UDK: 004 Struni rad

POVEĆANJE TAČNOSTI KLASIFIKACIJE SVM ALGORITMA KORIŠĆENJEM PCA METODE

INCREASING CLASSIFICATION ACCURACY OF SVM ALGORITHM USING PCA METHOD

Jasmina Novaković[1], Vladimir Veljović[2] Miloš Papić[3], Alempije Veljović[2]
[1]*Beogradska poslovna škola, Visoka škola strukovnih studija, Beograd*
[23]*Fakultet tehničkih nauka u Čačku Univerzitet u Kragujevcu*
[1]jnovakovic@sbb.rs, [2]veljo09@gmail.com, [3]milos.papic@ftn.ac.kg.rs

Apstrakt: *U ovom radu smo eksperimentalno istraživali performanse SVM algoritma sa PCA metodom u problemima klasifikacije. Klasifikacija je jedan od najčešćih zadatka mašinskog učenja, i predstavlja problem razvrstavanja nepoznate instance u jednu od unapred ponuđenih kategorija — klasa. 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. PCA predstavlja metodu formiranja novih, sintetskih varijabli koje su linearne složenice - kombinacije izvornih varijabli. Ovom metodom se redukuje dimenzionalnost, a maksimalni broj novih varijabli koji se može formirati jednak je broju izvornih, pri čemu nove varijable nisu međusobno korelisane. Eksperimentalna istraživanja su pokazala da je moguće da se poboljša tačnost klasifikacije SVM algoritma korišćenjem PCA metode.*
Ključne reči: *klasifikacija, mašinsko učenje, PCA, SVM.*

Abstract: *In this paper, we experimentally researched performance of SVM algorithm with PCA method in classification problems. Classification is one of the most common tasks of machine learning, and is a problem of unknown classification instance in one of the pre-given categories - classes. The advantage of SVM over other methods is to provide a better prediction of unseen test data, providing unique optimal solution to the problem in training and the existence of fewer parameters for optimization compared with other methods. PCA is a method of forming new synthetic variables that are linear compound - combinations of the original variables. This method reduces the dimensionality and the maximum number of new variables that can be obtained is equal to the original, with new variables are not correlated with each other. Experimental studies have shown that it is possible to improve the accuracy of SVM classification algorithm using PCA method.*
Key words: *classification, machine learning, PCA, SVM.*

1. UVOD

Klasifikacija je jedan od najčešćih zadatka mašinskog učenja [1], i predstavlja problem razvrstavanja nepoznate instance u jednu od unapred ponuđenih kategorija — klasa. U našoj prirodi da stvari oko sebe, kako bih ih bolje shvatili ili organizovali, klasifikujemo i kategorizujemo. Tako npr. klasifikacija se koristi u: dijagnostifikovanju bolesti, prognozi bolesti kod pacijenta, odabiru najbolje terapije za pacijenta od nekoliko mogućih, klasifikaciji kreditinih zahteva klijenata, proceni da li će i koji korisnici kupiti određeni proizvod, izboru ciljne grupe klijenata za marketinške kampanje, analizi slike, analizi glasa za biometrijska potrebe ili za potrebe analize zdravstvenog stanja osobe, prepoznavanju emotivnog stanja osoba na osnovu slike i glasa, dijagnostifikovanju zdravstvenog stanja biljaka ili životinja i slično. 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.

Proces klasifikacije se sastoji iz dve faze, pri čemu se u prvoj fazi gradi model na osnovu karakteristika objekata čija klasifikacija je poznata. Za izgradnju modela se koriste podaci koji se najčešće nalaze u tabelama. Svaka instanca uzima samo jednu vrednost atributa klase, a atribut klase može da ima konačan broj diskretnih vrednosti koje nisu uređene.

Klasifikacioni algoritam uči na osnovu poznatih klasifikacija tj. na osnovu instanci objekata čija klasifikacije je poznata. Pri tome, na osnovu vrednosti njihovih atributa i atributa klase, gradi se skup pravila na osnovu kojih će se kasnije vršiti klasifikacija. Metode klasifikacije su najčešće zasnovane na stablima odlučivanja, Bajesovim klasifikatorima, neuronskim mrežama, itd.

Nakon učenja, model se testira tj. procenjuje se njegova tačnost, pri čemu pod tačnošću podrazumevamo procenat instanci koje su tačno klasifikovane. Vrednost atributa klase svake testne instance poredi se sa vrednošću atributa klase koja je određena na osnovu modela. Važno je napomenuti da se za testiranje modela koriste instance koje nisu korišćene u fazi učenja.

