Poboljšanje tačnosti klasifikacije algoritama za induktivno učenje pravila primenom metoda prethodnog učenja

JASMINA Đ. NOVAKOVIĆ, Beogradska poslovna škola,
Visoka škola strukovnih studija, Beograd

Stručni rad
UDC: 159.953.5

U ovom radu istražujemo problem tačnosti klasifikacije algoritama mašinskog učenja primenom metoda prethodnog učenja. Za potrebe klasifikacije koriste se sledeći algoritmi: IBk, Naïve Bayes, SVM, J48 stablo odlučivanja i RBF mreža. Eksperimentalni rezultati pokazuju da se metodama prethodnog učenja mogu brzo identifikovati nevažni, redundantni atributi, kao i šum u podacima ako on postoji; kao i oni atributi koji su značajni za izučavanu pojavu. U radu se dokazuje da primenom metoda prethodnog učenja za redukciju dimenzionalnosti podataka je moguće znatno poboljšati performanse sistema za induktivno učenje pravila u problemima klasifikacije.

Ključne reči: *klasifikator, metode prethodnog učenja, selekcija atributa, tačnost klasifikacije, unakrsna validacija*

1. UVOD

Klasifikacija je jedan od najčešćih zadatka mašiskog učenja, i predstavlja problem razvrstavanja nepoznate instance u jednu od unapred ponuđenih kategorija - klasa. Važno zapažanje kod klasifikacije je da je ciljna funkcija u ovom problemu diskretna. U opštem slučaju, oznakama klasa se ne mogu smisleno dodeliti numeričke vrednosti niti uređenje. To znači da je atribut klase, čiju je vrednost potrebno odrediti, kategorički atribut.

Klasifikacija nekog objekta se zasniva na pronalaženju sličnosti sa unapred određenim objektima koji su pripadnici različitih klasa, pri čemu se sličnost dva objekta određuje analizom njihovih karakteristika. Pri klasifikaciji se svaki objekat svrstava u neku od klasa sa određenom tačnošću. Zadatak je da se na osnovu karakteristika objekata čija klasifikacija je unapred poznata, napravi model na osnovu koga će se vršiti klasifikacija novih objekata. U problemu klasifikacija, broj klasa je unapred poznat i ograničen.

U ovom radu korišćeni su sledeći algoritmi nadziranog učenja za izgradnju modela, a to je IBk, Naïve Bayes, SVM, J48 stablo odlučivanja i RBF mreža. Prednost IBk je da su oni u mogućnosti da uče brzo sa vrlo malim skupom podataka. Prednost Naïve Bayes klasifikatora je da zahteva malu količinu trening podataka za procenu parametara potrebnih za klasifikovanje. Prednost SVM nad drugim metodama je pružanje boljih predviđanja neviđenih test podataka, pružanje jedinstvenih optimalnih rešenja za problem u treniranju i postojanje manje parametara za optimizaciju u poređenju sa drugim metodama. J48 stablo odlučivanja ima razne prednosti: jednostavan za razumevanje i interpretaciju, zahteva malu pripremu podataka, robustan je, dobro radi i sa velikim brojem podataka u kratkom vremenu.

RBF mreže nude niz prednosti, uključujući i zahtevanje manje formalnih statističkih treninga, sposobnost da se implicitno detektuju složeni nelinearni odnosi između zavisnih i nezavisnih varijabli, sposobnost detektovanja svih mogućih interakcija između prediktorskih varijabli i dostupnost više algoritama za trening.

2. METODE PRETHODNOG UČENJA

Kod metoda prethodnog učenja koriste se određeni algoritmi za modeliranje kako bi se ocenili podskupovi atributa u odnosu na njihovu klasifikacijsku ili prediktivnu moć. Kod korišćenja ovih metoda u praksi se pojavljuju tri pitanja:

- kako pretražiti prostor svih mogućih podskupova atributa,
- kako proceniti uspešnost algoritma za modeliranje s obzirom na pretraživanje skupa atributa,
- koji postupak modeliranja koristiti kao crnu kutiju za metode prethodnog učenja.

Adresa autora: Jasmina Novaković, Beogradska poslovna škola, Visoka škola strukovnih studija, Beograd, Kraljice Marije 73

Rad primljen: 26.01.2015.
Rad prihvaćen: 02.02.2015.

