PRIMENA SVM METODA ZA KLASIFIKACIJU PACIJENATA OBOLELIH OD ASTME

APPLICATION OF SUPPORT VECTOR MACHINES FOR CLASSIFICATION OF PATIENTS SUFFERING FROM ASTHMA

Miloš Stojanović, Leonid Stoimenov, Milena Stanković, *Elektronski fakultet Niš,*

Svetlana Kamenov, *Dom Zdravlja Niš*

Sadržaj - *U ovom radu prikazana je primena metoda Support Vector Machines (SVM) za klasifikaciju pacijenata obolelih od astme. U radu je dato kratko upoznavanje sa SVM algoritmom za mašinsko učenje i klasifikaciju podataka. Za realizaciju klasifikatora korišćena je javno dostupna biblioteka LibSVM u verziji za C#. Izabrano je RBF jezgro sa odgovarajućim parametrima dobijenim postupkom kros-validacije. Testiranje je obavljeno nad raspoloživim podacima o pacijentima, koji se na osnovu odgovarajućih parametara klasifikuju u jednu od tri kategorije bolesti: laku, umerenu i tešku. Prikazani su rezultati testiranja: preciznost, odziv i F mera za svaku klasu pojedinačno, kao i ukupna uspešnost klasifikatora F^{macro} i F^{micro}, koji potvrđuju primenljivost ove metode.*

Abstract - *This paper presents the application of Support Vector Machines (SVM) for classification of patients with asthma. In the paper a brief introduction to SVM algorithm for machine learning and classification of data is given. For implementation of the classifier, a publicly accessible C# library LibSVM is used. The RBF kernel, with appropriated parameters obtained by using cross–validation, is applied. For testing a set of data about the patients, which are classified based on the appropriate parameters into one of three categories of disease: easy, moderate and severe is used. The results of testing: precision, recall, F measures for each class individually, and the overall performance of classifier F^{macro} and F^{micro} confirm the application of this method.*

1. UVOD

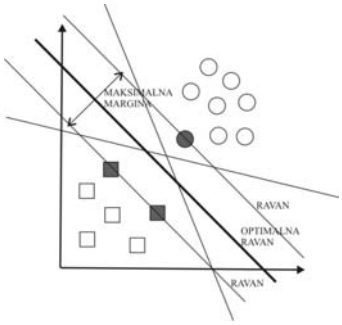
Support Vector Machines (SVM) metod za klasifikaciju podataka, razvijen od strane Vapnik-a i saradnika 1995 godine, uživa veliku popularnost u ovoj oblasti zbog veoma dobrih rezultata koji su slični ili bolji od onih koji se dobijaju primenom neuronskih mreža [1].

U ovom radu najpre će biti opisan princip funkcionisanja SVM-a (linearno razdvajanje, SVM sa "mekom" marginom i kernel funkcije), sa posebnim naglaskom na mere uspešnosti klasifikatora.

U nastavku rada opisan je generisani model klasifikatora koji vrši klasifikaciju pacijenata obolelih od astme u jednu od tri kategorije, kao i rezultati izvršenog testiranja sa raspoloživim podacima. Za modelovanje klasifikatora korišćena je javno dostupna biblioteka LibSVM u verziji za C#. Za potrebe pripreme podataka dobijenih od lekara napisan je jednostavan parser u C#-u koji iz ulaznog Excel fajla izdvaja i transformiše podatke u format koji prihvata LibSVM. Za klasifikaciju izabrano je RBF jezgro sa odgovarajućim parametrima dobijenim postupkom kros-validacije. Prikazani su rezultati testiranja kao i planirane aktivnosti za unapređenje dobijenih rezultata.

2. SVM KLASIFIKATORI

Kod primene SVM metoda za klasifikaciju osnovni zadatak je da se generiše model klasifikatora, koji je u ovom slučaju formula, a ne skup pravila. U vektorskom prostoru u kome su predstavljeni podaci treba naći optimalnu razdvajajuću hiper-ravan, tako da su svi podaci iz date klase sa iste strane ravni – što je zadatak trening faze. Optimalna hiper-ravan je ona sa maksimalnom marginom, tj maksimalnim rastojanjem od trenirajućih podataka, Slika 1. Jednačina te hiper-ravni predstavlja model na osnovu koga se obavlja klasifikacija. Rastojanje od hiper-ravni određuje izlaznu klasu – test faza. Pretpostavlja se da su podaci linearno razdvojivi (*Lineary separable*).



