Poređenje metoda mašinskog učenja za predikciju trenda promene finansijskih vremenskih serija

Miloš Stojanović
Visoka tehnička škola strukovnih studija Niš
Niš, Republika Srbija
milos.stojanovic@vtsnis.edu.rs

Stevica Cvetković, Goran Stančić
Elektronski fakultet, Univerzitet u Nišu
Niš, Republika Srbija
stevica.cvetkovic@elfak.ni.ac.rs
goran.stancic@elfak.ni.ac.rs

Sažetak— **U ovom radu je predstavljeno poređenje najčešće korišćenih tehnika mašinskog učenja za predikciju trenda promene finansijskih vremenskih serija. S obzirom na to da je cilj ovog rada predviđanje tendencija u promenama vrednosti indeksa, u radu su korišćeni isključivo tehnički indikatori za formiranje predikcionog modela. Upoređene su najčešće korišćene metode mašinskog učenja: SVM (linearni i RBF), Neuronske mreže i Random forest metoda. Uporedni rezultati testiranja na primeru Belex15 indeksa Beogradske berze su pokazali da najveću tačnost predikcije postižu nelinearna SVM metode kao i metoda Random Forest.**

Ključne riječi - predviđanje trenda; vremenske serije; mašinsko učenje; binarna klasifikacija

I. UVOD

Matematičko modeliranje finansijskih vremenskih serija posebno dobija na značaju poslednjih godina. Empirijski podaci koji se koriste za analizu finansijskog tržišta evidentiraju se u fiksnim vremenskim intervalima, uobičajeno na dnevnom nivou. Kako se količina podataka, koji se koriste u analizama ubrzano povećava, povećava se i potreba za razvojem odgovarajućih matematičkih alata za razumevanje i predikciju kretanja finansijskog tržišta. Jer, ukoliko bi se tržišni trendovi mogli predvideti preciznije, obezbedili bi se neophodni uslovi za maksimiranje prinosa na investicije u finansijsku aktivu.

Za razvoj efikasne strategije trgovanja, od posebnog značaja su precizna predviđanja kretanja indeksa cena akcija [1], [2]. Većina trgovinskih praksi usvojenih od strane finansijskih analitičara se oslanja na precizna predviđanja cene finansijskih instrumenata. Međutim, novije studije, predložene u [3], su pokazale da strategije trgovanja na osnovu predviđanja pravca tj. trenda promene cena mogu biti efikasnije i generisati veći prinos na investicije.

Predviđanja kretanja indeksa cena akcija je jedan od najvažnijih problema u oblasti analize finansijskih vremenskih serija.[2]. Namena indeksa je da meri promene cena akcija kojima se trguje metodom kontinuiranog trgovanja, a koje su prethodno zadovoljile kriterijum za uključivanje u indeksnu korpu. Belex 15 je vodeći indeks Beogradske berze, čiju vrednost određuju cene najlikvidnijih akcija, kojima se trguje na regulisanom tržištu Beogradske berze.

Pošto se ponašanje finansijskih sistema veoma dinamično menja u vremenu, čak i u slučajevima kada je moguće projektovati softver koji bi vršio predikciju, taj softver bi povremeno morao da se ažurira ili čak projektuje ponovo. Uz pomoć algoritama mašinskog učenja, moguća je stalno adaptiranje skupa parametara predikcionog modela kako bi na adekvatan način mogle da se proprate promene u ponašanju posmatranog sistema. Iz tog razloga u mnogim studijama algoritmi mašinskog učenja pokazali su se veoma efikasnim u predviđanju trenda berzanskih indeksa i na taj način doprineli uvećanju prinosa i smanjenju rizika trgovanja. Među najčešće korišćenim algoritmima mašinkog učenja za predviđanja u oblasti finansija, spadaju veštačke neuronske mreže (eng. Artificial Neural Networks - ANN) [2], [5], [6] i metode potpornih vektora (eng. Support Vector Machine - SVM) [1], [4], [7]. Osim ovih metoda, tehnika Random Forest (RF) [3], [8] je pokazala visok stepen tačnosti predviđanja prilikom primene na širokom spektru problema.

Iako mnoga istraživanja ukazuju da promene cena akcija nisu potpuno nasumične, posmatrano u dužem vremenskom intervalu promena cena može da se aproksimira slučajnim procesom (eng. random walk). Stoga se stepen preciznosti od oko 60% koji se dobija korišćenjem metoda mašinskog učenja često opisuje kao zadovoljvajući za predviđanja na tržištima kapitala [9].