Kod metoda prethodnog učenja vrednost određenog skupa atributa izražava se pomoću stepena ispravnosti klasifikacije koju postiže model konstruisan uz korišćenje tih atributa. To znači da su ove metode tesno vezane za odabrani algoritam mašinskog učenja. Za zadati podskup atributa, ispravnost klasifikacije se ocenjuje korišćenjem tehnika uzorkovanja, na primer unakrsnom validacijom (eng. cross-validation). Kod ovih metoda za svaki posmatrani podskup atributa izgrađuje se model i ocenjuju se njegove performanse, tako da bolje performanse nekog modela ukazuju na bolji izbor atributa iz kojih je model nastao.

Kod metoda prethodnog učenja postupak izbora atributa je računski vrlo zahtevan zbog učestalog izvođenja algoritma mašinskog učenja. Potrebno je dobiti ocenu performansi odgovarajućeg modela za svaki posmatrani podskup atributa, a metode ocene ispravnosti modela uglavnom zahtevaju usrednjavanje rezultata po većem broju izgrađenih modela. Kod ovih metoda za svaki posmatrani podskup atributa izgrađuje se više modela, a ukupan broj podskupova eksponencijalno raste s povećanjem broja atributa.

Iscrpno pretraživanje podskupova atributa se može sprovesti samo za mali broj atributa, budući da je taj problem NP-težak. Zato se koriste razne tehnike pretraživanja, kao što su: najbolji prvi (eng. best-first), granaj-pa-ograniči (eng. branch-and-bound), simulirano kaljenje (eng. simulated annealing), genetski algoritmi i sl. [1]. U praksi se pokazuje da pohlepne tehnike pretraživanja daju dobre rezultate, što znači da se nikad ne proveravaju već donesene odluke o tome da li da se atribut uključi (ili isključi) iz skupa.

Pohlepne tehnike prostor rešenja prelaze tako da u svakom koraku pregledaju lokalno dostupne alternative, pa proces pretraživanja nastavljaju od najbolje od njih (tehnika uspona na vrh). Pohlepne tehnike se dele na izbor atributa unapred (eng. forward selection) i eliminacija atributa unatrag (eng. backward elimination) [1]. Moguće je i pretraživanje u oba smera (eng. Bi-directional search).

Kod izbora atributa unapred postupak počinje sa praznim skupom atributa i u svakom koraku postupka dodaje se po jedan novi atribut. Eliminacija atributa unatrag je obrnuti postupak koji počinje sa punim skupom atributa i u svakom koraku se oduzima po jedan atribut. Opisani postupci su vrlo jednostavni, ali daju rezultate uporedive sa složenijim tehnikama pohlepnog pretraživanja kao što su zrakasto pretraživanje ili metoda najboljeg prvog.

Ako sa n označimo ukupan broj atributa, izbor atributa unapred i eliminacija atributa unatrag imaju složenost $O(n^2)$, i s obzirom da proizvode prihvatljive rezultate u razumnom vremenu, upravo ove dve tehnike pretraživanja se najčešće koriste u izboru atributa metodama prethodnog učenja.

Generalno, eliminacija atributa unatrag preferira veće podskupove atributa i može rezultirati nešto boljim klasifikacijskim performansama od odabira atributa unapred. S obzirom da se vrednost skupa atributa meri procenom ispravnosti klasifikacije, onda se zbog samo jedne optimistične procene oba postupka mogu preuranjeno završiti, i u tom slučaju eliminacija atributa unatrag će odabrati previše atributa, a odabir atributa unapred premalo. Usled nedostatka prognostičkih atributa može se ograničiti sposobnost zaključivanja, što će odraziti na nešto slabije klasifikacijske performanse. Manji broj izabranih atributa je poželjan u slučajevima kada je primarni cilj razumevanje međuzavisnosti i pravilnosti u podacima, jer su konstruisani modeli jednostavniji i naglašavaju najprediktivnije atribute.

Kod metoda prethodnog učenja najvažniji nedostatak je sporost pri izvođenju uslovljena pozivanjem ciljnog algoritma mašinskog učenja više puta, zbog čega ovim metodama ne odgovaraju obimni skupovi podataka za učenje sa većim brojem atributa.

3. OPIS IZABRANIH PROBLEMA UČENJA

Za potrebe eksperimentalnog istraživanja koristili smo 15 realnih skupova podataka i 3 veštačka, preuzeta iz UCI repozitorijuma [2], koji je namenjen istraživačima koji pročavaju probleme veštačke inteligencije.

Rak dojke (breast cancer – bc): zadatak ovog seta podataka je da predvidi da li ima ili nema povratka bolesti raka dojke kod pacijenata. Predviđanje se radi na osnovu godina, nastupanja menopauze, veličine tumora, veličine čvorova, stepena maligniteta, zahvaćene dojke tumorom, položaja tumora da li je vršeno zračenje ili ne kod pacijenta.

