Data Mining: Implikacije wrapper pristupa selekcije atributa na performanse klasifikacionog modela

Olivera Janković
ORAO a.d.
Bijeljina, Republika Srpska, BiH
janolja@yahoo.com

Sažetak— **Sa ciljem povećanja ukupne tačnosti klasifikacionog modela, eliminisanja negativnog uticaja irelevantnih atributa na performanse klasifikacije, u radu će biti prikazan način redukcije obima podataka selekcijom atributa, korištenjem metode omotača (wrapper). Eksperimentalna postavka wrapper pristupa za modeliranje u osnovi uključuje tri različita algoritma, te pohlepnu i metodu najbolji prvi za pretraživanje podskupova. Uticaj wrapper pristupa dat je kroz poređenje klasifikacione tačnosti, klasifikatora stablo odlučivanja J48 i algoritama najbližih susjeda IB1 i IBk, na originalnom i redukovanom skupu podataka (nastalom primjenom wrapper metode), u okruženju Weka alata za data mining.**

Ključne riječi- data mining; klasifikacija; selekcija atributa; wrapper metod;

I. UVOD

Već duže vrijeme svjedočimo sve moćnijim računarima, sve bržim i jeftinijim medijima za smještanje podataka, sve bržim komunikacionim mogućnostima što je pored ostalog (internet, WWW,…) velikom broju poslovnih subjekata omogućilo sve lakšu ali i jeftiniju pohranu velikih količina poslovnih podataka (baze podataka, slike, dokumenti, …) u elektronskoj formi. Cilj područja istraživanja podataka (*data mining*), sada već dobro poznate discipline, je pronalaženje, korištenjem metoda i implementacijom tehnika u okviru softverskih alata, potencijalno korisnih obrazaca u podacima koji postoje, kako bi se na određen način riješili aktuelni problemi boljim korištenjem dostupnih podataka [1], dobijanjem korisnih informacija, koje mogu biti iskorištene u kontekstu povećanja prihoda i/ili smanjnja troškova na primjer.

Kako bi se postigle što je moguće bolje performanse, bolji klasifikacioni rezultati, potrebno je nad postojećim podacima izvršiti određenu predobradu [2], kako bi se u moru prikupljenih pronašli oni podaci koji su u određenom kontekstu korisniji od drugih. Zahvaljujući predobradi podataka, moguće je prilagoditi podatke da ispune ulazne zahtjeve svakog data mining algoritma. Predobrada podataka uključuje i tehnike redukcije podataka, čiji cilj je smanjenje kompleksnosti podataka, otkrivanje i uklanjanje irelevantnih (nebitnih), kao i elemenata šuma iz podataka [3]. Jedna od važnih i najčešće korištenih tehnika u predobradi podataka za data mining je selekcija atributa (*feature selection*) [4],[5], koja predstavlja proces odabrira podskupa relevantnih atributa za izgradnju modela i koja je korisna, gledano iz perspektive procesa analize podataka, jer pokazuje koje su ulazne varijable ili atributi bitni za predikciju i u kojem su međusobnom odnosu. Krajnji cilj odabira atributa u kontekstu zadataka klasifikacije, jednog od uobičajenih problema data mininga, u biti je povećati tačnost klasifikacije [6], [7].