Slika 1. *Optimalna hiper-ravan, sa maksimalnom marginom.*

SVM određuje optimalnu razdvajajuću ravan maksimizujući rastojanje između hiper-ravni i tačaka koje su blizu potencijalne linije razdvajanja. Zatim SVM maksimizuje marginu oko razdvajajuće hiper-ravni. Razdvajajuća hiper-ravan je potpuno određena podskupom trening podataka koji se nazivaju *podržavajući-potporni vektori*, po čemu je metod i dobio naziv.

2.1 SVM kada su podaci linearno razdvojivi

Pretpostavimo da nam je dat skup trenig podataka koji se sastoji od n elemenata $\mathbf{x_i}, i = 1,\ldots,n$, gde je svaki element predstavljen d-dimenzionalnim vektorom $\mathbf{x_i} = (x_1, x_2, \ldots, x_d)$ (vektorski prostor sa dimenzijom d). Svakom podatku iz trening skupa je pridružena vrednost $y_i \in \{-1,1\}$ tj. klasa kojoj pripada (binarna klasifikacija). Jednačina hiper-ravni je:

$$\mathbf{w}^{\mathbf{T}}\mathbf{x} + b = 0$$

i ona je potpuno određena parametrima **w** i b. Parametar **w**, vektor težina određuje smer hiper-ravni, dok parametar b, pomeraj određuje udaljenost hiper-ravni od centra koordinatnog sistema.

Klasifikacija "nepoznatih" primera je zasnovana na znaku izraza: $\mathbf{w}^{\mathbf{T}}\mathbf{x} + b$ **,**[3]. Uslov razdvajanja, za svaku tačku $\{\mathbf{x_i}, y_i\}$ može da se formuliše sledećim uslovima:

$$\mathbf{w}^{\mathbf{T}}\mathbf{x}_i + b \geq 1 \; if \; y_i = 1 \text{ i}$$

$$\mathbf{w}^{\mathbf{T}}\mathbf{x}_i + b \leq -1 \; if \; y_i = -1$$

Rastojanje podržavajućih vektora od ravni razdvajanja biće $r = 1/\|\mathbf{w}\|$, a "debljina" margine biće $2/\|\mathbf{w}\|$. Da bi maksimizovali marginu treba minimizovati $\|\mathbf{w}\|$ tj. ½ $\mathbf{w}^{\mathrm{T}}\mathbf{w}$**.**

Potrebno je odrediti parametre **w** i b tako da se minimizuje $\Phi(\mathbf{w}) = ½\ \mathbf{w}^{\mathrm{T}}\mathbf{w}$ (uslov maksimalne margine) uz uslov $y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x_i} + b) \geq 1$ (razdvajanje).

Ovo je kvadratni optimizacioni problem uz linearne uslove, i jedan od načina za njegovo rešavanje je uz pomoć Lagranžeovih multiplikatora [4]. Nakon primene matematičkog aparata koji nije prikazan u ovom radu, za funkciju odluke (*Decision function*) se dobija:

$$f(\mathbf{x}) = sign(\sum a_i y_i \mathbf{x}_i^T \mathbf{x} + b)$$

Funkcija odluke - model klasifikatora je predstavljena skupom koeficijenata α_i i skupom podržavajućih vektora $\mathbf{x}_i$ Primećujemo da je za klasifikaciju svakog elementa **x** potrebno izračunati skalarni proizvod sa podržavajućim vektorima.

SVM može da se uopšti i na klasu problema kada podaci nisu linearno razdvojivi i to na dva načina [3].