Odobravanje kredita (credit approval - ca): ovaj set sadrži podatke koji se odnose na korišćenje kreditne kartice [3, 4]. Ovaj set podataka je interesantan za istraživanje jer postoji dobra mešavina atributa – kategoričkih i numeričkih vrednosti.

Kreditni podaci (Statlog german credit data - cg): ovaj skup podataka omogućava klasifikovanje potencijalnih korisnika kredita na one koji imaju mali ili visok rizik za odobravanje kredita.

Ultrazvuk (cardiography – ct): ovaj skup podataka sastoji se od atributa merenja fetalnog otkucaja srca i atributa kontrakcije materice na ultrazvuku koje su klasifikovali doktori [5]. Klasifikacija je urađena u odnosu na morfološke obrasce i na stanje fetusa.

Hepatitis (hepatitis – he): glavni cilj ovog skupa podataka je predvideti hoće li sa hepatitisom pacijenti

umreti ili ne. U ovom skupu podataka, postoje dve klase za predviđanje: prva klasa koja predviđa da će pacijent preživeti i druga klasa koja predviđa da će pacijent umreti.

Jetra (liver disorders - li): u skupu podataka pod nazivom jetra, prvih pet atributa su testovi krvi pacijenata; dok se druga dva atributa odnose na broj popijenih alkoholnih pića, i da li je pacijent svakodnevno pijan. Smatra se da ovi atributi ukazuju na bolesti jetre, koja bi mogla proizaći između ostalog i iz preteranog konzumiranja alkohola.

Rak pluća (lung cancer –lc): set podataka za rak pluća sadrži podatke koji opisuju tri vrste patološkog oblika raka pluća. Ove podatke su prvo koristili istraživači Hong i Young za ilustraciju dobrih performansi optimalno diskriminativnih ravni, čak i u loše datim postavkama [6].

Mamografska masa (mammographic mass - ma): zadatak ovog skupa podataka je da predvidi ozbiljnost (benigni ili maligni) mamografskih lezija na osnovu BI-RADS atributa i starosti pacijenta [7]. U skupu podataka, svaka instanca je povezana sa BI-RADS procenom koja se kreće u u rasponu od 1 (definitivno benigni) do 5 (vrlo sugestivna malignost) koja je dodeljena na osnovu procene dva radiologa.

MONK problemi: ovi problemi pripadaju klasi veštačkih (sintetičkih) domena, pri čemu svaki od tri problema koristi istu reprezentaciju podataka za upoređenje algoritama mašinskog učenja. Ovaj skup podataka ima 432 instance i ima 7 atributa. Postoje tri *Monk* problema: Monk1 (m1), Monk2 (m2) i Monk3 (m3).

Gljive (mushroom – mu): ovaj skup podataka uključuje opise hipotetičkih uzoraka koji odgovaraju 23 vrsti gljiva Agaricus i Lepiota familiji [8]. Svaka vrsta je identifikovana kao definitivno jestiva, definitivno otrovna ili nepoznatog jestivog sastava i ne preporučuje se za jelo.

Parkinson (Parkinson – pa): ovaj set podataka se sastoji od niza biomedicinskih merenja glasa kod 31 osobe, od toga 23 obolele od Parkinsonove bolesti [9]. Glavni cilj ovog skupa podataka je odvajanje zdravih ljudi od onih osoba koje su obolele od Parkinsa, na osnovu kolone „Status“ koja ima moguće vrednosti 0 za zdrave osobe i 1 za osobe sa Parkinsonovom bolešću.

Dijabetes (Pima Indijans dijabetes – pi): radi dijagnostifikovanja dijabetesa iz većeg skupa podataka izdvojeni su podaci za žene koje su starije od 21 godinu i pripadaju Pima Indijancima [10]. U ovom setu podataka dijagnostifikovano je da li pacijent pokazuje znakove dijabetesa prema kriterijima Svetske zdravstvene organizacije.

Segmentacija slike (image segmentation – se): slučajevi su izvučeni slučajnim izborom iz baze podataka 7 slika spoljnog okruženja [11]. Slike su ručno segmentirane kako bi se izvršila klasifikacija za svaki piksel. Svaka instanca u skupu podataka je 3x3 regija. Klasa ovog skupa podataka ima moguće vrednosti: površina cigle, nebo, lišće, cement, prozor, put i trava.

