Izbor klasifikatora za mali obučavajući skup obrazovnih podataka

Gabrijela Dimić, Dragana Prokin
gdimic@viser.edu.rs, dprokin@viser.edu.rs

Visoka škola elektrotehnike i računarstva strukovnih studija
Beograd, Srbija

Kristijan Kuk, Boško Bogojević
kkristijan@viser.edu.rs, bbosko@viser.edu.rs

Visoka škola elektrotehnike i računarstva strukovnih studija
Beograd, Srbija

Sadržaj — U radu je prikazan postupak izbora preciznih klasifikatora za mali obučavajući skup obrazovnih podataka. Korišćeni su podaci kursa Arhitektura i organizacija računara 1 realizovanog u okviru programa daljinskog učenja na Visokoj školi elektrotehnike i računarstva strukovnih studija u Beogradu. Testirana su četri klasifikatora (*OneR, J48, Naive Bayes, BayesNet TAN*) primenom metode unakrsne validacije. Za opisanu studiju slučaja utvrđeno je da algoritmi *Naive Bayes* i *J48* generišu model klasifikacije sa tačnošću većom od 70%.

Ključne riječi: Klasifikacija; Data mining; Daljinsko učenje; Predviđanje;Moodle

I. UVOD

Karakteristika koja izdvaja daljinsko učenje od ostalih obrazovnih metoda predstavlja činjenicu da student, upotrebom posebno dizajniranih materijala i komunikacionih kanala, savlada predviđeno gradivo bez fizičkog prisustva na nastavnom času [1]. Kod metode obrazovanja na daljinu, student se oseća izolovan pa su sve vrste komunikacionih kanala (student/ nastavnik, student/student) važan parametar za uspeh programa [2]. Tokom procesa učenja, suočavajući se sa teškoćama i nejasnoćama u gradivu, student može postati demoralisan. U nekim slučajevima ova činjenica usporava učenje a ponekad i dovodi do odustajanja od studiranja.

Prethodne istraživačke studije [3, 4] pokazale su da studenti pristupaju metodi obrazovanju na daljinu sa različitim očekivanjima vezanim za podršku, komunikaciju sa nastavnikom, dostupne resurse i materijale za učenje. Nastavnik kursa za daljinsko učenja ima značajnu ulogu posebno sa stanovišta predviđanja ponašanja i učinka studenata već na samom početku. Za postizanje navedenog cilja, upotreba *Data Mining* algoritama mašinskog učenja su perspektivna i obećavajuća oblast.

Data Mining odnosno „rudarenje podataka“ se definiše kao proces izdvajanja značajnih i razumljivih informacija sadržanih u velikim bazama podataka, a sve u cilju donošenja ispravnih poslovnih odluka [5].

Oblast koja se bavi problemima razvoja metoda za otkrivanje znanja iz podataka obrazovnih okruženja poznata je pod nazivom Obrazovni Data Mining (*eng. Educational Data Mining – EDM*) [6]. *Educational Data Mining* obuhvata proces pretvaranja sirovih podataka iz obrazovnih sistema u korisne informacije koje mogu imati veliki uticaj na obrazovno istraživanje i praksu. Ovaj proces se ne razlikuje od primene *Data Mining* metoda u drugim oblastima kao što su poslovanje, genetski inženjering, medicina zato što prati iste korake kao i opšti *Data Mining* proces (predprocesiranje, data mining i postprocesiranje).

U ovom radu korišćeni su podaci on-line kursa Arhitektura i organizacija računara 1 koji je realizovan na Moodle LMS sistemu [7] u okviru programa daljinskog učenja na Visokoj školi elektrotehnike i računarstva strukovnih studija u Beogradu. Studijski program učenja na daljinu Nove računarske tehnologije je akreditovan 2012. godine. Prva grupa studenata upisala je studije u školskoj 2012/13. godini. Od 40 upisanih, 20 studenata je izabralo da sluša kurs Arhitektura i organizacija računara 1.

Cilj opisanog istraživanja je usmeren na izbor klasifikacijskih algoritama mašinskog učenja koji daju što preciznije rezultate predviđanja na izuzetno malom obučavajućem skupu, podržavaju rad sa mešovitim podacima (kategorijskog i numeričkim tipa), nisu osetljivi na podatke koji su nepotpuni (*eng. missing attribute values*) ili se ne uklapaju u opšti model (*eng.outliers*). Poređenjem parametara za procenu efikasnosti, izdvojeni su algoritami sa tačnošću većom od 70%.