2.2 SVM sa "mekom" marginom

Prvi način je uvodjenje promenljivih ξ_i koje bi tolerisale "male" greške prilikom faze učenja klasifikatora i kasnije prlikom klasifikacije. Sada su uslovi minimizacije malo izmenjeni i potrebno je odrediti parametre **w** i b tako da se minimizuje:

$$\Phi(\mathbf{w}) = ½\ \mathbf{w}^{\mathrm{T}}\mathbf{w} + C\,\Sigma\xi_i$$

uz uslov:

$$y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x_i} + b) \geq 1 - \xi_i\ ,\ \xi_i \geq 0$$

gde je C konstanta koja pravi kompromis-balansira izmedju margine i greške prilikom učenja (minimizovaće grešku na račun ne baš maksimalne margine-razdvajanja). Za funkciju odluke se dobija:

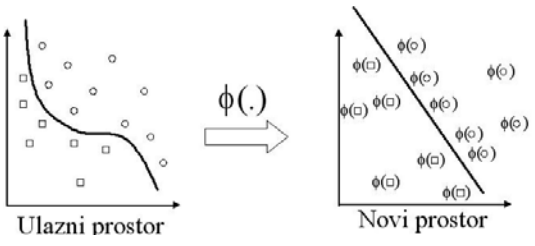
$$f(\mathbf{x}) = \Sigma\alpha_i y_i \mathbf{x_i}^{\mathrm{T}}\mathbf{x} + \boldsymbol{b}$$

Potrebno je izabrati prag odlučivanja p, ako je $f(\mathbf{x}) > p$ onda element pripada klasi, ako je $f(\mathbf{x}) < p$ onda ne pripada, dok je $f(\mathbf{x}) = p$ nedefinisan slučaj.

2.3 Kernel funkcije (funkcije jezgra)

Drugi način je da osnovni vektorski prostor u kome je trening skup linearno ne razdvojiv preslikamo u neki više dimenzionalni prostor u kome je trenirajući skup linearno razdvojiv pomoću preslikavanja Φ: $\mathbf{x} \rightarrow \varphi(\mathbf{x})$, Slika 2.

Umesto skalarnog proizvoda (mera sličnosti dva vektora) uvodi se kernel funkcija koja odgovara skalarnom proizvodu u preslikanom prostoru (prostru veće dimenzije) [5]. Ako svaku tačku preslikamo u prostor veće dimenzije, skalarni proizvod postaje $K(\mathbf{x_i}, \mathbf{x_j}) = \varphi(\mathbf{x_i})^{\mathrm{T}}\varphi(\mathbf{x_j})$, tj dovoljno je da znamo samo kako se izračunava skalarni proizvod u novom) prostoru.



Slika 2. *Ulazni prostor preslikavamo u više dimenzionalni prostor gde imamo razdvajanje.*

Problem koji se nameće je kako odabrati preslikavanje **K** tj. kernel koji odgovara skalarnom proizvodu u nekom novom prostoru. Postoji posebna matematička teorija o tome kako se konstruiše kernel za dati problem koja ovde neće biti opisivana. U ovom radu biće korišćeno

RBF jezgro: $K(\mathbf{x_i},\mathbf{x_j}) = \exp(-\gamma \left\| \mathbf{x_i} - \mathbf{x_j} \right\|^2)^d$, $\gamma > 0$ [6], gde su γ i d parametri jezgra.

2.4 Mere uspešnosti klasifikatora

Mere uspešnosti klasifikatora se definišu na osnovu svih mogućih ishoda prilikom predviđanja klase koji su dati u Tabeli 1.

Tabela 1. *Mogući ishodi prilikom predviđanja klase.*

	Stvarna klasa $y = +1$	**Stvarna klasa $y = -1$**
Predložena klasa $h(\mathbf{x}) = +1$	**SP-** stvarno pozitivni	**LP**-lažno pozitivni
Predložena klasa $h(\mathbf{x}) = -1$	**LN**-lažno negativni	**SN**-stvarno negativni
$\sum$	**Ukupno pozitivnih**	**Ukupno negativnih**

Dve osnovne mere koje se intuitivno nameću su tačnost i greška:

$$\text{Tačnost} = \frac{SP + SN}{SP + SN + LP + LN}$$

$$\text{Greška} = \frac{LP + LN}{SP + SN + LP + LN}$$

Ove dve mere uspešnosti klasifikatora nisu pogodne iz razloga zato što pridaju jednaku važnost i lažno pozitivnim i lažno negativnim ishodima klasifikacije, tj ne govore ništa o tome sa koliko uspeha klasifikator klasifikuje stvarno pozitivne ili stvarno negativne primere.