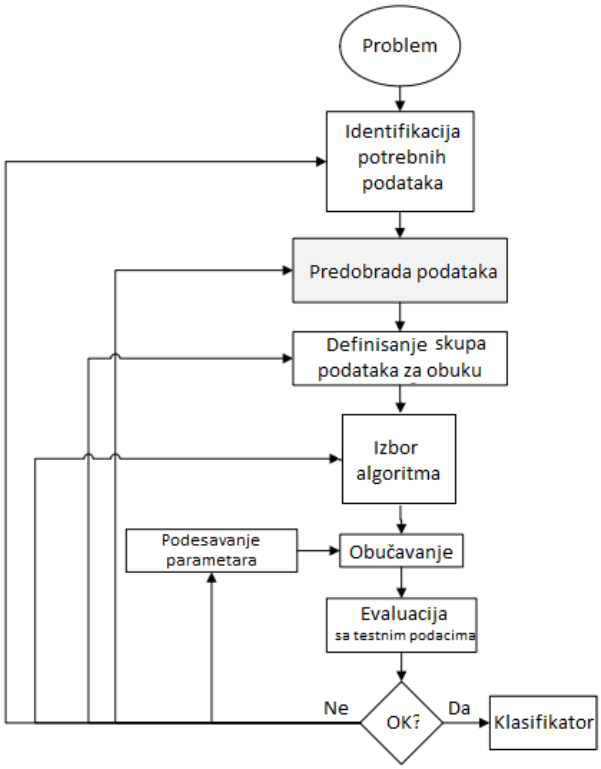
U radu će biti prikazani i diskutovani rezultati eksperimenata u kojima je primjenjen metod selekcije atributa korištenjem metode omotača (wrapper method), na atributima instanci seta podataka za procjenu automobila. Korišteni wrapper pristup ukjučuje tri različita algoritma te pohlepnu (*greedy*) i metodu najbolji prvi (*best first*), (sa različitim smjerovima pretraživanja) za pretraživanje podskupova. Za potrebe prezentovanja uticaja wrapper pristupa, putem rezultata klasifikacije na originalnom i redukovanom skupu podataka, korištena su klasifikatori: stablo odlučivanja J48 i algoritmi najbližih susjeda IB1 i IBk, sa pretpostavljenim ciljem povećanja ukupne tačnosti klasifikacionog modela, u okviru Weka softverskog okruženja za data mining.

II. NADZIRANO UČENJE

U procesu data mininga uobičajeno je eksperimentisanje sa raznim algoritmima mašinskog učenja. Tehnike nadziranog mašinskog učenja nalaze svoju primjenu u mnogim oblastima. Nadzirano mašinsko učenje je proces učenja skupa pravila iz primjera koji se nalaze u skupu podataka za obuku (trening) stvarajući kao rezultat procesa, klasifikator koji može biti korišten da generalizuje iz novih instanci. U nadziranom mašinskom učenju, ulaz za algoritam je obično predstavljen u formi seta instanci, svaka instanca je predstavljena određenim brojem atributa za obučavanje i oznakom klase. Zadatak algoritma je da nauči kako da novoj, neobilježenoj instanci podataka, dodijeli tačnu oznaku klase (ako je vrijednost nominalna radi se o klasifikaciji, a ako je numerička radi se o regresiji).

Sam proces praktične realizacije, prema [8], nadziranog mašinskog učenja, na realne probleme koji nas okružuju, dat je na Sl. 1. Prvi korak je prikupljanje potrebnih podataka, pri čemu izbor određenih atributa, odnosno atributa koji su najviše informativni, može biti sugerisan od strane adekvatnog stručnjaka, pod uslovom da je na raspolaganju (u suštini najbolji je ručni odabir). U startu se obično kreće od dostupnih podataka (često dostupnih iz različitih izvora), koji u sebi sadrže šum (neočekivane vrijednosti), nedostajuće varijable (nedostatak vrijednosti nekog atributa) i sl., tako da realizacija

postupaka predobrade podataka postaje neminovnost. U fazi pripreme i predobrade podataka, zavisno od okolnosti, postoje načini za rješavanje pomenutih problema, postoje načini za rješavanje nedostajućih podataka, tehnike za otkrivanje šuma, pri čemu svaki od načina/tehnika nose sa sobom određene prednosti i nedostatke. Jedan od mogućih postupaka procesa predobrade podataka je i proces selekcije atributa (koji je u fokusu i u osnovi eksperimentalne postavke ovoga rada) kao koristan proces za otkrivanje i uklanjanja irelevantnih i reduntantnih atributa u što većoj mjeri. Nakon što je odabran set podataka (definisan skup podataka za obuku) dolazi na red ključan korak izbora specifičnih algoritama za proces učenja (obučavanja) od dostupnih različitih klasifikacionih tehnika nadziranog mašinskog učenja. Sam kvalitet dobijenog klasifikatora se potvrđuje rezultatima nastalim testiranjem klasifikatora na neoznačenim instancama (instance kojima su poznate vrijednosti prediktorskih atributa ali je nepoznata vrijednost oznake klase).