Soja (soybean – so): zadatak je dijagnostifikovati bolesti u biljkama soje [12]. Vrednost atributa je merena posmatranjem svojstava lišća i različitih biljnih abnormalnosti.

Srce (Statlog heart - sh): zadatak je predvideti odsutnosti ili prisutnosti bolesti srca na osnovu starosti, pola, odgovarajućeg tipa bola u grudima, krvnog pritiska u mirovanju, nivoa holesterola i šećera u krvi, elektrokardiografskih rezultata, najvećeg broja otkucaja srca, promene pokazatelja prilikom napora i slično. Klasa za ovaj set podataka ima dve vrednosti: odsustvo i prisustvo bolesti srca.

Glasanje kongresmena (congressional voting records – vo): u ovom setu podataka stranačku pripadnost američkog Predstavničkog doma karakteriše kako su kongresmeni glasali na 16 ključnih pitanja kao što su trošenje na obrazovanje i imigracija [8].

U tabeli 1. prikazane su uporedne karakteristike posmatranih setova podataka. Da bi dobili referentne podatke tokom istraživanja u radu smo koristili i realne i veštačke skupove podataka za dokazivanje hipoteze. Možemo zaključiti da se u posmatranim skupovima podataka nalaze i skupovi sa izuzetno velikim brojem atributa, kao i oni skupovi koji imaju mali broj atributa, što je dobro sa stanovišta istraživanja. Posmatrani skupovi podataka su balansirani jer postoje skupovi koji sadrže samo ili kategoričke ili numeričke atribute, kao i skupovi podataka koji sadrže i kategoričke i numeričke podatke.

Što se tiče broja klasa u posmatranim skupovima podataka, samo dva skupa podataka imaju veći broj klasa od 3, i to *se* koji ima 7 klasa i *so* koji ima 19 klasa. Razlog za ovo je činjenica, što se u najvećem broju slučaja u problemima klasifikacije razvrstavanje postojećih instanci vrši u dve, eventualno tri klase, a ređe u veći broj klasa.

U tabeli 1. vidimo da broj instanci predviđen za treniranje varira od malog broja prikupljenih istanci što je slučaj sa *lc* koji ima samo 32 istance do skupova koji imaju mnogo veći broj istanci kao što je npr. slučaj sa *mu* koji ima 8124 istanci za trening. Što se tiče veličine skupa za testiranje, inicijalno kod svih realnih skupova podataka, imali smo pripremljen jedan skup podataka, iz koga smo metodom 10-struke unakrsne validacije izdvajali podatke koji će služiti za testiranje. Istraživači koji su kreirali veštačke skupove podataka *m1*, *m2* i *m3* su odvojili podatke u dve grupe i to one koji će služiti za treniranje i one koji će služiti za testiranje, pri čemu je manji broj podataka korišćen za trening (u

proseku oko 25%), a veći deo služi za testiranje tačnosti klasifikacije. U poslednjoj koloni tabele prikazana je referentna tačnost za realne i veštačke skupove podataka.

4. ESTIMACIJA TAČNOSTI KLASIFIKACIJE

U četvrtom delu, nakon razmatranja postavki eksperimentalnog istaživanja, biće prikazani rezultati istraživanja za različite metode prethodnog učenja, i to za svaki klasifikacioni algoritam posebno.

Eksperiment je rađen uz pomoć WEKA (Waikato Environment for Knowledge Analysis), alata za pripremu i istraživanje podataka razvijen na Waikato Univerzitetu na Novom Zelandu. Ovaj alat poseduje podršku za ceo proces istraživanja počevši od pripreme podataka preko procene i korišćenja različitih algoritama.

Kod metoda prethodnog učenja koriste se određeni algoritmi za modeliranje kako bi se ocenili podskupovi atributa u odnosu na njihovu klasifikacijsku ili prediktivnu moć. Kod ovih metoda vrednost određenog skupa atributa izražava se pomoću stepena ispravnosti klasifikacije koju postiže model konstruisan uz korišćenje tih atributa. Za klasifikaciju, za sve skupove podataka, korišćena je 10-struka unakrsna validacija, koja je pri tome bila uvek ponovljena 10 puta. Upoređivana je tačnost klasifikacije IBk, Naïve Bayes, SVM, J48 i RBF mreže na originalnom skupu podataka kao i na redukovanom skupu podataka dobijenom sa metodom prethodnog učenja.

Kod ovih metoda za svaki posmatrani podskup atributa izgrađuje se model i ocenjuju se njegove performanse, tako da bolje performanse nekog modela ukazuju na bolji izbor atributa iz kojih je model nastao. Postupak izbora atributa je računski vrlo zahtevan zbog učestalog izvođenja algoritma mašinskog učenja.

