

Univerzitet u Sarajevu
Elektrotehnički fakultet Sarajevo
Zmaja od Bosne bb, Sarajevo

Kandidat: mr.sci. Bruno Trstenjak

Datum: 30.1.2014.

Prijedlog teme doktorske disertacije

Sadržaj

1. Predloženi naslov teme doktorske disertacije	2
2. Tip istraživanja	2
3. Motivacija za istraživanje	2
4. Pregled stanja u oblasti istraživanja	4
5. Cilj istraživanja i fokus rješenja problema	11
6. Metode i plan istraživanja	18
7. Očekivani izvorni naučni doprinos disertacije	19
8. Polazna literatura	20

1. Predloženi naslov teme doktorske disertacije

“ Model predikcije klasa objekata baziran na višedimenzionalnim heterogenim podacima ”

“ Model prediction of class objects based on multidimensional heterogeneous data ”

2. Tip istraživanja

Doktorska disertacija će se temeljiti na razvoju hibridnog algoritma za predikciju klasa objekata na osnovu višedimenzionalnih heterogenih podataka.

Validacija i verifikacija rezultata bit će izvršena eksperimentalno u laboratorijskom okruženju i u polju primjene. Istraživanje pripada užim naučnim oblastima: softver inženjeringa, umjetne inteligencije i data mininga.

3. Motivacija za istraživanje

U današnjem vremenu masovnog interneta, pojave „cloud computinga“ i sve veće količine podataka, pojavio se sve veći problem obrade podataka, utvrđivanja korisnih informacija. Čovjek je po svojoj prirodi radoznalo biće. Oduvijek je pokušavao predvidjeti budućnost, događaje koji će uslijediti kao rezultat određenih aktivnosti. Razvojem interneta, računala i IT tehnologije radoznalost se nije promijenila. Međutim, promijenila se količina informacija i čimbenici koji utječu na ishod i pouzdanost predviđanja raznih događaja. Da bi taj problem riješio razvila se posebna grana računarstva koja se bavi umjetnom inteligencijom, simulacijom rada ljudskog mozga, razvojem ekspertnih sustava i računala koja imaju sposobnost samoučenja na bazi prijašnjih podataka.

Predviđanje ili predikcija pojam je koji je usko vezan uz područje data mininga i strojnog učenja. Odnosi se na predviđanje ponašanja i događaja ovisno o poznatim ulaznim veličinama i prijašnjem naučenom iskustvu. Uspješnost i točnost predviđanja klasa objekata može rezultirati raznim uštedama, povećanjem sigurnosti sustava, smanjenjem rizika, povećanjem pouzdanosti i sl. Predikcija se može susresti u mnogim porama ljudskog života. Predikcija uspješnosti klasa objekata usko je povezana s pojmom demografskih podataka. Demografski podaci bogati su informacijama koje opisuju karakter osobe, podatke o mjestu njihovog stanovanja, opisuju karakter njihovih roditelja, stupanj obrazovanja roditelja i tako dalje. Ti podaci mogu u značajnoj mjeri utjecati na konačan rezultat predikcije [1]. Određivanje važnosti pojedinog demografskog atributa u cijelom postupku predikcije može biti od iznimne važnosti na rezultat i kvalitetu klasifikacije. Proučavajući odnose između demografskih podataka, njihovu međusobnu regresiju te utjecaj na konačan rezultat predikcije, može se doći do iznenađujućih spoznaja. Dobiveni rezultati predikcijom mogu u prvi plan istaknuti neki demografski atribut za koji se na početku smatralo da je nebitan i da nema veliki utjecaj na rezultat. Velik broj demografskih atributa postavlja određena ograničenja u odabiru modela učenja. Javlja se potreba prilagodbe formata podataka, njegove prilagodbe algoritmu koji se koristi u postupku učenja i klasifikacije. Problem velikog broja demografskih atributa i dimenzija, mogućnost nestandardnog oblika podataka, mogu kod nekih tehnika učenja izazvati poteškoće u postupku predikcije. Jedno od rješenja koje se često koristi je preprocesiranje i smanjivanje atributa [2]. Međutim, smanjivanje broja atributa, smanjivanje dimenzija podatka često rezultira s gubitkom informacija. Gubitkom informacija smanjuje se mogućnost određivanja veza između pojedinih atributa. Smanjenjem broja atributa može doći do toga da se nakon izvršene predikcije ne može odrediti potpuno značenje dobivenih rezultata.

U ovoj oblasti provedena su razna istraživanja [3,4,5], međutim još uvijek ima velikog prostora za napredak i drugačiji pristup rješavanja istih ili sličnih problema.

Moj interes za istraživanjem te želja za razvojem i stvaranjem novih vrijednosti temeljni su pokretači raznih aktivnosti ostvarenih u području računarstva. Motivi za odabir ovog prijedloga teme doktorske disertacije su višestruki i raznoliki.

Magistarski studij uveo me u svijet umjetne inteligencije što je rezultiralo mojom težnjom za istraživanje područja koja svojom primjenom daju ljudima rezultate često nezamislive. Jedno takvo područje je i područje strojnog učenja, predikcije i predviđanja raznih događaja. Područje koje ljude ostavlja zamišljene kada vide ostvarene rezultate. Osnovno pitanje je zašto baš razvoj hibridnog algoritma predikcije.

Prvi od motiva predložene teme je želja za kreiranjem kompleksnog sustava predikcije kod kojeg će centralni dio biti novi hibridni algoritam. Algoritam koji će znati vršiti klasifikaciju objekata na temelju mnogobrojnih višedimenzionalnih demografskih podataka i koji će omogućiti pozitivni pomak prema kvalitetnijoj predikciji, pomak prema univerzalnosti u radu s višedimenzionalnim podacima.

Razvoj sustava i algoritma koji će se bazirati samo na teoretskim postavkama bez dodira s realnim životom izgleda kao nedovršen posao. Dodatni motiv u ovoj disertaciji je osmisliti hibridni algoritam koji će biti moguće implementirati korištenjem softverskih okruženja i koristiti u mnogim područjima primjene kao što su u obrazovanju za predikciju uspješnosti studenta, u bankama za predikciju uspješne realizacije kredita, u osiguravajućim društvima za predikciju uspješnosti sporazuma sa osiguranicima. Testiranje algoritma i usporedba sa standardnim tehnikama biće provedeno u jednom području primjene.

Jedan od posljednjih motiva, na neki način i vlastito dokazivanje, je potreba za fuzijom ukupnog vlastitog znanja iz raznih IT područja. Kompleksnost zadatka zahtijevat će upotrebu stečenog znanje iz područja umjetne inteligencije, data mininga, baza znanja, upotrebe razvojnih alata, statistike, programiranja. Ovakva kompleksnost i širina potrebnog predznanja daje osjećaj još veće vrijednosti ukoliko se postavljeni ciljevi ostvare.

Kao posljednje tu je i promocija ostvarenih rezultata disertacije i svih istraživanja te objavljivanje u žurnalima i/ili međunarodnim konferencijama.

4. Pregled stanja u oblasti istraživanja

Strojno učenje

Postoji potreba da se predvidi neki događaj na osnovi postavljenih kriterija, da se objekti i razni događaji svrstavaju u određene klase. Jedan od načina rješavanja ovakve problematike je upotreba strojnog učenja. Razlikuju se dva osnovna principa strojnog učenja: učenje pod nadzorom (*supervised learning*) i učenje bez nadzora (*unsupervised learning*) [6,7]. Kod učenja pod nadzorom metode koriste ulazne podatke, atribut i izlazni rezultat. Kod učenja bez nadzora metode koriste samo ulazne podatke. Cilj je pronaći regularnost u događajima, što će se dogoditi, a što neće ovisno o ulaznim podacima. U statistici se procjenjuje gustoća događaja (*density estimation*). Jedna od metoda koja koristi ovakav pristup je klastering (*clustering*). Cilj ovakvih metoda je određivanje klastera ili grupa na osnovi ulaznih podataka i njihovih svojstava. Za potrebe ove doktorske disertacije istraživanja će se provoditi upotrebom strojnog učenja pod nadzorom. Metode učenja pod nadzorom često se koriste za klasifikaciju (*classification*) objekata na osnovi ulaznih svojstava. Kod tehnika strojnog učenja provodi se trening, testiranje i učenje na osnovi ulaznih podataka i poznatih izlaznih rezultata. Kroz trening i testiranje s poznatim podacima nastoji se definirati model koji će vršiti klasifikaciju uz najmanju moguću pogrešku te uz najveći postotak vjerojatnoće pravilnog klasificiranja [6,8].

Područje predikcije i tehnike strojnog učenja već duže vrijeme tema je istraživanja u raznim područjima znanosti [8,9,10]. Oblast strojnog učenja dinamično je područje u kojem znanstvenici konstantno pokušavaju pronaći nove i bolje algoritme s kojima će razni modeli davati što bolje rezultate. Trenutno se koristi velik broj različitih tehnika strojnog učenja s implementiranim različitim algoritmima, posebno za potrebe predikcije. U tablici 1 prikazane su neke od tehnika strojnog učenja koje se učestalo koriste.

Tablica 1. Pregled tehnika strojnog učenja

Tehnika učenja	Radovi
Decision Trees (DT)	[11][12][13][15][17][22][24][25]
Support Vector Machines (SVM)	[18][19][22]
Multilayer Perceptron (MLP)	[19][22]
Neural Networks (NN)	[11][19][22]
Case Based Reasoning (CBR)	[30]
Logistic Regression (LR)	[26][31][32]
Naïve Bayes (NB)	[17][19]
Bayesian Network (BN)	[16][17]
Naïve Bayes Tree (NBTree)	[17]
Genetic Algorithms (GA)	[12][14][19][30]
Rough Sets (RS)	[19][20][21]
Fuzzy Sets (FS)	[23][27][28]
Self-Organizing Maps (SOM)	[28][29]

U posljednjoj dekadi pojavili su se razni unaprijeđeni algoritmi, vrlo često kao rezultat specifičnosti postavljenog problema koji se želi riješiti upotrebom strojnog učenja [33,34,35]. Svaki od njih ima svoje prednosti i slabosti što utječe na konačan odabir kod istraživanja i upotrebe.

Tijekom raznih istraživanja u području strojnog učenja i data mininga uočena su razna ograničenja kod pojedinih tehnika učenja; dobra i loša svojstva. U tablici 2 prikazane su tehnike učenja, njihova pozitivna ili negativna svojstva na osnovi postavljenih uvjeta [36-40].

Uvjeti:

- upotreba višedimenzionalnih podataka (MD),
- naučeno znanje transparentno korisniku (ZN),
- brzina učenja (BU),
- tipovi podataka (Diskretni(D)/Binarni(B)/Kontinuirani(K)) (TP),
- brzina predikcije (BP),
- potrošnja memorijskih resursa (MR),
- stupanj točnosti predikcije (SP).

Analiza uvjeta pojedinih algoritama je izvedena u cilju skupljanja smjernica i karakteristika hibridnog algoritma koji će se razviti u okviru disertacije.