Zbog toga se uvode jos tri mere: preciznost, odziv (pokrivanje) i F_b- mera [8].

Preciznost određuje tačnost klasifikacije, tj da element koji je klasifikovan u klasu $h(\vec{x}) = +1$ zaista i pripada toj klasi $y = +1$.

$$\text{Preciznost} = \frac{SP}{SP + LP}$$

Odziv ocenjuje "pokrivanje klase" tj. koliko primera iz date klase model može da prepozna.

$$\text{Odziv} = \frac{SP}{SP + LN}$$

F_b- mera efikasnosti je harmoniska sredina preciznosti i odziva sa težinom b kao parametrom koji naglašava važnost ili preciznosti ili odziva.

$$F_b = \frac{(1 + b^2) \cdot \text{Preciznost} \cdot Odziv}{b^2 \cdot \text{Preciznost} + Odziv}.$$

Ako je $b = 0$ tada ova mera postaje preciznost, za $b = 1$ i preciznost i odziv imaju jednak uticaj, dok kada b teži beskonačnosti ova mera postaje odziv.

Ako treba da se odredi uspešnost klasifikatora na više trening ili test skupova ili na više različitih klasa mora da se obuhvate svi dobijeni rezultati. Ovo se može postići na dva načina, makro-usrednjavanjem i mikro-usrednjavanjem.

Makro-usrednjavanje usrednjava rezultate eksperimenata tj. F_b mera. Mera efikasnosti se računa posebno za svaki eksperiment, a zatim se izračuna aritmetička sredina:

$$\mathbf{F}^{macro} = \frac{1}{n} \sum_{i=1}^{n} F_i$$

Mikro-usrednjavanje usrednjava polja iz tablice mogućih ishoda (Tabela 1.). Izračuna se aritmetička sredina elemenata iz tablica ishoda i zatim se izračuna mera efikasnosti:

$$\mathbf{F}^{micro} = \frac{2 \cdot SP^{avg}}{2 \cdot SP^{avg} + LP^{avg} + LN^{avg}}.$$

3. KLASIFIKACIJA PACIJENATA OBOLELIH OD ASTME

3.1 Opis problema

Testiranje smo obavili nad skupom test primera o pacijentima obolelim od astme. Na raspolaganju su nam bili podaci o 150 pacijenata, dobijeni na osnovu praćenja ove bolesti u Domu zdravlja u Nišu.

Pacijenti se prema težini bolesti, na osnovu podskupa parametara koje procenjuje lekar,"ručno" klasifikovani u tri kategorije: laku, srednju i tešku. Parametri na osnovu kojih lekar odlučuje o težini bolesti su numeričkog tipa, dobijeni na osnovu pregleda, laboratoriskih analiza, terapije i praćenja stanja pacijenata u vremenskom periodu od 5 i 15 dana. Lekar na osnovu iskustva iz prethodnih slučajeva i parametara procenjuje i određuje težinu bolesti, odnosno svrstava pacijente u jednu od tri navedene klase.

Ukupno, u posmatranom skupu pacijenata, od lakog oblika bolesti boluje 88 (58,7%) pacijenata, od umerenog oblika 32 (21,3%) pacijenta i od teškog oblika 30 (20,0%) pacijenata.

Primarni parametri koji učestvuju u odluci o težini bolesti tj. klasifikaciji su: `FVC_OST` - ostvareni forsirani vitalni kapacitet, `FEV1_OST` – ostvareni forsirani ekspiratorni volumen u prvoj sekundi i `FEF_OST` - forsirani srednji ekspiratorni protok.

Na osnovu pravilno klasifikovanih test primera SVM generiše model klasifikatora, koji kasnije u test fazi klasifikuje podatke u jednu od tri klase.

