Primjena stroja s potpornim vektorima za predviđanje kretanja na tržištima vrijednosnica

Hrga, Ingrid

Master's thesis / Diplomski rad

2015

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Pula / Sveučilište Jurja Dobrile u Puli**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:137:521520

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2024-03-28**

Image not found or type unknown

Repository / Repozitorij:

Digital Repository Juraj Dobrila University of Pula

zir.nsk.hr

Image not found or type unknown

Sveučilište Jurja Dobrile u Puli
Fakultet ekonomije i turizma
«Dr. Mijo Mirković»

INGRID HRGA

PRIMJENA STROJA S POTPORNIM VEKTORIMA ZA PREDVIĐANJE KRETANJA NA TRŽIŠTIMA VRIJEDNOSNICA

Diplomski rad

Pula, 2015.

Sveučilište Jurja Dobrile u Puli
Fakultet ekonomije i turizma
«Dr. Mijo Mirković»

INGRID HRGA

PRIMJENA STROJA S POTPORNIM VEKTORIMA ZA PREDVIĐANJE KRETANJA NA TRŽIŠTIMA VRIJEDNOSNICA

Diplomski rad

JMBAG: 0067212898, izvanredni student
Studijski smjer: Poslovna informatika

Predmet: Softversko inžinjerstvo
Mentor: doc. dr. sc. Krunoslav Puljić

Pula, rujan 2015.

IZJAVA O AKADEMSKOJ ČESTITOSTI

Ja, dolje potpisana Ingrid Hrga, kandidatkinja za magistru ekonomije ovime izjavljujem da je ovaj Diplomski rad rezultat isključivo mojega vlastitog rada, da se temelji na mojim istraživanjima te da se oslanja na objavljenu literaturu kao što to pokazuju korištene bilješke i bibliografija. Izjavljujem da niti jedan dio Diplomskog rada nije napisan na nedozvoljen način, odnosno da je prepisan iz kojega necitiranog rada, te da ikoji dio rada krši bilo čija autorska prava. Izjavljujem, također, da nijedan dio rada nije iskorišten za koji drugi rad pri bilo kojoj drugoj visokoškolskoj, znanstvenoj ili radnoj ustanovi.

Student:

U Puli, 16. 09. 2015. _____

SADRŽAJ

1. UVOD....1

2. TEORIJA UČENJA....4

2.1. Komponente problema učenja....4

2.2. Rizik....6

2.3. Empirijski rizik i princip minimizacije empirijskog rizika....6

2.4. Kompleksnost skupa hipoteza....7

2.5. Statistička teorija učenja i VC dimenzija....8

2.5.1. VC dimenzija....9

2.5.2. VC granica....10

2.6. Princip strukturne minimizacije rizika....11

3. STROJ S POTPORNIM VEKTORIMA....13

3.1. Sličnost....14

3.2. Klasifikator optimalne margine....15

3.2.1. Razdvajajuća hiperravnina....16

3.2.2. Optimalna hiperravnina....19

3.3. Metoda Lagrangeovih multiplikatora....20

3.4. Lagrangeova formulacija klasifikatora optimalne margine....22

3.5. Dualni problem optimalne hiperravnine....23

3.6. Veza između margine i VC dimenzije....25

3.7. Neseparabilni podaci....26

3.8. Linearni SVM s mekim marginama....27

3.9. Nelinearni SVM....30

3.9.1. Kernel trik....31

3.9.2. Kerneli....32

4. PREDVIĐANJE FINANCIJSKIH VREMENSKIH NIZOVA....35

4.1. Fundamentalna i tehnička analiza....36

4.1.1. Fundamentalna analiza.....36
4.1.2. Tehnička analiza.....37
4.2. Hipoteza efikasnog tržišta.....38
4.2.1. Predvidljivost prinosa.....39
4.2.2. Testovi i razine efikasnosti.....40
4.3. Predviđanje "gotovo" slučajnog niza.....41
5. SUSTAV ZA PREDVIĐANJE KRETANJA NA TRŽIŠTIMA VRIJEDNOSNICA.....43
5.1. Faze u procesu predviđanja.....43
5.2. Pretpostavke za korištenje aplikacije.....44
5.3. Funkcionalni zahtjevi.....44
5.4. Korištene tehnologije.....45
5.5. Shema aplikacije.....46
5.6. Sirovi podaci.....47
5.7. Priprema podataka za učenje.....48
5.7.1. Pretprocesiranje.....49
5.7.2. "Business time" pristup.....51
5.7.3. Testiranje efikasnosti.....51
Slučajni hod.....51
Autokorelacijska funkcija.....52
Parcijalna autokorelacijska funkcija.....53
Odnos varijanci.....53
5.7.4. Generiranje ulaznih i izlaznih varijabli.....54
Tehnički indikatori.....55
Izlazna varijabla.....60
Podjela podataka.....60
5.7.5. Odabir značajki.....62
Metode za odabir značajki.....63
Random forest.....65
5.7.6. Kreiranje LibSVM datoteka.....67
5.8. Odabir modela i učenje.....68
5.8.1. Grid-search pretraživanje.....69

5.8.2. Evaluacija klasifikatora tijekom odabira parametara....70
Unakrsna validacija....70
Metoda pomičnog prozora....71
5.8.3. Kompenzacija neravnoteže u podacima....73
5.9. Evalucija konačnog modela....76
5.9.1. Testiranje....76
5.9.2. Simulacija trgovanja....83
5.10. Korisničko sučelje i primjer upotrebe....85
5.10.1. Sučelje za pregled i podjelu podataka....85
5.10.2. Sučelje za testiranje efikasnosti....86
5.10.3. Sučelje za generiranje značajki....87
5.10.4. Sučelje za odabir značajki....88
5.10.5. Sučelje za odabir modela i treniranje....89
5.10.6. Sučelje za testiranje modela....92
5.10.7. Sučelje za simulaciju trgovanja....93
6. EKSPERIMENT I REZULTATI....95
6.1. Odabir burzovnih indeksa....96
6.2. Podaci....97
6.3. Prethodno testiranje i rezultati....97
6.4. Odabir značajki i rezultati....103
6.5. Rezultati treniranja i testiranja....105
6.5.1. Rezultati odabira modela i treniranja za različite kombinacije značajki....106
S&P500....107
CROBEXindustrija....109
SAX....111
6.5.2. Rezultati testiranja....114
CROBEXindustrija....114
SAX....120
S&P500....123
6.5.3. Ispitivanje utjecaja različitih kombinacija parametara na rezultat....124
CROBEXindustrija....125

SAX.....*128*

6.6. Simulacija trgovanja.....132

6.6.1. CROBEXindustrija.....*132*

6.6.2. SAX.....*135*

7. ZAKLJUČAK.....**138**

LITERATURA.....**140**

Knjige.....140

Članci.....141

Ostalo.....145

POPIS SLIKA.....**146**

POPIS TABLICA.....**151**

SAŽETAK.....**152**

SUMMARY.....**153**

1. UVOD

Ideja o stvaranju stroja dovoljno inteligentnog da čovjeka nadmaši u onome čime se najviše ponosi oduvijek je fascinirala ljude. Napori učinjeni u tom smjeru, s prvim priznatim rezultatima koji datiraju još iz 40-ih godina 20. stoljeća[1], doveli su do razvoja umjetne inteligencije kao znanstvene discipline. Cilj takvih napora bio je stvoriti strojeve koji će biti sposobni raditi ono što bi zahtijevalo inteligenciju kada bi to radili ljudi (Negnevitsky, 2005.). No, kako se ispostavilo, izuzetno je teško replicirati čak i najjednostavnije aspekte inteligentnog ponašanja. Ljudski mozak pokazuje izuzetnu sposobnost učenja iz iskustva, generaliziranja na temelju naučenih pravila, raspoznavanja uzoraka. Također, dobro koristi heuristiku i aproksimaciju, a uz sve to pokazuje i znatnu kreativnost. Naviknuti na takve sposobnosti, tek se pri pokušaju njihovog imitiranja uviđa koliko je ta ljudska nesavršenost zapravo kompleksna.

Zato, sve do kraja 1980-ih godina tehnologija, konceptualni alati, ali i pretpostavke na kojima su se temeljili nisu se pokazali prikladnima za suočavanje s navedenim izazovom. Stara škola umjetne inteligencije počivala je na simboličkoj manipulaciji i predikatnoj logici, što je stvaralo probleme s prikupljanjem i prikazom znanja, a inzistiranje na preciznosti, izvjesnosti i strogosti zahtijevalo je i mnogo računskog vremena. Međutim, ako se inteligencija definira kao sposobnost snalaženja u novim okolnostima, jasno je da stjecanje znanja o nekom prethodno nepoznatom konceptu i primjene naučenoga u još nepoznatim situacijama predstavlja njezin esencijalni dio. A ako se k tome doda još i to da se znanje stječe iz empirijskih podataka, nastaje jedan drugačiji pravac u stvaranju inteligentnih strojeva. Kako stvoriti stroj koji će biti sposoban rješavati zadatke bez da ga se eksplicitno programira, odnosno takav stroj koji će moći učiti, a onda i na temelju iskustva poboljšavati svoje performanse, postaje središnje pitanje strojnog učenja (Mitchell, 1997.).

Strojno učenje, kao jedno od područja umjetne inteligencije, oslanja se na koncepte i rezultate statistike, filozofije, teorije informacija, biologije, kognitivne znanosti, računske kompleksnosti i teorije kontrole (Mitchell, 1997.), a od 1990-ih godina, zahvaljujući dostupnosti sve veće snage računala i razvoju efikasnijih algoritama, pronalazi put uspješnoj

1 Prvi model umjetnog neurona stvoren je 1943. godine.

primjeni u raznim područjima, među kojima se posebno izdvaja svijet financija.

Financijski je sektor posljednjih nekoliko desetljeća dobio na važnosti više od bilo kojeg gospodarskog sektora. Unatoč periodičnim krizama i padovima, razina i raznovrsnost financijske aktivnosti konstantno se pojačava. Zahvaljujući otvorenosti prema tehnologijama te pod utjecajem inovacija, ubrzano se vrši transformacija tržišta kapitala koja se, od nekadašnjih tradicionalnih burzi smještenih na konkretnoj fizičkoj lokaciji, sve više pretvaraju u mreže računala. Pa iako je sam proces trgovanja danas znatno izmijenjen, njihove su temeljne funkcije i dalje jednake: osiguranje likvidnosti i formiranje cijena putem susretanja ponude i potražnje što zatim olakšava alokaciju kapitala te upravljanje rizikom. Međutim, unatoč tome, na ključno pitanje "kako pobijediti tržište?" još nitko nije dao konačan odgovor.

U nastojanju da se ovlada tržištem jednu mogućnost predstavlja usmjerenost na postizanje što veće brzine procesiranja i izvršavanja naloga, s obzirom da je od samih početaka jasno da se novac može zaraditi na komunikaciji niske latencije zbog informacijske prednosti koju ona pruža (The Government Office for Science, 2012.). Misao vodilja "biti brži od drugih" u konačnici je i dovela do pojave visokofrekventnog trgovanja, kojega karakteriziraju interakcije na brzinama u potpunosti izvan mogućnosti ljudi. Već u trajanju jednog treptaja, od oko 350 milisekundi, izvrši se stotinjak milijuna transakcija (Sajter, 2013.), što dovodi do situacije da sad brzina svjetlosti postaje ograničavajući faktor daljnjeg napretka.

Druga su, pak, nastojanja usmjerena prema pokušajima da se predvidi budućnost.

Cilj ovoga rada je istražiti mogućnost predviđanja kretanja na tržištima vrijednosnica primjenom algoritma strojnog učenja pri čemu se kao ulazne varijable koriste tehnički indikatori, dok izlaznu varijablu predstavlja smjer promjene cijena na određeni dan u budućnosti. Za potrebe provođenja eksperimenta izrađena je aplikacija u čijem se središtu nalazi stroj s potpornim vektorima (eng. *Support Vector Machine*, u daljnjem tekstu SVM) koji, kao prvi algoritam proistekao iz statističke teorije učenja, svoju uspješnost zahvaljuje implementaciji principa strukturne minimizacije rizika.

Rad je strukturiran na sljedeći način: nakon uvodnog razmatranja, u drugom se poglavlju prikazuju osnovne postavke teorije učenja. Nakon definiranja problema učenja, obrazlaže se utjecaj kapaciteta klase funkcija učećega stroja na sposobnost generalizacije.

Uvodi se pojam VC dimenzije te se pojašnjava princip strukturne minimizacije rizika.

Treće je poglavlje posvećeno stroju s potpornim vektorima. Prikazuje se kako SVM omogućava stvaranje nelinearnih granica odlučivanja u prostoru ulaznih podataka nelinearnim mapiranjem u visoko (čak i beskonačno) dimenzionalni prostor, te kako je računanje u takvom prostoru uopće moguće. Što je to kernel i kako široke margine povećavaju otpornost na šum, također se objašnjava u okviru ovog poglavlja.

Četvrtim se poglavljem pojašnjava što to čini predviđanje financijskih vremenskih nizova (gotovo) nemogućim zadatkom, dok se u petom poglavlju detaljno prikazuje programsko ostvarenje koje prati ovaj rad. Prikazane su glavne faze procesa predviđanja, nadopunjene prethodnim testiranjem predvidljivosti vremenskih nizova i naknadnom simulacijom trgovanja na temelju dobivenih rezultata.

U šestom se poglavlju daje opis te se prikazuju glavni rezultati eksperimenta. Provedeno je opsežno ispitivanje različitih faktora: duljine niza, kombinacija parametara, neravnoteže u podacima, tehničkih indikatora i samog procesa njihovog odabira, kombinacija klasifikatora, a sve s ciljem kako bi se moglo odgovoriti na glavno pitanje: uspijeva li SVM otkriti veze skrivene u podacima.

Sedmo se poglavlje odnosi na zaključak

U radu su korištene znanstvene metode deskripcije, kompilacije, analize i sinteze, komparativna metoda, metoda indukcije i dedukcije, te kvantitativna analiza. Korišteni su pretežno strani izvori podataka u obliku knjiga, članaka i resursa dostupnih na internetu.

2. TEORIJA UČENJA

"Ništa nije tako praktično kao dobra teorija."

Vladimir Vapnik

Kako strojevi uče i kako definirati problem učenja, koje koncepte je moguće naučiti i pod kojim uvjetima, koliko je dobro moguće naučiti te koncepte korištenjem algoritama, odnosno kako je učenje uopće moguće i kako znati je li nešto doista naučeno ili je samo memorirano? Odgovore na ta i slična pitanja daje teorija učenja čije su osnovne ideje izložene u nastavku.

2.1. Komponente problema učenja

Općenito se problem učenja, na primjeru klasifikacije[2], može opisati na sljedeći način: postoji nekakav nepoznati proces koji klasificira objekte tako da određeni objekt ili pripada ili ne pripada nekoj klasi. Mi ne znamo ništa o tom procesu. Jedino s čime raspolažemo jesu podaci o objektima i dodijeljenim im oznakama klasa koji predstavljaju očitovanje tog procesa. Na temelju raspoloživog uzorka podataka pokušavamo konstruirati model koji će aproksimirati taj nepoznati proces na najbolji mogući način i to tako da, jednom kad se suoči s novim, još neviđenim podacima, bude ih u stanju ispravno klasificirati. Formalnije se problem učenja može prikazati s tri osnovne komponente (Vapnik, 1999.):

1. generator ulaznih podataka[3] x izvučenih nezavisno iz nepoznate distribucije $P(x)$
2. funkcija koja za svaki x daje y prema uvjetnoj vjerojatnosti $P(y|x)$
3. učeći stroj (eng. l*earning machine*) sposoban implementirati skup funkcija

2 Ovdje se radi o problemu nadziranog učenja (eng. *supervised learning*) odnosno učenju uz "učitelja" s obzirom da se raspolaže s podacima koji se sastoje od parova ulaza i ispravnog izlaza (input, output). Ostale paradigme učenja s obzirom na prirodu podataka s kojima se raspolaže jesu:

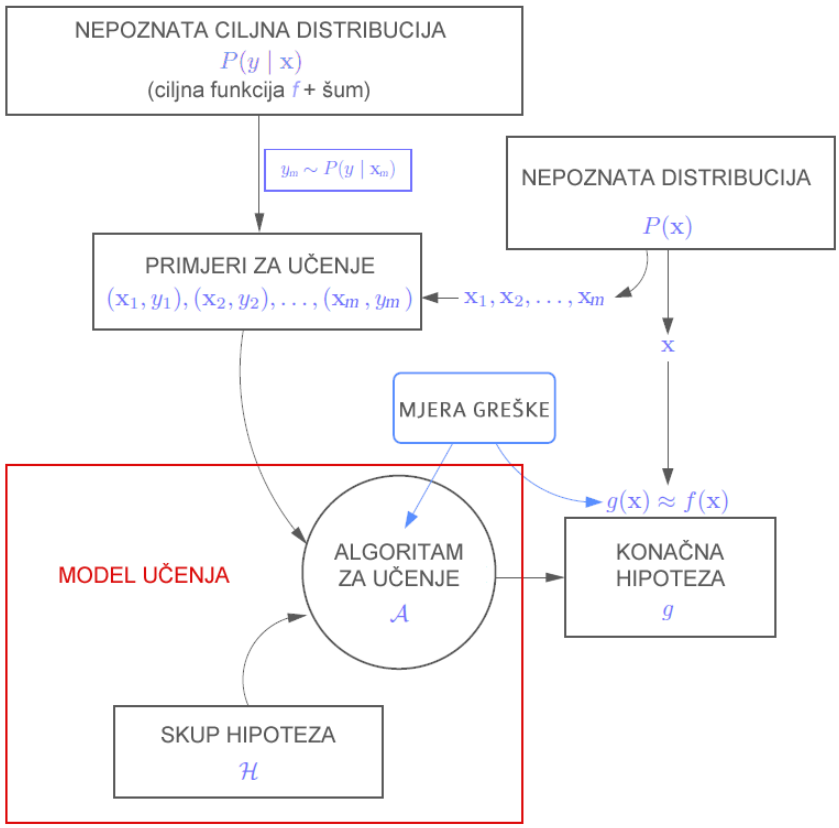
- Učenje s podrškom (eng. *reinforcement learning*) odnosno učenje uz "kritičara" koje se koristi kad se uz ulazne podatke ne pojavljuje ciljni izlaz već samo mogući izlaz popraćen određenom mjerom koja govori koliko je taj izlaz dobar s obzirom na dani ulaz (input, mogući output, mjera). Nije dostupna informacija o tome koliko bi neki drugi izlaz bio dobar s obzirom na dani ulaz.
- Nenadgledano učenje (eng. *unsupervised learning*) odnosno učenje bez nadzora koje se koristi kad podaci ne sadrže izlaz već samo ulaz (input). Može se shvatiti kao zadatak pronalaženja uzoraka i strukture u ulaznim podacima. (Abu-Mostafa, Magdon-Ismail i Lin, 2012.)

3 Za ulazne podatke koriste se i nazivi: uzorci, instance, opservacije, inputi. Izlazni podaci nazivaju se ponekad oznake, cilj, output.

$f(x, \alpha), \alpha \in \Lambda$.

Problem učenja tada se svodi na odabir iz danog skupa funkcija $f(x, \alpha), \alpha \in \Lambda$ onu koja najbolje predviđa odgovor ciljne funkcije. Odabir vrši algoritam učenja na temelju skupa od m primjera za učenje $(x_1, y_1), ..., (x_m, y_m)$. Standardna pretpostavka teorije učenja je da su primjeri za učenje i.i.d. (*independent and identically distributed*) odnosno da su generirani nezavisno iz neke nepoznate zajedničke distribucije $P(x, y) = P(x)P(y|x)$. [4]

Algoritam za učenje i skup funkcija, koje se u okviru strojnog učenja nazivaju hipotezama, čine model učenja (Abu-Mostafa, Magdon-Ismail i Lin, 2012.) kao što je prikazano na slici 1.



Slika 1. *Komponente problema nadziranog učenja*

Izvor: prilagođeno prema Abu-Mostafa, Magdon-Ismail i Lin (2012.)

4 To je jedina pretpostavka i ujedno bitna razlika između strojnog učenja i statistike koja postavlja puno restriktivnije pretpostavke.

2.2. Rizik

Nakon što je algoritam odabrao konačnu hipotezu, uspješnost učenja odredit će se na temelju uspješnosti generalizacije tj. sposobnosti predviđanja na novim primjerima. Diskrepancija između predviđanja dobivenih konačnom hipotezom i stvarne vrijednosti mjeri se funkcijom gubitka[5] (eng. *loss function*). U slučaju klasifikacije to će biti nula-jedan funkcija gubitka (eng. *zero-one loss function*) (Schoelkopf i Smola, 2002.):

$$c(x, y, f(x, \alpha)) = \begin{cases} 0 & \text{ako je } y = f(x, \alpha) \\ 1 & \text{ako je } y \neq f(x, \alpha) \end{cases} \tag{2.1}$$

Prikazani problem učenja zapravo se svodi na problem pronalaženja funkcije f koja minimizira očekivani gubitak ili rizik definiran kao (Schoelkopf i Smola, 2002.):

$$R(\alpha) = \int_{\mathcal{X} \times \mathcal{Y}} c(x, y, f(x, \alpha)) d\mathrm{P}(x, y) \tag{2.2}$$

gdje je c funkcija gubitka, a P nepoznata distribucija iz koje su izvučeni primjeri. Rizik se temelji na performansama nad cijelim ulaznim prostorom.

2.3. Empirijski rizik i princip minimizacije empirijskog rizika

Ako je učenje bilo uspješno, $R \approx 0$. Međutim, teškoća proizlazi iz toga što se pokušava minimizirati veličina koju nije moguće izračunati budući da je distribucija P nepoznata. Ali zato, ono što je poznato jesu podaci izvučeni iz P. Pokušat će se procijeniti funkcija f na temelju primjera za učenje, što bi trebalo biti blisko minimiziranju rizika (2.2). Rizik R zamjenjuje se empirijskim rizikom R_{emp} dobivenim na temelju primjera za učenje (Schoelkopf i Smola, 2002.):

$$R_{emp}(\alpha) = \frac{1}{m} \sum_{i=1}^{m} c(x_i, y_i, f(x_i, \alpha)). \tag{2.3}$$

To vodi principu indukcije poznatom kao princip minimizacije empirijskog rizika (eng. *empirical risk minimization (ERM) induction principle*) koji preporučuje da se među kandidatima hipoteza odabere $f(x, \alpha), \alpha \in \Lambda$ koja minimizira empirijski rizik R_{emp} (Schoelkopf i Smola, 2002.).

Dva su pitanja povezana s pokušajem da se na temelju empirijskog rizika procijeni

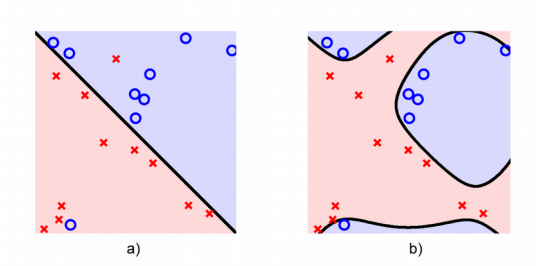
5 Gubitak zbog poduzimanja akcije $f(x)$ umjesto y za opaženi x.

rizik:

1. može li se R_{emp} učiniti dovoljno malim?
2. može li se osigurati da R_{emp} bude dovoljno blizu R?

2.4. Kompleksnost skupa hipoteza

Ako na raspolaganju imamo samo jednu hipotezu, ništa ne učimo nego provjeravamo je li određena hipoteza dobra ili nije. Zato se u procesu učenja pretražuje čitav skup hipoteza, tražeći onu koja daje malu grešku. Biranjem konačne hipoteze iz kompleksnijeg skupa omogućava se veća fleksibilnost algoritmu za učenje da pronađe onu koja će biti dobro prilagođen podacima i time postigne $R_{emp} = 0$. No, minimiziranje empirijskog rizika ne implicira nužno i postizanje male greške testiranja na primjerima iz iste distribucije (Abu-Mostafa, Magdon-Ismail, i Lin, 2012.).



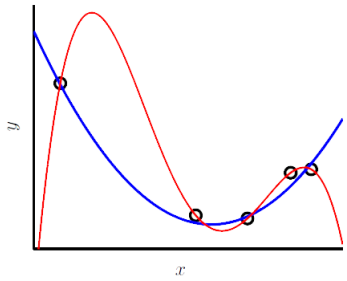
Slika 2. *Mogućnost savršenog razdvajanja podataka: a) linearnom, b) nelinearnom granicom odlučivanja*

Izvor: Abu-Mostafa, Magdon-Ismail, i Lin (2012.)

Na slici 2. prikazan je slučaj u kojem linearnom granicom odlučivanja nije postignuto savršeno razdvajanje podataka, dok nelinearnom to jeste. Međutim, ovako kompleksne granice imati će slabije generalizacijske performanse i dovesti do pojave *overfittinga,* odnosno situacije kad kompleksniji model koristi dodatne stupnjeve slobode da "nauči" šum u podacima.[6]

Slika 6. pokazuje situaciju u kojoj je, zbog ponešto šuma u podacima, algoritam učenja odabrao kompleksniji model (crvena linija) savršeno prilagođen podacima. Premda je ciljna funkcija polinom drugog stupnja (plava linija), zbog "naučenog" šuma odabran je polinom višega stupnja što vodi većoj grešci prilikom predviđanja na novim primjerima.

6 S druge strane, odviše jednostavan model neće dobro objasniti podatke što će predstavljati pojavu nazvanu *underfitting.*



Slika 3. *Overfitting: iako se kompleksnija funkcija (crvena linija) savršeno prilagođava podacima za učenje (označeni kružićima), njenim se odabirom postiže veća greška prilikom predviđanja na novim primjerima*

Izvor: Abu-Mostafa, Magdon-Ismail, i Lin (2012.)

Kako se kompleksnost skupa hipoteza povećava, povećava se i vjerojatnost da empirijski rizik R_{emp} neće biti dobar procjenitelj stvarnog rizika R. Ako se dopusti biranje hipoteze iz jako velike klase funkcija, uvijek je moguće pronaći neku funkciju f koja će dati jako malu vrijednost R_{emp} a da to istovremeno bude jako daleko od minimiziranja R. Ako se želi osigurati da R_{emp} bude dovoljno blizu, potrebno je voditi računa o kompleksnosti klase funkcija[7] iz kojeg se bira f. U protivnome, ne može se nadati nikakvome učenju (Schoelkopf i Smola, 2002.).

2.5. Statistička teorija učenja i VC dimenzija

Statistička teorija učenja (eng. *Statistical learning theoriy – SLT*) ili VC (Vapnik-Chervonenkis) teorija ima svoje začetke u kasnim 60-im godinama 20. stoljeća, a sve do 90-ih godina 20. stoljeća predstavljala je čisto teorijsku analizu koja se bavila estimacijom funkcija na temelju danog skupa podataka. Sredinom 90-ih godina razvijen je novi tip algoritma za učenje baziran na predloženoj teoriji tako da se, zahvaljujući njegovoj uspješnosti u rješavanju konkretnih problema iz prakse, pokazala ne samo kao alat za teorijsku analizu već i alat za stvaranje praktičnih algoritama za estimaciju višedimenzionalnih funkcija (Vapnik, 1999.).

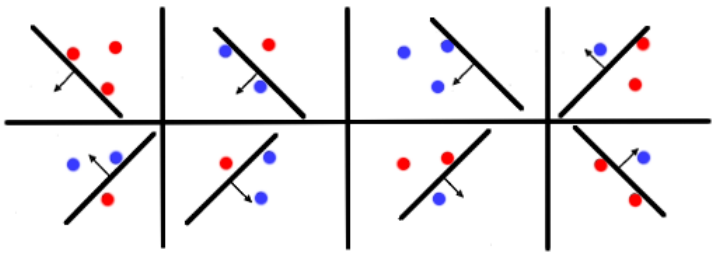
Statistička teorija učenja proučava matematička svojstva učećih strojeva, odnosno svojstva klasa funkcija koje učeći stroj može implementirati, a koja im omogućavaju da dobro generaliziraju na neviđenim podacima (Schoelkopf i Smola, 2002.).

7 Kompleksnost nepoznate ciljne funkcije f ne utječe na to koliko dobro R_{emp} aproksimira R (Abu-Mostafa, Magdon-Ismail, i Lin, 2012.)

2.5.1. VC dimenzija

Već je ukazana potreba postavljanja restrikcija na klasu funkcija iz kojih se odabire procjena ciljne funkcije f, odnosno potrebu uzimanja u obzir kompleksnosti ili kapaciteta klasa funkcija koje učeći stroj može implementirati. Najpoznatiji koncept kapaciteta iz VC teorije je VC dimenzija koja se može shvatiti kao mjera kapaciteta učećeg stroja izražena jednim brojem, a njeno objašnjenje je sljedeće (Schoelkopf i Smola, 2002.):

svaka klasa funkcija razdvaja uzorke na određen način što vodi određenom označavanju uzoraka. S obzirom da su u slučaju binarne klasifikacije moguće dvije oznake, npr. $\{\pm 1\}$, to je najviše 2^m mogućih načina označavanja m uzoraka. Jako bogata klasa funkcija moći će realizirati svih 2^m separacija, u tom slučaju kaže se da ta klasa funkcija "razbija" (eng. *shatter*) m točaka. Međutim, neće svaka klasa funkcija biti dovoljno bogata da "razbije" m točaka. VC dimenzija h neke klase funkcija definirana je kao najveći podskup od m točaka koje ta klasa funkcija može "razbiti". Ako je moguće rastaviti po volji velike skupove, tada je ona beskonačna.



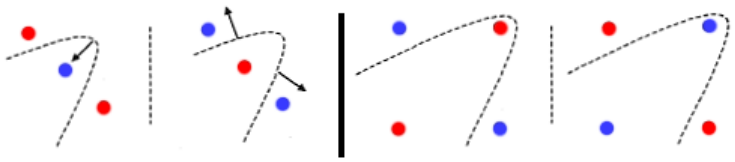
Slika 4. *VC dimenzija: linearnom funkcijom u dvodimenzionalnom prostoru moguće je "razbiti" tri točke*

Izvor: Kecman (2001.)

Slika 4. prikazuje razdvajanje triju točaka linearnom funkcijom u dvije klase što je moguće učiniti na osam načina (2^3). Vidljivo je da je u $\mathbb{R}^2$ moguće realizirati svih osam razdvajanja, što znači da VC dimenzija u ovome slučaju iznosi 3.

Ako VC dimenzija klase funkcija iznosi h, to znači da postoji barem jedan skup od h točaka koje mogu biti razdvojene tom klasom funkcija, iako može postojati i skup koji neće biti moguće razdvojiti. Na slici 5. lijevo prikazan je slučaj kolinearnih točaka i dva načina označavanja koja nije moguće realizirati linearnom funkcijom. Također, u $\mathbb{R}^2$ nije moguće razdvojiti četiri točke linearnom funkcijom (slika 5. desno). Međutim, u oba slučaja to je

moguće postići kvadratnom funkcijom.



Slika 5. *Nelinearnom funkcijom u dvodimenzionalnom prostoru moguće je "razbiti" tri kolinearne točke (lijevo) ili četiri točke (desno) što nije moguće postići linearnom funkcijom*

Izvor: Kecman (2001.)

VC dimenzija mjeri efektivne parametre ili stupnjeve slobode koji omogućavaju da model izrazi različiti skup hipoteza. Što više parametara model ima, to je raznovrsniji i skup hipoteza. Međutim, raznovrsnost nije nužno dobra u kontekstu generalizacije, navode Abu-Mostafa, Magdon-Ismail, i Lin (2012.). Za najveću raznovrsnost VC dimenzija je beskonačna i tada se ne može očekivati nikakva generalizacija.