Tablica 2 Prikaz tehnika učenja s karakteristikama rada za pojedini uvjet

Tehnika učenja	MD	ZN	BU	TP	BP	MR	SP
Decision trees (DT)	D	D	D	DBK	D	L	S
Support Vector Machines (SVM)	S	L	L	-BK	D	L	D
Rule Learner (RL)	-	D	S	DBK	S	-	S
Neural Networks (NN)	S	L	L	-BK	D	D	D
Case Based Reasoning (CBR)	S	D	L	DBK	L	L	S
k-Nearest Neighbor (kNN)	D	S	D	DBK	L	L	S
Naive Bayes	D	D	D	DB-	D	D	L
Bayesian Network (BN)	D	D	D	DB-	D	D	D

Objašnjenje oznaka svojstava: **D**- dobro, **S**-srednje, **L**-loše

Dodatno objašnjenje za kolone TP i MR u tablici 2. U koloni TP naznačeni su tipovi podataka koji su podržani kod pojedine tehnike klasifikacije. Oznakom "-" naznačeno je da određeni tip podataka nije primjeren za konkretnu tehniku učenja. U koloni MR korištena je obrnuta logika vrednovanja. Traži se od tehnike učenja da što manje troši resurse računala, što daje smjernice gdje je moguće koristiti pojedinu tehniku. To je izrazito važno svojstvo kada model predikcije na primjer djeluje u on-line okruženju. Iako pojedine tehnike učenja podržavaju rad s višedimenzionalnim podacima postoje određena ograničenja s obzirom na karakter algoritma [41,42].

Razvojem novog hibridnog algoritma u okviru disertacije pokušat će se riješiti ta ograničenja te omogućiti upotreba podataka s velikim brojem dimenzija. Korištenje demografskih podataka, njihova raznolikost glede tipova podataka ne bi smjela biti limitirajući element u radu budućeg hibridnog algoritma.

Hibridni algoritmi strojnog učenja

Kod svake navedene tehnike strojnog učenja razvijeno je po nekoliko različitih algoritama koji se mogu koristiti u različitim područjima. Svaka od navedenih tehnika u tablici 2 ima dobra i loša svojstva. Svojstva daju smjernice u njihovom odabiru za pojedina područja istraživanja. Upravo to je jedan od razloga uvođenja te korištenja hibridnog modela strojnog učenja. Hibridni modeli koriste nekoliko različitih tehnika i algoritama za potrebu predikcije. Modelima se nastoji od svake tehnike dati ono što je najbolje i time dobiti što bolji rezultat. Hibridni modeli mogu se grupirati u tri kategorije: kaskadni hibridni model/klasifikator (*cascaded hybrid classifiers*), klaster/pojedinačni klasifikator (*cluster + single hybrid classifiers*) i integrirani hibridni klasifikator (*integrated hybrid classifiers*) [43]. Niti za jednu od kategorija ne može se tvrditi da ima najbolji pristup rješavanju problema klasifikacije. Svaki hibridni model rezultat je nekih specifičnih zahtjeva koji su odredili smjernice u strukturi modela. Razlozi mogu biti raznoliki,

kao što je zahtjev za točnost klasifikacije, zahtjev za bržim učenjem, karakter ulaznih podataka, višedimenzionalni podaci, količina podataka i slično. Vrlo često u hibridnim modelima su implementirane minimalno dvije različite tehnike. Jednom tehnikom se nastoji unaprijediti kvaliteta predikcije te optimizirati ulazne podatke, smanjiti njihov broj, odrediti najvažnije od njih za proces predikcije, odrediti regresiju među njima. Druga tehnika zadužena je za klasifikaciju ovako pripremljenih podataka [23,29,44]. U tablici 3. prikazani su neki od hibridnih algoritama strojnog učenja te su naznačene tehnike koje su korištene.

O popularnosti hibridnih algoritama ukazuju mnogi objavljeni članci u raznim područjima istraživanja kao što su: bioinformatika, mrežne komunikacije, bankarstvo, obrazovanje, rad s tekstovima, digitalna obrada slika, prepoznavanje uzoraka, i tako dalje. Vjerojatno ne postoji područje gdje nije upotrijebljen neki od hibridnih algoritama.

Tablica 3. Pregled hibridnih algoritama

Kategorija	Tehnike	Radovi
Kaskadni hibridni algoritam	GA+CBR	[45][30]
	LR+MLP	[47][48]
Klaster/pojedinačni klasifikator	SOM + MLP	[49]
	DT+ k-means	[46]
	Neighborhood rough set + SVM	[60]
Integrirani hibridni klasifikator	SVM + LR	[53]
	GA + SVM	[50][51][52][55]
	Fuzzy + SVM	[54]
	Fuzzy + MLP/NN	[56][57]
	CBR + SVM	[58]
	m-Medoids, SVM, Gaussian Mixture Model(GMM)	[59]

Pristupi za rješavanje predikcije na osnovi višedimenzionalnih podataka upotrebom hibridnih algoritama su različiti. Veoma dobri rezultati u radu s višedimenzionalnim podacima dobiveni su kod hibridnog algoritma koji je koristio algoritam stabla odlučivanja (*Decision Tree*) za potrebu klasifikacije ulaznih podataka te nakon toga određivanjem klastera pomoću k-means algoritma. Dobiveni rezultati pokazali su zavidnu brzinu predikcije uz upotrebu velikog broja dimenzija [46]. Nešto drugačiji pristup prikazan je u hibridnom modelu koji je koristio Self-Organizing Map (SOM) i Neuro-Fuzzy system (ANFIS)[61]. U ovoj kombinaciji SOM algoritam zadužen je za redukciju kompleksnih višedimenzionalnih podataka na način da ih organizira u klastere. Nakon toga mapirani podaci se šalju u algoritam za klasificiranje. SVM tehnika učenja veoma je popularna u znanstvenim krugovima što je moguće zaključiti iz velikog broja objavljenih članaka. SVM se susreće u konstrukciji hibridnih algoritama predikcije. Razvijeni su sljedeći hibridni modeli bazirani na SVM klasifikatoru: SVM u kombinaciji s Genetskim algoritmom (GA), SVM u kombinaciji s Fuzzy algoritmom, SVM + Straight forward wrapper, SVM + Support Vector Regression (SVR), SVM+ Neural networks (NN), SVM + CBR[62].

Kod svih tih spomenutih hibridnih algoritama pristup glede višedimenzionalnih podataka donekle je sličan, a to je da se pokušava smanjiti broj atributa ili reducirati veličinu podataka koji se koriste za trening modela. Ovakav pristup smanjivanja dimenzija podataka nije u duhu ideje prijedloga ove doktorske disertacije.

Sljedeći primjer uspješno realiziranog hibridnog algoritma je model sastavljen od Self Organizing Maps (SOM) algoritma i Multivariate Adaptive Regression Splines (MARS) algoritma za klasifikaciju. Ova kombinacija algoritama upotrijebljena je za predviđanja bankrota poduzeća. Promatrajući višedimenzionalnost, u hibridnom algoritmu SOM algoritam koristi se za grupiranje podataka i pripremu za klasifikaciju. Podaci se grupiraju prema sličnosti, a u

postupak treninga šalju se predstavnici pojedinih grupa [63]. Također je zanimljiv pristup za rješavanje problema predikcije i procjene upada u računarsku mrežu upotrebom hibridnog modela. Model je sastavljen od dva algoritma, K-Medoid i Naïve Bayes. Prvi algoritam zadužen je za grupiranje podataka u klastere, a Naïve Bayes provodi klasifikaciju podataka. Dobiveni rezultati predikcija pokazali su zavidnu brzinu u radu [64]. Prije navedeni hibridni algoritmi pokazali su određena ograničenja, kao što su ograničenja vezana uz broj dimenzija podataka koje mogu obraditi i osjetljivost pojedinih algoritama na tip podataka koji se koristi u predikciji. Novi hibridni algoritam trebao bi smanjiti takva ograničenja, posebno vezano za upotrebu podataka s velikim brojem dimenzija.

Preprocesiranje podataka

Kod učenja pod nadzorom, prije definiranja modela potrebno je izvršiti analizu ulaznih podataka i ostvarenih rezultata. Zbog raznovrsnosti podataka, velikog broja atributa, nedostataka pojedinih ulaznih podataka, irelevantnih podataka, velika pažnja u području data mininga i strojnog učenja posvećuje se preprocesiranju podataka [65,66]. Preprocesiranje podataka podrazumijeva „čišćenje“ podataka, normalizaciju, transformaciju, izdvajanje atributa te njihovu selekciju [67,68]. Sve te aktivnosti ne moraju biti zastupljene prije početka učenja novog modela, što ovisi o kvaliteti pripremljenih ulaznih podataka kao i o samom karakteru podataka.

Selekcija atributa

Kompleksnost problema i zahtjevnost predikcije nameće pitanje o minimalnoj i optimalnoj veličini seta podataka za uspješno učenje i predikciju. Pojedina istraživanja na tu temu pokušala su dati preporuke za konkretne probleme [41,42]. Međutim, često ti podaci su veoma ovisni i o samom karakteru podataka koji ulaze u proces predikcije. Generalno, algoritmi za odabir atributa mogu se grupirati u dvije skupine: filter algoritmi i wrapper algoritmi [4,5]. Kod selekcije atributa moguće je koristiti algoritam koji je ugrađen u samu metodu učenja i treninga, [69]. Evaluacija važnosti atributa pomoću filtriranja izračunava važnost svakog atributa i nakon toga odabire attribute koji imaju najbolji skor.

Za filtriranje atributa koriste se mnoge metode, neke od njih su: Information Gain (IG), Document Frequency (DF), χ^2 statistic (CHI), Expected cross entropy (ECE), Weight of evidence for text (WET), Odds ratio (ODD). Odabir metode uvelike ovisi o tome koje metode i algoritam će se koristiti u strojnom učenju, svojstvo višedimenzionalnih podataka [70,71,72,73]. Wrapper metode (Sequential backward selection (SBS), Sequential forward selection (SFS), Plus-L minus-R selection (LRS), Genetic algorithms (GA)) [74,75] koriste klasifikatore za određivanje vrijednosti atributa ili skupa atributa izračunom pogreške klasifikatora. Metode su „umotane“ u algoritam klasifikacije, a grupiraju se u dvije kategorije, pohlepne algoritme i slučajne/stohastičke algoritme [76]. Te metode daju u pravilu bolje rezultate od filter metoda, ali su znatno sporije te ih je potrebno svaki puta ponavljati kod bilo kakvih promjena u strukturi podataka kao i kod promjene algoritma strojnog učenja [77]. Osim spomenutih algoritama u nekim istraživanjima koristile su se hibridne metode selekcije atributa (*hybrid filter/wrapper method*) čime se nastoji iskoristiti najbolje od svakog algoritma [78]. Mnoga istraživanja i dobiveni rezultati pokazala su poboljšanje u točnosti klasifikacije odabirom podskupa atributa [2,3].