S obzirom na to da je cilj ovog rada predviđanje tendencija u promenama vrednosti indeksa, u radu su korišćeni isključivo tehnički indikatori za formiranje predikcionog modela. Problem predviđanja promene trenda berzanskog indeksa se u praksi modeluje kao problem binarne klasifikacije. U prethodnim radovima autora [10], [11] postavljene su osnove za predviđanje promena trenda vrednosti Belex15 indeksa, tako da ovaj rad predstavlja dalja istraživanja i nastavak razvoja prethodnih modela.

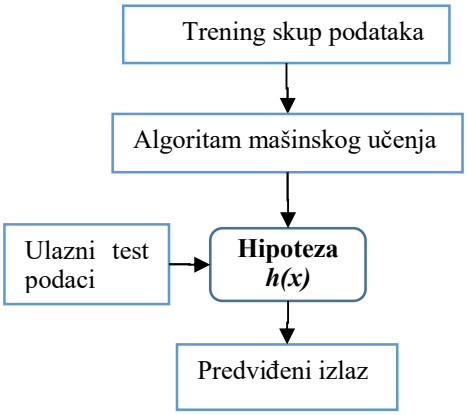
U nastavku rada je prvo dat kratak pregled tehnika mašinskog učenja koje su korišćene u radu. Zatim je opisan postupak za formiranje predikcionog modela, tj opisani su atributi finansijskih vremenskih serija koji su korišćeni za kreiranje vektora obeležja u testiranim metodama mašinskog

učenja. Nakon predstavljanja uporednih rezultata testiranja, na kraju su istaknuti pojedini zaključci i predstavljeni pravci budućih istraživanja.

II. PREGLED TEHNIKA MAŠINSKOG UČENJA

Mašinsko učenje razmatra iste oblasti istraživanja kao statistika i *data-mining*, ali sa drugačijeg aspekta. Statistika se fokusira na razumevanje procesa koji generišu podatke, često sa ciljem testiranja različitih hipoteza o tim procesima. *Data-mining* tehnike nastoje da pronađu šablone u podacima koji su razumljivi ljudima. Nasuprot tome, mašinsko učenje se prvenstveno bavi tačnošću i efikasnošću rezultujućeg sistema.

Zadatak algoritama nadgledanog mašinskog učenja je da na osnovu zadatog trening skupa S "nauče" funkciju predviđanja (tj. treniraju model) $h(x)$, tako da $h(x)$ bude optimalna aproksimacija za odgovarajuće vrednosti ciljnih promenljivih y, slika 1. Formalno definisano, $h \in H$ predstavlja jednu od hipoteza o funkciji koja treba da bude naučena na osnovu zadatog trening skupa S, gde H predstavlja konačan (u nekim slučajevima i beskonačan) skup hipoteza.



Slika 1. *Postupak mašinskog učenja*

Predikcioni model se formira pomoću trening algoritma, na osnovu izabranog trening skupa, koji se sastoji od atributa i vektora. U slučaju kada su promenljive koje je potrebno predvideti kontinualne, tada se problem definiše kao regresioni. Ukoliko predviđene vrednosti mogu da sadrže samo ograničen skup diskretnih vrednosti, tada se problem definiše kao klasifikacioni.

Zbog postignutih rezulatata u različitim oblastima primene, veštačke neuronske mreže spadaju u jedan od najpoznatijih i najkorišćenijih algoritama mašinskog učenja u poslednje dve decenije. Zasnivaju se na principu empirijske minimizacije rizika tako da minimizuju greške na trening primerima, što u nekim situacijama može dovesti do *overfitting-a*. Pored toga, optimizacioni problem koji se rešava kod ANN nije konveksan, tako da ne mora uvek doći do pronalaska optimalnog rešenja.

SVM tehnika mašinskog učenja rešava probleme nelinearne klasifikacije i regresije metodama konveksnog kvadratnog programiranja (*convex quadratic programming* - QP). Pored toga što SVM uvek pronalazi optimalno rešenje QP problema, u njemu figurišu samo primeri iz trening skupa koji mu najviše doprinose, takozvani potporni vektori (*support vectors*), odnosno formira se *sparse solution*. Čak i veličina QP problema ne zavisi od dimenzije ulaznog prostora, već samo od broja trening primera. Optimizacioni problem koji definiše SVM se zasniva na strukturnoj minimizaciji rizika, tako da se minimizuje gornja granica greške generalizacije, što u određenim situacijama može biti prednost u odnosu na ANN.