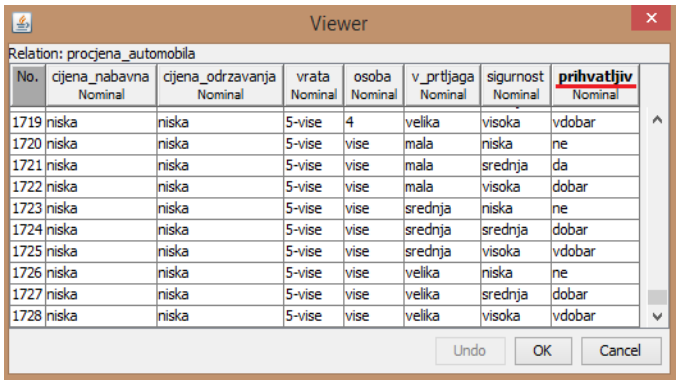


Slika 1. Proces nadziranog mašinskog učenja

Evaluacija klasifikatora najčešće se temelji na predviđanju tačnosti (procenat tačnih predviđanja podijeljen sa ukupnim brojem predviđanja). Uobičajene tehnike na osnovu kojih se računa klasifikaciona tačnost su *holdout* (u kome se vrši podjela raspoloživog skupa za trening u određenim odnosima na skup za trening, testiranje i validaciju), podjela skupa u određenim procentima (*percentage split*), i metoda unakrsne validacije (dijeli skup na N dijelova (*fold*)). Problem nastaje zato što u praksi često postoji samo jedan skup podataka određene veličini i sve procjene moraju biti dobiveni na osnovu tog skupa što neminovno dovodi do kolizije [9], obzirom da je istovremeno potreban što veći skup za obučavanje (da bi se dobio dobar klasifikator) i adekvatan skup za testiranje (za dobru procjenu o grešci) a sve na osnovu jednog početnog skupa podataka. Iz navedenog dovoljno je naglašen značaj obima i kvaliteta ulaznih podataka u kontekstu rezultata data mininga.

A. Ulazni skup podataka

Izbor ulaznog skupa podataka u uskoj je vezi i sa klasom problema, te je na samom početku korisno uzeti u obzir ulogu podataka sa kojima se raspolaže i prepoznati vrstu problema, što u krajnjem utiče i na izbor algoritma mašinskog učenja.



No.	cijena_nabavna Nominal	cijena_odrzavanja Nominal	vrata Nominal	osoba Nominal	v_prtljaga Nominal	sigurnost Nominal	prihvatljiv Nominal
1719	niska	niska	5-vise	4	velika	visoka	vdobar
1720	niska	niska	5-vise	vise	mala	niska	ne
1721	niska	niska	5-vise	vise	mala	srednja	da
1722	niska	niska	5-vise	vise	mala	visoka	dobar
1723	niska	niska	5-vise	vise	srednja	niska	ne
1724	niska	niska	5-vise	vise	srednja	srednja	dobar
1725	niska	niska	5-vise	vise	srednja	visoka	vdobar
1726	niska	niska	5-vise	vise	velika	niska	ne
1727	niska	niska	5-vise	vise	velika	srednja	dobar
1728	niska	niska	5-vise	vise	velika	visoka	vdobar

Slika 2. Dio seta podataka za procjenu automobila (zadnjih deset instanci)

U ovom radu korišten je javno dostupan (UCI Machine Learning Repository) set podataka za procjenu prihvatljivosti automobila (Car evaluation) koji ima 1728 instanci i 7 atributa (uključujući klasu (class)), čijih kranjih deset instanci, zajedno sa nazivima atributa, je prikazano na Sl.2, pri čemu je vidljivo da su svi nazivi atributa instanci i nominalne vrijednosti semantički prilagođeni adekvatnim terminima našeg jezika (radi lakše čitljivosti). Kao što se može vidjeti na Sl. 2, svakoj instanci raspoloživog seta podataka za procjenu automobila je pridružena klasa "prihvatljiv" (na Sl.2 se mogu vidjeti sve nominalne vrijednosti klase prihvatljiv – ne, da, dobar, vdobar) kojoj ta instanca pripada, stoga se može reći da ovaj ulazni skup podataka pripada grupi klasifikacionih problema, koji spadaju u domen nadziranog učenja.