2.5.2. VC granica

VC granica sposobnosti generalizacije (eng. *VC generalization bound*) najvažniji je matematički rezultat u teoriji učenja. Njome je postavljena gornja granica na rizik koja ovisi o empirijskom riziku i kapacitetu klase funkcija, te predstavlja univerzalni rezultat primjenjiv na sve skupove hipoteza, algoritme učenja, prostore ulaznih podataka, distribucije vjerojatnosti i ciljne funkcije (Abu-Mostafa, Magdon-Ismail i Lin, 2012.).

Ako je $h < m$ VC dimenzija klase funkcija koju može implementirati učeći stroj, tada za sve funkcije te klase, s vjerojatnošću barem $1 - \sigma$ vrijedi granica (Schoelkopf i Smola, 2002.):

$$R(\alpha) \leq R_{emp}(\alpha) + \phi(h, m, \sigma) \tag{2.4}$$

gdje je h VC dimenzija, m broj primjera za učenje, σ neki mali broj tako da $0 < \sigma < 1$.Član koji izražava pouzdanost (eng. *VC-confidence*) definiran je s (Schoelkopf i Smola, 2002.):

$$\phi(h, m, \sigma) = \sqrt{\frac{1}{m}\left(h(\ln\frac{2m}{h} + 1) + \ln\frac{4}{\sigma}\right)} \tag{2.5}$$

Minimizacija te granice vodi principu strukturne minimizacije rizika.

2.6. Princip strukturne minimizacije rizika

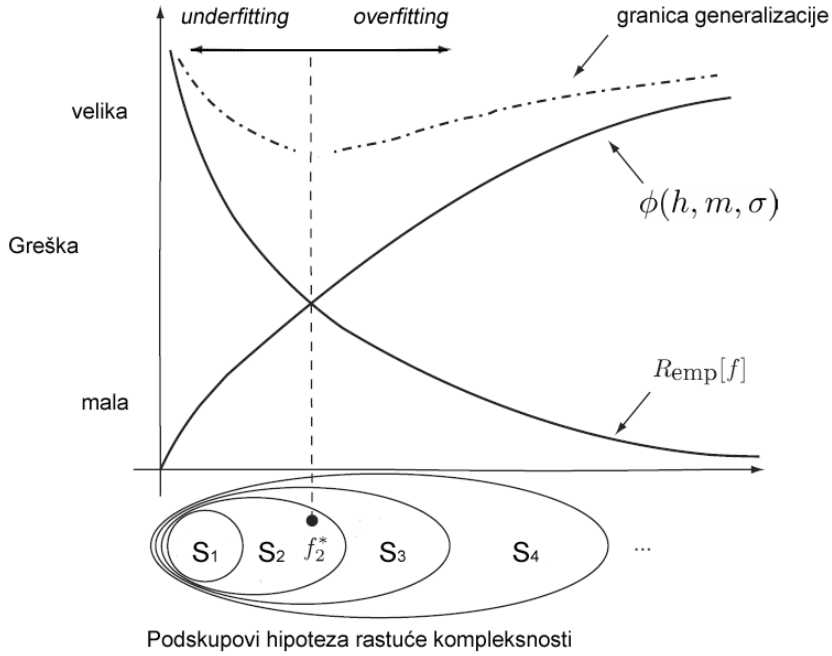
ERM princip namijenjen je velikim skupovima podataka. Ali, ako je odnos m/h mali, tada čak i mali empirijski rizik R_{emp} ne garantira malu vrijednost rizika R. Stoga je potrebno konstruirati princip indukcije koji uzima u obzir veličinu uzorka, odnosno princip za "mali" uzorak (mali uzorak smatra onaj za koji je $m/h < 20$) (Vapnik, 1999.).

Takav princip, baziran na simultanoj minimizaciji oba člana u (2.4), od kojih jedan ovisi o vrijednosti empirijskog rizika, a drugi ovisi o VC dimenziji skupa funkcija, nazvan je princip strukturne minimizacije rizika (eng. *structrural risk minimization (SRM) principle*).

Neka skup funkcija S ima strukturu koju čine podskupovi rastuće kompleksnosti:

$$S_1 \subset S_2 \subset ... \subset S_n ... \tag{2.6}$$

SRM princip sugerira da se za dani skup primjera za učenje odabere element strukture S_n gdje je $n = n(m)$ i odabere odgovarajuća funkcija iz S_n za koju je garantirani rizik (desna strana u 2.4) minimalan (slika 6) (Vapnik, 1999.).



Slika 6. *Odnos empirijskog rizika i VC-pouzdanosti: empirijski rizik opada s povećanjem kompleksnosti skupa hipoteza iz kojeg se bira f, ali povećava se i kazna koja se plaća za dodatnu kompleksnot. Zbrajanje empirijskog rizika i VC-pouzdanosti rezultira granicom generalizacije.*

Izvor: prilagođeno prema Hamel (2009.) i Kecman (2001.)

Dobre generalizacijske performanse postižu se kada se skup funkcija iz kojega se odabire f ograniči na onaj koji ima kapacitet prikladan za raspoloživu količinu podataka za učenje. To znači da za dobre rezultate predviđanja na novim podacima, klasa funkcija mora biti ograničena tako da kapacitet (VC dimenzija) bude dovoljno mali u odnosu na raspoloživu količinu podataka za učenje (pri čemu praktično pravilo kaže da m treba biti barem $10 \times h$ (Abu-Mostafa, Magdon-Ismail i Lin, 2012.)). Istovremeno, klasa funkcija treba biti dovoljno velika da pruži funkcije sposobne modelirati veze skrivene u $P(x, y)$. Time SRM princip sugerira potrebu postizanja kompromisa između kvalitete aproksimacije i kompleksnosti aproksimacijske funkcije (Vapnik, 1999.). Odabir odgovarajućeg skupa funkcija ključno je za učenje na temelju podataka.

3. STROJ S POTPORNIM VEKTORIMA

Stroj s potpornim vektorima (eng. *Support Vector Machine* - SVM) nelinearna je generalizacija algoritma nazvanog *Generalized Portrait,* razvijenog još 60-ih godina 20. stoljeća (Smola i Schoelkopf, 2004.). Premda se po prvi puta spominje u Vapnikovom radu iz 1979. godine, osnovni oblik SVM-a razvijen je nešto kasnije (Boser, Guyon i Vapnik, 1992.), (Cortes i Vapnik, 1995.) u okviru AT&T Bell Laboratories, zbog čega je, uz čvrsto uporište u statističkoj teoriji učenja, od početka bio usmjeren na praktičnu primjenu (Smola i Schoelkopf, 2004.).

Osnovni poticaj razvoju SVM-a proizašao je iz potrebe rješavanje problema kontrole kapaciteta ili *overfittinga*, na način kako je to objašnjeno u prethodnome poglavlju. Za razliku od nekih drugih algoritama učenja, VC dimenziju SVM-a moguće je izračunati (Burges, 1998.).

Kako navode Schoelkopf i Smola (2002.), SVM se zapravo može smatrati prvim "*spin-offom*" statističke teorije učenja koji spada u širu klasu algoritama učenja baziranih na kernelima (eng. *Kernel methods*), nastalih kombinacijom teorije strojnog učenja, optimizacijskih algoritama iz operacijskih istraživanja te funkcionalne analize. Kernel funkcija, koja se može shvatiti kao mjera sličnosti, omogućava konstruiranje računski efikasnih načina koji omogućuju korištenje visoko (čak i beskonačno!) dimenzionalnih nelinearnih transformacija.

Kombinacijom koncepata maksimalne margine, zahvaljujući kojoj pokazuje veću otpornost na šum i smanjenje pojave *overfittinga,* te kernela, dobiven je moćan nelinearni model s automatskom regularizacijom (Abu-Mostafa, Magdon-Ismail i Lin, 2012.) u čije se prednosti još ubrajaju (Vapnik, 1999.):

- jednostavnost korištenja,
- mali broj slobodnih parametara,
- dobro se snalazi u situacijama visoko dimenzionalnih problema popraćenih malim brojem primjera za učenje,
- optimizacijski problem ima jedinstveno rješenje,

- proces učenja je prilično brz,
- implementacija novog skupa decizijskih funkcija može se izvršiti promjenom samo jedne funkcije (kernela).

Inicijalni rad na SVM algoritmu bio je povezan s konstruiranjem klasifikatora za raspoznavanje strojno generiranih znakova (*Optical character recognition - OCR*), a kako su uskoro SVM klasifikatori postali usporedivi s najboljima, rad se nastavio na proširenjima kojima je omogućena uspješna primjena i na problemima regresije, a danas predstavlja jedan od standardnih algoritama korištenih za rješavanje najraznovrsnijih problema koji variraju od raspoznavanja govora, kategorizacije teksta, predviđanja vremenskih nizova, pa do detekcije kvarkova ili ekspresije gena, pri čemu u većini testova pokazuju jednake ili čak i bolje generalizacijske performanse u odnosu na ostale algoritme (Burges, 1998.).

3.1. Sličnost

Pretpostavimo da se radi o problemu binarne klasifikacije i da raspolažemo s primjerima za učenje (x, y), $x \in \mathcal{X}$, $y \in \{\pm 1\}$. Želimo moći ispravno klasificirati nove primjere i to tako da za određeni $x \in \mathcal{X}$ predvidimo odgovarajući $y \in \{\pm 1\}$. To će biti moguće ako postoji neka mjera sličnosti prema kojoj se odabire y tako da novi par (x, y) na neki način bude sličan skupu za učenje.

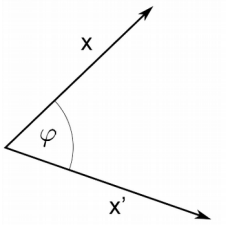
Sličnost izlaznih podataka uobičajeno se mjeri funkcijom gubitka koja kaže koliko se predviđena vrijednost y podudara s njezinom stvarnom vrijednošću. U tom smislu, u slučaju binarne klasifikacije, moguće su samo dvije situacije: dvije su oznake ili jednake ili različite. S druge strane, pitanje odabira mjere sličnosti ulaznih podataka središnja je tema strojnog učenja (Schoelkopf i Smola, 2002.).

Kao jednostavna mjera sličnosti može se koristiti unutarnji (skalarni) produkt. Skalarni produkt dvaju vektora $\mathbf{x}$ i $\mathbf{x}'$ definiran je s:

$$\langle \mathbf{x}, \mathbf{x}' \rangle := \|\mathbf{x}\| \|\mathbf{x}'\| \cos \varphi \tag{3.1}$$

pri čemu je $\|\mathbf{x}\|$ duljina (norma) vektora[8] $\mathbf{x}$, φ je kut među vektorima (slika 7.).

8 Iz 3.1 slijedi da je $\langle \mathbf{x}, \mathbf{x} \rangle = \|\mathbf{x}\|^2$ što daje $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$.



Slika 7. *Vektori*

Izvor: izradila autorica

Skalarni produkt dvaju vektora može biti jednak nuli onda i samo onda ako je jedan od njih nul-vektor[9] ili ako su međusobno okomiti. Tada se može reći da među njima nema sličnosti. Da bi se unutarnji produkt mogao koristiti kao mjera sličnosti, potrebno je najprije primjere za učenje reprezentirati kao vektore.

Kompleksnija mjera sličnosti oblika:

$$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$$

$$(x, x') \mapsto k(x, x') \tag{3.2}$$

naziva se kernel, o čemu će biti više riječi u poglavlju 3.9.1.

3.2. Klasifikator optimalne margine

Kako navode Schoelkopf i Smola (2002.), u nastojanju da pronađu klasu funkcija čiji se kapacitet može izračunati, Vapnik i suradnici razmatrali su klasu hiperravnina u nekom prostoru $\mathcal{H}$ na kojem je definiran unutarnji produkt:

$$\langle \mathbf{w}, \mathbf{x} \rangle + b = 0 \text{ gdje je } \mathbf{w} \in \mathcal{H}, b \in \mathbb{R} \tag{3.3}$$

što odgovara decizijskoj funkciji [10]

$$f(\mathbf{x}) = sgn(\langle \mathbf{w}, \mathbf{x} \rangle + b) \tag{3.4}$$

te su za linearno separabilne probleme predložili algoritam učenja koji omogućava konstruiranje f na temelju empirijskih podataka. Nazvali su ga *Generalized Portrait***,** a temelji se na sljedećem (Schoelkopf i Smola, 2002.):

- među svim hiperravninama koje razdvajaju podatke postoji jedinstvena optimalna

9 Nul-vektor je vektor koji ima početak i kraj u istoj točki. Označava se sa $\mathbf{0}$. Vrijedi $\|\mathbf{0}\| = 0$.

10 Decizijska funkcija je funkcija čiji predznak predstavlja klasu dodijeljenu nekom ulaznom podatku $\mathbf{x}$.

hiperravnina koju karakterizira maksimalna margina separacije između podataka i hiperravnine,

- kapacitet klase razdvajajućih hiperravnina smanjuje se s povećanjem margine, što znači da postoje teorijski argumenti koji podupiru dobru sposobnost generalizacije optimalne hiperravnine.

Pitanja koja se na temelju toga postavljaju:

- može li se efikasno pronaći maksimalna margina?
- na koji način margina utječe na kapacitet?
- što napraviti u situaciji kad podaci nisu linearno separabilni?

Odgovori na ta pitanja daju se u nastavku.

3.2.1. Razdvajajuća hiperravnina

Pretpostavimo da nam je dan skup primjera za učenje koji se sastoje od parova ulaznih podataka, reprezentiranih vektorima $\mathbf{x_1}, ..., \mathbf{x_m}$ u nekom prostoru $\mathcal{H}$, i pripadajućih oznaka klasa $y_1, ..., y_m$. Problem binarne klasifikacije tada se može prikazati kao (Schoelkopf i Smola, 2002.):

$$(\mathbf{x_1}, y_1), ..., (\mathbf{x_m}, y_m) \in \mathcal{H} \times \{\pm 1\} \tag{3.5}$$

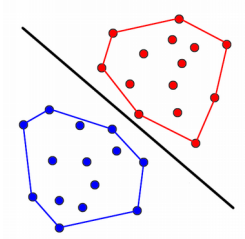
U slučaju linearno razdvojivih podataka, moguće je pronaći hiperravninu koja savršeno razdvaja pozitivne od negativnih primjera pri čemu se svaka hiperravnina u $\mathcal{H}$ može zapisati kao:

$$\{\mathbf{x} \in \mathcal{H} | \langle \mathbf{w}, \mathbf{x} \rangle + b = 0\}, \quad \mathbf{w} \in \mathcal{H}, b \in \mathbb{R} \tag{3.6}$$

gdje je $\mathbf{w}$ normala hiperravnine.

Na slici 8. prikazana su dva skupa podataka koji predstavljaju dvije različite klase u dvodimenzionalnom prostoru. Vidljivo je da se njihove konveksne ljuske ne preklapaju pa možemo reći da su podaci linearno razdvojivi, odnosno, hiperravnina razdvaja podatke ako postoje vektor $\mathbf{w}$ i skalar b tako da vrijedi:

$$|\langle \mathbf{w}, \mathbf{x_i} \rangle + b| > 0, \quad i = 1, ..., m \tag{3.7}$$



Slika 8. *Linearno separabilni podaci*

Izvor: izradila autorica

Razdvajajuća hiperravnina tada čini linearnu granicu odlučivanja te se koristi u konstruiranju decizijske funkcije $g(x)$ koja dodjeljuje pripadajuću klasu pojedinom podatku:

$$g(\mathbf{x}) = sgn(f(\mathbf{x})) \quad \text{gdje je} \quad f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b \quad \text{za} \quad \mathbf{w} \in \mathcal{H} \quad \text{i} \quad b \in \mathbb{R} \tag{3.8}$$

Međutim, nedostatak zapisa (3.7) je taj što postoji sloboda da se parametri $\mathbf{w}$ i b pomnože nekom konstantom različitom od nule čime se ponovo dobije ista hiperravnina. Zbog toga je uobičajena njihova normalizacija s $\|\mathbf{w}\|$ što vodi novoj definiciji razdvajajuće hiperravnine (Abu-Mostafa, Magdon-Ismail i Lin, 2012.):

hiperravnina razdvaja podatke ako i samo ako postoji par $(\mathbf{w}, b)$ za koje vrijedi :

$$\min_{i=1,\ldots,m} |\langle \mathbf{w}, \mathbf{x_i} \rangle + b| = 1. \tag{3.9}$$

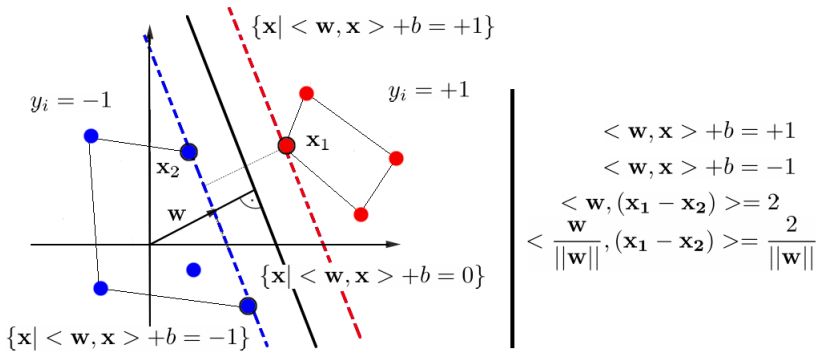
To znači da točka najbliža hiperravnini zadovoljava $|\langle \mathbf{w}, \mathbf{x_i} \rangle + b| = 1$ i njena udaljenost od hiperravnine iznosi $1/\|\mathbf{w}\|$. Ta udaljenost predstavlja marginu (slika 9.), odnosno geometrijska margina točke $(\mathbf{x}, y) \in \mathcal{H} \times \{\pm 1\}$ za hiperravninu $\{\mathbf{x} \in \mathcal{H} | \langle \mathbf{w}, \mathbf{x} \rangle + b = 0\}$ jest (Schoelkopf i Smola, 2002.):

$$\rho_{(\mathbf{w},b)}(\mathbf{x}, y) := \frac{y(\langle \mathbf{w}, \mathbf{x} \rangle + b)}{\|\mathbf{w}\|} \tag{3.10}$$

a minimalna vrijednost

$$\rho_{(\mathbf{w},b)} := \min_{i=1,\ldots,m} \rho_{(\mathbf{w},b)}(\mathbf{x_i}, y_i) \tag{3.11}$$

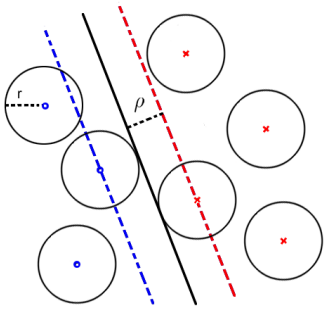
naziva se geometrijska margina skupa primjera za učenje ili jednostavno margina.



Slika 9. *Razdvajajuća hiperravnina i margina na primjeru binarne klasifikacije*

Izvor: prilagođeno prema Schoelkopf i Smola, (2002.)

Margina razdvajajuće hiperravnine i duljina vektora $\mathbf{w}$ imaju važnu ulogu u SVM algoritmu. Kao što je već spomenuto, kapacitet klase razdvajajućih hiperravnina smanjuje se s povećanjem margine, što znači da je za bolju sposobnost generalizacije potrebno pronaći hiperravninu sa širokom marginom.



Slika 10. *Široka margina i šum: unatoč prisutnosti šuma (predstavljenog kružnicama) podaci će biti ispravno klasificirani*

Izvor: Schoelkopf i Smola (2002.)

Najjednostavnije objašnjenje za opravdanje široke margine može se dobiti ako se pogleda slika 10. na kojoj su podaci prikazani kao 'o' ili 'x', dok je šum, inače često prisutan u podacima zbog npr. pogrešnog mjerenja ili greške prilikom ručnog unosa, prikazan kružnicama čiji je $r > 0$. Vidljivo je da će, unatoč tome što šum otežava ispravnu klasifikaciju, uz dovoljno široku marginu $\rho > r$ podaci ipak biti ispravno klasificirani. Osim toga, ako se podaci nalaze na udaljenosti od barem ρ od hiperravnine, tada niti male perturbacije parametara hiperravnine neće promijeniti klasifikaciju podataka za učenje.

O važnosti margine i veze s VC dimenzijom biti će više riječi u poglavlju 3.6.

3.2.2. Optimalna hiperravnina

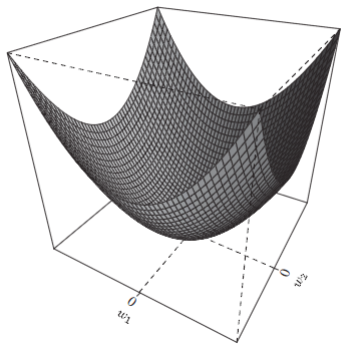
S obzirom da margina iznosi $1/\|\mathbf{w}\|$, široku marginu moguće je dobiti tako da se $\|\mathbf{w}\|$ učini malim. Stoga se problem konstruiranja optimalne hiperravnine svodi na optimizacijski probelm minimiziranja $\|\mathbf{w}\|$ (Schoelkopf i Smola, 2002.):

$$\min_{\mathbf{w},b} \ \tau(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 \tag{3.12}$$

uz ograničenja

$$y_i \left(\langle \mathbf{x_i}, \mathbf{w} \rangle + b\right) \geq 1, \forall i. \tag{3.13}$$

Budući da je funkcija cilja τ kvadratna uz linearna ograničenja, gore prikazani optimizacijski problem predstavlja problem kvadratnog programiranja kojeg karakterizira jedinstveno rješenje, odnosno globalni optimum (slika 11.).



Slika 11. *Graf ciljne funkcije*

Izvor: Hamel (2009.)

Navedeni optimizacijski problem predstavlja primalni problem. Međutim, kako svaki optimizacijski problem ima i svoj dualni problem, čijim se rješavanjem također dolazi do optimalnog rješenja, ako takvo rješenje postoji, obično se rješava onaj kojega je, iz određenih razloga, jednostavnije za riješiti. U ovom je slučaju uobičajeno rješavanje dualnog problema, čija svojstva olakšavaju kasniju modifikaciju kojom nastaje SVM, a zahvaljujući kojoj je omogućena primjena i u situacijama linearno neseparabilnih podataka. (poglavlje 3.9).

3.3. Metoda Lagrangeovih multiplikatora

Na temelju optimizacijskog problema oblika (Hamel, 2009.):

$$\min_{\mathbf{x}} \phi(\mathbf{x}) \tag{3.14}$$

uz ograničenja

$$g_i(\mathbf{x}) \geq 0, i = 1, ..., l \tag{3.15}$$

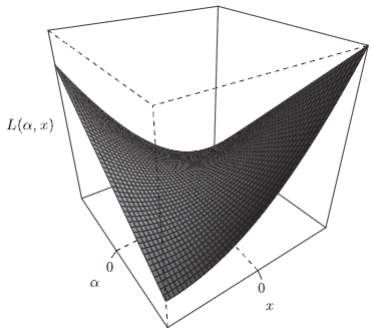
za $\mathbf{x} \in \mathbb{R}^n$ i uz pretpostavku da je funkcija cilja ϕ konveksna[11] uz linearna ograničenja g_i, konstruira se novi optimizacijski problem tako da nova funkcija cilja sadrži originalnu funkciju cilja ϕ i linearnu kombinaciju ograničenja g_i čime se dobiva Lagrangeova funkcija:

$$\max_{\boldsymbol{\alpha}} \min_{\mathbf{x}} L(\boldsymbol{\alpha}, \mathbf{x}) = \max_{\boldsymbol{\alpha}} \min_{\mathbf{x}} \left(\phi(\mathbf{x}) - \sum_{i=1}^{l} \alpha_i g_i(\mathbf{x}) \right) \tag{3.16}$$

uz ograničenja

$$\alpha_i \geq 0, i = 1, ..., l \tag{3.17}$$

$\mathbf{x} \in \mathbb{R}^n$ gdje su α_i Lagrangeovi multiplikatori. Rješenje ovog problema čini točka koja istovremeno maksimizira funkciju $L(\boldsymbol{\alpha}, \mathbf{x})$ s obzirom na dualnu varijablu $\boldsymbol{\alpha}$ i minimizira s obzirom na primalnu varijablu $\mathbf{x}$. To će biti sedlasta točka na grafu funkcije $L(\boldsymbol{\alpha}, \mathbf{x})$. S obzirom da je funkcija cilja konveksna uz linearna ograničenja, sedlasta točka bit će jedinstvena (slika 12.).



Slika 12. *Graf Lagrangeove funkcije*

Izvor: Hamel (2009.)

11 Funkcija $f : \mathbb{R}^n \to \mathbb{R}$ je konveksna ako vrijedi
$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda) f(y), \forall x, y \in \mathbb{R}^n$ i $\forall \lambda \in [0, 1]$.

Neka su $\boldsymbol{\alpha}^*$ i $\mathbf{x}^*$ rješenje Lagrangeove funkcije tako da (Hamel, 2009.)

$$\max_{\boldsymbol{\alpha}} \min_{\mathbf{x}} L(\boldsymbol{\alpha}, \mathbf{x}) = L(\boldsymbol{\alpha}^*, \mathbf{x}^*) = \left(\phi(\mathbf{x}^*) - \sum_{i=1}^{l} \alpha_i^* g_i(\mathbf{x}^*)\right) \tag{3.18}$$

tada je $\mathbf{x}^*$ rješenje primalnog problema (3.14)-(3.15) ako i samo ako su zadovoljeni Karush-Kuhn-Tuckerovi (KKT) uvjeti optimalnosti:

$$\frac{\partial L}{\partial \mathbf{x}}(\boldsymbol{\alpha}^*, \mathbf{x}^*) = 0 \tag{3.19}$$

$$\alpha_i^* g_i(\mathbf{x}^*) = 0 \tag{3.20}$$

$$g_i(\mathbf{x}^*) \geq 0 \tag{3.21}$$

$$\alpha_i^* \geq 0, \;\; i = 1, ..., l. \tag{3.22}$$

Uvjet (3.19) osigurava da se optimalno rješenje nalazi u sedlastoj točki. Uvjet (3.20) zahtijeva da izraz $\sum_{i=1}^{l} \alpha_i^* g_i(\mathbf{x}^*)$ iščezne tako da $L(\boldsymbol{\alpha}^*, \mathbf{x}^*) = \phi(\mathbf{x}^*)$. Taj se uvjet još se naziva KKT uvjet komplementarnosti. Uvjeti (3.21) i (3.22) su originalna ograničenja iz primalnog problema (3.15) i Lagrangeove formulacije (3.17).

Reformulacijom originalnog optimizacijskog problema u terminima dualne varijable dobije se dualni Lagrangeov problem (nazvan još i Wolfeov dual) (Hamel, 2009.):

$$\max_{\boldsymbol{\alpha}} \phi'(\boldsymbol{\alpha}) \tag{3.23}$$

uz ograničenja

$$\alpha_i \geq 0. \tag{3.24}$$

To znači da je moguće riješiti primalni optimizacijski problem pomoću Lagrangeovog duala

$$\max_{\boldsymbol{\alpha}} \phi'(\boldsymbol{\alpha}) = \phi'(\boldsymbol{\alpha}^*) = L(\boldsymbol{\alpha}^*, \mathbf{x}^*) = \phi(\mathbf{x}^*) \tag{3.25}$$

pri čemu $\mathbf{x}^*$ i $\boldsymbol{\alpha}^*$ moraju zadovoljiti KKT uvjete optimalnosti.

3.4. Lagrangeova formulacija klasifikatora optimalne margine

Na temelju optimizacijskog problema (3.12) i (3.13) konstruira se Lagrangeova funkcija (Schoelkopf i Smola, 2002.):

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{m} \alpha_i \left(y_i(\langle \mathbf{x_i}, \mathbf{w}\rangle + b) - 1\right) \tag{3.26}$$

s Langrangeovim multiplikatorima $\alpha_i \geq 0$.

Lagrangeovu funkciju L potrebno je maksimizirati s obzirom na dualnu varijablu $\boldsymbol{\alpha}$ i minimizirati s obzirom na primalne varijable $\mathbf{w}$ i b. Optimalno rješenje za $\boldsymbol{\alpha}$, $\mathbf{w}$ i b mora zadovoljiti KKT uvjete optimalnosti:

- u sedlastoj točki derivacije funkcije L s obzirom na primalne varijable moraju iščeznuti

$$\frac{\partial L}{\partial b}(\mathbf{w}, b, \boldsymbol{\alpha}) = 0 \tag{3.27}$$

$$\frac{\partial L}{\partial \mathbf{w}}(\mathbf{w}, b, \boldsymbol{\alpha}) = 0 \tag{3.28}$$

- uvjet komplementarnosti

$$\alpha_i[y_i\left(\langle \mathbf{x_i}, \mathbf{w}\rangle + b\right) - 1] = 0 \tag{3.29}$$

- originalna ograničenja

$$y_i\left(\langle \mathbf{x_i}, \mathbf{w}\rangle + b\right) - 1 \geq 0 \tag{3.30}$$

$$\alpha_i \geq 0, i = 1, ..., m. \tag{3.31}$$

(3.27) i (3.28) daju

$$\sum_{i=1}^{m} \alpha_i y_i = 0 \tag{3.32}$$

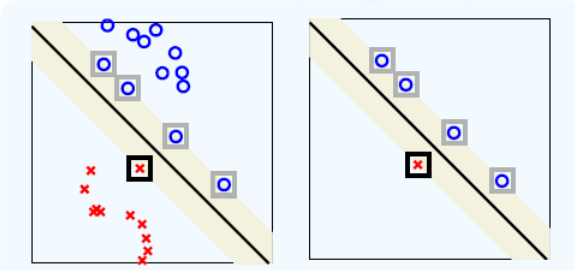
i

$$\mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x_i} \tag{3.33}$$

Rješenje $\mathbf{w}$ je jedinstveno zbog stroge konveksnosti[12] (3.12) i konveksnosti (3.13) .

12 Funkcija $f : \mathbb{R}^n \to \mathbb{R}$ je strogo konveksna ako $\forall x, y \in \mathbb{R}^n, x \neq y$ vrijedi $f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda)f(y), \forall \lambda \in (0,1)$.

Prema uvjetu komplementarnosti (3.29) samo primjeri $\mathbf{x_i}$ za koje vrijedi $\alpha_i > 0$ zadovoljavaju ograničenja (3.30) kao jednakost (aktivna ograničenja). Primjeri za koje je $\alpha_i > 0$ zovu se potporni vektori i leže točno na margini, na udaljenosti $1/\|\mathbf{w}\|$ od hiperravnine. Prema (3.29) svi preostali primjeri imaju $\alpha_i = 0$. Njihova je udaljenost veća od $1/\|\mathbf{w}\|$ i ne sudjeluju u (3.33). Za izračun optimalne hiperravnine potrebni su samo potporni vektori. Hiperravnina je u potpunosti određena primjerima koji su joj najbliže i rješenje ne ovisi o drugim primjerima. Ako se ukloni bilo koja druga točka koja nije potporni vektor, optimalna hiperravnina neće se promijeniti (slika 13.). To je važno svojstvo koje se iskorištava u implementaciji.



Slika 13. *Samo potporni vektori određuju hiperravninu*

Izvor: Abu-Mostafa, Magdon-Ismail, i Lin (2012.)

Kako navode Schoelkopf i Smola (2002.), naziv "potporni vektori" povezan je s teorijom konveksnih skupova i konveksnom optimizacijom. Za danu graničnu točku konveksnog skupa uvijek postoji hiperravnina, nazvana potporna hiperravnina, koja razdvaja tu točku od unutrašnjosti skupa (slika 9.). Potporni vektori leže na granicama konveksnih ljuski dviju klasa, oni "posjeduju" potporne hiperravnine. Optimalna hiperravnina leži u sredini između dviju paralelnih potpornih hiperravnina s maksimalnom udaljenošću. Isto tako, na temelju optimalne hiperravnine moguće je dobiti potporne hiperravnine za sve potporne vektore obiju klasa i to pomicanjem za $1/\|\mathbf{w}\|$ u oba smjera.