Višedimenzionalni demografski podaci

Višedimenzionalnost podataka povezana je s brojem varijabli/atributa kojima se opisuje objekt za koji se vrši klasifikacija. Velik broj atributa utječe na brzinu klasifikacije, brzinu kreiranja modela učenja, upotrebu resursa računala i slično. Zbog navedenih razloga provodi se reduciranje atributa. Višedimenzionalni podaci u pravilu su sastavljeni od različitih tipova atributa: nominalni, ordinarni, kontinuirani, podaci omjera, numerički [79], što prisiljava provođenje predradnje za transformaciju podataka pogodnih za postupak klasifikacije. Kada se spominje transformacija podataka neizostavno je potrebno spomenuti pojam diskretizacije. Diskretizacija je procedura koja se izvodi nad podacima, pretvaranje kontinuiranih tipova

atributa u kategorijski tipa podatka. Procedura dijeli raspon kontinuiranih vrijednosti u intervale te nastoji smanjiti broj mogućih vrijednosti atributa. Velik broj mogućnosti vrijednosti kontinuiranih atributa pridonosi sporom i neučinkovitom strojnom učenju. Manji broj intervala uvijek preferiraju tehnike induktivnog učenja [41]. Neke od tehnika diskretizacije: Class-Attribute Dependent Discretizar (CADD), Maximum Entropy (ME), Euqual Information Gain (EIG), Equal Interval Width (EIW), Equal Frequency Discretization(EFD), ... [80,81,82,83].

Mnogi algoritmi učenja primarno su orijentirani za rad s nominalnim atributima [84,85] ili mogu rukovati s diskretnim atributima[86], te imaju određena ograničenja u klasifikaciji kao što je navedeno u tablici 2.

Demografski podaci sastavni su dio u svakoj klasifikaciji koja uključuje podatke o ljudima, korisnicima, studentima i slično. Na koji način da se demografski podaci pripreme i prezentiraju neke od tehnika strojnog učenja? Podaci kao što je spol, godine starosti, mjesto rođenja, završeno obrazovanje, stupanj obrazovanja, nacionalnost i slično često se pretvaraju u nominalne vrijednosti pogodne za razne modele klasifikacije. Na taj način osiguravaju se diskretni tipovi podataka, tipovi podataka koji mogu koristiti velik broj tehnika učenja [87,88,89]. Novim hibridnim algoritmom pokušat će se umanjiti osjetljivost predikcije na tipove podataka. Takvo svojstvo hibridnog algoritma osigurala bi univerzalnost u upotrebi što će biti velika prednost kada se koriste demografski podaci koji obiluju raznim tipovima podataka.

Tijekom istraživanja proučavati će se dva pristupa prilagodbe atributa: statistički pristup upotrebom algoritama (*Data Driven Model*) i pristup u kojem se uključuje ekspert (*Domain Driven Model*). U većini istraživanja i modela učenja gdje klasifikator ima slabije rezultate u radu s velikim brojem atributa koriste se standardni pristupi optimiranja i smanjenja atributa [4]. Drugi pristup u određivanju važnosti atributa je pristup gdje se uključuje baza znanje eksperta iz određene domene problema. Ovaj pristup se susreće pod nazivom Domain - Driven Model. Modelom je moguće odrediti težinske vrijednosti pojedinih atributa i time na neki način utjecati na proces predikcije.

Predikcija klasa objekata

Predikcija klasa objekata u nekom području uvijek je bila od velikog interesa. Mogućnost predviđanja uspješnosti ljudi u raznim područjima potaknula je razna istraživanja što je dovelo do definiranja raznih modela klasifikacija. Jedno od područja koje je u centru pozornosti je sport te nastojanje da se odredi koji sportaši imaju predispozicije za najveće rezultate [90,91]. Drugo područje u kojem je u prvom planu čovjek i njegove karakteristike je financijsko područje, područje bankarstva, pr. određivanje rizika ulaganja, klasifikacija klijenata prema njihovim atributima [92,93].

Do sada je proveden relativno velik broj istraživanja koja se bave analizom učenja, pronalaženja odgovora na pitanje kako postići što bolje rezultate, povezanost između pojedinih varijabli koje imaju utjecaj na uspješnost određivanja klasa objekata. Dosadašnja provedena istraživanja o predviđanju uspješnosti studenta upotrebom strojnog učenja, često su bila usmjerena na upotrebu jedne od tehnika učenja. Jedno od takvih istraživanja provedeno je upotrebom neuronske mreže u predviđanju rizika uspješnosti studenata na Medicinskom fakultetu[94]; klasifikacija uspješnosti studenata u Novom Zelandu upotrebom CART (*Classification and Regression Tree*) metode[25]; određivanje uspješnosti za studente prve godine studija [95]; planiranje uspješnosti studenata upotrebom Neuro-Fuzzy sustava [96,97]; evaluacija uspjeha pomoću Fuzzy C-Means klastering[98]; evaluacija uspješnosti upotrebom Decision Tree[99], učenje na daljinu [100].

Relativno malo je istraživanja koja su na spomenutu temu pokušale spojiti predikciju uspješnosti studenata + on-line predikcija + hibridni model klasifikacije. Provedena su istraživanja na hibridnom modelu predikcije baziranom na neuronskoj mreži. Model koji omogućuje klasifikaciju studenata u više klasa [101]. Drugo zanimljivo istraživanje provedeno je u kreiranju klasičnog hibridnog modela u kombinaciji s genetskim algoritmom za optimiranje i ponovno neuronske mreže [102]. Upotreba hibridnog modela zasnovanog na proučavanju regresije

između varijabli može se koristiti za procjenu uspješnosti polaganja ispita[103]. Također, za predikciju je testiran hibridni model Bayesian mreža + neuronska mreža[104].

U predikciji klasa objekata često se koriste podaci iz raznih heterogenih izvora koja sadrže i demografske podatke. Razne analize pokazale su regresije i razne korelacije između pojedinih atributa. U analizama i klasifikacijama osim demografskih podataka znanstvenici su uzimali u obzir uspjeh institucije, njenu ocjenu kvalitete rada te razne vanjske faktore koji mogu pozitivno ili negativno utjecati na uspješnost [105,106,107].

Pregled referenci prema kriteriju metoda i tehnika i važnosti izvora objavljivanja

U tablici 4 prikazan je pregled referenci po kriteriju metode i tehnika.

Tablica 4. Pregled referenci po kriteriju metoda i tehnika

Metoda/Tehnika	Broj Referenci	Oznaka reference
Strojno učenje	47	[1][3][6] [7][8][11][12][13][14] [15][16][17][18][19] [20][21][22][23][24][25][26][27][28][29][30] [31][32][33][34][36] [37][38][39][40][42][66][79][84][86][91][111][112][116] [117][120][121][122]
Predikcija/klasifikacija	56	[1][5][6][12][14][16][17][18][19][20][21][22][23][24] [25][28][31][32][35][37][42][43][44][54][56][67][70][82] [84][87][88][90][91][92][93][94][95][96][97][98][99][1 00][102][103][104][105][106][107][109][110][113][114] [117][118][123][124]
Hibridni algoritam	35	[13] [14][16][23][29][30][33][39][44][45][46][47][48] [49] [50][51][52] [53] [54] [55][56][57][58][59][60][61][62][63][64][82][85][101][1 02][103][104]
Atributi/Selekcija atributa	39	[1] [2] [3] [4] [5] [8] [10] [12][18][21][24][25][34][37][54][55][58][61][62][65][69] [70][71][72][73][74][75][76][77][78][81][83][89][90][1 05][108][112][115][119]
Preprocesiranje	27	[6][7][8][14][20][39][40][41][43][48][49][65][66][67][68] [69][71][72][74][78][80][81][83][102][108][115][119]
Višedimenzionalnost, demografski podaci	32	[4][14][16][17][18][39][40][49][55][61][63][65][80][81][82][83][66][70][71][76][87][88][89][90][94][95][96][97] [99][100][101][102]

OBRAZLOŽENJE PRIJEDLOGA TEME DOKTORSKE DISERTACIJE

Literatura i istraživački radovi koji su korišteni u pripremi teme doktorske disertacije te su navedeni u spisku referenci imaju sljedeću relevantnost koja je prikazana u tablici 5.

Tablica 5. Pregled referenci prema važnosti izvora objavljivanja

Reference		Broj	Impact factor
Časopisi	IEEE Transactions on Systems	1	4,778
	Knowledge-Based Systems	3	4,104
	Pattern Recognition	4	3,219
	Decision Support Systems	1	3,037
	Journal of Biomechanics	1	3,031
	Data Mining and Knowledge Discovery	2	2,877
	Journal of Machine Learning Research (JMLR)	1	2,682
	Applied Soft Computing	6	2,526
	COMMUN ACM	3	2,511
	Expert Systems with Applications	9	2,339
	Magnetic Resonance Imaging (MRI)	1	2,286
	Accounting and Management Information Systems	1	2,274
	Artificial Intelligence	2	2,194
	IJARCSE	1	2,080
	European Journal of Operational Research	1	2,038
	IEEE Transactions...	4	1,890
	Medical Engineering & Physics	1	1,779
	International Journal of Engineering ... (IJERT)	1	1,760
	Medical Informatics and Decision Making	1	1,600
	Pediatric Exercise Science	1	1,570
	Machine Learning	2	1,454
	Computer Science & Information Technology	1	1,341
	Information Processing & Management	1	1,338
	International Journal ... (IJACSA)	3	1,324
	Neural Comput & Applic	1	1,168
	Educational and Psychological Measurement	1	1,070
	Journal of Educational Research	1	1,050
	IJSCE	2	1,000
	Computers & Electrical Engineering	1	0,928
	IJCA	6	0,821
	Journal of Intelligence Science	1	0,351
	International Journal of Info..	1	0,333
	IJCSI	5	0,242
	Journal of Data Science	2	-
Konferencije	IEEE Conference	4	-
	Conference – non IEEE	16	-
Ostalo	Knjige	5	-
	Workshope	6	-
	Publikacije	22	-
UKUPNO		124	

5. Cilj istraživanja i fokus rješenja problema

Cilj istraživanja

Razvoj novog hibridnog algoritma za predikciju klasa objekata na osnovu višedimenzionalnih heterogenih podataka.

Ovaj algoritam treba da zadovolji procesiranje maksimalnog broja atributa/dimenzija, da bude dovoljno adaptibilan heterogenom karakteru ulaznih višedimenzionalnih podataka a da pri tome postigne optimalnu brzinu i preciznost procesiranja.