III. SELEKCIJA ATRIBUTA

Iz prethodno navedenog može se zaključiti da neadekvatnost, loš kvalitet ulaznog skupa podataka na određen način utiču na preciznost postignutih rezultata mašinskog učenja. Takve situacije mogu da nastanu na primjer u slučajevima baza sa velikim brojem podataka, koje međusobno ne moraju biti povezane sa usko specifičnim oblastima. Stoga je prije korištenja podataka (zasnovanih na pomenutim izvorima podataka na primjer), sa ciljem veće efikasnosti i preciznosti primjenjenog algoritma, potrebno na određeni način - korištenjem metode selekcije atributa - eliminisati sve irelevantne atribute i naći najmanji podskup atributa koji zadovoljava određenu klasifikacionu tačnost (bolju od one na originalnom setu ili što bližu originalnoj raspodjeli (kada se koriste svi atributi)). Proces selekcije podskupova atributa u određenoj mjeri smanjuje dimenzionalnost podataka i omogućava da algoritmi data mininga rade brže (za ukupno računanje vremena potrebno je

uzeti u obzir i izračunavanja koja se odnose na izbor atributa) i djelotvornije. Kako je već pomenuto u kontekstu problema klasifikacije, sam cilj odabira atributa je u funkciji povećanja klasifikacione tačnosti

Broj podskupova atributa raste eksponencijalno sa povećanjem broja atributa kao i broja klasa, tako da za n atributa postoji 2^n mogućih podskupova atributa. U praksi pronalaženje optimalnog podskupa atributa je obično složen posao i mnogi zadaci vezani za selekciju atributa spadaju u klasu NP-teških (NP-hard) problema.

Selekcija atributa se vrši na osnovu heurističke procjene, a tehnike za selekciju atributa mogu se podijeliti na osnovu toga da li se analizom karakteristika ulaznog skupa vrednuje podskup atributa (*subset selection*) ili vrijednosti atributa pojedinačno (*feature ranking*). U prvom pristupu, (korišten u okviru eksperimentalnih postavki), pretražuje se prostor podskupova sa ciljem pronalaženja optimalnog podskupa, dok se u drugom slučaju vrši rangiranje atributa [11] na osnovu određenih kriterija i biraju oni koji prelaze definisani prag vrijednosti.

Uobičajene medode selekcije atributa su:

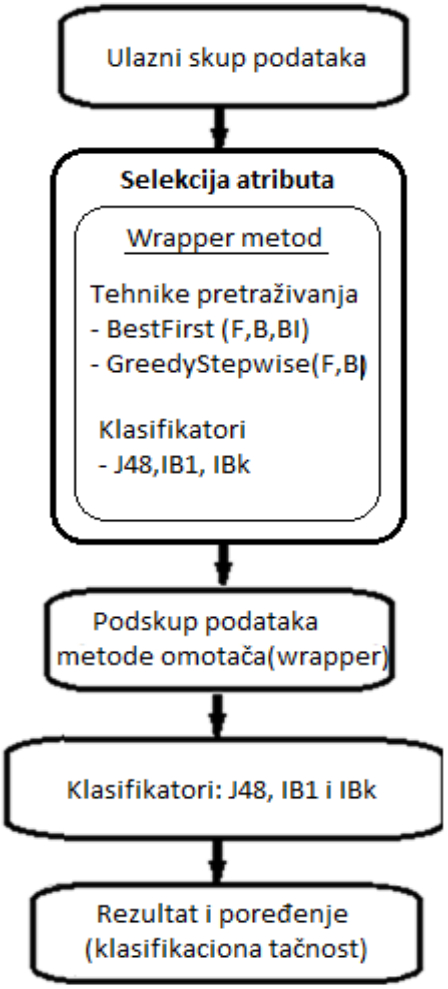
- Filter metoda – Filter pristupom prvo se selektuju atributi, nakon toga se koristi izabrani podskup za izvršenje klasifikacionog algoritma (izbor atributa nezavisan je od klasifikatora);
- Ugrađene (*embedded*) metode – Ove metode vrše odabir atributa za vrijeme učenja optimalnih parametara (slučaj neuronske mreže);
- Metoda omotača (*wrapper*) - poznat i kao metod prethodnog učenja, je metod koji uključuje algoritam mašinskog učenja u proces selekcije atributa (uzima u obzir zavisnost atributa podskupa od algoritama učenja).