Potrebno je dobiti ocenu performansi odgovarajućeg modela za svaki posmatrani podskup atributa, a metode ocene ispravnosti modela uglavnom zahtevaju usrednjavanje rezultata po većem broju izgrađenih modela. Kod ovih metoda za svaki posmatrani podskup atributa izgrađuje se više modela, a ukupan broj podskupova eksponencijalno raste s povećanjem broja atributa.

U ovom eksperimentalnom istraživanju metoda prethodnog učenja, kao metoda redukcije dimenzionalnosti podataka je koristila: različite klasifikatore za selekciju atributa, 5-struku unakrsnu validaciju i prag za ponavljanje unakrsne validacije ako standardna devijacija pređe ovu vrednost koji je podešen na 0.01.

Iscrpno pretraživanje podskupova atributa se može sprovesti samo za mali broj atributa, budući da je taj problem NP-težak. Zato se koriste razne tehnike pretraživanja, kao što su: najbolji prvi (eng. best-first), granaj-pa-ograniči (eng. branch-and-bound), simulirano kaljenje (eng. simulated annealing) i genetski algoritmi.

Kod metode prethodnog učenja, za pretraživanje prostora rešenja koristili smo heuristiku, kako bi ubrzali pretraživanje. Heuristika predstavlja iskustvena pravila o prirodi problema i osobinama cilja čija je svrha da se pretraživanje brže usmeri ka cilju. Heuristički ili usmereni postupak pretraživanja je onaj postupak pretraživanja koji koristi heuristiku kako bi suzio prostor pretraživanja. U ovom radu korišćen je heuristički postupak „pohlepnog najboljeg prvog“ (eng. greedy best-first), koji pretražuje podskup atributa koristeći algoritam uspona na vrh (eng. hill climbing). Postavljanje broja uzastopnih čvorova sa dozvoljenim ne-poboljšanjima kontroliše nivo praćenja unazad. Najbolji prvi može započeti sa praznim skupom atributa i pretraživati prema unapred, odnosno početi sa punim skupom atributa i pretraživati unazad, ili početi u bilo kojoj tački i pretraživati u oba smera (razmatranja svih mogućih pojedinačnih atributa za dodavanje ili brisanje u određenoj tački). U radu smo koristili smer pretraživanja unapred, što znači da smo započeli sa praznim skupom, a kao kriterijum za kraj pretraživanja postavili smo 5 uzastopnih čvorova sa dozvoljenim ne-poboljšanjima. Glavni razlog za izbor smera pretraživanja unapred je računski, jer je izgradnja klasifikatora sa nekoliko atributa mnogo brža nego kada ima više atributa. Iako u teoriji, pretraživanje unazad od punog skupa atributa, može lakše uhvatiti interakciju atributa, metoda je izuzetno računski skupa.

U eksperimentalnom istraživanju, kao i kod metoda filtriranja koristili smo uporedni t-test, gde je nivo značajnosti postavljen na vrednost 0.05. S obzirom da su postojali setovi podataka sa nedostajućim vrednostima, da bi mogli da koristimo SVM algoritam, bilo je neophodno zameniti nedostajuće vrednosti sa procenjenim vrednostima za dati skup, jer sam algoritam SVM nije mogao da se izbori sa nedostajućim vrednostima za pojedine atribute u nekim od istanci.

Slika 1. prikazuje broj atributa u originalnom skupu podataka i optimalan broj atributa dobijen metodama prethodnog učenja. Od 18 posmatranih setova podataka, u 15 setova podataka (svi osim li, ma i m2), tačno pola ili više od pola klasifikatora je smanjilo originalni broj atributa na pola. Najveću dobrobit od redukcije dimenzionalnosti podataka ima skup podataka lc, gde od 56 atributa, metodom prethodnog učenja smo izdvojili mali broj atributa relevantnih za posmatrani problem klasifikacije, čak isto ili manje od pet, za svaki od klasifikatora.

Koristeći metode prethodnog učenja za čak 7 skupova podataka, svi klasifikatori smanjuju broj atributa

na isto ili više od pola. Ti skupovi podatak su: bc, ct, lc, m3, mu, pa i vo. Možemo uočiti da su ove metode dovele do značajne redukcije dimenzionalnosti podataka.