Postoji više načina za izdvajanje testnih instanci, ali se najčešće izdvajaju slučajnim izborom, pre faze učenja, od instanci čija je klasifikacija poznata. Pri tome, ako je tačnost modela zadovoljavajuća onda se dalje koristi u klasifikaciji objekata čija vrednost atributa klase nije poznata.

Metode rangiranja rangiraju svaki atribut u skupu podataka. Rezultati se potvrđuju korišćenjem različitih algoritama za klasifikaciju. Širok raspon algoritama za klasifikaciju je na raspolaganju, svaki sa svojim prednostima i slabostima. Ne postoji takav algoritam učenja koji najbolje radi sa svim problemima nadziranog učenja. Mašinsko učenje uključuje veliki broj algoritama kao što su: veštačke neuronske mreže, genetski algoritmi, probabilistički modeli, indukcijska pravila, stabla odlučivanja, statističke ili metode raspoznavanje uzoraka, *k*-najbliži susedi, *Naïve Bayes* klasifikatori i diskriminatorna analiza.

U ovom radu korišćen je algoritam nadziranog učenja za izgradnju modela - SVM, jer pruža bolje predviđanje neviđenih test podataka, pruža jedinstveno optimalno rešenje za problem u treniranju i postoji manje parametara za optimizaciju u poređenju sa drugim algoritmima.

2. TEORIJSKI PREGLED ISTRAŽIVANJA METODE POTPORNIH VEKTORA

Metoda potpornih vektora (eng. *Support Vector Machine* - SVM) je binarni klasifikator koji konstrukcijom hiper-ravni u visoko-dimenzionalnom prostoru stvara model koji predviđa kojoj od dve klase pripada nova instanca. Ova metoda je razvijena od strane *Vapnik*-a i saradnika 1995. godine i uživa veliku popularnost zbog veoma dobrih rezultata koji se dobijaju.

U mašinskom učenju, metoda potpornih vektora je popularna zbog svojih dobrih performansi. Kao nadzirana metoda koja analizira podatke i prepoznaje obrasce, ona je strogo utemeljena na statističkim teorijama učenja i istovremeno smanjuje trening i test greške. Osnovna ideja ove metode je da se u vektorskom prostoru u kome su podaci predstavljeni, nađe razdvajajuća hiper-ravan tako da su svi podaci iz date klase sa iste strane ravni.

Kod korišćenja ove metode, postavlja se pitanje koje je rešenje bolje i na koji način definisati „bolje“ rešenje. Ako pretpostavimo da su podaci linerano razdvojivi, u fazi treniranja treba naći optimalnu razdvajajuću hiper-ravan, odnosno ravan sa maksimalnom „marginom“ (što pretstavlja rastojanje od trenirajućih podataka). U tom slučaju nađena hiper-ravan (tj. njena jednačina) je model. Potom, na osnovu modela izračunavamo rastojanje od hiper-ravni i na osnovu toga određujemo klasu (iznad/ispod ravni).

SVM određuje optimalno rešenje koje maksimizuje razdaljinu između hiper-ravni i tačaka koje su blizu potencijalne linije razdvajanja i predstavlja intuitivno rešenje: ako nema tačaka blizu linije razdvajanja, onda će klasifikacija biti relativno laka. U slučaju linearno ne-razdvajajućih problema, koristimo nelinearni SVM, pri čemu je osnovna ideja da se osnovni (ulazni) vektorski prostor preslika u neki više-dimenzioni prostor u kome je skup podataka za trening linearno razdvojiv.

SVM konstruiše hiper-ravan ili skup hiper-ravni u visokom dimenzionalnom prostoru, koji se može koristiti za klasifikaciju, regresiju, ili druge probleme. Mnoge hiper-ravni mogu služiti za klasifikovanje podataka, najbolja hiper-ravan je ona koja predstavlja

najveće razdvajanje, ili marginu između dve klase. Generalno govoreći, kada je veća margina onda je manja greška generalizacije klasifikatora. Bira se hiper-ravan sa maksimalnom marginom, za koju važi da je rastojanje od nje do najbliže tačke podataka na svakoj strani maksimalna.

2.1. LINEARNO ODVOJIVE KLASE

Ako u skupu za učenje imamo *L* vektora, odnosno tačaka u *D*-dimenzionalnom prostoru, gde svaki uzorak x_i ima *D* atributa, odnosno komponenti vektora i pripada jednoj od dve klase $y_i = -1 \; ili \; 1$, tada oblik jednog ulaznog podatka možemo prikazati izrazom:

$$\{x_i, y_i\} \text{ gde je } i = 1 \dots L \;,\; y_i \in \{-1,1\} x \in R^D, \tag{1}$$

gde se pretpostavlja da su podaci linearno odvojivi, što znači da možemo nacrtati pravac u koordinatnom sistemu sa osama x_1 i x_2 za slučaj *D* = 2, odnosno hiper-ravan za slučaj $D > 2$. Izrazom $w \cdot x + b = 0$ možemo opisati hiper-ravan pri čemu je *w* normala hiper-ravni i $\frac{b}{\|w\|}$ vertikalna udaljenost hiper-ravni od ishodišta koordinatnog sistema. Uzorci najbliži razdvajajućoj hiper-ravni su potporni vektori i zato se najteže klasifikuju. Cilj metoda potpornih vektora jeste da izabere hiper-ravan maksimalno udaljenu od najbližih uzoraka obe klase.