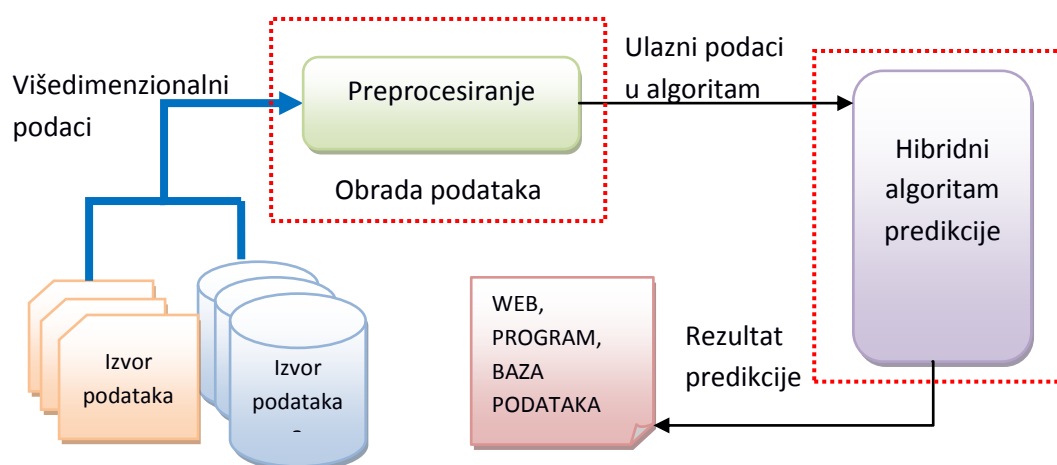
Da bi se ostvario cilj potrebno je uraditi sljedeće:

- Modelirati algoritam za predikciju klasa objekata na osnovu višedimenzionalnih heterogenih podataka.
- Definirati procedure preprocesiranja heterogenih ulaznih podataka.
- Odrediti načine određivanja težinskih vrijednosti atributa koji se koriste za predikciju.

Predloženi razvojni model

Prilikom razvoja hibridnog algoritma tijekom istraživanja dobiveni rezultati promatrat će se iz dva aspekta. Prvi aspekt bit će usmjeren na višedimenzionalnost, moguća ograničenja u budućem sustavu predikcije. Drugi aspekt usmjeren je na pojam demografskih podataka, njihovu prilagodbu hibridnom algoritmu, određivanje korelacije između atributa te utvrđenoj važnosti atributa za proces predikcije. Iz ove konstatacije, cijeli sustav u kojem će se odvijati istraživanje bit će sastavljen iz dva dijela: dio sustava klasifikacije s implementiranim hibridnim algoritmom te dio sustava zaduženog za pribavljanje višedimenzionalnih podataka, preprocesiranje i zapisivanje rezultata predikcije. Na slici 1. prikazan je cjeloviti razvojni sustav u kojem će djelovati hibridni algoritam.

Sustav predikcije za izvor višedimenzionalnih podataka moći će koristiti razne tipove datoteka, proračunskih tablica i razne baze podataka.



Slika 1. Razvojni sustav predikcije baziran na hibridnom algoritmu

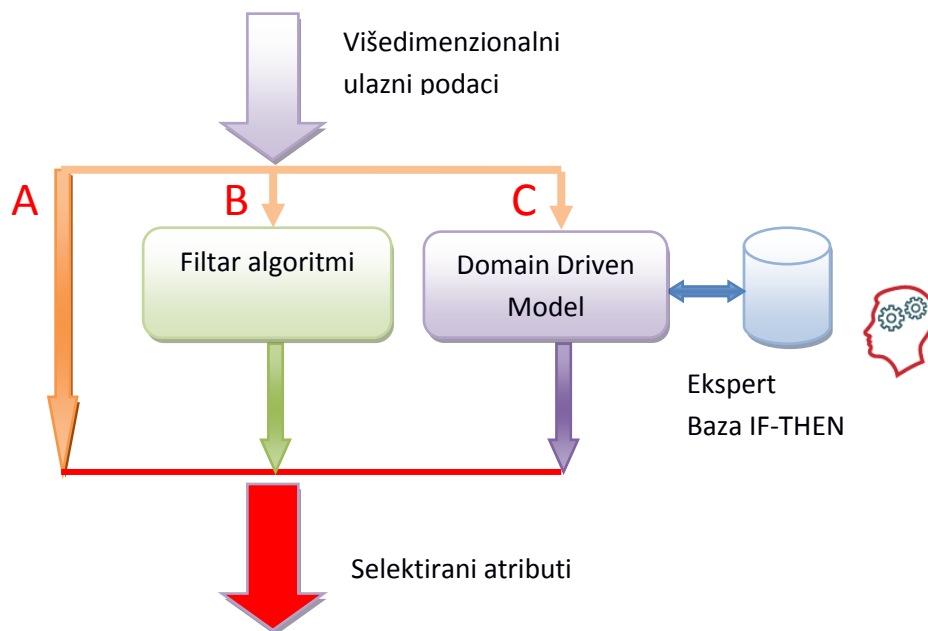
Prikupljeni višedimenzionalni podaci u sustavu prosljeđuju se na preprocesiranje. Preprocesiranje podataka podrazumijeva „čišćenje“ podataka, normalizaciju, transformaciju u oblik pogodan za model učenja i predikcije [67,68]. Kod preprocesiranja naglasak će biti na otklanjanju šuma u podacima, uklanjanju duplih podataka. Sve aktivnosti u pripremi podataka

bit će usmjerene na sprečavanje gubitka informacija u cilju kvalitetnije predikcije. Ovako pripremljeni podaci upućuju se u hibridni algoritam na predikciju.

Centralni i glavni dio sustava je hibridni algoritam koji će biti najkompleksniji u svojoj implementaciji. Tijekom istraživanja i razvoja algoritma odredit će se bazna tehnika učenja, ona koja će se pokazati kao najbolja za rad s višedimenzionalnim demografskim podacima.

Broj dimenzija ulaznih podataka, selekcija atributa, definiranje korelacije između atributa, sve su to elementi koji mogu utjecati na kvalitetu predikcije [108]. Zbog toga će se u prvom dijelu modela sustava nalaziti podmodul zadužen za optimizaciju procesa predikcije. Optimizacija će biti usmjerena prema što kvalitetnijoj selekciji atributa ukoliko će biti potrebno. Istraživanja će pokazati granične vrijednosti smanjivanja dimenzija ovisno o postignutoj točnosti predikcije. Ovaj podmodul pod nazivom „Selektor atributa“ usmjeravat će ulazne podatke ovisno o karakteru višedimenzionalnih podataka (broj dimenzija, vrsta modela). Struktura „Selektora podataka“ prikazana je na slici 2.

Ulazni podaci imaju mogućnost usmjeravanja u tri smjera. Smjer toka podataka pod oznakom A omogućava slanje izvornih podataka klasifikatoru bez smanjenja broja dimenzija. Istraživanja će odrediti optimalni broj dimenzija podataka za koje nema potrebe provoditi dodatnu selekciju atributa. Smjer toka podataka pod oznakom B šalje se odabranom algoritmu selektiranja atributa, a time i smanjenju dimenzija. Predviđeni su filter algoritmi, a istraživanja će pokazati koji od njih će ponuditi najbolje rezultate u radu s demografskim podacima. Ovaj dio selekcije pripada Data-Driven Modelu obrade podataka. Podaci jednake ili smanjene dimenzije prosljeđuju se u sljedeći dio hibridnog algoritma. Smjer podataka pod oznakom C odlazi u Domain Driven Model, model koji u svom odlučivanju koristi ekspertno znanje iz područja i domene predikcije te na osnovi iskustava provodi selekciju atributa. Model može ponuditi težinske vrijednosti pojedinih atributa kao polaznu točku u odlučivanju na osnovu formirane baze znanja sa if-then pravilima na osnovu domena problema.



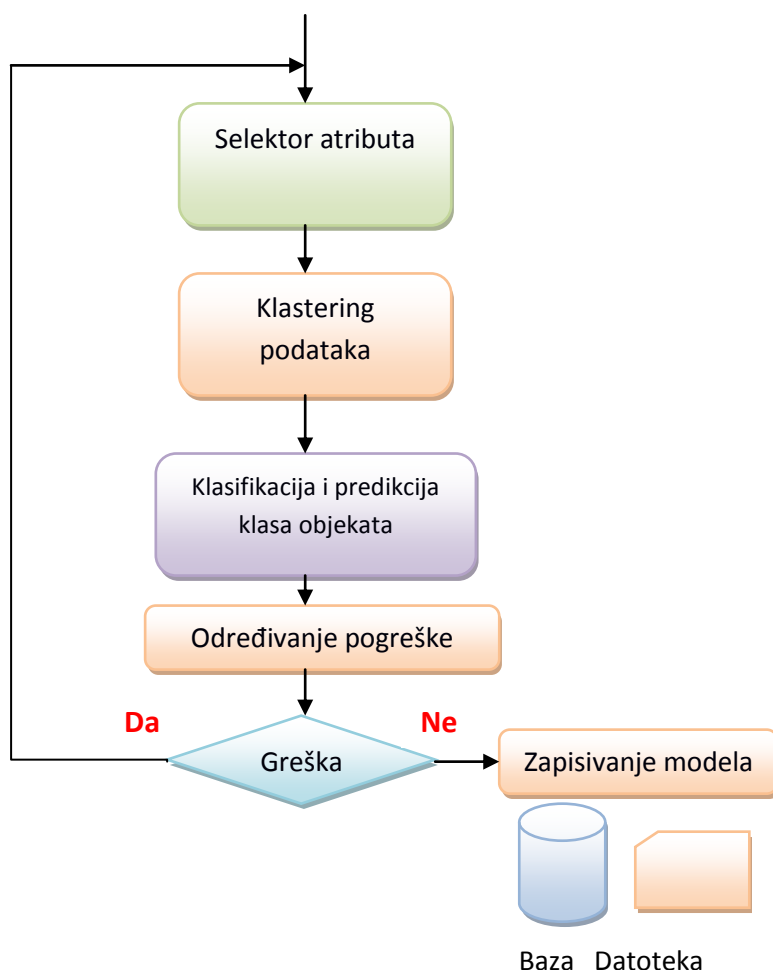
Slika 2. Struktura podmodula „Selektor atributa“

Odabrani atributi u višedimenzionalnom podatkovnom okruženju prosljeđuju se u sljedeću razinu sustava za predikciju. Prema planiranom konceptu taj dio u hibridnom algoritmu bio bi zadužen za klastering podataka. Klastering podataka je svrstavanje podataka i objekata prema njihovim zajedničkim svojstvima. Kao rezultat klasteringa može biti grupiranje skupine objekata

koji na prvi pogled nemaju mnogo toga zajedničko. Međutim, u procesu klasteringa objekti se promatraju u cijelosti, promatra se kompletna informacija koju objekti nose u sebi. Takve informacije karakteristične su kod demografskih podataka. Klasteringom višedimenzionalnih podataka istaknut će se klase podataka koja će se koristiti za trening modela. Takav pristup kod pojedinih kaskadnih hibridnih modela pokazao je dobre rezultate [61]. Nakon klasteringa podaci odlaze u klasifikator za predikciju. Tehnika učenja i algoritam klasifikacije odredit će se tijekom istraživanja.