Za razliku od ANN i SVM, koji pripadaju kategoriji "individualnih" algoritama nadgledanog mašinskog učenja, Random Forest spada u klasu ansambl metoda (eng. *ensemble methods*), koje na određeni način kombinuju rezultate više pojedinačnih metoda. Ovakav prisup ima za cilj dobijanje boljih rezultata predikcije u odnosu na bilo koji od individualnih metoda od kojih se sastoji endambl. Konkretno, kod RF formira se ensambl sačinjen od nekoliko stotina do nekoliko hiljada stabala odlučivanja (eng. *decision trees*). Prednosti stabala odlučivanja u odnosu na ostale metode mašinskog učenja (npr. ANN i SVM), su njihova jednostavnost implementacije i razumljivost postupka. Postoje pravila po kojima se stabla brzo i jednostavno formiraju a krajnji rezultat se može lako interpretirati. Pored toga, kod stabala odlučivanja atributi mogu imati i nepoznate vrednosti (*missing values*), što nije slučaj sa ANN i SVM. Međutim, jedan od nedostataka metode stabala odlučivanja je njihova nestabilnost. Mala promena ulaznih trening podataka može dovesti do velike promene topologije stabla. Nestabilnost se javlja zbog velikog broja mogućih podela koje često imaju približno isti značaj (eng. *competitor splits*). Zbog toga mala promena podataka može dovesti do sasvim drugačije podele, koja dalje unosi promene u sve grane stabla ispod sebe. RF prevazilazi ova ograničenja agregacijom rezultata predviđanja više stotina individualnih stabala.

III. FORMIRANJE MODELA

U ovom odeljku dat je opis atributa finansijskih vremenskih serija koji su korišćeni za kreiranje vektora obeležja u testiranim metodama mašinskog učenja.

S obzirom na to da je cilj ovog rada predviđanje tendencija u promenama vrednosti indeksa, u radu se oslanjamo na najčešće korišćene tehničke indikatore: EMA (engl. Exponential Moving Average – eksponencijalni pokretni prosek vrednosti indeksa na zatvaranju), RSI (engl. Relative Strength Index – indeks koji meri brzinu i promenu kretanja vrednosti indeksa), WILLR (engl. Williams %R – indikator kojim se predviđa tačka obrta poređenjem vrednosti indeksa na zatvaranju i vrednosti tokom određenog vremenskog perioda), MACD (engl. Moving Average Convergence-Divergence – indicator koji meri jačinu i pravac trenda), ROC (engl. Rate of Change – indicator koji pokazuje procentualnu promenu vrednosti indeksa na zatvaranju), CCI (engl. Commodity Channel Index – indikator koji se koristi za otkrivanje cikličnih promena vrednosti indeksa merenjem odstupanja vrednosti od njene statističke sredine), i SAR (engl. Parabolic Stop and Reverse – indikator koji otkriva pravac trenda promene vrednosti indeksa i koristi se za određivanje momenata za trgovanje). U Tabeli I je dat pregled tehničkih indikatora korišćeni za kreiranje predikcionih

modela, zajedno sa formulama za izračunavanje njihovih vrednosti.

TABELA I. TEHNIČKI INDIKATORI

Tehnički indikatori	Matematičke formule
CP_t - *Vrednost indeksa, t= 1,2, ... n,* LP_N / HP_N – *Najniža/najviša dnevna vrednost indeksa u poslednjih N dana trgovanja*	
EMA	$EMA_t = CP_t * k + EMA_{t-1} * (1-k), k = 2/(N+1)$
MACD	$MACD_t = EMA_{12,t} - EMA_{26,t}$
RSI	$RSI_t = 100 - \frac{100}{1+RS_t}$ $RS_t = \frac{\sum_{n=t}^{t-d} \max(0, CP_n - CP_{n-1})}{\sum_{n=t}^{t-d} \lvert\min(0, CP_n - CP_{n-1})\rvert}$
CCI	$CCI_t = (M_t - SM_t)/0.0015D_t$ $M_t = HP_t + CP_t + LP_t$ $SM_t = \sum_{i=t-m+i}^{t} M_i / m$ $D_t = \sum_{i=t-m+i}^{t} \lvert M_i - SM_t \rvert / m$
WILLR	$WILLR = \frac{HP_n - CP_t}{HP_n - LP_n} * 100$
SAR	$SAR_{t+1} = SAR_t + \alpha(EP - SAR_t)$ EP – ekstremna vrednost α – faktor ubrzanja
ROC	$ROC_t = 100((CP_t - CP_{t-n})/CP_{t-n})$