Na ovakav način se implementacija metode potpornih vektora svodi na izbor parametara *w* i *b*, takvih da ulazne podatke možemo opisati sledećim izrazima:

$$x_i \cdot w + b \geq +1 \text{ za } y_i = +1 \tag{2}$$

$$x_i \cdot w + b \leq -1 \text{ za } y_i = -1 \tag{3}$$

Kombinovanjem dva prethodna izraza dobijamo:

$$y_i(x_i \cdot w + b) \quad 1 \geq 0, \forall i \tag{4}$$

Ravni H_1 i H_2 na kojima leže potporni vektori možemo prikazati sledećim izrazima:

$$x_i \cdot w + b = +1 \text{ za } H_1 \tag{5}$$

$$x_i \cdot w + b = -1 \text{ za } H_2 \tag{6}$$

Ako definišemo vrednosti d_1 i d_2 kao rastojanje od H_1 i H_2 do hiper-ravni, ekvidistantnost hiper-ravni od H_1 i H_2 podrazumeva $d_1 = d_2 = \frac{1}{\|w\|}$, pri čemu vrednost d_1, односно d_2 nazivamo marginom. Da bi izabrali hiper-ravan maksimalno udaljenu od potpornih vektora, potrebno je maksimizirati marginu, što je ekvivalentno pronalaženju:

$$min\|w\| \text{ такав да } [y_i(x]_i \cdot w + b) - 1 \text{ за } \geq 0, \forall i \tag{7}$$

2.2. LINEARNO NEODVOJIVE KLASE

Da bi metode sa potpornim vektorima koristili i za na linearno neodvojive klase, potrebno je ublažiti uslove (1) i (2) uvođenjem nenegativne vrednosti ξ_i :

$$x_i \cdot w + b \geq +1 - \xi_i \text{ за } y_i = +1 \quad (8)$$

$$x_i \cdot w + b \leq -1 + \xi_i \text{ за } y_i = -1 \quad (9)$$

Kombinovanjem prethodna dva izraza dobijamo sledeći izraz:

$$[y_i(x]_i \cdot w + b) - 1 + \xi_i \geq 0, \xi_i \geq 0, \forall i \quad (10)$$

Primenjena metoda se naziva metoda meke margine (eng. *soft margin method* [2]), a izvorno je nastala sa idejom dozvoljavanja pogrešnog označavanja klasa pre samog postupka učenja. Mera rastojanja tog uzorka od pripadajućeg potpornog vektora je ξ . Izbor razdvajajuće hiper-ravni svodi se na pronalaženje:

$$\frac{\min 1}{2}\|w\|^2 + C\sum_{i=1}^{L}\xi_i \text{ такав да } [y_i(x]_i \cdot w + b) - 1 + \xi_i \geq 0, \forall i \quad (11)$$

gde vrednost *C* predstavlja faktor greške, kojim dozvoljavamo određene greške pri treniranju, bez čega pronalazak hiper-ravni ne bi bio moguć. Problem se može rešiti upotrebom metode Lagranžovih koeficijenata, što se može pokazati korisnim kod nelinearnih jezgara, i tada se dobija efikasan iterativan algoritam LSVM. Primer veoma jednostavne i efikasne implementacije je SMO (eng. *Sequential Minimal Optimization*), koji koristi razbijanje na najmanji podproblem gde se onda lako određuje vrednost jednog po jednog preostalog koeficijenta [3] umesto skupog numeričkog rešavanja problema kvadratnog programiranja.

3. METODOLOGIJA RADA I ORGANIZACIJA ISTRAŽIVANJA

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.

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 SVM na originalnom skupu podataka kao i na redukovanom skupu podataka. Problem dimenzionalnosti se može prevladati tako da se odabere samo podskup relevantnih atributa ili stvaranjem novih atributa koje sadrže najviše informacija o klasi. Prva metodologija se zove selekcija atributa, a druga se zove ekstrakcija atributa, a to uključuje linearnu (PCA, ICA i sl.) i ne-linearnu metodu ekstrakcije atributa.

