014.

[143] R. Singh, H. Kumar, and R.K. Singla, "TOPSIS Based Multi-Criteria Decision Making of Feature Selection Techniques for Network Traffic Dataset," *International Journal of Engineering and Technology (IJET)*, vol. 5, no. 6, pp. 4598-604, 2014.

[144] A. P. Sinha and H. Zhao, "Incorporating domain knowledge into data mining classifiers: An application in indirect lending," *Decision Support Systems*, vol. 46 , pp. 287–299, 2008.

[145] P. Somol, P. Pudil, J. Novovicova, and P. Paclõk, "Adaptive floating search methods in feature selection," *Pattern Recogniton Letter*, vol. 20, pp. 1157–1163, 1999.

[146] J. Z. Stanković, I. Marković, and O. Radović, "Предвиђање тренда belex15 индекса и његових конституената помоћу LS-SVM метода," *Анали Економског факултета у Суботици*, vol. 51, br. 3, str. 251-264, 2015.

[147] J. Stanković, Ivana Marković, and M. Stojanović, "Investment Strategy Optimization Using Technical Analysis and Predictive Modeling in Emerging Markets," in *Procedia Economics and Finance The Economies of Balkan and Eastern Europe Countries in the changed world, EBEEC 2014, Nis, Serbia*.: Elsevier B.V, 2014, vol. 19, pp. 51-62.

[148] M. B. Stojanović, Metodologija za izbor trening skupa zasnovana na konceptu

zajedničkih informacija kod predviđanja vremenskih serija metodama nadgledanog mašinskog učenja, Doktorska disertacija , 2013.

[149] B. M. Stojanović, M. M. Božić, M. M. Stanković, and Z. P. Stajić, "A methodology for training set instance selection using mutual information in time series prediction," *Neurocomputing*, vol. 141, no. 2, pp. 236–245, 2014.

[150] M. Sugiyama, M. Krauledat, and K-R. Muller, "Covariate Shift Adaptation by Importance Weighted Cross Validation," *Journal of Machine Learning Research* , vol. 8, pp. 985-1005, 2007.

[151] J.A.K. Suykens, J. De Brabanter, L. Lukas, and J. Vandewalle, "Weighted least squares support vector machines: robustness," *Neurocomputing*, vol. 48, pp. 85-105, 2002.

[152] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle, *Least Squares Support Vector Machines*. Singapore: World Scientific, 2002.

[153] J. Tang, S. Alelyani, and H. Liu, "Feature Selection for Classification: A Review," in *Data Classification Algorithms and Applications*.: Chapman and Hall/CRC Press, 2014, ch. II, pp. 37-64.

[154] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)* , pp. 267-288, 1996.

[155] V. N. Vapnik, *The nature of statistical learning theory*.: Springer, 1995.

[156] V. Vapnik and R. Izmailov, "Learning using privileged information: similarity control and knowledge transfer," *Journal of machine learning research*, vol. 16, no. 20232049, p. 55, 2015.

[157] V. Vapnik and A. Vashist, "A new learning paradigm: Learning using privileged information," *Neural Networks 22*, vol. 22, pp. 544-557, 2009.

[158] Vergara, J. R., and P. A. Estevez, "A review of feature selection methods based on mutual information," *Neural Computing and Applications*, vol. 24, pp. 175–186, 2014.

[159] Y. Wang, S. Chen, and H. Xue, "Support Vector Machine incorporated with feature discrimination," *Expert Systems with Applications*, vol. 38, pp. 12506-12513, 2011.

[160] L. Wang, P. Xue, and K. L. Chan, "Incorporating prior knowledge into SVM for image retrieval," in *In Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, Cambridge, UK, 2004, pp. 981-984.

[161] J. Wang, T. Yu, and W. Wang, "Integrating analytic hierarchy process and genetic

algorithm for aircraft engine maintenance scheduling problem.," in *InProceedings of the 6th CIRP-Sponsored International Conference on Digital Enterprise Technology*, 2010, pp. pp. 897-915.

[162] D. Wang and H. Zhang, "Group AHP and K-means cluster for a new segmentation of brand customer," *International Journal of Advancements in Computing Technology*, vol. 5, pp. 213-221, 2013.

[163] D. Wettschereck, D. W. Ahay, and T. Mohri, "A Review and Empirical Evaluation of Feature Weighting Methods for a Class of Lazy Learning Algorithms," *Artificial Intelligence Review*, vol. 11, no. 1–5, pp. 273–314, 1997.

[164] D. F. Wong, L. S. Chao, X. Zeng, M-I. Vai, and H.-L. Lam, "Time series for blind biosignal classification model," *Computers in Biology and Medicine*, vol. 54, no. 1 , pp. 32-36, 2014.