A. Metoda omotača

U osnovi wrapper pristupa [11] algoritam za selekciju podskupa atributa, selekciona procedura, postoji kao omotač oko algoritma mašinskog učenja. Algoritam za selekciju podskupa atributa vrši pretraživanje dobrog podskupa koristeći sam algoritam mašinskog učenja, kao dio funkcije evaluacije podskupa atributa. Ideja koja stoji iza pristupa metode omotača [12], u biti je jednostavna: algoritam mašinskog učenja se posmatra kao crna kutija (*black box*). Algoritam učenja se izvršava na setu podataka, obično podjeljen u interne trening i holdout setove, sa različitim skupovima atributa uklonjenim iz podataka. Podskup podataka koji je najbolje procjenjen se bira kao konačan set na kome će se izvoditi algoritam učenja. Nastali klasifikator se tada evaluira na nezavisnom testnom setu podataka, koji nisu prethodno korišteni.

IV. EKSPERIMENTALNE POSTAVKE I REZULTATI

Kako bi se pokazao uticaj wrapper metode, bazni eksperimentalni koraci prikazani su na Sl. 3, korištenjem pomenutog ulaznog skupa podataka za procjenu automobila, korišteno je Weka okruženje [13], koje u osnovi predstavlja kolekciju algoritama mašinskog učenja za zadatke data mininga (uz neophodne alate za predobradu, klasifikaciju, regresiju, klastering i ...).



Slika 3. Eksperimentalni koraci

Generalno posmatrano realizacija eksperimenta se odvija u dva dijela. U prvom dijelu su korištene različite metode sa ciljem selekcije najviše relevantnih atributa, a drugi dio se odnosi na primjenu – koristeći prethodno odabran podskup atributa – različitih algoritama mašinskog učenja, sa ciljem postizanja što bolje klasifikacione tačnosti za predviđanje (obzirom na ulazni skup, za predviđanje prihvatljivosti automobila). Kako bi se eliminisao negativan uticaj irelevantnih atributa na performanse klasifikacije, i u krajnjem postigla veća klasifikaciona tačnost korišten je wrapper pristup - izvršena je redukcija obima ulaznog seta podataka prije nego se proslijedi navedenim klasifikatorima. Pitanja koja se nameću u praksi [14] je kako pronaći prostor svih mogućih podskupova atributa, kako evaluirati predikcione performanse algoritma mašinskog učenja kako bi se upravljalo pretraživanjem i kako (kada) bi se isto zaustavilo, te koje prediktore koristiti. U principu mogla bi se izvršiti jedna iscrpna, sveobuhvatna pretraga u slučajevima kada broj atributa nije previše velik(obzirom da za n atributa postoji 2n mogućih podskupova).

TABELA I. REZULTATI WRAPPER SELEKCIJE ATRIBUTA ZA KORIŠTENE ALGORITME KLASIFIKACIJE I TEHNIKE PRETRAŽIVANJA

Algoritam	Metoda pretrazivanja -Smjer	Ispravnost klasifikacije (Merit)		Broj evaluiranih podskupova 5-fold / 10-fold	Broj selektovanih atributa
		5-fold	10-fold		
J48	BestFirst -F	0.919	0.929	36	5
	BestFirst -B	0.919	0.929	30	5
	BestFirst -BI	0.919	0.929	77	5
	GreedyStepwise-F	0.919	0.929	-	5
	GreedyStepwise-B	0.919	0.929	-	5
IB1	BestFirst -F	0.932	0.934	22	5
	BestFirst -B	0.932	0.934	29	5
	BestFirst -BI	0.932	0.934	60	5
	GreedyStepwise-F	0.932	0.934	-	5
	GreedyStepwise-B	0.932	0.934	-	5
IBk	BestFirst -F	0.933	0.939	36 /38	5
	BestFirst –B	0.933	0.939	28	5
	BestFirst -BI	0.933	0.939	77	5
	GreedyStepwise-F	0.933	0.939	-	5
	GreedyStepwise-B	0.933	0.939	-	5