Problem predviđanja pravca promene vrednosti berzanskog indeksa se u literaturi uobičajeno modeluje kao problem binarne klasifikacije, pri čemu se klase obeležavaju sa „*-1”* i „*1”*. Pravac (tj. trend) je, dakle, kategorička promenljiva koja u našem eksperimentu prikazuje pravac kretanja Belex15 indeksa u bilo kom trenutku vremena *t*. U ovom radu je kao indikator promene korišćena vrednost indeksa (engl. Closed price – CP). Klasa „1” označava da je vrednost indeksa za neredni dan trgovanja veća od vrednost indeksa na dan trgovanja *t*, dok klasa „-1” indikuje da je vrednost indeksa za naredni dan trgovanja manja od tekuće vrednosti indeksa.

IV. EVALUACIJA

Realni finanskijski podaci koji su korišćeni u ovom radu, preuzeti su sa zvaničnog sajta Beorgadske berze (www.belex.rs). Vremenska serija obuhvata zapise od oktobra 2005. do decembra 2015. godine, u ukupnom broju od 2549 zapisa tj. trgovinskih dana. Raspoloživi podaci su podeljeni u dva podskupa, za treniranje i testiranje. Za treniranje modela je korišćeno 2297 zapisa iz perioda od oktobra 2005. do decembra 2014. godine, dok je za testiranje korišćeno 252 zapisa.

Sve metode za izračunavanje atributa, kao i tehnike mašinskog učenja su implementirane u Matlab programskom paketu, uz korišćenje dodatnih biblioteka funkcija [12] za implementaciju određenih tehnika mašinskog učenja.

Neuronska mreža koja je korišćena za evaluaciju, sastoji se iz 3 sloja (ulazni, skriveni i izlazni sloj). Ulazni sloj se sastoji od 6 neurona, dok izlazni sloj ima samo jedan neuron sa sigmoidalnom tansfer funkcijom. Parametri linearnog SVM modela, kao i kernelizovanog SVM-RBF (Radial Basis Function) modela, određeni su procesom 10 unakrsnih validacija (engl. 10-fold cross validation), uz pomoć grid-search tehnike na trening skupu podataka. Za tehniku Random Forest, korišćen je Matlab funkcija *fitrensemble()*, pri čemu se ansambl sastoji od 200 stabala.

Kao mera za procenu performansi predviđanja na test skupu korišćena je tačnost predviđanja (engl. Hit Ratio - HR), koja se izračunava na osnovu broja pravilno predviđenih (tj. klasifikovanih) rezultata u okviru test skupa:

$$HR = \frac{1}{m}\sum_{i=1}^{m} PR_i \qquad PR_i = \begin{cases} 1 & PI_i = AI_i \\ 0 & PI_i \neq AI_i \end{cases}$$

Gde je PR_i predikcioni rezultat i-tog trgovinskog dana, pri čemu je AIi aktuelni izlaz za *i*-ti trening dan i PI_i je predviđena vrednost za i-ti dan trgovanja, dok je *m* broj podatka u test skupu [3]. Nakon trening faze, uspešnost modela je proverena nad test podacima.

TABELA II. UPOREDNA ANALIZA TAČNOSTI PREDIKCIJE TRENDA BERZANSKOG INDEKSA NA PRIMERU BELEX15 INDEKSA

Metoda	Tačnost predikcije (%)
Neuronske mreže	53.96
SVM-linearni	53.97
SVM-RBF	57.14
Random Forest	56.74

Na osnovu rezultata prikazanih u prethodnoj tabeli, može se zaključiti da, prema očekivanjima, i u skladu sa rezultatima prethodno predstavljenim u literaturi, nelinearna SVM-RBF metoda postiže najbolje rezultate. Koristan rezultat poređenja je da metoda Random Forest postiže izuzetno visoku tačnost u predikciji trenda finansijskih serija, iako nije često korišćena metoda za ovu namenu.