3.5. Dualni problem optimalne hiperravnine

Kao što je ranije navedeno, rješava se dualni problem zbog činjenice da se u toj formulaciji primjeri za učenje javljaju samo u obliku unutarnjeg produkta, a što olakšava generalizaciju postupka na nelinearni slučaj čime je omogućena primjena na puno većem rasponu problema (poglavlje 3.9).

Uvrštavanjem (3.32) i (3.33) u (3.26) dobije se Lagrangeov dualni optimizacijski problem (Schoelkopf i Smola, 2002.):

$$\max_{\alpha} \; W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \langle \mathbf{x_i}, \mathbf{x_j} \rangle \tag{3.34}$$

uz ograničenja

$$\sum_{i=1}^{m} \alpha_i y_i = 0 \tag{3.35}$$

i

$$\alpha_i \geq 0, \;\; i = 1, ..., m. \tag{3.36}$$

Dualni problem identificira potporne vektore na granici optimalne hiperravnine. Može se reći da, dok primalni problem traži hiperravninu čija je margina limitirana potpornim vektorima, dualni problem traži potporne vektore koji limitiraju veličinu margine (Hamel, 2009.).

Za razliku od primalnog problema koji ima ima $n + 1$ varijabli cilja i m ograničenja, dualni problem ne ovisi o dimenzionalnosti inputa što znatno olakšava njegovo rješavanje.

Na temelju rješenja za **w** (3.33) konstruira se decizijska funkciju koja koristi izraz za računanje unutarnjeg produkta između uzorka kojeg treba klasificirati i potpornih vektora.

$$f(\mathbf{x}) = sgn\left(\sum_{i=1}^{m} \alpha_i y_i \langle \mathbf{x}, \mathbf{x_i} \rangle + b \right) \tag{3.37}$$

Parametar b ne dobiva se direktno rješavanjem optimizacijskog problema, ali može se izračunati uvrštavanjem potpornih vektora (primjera za koje je $\alpha_i > 0$ i koji zadovoljavaju ograničenje (3.30) kao jednakost):

$$b = y_s - \sum_{i=1}^{m} y_i \alpha_i \langle \mathbf{x_s}, \mathbf{x_i} \rangle \tag{3.38}$$

gdje indeks *s* predstavlja potporne vektore.

Klasifikator optimalne hiperravnine naziva se još i **linearni SVM s tvrdim marginama** koji je primjenjiv samo u slučaju linearno separabilnih podataka.

Prije nego se prijeđe na situaciju s linearno neseparabilnim podacima, potrebno je detaljnije se osvrnuti na vezu između margine i sposobnosti generalizacije.

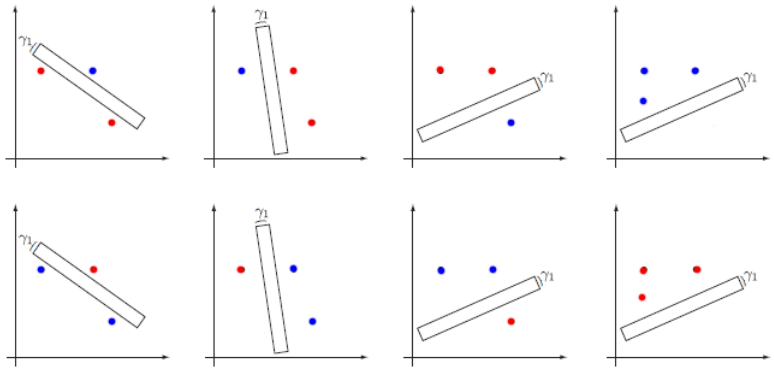
3.6. Veza između margine i VC dimenzije

Za implementaciju SRM principa indukcije u algoritmima učenja potrebno je kontrolirati dva faktora:

- empirijski rizik,
- kapacitet, odnosno potrebno je konačnu hipotezu birati iz takvog skupa funkcija čija VC dimenzija odgovara veličini skupa za učenje.

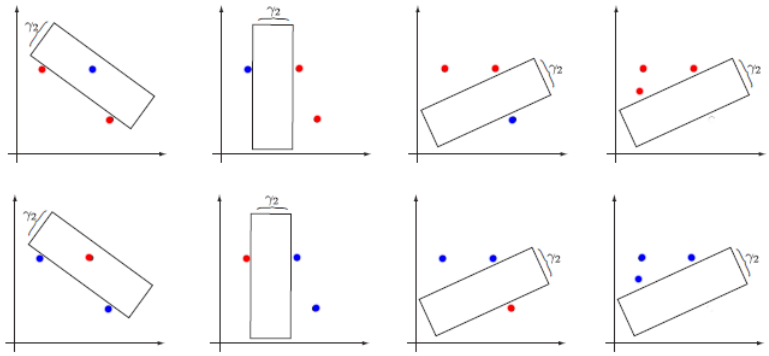
VC dimenzija skupa hiperravnina jednaka je $n + 1$, gdje je n dimenzionalnost ulaznog prostora. Međutim, VC dimenzija skupa razdvajajućih hiperravnina sa širokim marginama može biti manja od $n + 1$, ukazuje Vapnik (1999.).

Širokim marginama smanjuje se broj točaka koje mogu biti razdijeljene. Na slici 14. realizirano je svih 8 načina označavanja zahvaljujući uskim marginama. Međutim, na slici 15. vidljivo je da zbog širih margina to više nije moguće postići.



Slika 14. *VC dimenzija i uske margine*

Izvor: Hamel (2009.)



Slika 15. *VC dimenzija i široke margine*

Izvor: Hamel (2009.)

To sugerira da se margina ρ može koristiti za kontroliranje kompleksnosti modela. Razdvajajuća hiperravnina s empirijskim rizikom $R_{emp} = 0$ i s maksimalnom marginom ρ imati će najmanju VC dimenziju h što dovodi do sljedećeg teorema (Vapnik, 1999.):

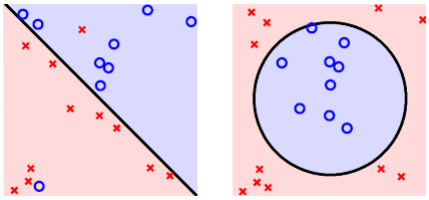
neka ulazni prostor čini sfera radijusa R u $\mathbb{R}^n$ tako da $\|\mathbf{x}\| \leq R$. Tada skup razdvajajućih hiperravnina s marginom ρ ima VC dimenziju h ograničenu s:

$$h \leq \min\left(\left[\frac{R^2}{\rho^2}\right], n\right) + 1 \tag{3.39}$$

Ono što je važno kod granice postavljene na ovakav način jest to što VC dimenzija više ne ovisi eksplicitno o dimenzionalnosti ulaznog prostora. To znači da, ako se podaci mapiraju čak i u beskonačno dimenzionalni prostor, dok god se koriste široke margine, rezultirati će dobrom generalizacijom. Time je uspostavljena presudna veza između margine i sposobnosti generalizacije koja ima važnu ulogu u konstruiranju SVM-a (Abu-Mostafa, Magdon-Ismail i Lin, 2012.).

3.7. Neseparabilni podaci

Podaci mogu biti neseparabilni na dva načina. Prva situacija (slika 16. lijevo) povezana je s problemom prisutnosti šuma u podacima. Kad bi se u ovakvoj situaciji inzistiralo na savršenoj razdvojivosti, optimizacijski problem optimalne hiperravnine ne bi imao rješenje. Međutim, ako bi se dopustilo određeno kršenje margine ili čak i to da određeni broj primjera bude pogrešno klasificiran, i dalje bi bilo moguće konstruirati linearnu granicu odlučivanja, ali na način da se klasifikator modificira iz onoga s tvrdim marginama u klasifikator s mekim marginama (poglavlje 3.8).



Slika 16. *Neseparabilni podaci: ako se dopusti određeno toleriranje greške i dalje je moguće konstruirati linearnu granicu odlučivanja (lijevo), situacija u kojoj nije moguće konstruirati linearnu granicu odlučivanja (desno)*

Izvor: Abu-Mostafa, Magdon-Ismail, i Lin (2012.)

U drugoj situaciji (slika 16. desno) linearni klasifikator nije primjenjiv na originalnim podacima, stoga se vrši njihova nelinearna transformacija i mapiranje u prostor veće dimenzionalnosti, nakon čega se je u tom novom prostoru ponovno omogućeno konstruiranje linearnog klasifikatora (poglavlje 3.9).

3.8. Linearni SVM s mekim marginama

Kako navode Schoelkopf i Smola (2002.) razdvajajuća hiperravnina u praksi možda neće postojati ako velika količina šuma u podacima dovede do preklapanja klasa. Također, ako takva hiperravnina i postoji, možda neće uvijek biti najbolje rješenje budući da bi zbog šuma rješavanje optimizacijskog problema moglo dovesti do uskih margina i pojave *overfittinga,* što smanjuje sposobnost generalizacije. Zbog toga je potrebno pronaći algoritam koji će tolerirati određenu količinu *outliera*. Rješenje je tzv. "*soft*" (meka) formulacija optimalne hiperravnine koja predstavlja pokušaj da se dobije hiperravnina sa širokim marginama, ali i dopusti određeno kršenje margine ili čak i poneki krivo klasificirani primjer.

Kad bi zadatak bio pronaći hiperravninu koja vodi minimalnom broju grešaka treniranja, to bi predstavljalo kombinatorijalni problem. Dokazano je da je problem pronalaženja hiperravnine, čija je greška treniranja veća za neki konstantni faktor u odnosu na optimalnu, *NP-hard*[13] (Schoelkopf i Smola, 2002.). Ali Cortes i Vapnik, (1995.) odabrali su drugi pristup za SVM. Uvode se dopunske (eng. *slack)* varijable:

$$\xi_i \geq 0, i = 1, ..., m \tag{3.40}$$

koje relaksiraju ograničenja (3.13).

Ako se ξ_i učini dovoljno velikim, odnosno ako se dopusti veliko kršenje margine, ograničenje na $(\mathbf{x_i}, y_i)$ uvijek može biti zadovoljeno. Kako svaki ξ_i ne bi poprimio previše veliku vrijednost, potrebno je kazniti kršenje margine dodavanjem ograničenja u funkciju cilja čime se dobiva sljedeći optimizacijski problem (Schoelkopf i Smola, 2002.):

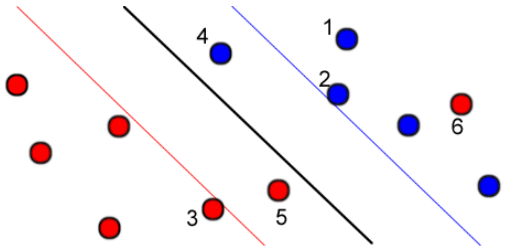
$$\min_{\mathbf{w},\boldsymbol{\xi}} \ \tau(\mathbf{w}, \boldsymbol{\xi}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\xi_i \tag{3.41}$$

uz ograničenja

$$y_i\left(\langle \mathbf{x_i}, \mathbf{w}\rangle + b\right) \geq 1 - \xi_i, \ \ i = 1, ..., m \tag{3.42}$$

13 NP-hard označava problem koji nije rješiv u polinomijalnom vremenu, odnosno za čije rješavanje nije pronađen efikasan algoritam.

$$\xi_i \geq 0, i = 1, ..., m \tag{3.43}$$



Slika 17. *Meke margine: primjeri označeni brojevima 4 i 5 krše marginu, ali i dalje su ispravno klasificirani. Primjer označen brojem 6 pogrešno je klasificiran.*

Izvor: izradila autorica

S obzirom na vrijednost varijabli *slacka*, moguće su sljedeće situacije:

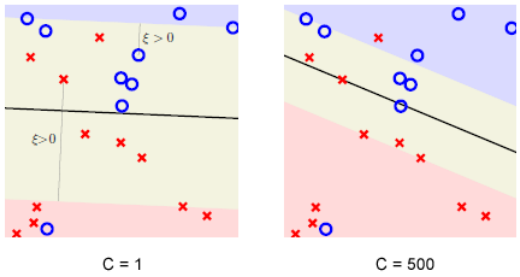
a) ako je $\xi_i = 0$, uzorak se nalazi s ispravne strane margine,

b) ako je $\xi_i > 0$, uzorak predstavlja grešku margine (eng. *margin error*),

c) ako je $\xi_i > 1$, tada se uzorak nalazi i s pogrešne strane hiperravnine.

Klasifikator koji dobro generalizira pronalazi se kontroliranjem kapaciteta (putem $\|\mathbf{w}\|$) i sume *slacka* $\sum_{i=1}^{m} \xi_i$, koja predstavlja gornju granicu na broj grešaka treniranja, pri čemu konstanta $C > 0$ određuje kompromis između dva konfliktna cilja: maksimizacije margine i minimizacije greške treniranja (Schoelkopf i Smola, 2002.). Za $C = \infty$ podaci će biti savršeno razdvojeni (uz pretpostavku da takva hiperravnina uopće postoji). Za konačne vrijednosti konstante C dopušta se određeno kršenje margine pri čemu veliki C daje veću važnost ispravnom klasificiranju svih podataka, dok manji C znači veću fleksibilnost (Burges, 1998.). No, s druge strane, veći C stvara uske margine i potencijalno vodi većoj opasnosti od *overfittinga*, dok će manji C rezultirati širim marginama i većim brojem potpornih vektora (što znači da će veći broj podataka bit će uključen u određivanje hiperravnine) (Hamel, 2009.). Vrijednost konstante C određuje korisnik uobičajeno putem unakrsne validacije[14].

14 Unakrsna validacija – vidi poglavlje 5.8.2.



Slika 18. *Utjecaj konstante C na širinu margine: mala vrijednost konstante C stvara široke margine (lijevo), velika vrijednost stvara uske margine (desno)*

Izvor: Abu-Mostafa, Magdon-Ismail, i Lin (2012.)

Problem se dalje rješava kao i separabilni problem. Dualni optimizacijski problem za hiperravninu s mekim marginama postaje (Schoelkopf i Smola, 2002.):

$$\max_{\alpha} \; W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j}^{m} \alpha_i \alpha_j y_i y_j \langle \mathbf{x_i}, \mathbf{x_j} \rangle \tag{3.44}$$

uz ograničenja

$$\sum_{i=1}^{m} \alpha_i y_i = 0 \tag{3.45}$$

i

$$0 \leq \alpha_i \leq C, \;\; i = 1, ..., m. \tag{3.46}$$

Optimalno rješenje zadovoljava KKT uvjete optimalnosti. Uvjet komplementarnosti sada je:

$$\alpha_i [y_i \left(\langle \mathbf{x_i}, \mathbf{w} \rangle + b \right) - 1 + \xi_i] = 0, \tag{3.47}$$

$$\xi_i \left(\alpha_i - C \right) = 0, \;\; i = 1, ..., m \tag{3.48}$$

Razlika u odnosu na slučaj s tvrdim marginama jest postavljanje gornje granice na Lagrangeove multiplikatore α_i. Time se ograničava utjecaj pojedinačnih primjera (koji bi mogli biti *outlieri*) na rezultat (Schoelkopf i Smola, 2002.). Slack varijable biti će $\xi_i \neq 0$ samo za $\alpha_i = C$. Uzorci za koje vrijede takve *slack* varijable su oni koje krše marginu tj. njihova je udaljenost od hiperravnine manja od $1/\|\mathbf{w}\|$. Uzorci za koje vrijedi $0 < \alpha_i < C$ jesu oni koje leže na udaljenosti od točno $1/\|\mathbf{w}\|$. Za uzorke na većoj udaljenosti vrijedi $\alpha_i = 0$. Samo potporni vektori imaju $\alpha_i > 0$ i samo oni sudjeluju u određivanju optimalne hiperravnine.

Rješenje za $\mathbf{w}$ dobije se jednako kao i u separabilnom problemu.

3.9. Nelinearni SVM

Razdvajajuće hiperravnine nisu dovoljno fleksibilne za postizanje niskog empirijskog rizika u mnogim stvarnim problemima, ali postoje dva načina kojima je moguće povećati njihovu fleksibilnost (Vapnik, 1999.):

- korištenjem bogatijeg skupa funkcija,
- mapiranjem ulaznih vektora u visoko dimenzionalni prostor značajki i konstruiranjem optimalne hiperravnine u tom prostoru.

Drugi način koristi SVM.

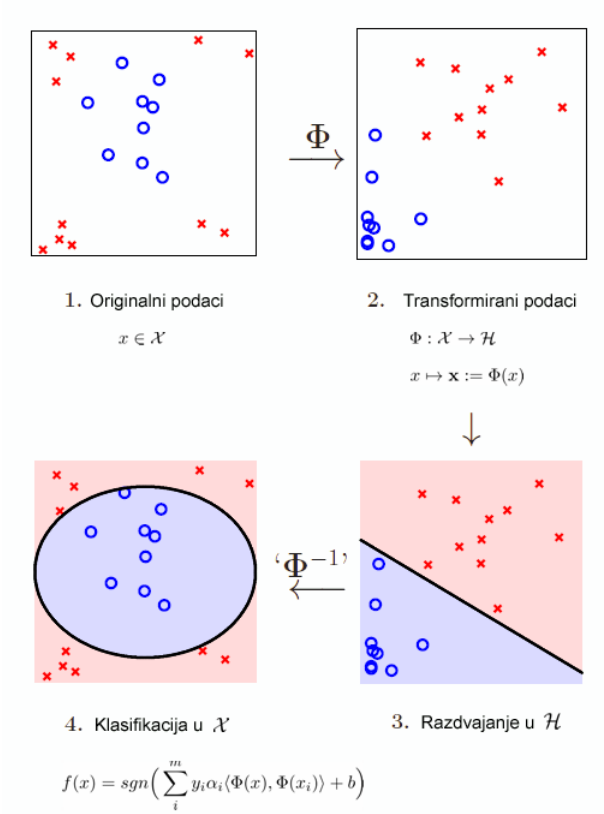
Zahvaljujući činjenici da za linearne modele nužnost predstavlja samo linearnost u parametrima, ali ne i u varijablama, omogućeno je korištenje nelinearnih transformacija i otvoren put konstruiranju SVM-a primjenom sljedeće ideje (Vapnik, 1999.) (slika 19.): najprije se vektori ulaznih podataka mapiraju u visoko dimenzionalni prostor $\mathcal{H}$ na kojem je definiran unutarnji produkt, pri čemu se nelinearno mapiranje odabire *a priori* i to na temelju poznavanja i razumijevanja prirode problema. Zatim se u tom novom prostoru konstruira razdvajajuća hiperravnina čija je VC dimenzija određena s R^2/ρ^2. Kako bi se postigla dobra sposobnost generalizacije, smanjuje se VC dimenzija konstruiranjem optimalne hiperravnine koja maksimizira marginu ρ. Upravo korištenje visoko dimenzionalnog prostora omogućava stvaranje većih margina.

Transformacija se vrši mapiranjem Φ:

$$\begin{aligned} \Phi &: \mathcal{X} \to \mathcal{H} \\ x &\mapsto \mathbf{x} := \Phi(x). \end{aligned} \tag{3.49}$$

Transformirani podaci nazivaju se značajke (eng. *feature*), a novi, prošireni prostor naziva se prostor značajki (eng. *feature space)*. Decizijska funkcija sada ima oblik:

$$f(x) = sgn\Big(\sum_{i}^{m} y_i \alpha_i \langle \Phi(x), \Phi(x_i) \rangle + b\Big) \tag{3.50}$$



Slika 19. *Mapiranje u visoko dimenzionalni prostor značajki: korištenje nelinearnih transformacija ulaznih podataka omogućava stvaranje linearnih granica odlučivanja u prostoru značajki koje odgovaraju nelinearnim granicama odlučivanja u ulaznome prostoru*

Izvor: prilagođeno prema (Abu-Mostafa, Magdon-Ismail i Lin, 2012.)

Nelinearnim transformacijama omogućeno je stvaranje sofisticiranih granica odlučivanja u prostoru ulaznih podataka s ciljem smanjenja empirijskog rizika. Međutim, ovdje se javlja jedan drugi problem, a to je problem računanja u visoko dimenzionalnom prostoru. Pitanje koje se sad postavlja jest kako izvršiti transformaciju i računati unutarnji produkt na efikasan način, a da to bude neovisno o dimenzionalnosti prostora značajki?

3.9.1. Kernel trik

Na rješenje problema ukazali su Cortes i Vapnik, (1995.). Kako se za opisivanje optimalne hiperravnine u prostoru značajki i estimaciju odgovarajućih koeficijenata razdvajajuće hiperavnine koristi unutarnji produkt dvaju vektora $\Phi(\mathbf{x_1})$ i $\Phi(\mathbf{x_2})$, koji su slike ulaznih vektora $\mathbf{x_1}$, $\mathbf{x_2}$ u prostoru značajki, njihov se unutarnji produkt u tom prostoru može računati kao funkcija dviju varijabli u ulaznome prostoru. Funkcija koja kombinira

transformaciju i unutarnji produkt naziva se kernel funkcija ili jednostavno kernel:

$$\begin{aligned} \mathrm{k}\colon \mathrm{X} \times X &\to \mathbb{R} \\ (x, x') &\mapsto k(x, x') \end{aligned}$$

Korišenjem kernela k svugdje gdje se u algoritmu za učenje pojavljuje unutarnji produkt tako da (Schoelkopf i Smola, 2002.) :

$$k\left(x, x_i\right) := \langle \mathbf{x}, \mathbf{x_i} \rangle = \langle \Phi\left(x\right), \Phi\left(x_i\right) \rangle \tag{3.51}$$

vodi decizijskoj funkciji oblika

$$f\left(x\right) = sgn\left(\sum_{i=1}^{m} y_i \alpha_i \langle \Phi\left(x\right), \Phi\left(x_i\right) \rangle + b\right) = sgn\left(\sum_{i=1}^{m} y_i \alpha_i k(x, x_i) + b\right). \tag{3.52}$$

Optimizacijski problem tada postaje

$$\max_{\boldsymbol{\alpha}} \; W(\boldsymbol{\alpha}) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j k(x_i, x_j) \tag{3.53}$$

uz ograničenja

$$0 \leq \alpha_i \leq C, \;\; i = 1, ..., m \tag{3.54}$$

i

$$\sum_{i=1}^{m} \alpha_i y_i = 0. \tag{3.55}$$

Primjena kernela, kojima se vrši implicitno mapiranje u visokodimenzionlani prostor te se omogućava konstruiranje nelinearnih decizijskih funkcija u ulaznom prostoru ekvivalentnima linearnim decizijskim funkcijama u prostoru značajki, i to iskorištavajući činjenicu da se u dualnom optimizacijskom problemu ulazni podaci javljaju samo u obliku unutarnjeg produkta, u literaturi je poznata pod nazivom "kernel trik".

3.9.2. Kerneli

Da bi neka funkcija bila ispravni kernel, mora zadovoljiti Mercerov uvjet[15]. Tada postoji mapiranje Φ i vrijedi $k\left(x, x_i\right) = \langle \Phi\left(x\right), \Phi\left(x_i\right) \rangle$, a rezultirajući je problem konveksan. Ako se

15 Kernel se može zapisati u obliku matrice, nazvane Gram matrica, $K_{ij} = \langle \Phi\left(x\right), \Phi\left(x'\right) \rangle = k(x, x')$ koja se dobije evaluacijom kernela na svim primjerima za učenje. Prema Mercerovom uvjetu $k(x, x')$ je ispravni kernel ako i samo ako je Gram matrica K simetrična i pozitivno semidefinitna za svaki $x \in \mathcal{X}$ (Abu-Mostafa, Magdon-Ismail i Lin, 2012.).

koristi kernel koji ne zadovoljava Mercerov uvjet, za neke primjere problem kvadratnog programiranja neće imati rješenje.

U praksi se najčešće koriste sljedeći kerneli:

- polinomijalni

$$k(x, x_i) = \langle x, x_i \rangle^d \tag{3.56}$$

gdje je d slobodni parametar koji određuje stupanj polinoma,

- Gaussova radijalna bazna funkcija (eng. *Radial basis function* - RBF)

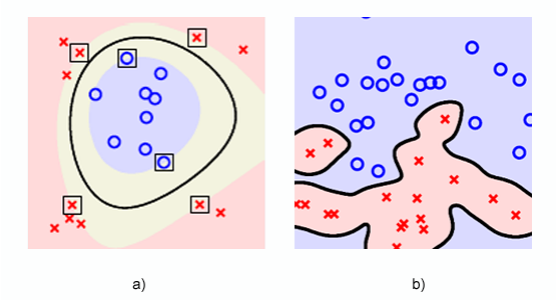
$$k(x, x_i) = \exp\left(-\frac{\|x - x_i\|^2}{2\sigma^2}\right) \tag{3.57}$$

gdje je $\sigma > 0$ slobodni parametar kojim se određuje širina (u zapisima se koristi i $\gamma = -1/2\sigma^2$),

- dvoslojna sigmoidalna neuronska mreža

$$k(x, x_i) = \tanh(\kappa \langle x, x_i \rangle + \theta) \tag{3.58}$$

sa slobodnim parametrima κ i θ koji određuju prirast (eng. *gain*) i pomak.



Slika 20. *Granice odlučivanja u ulaznom prostoru dobivene primjenom različitih kernela: a) polinomijalnim, b)RBF kernelom.*

Izvor: Abu-Mostafa, Magdon-Ismail, i Lin (2012.)

Iako neki autori naglašavaju važnost odabir kernela za postizanje dobrih performansi u rješavanju problema, Schoelkopf i Smola (2002.) navode da u tome ipak presudnu ulogu ima podešavanje njihovih slobodnih parametara. Empirijski je utvrđeno da, uz dobro podešene parametre, različiti kerneli postižu sličnu razinu točnosti.

Korištenje kernela općenito predstavlja korištenje veće klase funkcija, odnosno

povećanje kapaciteta učećega stroja, čime problem postaje separabilan. Pa iako neki kerneli imaju beskonačnu VC dimenziju (kao što je to npr. RBF kernel), dobra generalizacijska sposobnost SVM-a proizlazi iz mogućnosti kontrole oba člana u (2.4).

4. PREDVIĐANJE FINANCIJSKIH VREMENSKIH NIZOVA

"Predviđati je teško. Pogotovo budućnost."

Niels Bohr

Financijska tržišta[16] kompleksni su sustavi podložni raznim ekonomskim i neekonomskim utjecajima poput općih gospodarskih kretanja, političkih događaja, ali i ljudskih očekivanja te raznih psiholoških stanja njihovih sudionika. Upravo iz nemogućnosti modeliranja svih ovih faktora od utjecaja izvire jedan od glavnih problema predviđanja. Drugome pridonosi konstantna evolucija tržišta.

Velike količine šuma u podacima povezane s nestacionarnošću[17] financijskih vremenskih nizova čine zadatak predviđanja iznimno zahtjevnim, a ponekad i nemogućim, složit će se mnogi autori (Cao i Tay, 2001; Kara et al., 2011; Hellstroem i Holmstroem, 1998.). Stoga se za uspješnost predviđanja uobičajeno naglašava nužnost pažljivog odabira odgovarajućih inputa, ali ona u još većoj mjeri ovisi o mogućnosti da se utvrdi:

- postoje li uzorci u prošlim podacima,
- ako postoje, može li ih se pronaći,
- ako se mogu pronaći, hoće li se ponoviti u budućnosti.

Međutim, poteškoće u pronalaženju takve jasne veze između prošlosti i budućnosti nagnale su brojne ekonomiste da prihvate hipotezu efikasnog tržišta koja u svom najstrožem obliku kaže da je svaki pokušaj predviđanja cijena na financijskim tržištima osuđen na propast. Dok, s druge strane, neki od njih ipak pružaju ponešto nade smatrajući da je "*standardna*

16 Financijska tržišta su tržišta za trgovinu financijskim instrumentima. Ona utvrđuju obujam raspoloživih sredstava, mobiliziraju uštede, uspostavljaju cijene vrijednosnica i razinu kamatnih stopa.
Financijski instrument je isprava koja za imaoca inkorporira potraživanje prema zaradi ili imovini izdatnika – poduzeća, financijskih institucija, kućanstva ili države, a za izdatnika predstavlja obvezu.
Tržište kapitala je tržište na kojem se trguje financijskim instrumentima s rokom dospijeća preko godinu dana. Na njemu se pribavljaju sredstva za financiranje dugoročnih ulaganja.
Tržište vrijednosnica je tržište za emisiju i kasnije obično višestruke kupnje i prodaje financijskih instrumenata.
Vrijednosnica je financijska imovina koja se definira kao potraživanje na gotovinske tijekove od realne imovine – poslovnih i proizvodnih objekata, zemljišta, opreme i sl. (Vidučić, 2006.)

17 Vremenski niz je stacionaran ako ne sadrži trend, ako u njemu nisu prisutne periodske varijacije, te ako varijanca i korelacija ne ovise o vremenu.

pretpostavka slučajnog hoda cijena vrijednosnica samo veo slučajnosti koji prekriva nelinearni proces s mnoštvo šuma" (Tsaih et al., 1998.).

U nastavku se daje prikaz dvaju osnovnih pristupa u analizi tržišnih kretanja, rezultati kojih ujedno služe i kao izvori ulaznih podataka, nakon čega se daje osvrt na hipotezu efikasnog tržišta, te implikacija navedenih dvaju problema na izgradnju modela.

4.1. Fundamentalna i tehnička analiza

U analizi tržišnih kretanja investitoru stoje na raspolaganju dva osnovna pristupa: fundamentalna i tehnička analiza. Premda oba pristupa pokušavaju riješiti isti problem, a to je odrediti smjer u kojem će se u budućnosti najvjerojatnije kretati cijene, među njima postoje i znatne razlike.

4.1.1. Fundamentalna analiza

Najopćenitije rečeno, fundamentalna se analiza bavi proučavanjem uzroka tržišnih kretanja, odnosno ekonomskim silama ponude i potražnje koje dovode do promjena cijena vrijednosnica kako bi se odredila njihova intrinzična vrijednost (Murphy, 1999.). Ako je ta vrijednost ispod trenutne tržišne cijene, vrijednosnica je precijenjena i trebalo bi ju prodati. Ako je tržišna cijena ispod intrinzične vrijednosti, tada je podcijenjena i trebalo bi kupovati.

Tradicionalni pristup fundamentalnoj analizi temeljio se prvenstveno na analizi financijskih izvještaja, nadopunjenoj makroekonomskim faktorima (poput zaposlenosti, BDP-a, indeksa cijena,....) te analizi vijesti, dok se u novije vrijeme sve više koriste masivni izvori podataka nastali kao rezultat interakcija putem interneta poput blogova, recenzija ili društvenih mreža. Tako, na primjer, Preis, Moat i Stanley (2013.) analiziraju promjene u broju pretraživanja određenih financijskih pojmova koristeći *Google Trends*[18], te na temelju pronađenih uzoraka, koji se mogu interpretirati kao rani znakovi upozorenja promjena kretanja na tržištu, zaključuju da *Google Trends*, osim što reflektira trenutno stanje gospodarstva, može pružiti uvid u buduće trendove u ponašanju gospodarskih subjekata.

18 Google Trends je servis baziran na Googleovom pretraživaču koji omogućava prikaz promjene broja pretraživanja nekog pojma tijekom vremena.

Slično tome Si et al. (2013.) koriste analizu sentimenta[19] *tweetova*[20] kao pomoć u predviđanju tržišnih kretanja, a analiza sentimenta kombinirana s generiranjem signala za kupnju ili prodaju već se nudi kao komercijalna usluga.