Za generisanje modela klasifikatora iskorišćena je dostupnu biblioteku LibSVM - A Library for Support Vector Machines [9] u verziji za C#. Celokupan proces klasifikacije podataka je podeljen u sledećih nekoliko faza:

1. Priprema podataka za LibSVM
2. Skaliranje podataka
3. Izbor jezgra i određivanje najboljih parametara C i γ postupkom kros-validacije
4. Treniranje klasifikatora
5. Testiranje

3.2 Priprema podataka

Prvi koak u procesu automatske klasifikacije je priprema (parsiranje) podataka u format koji prihvata LibSVM. Ulazni podaci dobijeni od lekara su u obliku MS Excel tabele i treba ih izdvojiti i transformisati u txt fajl koji ima sledeću strukturu:

[label] [index1]:[value1] [index2]:[value2] ...
[label] [index1]:[value1] [index2]:[value2] ...

gde su:

label - klasa kojoj element pripada – tipa integer

index - na osnovu njega se jednoznačno pristupa elementima vektora – tipa integer

value – podatak za treniranje ili testiranje – tipa real

Na primer jedna linija u ulaznom txt fajlu bi bila:

+1 1:0.708 2:1 3:1 4:-0.320 5:-0.105 6:-1

Za potrebe pripreme podataka u LibSVM format napisan je jednostavan parser u C#-u koji iz ulaznog fajla izdvaja i transformiše podatke u gore navedeni format.

3.3 Skaliranje podataka

Skaliranje podataka se obavlja iz razloga da atributi sa većom numeričkom vrednošću ne bi bili dominantni u odnosu na one sa manjom. Drugi razlog je zbog ubrzavanja numeričkih izračunavanja. Vrši se linearno skaliranje podataka na opseg [-1,1].

3.4 Izbor jezgra i određivanje parametara

U slučajevima kada je broj primera za klasifikaciju mnogo veći od broja atributa (dimenizja vektora - prostora) preporučuje se korišćenje RBF kernela [6]. To se radi zbog toga što u slučaju kada je dimenzija prostora mala veća je verovatnoća da su podaci linearno ne radvojivi. RBF kernel mapira ulazni prostor u prostor sa više dimenzija gde je verovatnoća linearnog razdvajanja veća. Za generisanje klasifikatora korišćeno je RBF jezgro.

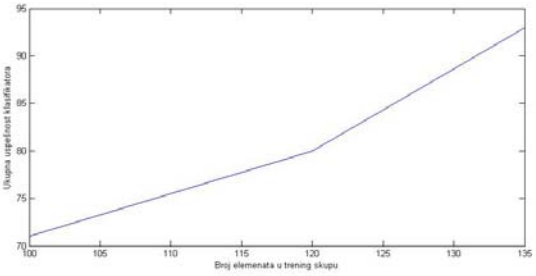
RBF jezgro koristi dva parametra: C i γ. Unapred nije poznato koje vrednosti ovih parametara su najbolje za dati problem. Treba odabrati takvu vrednost parametara koja će najbolje klasifikovati nepoznate (test) podatke. Parametri se mogu odrediti korišćenjem postupka kros-validacije *(Cross-Validation)*.

Skup ispravno klasifikovanih podataka (za treniranje) se na slučajan način podeli na trenirajuće i test delove, napr. u odnosu 1:9. Zatim se primeni algoritam za učenje na trenirajućem delu i proceni se kvalitet naučenog klasifikatora na test delu. Ovaj postupak se ponovlja *n* puta npr. 10 i izaberu se C i γ sa kojima se postiže najbolja preciznost.

3.5 Rezultati testiranja

Testiranje je obavljeno nad raspoloživim skupom od 150 pacijenata, korišćenjem 3 atributa za klasifikaciju. Za treniranje je korišcen podskup od 100, 120 i 135 slučajno odabranih uzoraka, tj. 50, 30 i 15 za testiranje respektivno. U Tabeli 2. prikazani su preciznost, odziv i F mera za svaku klasu pojedinačno, kao i ukupna uspešnost klasifikatora F^{macro} i F^{micro}.

Iz tabele se vidi da bez obzira na veličinu trening skupa klasifikator najbolje klasifikuje instance iz kategorije L. To je posledica toga što najveći broj pacijenata u ukupnom uzorku ima tip astme koji se svrstava u laku kategoriju, tako da povećanje trening skupa ne utiče u toj meri na uspešnost klasifikacije kategorije L. Što se tiče druge dve kategorije T i S uspešnost klasifikatora se znatno poboljšala sa povećanjem trening skupa.