Konačan oblik hibridnog algoritma odredit će evaluacija rezultata između algoritama klasičnih tehnika učenja i hibridnog algoritma. Istraživanje će također definirati kategoriju hibridnog modela: kaskadni hibridni model/klasifikator (*cascaded hybrid classifiers*), klaster/pojedinačni klasifikator (*cluster + single hybrid classifiers*) i integrirani hibridni klasifikator (*integrated hybrid classifiers*)[43].

Višedimenzionalnost je osnovni faktor u predikciji koji će imati najveći utjecaj u odabiru tehnike učenja. U tablici 2 navedene su tehnike učenja i njihova osnovna svojstva. U prvoj koloni nalaze se podaci o tome koliko je tehnika učenja uspješna u korištenju višedimenzionalnih podataka. Nakon izvršene klasifikacije provest će se evaluacija i mjerenje pogreške predikcije. U slučaju dobivenih loših rezultata hibridnog algoritma postupak treninga i testiranja će se ponoviti. Slika 3 prikazuje strukturu hibridnog algoritma.



Slika 3. Struktura hibridnog algoritma

Evaluacijska metrika kvalitete hibridnog algoritma

Uvijek u procesu razvoja potrebno je izmjeriti rezultate i izvršiti komparaciju dobivenih rezultata s prijašnjim. Istraživanje novog hibridnog algoritma vršit će se mjerenjem te usporedbom rezultata ostvarenih upotrebom baznih tehnika učenja. Tijekom istraživanja provodit će se trening i testiranje na dvije platforme. U prvoj će demografski višedimenzionalni podaci biti proslijeđeni odabranoj tehnici učenja. Nakon toga isti postupak ponavljat će se upotrebom hibridnog algoritma uz razne podvarijante vezano za selekciju atributa, broj klasa, način dijeljenja podataka tijekom treninga i testiranja sa aspekta sljedećih kriterija:

- stupanj točnosti predikcije (*accuracy*) i očekivana pogreška (*estimated error*),
- pretreniranost (*overfitting*),
- brzina učenja modela (*training/testing time*),
- potrošnja memorijskih i CPU resursa (*memory and CPU consumption*).

Stupanj točnosti predikcije i očekivana pogreška veoma su važni pokazatelji koji određuju da li je neki algoritam primjenjiv za neko područje istraživanja ili nije. Već duže vrijeme to je stalna tema istraživanja. Točnost predikcije, osim o tehnici strojnog učenja, ovisi o karakteru algoritma koji se koristi, načinu treniranja modela, karakteru trening podataka, njegovoj veličini kao i o specifičnosti kompozicije podataka koji se obrađuju tijekom predikcije. U praksi nije moguće stvarno izmjeriti stvarnu pogrešku u predikciji, zbog toga se koriste različiti pristupi procjene pogreške kao što su: Bootstrapping, Hold-out, Cross-validation [109]. Razlika između pojedinih metoda je u načinu dijeljenja podatka za trening i testiranje modela. Točnost predikcije može se izraziti jednostavnom jednadžbom[110]:

$$accuracy = \frac{\text{number of true positives} + \text{number of true negatives}}{\text{number of true positives} + \text{false positives} + \text{false negatives} + \text{true negatives}}$$

Prilikom analiza kvalitete predikcije pojedinog algoritma u istraživanjima često se koristi metoda Receiver Operating Characteristics (ROC). Metoda koristi vizualan način prikazivanja performansi klasifikatora pomoću grafova. ROC grafovi se također upotrebljavaju za vizualizaciju i analiziranje ponašanja dijagnostičkih sustava. ROC grafovi dopunjeni krivuljama, posebno za više klasnu predikciju relativno često se upotrebljavaju [111]. Osim ROC metode prikazivanja kvalitete klasifikatora u istraživanjima također se koristi mjerenje površine ispod ROC krivulje (*Area Under the ROC Curve (AUC)*), H-mjere i statistička mjere P-value, Mean Squared Error (MSE) [112,113,114]. Drugi važan pojam kod strojnog učenja i predikcije je očekivana pogreška. Ona je usko povezana s pojmom treniranja modela, treniranja podataka i tehnike koju koristimo za učenje modela predikcije. Očekivana pogreška predstavlja ukupni udio pogrešaka nastalih tijekom svih prolaza nad testnim podacima. U cilju smanjenja očekivane pogreške često se isprobavaju različiti načini dijeljenja testnih podataka kao što su spomenute metode Cross-validation, Repeated Hold-out, Bootstrap [115]. Cross-validation je tradicionalna metoda, često nazivana k-fold metoda. Ulazni podaci za trening i testiranje dijele se na k jednakih dijelova. Svaki dio barem jednom se ostavlja u procesu učenja za potrebu testiranja modela, a ostali dijelovi za treniranje modela. Kod Hold-out metode ulazni podaci se dijele na dvije grupe koje ne trebaju biti jednake veličine. Prva grupa koristi se za treniranje modela učenja, a druga za testiranje greške predikcije. Kod bootstrap metode ulazni skup podataka koristi se za izrađivanje novih skupina podataka na osnovi slučajnog odabira. Na taj način dobivaju se izmijenjene kopije podataka koje se šalju u model te se pokušava dobiti srednja pogreška modela.

Pretreniranost je pojava prenaučivosti modela učenja, a najčešći uzrok je prevelik broj podataka koji se koristi kod učenja i treniranja modela. Dolazi i kod forsiranja modela za što manjom pogreškom, tada dolazi do loše generalizacije. Generalizacija predstavlja interpolaciju podataka

najjednostavnijom krivuljom, svojstvo dobre klasifikacije za nepoznate ulazne podatke. Kod analize overfittinga koriste se dva parametra: opća pogreška na pristranost i varijanca. Pristranost se opisuje kao tendencija učenika da zbog dosljednog učenja radi iste pogrešne stvari. Prenaučenost se može smanjiti upotrebom određenih metoda dijeljenja ulaznih podataka kao što su cross-validation metode. Drugi način je korištenje statističkih testova značajnosti kao što je chi-square prije dodavanju nove strukture u procesu klasifikacije [116,117,118,119].

Veoma često za pojedini model učenja spominje se brzina učenja modela, pritom se razmatra vrijeme potrebno za trening budućeg modela i vrijeme potrošeno za testiranje. Kod pojedinih modela koji se koriste u on-line okruženju, koriste se modeli čiji se ulazni atributi dinamično mijenjaju. Modeli koji trebaju reagirati u realnom vremenu, brzina treninga i testiranja predstavljaju parametar koji definira upotrebljivost modela. Računarska složenost algoritama učenja postepeno postaje kritična i ograničavajući faktor u predikciji koja koristi veoma složene skupove podataka [120,121]. Kod ponekih tehnika učenja potrebno je uzeti u obzir specifičnost sustava na kojem se aplikacija izvršava (CPU, veličina i brzina memorije i skladišta podataka,...)[122]. U tablici 2. prikazana su određena svojstva tehnika učenja, između kojih je navedena i brzina učenja.

Današnja računala, u odnosu na računala iz bliske prošlosti, koriste znatno veće resurse. To se prvenstveno odnosi na procesorsku snagu te veličinu memorije. Veći memorijski kapaciteti donekle uklanjaju ograničenja pojedinih tehnika učenja za njihovu potrebu kada se koristi veći broj atributa i podataka. Zbog specifičnosti pojedini modeli učenja koriste modificirane algoritme. Karakteristični primjer je Decision trees tehnika učenja, često korištena u raznim područjima istraživanja. Kod nje u procesu učenja može se limitirati dubina stabla ili se izvršiti rezanje pojedinih dijelova stabla [123,124].

U procesu istraživanja u doktorskoj disertaciji provest će se određena metrika. Metrikom će se evaluirati odabrani standardni algoritmi za klasifikaciju pogodni za rad s višedimenzionalnim podacima, odabrani primjer već poznatog hibridnog algoritma i novi hibridni algoritam. Cilj usporedbe je definiranje dobrih i loših svojstava budućeg hibridnog algoritma te definiranje smjernica za njegovo poboljšanje. Višedimenzionalnost podataka kao osnovni kriterij istraživanja donekle određuje algoritme predikcije koji će se koristiti u evaluaciji. Analizom svojstava tehnika učenja, odabrana je Bayesian Network tehnik. Ova tehnika učenja podržava rad s višedimenzionalnim podacima, naučen model klasifikacije ima relativno kratko vrijeme predikcije i naučeno znanje je transparentno što znači da se kasnije može transformirati i prenijeti u neku bazu podatak. Od hibridnih algoritama, testiranje predikcije provest će se na hibridnom algoritmu sastavljenom od Decesion Tree + K-means. Definiranje klastera prije procesa predikcije ključni je element za predikciju podataka s većim brojem dimenzija ukoliko ne želimo smanjivati broj atributa. Nakon grupiranja podataka oni se šalju u algoritam za klasifikaciju. Dobiveni rezultati pokazali su zavidnu brzinu predikcije uz upotrebu velikog broja dimenzija [46].

U evaluaciji i usporedbi u disertaciji će se koristiti sljedeći kriteriji i metrika [110,111]:

1. Stupanj točnosti (*engl. Accuracy*)

Matematički izraz:

Klase	Klasificirano kao pozitivno	Klasificirano kao negativno
Predikcija pozitivnih (pos)	true pozitivno (tp)	false negativno (fn)
Predikcija negativnih (neg)	false pozitivno (fp)	true negativno (tn)

$$Accuracy = \frac{tp + tn}{tp + fn + fp + tn}$$

Metrika koja prikazuje omjer između točno klasificiranih pozitivnih i negativnih klasa u odnosu na ukupan broj pozitivnih i negativnih, točnih i netočnih klasifikacija.

2. Srednja točnost (*engl. Average Accuracy*)

Matematički izraz:

$$Average_accuracy = \frac{\sum_{i=1}^l \frac{tp_i + tn_i}{tp_i + fn_i + fp_i + tn_i}}{l}$$

Srednja točnost koristi za određivanje efikasnosti algoritma koji radi s višeklasnim podacima (*engl. Multi-class classification*). Efikasnost se odnosi na srednju vrijednost točnosti po klasi prilikom klasifikacije.

3. Preciznost (*engl. Precision*)

Matematički izraz:

$$Precision = \frac{\sum_{i=1}^l tp_i}{\sum_{i=1}^l (tp_i + fp_i)}$$

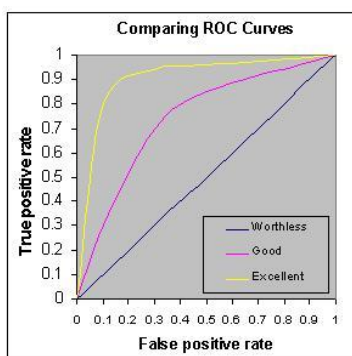
Kod predikcije u više klasa, ova mjera označava omjer sume svih točnih pozitivnih predikcija u odnosu na broj svih točnih i netočnih pozitivnih predikcija.