[165] W. K. Wong, M. Manzur, and B. K. Chew, "How Rewarding is Technical Analysis? Evidence from Singapore Stock Market," *Applied Financial Economics*, vol. 13, no. 7, pp. 543-551, 2010.

[166] X. Wu and R Srihari, "Incorporating prior knowledge with weighted margin support vector machines," in *KDD '04 Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, Seattle, WA, USA, 2004, pp. 326-333.

[167] H. Xing, M. Ha, B. Hu, and D. Tian, "Linear feature-weighted support vector machine," *Fuzzy Information and Engineering*, vol. 1, no. 3, pp. 289-305, 2009.

[168] B. Xue, M. Zhang, W. N. Browne, and X. Yao, "A Survey on Evolutionary Computation Approaches to Feature Selection," *IEEE Transactions on Evolutionary Computation*, vol. 20, no. 4, pp. 606 - 626, 2016.

[169] J. Yang and S. Olafsson, "Optimization-based feature selection with adaptive instance sampling," *Computers & Operations Research*, vol. 33, no. 11, pp. 3088-3106, 2006.

[170] J. Yang and S. Olafsson, "Optimization-based feature selection with adaptive instance sampling," *Computers & Operations Research*, vol. 33, no. 11, pp. 3088-3106, 2006.

[171] X. Yang, Q. Song, and A. Cao, "Weighted support vector machine for data classification," in *Neural Networks, 2005. IJCNN '05. Proceedings. 2005 IEEE International Joint Conference on*, Montreal, Que., Canada, 2005.

[172] J. Yao, S. Zhao, and L. Fan, "Enhanced Support Vector Machine Model for Intrusion Detection," *Rough Sets and Knowledge Technology LNCS*, vol. 4062, pp. 538-543, 2006.

[173] T. Yilmaz, A. Yazici, and M. Kitsuregawa, "Non-linear weighted averaging for multimodal information fusion by employing Analytical Network Process," in *Pattern Recognition (ICPR), 2012 21st International Conference on*, Tsukuba, Japan, 2012.

[174] P. D. Yoo, M.H. Kim, and Jan T., "Machine Learning Techniques and Use of Event Information for Stock Market Prediction: A Survey and Evaluation," in *Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce*, 2005, pp. 835 – 841.

[175] L. Yu, H. Chen, S. Wang, and K. K. Lai, "Evolving Least Squares Support Vector Machines for Stock Market Trend Mining," *Evolutionary Computation, IEEE Transactions on*, vol. 13, no. 1, 2009.

[176] L. Yuling, H. Guo, and J. Hu, "An SVM-based Approach for Stock Market Trend Prediction," in *Neural Networks (IJCNN), The 2013 International Joint Conference on*, Dallas, TX, USA, 2013, pp. 1-7.

[177] L. Yu and H. Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution," in *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)*, Washington DC, 2003.

[178] S. C. Yusta, "Different metaheuristic strategies to solve the feature selection problem," *Pattern Recognition Letters*, vol. 30, no. 5, pp. 525-534, 2009.

[179] L. Yu, S. Wang, and K. K. Lai, "Mining Stock Market Tendency Using GA-Based Support Vector Machines," in *WINE 2005, LNCS*, X. Deng and Y. Ye, Eds.: Berlin Heidelberg: Springer-Verlag., 2005, vol. 3828, pp. 336-345.

[180] Y., Hsu, A. Zhai and S. K Halgamuge, "Combining News and Technical Indicators in Daily Stock Price Trends Prediction," in *ISNN 2007, Part III, LNCS*, D. Liu et al., Ed.: Berlin Heidelberg: Springer-Verlag, 2007, vol. 4493, pp. 1087-1096.

[181] J. Zhao, K. Lua, and X. He, "Locality sensitive semi-supervised feature selection," *Neurocomputing*, vol. 71, pp. 1842–1849, 2008.

[182] H. Zhao, A. P. Sinha, and W. Ge, "Effects of feature construction on classification performance: An empirical study in bank failure prediction," *Expert Systems with Applications*, vol. 36, no. 2, pp. 2633-2644, 2009.

[183] L. P. Zhu, L. Lexin, L. Runze, and L. X. Zhu, "Model-Free Feature Screening for Ultrahigh-Dimensional Data," *Journal of the American Statistical Association*, vol. 106, no. 496, pp. 1464-1475., 2011.

[184] F. Zhu, L. Shao, J. Xie, and Y. Fang, "From handcrafted to learned representations for human action recognition: A survey," *Image and Vision Computing*, vol. 55, pp. 42-52, 2016.