Međutim, bez obzira na napredak tehnologije, još je uvijek teško trgovati isključivo korištenjem takvih informacija. Financijski izvještaji često nisu dovoljno pouzdani, makroekonomski pokazatelji nisu prikladni za velike brzine trgovanja, a masivni podaci nisu svima dostupni budući da su njihovo pribavljanje i analiza skupi .

4.1.2. Tehnička analiza

Tehnička je analiza, s druge strane, usmjerena na proučavanje posljedica djelovanja ekonomskih sila ponude i potražnje, odnosno bavi se proučavanjem povijesnih kretanja cijena i volumena trgovanja[21]. Analitičari koji primjenjuju tehničku analizu ne zamaraju se njihovim uzrocima, već svoju analizu baziraju na sljedećim premisama (Murphy, 1999.):

- Svi fundamentalni faktori sadržani su u cijeni. Analitičari koji se oslanjaju na tehničku analizu vjeruju da je u tržišnoj cijeni već reflektirano sve što na nju može utjecati. Stoga je dovoljno proučavati kretanje cijena.

- Cijene se kreću u trendovima. Trend je opći smjer kojim se kreće cijena neke vrijednosnice. Koncept trenda zauzima važno mjesto u tehničkoj analizi budući da je većina njezinih tehnika "*trend-following*", odnosno njihov je smisao identificirati i slijediti postojeći trend sve dok ne pokaže znakove zaokreta.

- Povijest se ponavlja. Ključ razumijevanja budućnosti leži u proučavanju prošlosti. Da bi se budućnost mogla predviđati na temelju prošlosti, u povijesnim podacima moraju postojati uzorci koji će se nastaviti ponavljati i u budućnosti.

Osnove tehničke analize postavio je Charles Dow na prijelazu iz 19. u 20. stojeće. U svojim začecima oslanjala se prvenstveno na analizu grafikona i tehničkih indikatora, dok se

19 Analizom sentimenta određuje se sentiment nekog tekstualnog sadržaja, odnosno ima li neki tekst pozitivnu ili negativnu konotaciju. Analizom velikih količina takvih sadržaja nastoji se odrediti npr. opći stav tržišta prije nego se on odrazi u cijenama vrijednosnica.

20 *Tweeter* je društvena mreža namijenjena slanju i čitanju kratkih poruka. Jedna takva poruka naziva se *tweet*.

21 i otvorenog interesa koji se odnosi samo na *futures* ugovore i opcije.

danas pristaše tehničke analize mogu podijeliti u dvije grupe (Murphy, 1999.):

- Prvu grupu čine oni koji i dalje preferiraju tradicionalni pristup analize grafikona[22] no, ovakav se pristup smatra više umjetnošću nego znanošću zbog velike doze subjektivnosti.

- Drugu grupu čine kvantitativni analitičari[23] koji razvijaju algoritamske sustave trgovanja s osnovnim ciljem eliminiranja ljudske subjektivnosti. Dalje se dijele na:

    - one koji koriste tehnologiju za razvijanje boljih tehničkih indikatora što omogućava zadržavanje kontrole nad interpretacijom indikatora i samog procesa odlučivanja,

    - one koji preferiraju "*black box*"[24] pristup.

Na potonjem se pristupu temelji i ovaj rad.

4.2. Hipoteza efikasnog tržišta

Već je prethodno naglašen značaj financijskih tržišta za dobro funkcioniranje gospodarstva neke zemlje. Osim što pružaju važne usluge investitorima, svojom stabilnošću ili poremećajima mogu utjecati i na dugoročnu sigurnost čitave regije. Međutim, da bi neko tržište moglo ispunjavati sve svoje funkcije, tržišne cijene trebaju u potpunosti odražavati sve raspoložive informacije. Na takvom, kako ga je Fama (1970.) nazvao, efikasnom tržištu, poduzeća mogu donositi ispravne odluke o proizvodnji i investicijama, a investitori mogu donositi odgovarajuće odluke o sastavljanju portfelja vrijednosnica.[25]

Na efikasnom tržištu cijene vrijednosnica mijenjaju se samo s dotokom novih informacija. Nove informacije šire se velikom brzinom, a podešavanje cijena, kao reakcija na njih, je trenutno.[26] To znači da niti tehnička analiza, niti fundamentalna analiza ne mogu

22 Tzv. *"chartist"* što dolazi od riječi *chart* koja na engleskom jeziku znači grafikon.
23 Popularno nazvani *"quants"*.
24 Black box – znači bez poznavanja unutarnjeg funkcioniranja sustava.
25 Ovakva efikasnost temelji se na informacijskoj efikasnosti.
26 Podešavanje je trenutno i zbog toga što jaki oblik tržišne efikasnosti pretpostavlja nepostojanje transakcijskih troškova, troškova pribavljanja informacija ili troškova prilagodbe cijena informacijama, dok prema slabijem obliku efikasnosti cijene odražavaju informacije do točke u kojoj marginalne koristi njihove upotrebe za stvaranje profita ne prelaze marginalne troškove njihovog pribavljanja (Fama, 1991.).

pomoći investitoru da otkrije podcijenjene dionice koje bi mu omogućile ostvarivanje natprosječnog profita bez da ujedno prihvati i natprosječni rizik (Malkiel, 2003.b).

Kako taj mehanizam djeluje može se vidjeti iz sljedećega. Na tržištima sudjeluje veliki broj investitora koji pokušavaju ostvariti profit temeljem informacija koje posjeduju. U nastojanju da iskoriste i najmanju informacijsku prednost, oni inkorporiraju informacije u tržišne cijene, što zatim dovodi do eliminacije inicijalne prilike za ostvarivanje ekstra profita. Ako se informacije trenutno inkorporiraju u cijene, sutrašnje promjene cijena reflektirati će samo sutrašnje vijesti i biti će nezavisne od današnjih promjena. Stoga su na potpuno efikasnim tržištima cijene potpuno slučajne i nepredvidive.[27] Upravo zbog toga hipoteza efikasnog tržišta povezana je s idejom "slučajnog hoda" prema kojoj niz cijena karakterizira svojstvo da svaka sljedeća promjena predstavlja slučajno odstupanje od one prethodne. Ovakva ekstremna hipoteza efikasnog tržišta predstavlja svojevrsnu idealizaciju koja nije ekonomski ostvariva, ali služi kao *benchmark* za mjerenje relativne efikasnosti usporedbom različitih tržišta (Lo, 2007.).

4.2.1. Predvidljivost prinosa

Hipotezu efikasnog tržišta razvili su, nezavisno jedan od drugoga, Paul A. Samuelson i Eugene F. Fama 60-ih godina 20. stoljeća, a od tada izaziva brojne kontroverze. Unatoč brojnim provedenim istraživanjima i objavljenim radovima, među ekonomistima još nije postignut konsenzus jesu li tržišta efikasna ili nisu (Lo, 2007.). Upravo brojne anomalije[28] uočene u kretanjima prinosa (poput kalendarskih učinaka[29]) ukazuju na predvidljivost prinosa na temelju povijesnih podataka, što je u suprotnosti s ranijim radovima u kojima je uobičajena hipoteza da je očekivani prinos konstantan u vremenu, što tada implicira da je najbolje predviđanje njihova povijesna sredina (Fama, 1991.).

Početkom 21. stoljeća među mnogim ekonomistima i statističarima javlja se uvjerenje o

27 U svojoj znamenitoj knjizi, koja je popularizirala hipotezu slučajnog hoda prinosa, Malkiel (2003.a) uspoređuje majmuna i brokera koji s jednakim uspjehom biraju dionice, aludirajući na nemogućnost pobjeđivanja tržišta primjenom aktivne ulagačke strategije.

28 Anomalije u prinosima dionica mogu se definirati kao odstupanja ponašanja prinosa dionica od uobičajenog ponašanja kojeg predviđa i opisuje neka općeprihvaćena teorija (Škrinjarić, 2012.)

29 Kalendarski učinci predstavljaju vremenske anomalije u prinosima dionica. Jedna takva anomalija je i učinak siječnja kojeg karakteriziraju natprosječno veliki prinosi u mjesecu siječnju u odnosu na prinose u ostalim mjesecima (Škrinjarić, 2012.).

barem djelomičnoj predvidljivosti prinosa, posebno zahvaljujući promjeni u shvaćanju ponašanja tržišnih sudionika (Malkiel, 2003.b). Za razliku od klasične financijske teorije koja pretpostavlja racionalnog investitora, bihevioralni ekonomisti naglašavaju psihološke elemente u ponašanju koji utječu na proces formiranja cijena, poput pohlepe ili straha (Lo et al., 2005.) koji onda dovode do pretjeranih reakcija, pretjeranog samopouzdanja, pogrešne procjene rizika, averzije prema gubitku te se u konačnici mogu u znatnoj mjeri negativno odraziti na ekonomsko blagostanje pojedinca (Lo, 2007.). Uz to, poznate su strategije koje se baziraju na kupovanju gubitnika, a prodaji dobitnika .

No unatoč tome, čak i kada se uzme u obzir iracionalnost pojedinca, Malkiel (2003.b) zaključuje da su tržišta ipak efikasnija i manje predvidiva nego što se stječe dojam na temelju recentnijih akademskih istraživanja. Čak i kada postoje anomalije u kretanju cijena, one ne stvaraju prilike koje bi omogućile investitoru ostvarivanje natprosječnog profita korigiranoga za rizik[30]. Ako se i utvrdi neka anomalija koja bi bila dovoljno predvidljiva i postojana, ubrzo dolazi do njezinog samouništenja. Čim se vijest o njoj objavi i ona postane široko poznata, izgubi se mogućnost njezinog iskorištavanja.

Barbić (2010.a) ističe dodatne specifičnosti tržišta kapitala u razvoju koje karakterizira slabija likvidnost, viši transakcijski troškovi i manji volumen trgovanja što otvara prostor manipulacijama velikih igrača. Osim toga, sudionici na takvim tržištima lošije su informirani, te skloniji neracionalnim odlukama. Premda testiranja takvih tržišta[31] uglavnom ukazuju na odbacivanje hipoteze o slučajnom hodu, njihovi sudionici, zbog prethodno navedenih razloga, uglavnom nisu u mogućnosti iskoristiti uočene neefikasnosti.

4.2.2. Testovi i razine efikasnosti

Testiranje efikasnosti financijskih tržišta provodi se s obzirom na određeni informacijski skup pri čemu se razlikuju (Fama, 1970.)[32]:

- Testovi slabog oblika. Informacijski skup predstavljaju povijesni podaci o cijenama, a

30 Prinos i rizik uvijek treba promatrati zajedno.

31 Testiranja tržišta u razvoju ograničena su na testove slabog oblika efikasnosti.

32 U naknadnom radu (Fama, 1991.) ova je podjela ponešto izmijenjena. Testovi slabog oblika nazvani su općenitije testovi predvidljivosti prinosa, pri čemu se, osim povijesnih cijena uzimaju u obzir i dodatne varijable poput dividendi i kamatnih stopa. Testovi polu-jakog oblika nazvani su studijama događaja (eng. *event study*), dok su testovi jakog oblika preimenovani u testove privatnih informacija.

ovaj oblik efikasnosti pobija mogućnost ostvarivanja natprosječnih prinosa na temelju tehničke analize.

- Testovi polu-jakog oblika. Bave se pitanjem brzine podešavanja cijena s obzirom na sve javno raspoložive informacije (objave godišnjih zarada, spajanja ili podjele dionica,...). Prema ovom obliku nije moguće ostvariti natprosječne prinose niti pomoću fundamentalne analize.

- Testovi jakog oblika. Bave se pitanjem posjeduju li pojedinci ili grupe investitora monopolski pristup informacijama relevantnima za formiranje cijena. Ovaj oblik efikasnosti podrazumijeva da nije moguće ostvariti natprosječni profit čak niti temeljem povlaštenih informacija.

Kategorizacija testova efikasnosti služi kako bi se ustanovila razina informacija na kojoj se hipoteza efikasnog tržišta odbacuje. Testiranje započinje testovima slabog oblika. Ako se utvrdi da testovi podržavaju hipotezu na toj razini, nastavlja se s testovima polu-jakog oblika (Barbić, 2010.a). Testiranje jakog oblika nije realno izvedivo.

4.3. Predviđanje "gotovo" slučajnog niza

Kako se prethodno navedeno odražava na pokušaje izgradnje i ocjene modela za potrebe predviđanja financijskih vremenskih nizova ukazuju Abu-Mostafa i Attya (1996.). Prilikom modeliranja nekog sustava svi faktori izostavljeni iz modela predstavljaju šum. S obzirom na brojnost faktora od utjecaja na cijene dionica, odnosno na nemogućnost njihovog potpunog obuhvaćanja, te da se modeliranje uglavnom vrši na temelju ograničenog skupa podataka, velika količina šuma dovodi do situacije u kojoj se razina podudaranja između stvarnog izlaza i onoga predviđenoga modelom kreće tek ponešto iznad 50%. Abu-Mostafa et al. (2012.) stoga navode da se kod takve vrste problema već i rezultat od 54% može smatrati prihvatljivim pa čak i dobrim.

Potkrijepljeno jednim primjerom (Hellstroem i Holmstroem, 1998): ako se predviđa kretanje prinosa dionice i uz pretpostavku da se u podacima nalazi jednak broj pozitivnih i negativnih vrijednosti u razdoblju od jedne godine, odnosno 250 dana trgovanja, vjerojatnost da će x predviđanja algoritma baziranog u potpunosti na slučaju biti točna dana je sa:

$$P(hit\ \ rate = x) = \binom{250}{x} 0.5^x * 0.5^{250-x} \tag{4.1}$$

Ako se zahtijeva da podudaranje bude iznad 54%, uvrštavanjem $x = 0,54 * 250 = 135$ u

$$P(hit\ \ rate > x) = 1 - P(hit\ \ rate <= x) \tag{4.2}$$

dobije se $P(hit\ \ rate > 135) \approx 0.092$ kao vjerojatnost da će slučajni algoritam imati *hit rate*[33] iznad 54%, odnosno rizik da će slučajni algoritam biti klasificiran kao korisni prediktivni model iznosi 9%. Ako se zahtijevana razina podudaranja spusti samo malo, na primjer na 52% taj će rizik biti znatno veći i iznositi će 24,33%.

Kako dalje Abu-Mostafa i Attya (1996.) objašnjavaju, ono što predstavlja glavnu poteškoću kod tako niske zahtijevane razine podudaranja u samo $(1/2 + \varepsilon) * m$ primjera jest činjenica da je broj funkcija koje to postižu zapravo jako velik. Previše je slučajnih funkcija koje djeluju kao dobri kandidati hipoteza na ograničenom skupu primjera. Međutim, u rangu performansi od 100%, odnosno tamo gdje se zahtijeva podudaranje u $(1 - \varepsilon) * m$ primjera, broj funkcija je ipak ograničen. Takva se razina očekuje npr. kod problema raspoznavanja znakova (OCR). Zbog toga nije rijetkost da poneka strategija trgovanja bazirana isključivo na slučaju daje određeni period dobre rezultate, ali je prilično neizgledno da će nešto slično biti moguće postići i kod raspoznavanja znakova.

Ipak, dovoljno dugim nizom moguće je donekle kompenzirati probleme uzrokovane prisutnošću velike količine šuma, no tada do izražaja dolazi problem nestacionarnosti financijskih vremenskih nizova. Ukupno uzevši, ova dva problema znače da podaci za treniranje uglavnom neće sadržavati dovoljno informacija potrebnih za uspješno učenje (Abu-Mostafa i Attya, 1996.).

33 Vidi poglavlje 5.9.1.

5. SUSTAV ZA PREDVIĐANJE KRETANJA NA TRŽIŠTIMA VRIJEDNOSNICA

U ovom poglavlju dan je opis aplikacije koja se naslanja na teorijsku osnovu predstavljenu u prethodnim poglavljima, popraćeno detaljnijom argumentacijom pozadine pojedinih odluka. Dva su osnovna cilja koja su vodila razvoj aplikacije: prvi se odnosi na pružanje potrebnih funkcionalnosti koje će zatim omogućiti da se spoznaje iz prethodnih poglavlja iskoriste za predviđanje kretanja na tržištima vrijednosnica, dok se drugi odnosi na izradu sučelja koje će biti intuitivno za korištenje i pružiti dovoljnu udobnost pri provođenju eksperimenta.

5.1. Faze u procesu predviđanja

Iako strojno učenje započinje izgradnjom modela, konačna uspješnost predviđanja u znatnoj mjeri ovisi o odgovarajućoj pripremi podataka. Stoga se ukupan proces obuhvaćen aplikacijom može podijeliti u dvije glavne faze:

1. priprema podataka za učenje koja obuhvaća:
    - prikupljanje podataka,
    - transformaciju podataka u ulazne i izlazne varijable,
    - odabir značajki te
2. učenje, odnosno izgradnja modela i njegovo testiranje.

Međutim, s obzirom da zahtjevnost problema traži intenzivno eksperimentiranje, bilo je nužno pronaći način za minimiziranje uzaludnih pokušaja. Chen i Navet (2007.) predlažu da se u situaciji velike neizvjesnosti i kada su troškovi pogrešne odluke veliki provede prethodno testiranje, što se, kako autori navode, često izostavlja. Korisnost provođenja određenih testova prije primjene algoritma za učenje sastoji se u tome što, uz uštedu vremena, prethodno testiranje omogućava razlikovanje dvaju mogućih razloga neuspjeha: prvi je povezan s mogućnošću da se u podacima pokušavaju pronaći uzorci koji u njima zapravo niti ne postoje, kao što bi to bilo u slučaju efikasnosti tržišta, a drugi je pogrešna primjena samog algoritma. Autori se u radu bave primjenom genetskog programiranja u oblikovanju strategija trgovanja,

za što predlažu četiri vrste prilagođenih testova, Zemke (2002.) dalje predlaže dodatne testove bazirane na teoriji informacija ili teoriji kaosa, dok je za potrebe ove aplikacije odabrano testiranje slabog oblika hipoteze efikasnog tržišta.

Drugo potrebno proširenje odnosi se na naknadnu simulaciju trgovanja na temelju rezultata predviđanja što bi trebalo pridonijeti dobivanju realističnije ocjene izgrađenoga modela. Općenito je namjena sličnih sustava za predviđanje njihovo korištenje u okviru kompleksnijih sustava algoritamskog trgovanja, s jasnim ciljem ostvarenja profita. Stoga uobičajene mjere kojima se ocjenjuju klasifikatori, poput točnosti predviđanja, nisu dostatne već ih, kako Zemke (2002.) predlaže, treba nadopuniti s onim financijskima, poput prinosa od trgovanja.

5.2. Pretpostavke za korištenje aplikacije

Inicijalna zamisao aplikacije uključivala je modul za preuzimanje podataka direktno s tržišta, no s obzirom da se čišćenje podataka ne može u potpunosti automatizirati, a naročito ako se podaci preuzimaju iz neprovjerenih izvora ili onih nad kojima se nema potpuna kontrola, od toga se ipak odustalo. Stoga se glavna pretpostavka aplikacije sastoji u tome da su sirovi podaci već prikupljeni i očišćeni te da se nalaze u bazi podataka spremni za upotrebu, što ujedno znači da je iz aplikacije izostavljen prethodno navedeni prvi korak.

5.3. Funkcionalni zahtjevi

Uzevši u obzir gore navedeno, glavne funkcionalne zahtjeve aplikacije, prema obuhvaćenim fazama procesa, čine:

1. U okviru priprema podataka za učenje potrebno je omogućiti:
   - pregled podataka te odabir njihove odgovarajuće podjele na skupove za treniranje i testiranje,
   - testiranje predvidljivosti vremenskog niza u ukupnom razdoblju te prema odabranim podjelama,
   - kreiranje i brisanje ulaznih i izlaznih varijabli, njihov grafički prikaz te nadopunu podataka ako se u bazu unesu novi sirovi podaci,

- odabir ulaznih varijabli (eng. *feaure selection*) koji treba obuhvatiti:
  - grafički prikaz rezultata,
  - mogućnost skraćivanja niza podataka radi bržeg provođenja preliminarnog testiranja,
  - odabir najbolje ili najgore rangiranih varijabli,
  - ponavljanje postupka odabira kroz nekoliko iteracija vođeno određenim kriterijem za eliminaciju varijabli,
  - kreiranje LibSVM datoteka.

2. U okviru odabira modela i učenja potrebno je omogućiti:
   - odabir kernela,
   - odabir parametara modela,
   - odabir metode evaluacije,
   - korištenje kompenzacije neravnoteže u podacima,
   - grafički prikaz rezultata pretraživanja parametara.

3. Evaluacija modela treba obuhvatiti:
   - testiranje klasifikatora na izdvojenom skupu podataka,
   - prikaz rezultata testiranja prema različitim mjerama evaluacije,
   - simulaciju trgovanja koja obuhvaća:
     - odabir strategije i veličine uloga,
     - mogućnost kombinacije nekoliko klasifikatora,
     - grafički i tabelarni prikaz rezultata.

5.4. Korištene tehnologije

Aplikacija je izrađena upotrebom MATLAB programskog jezika i razvojnog okruženja verzije R2014a uz dodatke MATLAB Statistics Toolbox, MATLAB Econometrics Toolbox, MATLAB Financial Toolbox, MATLAB Database Toolbox.

Iako je MATLAB komercijalni proizvod za čije je korištenje potrebno posjedovanje licence, odabran je iz razloga što omogućava brzo prototipiranje te veću posvećenost rješavanju problema, a manje brizi oko same implementacije. S obzirom da je u ovome radu naglasak stavljen na eksperiment, a manje na sam razvoj, to svakako predstavlja cijenjenu karakteristiku. Dodatna prednost proizlazi iz toga što je postignuta ujednačenost među dodatcima koji slijede sličnu logiku i sintaksu, čime se dodatno ubrzava rad i smanjuje potreba za privikavanjem. S druge strane, izrada sučelja ponešto je otežana i usporena, premda i dalje postoji dovoljno slobode u njegovoj realizaciji, uz poneke zamjerke s dizajnerskog aspekta. Osim toga, jednom dobiveno zadovoljavajuće rješenje, jednostavnije se kasnije implementira i upotrebom drugih tehnologija.

Zbog potrebe manipulacije velikim količinama podataka, aplikacija koristi ORACLE XE 11g RDBMS. Odabir sustava za upravljanje bazom podataka u ovom slučaju nije presudan, već jednim dijelom ovisi o mogućnosti usklađivanja s MATLAB-om, a mnogo više o osobnim preferencijama.

Implementacija SVM algoritma realizirana je korištenjem LibSVM biblioteke (Chang i Lin, 2001.) verzije 3.18 koja je odabrana zbog dobrih performansi i široke prihvaćenosti. Osim što nudi sučelje za brojne jezike kao što su, osim MATLAB-a, to još i Python, R, Perl, Ruby, PHP i drugi, sastavni je dio raznih okruženja za rudarenje podacima poput RapidMiner-a ili LionSolver-a. Za korištenje u okviru MATLAB-a potrebno ju je najprije kompajlirati, nakon čega se njena upotreba svodi na pozivanje funkcija:

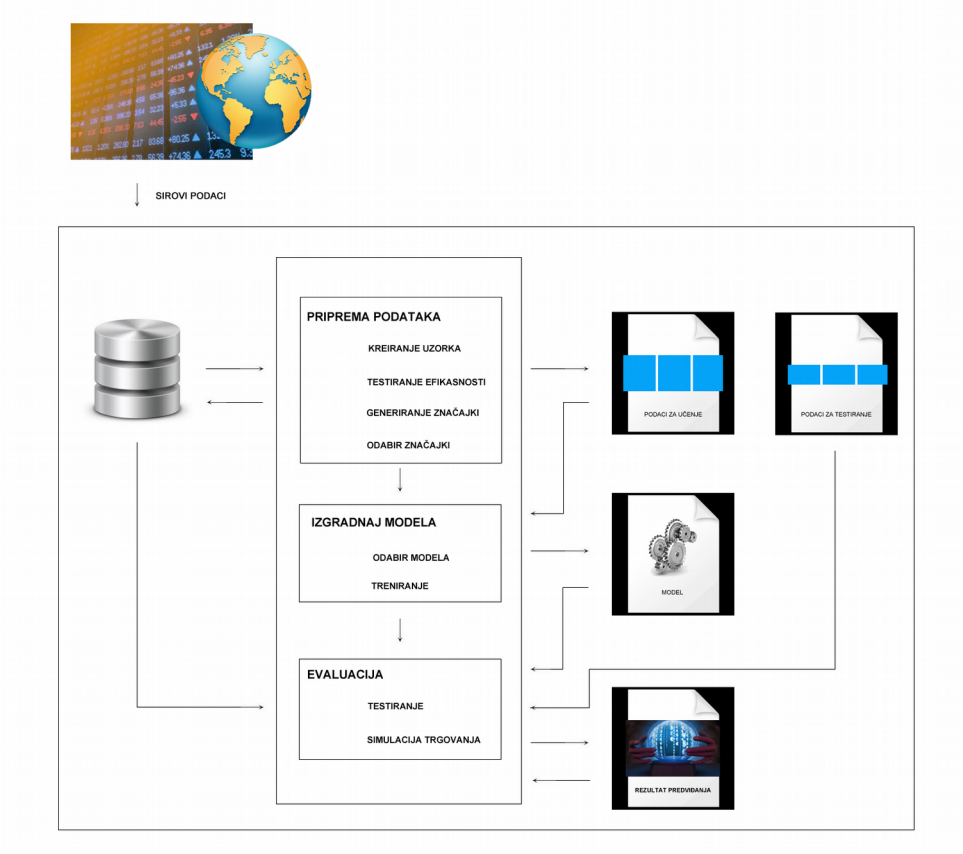
- libsvmwrite - za kreiranje LibSVM formata datoteke,
- libsvmread - za učitavanje datoteke u LibSVM formatu,
- svmtrain - za kreiranje modela,
- svmpredict - za klasifikaciju novih podataka.

S obzirom na otvoreni kod, moguće ju je dodatno prilagoditi vlastitim potrebama.

5.5. Shema aplikacije

Aplikacija je u svojoj osnovi jednostavna budući da glavna zahtjevnost proizlazi iz poteškoća u postizanju dobrih rezultata predviđanja. Stoga se prilikom izrade ograničilo samo

na pružanje nužnih funkcionalnosti tako da aplikaciju treba promatrati prvenstveno u kontekstu analize problema. Također, i sama baza podataka slijedi izrazito jednostavnu strukturu. Iako se radi o velikim količinama podataka, sastoji se od samo nekoliko tablica. Nužnost za takvim pristupom velikim dijelom proizlazi i iz činjenice da se razvoj vršio na *netbook* računalu skromnih mogućnosti, te se time nastojalo smanjiti suvišno opterećenje kako bi se resursi sačuvali za računski zahtjevnije operacije. Shema aplikacije, koja prati osnovne faze procesa, prikazana je na slici 21. dok je opis funkcionalnosti i njihove implementacije dan u nastavku teksta.



Slika 21. *Shematski prikaz aplikacije*

Izvor: izradila autorica

5.6. Sirovi podaci

Sirove podatke predstavljaju dnevni podaci preuzeti s tržišta vrijednosnica koje, u skladu s uobičajenom burzovnom kotacijom (slika 22.), čine (Achelis, 2001.) :

- Prva. To je cijena po kojoj je obavljeno prvo trgovanje određenog dana. Ima poseban

značaj u analizi budući da su investitori imali priliku "prespavati" i odreagirati na događaje koji su se dogodili nakon zatvaranja tržišta prethodnoga dana.

- Najviša. Najviša cijena koju je dionica postigla u toku promatranoga perioda (u ovom slučaju jednog dana). To je cijena pri kojoj je bilo više onih spremnih prodati nego kupiti. Iako uvijek postoje prodavači spremni prodati po višoj cijeni, to je najviša cijena pri kojoj su kupci bili spremni kupiti.
- Najniža. Najniža cijena koju je dionica postigla u toku promatranoga perioda. To je cijena pri kojoj je bilo više kupaca nego prodavača, odnosno najniža cijena koju su prodavači bili spremni prihvatiti.
- Zaključna. Zadnja cijena po kojoj je izvršeno trgovanje u određenom danu. Najčešće se u analizama koristi upravo ta cijena.
- Volumen. Broj dionica kojima se trgovalo tijekom dana.

Sve | Službeno tržište | Redovito tržište

Simbol	Sektor	Zaključna	Zadnja	Promjena %	Prva	Najviša	Najniža	Prosječna	Količina	Promet
ADPL-R-A	CL	84,90	84,10	-1,66 %	85,28	85,97	84,00	84,90	1.023	86.848,89
ADRS-P-A	MA	342,41	344,90	+1,44 %	338,78	344,90	338,50	342,41	127	43.486,56
ADRS-R-A	MA	375,36	384,99	+1,31 %	374,20	384,99	373,00	375,36	239	89.710,34
ATGR-R-A	G	927,72	926,00	0,00 %	929,44	929,44	926,00	927,72	2	1.855,44
ATLN-R-A	L	98,63	102,00	+13,33 %	94,29	102,00	94,27	98,63	3.035	299.333,62
ATPL-R-A	H	315,44	315,00	-0,54 %	316,70	319,00	312,41	315,44	355	111.981,55
BLJE-R-A	A	34,52	34,50	-0,29 %	34,50	34,53	34,50	34,52	977	33.721,45
CKML-R-A	CA	3.801,08	3.801,08	-4,97 %	3.801,08	3.801,08	3.801,08	3.801,08	3	11.403,24
CROS-P-A	K	7.412,25	7.412,25	-0,01 %	7.412,25	7.412,25	7.412,25	7.412,25	1	7.412,25
DDJH-R-A	MA	38,06	37,20	-0,35 %	37,50	39,50	37,20	38,06	12.722	484.233,54
DLKV-R-A	F	16,22	16,29	+2,32 %	16,00	16,29	15,97	16,22	4.789	77.663,36

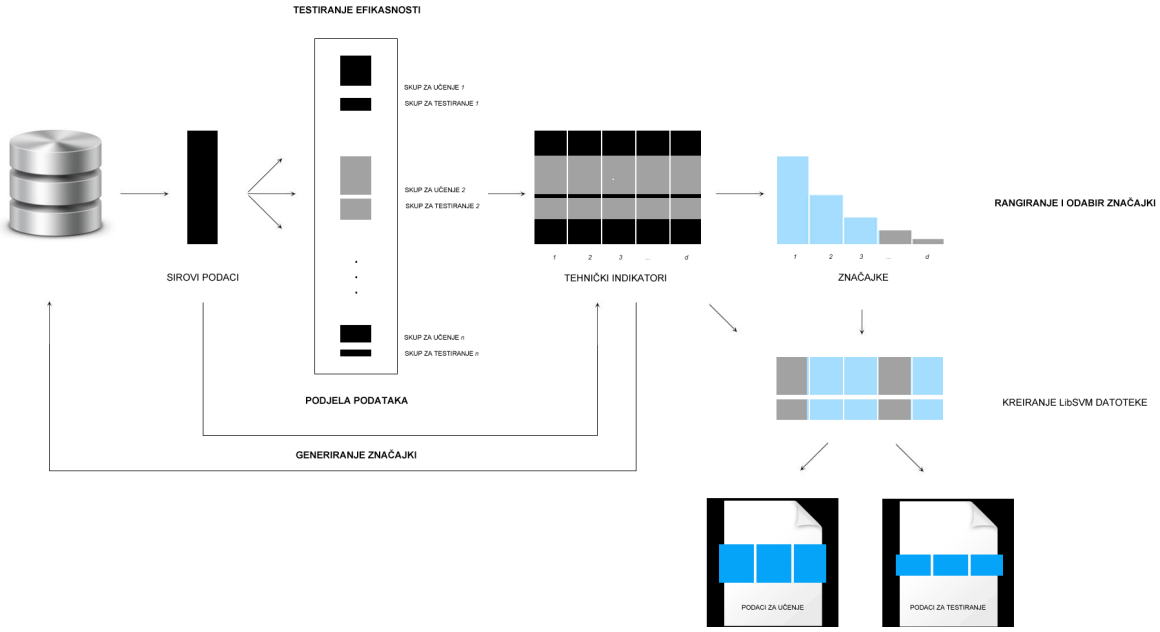
Slika 22. *Kotacija Zagrebačke burze*

Izvor: Zagrebačka burza [Online] Dostupno na: www.zse.hr [Pristupljeno: 21. veljače 2015.]