4. Receiver Operating Characteristics (ROC)

ROC krivulja grafički prikazuje odnos između osjetljivosti i specifičnosti klasifikacije dvaju razreda odabrane domene problema. Analiza osjetljivosti i specifičnosti testa ovisno o postavljanju granice koja odvaja "test-pozitivne" od "test negativnih" naziva se ROC analiza. Krivulja prikazuje rad klasifikatora. Pomoću koordinatnog sustava prikazuje rezultate: na apscisi – odnos lažno pozitivnih (1-specifičnost); na ordinati – odnos ispravno pozitivnih (osjetljivost).

$$Sensitivity = \frac{tp}{tp + fn}, \quad Specificity = \frac{tn}{tn + fp}$$

Što je klasifikator bolji (bolje naučen) to se njegova ROC krivulja približava gornjem lijevom uglu koordinatnog sustava. Na slici 4. prikazana je ROC krivulja.



Slika 4. Primjer ROC krivulje. Izvor slike: <http://gim.unmc.edu/dxtests/roc3.htm>

5. Area Under the ROC Curve (AUC)

Matematički izraz:

$$AUC = \frac{1}{2} \left(\frac{tp}{tp + fn} + \frac{tn}{tn + fp} \right)$$

AUC krivulja prikazuje odnos između udjela lažnih predviđanja (*engl. False Positive Rate*) na apscisi i točnih predviđanja (*engl. True Positive Rate*) na ordinati. Opisujući tako relativan odnos između koristi (točnog predviđanja, *engl. True Positive*) i troška (netočnog predviđanja, *engl. False Positive*). AUC je ekvivalentan vjerojatnosti da će klasifikator bolje rangirati na pozitivan primjer od slučajno odabranog negativnog primjera.

6. Učestalost pogreške klasifikacije po klasi (*engl. Error Rate*)

Matematički izraz:

$$Error_rate = \frac{\sum_{i=1}^l \frac{fp_i + fn_i}{tp_i + fn_i + fp_i + tn_i}}{l}$$

Ova mjera prikazuje srednju vrijednost pogrešnog klasificiranja po klasi predikcije. Učestalost pogreške predstavlja odnos broja pogrešnih klasificiranja u odnosu na ukupan broj podataka.

7. Brzina učenja (*engl Training & Testing time*)

Ovom metrikom odredit će se potrebno vrijeme učenje za određeni model. Ovaj podatak dosta je važan ukoliko očekujemo da se predikcije i proces učenja ostvaruje u internetskom okruženju. Kako se planira kasnija implementacija hibridnog algoritma nekim od programskih jezika ovaj parametar dat će informaciju da li će model predikcije zadovoljavati neke standarde koji su karakteristični za web okruženje. Tu se prvenstveno misli na vrijeme izvršavanja web aktivnosti. Ukoliko proces učenje modela trajao duže vrijeme to bi se moglo reflektirati na preopterećenja serverskih resursa, a kod korisnika izazvati osjećaj općeg zastoja web servisa.

Istraživanje u evaluacija rezultata bi trebalo dati odgovor o eventualnim ograničenjima hibridnog algoritma.

6. Metode i plan istraživanja

Metode istraživanja

U istraživanju će se koristiti osnovne, znanstvene metode te tehnike prikupljanja podataka. Jedna od metoda koja će koristiti je teorijsko proučavanja i istraživanja objavljenih stručnih i znanstvenih članaka iz oblasti umjetne inteligencije, tehnika strojnog učenja, klasifikacije, data mininga, prilagodbe podataka za potrebu učenja te njihovim optimiranjem. Statistička obrada podataka te analiza dobivenih rezultata.

U okviru ove teze u svrhu postizanja ciljeva koristiti će se sljedeće metode:

- Analiza tehnika, metoda i algoritama za mašinsko učenje i data mining.
- Modeliranje hibridnog algoritma predikcije klasa objekata zasnovanog na višedimenzionalnim heterogenim podacima.
- Prikupljanje statističkih podataka iz postojećih baza podataka. Izvršit će se grupiranje podataka te se formirati višedimenzionalni skup pogodan za upotrebu u predikciji.
- Transformacija podataka u oblik pogodan za upotrebu u hibridnom algoritmu. Oblik transformacije ovisit će o tehnici koja će se koristiti u hibridnom algoritmu.
- Anketiranje eksperata, kreiranje baze znanja za odabranu domenu predikcije.
- Računarska simulacija rada sustava predikcije. Trening i testiranje hibridnog algoritma nad stvarnim podacima.
- Mjerenje rezultata predikcije uslijed mijenjanja ulaznih atributa i njihovih vrijednosti. Određivanje optimalnog i maksimalnog broja dimenzija za uspješan rad algoritma.
- Mjerenje performansi rada hibridnog algoritma implementiranog u nekom programskom okruženju.
- Usporedba rezultata predikcije hibridnog algoritma i drugih odabranih algoritama.

Kod istraživanja i modeliranja koristit će se razvojno okruženje SPSS i WEKA za povezivanje, trening, testiranje i prezentaciju rezultata kod različitih tehnika učenja koje će se koristiti u razvoju hibridnog modela predikcije.

Plan istraživanja

Kako bi se ostvarili postavljeni ciljevi, istraživanje će biti koncipirano prema sljedećem planu:

1. Analiza modela
Definiranje komponenti sustava predikcije, određivanja načina povezivanja komponenti sustava.
2. Prikupljanje podataka
Iz postojeće baze podataka izdvojiti će se relevantni podaci nad kojima će se provoditi trening i testiranje budućeg modela. Podaci će biti višedimenzionalni s velikim brojem demografskih podataka. Preprocesiranje podataka i priprema za model klasifikacije.
3. Modeliranje hibridnog algoritma
Izraditi prijedlog konstrukcije hibridnog algoritma. Definirati tehniku učenja, način optimiranja podataka te način međusobnog povezivanja.
4. Implementacija algoritma u odbranom programskom okruženju.
5. Testiranje i validacija u laboratoriju i polju primjene.
6. Analiza svih dobivenih rezultata, komparacija rezultata s rezultatima ostvarenim odabranim baznim i hibridnim algoritmima.
7. Iterativno poboljšanje hibridnog algoritma predikcije klasa objekata.

7. Očekivani izvorni naučni doprinos disertacije

Doprinosi studije bit će vidljivi u naučnom i praktičnom polju.

Doprinos u naučnom polju je razvoj hibridnog algoritma za predikciju klasa objekata na osnovu višedimenzionalnih heterogenih podataka. Poboľšan po navedenim kriterijima u odnosu na odabrane referentne klasifikatore prikazane u sekciji „Evaluacijska metrika kvalitete hibridnog algoritma“.

U praktičnom polju algoritam treba da bude pogodan za implementaciju u različitim programskim okruženjima sa mogućnošću primjene na različite klase problema.

Elementi koji treba da potvrde doprinos disertacije su:

- Dokazan i istražena postojanje ograničenja prilikom klasifikacije višedimenzionalnih demografskih podataka kod standardnih tehnika učenja.
- Istražen i razvijen koncept ekspertne baze znanja za određenu domenu predikcije.
- Dokazan i istražen utjecaj određenih demografskih atributa na rezultat predikcije.
- Modeliran novi hibridni algoritam predikcije klasa objekata na osnovu višedimenzionalnih heterogenih podataka.
- Implementiran sustav predikcije sa novim hibridnim algoritmom za rad u realnom okruženju.

8. Polazna literatura

- [1] M. Jezewski, R.Czabanski, D. Roj, "Influence of Input Data Modification of Neural Networks Applied to the Fetal Outcome Classification", in Latest Trends on Computers, vol. 1, 2010,
- [2] M.R.Hossain, A.M. Than Oo, A.B.M.Shawkat Ali, "The Effectiveness of Feature Selection Method in Solar Power Prediction", in Power Engineering Research Group, 2013,
- [3] M.A.Jayaram, A.G. Karegowda, A.S. Manjunath, "Feature Subset Selection Problem using Wrapper Approach in Supervised Learning", in International Journal of Computer Applications, 2010,
- [4] Pablo Bermejo, Luis de la Ossa, José A. Gámez, José M. Puerta, "Fast wrapper feature subset selection in high-dimensional datasets by means of filter re-ranking", Knowledge-Based Systems, vol.25, 2012, pp. 35-44,
- [5] Sérgio Francisco da Silva, Marcela Xavier Ribeiro, João do E.S. Batista Neto, Caetano Traina-Jr., Agma J.M. Traina, "Improving the ranking quality of medical image retrieval using a genetic feature selection method", Decision Support Systems, vol.51, 2011, pp.810-820,
- [6] E.Alpaydm, "Introduction to Machine Learning", in MIT Press Cambridge, Massachusetts, London, 2004,
- [7] R.Dash, R.L.Paramguru, R.Dash, "Comparative Analysis of Supervised and Unsupervised Discretization Techniques", in International Journal of Advances in Science and Technology, vol. 2, No. 3, 2011,
- [8] Tom M. Mitchell, "Machine Learning", in McGraw-Hill Science, 1997,
- [9] L.David, O.D.Delen, "Advanced Data Mining Techniques", Springer, 2008,
- [10] P. Domingos, "A Few Useful Things to Know about Machine Learning", in Communications of the ACM 2012, vol. 55, 2012, pp.78-87,
- [11] C. Romero, S. Ventura, P. G. Espejo, C. Hervás, "Data mining algorithms to classify students", in 1st Int. Conf. on Educational Data Mining, 2008, pp. 187-191,
- [12] D.Kalles, C.Pierrakeas, "Analyzing student performance in distance learning with genetic algorithms and decision trees", in Artificial Intelligence, 2008, pp. 655-674,
- [13] R.R.Kabra, R.S.Bichkar, "Performance Prediction of Engineering Students using Decision Trees", in International Journal of Computer Applications, vol. 36, No.11, 2011,
- [14] E.Kızılkaya Aydogana, I.Karaoglanb, P.M. Pardalos, "Hybrid genetic algorithm in fuzzy rule-based classification systems for high-dimensional problems", Applied Soft Computing, 2012, pp. 800-805,
- [15] Barak Aviad, Gelbard Roy, "Classification by clustering decision tree-like classifier based on adjusted clusters", Expert Systems with Applications, vol.38, 2011, pp. 8220-8228,
- [16] L. Enrique Sucar, C.Bielza, E.F. Morales, P.Hernandez-Leal, J.H.Zaragoza, P.Larrañaga, "Multi-label classification with Bayesian network-based chain classifiers", Pattern Recognition Letters, 2013
- [17] D.Garcia-Saiz, M.Zorrilla, "Comparing classification methods for predicting distance students' performance", in JMLR: Workshop and Conference Proceedings, 2011, pp. 26-32,
- [18] J.Zheng, F.Shen, H.Fan, J.Zhao, "An online incremental learning support vector machine for large-scale data", in Neural Comput & Applic, 2013, pp.1023-1035,
- [19] W.A.Awad, S.M.ELseuofi, "Machine Learning methods for E-mail Classification", in IJCA, vol. 16, No.1, 2011,
- [20] Li Xiang-Wei, Qi Yian-Fang, "A Data Preprocessing Algorithm for Classification Model Based On Rough Sets", Physics Procedia, vol.25, 2012, pp.2025-2029,
- [21] H.Al-Qaheri, A.Ella Hassanien, A.Abraham, S. Zamoon, "Rough Set Generating Prediction Rules for Stock Price Movement", Second UKSIM European Symposium on Computer Modeling and Simulation, 2008,
- [22] S.Pandya, Dr.Paresh V. Virparia, "Comparing the Applications of Various Algorithms of Classification Technique of Data Mining in an Indian", in IJARCSSE, vol. 3, 2013,
- [23] A.Martin, V.Gayathri, G.Saranya, P.Gayathri, Dr.P. Venkatesan, "A HYBRID MODEL FOR BANKRUPTCY PREDICTION USING GENETIC ALGORITHM, FUZZY C-MEANS AND MARS", in International Journal on Soft Computing (IJSC), Vol.2, No.1, 2011,