PCA predstavlja tehniku formiranja novih, sintetskih varijabli koje su linearne složenice - kombinacije izvornih varijabli. Ovom tehnikom se redukuje dimenzionalnost, a maksimalni broj novih varijabli koji se može formirati jednak je broju izvornih, pri čemu

nove varijable nisu međusobno korelisane. Očekuje se da će većina novih varijabli činiti šum, i imati tako malu varijansu da se ona može zanemariti. Većinu informacija će poneti prvih nekoliko varijabli - glavnih komponenti, čije su varijanse značajne veličine. Na taj način, iz velikog broja izvornih varijabli kreirano je tek nekoliko glavnih komponenti koje nose većinu informacija i čine glavni oblik. Naravno, ima situacija kada to nije tako, i to u slučaju kada su izvorne varijable nekorelisane, tada analiza ne daje povoljne rezultate.

U analizi glavnih komponenata osnovni koraci su: standardizacija varijabli, izračunavanje matrice korelacije, pronalaženje svojstvenih vrednosti glavnih komponenti i odbacivanje komponenti. Najpre, potrebno je standardizovati varijable tako da im je prosek 0, a varijansa 1 kako bi sve bile na jednakom nivou u analizi, jer je većina setova podataka konstruisana iz varijabli različitih skala i jedinica merenja. Potom, potrebno je izračunati matrice korelacija između svih izvornih standardizovanih varijabli, a nakon toga, pronaći svojstvene vrednosti glavnih komponenata. Na kraju, potrebno je odbaciti one komponente koje imaju proporcionalno mali udeo varijanse (obično prvih nekoliko komponenti ima 80% - 90% ukupne varijanse). U eksperimentalnom istraživanju koristili smo za prag odbacivanja vrednost od 95%.

4. REZULATI EKSPERIMENTALNOG ISTRAŽIVANJA I DISKUSIJA DOBIJENIH REZULTATA

Za potrebe eksperimentalnog istraživanja koristili smo 15 realnih skupova podataka (*bc, ca, cg, ct, he, li, lc, ma, mu, pa, pi, se, so, sh, vo*) i 3 veštačka (*m1, m2, m3*), preuzeta iz UCI repozitorijuma [4], koji je namenjen istraživačima koji pročavaju probleme veštačke inteligencije.

Tabela 1. Tačnost klasifikacije SVM algoritma uz pomoć PCA [5]

Skup	SVM	SVM_P
bc	72.18	71.50
ca	55.88	85.49 +
cg	70.00	75.53 +
ct	81.01	98.97 +
he	79.38	85.90 +
li	59.37	62.29
lc	72.67	71.67
ma	80.28	81.97
m1	91.37	100.00 +
m2	65.44	61.20 -
m3	96.39	98.92 +
mu	100.00	99.94 -
pa	79.36	82.89

pi	65.11	76.38 +
se	63.98	91.68 +
so	93.63	92.94
sh	55.93	83.37 +
vo	95.63	94.25

U tabeli 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.

PCA je kod SVM algoritma u dve trećine skupova podataka (12 skupova) pokazao iste ili bolje rezultate od SVM algoritma na osnovnom skupu podataka. U 9 skupova podataka rezultati za tačnost klasifikacije su bili i statistički bolji.

5. ZAKLJUČAK

Možemo zaključiti da je moguće da se poboljša tačnost klasifikaciji SVM algoritma, koristeći PCA metod za smanjenje dimenzionalnosti podataka. Da bi to dokazalo, realizovali smo i empirijski testirali metod prethodnog učenja za smanjenje dimenzionalnost podataka. Eksperimentalni rezultati pokazuju da PCA metoda doprinosi otkrivanju i otklanjanju nevažnih i suvišnih podataka, kao i šuma u podacima. U mnogim slučajevima PCA metoda dovodi do veće preciznosti klasifikacije.

ZAHVALNICA

Autori se zahvaljuju za podršku Ministarstvu prosvete, nauke i tehnološkog razvoja Republike Srbije - Projekti TR 34009 i TR1653014.

LITERATURA

[1] Janičić, P., Nikolić, M. (2010). *Veštačka inteligencija*, Matematički fakultet u Beogradu.

[2] Fletcher, T. (2009). *Support Vector Machines Explained*, http://www.tristanfletcher.co.uk/SVM%20Explained.pdf.

[3] Platt, J.C. (1999). *Fast training of Support Vector Machines using sequential minimal optimization, Advances in kernel methods*, Pages 185-208, MIT Press Cambridge, MA, USA.

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

[5] Novaković, J. (2013). *Redukcija dimenzionalnosti podataka u klasifikacionim problemima veštačke inteligencije*, doktorska disertacija, Univerzitet u Kragujevcu, Fakultet tehničkih nauka Čačak.