5.7. Priprema podataka za učenje

Priprema podataka predstavlja obvezni korak koji prethodi procesu učenja. Bez adekvatne pripreme, algoritam ili neće raditi ili će davati netočne rezultate. Npr. neki su algoritmi posebno osjetljivi na šum, dok drugi ne prihvaćaju nominalne atribute. Stoga je važno prethodno neupotrebljive podatke pretvoriti u one prikladne za upotrebu s odabranim algoritmom. Za potrebe SVM algoritma i LibSVM biblioteke priprema podataka za učenje vrši se na sljedeće načine (slika 23.):

- najprije se vrši pretprocesiranje podataka koje obuhvaća čišćenje i tretiranje nedostajućih vrijednosti, nakon čega slijedi
- podjela podataka na skupove za treniranje (koji služi za izgradnju modela) i testiranje (koji služi za evaluaciju modela) pri čemu kao pomoć u odabiru vremenskog intervala mogu poslužiti rezultati testiranja efikasnosti tržišta,
- na temelju sirovih podataka izračunavaju se tehnički indikatori koji predstavljaju ulazne varijable,
- na kraju se vrši odabir značajki primjenom Random forest algoritma, te se kreiraju LibSVM datoteke koje služe kao ulaz u proces učenja i testiranja.



Slika 23. *Shema pripreme podataka za učenje*

Izvor: izradila autorica

5.7.1. Pretprocesiranje

Pretprocesiranjem se nastoji poboljšati kvaliteta podataka kako bi se povećala efikasnost naknadnih faza analize i predviđanja. Kvalitetu podataka čine točnost, potpunost i konzistentnost (Garcia, Luengo, i Herrera, 2015.), dok su stvarni podaci često prljavi, nekompletni i nekonzistentni. Prljavi podaci su takvi podaci među kojima postoje nedostajuće vrijednosti, greške ili nestandardna reprezentacija istih podataka. Na primjer, podaci dobiveni s tržišta obično nisu pogodni za direktnu upotrebu. Zbog praznika ili slabe likvidnosti moguća

je pojava nedostajućih vrijednosti, a uz to i grešaka nastalih tijekom njihovog preuzimanja. Stoga je prije poduzimanja naknadnih koraka nužno izvršiti (Garcia, Luengo, i Herrera, 2015.):

- čišćenje, koje obuhvaća ispravljanje nepreciznosti, filtriranje ili ispravljanje pogrešnih podataka.
- tretiranje nedostajućih vrijednosti koje može obuhvaćati:
  - brisanje, što je prihvatljiv postupak samo u slučaju malog broja takvih primjera, dok inače može dovesti do gubitka važnih informacija,
  - imputaciju, koja obuhvaća postupke zamjene nedostajućih vrijednosti procijenjenima,
  - interpolaciju, koja se odnosi na umetanje novih podataka između postojećih.

Iako nedostajuće vrijednosti otežavaju analizu, njihovo se neodgovarajuće tretiranje također može negativno odraziti na konačan ishod zbog uvođenja pristranosti u podatke, što zatim vodi pogrešnim zaključcima.

Na važnost odabira odgovarajuće tehnike tretiranja nedostajućih vrijednosti u financijskim vremenskim nizovima upozorava i Leonardelli (2012.) koji ukazuje da primjena tehnika interpolacije rezultira manjim procijenjenim vrijednostima mjera rizika, dok s druge strane, primjena tehnika imputacije rezultira većim vrijednostima mjera rizika, što dovodi i do stvaranja financijski konzervativnijih procjena.

U opsežnom pregledu primjene strojnog učenja u predviđanju financijskih vremenskih nizova, koji daju Atsalakis i Valavanis (2009.), uobičajen tretman nedostajućih vrijednosti većine autora njihova je zamjena srednjom vrijednošću ili zadnjom opaženom cijenom, dok dio autora ne koristi nikakvo pretprocesiranje.

Iako ovaj korak u okviru aplikacije nije automatiziran, važan je, uz sve prethodno navedeno, i s obzirom da, zbog zapisa rijetke matrice, LibSVM biblioteka nedostajuće vrijednosti interpretira kao nule, što u ovom slučaju nije prihvatljivo te je potrebno primijeniti jedan od navedenih postupaka.[34]

34 Npr. prinos = 0 znači da nije bilo promjene cijene, ali ako takav podatak nedostaje, nije svejedno hoće li biti zamijenjen nulom ili nekom drugom vrijednošću.

5.7.2. "Business time" pristup

Kako navodi Cont (1999.), većina tehnika analize vremenskih nizova dizajnirane su tako da se bave podacima pravilno razmaknutima u vremenu. Međutim, većina financijskih vremenskih nizova ne slijedi takvu pravilnost ako se promatranje vrši u fizičkom vremenu (npr. na većini se tržišta vikendom i praznicima ne trguje). Najjednostavniji način da se prevaziđe takva neusklađenost jest postupak deformacije vremena odnosno uvođenje nove varijable vremena koju neki autori nazivaju "*business time*" ili "*intrinsic time*". Upravo se taj pristup koristi i u okviru aplikacije, što znači da se kao sukcesivne vrijednosti uzimaju samo dani na koje je bilo trgovanja, a zanemaruju ostali dani. Ovime se gubi mogućnost direktne usporedbe svjetskih tržišta, ali s druge strane, takav pristup oslobađa od problema odabira odgovarajuće metode interpolacije ili potrebe za nekom drugom vrstom složenije transformacije.

5.7.3. Testiranje efikasnosti

U svrhu ispitivanja slabog oblika efikasnosti vrši se testiranje RW3 verzije slučajnog hoda.

Slučajni hod

Slučajni hod (eng. *Random Walk - RW*) daje statistički opis nepredvidljivosti cijena. U verziji nazvanoj RW1 to je slučajni proces koji se može izraziti kao (Lo i Mackinlay, 1989.) :

$$P_t = c + P_{t-1} + \varepsilon_t, \varepsilon_t \sim IID(0, \sigma^2), \tag{5.1}$$

gdje je P_t cijena u vremenu t, c očekivana promjena cijene a $IID(0, \sigma^2)$ označava da su slučajne pogreške ε_t nezavisno i jednako distribuirane s očekivanom vrijednošću 0 i varijancom σ^2. Ova verzija slučajnog hoda podrazumijeva ne samo nezavisnost slučajnih pogrešaka, nego i to da su njihove nelinearne funkcije nekorelirane.

S obzirom da pretpostavka o jednako distribuiranim slučajnim pogreškama nije realna za duže vremenske serije, verzija slučajnog hoda RW2 relaksira pretpostavku o IID i uvodi proces s nezavisnim, ali ne i jednako distribuiranim (INID) slučajnim pogreškama ε_t. Za razliku od RW1 procesa, RW2 dozvoljava heteroskedastičnost[35] slučajne pogreške ε_t.

35 Problem heteroskedastičnosti prisutan je kad je narušena pretpostavka o nepromjenjivosti varijance slučajnih

Treća, najopćenitija i ujedno najslabija, verzija slučajnog hoda RW3 zamjenjuje pretpostavku o nezavisnosti s pretpostavkom da su slučajne pogreške ε_t nekorelirane. Primjer procesa RW3 je proces za koji vrijedi $Cov[\varepsilon_t, \varepsilon_{t-k}] = 0$, za svaki $k \neq 0$, ali $Cov[\varepsilon_t^2, \varepsilon_{t-k}^2] \neq 0$.

U okviru aplikacije omogućeno je testiranje slučajnosti vremenskog niza primjenom nekoliko testova odabranih na temelju tablice 1.: uz grafičke prikaze autokorelacijske funkcije i funkcije parcijalne autokorelacije, koristi se i parametarski test omjera varijanci, odabran iz razloga što, kako navode njegovi autori Lo i MacKinlay (1989.), u slučaju heteroskedastičnog RW daje pouzdanije rezultate od inače često korištenoga Dickey-Fullerovog testa.

Tablica 1. *Testiranje varijanti hipoteze slučajnog hoda*

Varijanta RW	Testovi
RW1	• Statistički testovi • Sljedovi, obrati i *runs* testovi
RW2	• Pravila filtra • Tehnička analiza
RW3	• Autokorelacija • Portmanteau statistika • Odnos varijanci

Izvor: Barbić (2010.a)

Autokorelacijska funkcija

Prema hipotezi slučajnog hoda promjene cijena su nekorelirane. Stoga se testiranje vrši ispitujući nultu hipotezu da su autokorelacijski koeficijenti različitih pomaka jednaki nuli. (Barbić, 2010.a).

Koeficijent autokorelacije reda k je koeficijent linearne korelacije između članova stohastičkog procesa razmaknutih za k-vremenskih razdoblja (Bahovec i Erjavec, 2009.) :

$$\rho(k) = \frac{Cov(Y_t, Y_{t+k})}{\sigma_{Y_t}\sigma_{Y_{t+k}}} \tag{5.2}$$

Za $k = 0, \pm1, \pm2, ...$, niz koeficijenata autokorelacije $\rho(0), \rho(1), \rho(2), .,$ kao funkcije vremenskog pomaka k čini autokorelacijsku funkciju stohastičkog procesa (engl. *AutoCorrelation Function – ACF*). Grafički prikaz autokorelacijske funkcije zove se korelogram.

varijabli u linearnom regresijskom modelu.

Parcijalna autokorelacijska funkcija

Parcijalna autokorelacija je korelacija između članova stohastičkog procesa udaljenih međusobno za k razdoblja, Y_t i Y_{t+k} uz uvjet da se ukloni njihova linearna zavisnost o varijablama $Y_{t+}, Y_{t+2}, ..., Y_{t+k-1}$. Koeficijenti parcijalne autokorelacijske funckije definiraju se formulom (Bahovec i Erjavec, 2009.) :

$$\phi_{kk} = Corr(Y_t, Y_{t+k} | Y_{t+1}, Y_{t+2}, ..., Y_{t+k-1}) \tag{5.3}$$

Niz koeficijenata parcijalne autokorelacije procesa $\phi_{kk}, k = 1, 2, ...$ kao funkcija pomaka k, definira parcijalnu autokorelacijku funkciju (eng. *Partial AutoCorrelation Function - PACF*).

Odnos varijanci

Lo i MacKinlay (1989), potaknuti svojstvom slučajnog hoda da je varijanca slučajne pogreške linearna funkcija vremena, što znači da je varijanca dvotjednog prinosa jednaka dvostrukoj varijanci tjednog prinosa, predložili su test za testiranje slučajnog hoda koji se bazira na odnosu varijanci uzastopnih prinosa: uspoređuje se varijanca od $r_t + r_{t-1}$ s dvostrukom varijancom od r_t, gdje je r_t prinos u vremenu t. Njihov odnos treba biti statistički nerazlučiv od 1 (Barbić, 2010.a).

Testiranje se u aplikaciji vrši pomoću funkcija u okviru MATLAB Econometrics Toolbox-a:

- za prikaz korelograma koriste se funkcije **autocorr** i **parcorr** koje kao argument primaju logaritme prinosa ili njihove transformacije (primjer 1.)
- za provođenje testa omjera varijanci koristi se funkcija **vratiotest** koja kao argument prima logaritam cijene (primjer 2.).

Primjer 1. *Prikaz korelograma*

```
% računanje logaritma prinosa
% p je vektor cijena
r = diff(log(p));
% prikaz autokorelacijske funkcije logaritma prinosa za 20 pomaka
f = figure;
```

```
autocorr(r, 20);
```

Primjer 2. *Test omjera varijanci*

```
% testiranje nul hipoteze da cijene slijede slučajni hod
% za h = 0 ne odbacuje se nul hipoteza, za h = 1 odbacuje se nul hipoteza
h = vratiotest(log(p));
```

5.7.4. Generiranje ulaznih i izlaznih varijabli

Za automatsko raspoznavanje uzoraka svi objekti od interesa trebaju biti predstavljeni informacijama koje se odnose na njihove bitne karakteristike. U slučaju kad te informacije nisu kompletne ili su nepouzdane, kad su mnoge od njih irelevantne ili redundantne, otkrivanje uzoraka u podacima bit će otežano, a rezultirajuća prediktivna točnost izgrađenog sustava biti će nezadovoljavajuća, čineći opservacije i zaključke nepouzdanima (Stanczyk i Jain, 2015.). Stoga, kako navode Guyon i Elisseeff, (2003.), umjetnost strojnog učenja započinje odabirom odgovarajuće reprezentacije podataka.

Podaci se za potrebe strojnog učenja reprezentiraju kao n-dimenzionalni vektori konačnog broja značajki $\mathbf{x} = (x_1, ...x_n)$ čije elemente čine vrijednosti tih značajki. Značajka predstavlja distinktivni atribut ili aspekt nečega i koristi se kao sinonim za karakteristiku, kvalitetu ili svojstvo (Stanczyk i Jain, 2015.).

Pronalaženje dobre reprezentacije podataka usko je vezano za domenu problema koji se rješava i ovisno je o konačnoj primjeni. Stoga stvaranje odgovarajuće reprezentacije podataka predstavlja priliku da se inkorporira prethodno znanje o problemu, a upravo to i doprinosi boljim performansama koje se postižu korištenjem značajki izvedenih iz originalnih atributa (Guyon i Elisseeff, 2003.). Međutim, teško je unaprijed predvidjeti koja će transformacija najviše pridonijeti poboljšanju performansi modela.

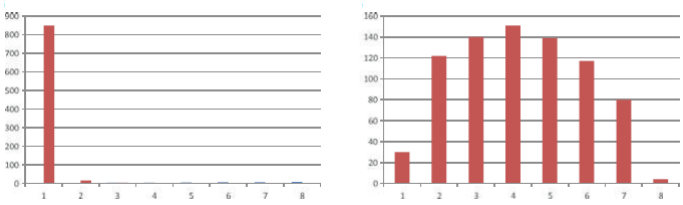
Uobičajeno se transformacija podataka u značajke vrši na način da se kombiniraju originalni sirovi podaci korištenjem raznih matematičkih formula, koje mogu biti utemeljene u poslovnim modelima, kao što su to tehnički indikatori u analizi dionica, ili pak čiste matematičke formule. Na primjer, mnogi statistički postupci pretpostavljaju normalnu distribuciju podataka. Iako stvarni podaci često nisu tako distribuirani, moguće ih je transformirati pomoću Box-Cox transformacije tako da njihova distribucija bude blizu normalne (Garcia, Luengo, i Herrera, 2015.) (slika 24):

$$y = \begin{cases} x^{\lambda-1}/\lambda, & \lambda \neq 0 \\ log(x), & \lambda = 0 \end{cases} \tag{5.4}$$

Sve linearne, kvadratne i slične transformacije posebni su slučajevi Box-Cox transformacija. Logaritmiranje prinosa vrijednosnica česta je transformacija koja se uglavnom koristi u akademskim istraživanjima[36] a računa se prema formuli:

$$r_t = log\left(\frac{p_t}{p_{t-1}}\right) \tag{5.5}$$

gdje je p_t cijena u vremenu t.



Slika 24. *Box-Cox transformacija: a) originalni podaci, b) transformirani podaci i širenje histograma kao posljedica transforamcije*

Izvor: Garcia, Luengo, i Herrera (2015.)

Tehnički indikatori

Tehnički indikatori numerički izražavaju određena svojstva vremenskog niza, a izračunavaju se na temelju podataka o cijenama i volumenu. Unatoč vječnoj debati oko smislenosti njihove direktne primjene u predviđanju tržišnih kretanja (Murphy, 1999.), ipak mogu pružiti zanimljiv pogled na dinamiku vremenskog niza, navodi Torgo (2010.), dok Zemke (2002.) savjetuje njihovu primjenu u automatiziranim sustavima predviđanja jer:

- usrednjavanjem podataka, koje je prisutno u mnogim formulama, doprinose smanjenju šuma,
- ističu karakteristike podataka koje su pogodne za predviđanje.

Tehnički indikatori mogu se podijeliti u dvije glavne grupe: sljedbenike trenda ("*trend-follower*") čija je svrha identifikacija signala početka novog ili završetka starog trenda, te oscilatore, korisne u situacijama kad nema izraženog trenda pa tad većina "*trend-following*"

36 To je tzv. složeni kontinuirani prinos., dok se u praksi češće koristi jednostavni diskretni prinos: $R_t = \frac{p_t - p_{t-1}}{p_{t-1}} \times 100$

indikatora ne radi dobro, a korisni su i kad je trend prisutan te tada mogu poslužiti za upozorenje na kratkoročne ekstreme ili mogu signalizirati da je postojeći trend blizu kraja (Murphy, 1999.).

Primjer sljedbenika trenda, a ujedno u praksi i najčešće korištenih indikatora, jesu pomični prosjeci (slika 25.) koji predstavljaju prosjek cijene određenog razdoblja. Računaju se tako da se podaci pomiču unaprijed sa svakim danom trgovanja. Kad se doda novi dan, posljednji se ukloni te se izračuna novi prosjek. Korisni su zbog efekta zaglađivanja, a najčešće se koriste:

- Jednostavni pomični prosjeci. Predstavljaju aritmetičku sredinu n podataka. Mana im je što daju jednaku važnost svakom podatku te što se izračunavaju samo na temelju određenog broja dana. Neki analitičari smatraju da bi veću važnost trebalo dati novijim podacima.

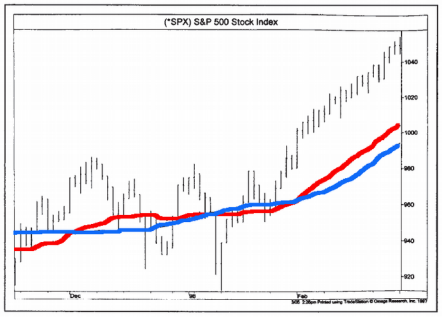
$$SMA_{t,n} = \frac{p_t + p_{t-1} + ... + p_{t-n+1}}{n} \tag{5.6}$$

  gdje je p_t cijena u vremenu t, n broj dana uključenih u izračun.

- Eksponencijalni pomični prosjeci. Rješavaju oba problema: daju veću važnost novijim podacima, a u izračunu sudjeluju svi podaci pri čemu se za veći α brže smanjuje utjecaj starijih podataka.

$$EMA_t = \alpha p_t + (1 - \alpha) EMA_{t-1} \tag{5.7}$$

Primjere oscilatora predstavljaju npr. RSI, %K, %D prikazani u tablici 2.



Slika 25. *Pomični prosjeci: jednostavni (plava linija), eksponencijalni (crvena linija) za 40 dana. Eksponencijalni prosjek pokazuje nešto veću osjetljivost.*

Izvor: Murphy (1999.)

Atsalakis i Valavanis (2009.) daju opsežni pregled korištenja tehničkih indikatora u strojnom učenju: oko 30% autora koristi kao ulazne varijable samo cijene, oko 20% koristi tehničke indikatore (između dva i 25 indikatora) i to često u kombinaciji s dnevnim cijenama prethodnoga dana, dok većina autora u prikazu kombinira tehničke indikatore s fundamentalnima.

Za potrebe aplikacije na temelju sirovih podataka omogućeno je kreiranje tehničkih indikatora. Uz indikatore za koje postoje funkcije u okviru MATLAB Financial Toolboxa-a (prikazanih u tablici 2.), dodatno je omogućeno kreiranje jednostavnih i eksponencijalnih pomičnih prosjeka za različite vremenske intervale (4,5,9,10,18,20 dana). Tome su pridodane relativne promjene cijena (5,10,15,20 dana) (Tay i Cao, 2001.b), transformirana zaključna cijena od koje je oduzet 15-dnevni eksponencijalni pomični prosjek čime se uklanja trend (Tay i Cao, 2001.b) te varijabla "ponedjeljak", koja je, iako ne pripada tehničkim indikatorima, dodana kako bi se ispitali potencijalni kalendarski učinci.

Tablica 2. *Tehnički indikatori korišteni u aplikaciji*

Naziv	Formula	Tumačenje
Acceleration between times	$ACC_t = MOM_t - MOM_{t-n}$ gdje je MOM momentum.	Razlika dvaju momentuma razmaknutih za $n$perioda.
Accumulation / Distribution line	$ADL_t = ADL_{t-1} * MFV_t$ gdje je $MFV_t = MFM_t \times V_t$, $MFM_t = [(C_t - L_t) - (H_t - C_t)] / (H_t - L_t)$	Povezuje promjenu cijene i volumena. Bazira se na premisi da značajnije promjene cijena prati i veći volumen.
Accumulation / Distribution oscillator	$ADOSC_t = \frac{H_t - C_{t-1}}{H_t - L_t}$	Povezuje promjene u cijenama.
Bollinger band	$MB = \frac{\sum_{i=1}^{n} C_i}{n}$ $UB = MB + \left[D * \sqrt{\sum_{i=1}^{n} (C_i - \mathrm{MB})^2 / n} \right]$ $LB = MB - \left[D * \sqrt{\sum_{i=1}^{n} (C_i - \mathrm{MB})^2 / n} \right]$	Mjeri volatilnost cijena. Sastoji se od tri krivulje: gornja (UB) i donja krivulja (LB) nalaze se na udaljenosti od $\pm$ D standardnih devijacija od srednje krivulje (MB) koju čini jednostavni pomični prosjek SMA.
Chaikin oscillator	$CHO_t = EMA_3(ADL_t) - EMA_{10}(ADL_t)$	Indikator volumena. Bazira se na eksponencijalnim pomičnim prosjecima indikatora akumulacije/distribucije (*ADL*).
Chaikin volatility	$CHV_t = \left(\frac{\overline{H-L}_t - \overline{H-L}_{t-n}}{\overline{H-L}_{t-n}} \right) \times 100$ gdje je $\overline{H-L} = EMA(H-L)$	Uspoređuje razmak (*spread*) najviše i najniže cijene.
Highest high	$HH_t = max(H_t, H_{t-1}, ...H_{t-n})$	Makismalna cijena u razdoblju

Naziv	Formula	Tumačenje
		od n dana.
Lowest low	$LL_t = min(L_t, L_{t-1}, ...L_{t-n})$	Minimalna cijena u razdoblju od n dana.
Moving Average Convergence / Divergence (MACD)	$MACD = EMA_{12} - EMA_{26}$	Uspoređuje dva eksponencijalno zaglađena prosjeka.
Median price	$MP_t = \frac{H_t + L_t}{2}$	Srednja vrijednost dnevne cijene.
Momentum between times	$MOM_t = C_t - C_{t-n}$	Mjeri brzinu promjene cijene.
Negative volume index	Ako je današnji volumen manji od jučerašnjeg volumena: $NVI_t = NVI_{t-1} + \left(\frac{C_t - C_{t-1}}{C_{t-1}} \times NVI_{t-1}\right)$ Ako je današnji volumen veći ili jednak jučerašnjem volumenu: $NVI_t = NVI_{t-1}$	Usmjeren na dane u kojima volumen pada. Bazira se na premisi da informirani investitori zauzimaju pozicije kad volumen pada, dok neinformirana masa zauzima pozicije kad volumen raste.
On-Balance Volume (OBV)	Ako je današnja zaključna cijena veća od jučerašnje zaključne cijene: $OBV_t = OBV_{t-1} + V_t$ Ako je današnja zaključna cijena manja od jučerašnje zaključne cijene: $OBV_t = OBV_{t-1} - V_t$ Ako je današnja zaključna cijena jednaka jučerašnjoj zaključnoj cijeni: $OBV_t = OBV_{t-1}$	Povezuje promjenu volumena s promjenom cijene.
Positive volume index	Ako je današnji volumen veći od jučerašnjeg volumena: $PVI_t = PVI_{t-1} + \left(\frac{C_t - C_{t-1}}{C_{t-1}} \times PVI_{t-1}\right)$ Ako je današnji volumen manji ili jednak jučerašnjem volumenu: $PVI_t = PVI_{t-1}$	Slično kao i indeks negativnog volumena (*NVI*), ali usmjeren na dane kada volumen raste.
Price rate of change	$ROC_t = \frac{C_t - C_{t-n}}{C_{t-n}} \times 100$	Stopa promjene cijene.
Price-Volume Trend (PVT)	$PVT_t = PVT_{t-1} + \left(\frac{C_t - C_{t-1}}{C_{t-1}} \times V_t\right)$	Slično kao i indikator ravnotežnog volumena (*OBV*) povezuje promjene volumena i promjene cijene.
Relative Strength Index (RSI)	$RSI_t = 100 - \frac{100}{1 + \left(\sum_{i=0}^{n-1} UP_{t-1}/n\right) / \left(\sum_{i=0}^{n-1} DW_{t-1}/n\right)}$ gdje je UP_t– pozitivna promjena, DW_t je negativna promjena u vremenu t.	Uspoređuje snagu porasta i snagu pada cijene u određenom razdoblju.
Stochastic oscillator	$\%K = \frac{C_t - LL_{t-n}}{HH_{t-n} - LL_{t-n}} \times 100$	*%K* uspoređuje kretanje zadnje cijene u odnosu na raspon

Naziv	Formula	Tumačenje
		cijena u danom periodu.
	$\%D = \frac{\sum_{i=0}^{n-1} \%K_{t-i}}{n}$	**%D** - pomični prosjek od %K.
	$slow\%D = \frac{\sum_{i=0}^{n-1} \%D_{t-i}}{n}$	Slow %D - pomični prosjek od %D.
Typical price	$TP_t = \frac{H_t + L_t + C_t}{3}$	Tipična, prosječna cijena.
Volume rate of change	$VROC_t = \frac{V_t - V_{t-n}}{V_{t-n}} \times 100$	Stopa promjene volumena.
Weighted close	$WC_t = \frac{2C_t + H_t + L_t}{4}$	Srednja vrijednost cijena pri čemu je zaključnoj cijeni dana veća težina.
Williams Accumulation / Distribution line	$WA/D = A/D_t + WA/D_{t-1}$ Ako je današnja zaključna cijena veća od jučerašnje zaključen cijene: $A/D_t = C_t - TRL_t$ Ako je današnja zaključna cijena manja od jučerašnje zaključen cijene: $A/D_t = C_t - TRH_t$ Ako je današnja zaključna cijena jednaka jučerašnjoj zaključnoj cijeni: $A/D_t = 0$ gdje je $TRL_t = min(C_{t-1}, L_t)$ $TRH_t = max(C_{t-1}, H_t)$	Akumulacija opisuje tržište koje kontroliraju kupci, a distribucija tržište koje kontroliraju prodavači. Na distribuciju upućuje situacija u kojoj vrijednosnica postiže najvišu cijenu, što A/D indikator ne prati. Tada treba prodati. Akumulacija je suprotna situacija.
Williams %R	$W\%R_t = \frac{HH_t - C_t}{HH_t - LL_t} \times 100$	Mjeri pretjeranu potražnju ili ponudu.

C_t– zadnja cijena u vremenu t, H_t – najviša u vremenu t, L_t- najniža u vremenu t, V_t– volumen u vremenu t, EMA– eksponencijalni pomični prosjek ,SMA– jednostavni pomični prosjek

Izvor: izradila autorica prema Achelis (2001.)

Cont (2001.) navodi odsustvo autokorelacije cijena na likvidnim tržištima kao široko dokumentiranu i često citiranu empirijsku potporu hipotezi efikasnog tržišta i modelu slučajnog hoda. Međutim, samo odsustvo serijske autokorelacije ne implicira i nezavisnost prinosa. Nezavisnost implicira da će bilo koja nelinearna funkcija prinosa također pokazivati odsustvo autokorelacije. Međutim, Cont (2001.) ukazuje da to ipak ne vrijedi. Jednostavne nelinearne funkcije prinosa, kao što je apsolutni prinos ili kvadratna funkcija prinosa, pokazuju signifikantnu pozitivnu autokorelaciju. Taj je fenomen poznat kao stvaranje klastera volatilnosti. (eng. *volatility clustering*): iza velikih varijacija u cijenama uslijediti će ponovno

velike varijacije u cijenama**,** ne nužno istog predznaka. Tada se ne može reći da cijene slijede slučajni hod. Mjera za *volatility clustering* je (Cont, 2001.)*:*

$$C_2(k) = corr(|r_{t+k}|^2, |r_t|^2). \tag{5.8}$$

gdje je r_t prinos u vremenu t, k je pomak.

Emprijska istraživanja prinosa pokazuju da ta korelacijska funkcija sporo opada i ostaje signifikantno pozitivna nekoliko dana, ponekad i tjedana što ukazuje na određen stupanj predvidljivosti. Osim kvadrata mogu se promatrati i druge transformacije, pri čemu je najveća predvidljivost utvrđena za apsolutne prinose (Cont, 2001.). Stoga su, uz logaritam prinosa, kao ulazne varijable u okviru aplikacije ponuđene i njegove transformacije – apsolutna vrijednost, kvadrat i kub kako bi se ispitalo doprinose li takve transformacije boljim rezultatima predviđanja.

Izlazna varijabla

Izlaznu varijablu predstavlja predznak promjene prinosa na sljedeći dan i to u tri varijante - s pomakom od jednog, dva ili tri dana:

$$y_t = \begin{cases} 1 & \text{ako je } r_{t+k} > 0 \\ 0 & \text{ako je } r_{t+k} \leq 0 \end{cases} \tag{5.9}$$

$$r_{t+k} = log(p_{t+k}) - log(p_{t+k-1}), \;\; k = 1, 2, 3 \tag{5.10}$$

gdje je p_t cijena u vremenu t, r_t prinos u vremenu t.