- [24] S. Anupama Kumar, Dr. Vijayalakshmi M.N, "EFFICIENCY OF DECISION TREES IN PREDICTING STUDENT'S ACADEMIC PERFORMANCE", in CS & IT 02, 2011, pp. 335-343,
- [25] Zlatko J. Kovačić, "Early Prediction of Student Success Mining Students Enrolment Data", in Proceedings of Informing Science & IT Education Conference (InSITE), 2010,
- [26] Chao-Ying J. Peng, K. L. Lee, Gary M. Ingersoll, "An Introduction to Logistic Regression Analysis and Reporting", in The Journal of Educational Research, 2010,
- [27] E. Hullermeier, "Fuzzy Sets in Machine Learning and Data Mining", in IJACSA, Vol. 2, No. 6, 2011, pp 63-69,
- [28] Hai-Jun Rong, Guang-Bin Huang, N. Sundararajan, P. Saratchandran, "Online Sequential Fuzzy Extreme Learning Machine Function Approximation and Classification Problems", in IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART B: CYBERNETICS, VOL. 39, NO. 4, 2009,
- [29] S.Ismail,A.Shabri,R.Samsudin, "A hybrid model of self-organizing maps (SOM) and least square support vector machine (LSSVM) for time-series forecasting", in Expert Systems with Applications , 2011, pp. 10574-10578,
- [30] H.Ahn, Kyoung-jae Kim, "Bankruptcy prediction modeling with hybrid case-based reasoning and genetic algorithms approach", in Applied Soft Computing, vol.9, 2009, pp. 599-607,
- [31] G.Jibing, S. Shengtao, "A New Approach of Stock Price Trend Prediction Based on Logistic Regression Model", in International Conference on New Trends in Information and Service Science, 2009, pp.1366 - 1371
- [32] L.Weiguo, L. Cuiying, D.Xiaoping, Q.Kun, Z. Hanjie, H.Dezao, " A Traffic Flow Prediction Model Based on Ordered Logistic Regression", in Digital Content, Multimedia Technology and its Applications (IDC), 2010, pp. 213 - 216,
- [33] S. Yuan-Hai, Z.Chun-Hua, W.Xiao-Bo, D.Nai-Yang, " Improvements on twin support vector machines, Neural Networks", in IEEE Transactions, 2011, pp. 962 - 968,
- [34] J.Snoek, H.Larochelle, R.P.Adams, "Practical Bayesian Optimization of Machine Learning Algorithms", Carnell University Library, 2012,
- [35] J. Domenech, B. de la Ossa, J.Sahuquillo, J.A.Gil, A.Pont, "A taxonomy of web prediction algorithms", Expert Systems with Applications, vol.9,2012, pp.8496-8502
- [36] Y. Zhang, J.Wen, X Wang, Z.Jiang, "Semi-supervised learning combining co-training with active learning", Expert Systems with Applications, vol. 44, 2014, pp.2372-2378
- [37] L.Vanneschi, A.Farinaccio, G.Mauri, M.Antoniotti, P.Provero, M.Giacobini, " A comparison of machine learning techniques for survival prediction in breast cancer", Vanneschi et al. BioData Mining, 2011,
- [38] Hetal Bhavsar, Amit Ganatra, "A Comparative Study of Training Algorithms for Supervised Machine Learning", in International Journal of Soft Computing and Engineering, vol.2, 2012,
- [39] J.H. Zaragoza, L.E.Sucar, E.F. Morales, C.Bielza, P.Larranaga, "Bayesian Chain Classifiers for Multidimensional Classification", in International joint conference on Artificial Intelligence, vol.3 , 2011, pp.2192-2197,
- [40] Y.Zhang, P.Louvieris, M.Petrou, " Case-Based Reasoning Adaptation for High Dimensional Solution Space", in Transactions on Case-Based Reasoning for Multimedia Data, vol.1, No 1, 2008, pp. 21-36,
- [41] Ali Al-Ibrahim, " Discretization of Continuous Attributes in Supervised Learning algorithms", in International Journal of ACM Jordan, vol. 2, 2011,
- [42] G.T.Knofczynski, D.Mundfrom, " Sample Sizes When Using Multiple Linear Regression for Prediction", Educational and Psychological Measurement, vol. 68, No 3, 2008, pp.431-442,
- [43] Wei-Yang Lin, Ya-Han Hu, C.F. Tsai, " Machine Learning in Financial Crisis Prediction: A Survey", in IEEE Transaction on Systems, vol.42, No.4, 2012,
- [44] S.Fong, Yain-Whar Si, R.P.Biuk-Aghai, " Applying a Hybrid Model of Neural Network and Decision Tree Classifier for Predicting University Admission", in IEEE ICICS'09, 2009, pp. 1-5,

- [45] G. Paul Suthan, Lt. Dr. Santhosh Baboo, "Hybrid CHAID a key for MUSTAS Framework in Educational Data Mining", in IJCSI International Journal of Computer Science Issues, vol. 8, January 2011,
- [46] S. S. Kumar, E. Ramaraj, "A Hybride Model For Mining Multi Dimensional Data Sets", in International Journal of Computer Applications Technology and Research, vol. 2, 2013, pp.214 – 217,
- [47] C.F. Tsai, J.W. Wu, "Credit rating by hybrid machine learning techniques", App.Soft Comp., vol. 10, 2010, pp. 374-380,
- [48] Shu Ling Lin, "A new two-stage hybrid approach of credit risk in banking industry", in Expert Systems with Applications 36, 2009, pp.8333–8341
- [49] M.Sustersic, D.Mramor, J.Zupan, "Consumer credit scoring models with limited data", in Expert Syst. Appl., vol. 36, 2009, pp. 4736–4744,
- [50] W.Chen, C.Ma, L.Ma, "Mining the customer credit using hybrid support vectormachine technique", in Expert Syst. Appl., vol. 36, 2009, pp. 7611–7616,
- [51] F.Kuang, W.Xu, S.Zhang, "A Novel Hybrid KPCA and SVM with GA Model for Intrusion Detection", Applied Soft Computing, 2014,
- [52] Chih-Hung Wu, Y.Ken, T.Huang, "Patent classification system using a new hybrid genetic algorithm support vector machine", Applied Soft Computing, vol.10, 2010, pp. 1164-1177
- [53] M.K.Elbashir, W.Jianxin, "A hybrid approach of support vector machines with logistic regression for β -turn prediction", in Bioinformatics and Biomedicine Workshops (BIBMW), 2012,
- [54] Qi Xu, H.Zhou, Y.Wang, J.Huang, "Fuzzy support vector machine for classification of EEG signals using wavelet-based features", in Medical Engineering & Physics 31, 2009, pp. 858–865,
- [55] I.Aydin, M.Karakose, E.Akin, "A multi-objective artificial immune algorithm for parameter optimization in support vector machine", in Applied Soft Computing 11, 2011, pp. 120–129,
- [56] S.Alghowinem, R.Goecke, M.Wagner, J.Epps, T.Gedeon, M.Breakspear, G.Parker, "A COMPARATIVE STUDY OF DIFFERENT CLASSIFIERS FOR DETECTING DEPRESSION FROM SPONTANEOUS SPEECH", in ICASSP SP-P18.2, 2013,
- [57] E. A. Sodre, W. S. Motta, B. S. Alencar, "A hybrid intelligent system for power system security assessment", in XIII ERIAC, 2009,
- [58] Pei-Chann Chang, Chiung-Hua Huang, Chi-Yang Tsai, "A hybrid system by the integration of Case-Based-Reasoning with Support Vector Machine for prediction of financial crisis", in ICIC International, vol. 9, No 6, 2013,
- [59] S. Khalid, S. Arshad, "Framework for Constructing Hybrid Classifier using Weight Learning to Combine Heterogeneous Classifiers", in International Conference on Computational Intelligence, Modelling and Simulation, 2013, pp.163 – 168,
- [60] P.Yao, "Hybrid Classifier Using Neighborhood Rough Set and SVM for Credit Scoring", in International Conference on Business Intelligence and Financial Engineering, 2009, pp. 138-142,
- [61] M.Nilashi, O. bin Ibrahim and N. Ithnin, "Hybrid recommendation approaches for multi-criteria collaborative filtering", in Expert Systems with Applications, 2014, pp. 3879-3900,
- [62] J.Jayanthi, K.Suresh Joseph and J.Vaishnavi, "Bankruptcy Prediction using SVM and Hybrid SVM Survey", in International Journal of Computer Applications, 2011,
- [63] J.DE Andes, F.Sanchez-Lasheras, P.Lorca, F.J. DE Cos Juez, "A Hybrid Device Of Self Organizing Maps (SOM) and Multivariate Adaptive regression Splines for The Forecasting of Firms' Bankruptcy", in Accounting and Management Information Systems, vol. 10, 2011, pp. 351–374,
- [64] D.Upadhyaya, S. Jain, "Hybrid Approach for Network Intrusion Detection System Using K-Medoid Clustering and Naïve Bayes Classification", in International Journal of Computer Science Issues, 2013,