II. IZBOR KLASIFIKATORA

Klasifikacijam, jedan od najčešće proučavanih problema u oblasti Data Mininga (DM) i Mašinskog Učenja (ML), smatra se zadatkom nadgledanog učenja (*eng. supervised learning*). Skup podataka podeljen je u klase tj. svaka instanca skupa ima oznaku koja identifikuje klasu kojoj pripada. Algoritmi nadgledanog mašinskog učenja koriste se da podstaknu klasifikator iz skupa pravilno klasifikovanih instanci odnosno skupa obučavanja. Skup testiranja, skup pravilno klasifikovanih instanci podataka, koristi se za merenje kvaliteta klasifikatora dobijenog nakon primenjenog procesa obučavanja. Različiti tipovi modela se mogu koristiti da

[1] Ovaj rad je realizovan pod delimičnim pokroviteljstvom Ministarstva zaobrazovanje i nauku Republike Srbije, projekati br. III47016

predstave klasifikatore, a postoji i veliki broj algoritama dostupnih za indukovanje klasifikatora iz podataka.

U slučaju izbora klasifikatora za obrazovne skupove podataka, situacija je veoma jasna. Obrazovni skupovi podataka su veoma mali; postupak izvođenja faze eksperimentisanja je teži zbog činjenice da su podaci dinamički promenljivi. Obrazovni podaci su većinom numeričkog i kategorijskog tipa, a s obzirom da se izdvajaju iz baza podataka zahtevaju manje čišćenja u fazi pretprocesiranja. U slučaju ove vrste podataka česta je pojava nepotpunih podataka i podataka koji se ne uklapaju u opšti model.

Neke od primena različitih klasifikatora, na podatke izdvojene iz obrazovnog okruženja, opisane su u radovima koji su navedeni u okviru navedenih referenci. U radu [8] autori su koristili neuronske mreže sa RBF funkcijom za predviđanje konačne ocene studenata (položio/pao) na osnovu zapisa u Moodle log fajlovima. Analizirani skup podataka je sadržao 240 redova pri čemu je svaki red predstavljao zapis za jednog studenta. Informacije su kombinovane sa konačnim ocenama studenata. Razvijen je model za predviđanje koji će od studenata uspešno da završe kurs. Bayesian mreže korišćene su u slučaju modelovanja znanja i predviđanje performansi studenata u okviru tutorskog sistema za podučavanje [9]. Za predviđanje učinka studenata u e-learning okruženju autori [10] koriste fuzzy pravila udruživanja. U [11] prikazana je upotreba stabla odlučivanja za predviđanje učinka studenata u e-learning sistemima. Za predviđanje rezultata ispita na kursevima daljinskog učenja u [12] korišćena je tehnika linearne regresije. Primena tehnika linearne regresija, stabla odlučivanja, neuronskih mreže za predviđanje ocena studenata u programu daljinskog učenja na Univerzitetu Hellenic Open opisana je u radu [13]. U [14] opisan je model predviđanja uspešnosti studenata na prvom kolokvijumu uzimajući u obzir promenljive koje se odnose na aktivnosti studenata u okviru posmatranog kursa. Za kreiranje modela korišćeni su *Naive Bayes* i *Decision Trees* klasifikatori. Analiziran je uticaj ulaznih atributa na performanse modela kako bi se ostvarila veća tačnost generisanih modela.

U ovom radu izabrani su *Stabla odluke (J48), Pravila klasifikacije* (*OneR)* i *Bayesian klasifikatori (Naıve Bayes, BayesNet TAN)* za primenu na obučavajući skup podataka. Postupak analize izvršen je upotrebom Weka, *open- source data mining* alata [15].

Stabla odluke (*eng. Decission trees*) [16] predstavljaju skup uslova organizovanih u hijerarhijskoj strukturi od nula ili više unutrašnjih čvorova i jedan ili više listova čvorova. Svi unutrašnji čvorovi imaju dva ili više dete čvora, izraz (kriterijum) kojim se testiraju vrednosti atributa i izvršava dalja podela odnosno grananje stabla. Grane stabla, veze između unutrašnjih i njihove dece čvorova, obeležene su rezultatima testiranja. Svaki čvor list ima oznaku pripadajuće klase. Stablo odluke je prediktivni model u kome se instance klasifikuju tako što slede putanju zadovoljenih uslova od korena stabla pa do lista odgovarajuće klase. Neki od najpoznatijih algoritama stabla odluke su *ID3* [17], *C4.5, J48* [18]. *J48* predstavlja poboljšanje i unapređenje algoritma *C4.5*. Prednosti stabla odluke su: jednostavnost i lakoća razumevanja, mogućnost rada sa numeričkim i kategorijskim vrednostima, brzo razvrstavanje novih instanci, fleksibilnost, kao i mogućnost vizuelne reprezentacije.