Podjela podataka

Odabranu podjelu sirovih podataka potrebno je ponoviti i s indikatorima prije njihovog korištenja u procesu odabira značajki. Primjer funkcije koja vrši podjelu podataka i ujedno primjer pristupanja bazi korištenjem funkcija iz MATLAB Database Toolbox-a:

Primjer 3. *Funkcija za podjelu podataka indikatora*

```
function status = podijeliIndikatore(indeksSifra)
      % vrati podatke iz baze u obliku tablice
      pref.DataReturnFormat = 'table';
      setdbprefs(pref);
```

```
% formiranje upita
query = strcat('select min(datum) as min, max(datum) as max, ''train'' as
tip'...
' from stx_idx_vr where indeks_sifra = ''', indeksSifra, ''' and' ...
' skup = ''train''',...
' union select min(datum) as min, max(datum) as max, ''test'' as tip' ...
' from stx_idx_vr where indeks_sifra = ''', indeksSifra, ''' and' ...
' skup = ''test''');

% preuzimanje podataka
conn = dbConnect();
e = exec(conn, query);
e = fetch(e);
datumi = e.Data

% ako podaci cijena još nisu podijeljeni, prekini
if strcmp(datumi.MIN(1), 'null') | strcmp(datumi.MIN(2), 'null') |
strcmp(datumi.MAX(1), 'null') | strcmp(datumi.MAX(2), 'null')
       close(conn);
       status = 0;
       return;
end

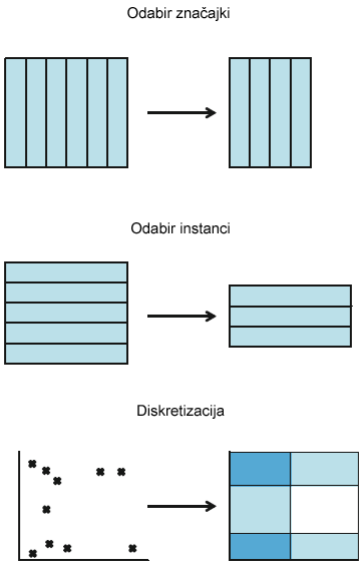
% pripremi where uvjet
nullWhere    = strcat('WHERE INDEKS_SIFRA = ''', indeksSifra, '''');
treningWhere = strcat('WHERE INDEKS_SIFRA = ''', indeksSifra, '''', ...
{' '}, 'AND DATUM BETWEEN ''', datestr(datumi.MIN(1), 'dd.mm.yyyy'), ...
'''', {' '}, ' AND ',{' '}, '''', ...
datestr(datumi.MAX(1),'dd.mm.yyyy'), '''');
testWhere    = strcat('WHERE INDEKS_SIFRA = ''', indeksSifra, '''', ...
{' '}, 'AND DATUM BETWEEN ''', datestr(datumi.MIN(2), 'dd.mm.yyyy'), ...
'''', {' '}, ' AND ',{' '}, '''', ...
datestr(datumi.MAX(2),'dd.mm.yyyy'), '''');

% izvrši podjelu u bazi
data{1} = '';
update(conn, 'stx_ind_vr', {'SKUP'}, data, nullWhere);
data{1} = 'train';
update(conn, 'stx_ind_vr', {'SKUP'}, data, treningWhere);
data{1} = 'test';
update(conn, 'stx_ind_vr', {'SKUP'}, data, testWhere);
close(conn);
status = 1;
```

```
end
```

5.7.5. Odabir značajki

Redukcija podataka obuhvaća skup tehnika kojima se dobiva reducirana reprezentacija originalnog skupa podataka. Za razliku od pripreme podataka, redukcija nije obvezna, ali u slučaju velike dimenzionalnosti podataka vrijeme potrebno za izvršavanje algoritma učenja može biti prohibitivno te tada ona postaje nužan korak, navode Garcia, Luengo, i Herrera (2015.). Općenito se redukcijom podataka smanjuje broj atributa ili instanci pri čemu je nužno zadržati cjelovitost informacije originalnih podataka, odnosno cilj je postići barem jednak rezultat primjenom algoritma na reduciranima kao što bi to bilo primjenom na originalnim podacima.



Slika 26. *Oblici redukcije podataka: odabir značajki (gore), odabir instanci (u sredini), diskretizacija (dole)*

Izvor: Garcia, Luengo, i Herrera (2015.)

Oblike redukcije podataka (slika 26.) čine:

- odabir značajki – koristi se za reduciranje dimenzionalnosti podataka,
- odabir instanci – koristi se za eliminaciju redundantnih i/ili konfliktnih podataka (npr. uzorkovanjem),
- diskretizacija – koristi se za pojednostavljenje domene atributa pri čemu se vrši transformacija numeričkih atributa u diskretne vrijednosti nakon čega se podaci

mogu tretirati kao nominalni.

Odabirom značajki reducira se skup podataka na način da se eliminiraju irelevantne ili redundantne značajke čime se postiže smanjenje dimenzionalnosti. Teoretski, više značajki trebalo bi povećati njihovu diskriminatornu snagu, ali u praksi nije uvijek tako, navode Karagiannopoulos et al. (2007.). Među varijablama može biti mnogo redundantnih, odnosno onih koje su međusobno korelirane pa ih nije potrebno uključiti u modeliranje, ili pak irelevantnih, onih koje nemaju utjecaja na izlazne varijable. Postojanje prevelikog broja značajki može biti problem čak i kad su one relevantne jer to vodi problemu "prokletstva dimenzionalnosti"[37]. Stoga, u sustavima strojnog učenja prednost je imati što manji skup ulaznih varijabli, a proces odabira značajki (eng. *Feature Selection* - *FS*) jedan je od načina da se to postigne.

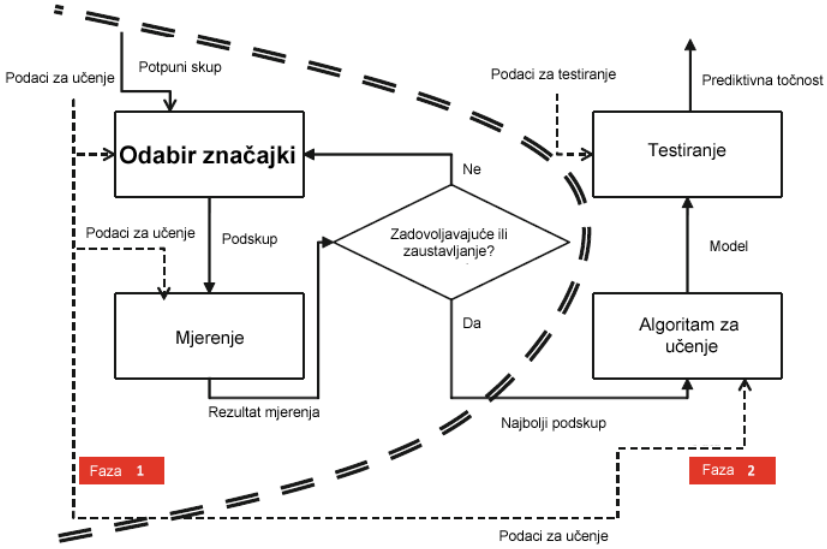
Odabir značajki može se provesti na temelju poznavanja problema kojeg se rješava, ali i primjenom odgovarajućih metoda koje se koriste kada znanje o problemu nije dostupno ili je nedostatno za donošenje informirane odluke, a mogu poslužiti i kao potpora postojećem ekspertnom znanju (Stanczyk i Jain, 2015.).

Metode za odabir značajki

Metode za odabir značajki uobičajeno se dijele na (Garcia, Luengo, i Herrera, 2015.):

1. Metode filtra – djeluju kao vanjski proces koji filtrira nepoželjne varijable prije samog učenja, neovisno od algoritma za učenje (slika 27.). Upravo zbog takve neovisnosti imaju univerzalnu primjenjivost, što s druge strane rezultira i ponešto lošijim rezultatom. Prikladne su u slučaju velike količine podataka i atributa zbog manje računske zahtjevnosti.

37 Problem poznat pod nazivom "prokletstvo dimenzionalnosti" (eng. *curse of dimensionality*) je problem velikih skupova podataka s velikim brojem potencijalnih prediktora što predstavlja poteškoće algoritmima za učenje zbog računske kompleksnosti.

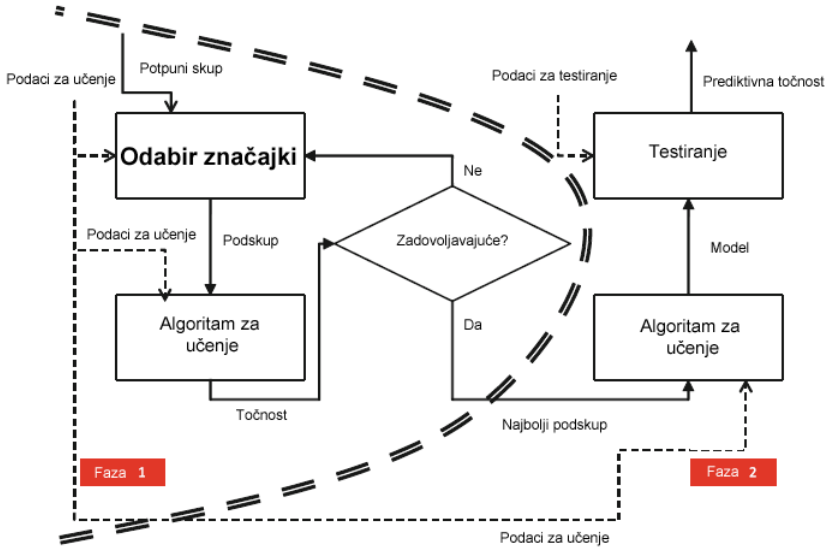


Slika 27. *Shema metode filtra za odabir značajki*

Izvor: Garcia, Luengo, i Herrera (2015.)

Neki autori posebnom vrstom metoda filtra smatraju metode koje rangiraju značajke prema nekom unaprijed odabranom kriteriju. Nakon tako dobivenog rangiranja, o broju korisnih značajki odlučuju ili korisnik ili algoritam za učenje koji se naknadno primijeni.

2. Metode omotača (eng. W*rappers*) - u tom pristupu algoritam za učenje sudjeluje u odabiru podskupa varijabli pri čemu se procijenjena prediktivna točnost smatra najvažnijim indikatorom relevantnosti nekog atributa (slika 28.). Omotači nastoje konstruirati skup atributa koji se prilagođava određenom zadatku i određenom sustavu. Ovakva ovisnost o algoritmu za učenje znači gubitak univerzalnosti i određenu pristranost, ali pridonosi postizanju boljih performansi. Dodatna mana ovog pristupa je računska zahtjevnost.



Slika 28. *Shema metode omotača za odabir značajki*

Izvor: Garcia, Luengo, i Herrera (2015.)

Metode filtera i omotača najviše se razlikuju prema kriteriju evaluacije. Filteri koriste kriterij koji nije uključen u učeći stroj, dok omotači koriste mjere performansi učećeg stroja treniranog korištenjem određenog podskupa značajki.

3. Ugrađene metode (eng. E*mbedded*) – algoritam za odabir značajki dio je sustava za učenje. Neki algoritmi za učenje imaju ugrađeni mehanizam za odabir značajki (npr. stabla odlučivanja).

Random forest

Kako bi se izbjegla potpuna subjektivnost u odabiru ulaznih varijabli, u okviru aplikacije odabir značajki realiziran je kao metoda filtra koja koristi Random forest algoritam (u nastavku RF) (Breiman, 2001.) koji, uz sposobnost rješavanja problema klasifikacije ili regresije, nudi i mogućnost rangiranja varijabli na temelju mjere definirane kao njihov značaj u predviđanju izlazne varijable (*Variable Importance Measure* - VIM). Sličan pristup, kombiniranja RF algoritma za odabir značajki i SVM algoritma za učenje, može se pronaći i kod Chen i Lin (2006.).

RF se bazira na agregaciji velikog broja nekoreliranih stabala odluke[38] kako bi se

38 Općenito se stabla odluke stvaraju rekurzivnom podjelom ulaznih podataka na disjunktne podskupove. Počevši od korijena, pronalazi se podjela koja najviše doprinosi smanjenju odabrane mjere greške, što se nastavlja sve dok mjera greške ne padne ispod unaprijed određene razine ili smanjenje greške koje rezultira

smanjila varijanca pojedinačnih stabala i povećala točnost predikcije. Klasifikacija nove instance vrši se principom glasanja: svako stablo u "šumi" daje svoj glas za određenu klasu, a većinska klasa pobjeđuje. Stabla iz ansambla konstruiraju se na temelju *bootstrap*[39] uzoraka čime se postiže da svako stablo uči na svojim podacima, a od n mogućih varijabli, najprije se slučajnim odabirom odabire njihov podskup od $k \leq n$ iz kojeg se zatim odabire ona koja daje najbolju podjelu. Broj varijabli u svakoj podjeli ima važnu ulogu u određivanju stope greške čitave "šume": manji k smanjuje korelaciju među stablima, ali i snagu[40] svakog stabla.

Primjeri iz originalnog uzorka koji nisu korišteni u izgradnji pojedinog stabla nazivaju se *out-of-bag (OOB)* i na njima se vrši testiranje svakog stabla, a značaj varijable određuje se na temelju povećanja greške klasifikacije do kojega dolazi nakon što se izvrši permutacija njezinih vrijednosti na OOB instancama.

Prednost primjene RF algoritma za odabir značajki predstavlja njegova sposobnost da uspješno identificira jako korelirane varijable uključene u nelinearne interakcije, što je važno u slučaju tehničkih indikatora koji se izračunavaju na temelju istih podataka o cijenama i volumenu. Također, prednost RF algoritma predstavlja i to što se dobro snalazi u slučaju velikog broja irelevantnih varijabli (Boulesteix et al., 2012.)

Međutim, prepuštanje tog zadataka algoritmu ipak nije u potpunosti eliminiralo subjektivnost konačnog odabira, s obzirom da, uz rangiranje, RF ne daje i odgovor koje varijable odabrati. Stoga je konačan odabir prepušten korisniku. Neki od prijedloga o načinu rješavanja tog problema odnose se na eliminaciju varijabli temeljem rezultata unakrsne validacije[41] (Chen i Lin, 2006.) ili na ubacivanje lažne varijable sa slučajno generiranim vrijednostima iz Gaussove distribucije, a zatim odbacivanjem svih onih koje se pokažu manje relevantnima od lažne (Guyon i Elisseeff, 2003.). Potonji se pristup i ovdje koristi.

MATLAB-ovu implementaciju Breimanovog algoritma predstavlja funkcija **TreeBagger.**

nakon sljedećih podjela više ne prelazi određenu unaprijed definiranu vrijednost.

39 Bootstrap je statistička metoda ponovnog uzorkovanja iz istoga skupa za učenje što je praktično kad nije moguće kreirati različite uzorke iz populacije.

40 Stablo s niskom stopom greške je jaki klasifikator.

41 Unakrsna validacija – vidi poglavlje 5.8.2.

Primjer 4. *Rangiranje varijabli po značaju primjenom Random forest algoritma*

```
% izgradnja RF modela
model = TreeBagger( brojStabala, X, Y, 'OOBVarImp', 'On', ...
'MinLeaf', velicinaLista 'NvarToSample', brojVarijabli);

% povećanje greške klasifikacije ako se izvrši permutacija vrijednosti varijable
varerror = model.OOBPermutedVarDeltaError;

% priprema za prikaz varijabli po značaju
% soritranje
[sortirano indeks] = sort(model.OOBPermutedVarDeltaError, 2);

% prikaz varijabli rangiranih po značaju pomoću horizontalnih stupaca
f = figure;
barh(sortirano);
```

Breiman (2001.) preporučuje da se za probleme klasifikacije odabere $k = \sqrt{n}$ gdje je n ukupan broj varijabli, dok Boulesteix et al. (2012.) smatraju da to može biti premalo u slučaju velikog broja neinformativnih prediktora. Minimalni broj opservacija po listu preporučuje se u slučaju klasifikacije postaviti na 1.

Omogućeno je ponavljanje postupka s istim varijablama kroz više iteracija pri čemu se iz svakog sljedećeg kruga izbacuju one s negativnim značajem i one lošije rangirane od lažne, te se postupak izgradnje šume i rangiranja ponavlja s njihovim smanjenim brojem. Također je omogućen i proizvoljan odabir varijabli i to redom najboljih ili najgore rangiranih.

Važno je pri tome napomenuti da se odabir varijabli provodi isključivo na temelju podataka za treniranje.

5.7.6. Kreiranje LibSVM datoteka

Nakon provedenoga odabira kreiraju se LibSVM datoteke s podacima za učenje i testiranje. Prethodno se podaci normaliziraju.

Svrha normalizacije je reduciranje razlika u rasponu varijabli kako atributi s većim rasponom ne bi dominirali nad onima s manjim, odnosno kako bi svi atributi imali jednaku težinu. Time se ujedno i ubrzava rad algoritma za učenje.

Najčešće se koristi *min-max* normalizacija kod koje se vrši skaliranje vrijednosti v

numeričkog atributa A na novi raspon $[new_{minA}, new_{maxA}]$ čime se dobiva nova vrijednost v' (Garcia, Luengo, i Herrera, 2015.):

$$v' = \frac{v - min_A}{max_A - min_A}(new_{maxA} - new_{minA}) + new_{minA} \tag{5.11}$$

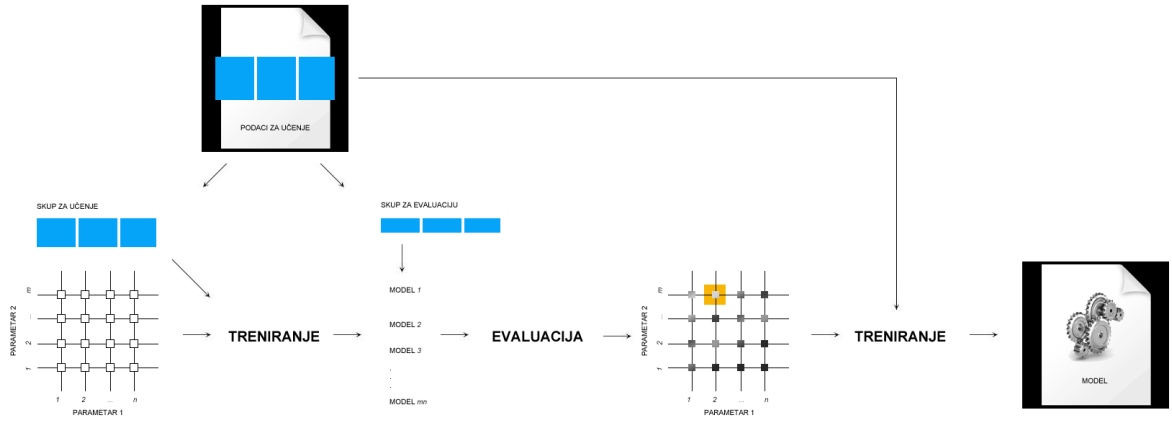
gdje su max_A i min_A originalna maksimalna i minimalna vrijednost atributa. Uobičajeno je svođenje vrijednosti na raspon $[-1, 1]$ ili $[0, 1]$, što se i ovdje koristi.

Za potrebe LibSVM-a najprije se na raspon [0,1] skaliraju podaci za treniranje, a zatim se korištenjem istog faktora skaliraju i oni za testiranje. Ovakvo razdvajanje nužno je provesti kako u podacima za treniranje ne bi bile sadržane informacije o kojima se u trenutku učenja ne zna ništa. Podaci za testiranje predstavljaju budućnost koju se pokušava predvidjeti, stoga je izrazito važno da podaci za testiranje ne sudjeluju ni na kakav način ni u kojem koraku prije samoga testiranja jer bilo kakva kontaminacija podataka za treniranje poništava učinke testiranja. Ono na što dodatno treba obratiti pozornost kod predviđanja unaprijed jest i pomak u podacima koji ne smije dovesti do preklapanja s testnim skupom.

Prije kreiranja datoteka izbacuje se lažna varijabla.

5.8. Odabir modela i učenje

Da bi se dobio prikladan model za rješavanje određenog problema, potrebno je eksperimentalno odrediti vrijednosti parametara koji će osigurati da se izgrađeni model ponaša na željeni način. U ovom slučaju to znači da je najprije potrebno odabrati kernel funkciju i njezine parametre iza čega slijedi proces učenja odnosno izgradnje binarnog kalsifikatora (slika 29.). Omogućen je odabir linearnog, polinomijalnog, RBF i sigmoidalnog kernela. Neovisno o odabranom kernelu postavlja se vrijednost konstante C, dok se od parametara kernel funkcija postavljaju γ za RBF i sigmoidalni kernel te d za polinomijalni. Ostali parametri ostavljeni su, prema preporukama, na pretpostavljenim vrijednostima.



Slika 29. *Shema odabira modela i učenja*

Izvor: izradila autorica

Odabir modela, odnosno najboljih parametara realizirano je tako da se može izvršiti na dva načina:

1. direktnim unosom

2. *grid-serach* pretraživanjem.

5.8.1. Grid-search pretraživanje

Grid-serach pretraživanje provodi se na način da se stvara mreža eksponencijalno rastućih kombinacija vrijednosti parametara pri čemu se za svaku kombinaciju kreira model, vrši njegova evaluacija primjenom jedne od ponuđenih metoda, izračunava se vrijednost odabrane mjere te se kao optimalna kombinacija parametara odabire ona za koju odabrana mjera evaluacije daje najbolju vrijednost. Ponuđena je mogućnost određivanja raspona vrijednosti parametara i veličine koraka čime se određuje gustoća mreže, a kreće se kao potencija broja 2. Npr. odabir raspona od 0 do 4 za C i od -2 do 2 za γ uz korak 2 znači da se stvara i ispituje devet modela sa sljedećim kombinacijama parametara (tablica 6.):

Tablica 3. *Primjer kombinacija parametara za grid search pretraživanje*

$\gamma = 2^{-2} = 0.25,\ C = 2^0 = 1$	$\gamma = 2^0 = 1,\ C = 2^0 = 1$	$\gamma = 2^2 = 4,\ C = 2^0 = 1$
$\gamma = 2^{-2} = 0.25,\ C = 2^2 = 4$	$\gamma = 2^0 = 1,\ C = 2^2 = 4$	$\gamma = 2^2 = 4,\ C = 2^2 = 4$

$\gamma = 2^{-2} = 0.25,\ C = 2^4 = 16$	$\gamma = 2^0 = 1,\ C = 2^4 = 16$	$\gamma = 2^2 = 4,\ C = 2^4 = 16$

Izvor: izradila autorica

Pretraživanje je moguće provoditi kroz više iteracija pri čemu se u svakoj sljedećoj ono vrši u okolini do tada pronađenih najboljih vrijednosti, dok se korak povećanja smanjuje na polovicu prethodne veličine.

5.8.2. Evaluacija klasifikatora tijekom odabira parametara

Tijekom pretraživanja najboljih parametara vrši se evaluacija modela koja vodi njihov odabir pri čemu se podaci za učenje dodatno dijele na skup za izgradnju modela i skup za njegovu evaluaciju. U kojem će to omjeru biti, ovisi o odabranoj metodi.

Unakrsna validacija

U okviru LibSVM biblioteke kao metoda evaluacije implementirana je *k*-struka unakrsna validacija (eng. *k-fold Cross Validation – CV*) koja kao evaluacijsku mjeru koristi točnost (eng. a*ccuracy*)[42]. CV je najčešće korištena metoda za procjenu greške u strojnom učenju. Provodi se tako da se skup podataka S od m primjera podijeli na k podskupova jednake veličine. Algoritam za učenje trenira se na $k-1$ podskupova, a testira na jednom preostalom. To se ponavlja k puta, svaki put testirajući na drugom podskupu. Na taj se način dobije k pojedinačnih procjena, a njihov prosjek predstavlja konačnu procjenu greške klasifikatora. Primjer 5. prikazuje pseudokod unakrsne validacije (Japkowitz i Shah, 2011.).

Primjer 5. *Pseudokod unakrsne validacije*

```
Podijeli raspoloživi skup podataka S veličine m na k podskupova S_i, i = 1,2,...,k
veličine otprilike m/k i bez preklapanja;
Inicijaliziraj i = 1;
Ponavljaj dok je i <= k;
      Označi i-ti podskup S_i kao testni skup;
      Za testni skup S_i generiraj komplement trening skupa S_î koji sadrži sve
      primjere iz S osim onih u S_i;
      Treniraj algoritam učenja na S_î i testiraj na S_i;
      Izračunaj empririjski rizik R_emp(f_i) klasifikaotra f_i;
      Povećaj za 1;
```

42 O evaluacijskim mjerama više u poglavlju 5.9.1

```
Izračunaj prosječni R_emp(f_i) za sve i kako bi se dobio R_emp(f), prosječni
epmirijski rizik k-struke CV;
Vrati R_emp(f);
```

Najčešće korištene vrijednosti za k su 5 ili10, a za $k = m$ predstavlja poseban slučaj nazvan *leave-one-out* (*LOO*), pogodan samo za male skupove podataka.

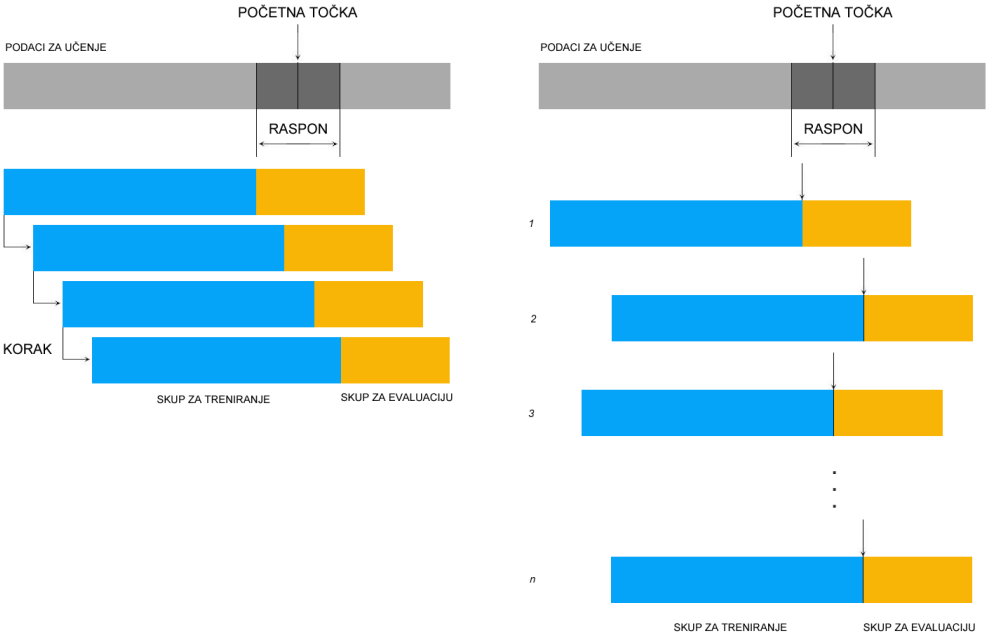
Metoda pomičnog prozora

Međutim, unakrsna validacija nije optimalna metoda kad se radi s vremenskim nizovima jer remeti njihov poredak te, kako navodi Zemke (2002.), daje previše optimističnu procjenu greške. Ako se izvrši podjela na k dijelova, u barem $k - 2$ slučaja podaci za učenje istovremeno će prethoditi i slijediti podatke za evaluaciju, što je u suprotnosti s predviđanjem budućnosti kod kojega je poznata samo prošlost. Stoga je ponuđena i dodatna metoda pomičnog prozora (eng. *window approach*, *windowing*, *rolling forecasting origin*) kod koje podaci za evaluaciju uvijek slijede podatke za treniranje, a provodi se tako da se prozor određene veličine pomiče prema naprijed te se tijekom tog pomicanja gradi model i vrši njegova evaluacija nad obuhvaćenim podacima.

Što se tiče određivanja optimalnog omjera podataka za treniranje i evaluaciju, Zemke (2002.) savjetuje da, zbog nestacionarnosti financijskih vremenskih nizova, udio primjera za evaluaciju ne iznosi više od 20% ukupnog broja primjera unutar prozora.

Ponuđene su dvije varijante pomicanja prozora koje predstavljaju određene modifikacije prema Hyndman i Athanasopoulos (2013.) i Torgo (2010.) prikazane na slici 30.

Kod prvog se načina (slika 30. lijevo) pomicanje vrši slijedno. Odabirom početne točke i raspona određuje se veličina prozora i udio primjera za učenje i testiranje, a zatim se prozor pomiče unaprijed za odabranu veličinu koraka. Kretanje započinje od samog početka raspoloživih podataka pa sve do njihovog kraja. Kod drugog se načina, za odabranu početnu točku i unutar odabranog raspona, slučajnim odabirom određuje točka iza koje se u prethodno odabranom omjeru uzimaju primjeri za trening, a ispred te točke primjeri za evaluaciju. Konačna vrijednost odabrane mjere računa se kao prosjek dobivenih pojedinačnih vrijednosti.



Slika 30. *Evaluacija metodom pomičnog prozora: slijedno pomicanje (lijevo), slučajni odabir početne točke (desno)*

Izvor: izradila autorica

Određena mana ovakvog pristupa je ta što će primjeri s početka i kraja niza biti rjeđe obuhvaćeni, dok će oni u sredini niza uvijek biti uključen u barem jedan od dva skupa (za treniranje ili evaluaciju).

Primjer 6. *Evaluacija pomoću pomičnog prozora sa slijednim pomicanjem*

```
function accTotal = windowingSeq(oznakeKlasa, podaciTrening, cmd, opcije)
       podjela         = opcije{1,1}/100;
       raspon          = opcije{2,1}/100;
       korak           = opcije{3,1};
       brojPrimjera    = size(podaciTrening, 1);
       granica         = brojPrimjera * (1 - raspon);
       min             = round(granica*podjela);
       velicinaTrening = min;
       velicinaEval    = round(granica*(1-podjela));
       max             = brojPrimjera - velicinaEval;
       maxRaspon       = round(brojPrimjera*raspon);

       brojPonavljanja = floor(maxRaspon/korak);
       accTotal = zeros(brojPonavljanja,1);
       for i = 1:brojPonavljanja
```

```
            % odaberi tocke iz raspona
            prviTrening    = 1+(i-1)*korak;
            zadnjiTrening  = velicinaTrening + (i-1)*korak - 1;
            prviEval       = zadnjiTrening +1;
            zadnjiEval     = prviEval + velicinaEval;

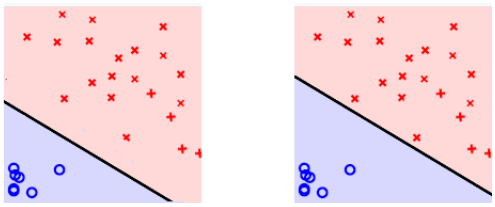
            %podjela podataka
            treningX = podaciTrening(prviTrening:zadnjiTrening);
            treningY = oznakeKlasa(prviTrening:zadnjiTrening);
            evalX    = podaciTrening(prviEval:zadnjiEval);
            evalY    = oznakeKlasa(prviEval:zadnjiEval);
            treningX = reshape(treningX,size(treningX,2), size(treningX,1));
            evalX    = reshape(evalX,size(evalX,2), size(evalX,1));

            %napravi model
            model = svmtrain(treningY, treningX, cmd);
            %testiraj
            [predictLbl, acc, score] = svmpredict(evalY, evalX, model);
            accTotal(i,1) = acc(1)
        end
        accTotal = sum(accTotal) / brojPonavljanja
end
```

5.8.3. Kompenzacija neravnoteže u podacima

Cont (2001.) ukazuje na uočenu asimetriju dobitaka i gubitaka u financijskim vremenskim nizovima stoga se među prikupljenim podacima može očekivati određena razina neravnoteže u odnosu pojedinih klasa. Akbani, Kwek i Japkowicz (2004.) upozoravaju da se situacija znatne neravnoteže u podacima može izrazito negativno odraziti na performanse SVM algoritma. Do smanjenja performansi općenito kod algoritama strojnog učenja dolazi zbog induktivne pristranosti ugrađene u same algoritme, odnosno iz razloga što su dizajnirani tako da među kandidatima hipoteza odabiru onu najjednostavniju koja se dobro prilagođava podacima. U slučaju velike neravnoteže to će biti ona koja klasificira sve instance u brojniju klasu. Konkretno se to kod SVM algoritma manifestira na način da će, s obzirom da SVM stvara stvara granicu odluke koja je na polovici udaljenosti između dviju klasa, naučena hiperravnina biti pomaknuta bliže primjerima malobrojnije klase (slika 31. lijevo), iako neravnoteža u podacima može ukazivati da primjeri malobrojnije klase zapravo leže na većoj udaljenosti od idealne granice (slika 31. desno), što će u konačnici prouzročiti smanjenja

performansi prilikom testiranja na novim instancama.



Slika 31. *Neravnoteža u podacima i optimalna hiperravnina: naučena(lijevo), idealna (desno)*

Izvor: izradila autorica

Osim toga, ako konstanta C nije velika, čime bi se forsiralo ispravno klasificiranje svih primjera, SVM jednostavno nauči klasificirati sve primjere u mnogobrojniju klasu jer se time stvara najveća margina, s nula greške na brojnijim primjerima i samo malom greškom na ostalim primjerima.