- [65] T.Ruuckstieß, C.Osendorfer, P. van der Smagt, "Minimizing data consumption with sequential online feature selection", in *Int. J. Mach. Learn. & Cyber.* DOI 10.1007/s13042-012-0092-x, Springer-Verlag, 2012,
- [66] T.Karunaratne, H.Boström, U.Norinder, "Pre-Processing Structured Data for Standard Machine Learning Algorithms by Supervised Graph Propositionalization - a Case Study with Medicinal Chemistry Datasets", in *Machine Learning and Applications (ICMLA)*, 2010,
- [67] G.Qiang, W.Xinmin, D.Chao, Li Chenguang, "Data preprocessing for prediction of rerecirculating water chemistry faults", in *International Conference on Computer Application and System Modeling*, 2010,
- [68] W.Jianping, "Research on Data Preprocessing in Supermarket Customers Data Mining", *Information Engineering and Computer Science*, 2010, pp.1-4,
- [69] S.Beniwal, J.Arora, "Classification and Feature Selection Techniques in Data Mining", in *International Journal of Engineering Research & Technology*, 2012,
- [70] Yan Xu, "A comparative study on feature selection in Chinese Spam Filtering", in *Application of Information and Communication Technologies (AICT)*, 2012, pp. 1-6,
- [71] C.Shang, M.Li, S.Feng, Q.Jiang, J.Fan, "Feature selection via maximizing global information gain for text classification", *Knowledge-Based Systems*, vol. 54, 2013, pp. 298-309,
- [72] J.M. Carmona-Cejudo, G. Castillo, M.Baena-Garcia, R. Morales-Bueno, "A Comparative Study on Feature Selection and Adaptive Strategies for Email Foldering", in *Intelligent Systems Design and Applications (ISDA)*, 2011,
- [73] Kavita Das, O. P. Vyas, "A Comparative Study of Four Feature Selection Methods for Associative Classifiers", in *Computer and Communication Technology*, 2010, pp. 431-435,
- [74] S.F. Pratama, A.K. Muda, "Feature Selection Methods for Writer Identification: A Comparative Study", in *International Conference on Computer and Computational Intelligence*, 2010,
- [75] X. Liu, L.Shang, "A Fast Wrapper Feature Subset Selection Method Based On Binary Particle Swarm Optimization", in *IEEE Congress on Evolutionary Computation*, 2013, pp. 3347-3353,
- [76] I.A.Gheyas, L.Smith, "Feature subset selection in large dimensionality domains", in *Pattern Recognition*, 2009,
- [77] Hui-Yan Wang, "A Comparative Study Of The Feature Selection Influence On Diagnosis In Traditional Chinese Medicine", in *Proceedings of the Ninth International Conference on Machine Learning and Cybernetics*, Qingdao, 2010,
- [78] P.A. Estévez, M.Tesmer, C.A.Perez, J.M.Zurada, "Normalized Mutual Information Feature Selection", in *Transformation on Neural Networks*, 2009,
- [79] Ian H. Witten, E. Frank, M.A. Hall, "Data Mining – Practical Machine Learning Tools and Techniques", in *Morgan Kaufmann Pub., Burlington USA*, 2011,
- [80] S.Garcia, J.Luengo, J.Antonio Saez, V.Lopez, F.Herrera, "A Survey of Discretization Techniques: Taxonomy and Empirical Analysis in Supervised Learning", in *Transaction on Knowledge and Data Engineering*, 2013,
- [81] A.J. Ferreira, M.A.T. Figueiredo, "An unsupervised approach to feature discretization and selection", *Pattern Recognition*, vol. 45, 2012, pp. 3048-3060
- [82] Tzu-Tsung Wong, "A hybrid discretization method for naïve Bayesian classifiers", *Pattern Recognition*, vol.45, 2012, pp. 2321-2325
- [83] Y. Yang, G.I. Webb, X. Wu, "Discretization Methods", in *Data Mining and Knowledge Discovery Handbook*, 2010, pp. 101-116,
- [84] Steven C.H. Hoi, Rong Jin, Peilin Zhao, Tianbao Yang, "Online Multiple Kernel Classification", *Machine Learning*, Springer, 2012,
- [85] Y.Bodyanskiy, O.Kharchenko, O.Vynokurova, "Hybrid cascade neural network based on Wavelet-Neuron", in *International Journal "Information Theories and Applications"*, Vol. 18, No. 4, 2011,
- [86] *The Top Ten Algorithms in Data Mining*, Chapman & Hall/CRC Data Mining and Knowledge Discovery, X. Wu and V. Kumar. CRC Press. 2009,

- [87] Sara Speltdoorn, "Predicting demographic characteristics of web users using semi-supervised classification techniques", in Master in Applied Economic Sciences, Ghent University Faculty of Economics and Business Administration, 2010,
- [88] H. Rasis, A. Erwin, J. Purnama, M. Galinium, "Automatic Demographic Classification of Indonesian Twitter Users", in Information Systems International Conference, 2013,
- [89] B. Lupin, E.M. Rodriguez, "Quality attributes and socio-demographic factors affecting channel choices", in International Association of Agricultural Economists, 2012,
- [90] J.M. Saavedra, Y. Escalante, "A Multivariate Analysis of Performance in Young Swimmers", in Pediatric Exercise Science, 2010, pp. 135-151,
- [91] R. K. Fukuchi, B.M. Eskofier, M. Duarte, R. Ferber, "Support vector machines for detecting age-related changes in running kinematics", in Journal of Biomechanics, 2011, pp. 540-542,
- [92] Xiao-Lin Li, Yu Zhong, "An Overview of Personal Credit Scoring: Techniques and Future Work", in International Journal of Intelligence Science, 2012, pp. 181-189,
- [93] H. S. Kim and S. Y. Sohn, "Support Vector Machines for Default Prediction of SMEs Based on Technology Credit", in European Journal of Operational Research, vol. 201, No. 3, 2010, pp. 838-846,
- [94] J.K. Alenezi, M.M. Awany, Maged M.M. Fahmy, "Effectiveness of Artificial Neural Networks in Forecasting Failure Risk for Pre-Medical Students", in Computer and Information Science, 2011,
- [95] A. Olani, "Predicting First Year University Students' Academic Success", in Electronic Journal of Research in Educational Psychology, 2009, pp. 1053-1072,
- [96] U. Rahul Saxena, S.P. Sinh, "Integrating Neuro-Fuzzy Systems to Develop Intelligent Planning Systems for Predicting Students' Performance", in International Journal of Evaluation and Research in Education, 2012, pp. 61-66,
- [97] N. Mymoon Zuviria, S. Letishia Mary, V. Kuppammal, "SAPM: ANFIS Based Prediction of Student Academic Performance", in International Conference on Computing, Communications and Networking Technologies, 2012,
- [98] R.S. Yadav, V.P. Singh, "Modeling Academic Performance Evaluation using Fuzzy C-Means Clustering Techniques", in International Journal of Computer Applications, 2012,
- [99] M. Wook, Y.H. Yahaya, N. Wahab, M.R.M. Isa, N.F. Awang, Y.S. Hoo, "Predicting NDUM Student's Academic Performance Using Data Mining Techniques", in Computer and Electrical Engineering, 2009, pp. 357-361,
- [100] S. Kotsiantis, K. Patriarcheas, M. Xenos, "A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education", in Knowledge-Based Systems, 2010, pp. 529-535,
- [101] Qu Jin, P.K. Imbrie, X. Chen, "A Multi-Outcome Hybrid Model For Predicting Student Success in Engineering", in American Society for Engineering Education, 2011,
- [102] R. Alkhasawneh, "Developing a Hybrid Model to Predict Student First Year Retention and Academic Success in STEM Disciplines Using Neural Networks", in Virginia Commonwealth University, 2011,
- [103] J. Gamulina, O. Gamulinb, D. Kermek, "Data mining in hybrid learning- possibility to predict the final exam result", in MIPRO, 2013,
- [104] A.I.Z. Abidin, I.A. Setu, S.P. Yong, O.M. Foong, J. Ahmad, "Classifying Student Academic Performance: A Hybrid Approach", in International MultiConference of Engineers and Computer Scientists, 2008,
- [105] S.B. Kotsiantis, "Use of machine learning techniques for educational purposes: a decision support system for forecasting students' grades", in Artif Intell Rev, 2012, pp. 331-344,
- [106] M. Ramaswami, R. Bhaskaran, "A CHAID Based Performance Prediction Model in Educational Data Mining", in IJCSI International Journal of Computer Science Issues, vol. 7, No. 1, 2010,
- [107] B. Kumar Baradwaj, S. Pal, "Mining Educational Data to Analyze Students Performance", in International Journal of Advanced Computer Science and Applications, vol. 2, No. 6, 2011,

- [108] A.Arauzo-Azofra, J.L.Aznarte, J.M.Benítez, "Empirical study of feature selection methods based on individual feature evaluation for classification problems", *Expert Systems with Applications*, vol.38, 2011, pp. 8170-8177,
- [109] J.E.Smith, P.Caleb-Solly, M.A.Tahir, D.Sannen, H.van-Brussel, "Making Early Predictions of the Accuracy of Machine Learning Applications", Cornell University Library, Computer Science, 2012,
- [110] M. Sokolova, G. Lapalme, "A systematic analysis of performance measures for classification tasks", in *Information Processing and Management*, 2009, pp. 427-437,
- [111] José Hernández-Orallo, "ROC curves for regression", *Pattern Recognition*, vol.46, 2013, pp. 3395-3411,
- [112] J.Juntu, J.Sijbers, S. De Backer, J.Rajan, D.Van Dyck, "Machine Learning Study of Several Classifiers Trained With Texture Analysis Features to Differentiate Benign from Malignant Soft-Tissue Tumors in T1-MRI Images", in *Journal Of Magnetic Resonance Imaging*, 2010, pp.680-689,
- [113] C Anagnostopoulos, D.J.Hand, N.M.Adams, "Measuring classification performance: the hmeasure package", Department of Mathematics, South Kensington Campus, Imperial College London, 2012,
- [114] J.M. Santos, M. Embrechts, "On the Use of the Adjusted Rand Index as a Metric for Evaluating Supervised Classification", in *Artificial Neural Networks – ICANN*, 2009, pp. 175-184,
- [115] Ji-Hyun Kim, "Estimating Classification Error Rate: Repeated Cross-validation, Repeated Hold-out and Bootstrap", Department of Statistics and Actuarial Science, Soongsil University, 2009,
- [116] P.Domingos, "A Few Useful Things to Know about Machine Learning", in *Communications of the ACM*, 2012, pp.78-87,
- [117] S.Whiteson, B.Tanner, M.E.Taylor, P.Stone, "Protecting Against Evaluation Overfitting in Empirical Reinforcement Learning", in *Adaptive Dynamic Programming And Reinforcement Learning (ADPRL)*, 2011, pp. 120 – 127,
- [118] J.D. Rodriguez, A. Perez, J.A. Lozano, "Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation", in *Transaction on Pattern Analysis and Machine Intelligence*, 2010,
- [119] G.C. Cawley, N.L.C. Talbot, "On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation", in *Journal of Machine Learning Research* 11, 2010, pp. 2079-2107,
- [120] L.Bottou, "Large-Scale Machine Learning with Stochastic Gradient Descent", in *Proceedings of COMPSTAT' 2010*, 2010, pp. 177-186,
- [121] P.Domingos, "A few useful things to know about machine learning", in *Communications of the ACM*, vol.55, 2012, pp.78-87,
- [122] A.Matsunaga, J. Fortes, "On the use of machine learning to predict the time and resources consumed by applications", in *ACM International Conference on Cluster, Cloud and Grid Computing*, 2010, pp. 495 – 504,
- [123] J. Sylvester, S. Weems, J. Reggia, "Predicting Improvement on Working Memory Tasks with Machine Learning Techniques", Technical Report, 2011,
- [124] Yuh-Jyh Hu, Tien-Hsiung Ku, Rong-Hong Jan, Kuochen Wang, Yu-Chee Tseng, Shu-Fen Yang, "Decision tree-based learning to predict patient controlled analgesia consumption and readjustment", in *Medical Informatics and Decision Making*, 2012.