TABELA II. REZULTATI KLASIFIKACIJE NA ORIGINALNOM I REDUKOVANOM SETU PODATAKA (KORIŠTENJEM WRAPPER PRISTUPA)

	J48	J48-W	IB1	IB1-W	IBk	IBk-W
Klasifikaciona tačnost (%)	92.36	93.22	77.25	93.34	93.51	94.21

Osnovni elementi korištenog wrapper pristup dati su u Tabeli I, iz koje se može vidjeti da su u okviru eksperimentalnih postavki kao algoritmi mašinskog učenja odabrani stablo odlučivanja J48 (baziran na poznatom C4.5 algoritmu koji generiše stablo odlučivanja) i algoritmi najbližih susjeda IB1 (prosti algoritam najbližeg susjeda) i IBk (algoritam k najbližeg susjeda), u kombinaciji sa strategijama za pretraživanje prostora podskupova. Za pretraživanje podskupova korištena je pohlepna (GreedyStepwise) i metoda najbolji prvi (BestFirst), sa različitim opcijama smjerova pretraživanja (unaprijed F(forward), unazad B(backward) i u oba smjera BI (Bidirectional)), uz korištenje opcija 5 fold i 10 fold unakrsne validacije za evaluaciju performansi. Dakle, korištenjem wrapper pristupa navedeni klasifikacioni algoritmi J48, IB1 i IBk, su primjenjeni nad svakim kandidatom podskupa atributa, nakon čega je evaluiran podskup atributa (uz Weka default vrijednosti: prag vrijednosti *threshold* i kriterijum zaustavljanja *searchTermination*).

Na osnovu postignutih rezultata (Tabela I) korištenjem wrapper pristupa, u okviru datih eksperimentalnih postavki, može se zaključiti da:

- postignuti rezultati ispravnosti klasifikacije, za određeni klasifikator, nisu zavisili od korištene strategije pretraživanja (isti rezultati su postignuti i za medodu najbolji prvi i za pohlepnu metodu),
- rezultati postignuti za ispravnost klasifikacije, za određeni klasifikator, nisu zavisili od od smjera (*direction*) korištene strategije pretraživanja (isti rezultati su postignuti za sve korištene smjerove metode najbolji prvi i pohlepne metode),
- ispravnost klasifikacije, za sve korištene klasifikatore, je zavisna od korištene opcije unakrsne evaluacije (bolji rezultati su postignuti za 10-fold unakrsnu validaciju),
- broj evaluiranih podskupova je različit zavisno od korištenog klasifikatora i strategije pretraživanja, pri čemu je isti, osim u jednom slučaju (IBk sa BestFirst-F), za obe korištene opcije unakrsne validacije,
- najbolje rezultate, ispravnost klasifikacije postigao je IBk, algoritam korištenjem 10-fold unakrsne validacije,
- podskup selektovanih atributa bez obzira na korištene klasifikatore i strategije pretraživanja je isti (5 selektovanih atributa (isključen je atribut “vrata”)).

U drugom dijelu realizacije eksperimenta primjenjeni su klasifikatori J48, IB1 i IBk nad redukovanim skupom podataka, nad podskupom od 5 selektovanih atributa, dobijenih wrapper metodom. Postignuti rezultati klasifikacije nad originalnom i redukovanom setu podataka nastalom korištenjem wrapper pristupa, dati su u Tabeli II. Na osnovu uvida u rezultate date Tabelom II, može se zaključiti da su svi klasifikatori, u okviru datih eksperimentalnih postavki, postigli veću klasifikacionu tačnost nad redukovanom skupu podataka u odnosu na originalni, ulazni skup podataka, pri

čemu je najveći benefit, od korištenog wrapper pristupa selekcije atributa na performanse klasifikacionog modela postigao IB1, odnosno IB1-W (77.25% klasifikacione tačnosti IB1 na originalnom i čak 93.34% na redukovanom skupu korištenjem wrapper pristupa (IB1-W)), dok je u cjelini posmatrano najbolji rezultat postigao algoritam IBk sa 94.21% klasifikacione tačnosti.