U nastavku eksperimentalnog istraživanja, za izabrani optimalan broj atributa, za svaki skup podataka i klasifikator, proveravana je tačnost klasifikacije korišćenjem različitih algoritama, i to: IBk, Naïve Bayes, SVM, J48 i RBF mreže. U tabeli koja sledi za tačnost klasifikacije različitih klasifikatora su prikazane oznake „+“ i „-“, koje označavaju da je određeni rezultat statistički bolji (+) ili lošiji (-) od osnovnog klasifikatora na nivou značajnosti koji je specificiran na vrednost od 0,05.

Tabela 2. prikazuje tačnost klasifikacije različitih klasifikatora za originalni i redukovani skup podataka uz pomoć metoda prethodnog učenja. Možemo uočiti da u svim setovima podataka imamo dobijene rezultate za bar jednu od metoda prethodnog učenja koji su statistički bolji od osnovnog klasifikatora. Samo u dva seta podataka m2 i vo, imamo značajno lošije podatke za neku od metoda prethodnog učenja.

Metod prethodnog učenja sa IBk klasifikatorom je u više od pola skupova podataka (10 skupova) pokazao iste ili bolje rezultate od IBk algoritma na osnovnom skupu podataka, a u 5 skupova podataka rezultati su bili i statistički bolji. Metod prethodnog učenja sa Naïve Bayes klasifikatorom je u više od dve trećine skupova podataka (13 skupova) pokazao iste ili bolje rezultate od Naïve Bayes algoritma na osnovnom skupu podataka, a u 7 skupova podataka rezultati su bili i statistički bolji.

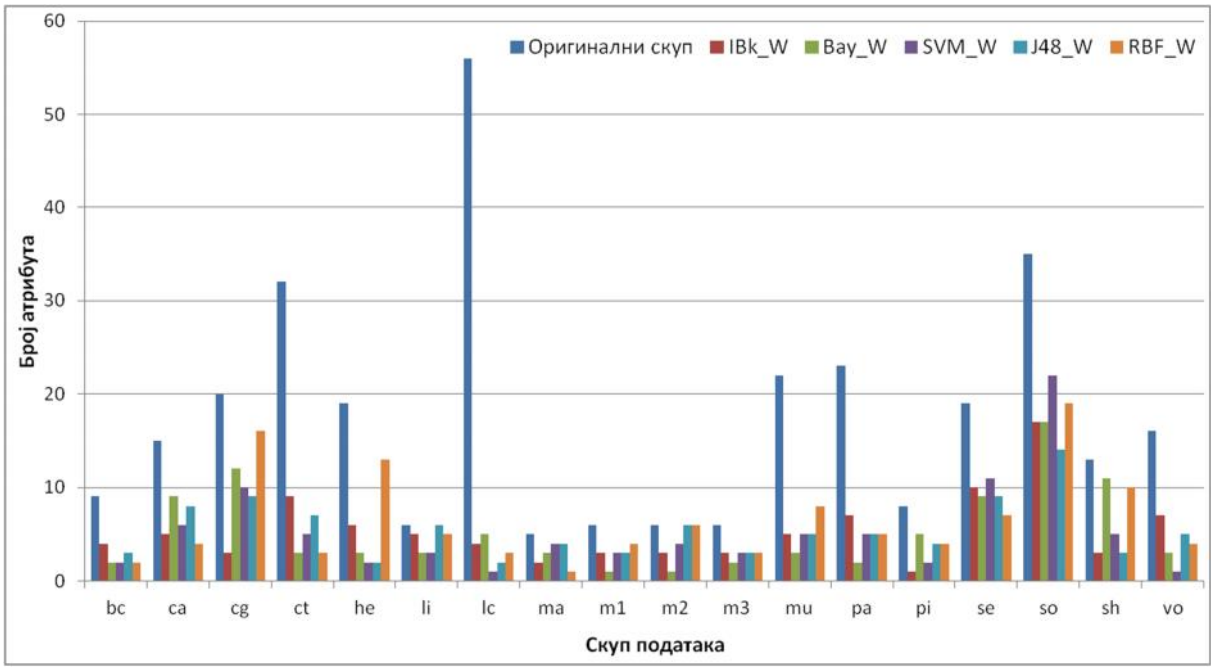
Metod prethodnog učenja sa SVM klasifikatorom je u skoro svim skupovima podataka (16 skupova) pokazao iste ili bolje rezultate od SVM algoritma na osnovnom skupu podataka. U 9 skupova podataka rezultati su bili i statistički bolji. Metod prethodnog učenja sa J48 klasifikatorom je u više od pola skupova podataka (11 skupova) pokazao iste ili bolje rezultate od J48 algoritma na osnovnom skupu podataka, ali ne postoji rezultat koji bi bio i statistički bolji.