Slika 3. *Zavisnost uspešnosti klasifikacije od veličine trening skupa*

4 ZAKLJUČAK

Rezultati testiranja prikazani u radu, iako dobijeni na veoma malom skupu podataka za treniranje/testiranje pokazuju primenljivost SVM-a u jednom automatskom sistemu za klasifikaciju pacijenata obolelih od astme.

Tabela 2. *Rezultati tesiranja*

Broj atributa	Broj trening primera	Broj test primera	Klasa	Preciznost	Odziv	F_b	F^{macro}	F^{micro}
3	100	50	L	86%	76%	81%	47%	71%
			T	100%	18%	30%		
			S	20%	60%	30%		
	120	30	L	85%	94%	89%	69%	80%
			T	80%	50%	61%		
			S	50%	66%	57%		
	135	15	L	91%	100%	95%	91%	93%
			T	100%	66%	80%		
			S	100%	100%	100%		

Ukupna uspešnost klasifikatora koje se kreće od 71% do 93% predstavalja jako dobar rezultat, s obzirom na to da je testiranje vršeno sa samo 150 instanci. Da je na raspolaganju bilo više podataka imali bi veći trening skup, što bi impliciralo uniformniju raspodelu trening skupa po klasama, tj bolje rezultate klasifikacije. Na Slici 3. prikazana je zavisnost uspešnosti klasifikacije od veličine trening skupa.

Može se zaključiti da realizovani klasifikator uspešno obavlja klasifikovanje pacijenata obolelih od astme u jednu od tri klase.

Automatska klasifikacija mogla bi da služi ne samo kao asistencija lekaru tj. predlog težine bolesti, već bi mogla da vrši statističku analizu uspešnosti lečenja. Vršila bi se automatska klasifikacija pacijenata pre i nakon završetka terapije i na osnovu tih podataka procenjivala uspešnost lečenja.

Planirane aktivnosti za unapređenje dobijenih rezultata podrazumevaju uvođenje dodatnih parametara (atributa) za klasifikaciju, kao i proširenje skupa trening/test primera sa novim podacima o pacijentima. Očekuje se da će proširenje trening/test skupa u značajnoj meri uticati na poboljšanje rezultata klasifikacije.

LITERATURA

[1] S. Čabarkapa, N. Kojić, V. Radosavljević, B. Reljin, "Jedna implementacija SVM u CBIR sistemu", Telfor, Beograd, 2008.

[2] S. R. Gunn, "Support Vector Machines for Classification and Regression", Technical Report, Image Speech and Intelligent Systems Research Group, University of Southampton, 1997.

[3] I. Pilaszy, "Text Categorization and Support Vector Machines", The Proceedings of the 6th International Symposium of (2005) Volume: 1.

[4] P. H. Chen, C. J. Lin, B. Scholkopf, "A Tutorial on ν-Support Vector Machines", 2002, http://www.csie.ntu.edu.tw/ ~cjlin/papers/nusvmtutorial.pdf.

[5] F. Rossi, N. Villa, "Classification in Hilbert Spaces with Support Vector Machines", Proceedings of XIth International Symposium on Applied Stochastic Models and Data Analysis, ASMDA 2005, Brest, France.

[6] C. W. Hsu, C. C. Chang, C. J. Lin, "A Practical Guide to Support Vector Classification", Department of Computer Science National Taiwan University, Taipei 106, Taiwan

http://www.csie.ntu.edu.tw/~cjlin Last updated: May 19, 2009.

[7] K. B. Duan, S. S. Keerthi, "Which Is the Best Multiclass SVM Method? An Empirical Study", Proceedings of the Sixth International Workshop, MCS 2005, Seaside, CA, USA, June 13-15, 2005, pp. 278-285.

[8] C. Goutte, E. Gaussier, "A Probabilistic Interpretation of Precision, Recall and F-score, with Implication for Evaluation", Xerox Research Centre Europe, 6, chemin de Maupertuis, F-38240 Meylan, France.

[9] C. C. Chang, C. J. Lin, "LibSVM: a library for support vector machines", Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/.