OneR algoritam predstavlja klasifikator koji generiše pravila u formi stabla odluke sa jednim nivoom *(eng. 1-level decision tree*) [19]. Može da generiše više stabala odluke, gde svako stablo predstavlja test jednog određenog atributa. Na kraju se izdvaja stablo sa najmanjom greškom (*eng.error rate*). Greška predstavlja odnos primera koji ne pripadaju većinskoj klasi u granama stabla.

Bayesian klasifikatori pretpostavljaju da je znanje o nekom događaju opisano verovatnoćom pojave tog događaja. Statističke zavisnosti ovih klasifikatora predstavljaju se vizuelnom strukturom grafa [20]. U obrazovnom domenu pretpostavka o uslovnoj nezavisnosti se često ignoriše i narušava. S obzirom da su promenljive međusobno povezane, *Naıve Bayes* klasifikator može da toleriše snažne iznenađujuće zavisnosti između nezavisnih promenljivih. Smatra se da su *Naıve Bayes* klasifikatori nadmašili sofisticiranije klasifikatore kao što su *Stabla odluke* i opšte *Bayesian klasifikatore*, posebno u slučajevima skupova podataka sa manjim brojem zapisa [21]. *BayesNet TAN* klasifikator [22] predstavljaju proširenje *Naive Bayes* modela dajući i mogućnost dodatnih zavisnosti. Struktura *TAN* modela je ista kao i kod *Naıve Bayes* mreže jedino što čvorovi listova mogu biti međusobno zavisni, pored zavisnosti prema klasnoj promenljivoj. *TAN* model je često dobar kompromis između *Naıve Bayes* i opšte *Baiesove* mreže: struktura modela je dovoljno jednostavna da se izbegne problem detaljisanja (*eng. overfitting*), ali bi trebalo uzeti u obzir postojanje jake zavisnosti među atributima.

III. STUDIJA SLUČAJA

U zimskom semestru školske 2012/13. godine na Visokoj školi elektrotehnike i računarstva strukovnih studija u Beogradu, održan je on-line kurs (predavanja i laboratorijske vežbe) Arhitektura i organizacija računara 1 u okviru programa daljinskog učenja. Kurs je kreiran na *Moodle open-source LMS* (*eng. Learning Management System*) platformi [23] za koncept daljinskog učenja.

Iz baze podataka *LMS Moodle* sistema, izdvojen je obučavajući skup od 20 zapisa. Aktivnost studenata na kursu merena je pomoću sledećih atributa:

- ukupno provedeno vreme u okviru kursa,
- broj ostvarenih sesija,
- prosečan broj sesija nedeljno,
- prosečno vreme u okviru jedne sedmice,
- prosečno vreme po sesiji,
- broj pročitanih poruka sa foruma,
- broj poslatih poruka na forum
- bodovi ostvareni na testovima za laboratorijske vežbe,
- bodovi ostvareni na domaćim zadacima,
- prosečni bodovi ostvareni na testovima za samostalnu proveru znanja,

- bodovi ostvareni na prvom i drugom kolokvijumu,
- bodovi ostvareni na završnom ispitu
- konačna ocena.

Primenom odgovarajućih metoda diskretizacije [24], atributi numeričkog tipa vrednosti transformisane su u nominalne (kategorijske). Numerički domen vrednosti atributa *konačna ocena (5,6,7,8,9,10)* diskretizovan je ručnom metodom na tri kategorijske vrednosti (*pali, polozili, odlicni)* i to:

- *ocena 5 - pali*
- *ocene 6,7 - polozili*
- *ocene 8, 9, 10 - odlicni*

Atribut *konačna ocena* je označen kao klasni atribut odnosno promenljiva predviđanja.

IV. POREĐENJE ALGORITAMA KLASIFIKACIJE

Za izabrane klasifikatorske algoritme postavljeni su podrazumevani parametri. Testiranje je izvršeno metodom unakrsne validacije poznate i kao rotaciona estimacija. Ovom metodom izvršena je slučajna podela skupa instanci na *k=5* međusobno isključivih podskupova približno iste veličine. Postupak ocenjivanja se ponavljao 2 puta, svaki put koristeći različit podskup kao testni skup.