V. ZAKLJUČAK

Uporednom analizom modela pokazuje se visok stepen tačnosti nelinearne SVM metode i tehnike Random Forest. I ostale metode su postigle dobre rezultate ako se uzme u obzir da se u ovom radu ispituje stepen predvidljivosti kretanja indeksa cena akcija na tržištu u razvoju, kao što je tržište kapitala Republike Srbije, nasuprot većini radova iz ove oblasti, koji se bave predviđanjem indeksa cena akcija na razvijenim tržištima.

U narednom istraživanja bi se moglo definisati više pravaca koji bi mogli dovesti do poboljšanja preciznosti modela. Od interesa je formiranje takozvanih kombinovanih modela predviđanja, gde bi se izlazi iz više modela kombinovali u finalnom modelu. Na kraju, važno je napomenuti da se u predstavljenom modelu predviđanja, svako povećanje tačnosti od minimum jednog procenta u odnosu na aktuelne rezultate smatra relevantnim doprinosom, pošto se na tržištu preslikava u visoke novčane dobitke [7].

ZAHVALNICA

Ovaj rad je podržan od strane Ministarstva prosvete, nauke i tehnološkog razvoja Republike Srbije, projekat broj TR33035.

LITERATURA

[1] W. Huang; Y. Nakamori and SY. Wang, "Forecasting stock market movement direction with support vector machine," Computers & Operations Research vol. 32, 2005, pp. 2513–2522

[2] Y. Kara, M. A. Boyacioglu b, Ö. K. Baykan, „Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange",Expert systems with Applications, vol. 38, no. 5, 2011, pp. 5311-5319

[3] M. Kumar, M. Thenmozhi, "Forecasting stock index movement: a comparasion of support vector machines and random forest" Indian Institute of Capital Markets 9th Capital Markets Conference Paper, 2006, available at SSRN: http://ssrn.com/abstract=876544 or http://dx.doi.org/10.2139/ssrn.876544, 2006.

[4] O. Phichhang, H. Wang, "Prediction of Stock Market Index Movement by Ten Data Mining Techniques" Modern Applied Science, Vol. 3, No. 12, 2009, pp. 28-42.

[5] W. Dai, J-Y. Wu, & C-J. Lu, (2012). Combining nonlinear independent component analysis and neural network for the prediction of Asian stock market indexes. *Expert Systems with Applications*, 39(4), 4444–4452. doi:10.1016/j.eswa.2011.09.145

[6] Wei, L-Y. & Cheng, C-H. (2012). Hybrid recurrent neural networks model based on synthesis features to forecast the Taiwan stock market. *International Journal of Innovative Computing, Information and Control*, 8 (8), 5559-5571.

[7] L. Yuling, H. Guo and J. Hu, "An SVM-based Approach for Stock Market Trend Prediction", Neural Networks (IJCNN), The 2013 International Joint Conference on. IEEE, 2013, pp. 1-7.

[8] L. Breiman. "Random Forests". Machine Learning. vol. 45 (1): 5–32, 2001. doi:10.1023/A:1010933404324.

[9] S. Lahmiri, "A Comparison of PNN and SVM for Stock Market Trend Prediction using Economic and Technical Information", International Journal of Computer Applications, vol. 29, no.3, 2011.

[10] I. Marković, J. Stanković, M. Stojanović, M. Božić, „Predviđanje promene trenda vrednosti berzanskog indeksa Belex15 pomoću LS-SVM klasifikatora, Zbornik radova Međunarodne konferencije *INFOTEH Jahorina 2014*, Bosna i Hercegovina, 2014a, 739–782.

[11] I. Marković, J. Stanković, M. Stojanović, M. Božić, „Prediction of the stock market trend using LS-SVMs based on technical indicators", Zbornik radova Međunarodne konferencije *ICEST2014*, Republika Srbija, 2014b, 61–64.

[12] C. C. Chang, C. J. Lin, "LibSVM: a library for support vector machines," [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm/

ABSTRACT

We presented a comparative evaluation of commonly used machine learning methods for trend prediction of financial time series. Since the aim of this paper is the trend prediction of changes of the stock market index values, only the technical indicators for the formation of the prediction model were used. Following machine learning techniques are compared: SVM (linear and RBF), Neural Networks and Random Forest. Comparison of accuracy demonstrated that the SVM-RBF and Random Forest are the most appropriate methods in terms of accuracy for stock market trend prediction.

COMPARISON OF MACHINE LEARNING METHODS FOR TREND PREDICTION OF FINANCIAL TIME SERIES

Miloš Stojanović, Stevica Cvetković, Goran Stančić