Metod prethodnog učenja sa RBF klasifikatorom je u dve trećine skupova podataka (12 skupova) pokazao iste ili bolje rezultate od RBF algoritma na osnovnom skupu podataka, a u 5 skupova podataka rezultati su bili i statistički bolji.

Korišćenjem metode prethodnog učenja možemo da zaključimo da je SVM klasifikator u najvećem broju slučaja doveo do statistički boljih rezultata na posmatranim skupovima podataka.

Tabela 1. Prikaz setova podataka. „CV“ označava 10-struku unakrsnu validaciju

Skup	Atributi			Broj klasa	Veličina za treniranje	Veličina za testiranje	Referentna tačnost
	ukupno	kategorički	numerički				
bc	9	9	0	2	286	CV	70.30
ca	15	9	6	2	690	CV	55.50
cg	20	13	7	2	1000	CV	50.10
ct	23	0	23	3	2126	CV	95.00
he	19	13	6	2	155	CV	78.10
li	6	0	6	2	345	CV	58.10
lc	56	0	56	3	32	CV	26.80
ma	5	0	5	2	961	CV	84.00
m1	6	6	0	2	124	308	50.00
m2	6	6	0	2	169	263	67.13
m3	6	6	0	2	122	310	52.78
mu	22	22	0	2	8124	CV	51.80
pa	23	0	23	2	195	CV	76.00
pi	8	0	8	2	768	CV	65.10
se	19	0	19	7	2310	CV	14.30
so	35	35	0	19	683	CV	13.47
sh	13	3	10	2	270	CV	55.00
vo	16	16	0	2	435	CV	61.40



Slika 1 - Broj atributa u originalnom skupu i optimalan broj atributa dobijen metodama prethodnog učenja

Tabela 2. Tačnost klasifikacije različitih klasifikatora za originalni i redukovani skup podataka uz pomoć metoda prethodnog učenja

Skup	IBk	IBk_W	Bay	Bay_W	SVM	SVM_W	J48	J48_W	RBF	RBF_W
bc	72.85	69.81	72.70	72.26	72.18	73.47	74.28	72.95	71.41	74.01
ca	81.57	85.22 +	77.86	85.67 +	55.88	85.86 +	85.57	84.43	79.55	85.91 +
cg	71.88	71.70	75.16	74.00	70.00	72.62 +	71.25	71.72	73.58	73.93
ct	98.85	98.42	87.30	98.49 +	81.01	98.38 +	98.57	98.88	97.93	98.67 +
he	81.40	81.85	83.81	82.21	79.38	83.90	79.22	81.90	85.29	82.12
li	62.22	59.66	54.89	59.46	59.37	60.62	65.84	66.36	65.06	62.86
lc	68.75	70.67	78.42	79.33	72.67	77.42	79.25	78.83	76.00	76.08
ma	75.60	83.02 +	82.64	82.01	80.27	82.03	82.19	82.47	77.31	81.11 +
m1	99.87	100.00	74.64	74.64	91.37	97.83 +	97.80	100.00	75.36	88.16 +
m2	79.08	65.72 -	62.79	65.72 +	65.44	65.72	63.48	65.72	67.82	65.67
m3	97.46	98.92 +	96.39	96.39	96.39	98.92 +	98.92	98.92	96.54	97.49
mu	100.00	100.00	95.76	99.63 +	100.00	100.00	100.00	100.00	98.61	97.12
pa	95.91	93.40	69.98	82.04 +	79.36	97.64 +	84.74	86.24	81.22	87.47 +
pi	70.62	67.76	75.75	76.11	65.11	71.76 +	74.49	73.44	74.04	75.79
se	97.15	97.08	80.17	89.83 +	64.76	90.91 +	96.79	96.73	87.88	91.82
so	91.20	94.77 +	92.94	92.67	90.04	89.17	91.78	91.74	84.48	84.41
sh	76.15	78.56	83.59	84.30	55.93	81.74 +	78.15	81.74	83.11	82.59
vo	92.58	94.92 +	90.02	95.75 +	95.63	95.54	96.57	95.24 -	93.73	94.94

5. DISKUSIJA REZULTATA I DALJA ISTRAŽIVANJA

Osnovna hipoteza rada je da je moguće znatno poboljšati performanse sistema za induktivno učenje pravila u problemima klasifikacije, primenom metoda prethodnog učenja za redukciju dimenzionalnosti podataka.