Za poređenje klasifikatora razmatrani su sledeći parametri [14]:

- *TP (eng.True Positive)* predstavlja broj tačnih pozitivnih predikcija (stopa pozitivno klasifikovanih instanci)
- *FP (eng. False Postive)* predstavlja broj pogrešnih pozitivnih predikcija (stopa negativno klasifikovanih instanci)
- *Preciznost (eng. Precision)* izračunava se po formuli Pr= TP/(TP+FP) (tačnost klasifikacije).

Rezulati primenjenih algoritama prikazani su u Tabeli 1.

TABELA I. TAČNOST KLASIFIKATORA, STOPE POZITVNO I NEGATIVNO KLASIFIKOVANIH INSTANCI

Algoritmi	*TP*	*FP*	*Tačnost (%)*
BayesNet TAN	*0.622*	*0.345*	*62.22*
J48*	*0.804*	*0.187*	*80.44*
Naive Bayes*	*0.814*	*0.196*	*81.44*
OneR	*0.511*	*0.266*	*51.11*

Iz Tabele 1 vidi se da su *Naive Bayes* i *J48* algoritmi sa najvećom tačnošću. Mali obučavajući skup, a veći broj atributa uslovio je da *Naive Bayes* algoritam generiše klasifikatorski model najveće tačnosti uprkos postojanju nepotpunih podataka. Rezultati ove studije ukazuju na činjenicu da je *Naive Bayes* algoritam pogodan za obučavajuće skupove manje od 50 instanci. Međutim, pored *Naive Bayesa* ne treba zanemariti rezulate *J48* algoritma koji je generisao klasifikatorski model sa neznatno manjom preciznošću. Dobri rezultati *J48* algoritma mogu se povezati sa kategorijskim tipom atributa i sa diskretizacijom klasnog atributa na 3 klasne oznake.

V. ZAKLJUČAK

Dobar izbor klasifikatora, naročito za izuzetno mali obučavajući skup (što je i slučaj u ovoj studiji), je od izuzetne važnosti. Prvo, pre same implementacije odabranog klasifikatorskog algoritma treba znati kakvi su podaci izdvojeni za analizu. Obrazovni podaci su poprilično čisti, ne sadrže pogrešne vrednosti u velikoj meri, jer se sakupljaju iz baza podataka, a vrednosti ovih podataka mogu biti numeričkog, kategorijskog tipa.

U ovoj studiji slučaja utvrđeno je da algoritmi *Naive Bayes* i *J48* generišu precizne klasifikatorske modele na izuzetno malom obučavajućem skupu sa tačnošću predviđanja većom od 70%. U slučaju malog obučavajućeg skupa (manje od 50 instanci) i atributa numeričkog tipa, *Naive Bayes* se pokazao kao dobar izbor. Na osnovu naše analize, za *J48* algoritam može se zaključiti da u slučaju malog obučavajućeg skupa podataka daje dobre rezultate predviđanja ako su podaci kategorijskog tipa, a klasni atribut može da ima jednu od tri klasne vrednosti (*pali, polozili, odlicni*). *OneR* algoritam nije dao značajne rezultate, a *BayesNet TAN* bi bio bolja alternativa za veći obučavajući skup.

Za nastavak istraživanja planirano je proširenje eksperimenta sa drugim skupovima obrazovnih podataka programa za daljinsko učenje kako bi se navedeni zaključci proverili i utvrdili. U cilju povećavanja tačnosti klasifikatorskih modela malih obučavajućih skupova predstoji i proučavanje meta - algoritama koji koriste metod kombinovanja najpreciznijih klasifikatora.

LITERATURA

[1] K. Patriarcheas, M. Xenos, "Modelling of distance education forum: formal languages as interpretation methodology of messages in asynchronous text based discussion", Computers and Education, Vol.52, pp. 438–448, 2009.

[2] T. Tooth, "The Use of Multimedia in Distance Education", Vancouver Communication of Learning, Vancouver, 2000.

[3] K. Stevenson, P. Sander, P. Naylor, "Student perceptions of the tutor's role in distance learning", Open Learning 11(1), pp. 22–30, 1996.

[4] K. Stevenson, P. Sander, "How do Open University students expect to be taught at tutorials?", Open Learning 13 (2), pp. 42–46, 1998.