BIOGRAFIJA AUTORA

Ivana Marković rođena je 20. decembra 1979. godine u Nišu, gde je završila osnovnu školu i gimnaziju sa odličnim uspehom. Diplomirala je 2005. godine na Elektronskom fakultetu Univerziteta u Nišu, na smeru Računarska tehnika i informatika, sa prosečnom ocenom u toku studija 9.34. Doktorske akademske studije upisala je takođe na Elektronskom fakultetu Univerziteta u Nišu, na studijskom programu Elektrotehnika i računarstvo. Sve ispite predviđene nastavnim planom i programom na doktorskim akademskim studijama položila je sa prosečnom ocenom 10.00. Temu doktorske disertacije pod naslovom „Izbor atributa primenom metoda odlučivanja kod prediktivnog modelovanja vremenskih serija nadgledanim mašinskim učenjem“ prijavila je 2016. godine.

Zaposlena je na Ekonomskom fakultetu Univerziteta u Nišu od oktobra 2008. godine, najpre kao saradnik u nastavi, a od 2011. godine u zvanju asistenta za užu naučnu oblast Informatika, Informatika i kibernetika u ekonomiji, pri Katedri za računovodstvo, matematiku i informatiku. Angažovana je u nastavi na predmetima Informatika i Elektronsko poslovanje.

Autor ili koautor je 23 naučna rada publikovanih u međunarodnim časopisima, tematskim zbornicima nacionalnog značaja, časopisima nacionalnog značaja, saopštenih na skupovima međunarodnog značaja i saopštenih na skupovima nacionalnog značaja i publikovanih u odgovarajućim zbornicima radova.

Univerzitet u Nišu
Elektronski fakultet

IZJAVA O AUTORSTVU

Izjavljujem da je doktorska disertacija, pod naslovom **Izbor atributa integracijom znanja o domenu primenom metoda odlučivanja kod prediktivnog modelovanja vremenskih serija nadgledanim mašinskim učenjem**, koja je odbranjena na Elektronskom fakultetu Univerziteta u Nišu:

- rezultat sopstvenog istraživačkog rada;
- da ovu disertaciju, ni u celini, niti u delovima, nisam prijavljivala na drugim fakultetima, niti univerzitetima;
- da nisam povredila autorska prava, niti zloupotrebila intelektualnu svojinu drugih lica.

Dozvoljavam da se objave moji lični podaci, koji su u vezi sa autorstvom i dobijanjem akademskog zvanja doktora nauka, kao što su ime i prezime, godina i mesto rođenja i datum odbrane rada, i to u katalogu Biblioteke, Digitalnom repozitorijumu Univerziteta u Nišu, kao i u publikacijama Univerziteta u Nišu.

U Nišu,14. novembar 2017. godine

Autor disertacije

Ivana P. Marković

Univerzitet u Nišu
Elektronski fakultet

IZJAVA O ISTOVETNOSTI ŠTAMPANOG I ELEKTRONSKOG OBLIKA DOKTORSKE DISERTACIJE

Ime i prezime autora: **Ivana Marković**

Naslov disertacije: **Izbor atributa integracijom znanja o domenu primenom metoda odlučivanja kod prediktivnog modelovanja vremenskih serija nadgledanim mašinskim učenjem**

Mentor: **Dr Milena Stanković, redovni profesor**

Izjavljujem da je štampani oblik moje doktorske disertacije istovetan elektronskom obliku, koji sam predala za unošenje u Digitalni repozitorijum Univerziteta u Nišu.

U Nišu, 14. novembar 2017. godine

Autor disertacije

Ivana P. Marković

Univerzitet u Nišu
Elektronski fakultet

IZJAVA O KORIŠĆENJU

Ovlašćujem Univerzitetsku biblioteku „Nikola Tesla“ da, u Digitalni repozitorijum Univerziteta u Nišu, unese moju doktorsku disertaciju, pod naslovom: **Izbor atributa integracijom znanja o domenu primenom metoda odlučivanja kod prediktivnog modelovanja vremenskih serija nadgledanim mašinskim učenjem**.

Disertaciju sa svim prilozima predala sam u elektronskom obliku, pogodnom za trajno arhiviranje.

Moju doktorsku disertaciju, unetu u Digitalni repozitorijum Univerziteta u Nišu, mogu koristiti svi koji poštuju odredbe sadržane u odabranom tipu licence Kreativne zajednice (Creative Commons), za koju sam se odlučila.

1. Autorstvo **(CC BY)**
2. Autorstvo – nekomercijalno **(CC BY-NC)**
3. Autorstvo – nekomercijalno – bez prerade **(CC BY-NC-ND)**
4. Autorstvo – nekomercijalno – deliti pod istim uslovima **(CC BY-NC-SA)**
5. Autorstvo – bez prerade **(CC BY-ND)**
6. Autorstvo – deliti pod istim uslovima **(CC BY-SA)**

U Nišu, 14. novembar 2017. godine

Autor disertacije

Ivana P. Marković