Eksperimentalna istraživanja su sprovedena uz korišćenje veštačkih i prirodnih skupova podataka. Eksperimentalni rezultati pokazuju da primenjene me-

tode efikasno doprinose otkrivanju i eliminisanju nebitnih, redundatnih podataka, kao i šuma u podacima. U mnogim slučajevima opisane metode prethodne selekcije atributa odabiraju relevantne atribute u skupovima podataka, i doprinose većoj tačnosti klasifikacije.

Smatra se da metode prethodnog učenja omogućuju postizanje nešto boljih performansi klasifikacije, zbog tesne povezanosti s ciljnim algoritmom mašinskog učenja. Ovo ujedno može predstavljati i opasnost jer preterano prilagođavanje skupa za učenje ciljnom algoritmu može naglasiti njegove nedostatke.

Sprovedena istraživanja u nadgledanom učenju, pokušavaju dati uvid u prednosti i ograničenja različitih metoda prethodne selekcije atributa. Sa ovakvim uvidom i predznanjem za određeni konkretni problem, stručnjaci mogu odabrati koje metode treba primeniti. Takav je slučaj sa nekim od metoda prethodne selekcije atributa, koje mogu da poboljšaju (ili da ne degradiraju) izvršenje algoritama mašinskog učenja, dok u isto vreme postižu smanjenje broja atributa koji se koriste u učenju. Neke od prikazanih metoda, imale su problem kod izbora relevantnih atributa, kada u podacima postoji snažna interakcija između atributa, ili kada imamo skupove podataka sa oskudnim brojem instanci.

LITERATURA

[1] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artificial Intelligence - Special issue on relevance archive, Volume 97 Issue 1-2, pp. 273 – 324, Dec. 1997.

[2] A. Frank, A. Asuncion, UCI Machine learning repository [http://archive.ics.uci.edu/ml], Irvine, CA: University of California, School of Information and Computer Science, 2010.

[3] J. R. Quinlan, Simplifying decision trees, Int J Man-Machine Studies 27, pp. 221-234, Dec 1987.

[4] J. R. Quinlan, C4.5: Programs for machine learning, San Mateo, Morgan Kaufman, 1993.

[5] D. Ayres de Campos et al., SisPorto 2.0 A Program for automated analysis of cardiotocograms, J Matern Fetal Med 5:311-318, 2000.

[6] Z. Q. Hong, J. Y. Yang, Optimal discriminant plane for a small number of samples and design method of classifier on the plane, Pattern Recognition, Vol. 24, No. 4, pp. 317-324, 1991.

[7] M. Elter, R. Schulz-Wendtland, T. Wittenberg, The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process, Medical Physics 34(11), pp. 4164-4172, 2007.

[8] J. S. Schlimmer, Concept acquisition through representational adjustment (Technical Report 87-19), Doctoral disseration, Department of Information and Computer Science, University of California, Irvine, 1987.

[9] M. A. Little, P. E. McSharry, S. J. Roberts, D. A. E. Costello, I. M. Moroz, Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection, BioMedical Engineering OnLine 2007, 6:23, 26 June 2007.

[10] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, R. S. Johannes, Using the ADAP learning algorithm to forecast the onset of diabetes mellitus, In Proceedings of the Symposium on Computer Applications and Medical Care, pp. 261-265, IEEE Computer Society Press, 1988.

[11] J. H. Piater, E. M. Riseman, P. E. Utgoff, Interactively training pixel classifiers, Published in the International Journal of Pattern Recognition and Artificial Intelligence 13(2), 1999.

[12] R. S. Michalski, R. L. Chilausky, Learning by being told and learning from examples: an experimental comparison of the two methods of knowledge acquisition in the context of developing an expert system for soybean disease diagnosis, International Journal of Policy Analysis and Information Systems, Vol. 4, No. 2, 1980.

SUMMARY

IMPROVING THE ACCURACY OF CLASSIFICATION ALGORITHMS FOR INDUCTIVE LEARNING RULES USING WRAPPER METHODS

In this paper we investigate the problem of the accuracy of classifier using wrapper methods. For the purposes of classification is used a large number of algorithms: IBK, Naïve Bayes, SVM, J48 decision tree and RBF networks. Experimental results show that wrapper methods can rapidly identify irrelevant, redundant attributes, as well as the noise in the data, if any; and those attributes which are important for the studied phenomenon. The paper prove that applying wrapper methods for reducing the dimensionality of the data it is possible to significantly improve system performance for inductive learning rules in classification problems.

Key words: *classifier, wrapper methods, attribute selection, classification accuracy, cross-validation*