Međutim, kako Akbani, Kwek i Japkowicz (2004.) dalje navode, SVM ne pokazuje loše performanse na umjereno neuravnoteženim podacima. S povećanjem neravnoteže među podacima, povećava se i neravnoteža u omjeru potpornih vektora jedne i druge klase. Moglo bi se očekivati da će testne instance biti klasificirane u mnogobrojniju klasu iz razloga što će tada okolinom instanci, koje se nalaze blizu granice odluke, dominirati potporni vektori mnogobrojnije klase. Ali zbog KKT uvjeta:

$$0 \leq \alpha_i \leq C, \quad i = 1, ..., m \tag{5.12}$$

$$\sum_{i=1}^{m} \alpha_i y_i = 0 \tag{5.13}$$

zbroj α_i potpornih vektora jedne klase mora biti jednak zbroju α_i povezanih s potpornim vektorima druge klase. Ako postoji manje primjera jedne klase, to će biti i manje pripadajućih α_i, što znači da u prosjeku moraju biti veći od α_i mnogobrojnije klase. Kako α_i predstavljaju težine u finalnoj decizijskoj funkciji

$$f(x) = sgn\left(\sum_{i=1}^{m} y_i \alpha_i k(x, x_i) + b\right) \tag{5.14}$$

kao rezultat toga, zbog većeg α_i, malobrojniji primjeri dobivaju veću težinu nego mnogobrojniji, čime se donekle kompenzira učinak neravnoteže.

Jedan od načina da se izbjegne pogoršanje performansi u slučaju neravnoteže u podacima jest korištenje različitih vrijednosti konstante C za svaku od klasa. Na taj se način više kažnjava pogrešno klasificiranje primjera malobrojnije klase te se granica odluke pomiče dalje od njih. Tada Lagrangeov primal postaje (Akbani, Kwek i Japkowicz, 2004.):

$$\min_{\mathbf{w},\boldsymbol{\xi}} \tau(\mathbf{w},\boldsymbol{\xi}) = \frac{1}{2}\|\mathbf{w}\|^2 + C^+ \sum_{i|y_i=+1}^{m} \xi_i + C^- \sum_{j|y_j=-1}^{m} \xi_j - \sum_{i=1}^{m} \alpha_i [y_i \left(\langle \mathbf{x_i}, \mathbf{w} \rangle + b\right) - 1 + \xi_i] \tag{5.15}$$

a ograničenja postaju:

$$0 \leq \alpha_i \leq C^+ \text{ ako je } y_i = +1 \text{ i } 0 \leq \alpha_i \leq C^- \text{ ako je } y_i = -1.$$

LibSVM biblioteka omogućava postavljanje različitih vrijednosti konstante C za svaku od klasa pomoću opcije *-wi* prilikom poziva funkcije za treniranje **svmtrain.** Time se postavlja vrijednost parametra C za klasu *i* na vrijednost $w \times C$. U primjeru 7. to znači da će više biti kažnjena pogrešna klasifikacija primjera klase 0 $(C \times w0 = 10 \times 5 = 50)$, od onih klase 1 $(C \times w1 = 10 \times 1 = 10)$.

Primjer 7. *Postavljanje različitih vrijednosti konstante C za svaku od klasa*

```
svmtrain( Y, X, -c 10 -w1 1 -w0 5)
```

Optimalne vrijednosti parametra za uravnoteženje također se mogu odrediti pomoću unakrsne validacije ili se mogu koristiti u omjeru obrnuto proporcionalnom omjeru klasa, što je preporuka koja uglavnom daje dobre rezultate te se i ovdje prihvaća:

$$m_0 \times w_0 = m_1 \times w_1 \tag{5.16}$$

gdje je m_x broj primjera klase x.

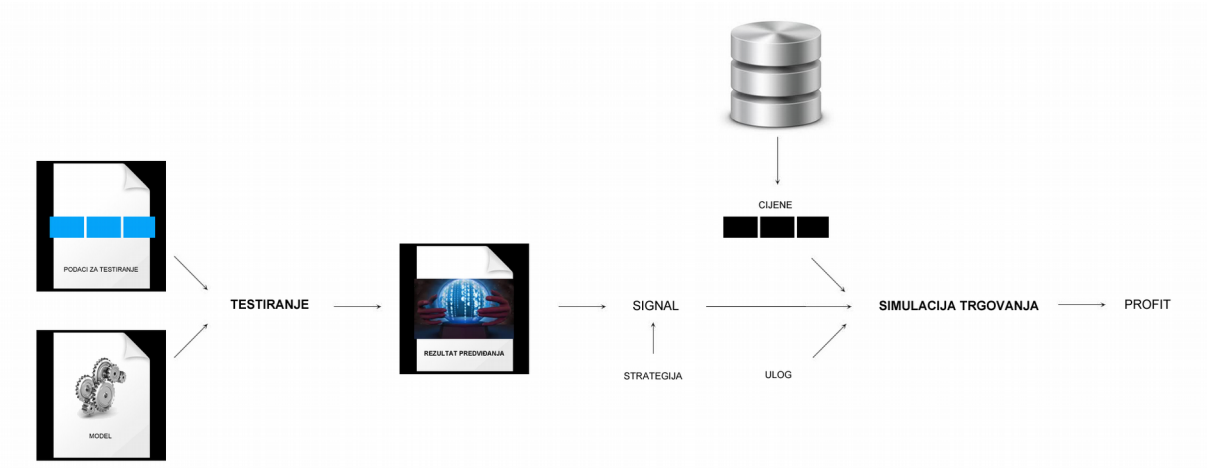
Nakon odabira odgovarajućih parametara trenira se model korištenjem čitavog skupa podataka za učenje.

5.9. Evalucija konačnog modela

Važan korak u strojnom učenju predstavlja ocjena izgrađenog modela prilikom čega je potrebno:

- odabrati odgovarajuću mjeru za ocjenu performansi,
- odabrati metodu za dobivanje nepristrane procjene odabrane evaluacijske mjere.

Evaluacija konačnog modela u okviru aplikacije realizirana je na način da se za ocjenu performansi prema mjerama uobičajenima u evaluaciji klasifikatora koristi testiranje na izdvojenom skupu podataka, dok se za dobivanje ocjene klasifikatora temeljem financijske mjere, u vidu ostvarenog prinosa, rezultati predviđanja na skupu za testiranje koriste dalje u simulatoru trgovanja (slika 32.).

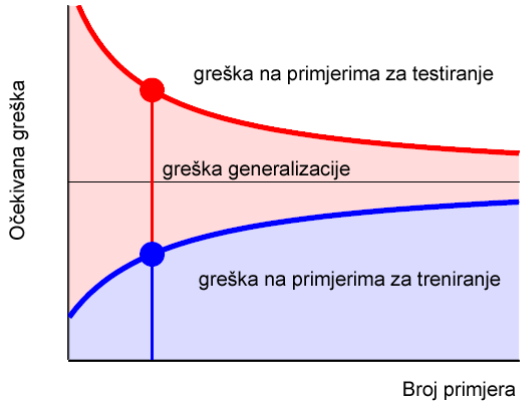


Slika 32. *Shema evaluacije konačnog modela*

Izvor: izradila autorica

5.9.1. Testiranje

Idealno bi bilo raspolagati čitavom populacijom, ali u stvarnosti se najčešće raspolaže samo s ograničenom količinom primjera za učenje. Kad bi ta količina bila dovoljno velika, za procjenu greške bilo bi dovoljno testirati klasifikator na istim podacima na kojima je on i treniran s obzirom da se s povećanjem količine primjera za učenje empirijski rizik približava stvarnom riziku, kao što to prikazuje krivulja učenja na slici 33.



Slika 33. *Krivulja učenja: s povećanjem količine primjera za učenje empirijski se rizik približava stvarnom riziku*

Izvor: Abu-Mostafa, Magdon-Ismail, i Lin (2012.)

Međutim, najčešće ta količina nije dovoljno velika, tako da empirijski rizik nije dobar procjenitelj stvarnog rizika. Stoga je potrebno primijeniti jedan od sljedeća dva pristupa (Japkowitz i Shah, 2011.):

- prvi se temelji na tehnikama probira (eng. *resampling*), odnosno ponovnoj upotrebi iste ograničene količine podataka u stvaranju procjene greške,
- drugi se temelji na testiranju na skupu ranije neviđenih podataka, što se naziva *holdout* ili *train & test* metoda.

Primjer prvog pristupa je *k*-struka unakrsna validacija ili njena verzija LOO (*leave-one-out*) opisana u poglavlju 5.8.2. u kontekstu odabira modela. Kod drugog se pristupa dio podataka namijenjenog procjeni greške odvaja prije procesa učenja te se iz njega u potpunosti izuzima. Tada performanse klasifikatora na skupu za testiranje obično predstavljaju dobar indikator sposobnosti generalizacije (Japkowitz i Shah, 2011.).

Najčešće korištena mjera za ocjenu klasifikatora je točnost (eng. *accuracy*) koja daje ocjenu cjelokupnih performansi klasifikacije uzimajući u obzir sve klase. Točnost se računa kao broj pogrešno klasificiranih primjera podijeljen s ukupnim brojem primjera:

$$tocnost = \frac{broj\ \ ispravno\ \ klasificiranih\ \ primjera}{ukupan\ \ broj\ \ primjera} \tag{5.17}$$

Međutim, ona sama često nije dostatna za objektivnu ocjenu performansi. Dva klasifikatora mogu imati jednaku točnost klasifikacije, ali jedan može davati bolje rezultate u

klasificiranju pozitivnih primjera, dok drugi može pokazivati bolje rezultate na negativnim primjerima. Koji je od njih bolji moći će se ocijeniti tek u kontekstu samog problema, odnosno uzimanjem u obzir mogućeg različitog troška pogrešne klasifikacije.

Uzmimo za primjer ulaganje u dionice. Ako netko pogrešno predvidi pad cijena te se zbog toga suzdrži od ulaganja, zapravo samo propušta priliku za zaradu. Međutim, ako pogrešno predvidi porast cijena te kupi, a cijena nakon toga padne, tada doista ostvaruje gubitak.

Drugi problem koji se može pojaviti uzrokovan je neravnotežom u zastupljenosti pojedinih klasa. Npr. ako je 90% primjera pozitivno, a 10% negativno već i trivijalni klasifikator koji bi samo predviđao dominantnu klasu imao bi točnost od 90%, ali u klasificiranju novih primjera bio bi beskoristan. Stoga točnost nije adekvatna mjera niti u slučaju neravnoteže u podacima.

Postoje različite mjere za ocjenu performansi klasifikatora, a njihov odabir može znatno utjecati na donošenje konačne odluke o njihovoj upotrebi. S obzirom da svaka mjera stavlja naglasak samo na određeni aspekt performansi, a niti jedna ne obuhvaća njih sve, bitno je odrediti one aspekte koji su od najvećeg interesa, a odabrane mjere trebaju biti one koje se fokusiraju na te aspekte (Japkowitz i Shah, 2011.).

Konfuzijska matrica daje bolji pregled vrste grešaka koje klasifikator može učiniti. Konfuzijska matrica $\mathbf{C} = \{c_{ij}\}, i, j, \in \{1, 2, ..., l\}$, je kvadratna lxl matrica gdje je i indeks retka, j indeks stupca, l broj klasa. Svaki element $c_{ij}(f)$ konfuzijske matrice označava broj primjera klase i koje klasifikator f dodjeljuje klasi j. Za određeni skup primjera za testiranje T i klasifikator f, konfuzijska matrica $\mathbf{C}(f)$ može se definirati (Japkowitz i Shah, 2011.):

$$\mathbf{C}(f) = \{c_{ij}(f) = \sum_{\mathbf{x} \in T} [(y = i) \wedge (f(\mathbf{x}) = j)]\} \tag{5.18}$$

gdje je $\mathbf{x}$ je primjer za testiranje, a y je pripadajuća oznaka klase tako da je $y \in \{1, 2, ..., l\}$. Sve vrijednosti na glavnoj dijagonali c_{ii} označavaju ispravno klasificirane primjere klase:

$\sum_{i=1}^{l} c_{ii}(f)$ je ukupan broj primjera koje je klasifikator ispravno klasificirao.

Svi elementi koji nisu na glavnoj dijagonali označavaju pogrešnu klasifikaciju:

$\sum_{i,j;i\neq j} c_{ij}(f)$ označava ukupan broj primjera koje je klasifikator f dodijelio pogrešnoj klasi.

U slučaju binarne klasifikacije konfuzijska je matrica veličine 2x2 i sadrži četiri karakteristične vrijednosti prikazane u tablici 1.:

Table 1. *Konfuzijska matrica*

		Predviđeno klasifikatorom		
		Negativna klasa	Pozitivna klasa	
Stvarna klasa	Negativna klasa	Stvarno negativni (eng. t*rue negative - TN*)	Lažno pozitivni (eng. *false positive - FP*)	N = TN + FP
	Pozitivna klasa	Lažno negativni (eng. f*alse negative - FN*)	Stavrno pozitivni (eng. *true positive - TP*)	P = FN + TP

Izvor: prilagođeno prema Japkowitz i Shah (2011.)

Prema konfuzijskoj matrici točnost se računa:

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{5.19}$$

Dodatne mjere temeljene na konfuzijskoj matrici i samo jednoj klasi jesu (Japkowitz i Shah, 2011.):

Stopa stvarno pozitivnih (eng. *True positive rate – TPR)*, nazvana još i odziv *(eng. Recall),* osjetljivost[43] *(eng. Sensitivity)*, stopa pogodaka (eng. *hit rate),* mjeri udio primjera klase i koje je klasifikator doista dodijelio klasi i. Kod binarne klasifikacije to je:

$$TPR(f) = \frac{c_{22}(f)}{c_{22}(f) + c_{21}(f)} = \frac{TP}{TP + FN}. \tag{5.20}$$

Stopa lažno pozitivnih (eng. *False positive rate – FPR)*, nazvana još i *fallout* ili stopa lažnog alarma, mjeri udio primjera koji su pogrešno dodijeljeni klasi i. Kod binarne klasifikacije to je:

$$FPR(f) = \frac{c_{12}(f)}{c_{12}(f) + c_{11}(f)} = \frac{FP}{FP + TN}. \tag{5.21}$$

43 Mjera se koristi u medicini za ocjenu osjetljivosti kliničkih testova u otkrivanju bolesti tj. koliko je stvarnih slučajeva bolesti moguće uspješno detektirati, pa otuda i naziv (Japkowitz i Shah, 2011.).

Stopa stvarno negativnih (eng. *True negative rate – TNR)*, nazvana još i specifičnost (eng. *Specificity*):

$$TNR(f) = \frac{c_{11}(f)}{c_{11}(f) + c_{12}(f)} = \frac{TN}{TN + FP} \tag{5.22}$$

Stopa lažno negativnih (eng. *False negative rate – FNR*):

$$FNR(f) = \frac{c_{21}(f)}{c_{21}(f) + c_{22}(f)} = \frac{FN}{FN + TP} \tag{5.23}$$

Pri čemu vrijedi:

$$TPR(f) = 1 - FNR(f) \tag{5.24}$$

Pozitivna prediktivna vrijednost (eng. *Positive predictive value – PPV)*, nazvana još preciznost (eng. *Precision*) predstavlja udio primjera koji stvarno pripadaju klasi i među svim primjerima koji su klasificirani kao klasa i, odnosno mjeri koliko je algoritam precizan u identificiranju primjera određene klase. U slučaju binarne klasifikacije:

$$PPV(f) = \frac{TP}{TP + FP} \tag{5.25}$$

Negativna prediktivna vrijednost (eng. *Negative predictive value* - *NPV*), koja mjeri udio ispravno dodijeljenih negativnih primjera, računa se kao:

$$NPV(f) = \frac{TN}{TN + FN} \tag{5.26}$$

Pozitivna i negativna prediktivna vrijednost daju uvid u to koliko su pouzdana predviđanja klasifikatora s obzirom na pojedinu klasu. Na primjer, ako je PPV= 0.3 to znači da pozitivna predviđanja treba uzeti s rezervom jer su točna samo u 30% slučaja.

Za razliku od prethodno navedenih mjera, **geometrijska sredina** osjetljivosti i specifičnosti (TPR, TNR) uzima u obzir relativnu ravnotežu u performansama klasifikatora kako na pozitivnim tako i na negativnim primjerima.

$$G - mean_1(f) = \sqrt{TPR(f) \times TNR(f)} \tag{5.27}$$

Geometrijska sredina postaje 1 samo kad je $TPR(f) = TNR(f) = 1$

Druga verzija geometrijske sredine uzima u obzir osjetljivost i preciznost, odnosno

uzima u obzir udio stvarno pozitivnih primjera koji su kao takvi i označeni, te udio primjera označenih kao pozitivni koji su doista pozitivni.

$$G - mean_2(f) = \sqrt{TPR(f) \times PPV(f)} \tag{5.28}$$

F-mjera je ponderirana harmonijska sredina preciznosti i odziva. Za $\alpha \in \mathbb{R}, \alpha > 0$, općenita formulacija F-mjere je dana sa:

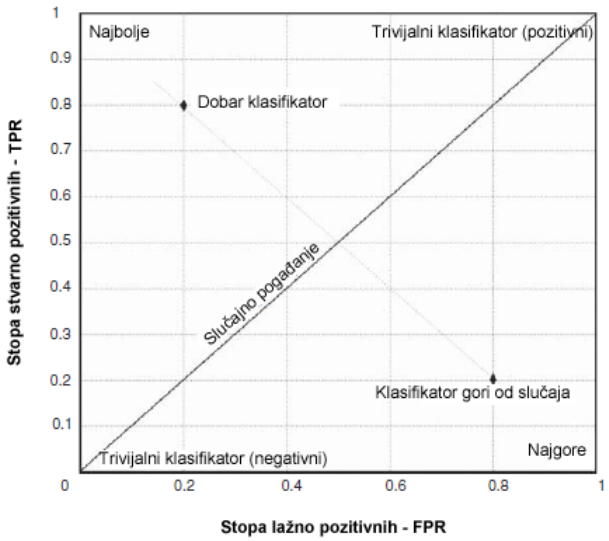
$$F_\alpha = \frac{(1+\alpha)[Prec(f) \times Rec(f)]}{[\alpha \times Prec(f)] + Rec(f)} \tag{5.29}$$

Najčešće prikladne težine nisu poznate pa se tada koristi uravnotežena F-mjera kod koje je $\alpha = 1$:

$$F_1 = \frac{2[Prec(f) \times Rec(f)]}{[Prec(f) + Rec(f)]} \tag{5.30}$$

ROC (*Receiver Operating Characteristic)* **krivulja** ima svoje porijeklo u teoriji detekcije signala, a u kontekstu strojnog učenja koristi se za vizualizaciju performansi algoritama za učenje uobičajeno kod binarne klasifikacije.

ROC analiza proučava vezu između osjetljivosti i specifičnosti klasifikatora. U prikazu ROC krivulje na horizontalnoj se osi prikazuje FPR (odnosno 1-specifičnost), a na vertikalnoj osi TPR (odnosno osjetljivost). ROC prostor je jedinični kvadrat sa $0 <= TPR <= 1$ i $0 <= FPR <= 1$ prikazan na slici 34.



Slika 34. *ROC prostor: stavlja u odnos stopu stvarno pozitivnih i stopu lažno pozitivnih primjera*

Izvor: Japkowitz i Shah (2011.)

Točka (0,0) označava trivijalni klasifikator koji klasificira sve instance u negativne. Točka (1,1) označava trivijalni kalsifikator koji sve instance klasificira u pozitivne. Dijagonala koja povezuje (0,0) i (1,1) ima TPR = FPR u svakoj točki. Klasifikator koji se nalazi na dijagonali može se smatrati klasifiktaotrom koji slučajnim odabirom dodjeljuje oznake klasa. Klasifikatori koji se nalaze iznad dijagonale bolji su od slučaja, dok su oni ispod dijagonale gori od slučaja. Točke (1,0) i (0,1) dva su ekstrema u ROC prostoru. Točka (1,0) ima FPR = 1, a TPR = 0 što znači da su sva predviđanja pogrešna. S druge strane, (0,1) označava idealni klasifikator, koji klasificira točno sve pozitivne instance i ne griješi na negativnima. Dijagonala koja povezuje te točke ima TPR = 1-FPR. 1-FPR je isto što i TNR. Na toj su dijagonali jednako dobre performanse na negativnim i pozitivnim instancama.

Klasifikator koji se nalazi bliže lijevoj strani na grafu je konzervativniji u klasificiranju pozitivnih instanci, u smislu da preferira grešku u prepoznavanju pozitivnih nego da riskira pogrešno klasificiranje negativnih instanci. Ako se nalazi bliže desnoj strani onda je liberalniji u klasificiranju pozitivnih instanci, preferira pogrešno klasificirati negativne nego da promaši prepoznati pozitivne.

U slučaju determinističkog klasifikatora, kao što je to SVM koji kao rezultat predviđanja daje samo oznake klasa, on će u ROC prostoru bit će prikazan samo jednom točkom, dok se u slučaju probabilističkih klasifikatora dobije krivulja i tada se može računati površina ispod ROC krivulje što daje mjeru AUC. Kod slučajnog klasifikatora AUC = 0.5, dok je kod savršenog klasifikatora AUC = 1.

Prednost ROC krivulje i AUC mjere jest to što su pogodni i u slučaju neravnoteže klasa.

U okviru aplikacije, uz točnost, izračunavaju se sve prethodno navedene mjere: AUC, TPR, NPR, FPR, FNR, TPV, NPV, geometrijske sredine, F-mjera, te se nude grafički prikazi ROC krivulje i konfuzijske matrice.

Iako je SVM daje deterministički klasifikator LibSVM bibilioteka omogućava da se, uz oznake klasa, kao izlaz dobije i vjerojatnost pripadnosti primjera određenoj klasi što se koristi kod izrade ROC krivulje i za računanje AUC mjere.

5.9.2. Simulacija trgovanja

Rezultati predviđanja na testnim podacima predstavljaju ulaz u simulator trgovanja. Velika ostvarena točnost predviđanja (ili dobra vrijednost neke druge mjere) ne mora rezultirati i jednako dobrim ponašanjem u stvarnim uvjetima. S obzirom da je problem u ovom radu sveden na binarnu klasifikaciju koja ne uzima u obzir magnitudu promjene prinosa, moguća je situacija u kojoj se postiže velika točnost predviđanja predznaka na malim promjenama, ali se, zbog samo nekoliko netočnih predviđanja velikih negativnih promjena, u konačnici ipak završi s gubitkom.

Kako bi se testiralo ponašanje klasifikatora pri trgovanju i omogućilo isprobavanje različitih strategija, omogućeno je kombiniranje nekoliko klasifikatora, od kojih svaki može učiti rješavati različiti zadatak.

Podržano je nekoliko strategija:

- *"Tri plus tri minus"* - ako se predviđa da će cijena rasti/padati tri dana zaredom, kupi/prodaj, inače čekaj (koristi tri klasifikatora),
- *"Dva plus dva minus"* – kao prethodna, ali ograničeno na dva dana i dva klasifikatora,
- trgovanje na dnevnoj bazi (koristi jedan klasifikator).

Na temelju rezultata predviđanja i odabrane strategije generira se signal za kupnju ili prodaju. Nije dopuštena kratka prodaja[44] niti drugo zaduživanje u novcu tako da se sve transakcije realiziraju u okviru raspoloživih sredstava pri čemu transakcijski troškovi nisu uzeti u obzir.

Radi bolje procjene rezultata, omogućen je odabir i slučajno generiranog signala (sa ili bez opcije držanja), te obrnuta strategija kod koje se stvara signal suprotan rezultatu primjene prethodno navedenih strategija.

44 Kratka prodaja je špekulativna operacija u kojoj investitor, očekujući pad cijena, posuđuje dionicu te ju odmah prodaje. Nakon određenog vremena kupuje dionicu s namjerom da je vrati. Ako je doista došlo do pada cijene, investitor je zaradio na razlici cijena umanjenoj za cijenu posudbe.

Primjer 8. *Strategija "Tri plus tri minus"*

```
% ako će cijena rasti tri dana zaredom, kupi
% ako će cijena padati tri dana zaredom, prodaj
% inače čekaj
function rezultat = triPlusTriMinus(podaci, ulog)
 signal_kupi   = podaci.R1 == 1 & podaci.R2 ==1 & podaci.R3 == 1
 signal_prodaj = podaci.R1 == 0 & podaci.R2 ==0 & podaci.R3 == 0
 idx_k = find(signal_kupi   == 1);
 idx_p = find(signal_prodaj == 1);

 signal = zeros(size(podaci,1),1);
 signal(idx_k) = 1;
 signal(idx_p) = -1;

 novac    =  zeros(size(signal,1)+1, 1);
 dionica  =  zeros(size(signal,1)+1, 1);
 prinos   =  zeros(size(signal,1)+1, 1);
 novac(1) = ulog;

 for i = 1:size(signal,1)
     if signal(i) == 1 %kupi, ako ima dovoljno novca
         if novac(i)>= podaci.ZADNJA(i)
             % kupi koliko god ima novca
             kolicina = floor(novac(i)/podaci.ZADNJA(i))
             novac(i+1) = novac(i) - kolicina*podaci.ZADNJA(i)
             dionica(i+1) = dionica(i) + kolicina
         else
             % ne radi ništa
             novac(i+1) = novac(i)
             dionica(i+1) = dionica(i)
         end
     elseif signal(i) == 0 % ne radi ništa
         novac(i+1) = novac(i)
         dionica(i+1) = dionica(i)
     elseif signal(i) == -1 % prodaj, ako ima dionica
         if dionica(i) > 0
             novac(i+1) = novac(i) + dionica(i) * podaci.ZADNJA(i)
             dionica(i+1) = 0
         else
             % ne radi ništa
             novac(i+1) = novac(i)
             dionica(i+1) = dionica(i)
```

```
        end
    end
end
rezultat = table(podaci.DATUM, signal, novac(2:end), dionica(2:end),...
'VariableNames',{'DATUM', 'SIGNAL', 'NOVAC', 'DIONICA'});
end
```

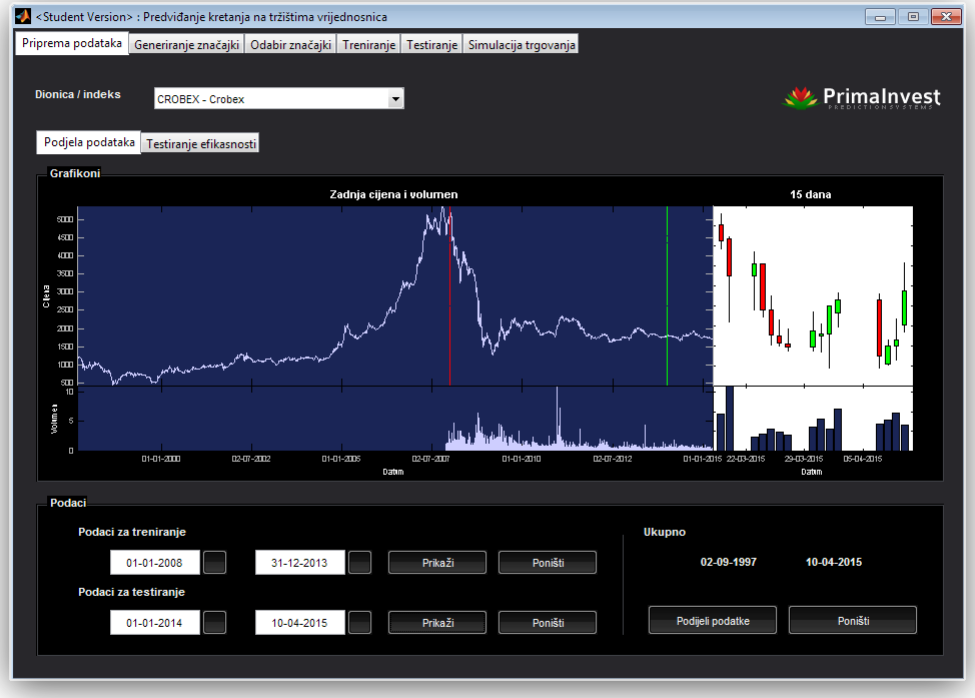
5.10. Korisničko sučelje i primjer upotrebe

Prilikom izrade sučelja vodilo se nastojanjem da se brojnim vizualizacijama doprinese boljem razumijevanju kvantitativnih informacija, uvažavajući pritom temeljna pravila upotrebljivosti (eng *Usability*)[45].

5.10.1. Sučelje za pregled i podjelu podataka

Sučelje za pripremu podataka (slika 35.) omogućava prikaz svih podataka sadržanih u bazi odabranog indeksa ili dionice te njihovu podjelu na odgovarajuće skupove za treniranje i testiranje. Na većem grafikonu (lijevo) prikazuju se zadnja cijena i volumen, dok se desno prikazuju volumen i grafikon svijeća (eng. *Candlestick chart*) za posljednjih 15 dana.

45 Upotrebljjivost se odnosi na lakoću kojom korisnici koriste aplikaciju prilikom postizanja svojih ciljeva. Obuhvaća sljedeće elemente: efektivnost, efikasnost, sigurnost, korisnost, jednostavnost učenja, lakoću pamćenja (Sharp, Rogers i Preece, 2002).



Slika 35. *Sučelje za pregled i podjelu podataka*

Izvor: izradila autorica

Podjela podataka označava se uspravnim linijama: crvenim linijama označen je skup za treniranje, a zelenima skup za testiranje. Ako podjela podataka još nije izvršena ili je ona poništena, na grafikonu se sugerira podjela cjelokupnog skupa u omjeru 70% za treniranje i 30% za testiranje.

5.10.2. Sučelje za testiranje efikasnosti

Na drugom se *tabu* u sklopu početne pripreme podataka (slika 36.) prikazuju logaritmi prinosa i korelogrami za cjelokupno razdoblje, te zasebno prema prethodno odabranoj podjeli podataka. U početnom se odabiru prikazuje zadnja cijena, ali omogućen je i prikaz ostalih cijena (prve, najviše, najniže) te transformacija zadnje cijene – apsolutne vrijednosti, kvadrata i kuba.



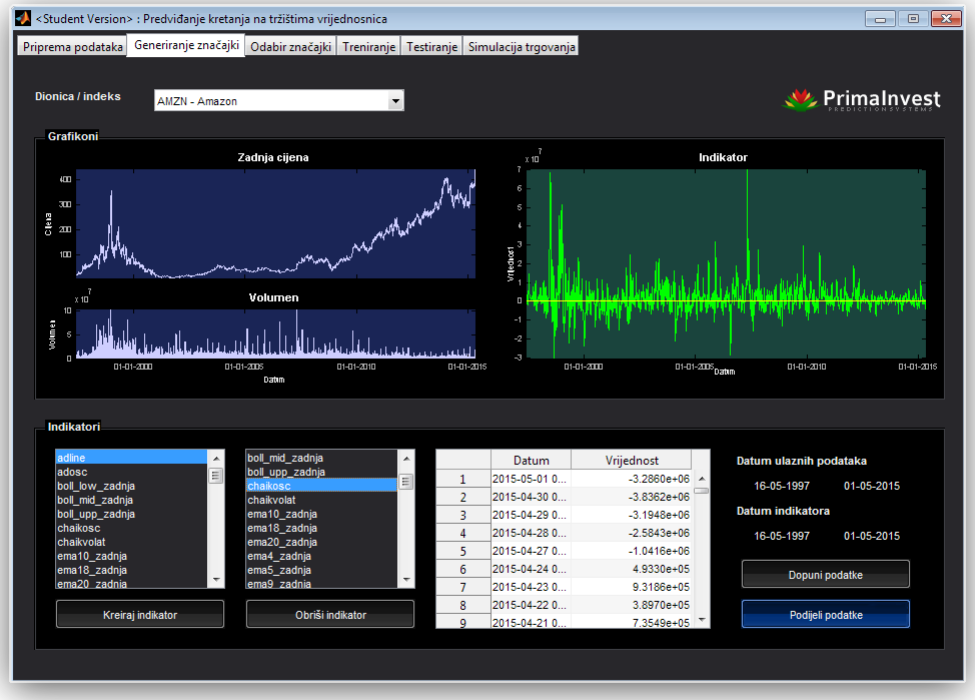
Slika 36. *Sučelje za testiranje efikasnosti*

Izvor: izradila autorica

U donjem dijelu prikazuju se rezultati testa omjera varijanci. Prvi prikazani rezultat odnosi se na slučaj koji uzima u obzir heteroskedastičnost, dok drugi testira i.i.d. RW.