ZAKLJUČAK

U radu je korištenjem navedenih eksperimentalnih faza i navedenih opcija za iste, uz neophodan teorijski okvir, prikazan način i diskutovani rezultati eksperimenata u kojima je primjenjen metod selekcije atributa korištenjem metode omotača na atributima instanci seta podataka za procjenu automobila.

Postignuti rezultati korištenja wrapper metode, nad navedenim skupom podataka procjena automobila ukazuju na pozitivan uticaj wrapper pristupa selekcije atributa na performanse klasifikacije, povećanjem klasifikacione tačnosti svih korištenih klasifikatora, a sam način realizacije se potencijalno može iskoristiti za testiranje uticaja wrapper pristupa i u kontekstu drugih setova ulaznih podataka.

LITERATURA

[1] M.R. Berthold, C.Borgelt, F.Höppner, F. Klawonn, Guide to Intelligent Data Analysis - How to Intelligently Make Sense of Real Data, Springer, 2010.

[2] D. Pyle, Data Preparation for Data Mining, Morgan Kaufmann Publishers, Inc, 1999.

[3] S. García, J.Luengo, F. Herrera, Data Preprocessing in Data Mining, Springer, 2015.

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

[5] M. Dash, H. Liu, "Feature Selection for Classification", Intelligent Data Analysis, Vol. 1, pp. 131–156, 1997.

[6] A.Blum, P. Langley, "Selection of relevant features and examples in machine learning", Artificial Intelligence, 97, 245-271, 1997.

[7] D. Koller, M. Sahami, "Toward Optimal Feature Selection", Proceedings of the Thirteenth International Conference on Machine Learning (ICML), pp. 284-292, 1996.

[8] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques", Informatica 31, pp. 249-268, 2007.

[9] O. Janković, "Mašinsko učenje: Evaluacija performansi klasifikatora u kontekstu ograničene raspoloživosti podacima", XLI Simpozijum o operacionim istraživanjima – SYM-OP-IS 2014., Zbornik radova, str. 232-237, 2014.

[10] O. Janković, Data Mining: Implikacije filterovanja rangiranjem na performanse klasifikacionog modela, XXII naučna i biznis konferencijaYU INFO 2016, prihvaćen za objavljivanje

[11] G. John, R. Kohavi and K. Pfleger, Irrelevant features and the subset selection problem, Fifth International Conference on Machine Learning, New Brunswick, NJ (Morgan Kaufmann, Los Altos, CA, 1994) 121-129.

[12] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artificial Intelligence 97, pp.273-324, 1997.

[13] A. Bifet, G. Holmes, R. Kirkby, B. Pfahringer, DATA STREAM MINING A Practical Approach, The University of Waikato, COSI, 2011

[14] B. Jantawan, C. Tsai, "A Comparison of Filter and Wrapper Approaches with Data Mining Techniques for Categorical Variables Selection", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue 6, 2014.

ABSTRACT

In order to increase the overall accuracy of the classification model, eliminating the negative impact irrelevant attributes on the performance classification, in this paper will be shown a way to reduce a data volume by selection of attributes, using the wrapper method. Experimental setting, the wrapper approach for modeling basically involves three different algorithms and greedy and the method best first for subsets searching. Influence of wrapper approach is provided by comparing the classification accuracy, J48 decision tree classifier and algorithms nearest neighbors: IB1 and IBk, on the original and a reduced set of data (resulting from application wrapper methods), using Weka data mining tool.

DATA MINING: IMPLICATION WRAPPER APPROACH ON THE SELECTION OF ATTRIBUTES ON THE PERFORMANCE OF CLASSIFICATION MODEL

Olivera Janković