[5] U. Fayyad, G.Piatetsky-Shapiro, P.Smyth, "From Data Mining to Knowledge Discovery in Databases", American Association for Artificial Intelligence, 17, pp.37-54, 1996.

[6] C. Romero, S. Ventura, "Educational Data Mining: a Survey from 1995 to 2005", Expert Systems with Applications, 33(1), pp.135-146, 2007.

[7] D. Prokin, G. Dimic, K. Kuk, M.Prokin, "Moodle kao platforma za realizaciju nastavnih aktivnosti iz predmeta Arhitektura i organizacija računara 1", INFOTEH-JAHORINA Vol. 11, pp.857–862, March 2012.

[8] M.Delgado, E.Gibaja, M.C.Pegalajar, O.Pérez, "Predicting Students' Marks from. Moodle Logs using Neural Network Models", In International Conference on Current Developments in Technology-Assisted Education, Sevilla, Spain, pp. 586-590, 2006.

[9] Z.Pardos, N.Heffernan, B. Anderson, C.Heffernan, "The Effect of Model Granularity on Student Performance Prediction Using Bayesian Networks", In International Conference on User Modeling, Corfu, Greece, pp.435-439, 2007.

[10] A.Nebot, F.Castro, A.Vellido, F. Mugica, "Identification of fuzzy models to predict students perfornance in an e-learning environment", In International Conference on Web-based Education, Puerto Vallarta, pp.74-79, 2006.

[11] C.C. Chan, "A Framework for Assessing Usage of Web-Based e-Learning Systems", In International Conference on innovative Computing, Information and Control, Washington, DC, pp. 147- 151, 2007.

[12] N. Myller, J.Suhonen, E.Sutinen, "Using Data Mining for Improving Web-Based Course Design", In International Conference on Computers in Education, Washington, pp.959- 964, 2002.

[13] S.B.Kotsiantis, P.E.Pintelas, "Predicting Students' Marks in Hellenic Open University", In IEEE international Conference on Advanced Learning Technologies, Washington, DC, pp.664-668, 2005.

[14] G. Dimic, D. Prokin, K. Kuk, M.Micalovic, "Primena Decision Trees i Naive Bayes klasifikatora na skup podataka izdvojen iz Moodle kursa", INFOTEH-JAHORINA Vol. 11, pp.878–882, March 2012.

[15] Weka, University of Waikato, New Zealand, http://www.cs.waikato.ac.nz/ml/weka

[16] H. I. Witten, F. Eibe, "Data mining: Practical machine learning tools and techniques", Second edition, San Francisco: Morgan Kaufmann, chapter 6.1, 2005.

[17] J. R. Quinlan, "Induction of decision trees", Machine Learning, 1:81-106, 1986.

[18] J. R. Quinlan, "C4.5:Programs for Machine Learning", San Mateo, CA:Morgan Kaufmann, 1993.

[19] R.C. Holte, "Very simple classification rules perform well on most commonly used datasets", Machine Learning. 11:63-91., 1993

[20] J. Pearl, "Probabilistic reasoning in intelligent systems: networks of plausible inference", San Mateo, California: Morgan Kaufman Publishers, 1988.

[21] P. Domingos, M. Pazzani, " On the optimality of the simple Bayesian classifier under zero-one loss", *Machine Learning*, 29:103–130, 1997.

[22] N. Friedman, D. Geiger, M. Goldszmid, "Bayesian network classifiers", Machine Learning, 29(2-3):131–163, 1997.

[23] Moodle, http://moodle.org/

[24] G. Dimic, D. Prokin, K. Kuk, P. Spalevic, " The Use of Data Mining Methods for Analyzing and Evaluating Course Quality in the Moodle System", Международна научна конференция "УНИТЕХ'10" – Габрово, P. 309-315, 2010.

ABSTRACT

This paper describes the selection procedure accurate classifiers for small training dataset. We have used data course Computer architecture and organization realized in the distance learning program at the School of Electrical Engineering and Computer Science Applied Studies in Belgrade. For testing four classifiers (OneR, J48, Naive Bayes, BayesNet TAN) we have used the method of cross-validation. Ours results showed that Naive Bayes and the J48 classification algorithm generated model with a precision better than 70%.

SELECTION OF CLASSIFIERS FOR A SMALL EDUCATIONAL TRAINING DATASET

Gabrijela Dimić, Dragana Prokin,
Kristijan Kuk, Boško Bogojević