5.10.3. Sučelje za generiranje značajki

Za odabranu dionicu ili indeks, čija se zadnja cijena i volumen prikazuju na grafikonu lijevo (slika 37.), željeni se indikator kreira tako da se odabere s predefiniranog popisa indikatora za koje već postoje gotove formule, a njegovi se detalji prikazuju tabelarno i grafički na grafikonu desno. Već kreirani indikatori za određenu dionicu/indeks vidljivi su na zasebnom popisu.



Slika 37. *Sučelje za generiranje značajki*

Izvor: izradila autorica

Nakon kreiranja indikatora potrebno je podijeliti i njihove podatke u skladu s prethodno odabranom podjelom sirovih podataka na trening/test skup. Također, nakon dodavanja novih sirovih podataka u bazu, potrebno je preračunati indikatore. Datumi sirovih podataka i datumi indikatora prikazuju se krajnje desno kako bi se mogli lakše uskladiti.

5.10.4. Sučelje za odabir značajki

Sučelje omogućava da se proces odabira značajki započne sa svim kreiranim varijablama, samo onima čiji se izračuni baziraju na zadnjoj cijeni (npr. za slučaj kad nisu dostupni podaci o svim cijenama ili volumenu) ili izvrši njihov pojedinačni odabir (slika 38.). Omogućeno je postavljanje osnovnih opcija RF algoritma (broj stabala, minimalni broj instanci u listu, broj varijabli u svakoj podjeli). Preostale opcije ostavljene su na pretpostavljenim vrijednostima. Kako bi se ubrzalo preliminarno testiranje, omogućeno je skraćivanje niza podataka.



Slika 38. *Sučelje za odabir značajki*

Izvor: izradila autorica

Rezultati primjene RF algoritma prikazuju se grafički i to:

- greška klasifikacije,
- matrica blizina[46] pri čemu je izvršeno višedimenzionalno skaliranje[47]
- varijable rangirane po značaju.

Nakon rangiranja na korisniku preostaje da izvrši konačan odabir varijabli i njihovog broja ili nastavi postupak. Na temelju odabira kreiraju se LibSVM datoteke.

5.10.5. Sučelje za odabir modela i treniranje

U gornjem dijelu sučelja prikazuju se osnovne informacije o podacima te njihove vizualizacije prema klasama i značajkama – negativni primjeri u gornjem dijelu grafičkog

46 Matrica blizina - mjera sličnosti između parova instanci. Jednaka je proporciji stabala za koja dvije različite instance završe u istom listu (terminalnom čvoru) stabla.

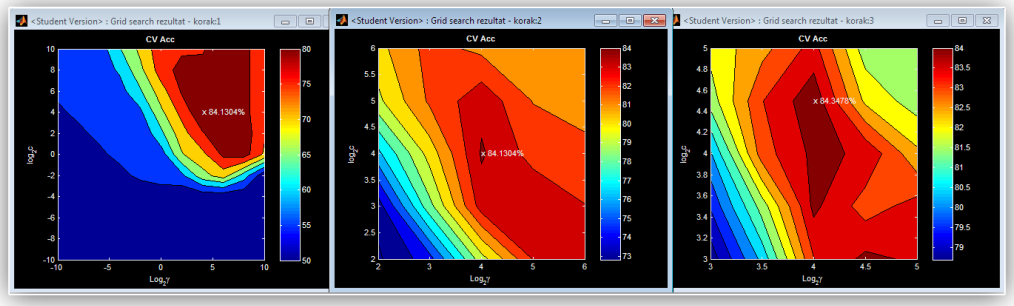
47 Višedimenzionalno skaliranje (eng. m*ulti-dimensional scaling - MDS*) - pronalaženje konfiguracije točaka u niže dimenzionalnom prostoru čije međusobne udaljenosti korespondiraju sličnosti u višedimenzionalnom prostoru.

prikaza, a pozitivni u donjem dijelu. Drugi dio sučelja sastoji se od tri *taba* (na slici 40. prikazani jedan ispod drugoga):

- Na prvome *tabu* bira se način odabira modela te postavke za provođenje *grid-search* pretraživanja. Rezultati pretraživanja prikazuju se grafički pri čemu se posebno označava najbolja postignuta točnost za ispitane kombinacije parametara (slika 39.).

- Na drugome se *tabu* odabire metoda evaluacije koja se primjenjuje prilikom pretraživanja optimalnih parametara. Pretpostavljeni odabir je unakrsna validacija.

- Treći *tab* namijenjen je treniranju. Podaci s označenim potpornim vektorima prikazuju se u prikazu smanjene dimenzionalnosti. Omogućeno je spremanje modela.



Slika 39. *Grafički prikaz rezultata pretraživanja optimalnih parametara modela*
Izvor: izradila autorica



Slika 40. *Sučelje za odabir modela i treniranje*

Izvor: izradila autorica

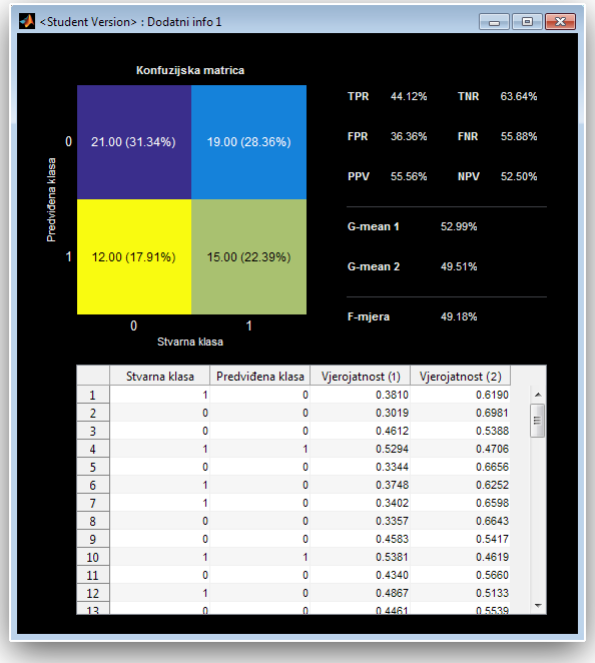
5.10.6. Sučelje za testiranje modela

Kao i kod sučelja za treniranje, omogućen je prikaz osnovnih informacija o podacima za testiranje (slika 41.). Model koji se želi testirati odabire se s popisa kreiranih za određenu dionicau/indeks pri čemu se prikazuju i osnovne informacije o treniranju (procijenjena točnost i vrijednosti parametara). Rezultati testiranja prikazuju se u obliku evaluacijskih mjera: točnost, AUC i ROC krivulja, a u dodatnom prozoru konfuzijska matrica s preostalim mjerama (slika 42.). Rezultate testiranja moguće je pohraniti radi naknadne upotrebe u simulatoru trgovanja.



Slika 41. *Sučelje za testiranje modela*

Izvor: izradila autorica



Slika 42. *Konfuzijska matrica i dodatne evaluacijske mjere*

Izvor: izradila autorica

5.10.7. Sučelje za simulaciju trgovanja

Omogućeno je kombiniranje dvaju ili triju klasifikatora ili korištenje samo jednoga, ovisno o odabranoj strategiji trgovanja. Tijekom simulacije na grafikonu se u gornjem dijelu prikazuju cijena i generirani signali za kupnju (zelene vertikalne linije) ili prodaju (crvene linije), dok se u donjem dijelu prikazuje evolucija financijskog statusa.



Slika 43. *Sučelje za simulaciju trgovanja*

Izvor: izradila autorica

Konačan rezultat simulacije predstavlja ostvareni prinos na odabrano početno ulaganje, dok je detaljan pregled izvršenih transakcija dostupan u zasebnom prozoru.

6. EKSPERIMENT I REZULTATI

"Ako mučite podatke dovoljno dugo, priznat će!"

Ronald Coase

Testirana je mogućnost predviđanja smjera kretanja na nekoliko tržišta vrijednosnica koristeći SVM algoritam i podatke dobivene temeljem tehničke analize. Predviđala se promjena predznaka prinosa na određeni dan u budućnosti. Na temelju rezultata predviđanja vršila se simulacija trgovanja čime je dobivena konačna ocjena izgrađenoga modela. Zbog dugotrajnosti čitavog postupka, prije samoga početka provedeni su testovi predvidljivosti vremenskih nizova kako bi se opravdavalo poduzimanje naknadnih koraka. Glavni koraci u eksperimentu prikazani su na slici 44.



Slika 44. *Glavni koraci u eksperimentu*

.Izvor: izradila autorica

Pitanja na koja je eksperiment trebao dati odgovore:

- doprinosi li veća količina podataka boljim rezultatima predviđanja,
- posjeduju li tehnički indikatori dovoljnu prediktivnu moć,
- doprinose li boljim rezultatima predviđanja dodatne informacije sadržane u volumenu i ostalim cijenama (osim zaključne cijene),
- doprinosi li boljim rezultatima predviđanja odabir značajki,
- pomaže li kompenzacija neravnoteže u podacima,
- može li se kombinacijom klasifikatora poboljšati rezultat,
- postoje li kombinacije parametara koje općenito daju bolje rezultate na različitm skupovima podataka,
- uspijeva li SVM otkriti veze između podataka i oznaka klasa.

6.1. Odabir burzovnih indeksa

Zbog iznesenoga u poglavlju 4.2 o dvojbenoj korisnosti tehničke analize u predviđanju budućih kretanja, za eksperiment su odabrana samo ona tržišta kojima je utvrđena određena razina neefikasnosti i to na temelju prethodno objavljenih rezultata istraživanja svjetskih tržišta kapitala. Kristoufek i Vosvrda (2013.) istražili su 41 svjetsko tržište analizirajući fraktalnu dimenziju, Hurstov eksponent i entropiju te su ista rangirali s obzirom na predloženu mjeru efikasnosti definiranu kao udaljenost od idealnog stanja, odnosno efikasnog tržišta. U promatranome razdoblju od 2000. godine do 2011. godine slovačko se tržište pokazalo kao jedno od najneefikasnijih. Do sličnog je rezultata došla i Baciu (2014.) usmjeravajući svoje istraživanje isključivo na europska tržišta. U razdoblju od 1999. godine do 2013. godine, prema rezultatima nekoliko mjera efikasnosti, ponovno se slovačko pozicioniralo među najneefikasnijima. Stoga je slovačko tržište, predstavljeno burzovnim indeksom SAX[48], odabrano za prvog kandidata daljnjih koraka eksperimenta.

Za predstavnika efikasnijeg, ali ne i potpuno efikasnog tržišta (Kristoufek i Vosvrda, 2013.) uvršteno je američko tržište predstavljeno indeksom S&P500[49] koji je ujedno i jedan od najčešće istraživanih indeksa (u Tsaih et al. (1998.) autori predviđaju njegovo kretanje pomoću hibridnog sustava, u Huang et al. (2005.) se koristi kao ulazna varijabla za predviđanje japanskog indeksa NIKKEI, a Atsalakis i Valavanis (2009.) daju detaljniji popis radova u kojima se taj indeks istražuje).

Barbić (2010.b) je, testirajući slabi oblik EMH na hrvatskom tržištu kapitala u razdoblju od 1997. godine do 2007. godine pomoću proširenog Dickey-Fullerovog ADF testa i testa autokorelacije, utvrdila da prinosi indeksa CROBEX[50] ne slijede slučajni hod. Stoga je u inicjalni odabir uključen i indeks Zagrebačke burze, a njemu su pridodana još i tri novija

48 SAX (***S****lovenský* ***a****kciový inde****x***) – službeni dionički indeks slovačke burze Bratislava Stock Exchange. Uspoređuje tržišnu kapitalizaciju skupa dionica u odnosu na bazni datum (14.09.1993.). Osim fluktuacija cijena, uključuje i dividende te zarade povezane s povećanjem kapitala. Udio pojedine sastavnice iznosi maksimalno 20%.

49 S&P500 – dionički indeks u čijem su sastavu dionice 500 najvećih kompanija kojima se trguje u SAD-u pri čemu je udio svake dionice određen njezinom tržišnom vrijednošću. S obzirom na široku bazu dionica u sastavu indeksa, predstavlja jedan od glavnih indikatora stanja američkog tržišta. Ime je dobio prema financijsko-konzultantskoj kompaniji Standard & Poor's koja ga je kreirala, a u ovom obliku postoji od 1957. godine.

50 CROBEX – najstariji službeni dionički indeks Zagrebačke burze koji se počeo objavljivati 1. rujna 1997. godine. Za ulazak u sastav indeksa uzimaju se u obzir samo dionice kojima se trgovalo više od 90% ukupnog broja trgovinskih dana u određenom razdoblju pri čemu maksimalna težina pojedine sastavnice iznosi 10%. Bazni datum je 01.07.1997. Dividende se ne uključuju u izračun indeksa Zagrebačke burze.

indeksa, CROBEXindustrija, CROBEXturist[51], CROBEX10[52].

6.2. Podaci

Sirovi podaci prikupljeni su s web sjedišta Zagrebačke burze (za burzovne indekse CROBEX, CROBEXturist, CROBEXindustrija, CROBEX10), s web sjedišta slovačke burze Bratislava Stock Exchange (za burzovni indeks SAX), dok je za S&P500 korišten Yahoo Finance servis. Podatke čine dnevne cijene: prva, zadnja, najniža i najviša, te volumen. Za indeks SAX dostupni su bili samo podaci o njegovoj vrijednosti, dok su za indekse CROBEX, CROBEXturist, CROBEXindustrija svi podaci bili dostupni samo za ograničeni period[53].

6.3. Prethodno testiranje i rezultati

Ukupna duljina razdoblja, prema pojedinom indeksu, za koja su provedena testiranja efikasnosti prikazana su u tablici 4. Za indekse Zagrebačke burze i indeks SAX to su ujedno i razdoblja od početaka njihove primjene.

Tablica 4. *Indeksi i razdoblja podvrgnuta testiranju efikasnosti tržišta*

Indeks	Početni datum	Završni datum	Broj dana
CROBEX	02.09.97	10.04.15	4309
CROBEXindustrija	30.12.11	10.04.15	816
CROBEXturist	30.12.11	10.04.15	816
CROBEX10	31.07.09	10.04.15	1423
S&P500	03.01.50	01.05.15	16438
SAX	03.07.95	04.05.15	4855

Izvor: izradila autorica

U tablici 5. prikazane su neke od mogućih podjela podataka, za koje su provedeni testovi, zajedno s karakterističnim rezultatima testiranja[54]. Za svako razdoblje provedena su

51 CROBEXindustrija (industrijska proizvodnja) i CROBEXturist (turizam) – dionički indeksi kod kojih je uvjet za uključenje u njihov sastav 70% dana trgovanja dionicom od ukupnog broja trgovinskih dana u promatranom razdoblju. Svaka dionica ima jednaku težinu, a broj sastavnica je neograničen. Bazni datum je 21.02.2013.

52 CROBEX10 – sastoji se od deset najlikvidnijih dionica iz indeksa CROBEX s najvećom *free float* tržišnom kapitalizacijom i najvećim ostvarenim prometom na Zagrebačkoj burzi. Težina pojedine dionice ograničena je na 20%. Bazni datum je 21.02.2013.

53 Datumi od kojih su dostupni svi podaci: CROBEX od 23.11.2007., CROBEXindustrija od 22.02.2013., CROBEXturist od 22.02.2013.

54 Testirane su i druge moguće podjele za koje su testovi pokazali da nije moguće odbaciti hipotezu o slučajnom hodu. Budući da su za eksperiment zanimljiva samo ona razdoblja za koja je utvrđena određena razina predvidljivosti, njihov je prikaz izostavljen.

po dva testa omjera varijanci: jedan testira i.i.d. slučajan hod (test 2 u tablici 5.), dok drugi uzima u obzir heteroskedastičnost (test 1 u tablici 5.). Crvenom bojom označena su razdoblja za koja je test omjera varijanci pokazao da nije moguće odbaciti hipotezu o slučajnom hodu[55], dok su zelenom bojom označena razdoblja za koja se ta hipoteza odbacuje. Ako test 1 pokaže da nije moguće odbaciti hipotezu o RW, a test 2 ju odbacuje, radi se o heteroskedastičnom RW.

Tablica 5. *Rezultati testa omjera varijanci*

	Skup za treniranje		Skup za testiranje	
	Test 1	Test2 (i.i.d.)	Test 1	Test2 (i.i.d.)
CROBEX	2.9.1997-31.12.2007		1.1.2008 - 10.4.2015	
	0	1	0	1
	1.1.2008 -31.12.2013		1.1.2014 - 10.4.2015	
	0	1	0	0
	23.11.2007 - 31.12.2014		1.1.2015 - 10.4.2015	
	0	1	0	0
	2.9.1997 - 31.12.2014		1.1.2015 - 10.4.2015	
	0	0	0	0
CROBEX10	31.7.2009 - 31.12.2014		1.1.2015 - 10.4.2015	
	0	0	0	0
CROBEXindustrija	30.12.2011 - 15.4.2014		16.4.2014 - 10.4.2015	
	0	0	0	0
	30.12.2011 - 31.12.2014		1.1.2015 -10.4.2015	
	0	0	1	1
	22.2.2013 -31.12.2014		1.1.2015 -10.4.2015	
	0	1	1	1
CROBEXturist	30.12.2011 - 31.12.2014		1.1.2015 - 10.4.2015	
	0	0	0	0
	22.2.2013 - 31.12.2014		1.1.2015 - 10.4.2015	
	0	1	0	0
S&P500	3.1.1950 - 25.9.1995		26.9.1995 - 1.5.2015	
	1	1	1	1
	1.1.2008 - 31.12.2014		1.1.2015 - 10.4.2015	
	1	1	0	0
SAX	3.7.1995 - 21.5.2009		22.5.2009 -4.5.2015	
	0	0	1	1
	1.1.2008 - 31.12.2013		1.1.2014 - 10.4.2015	
	1	1	1	1
	1.1.2000 -31.12.2013		1.1.2014 -10.4.2015	
	0	1	1	1

Izvor: izradila autorica

55 Na razini značajnosti od 5%.

Iz priložene tablice vidljivo je da slučajnost određenog niza nije konstantna već se mijenja tijekom vremena, što znači da potencijalne prilike nastaju i iščezavaju. Kao što je bilo i očekivano, indeks SAX pokazao je predvidljivost u većini testiranih intervala, iako ne i u ukupnom razdoblju, dok indeks SP&500 predvidljivost pokazuje u ukupnom razdoblju[56] te, ako se promatra kraće razdoblje, u onom dijelu koji obuhvaća razdoblje globalne financijske krize[57]. Zanimljivo je uočiti kako razlika od samo nekoliko mjeseci utječe na drugačiji ishod testiranja, što je posebno izraženo kod indeksa CROBEXindustrija.

Indeksi odabrani za provođenje eksperimenta, zajedno s pripadajućim podjelama podataka, prikazani su u tablici 6.

Tablica 6. *Odabrani indeksi i podjele podataka na skupove za učenje i testiranje*

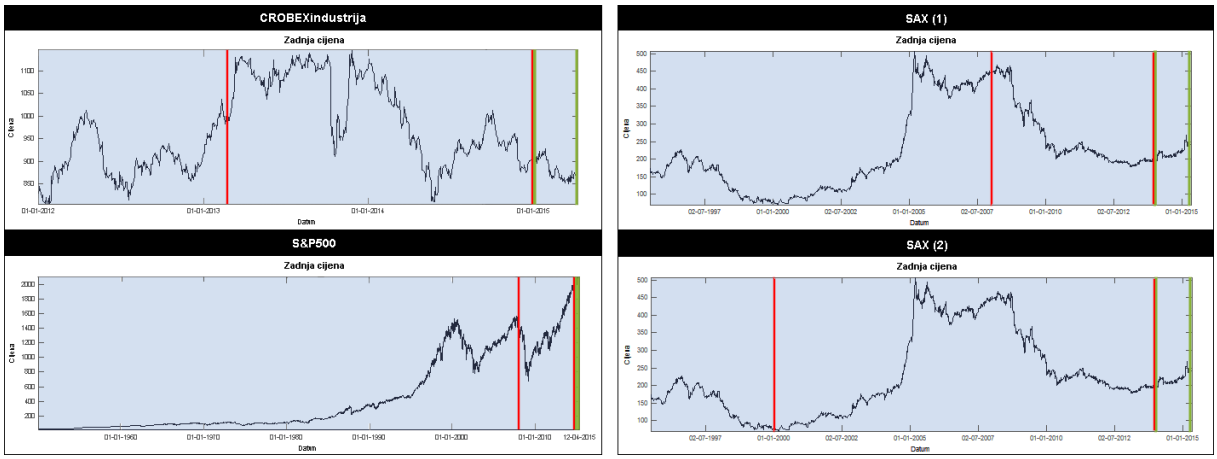
Indeks	Početak	Kraj	Broj dana	Skup
CROBEXindustrija	22.02.13	31.12.14	460	učenje
CROBEXindustrija	02.01.15	10.04.15	68	testiranje
S&P500	02.01.08	31.12.14	1763	učenje
S&P500	02.01.15	10.04.15	68	testiranje
SAX (1)	07.01.08	31.12.13	1495	učenje
SAX (1)	02.01.14	10.04.15	316	testiranje
SAX (2)	07.01.00	31.12.13	3435	učenje
SAX (2)	02.01.14	10.04.15	316	testiranje

Izvor: izradila autorica

Iako se općenito bolji rezultati strojnog učenja postižu s većim skupovima za učenje, kod financijskih vremenskih nizova to ne mora biti slučaj. Vremenski nizovi mogu sadržavati uzorke koji se u budućnosti više neće ponavljati, što znači da njihova identifikacija ne bi bila korisna za ostvarenje profita, a što je krajnji cilj ovakvog sustava. Kako bi se ispitalo doprinosi li veća količina podataka boljim rezultatima predviđanja, za indeks SAX odabrane su dvije podjele s različitim rezultatom prethodnog testiranja: kraća serija koja obuhvaća razdoblje od šest godina i za koju je odbačena hipoteza slučajnog hoda, te produljenje tog razdoblja na 14 godina, za koje je utvrđen heteroskedastičan slučajni hod.

56 Iako bi bilo zanimljivo provesti eksperiment na ukupnom razdoblju, zbog ograničenja računala, to nije bilo izvedivo.

57 Financijska kriza započela je 2007. godine u SAD-u kao kriza tržišta nekretnina, ali ubrzo se proširila na ostatak svijeta te je dovela do propasti mnogih financijskih, a kasnije i nefinancijskih kompanija.



Slika 45. *Kretanja zaključne cijene odabranih indeksa s označenim podjelama podataka na skupove za učenje i testiranje*

Izvor: izradila autorica

Slika 45. prikazuje kretanje cijena u odabranim razdobljima. Dodatno su na pojedinim grafikonima označene podjele podataka – crvenim linijama skup za treniranje, zelenim linijama skup za testiranje. Može se vidjeti da indeksi CROBEXindustrija i SAX pokazuju slično kretanje: oba najprije pokazuju veći rast, zatim značajniji pad, te vraćanje na početnu razinu. Kretanje S&P500 indeksa pokazuje ponešto drugačiji tijek s dva uočljiva loma – prvi se odnosi na "*dot.com bubble*"[58] oko 2000. godine, a drugi na financijsku krizu s kraja protekloga desetljeća. Za razliku od prethodna dva, ovaj indeks pokazuje kontinuirani rast posljednjih nekoliko godina koji ujedno nadmašuje sve prethodne vrijednosti.

Tablica 7. *Pregled deskriptivne statistike za logaritme prinosa*

	CROBEXIndustrija		SAX (1)		SAX (2)		S&P500	
	Trening	Test	Trening	Test	Trening	Test	Trening	Test
Sredina (%)	-0,019	-0,055	-0,055	0,080	0,027	0,080	0,020	0,032
Standardna devijacija (%)	1,384	1,051	1,228	1,385	1,213	1,385	1,452	0,860
Asimetrija	-0,664	0,096	-1,824	-0,002	-0,883	-0,002	-0,307	-0,046
Zaobljenost	11,372	2,661	31,421	16,175	19,107	16,175	12,216	2,383
Minimum (%)	-10,023	-2,315	-14,810	-9,329	-14,810	-9,329	-9,470	-1,845
Maksimum (%)	7,144	2,587	11,880	9,118	11,880	9,118	10,960	1,773
Jarque-Bera	1	0	1	1	1	1	1	0
N	459	67	1494	315	3434	315	1762	67

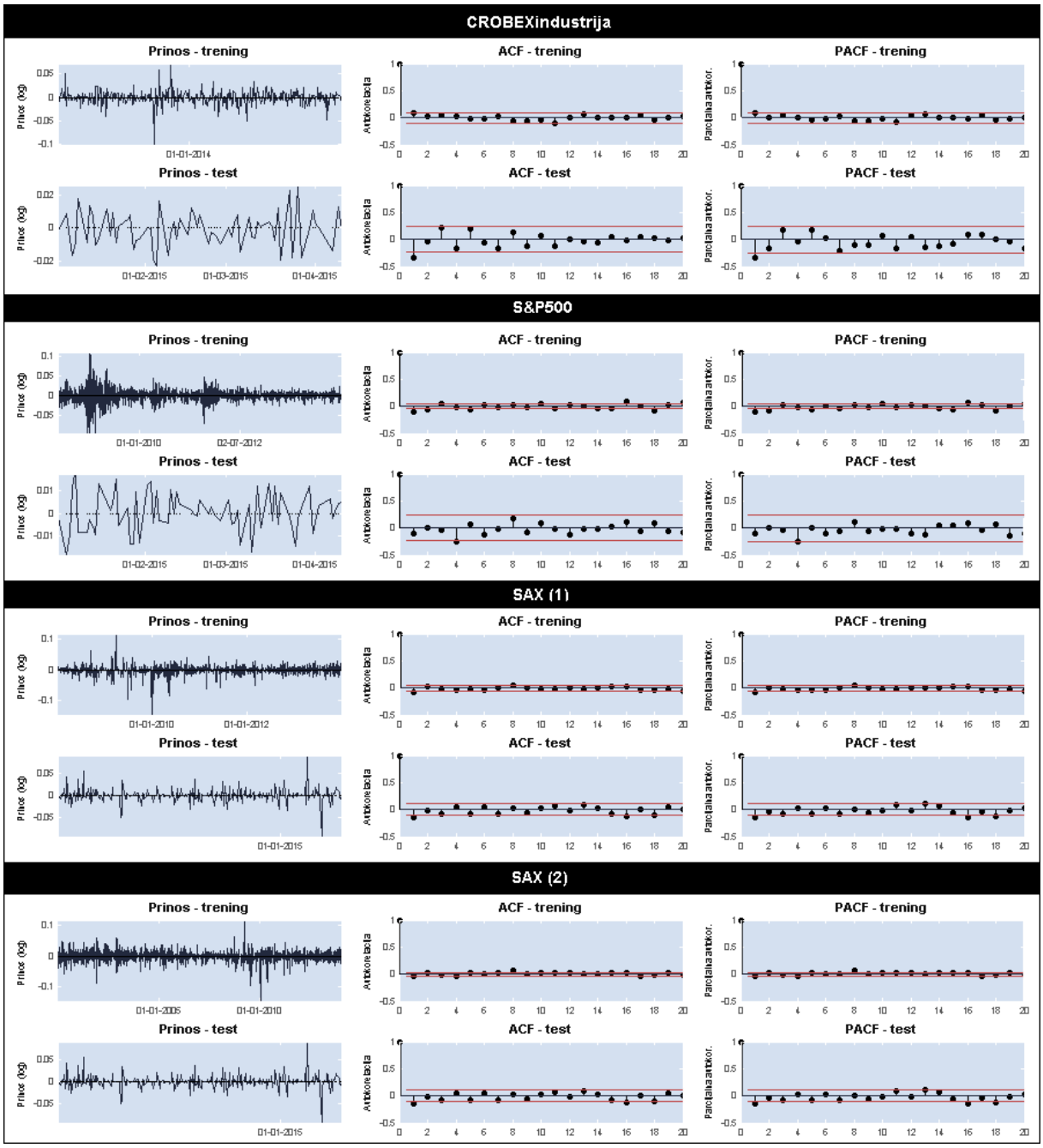
Izvor: izradila autorica

Iz tablice 7., koja daje pregled deskriptivne statistike za logaritme prinosa, može se

58 dot.com. bubble – odnosi se na propast brojnih internetskih kompanija iza čijih se nerealno visokih cijena dionica uglavnom nije nlazilo ništa osim prefiksa "e" ili sufiksa ".com" u njihovom nazivu.

primijetiti da je volatilnost općenito veća u razdoblju koje obuhvaća podatke za treniranje nego u razdoblju podataka za testiranje što je i razumljivo zbog zahvaćanja razdoblja turbulentnih događaja u svjetskoj ekonomiji (indeksi SAX i S&P500). Ako se promatra koeficijent asimetrije (eng. *skewness*), koji u slučaju normalne distribucije iznosi 0, indeksi CROBEXindustija i SAX (2) pokazuju umjerenu negativnu asimetriju, dok SAX (1) pokazuje izraženu negativnu asimetriju, što ukazuje na teški lijevi rep (eng. *fat tail*) odnosno na povećanu vjerojatnost ekstremnih događaja, u ovom slučaju velikih gubitaka. Koeficijent zaobljenosti (eng. *kurtosis*) za normalnu distribuciju iznosi 3. Vidljivo je da u svim razdobljima, osim u testnim razdobljima indeksa CROBEXindustrija i S&P500 za koja ujedno Jarque-Bera test ne odbacuje nultu hipotezu o normalnosti distribucije, prelazi tu vrijednost što znači da postoji veći broj stopa prinosa koje su približno jednake 0.

Slika 46. prikazuje korelograme koji su u skladu su s rezultatima testa omjera varijanci. Može se vidjeti da indeksi u svim testiranim razdobljima (osim S&P500 u skupu za testiranje) pokazuju statistički značajan koeficijent autokorelacije na prvom pomaku i to negativan, dok je kod indeksa CROBEXindustrija u skupu za treniranje on pozitivan. Na ostalim pomacima najviše je statistički značajnih koeficijenata autokorelacije utvrđeno kod indeksa S&P500.

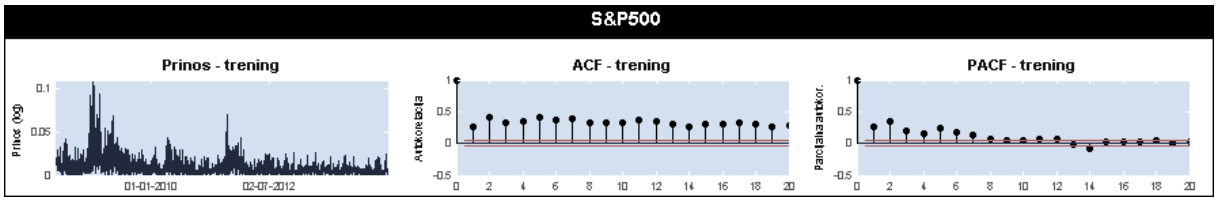


Slika 46. *Logaritam prinosa, ACF i PACF: a) CROBEXindustrija, b) S&P500, c) SAX (1), d) SAX (2)*

Izvor: izradila autorica

Usporedno je prikazan i logaritam prinosa. Može se primijetiti stvaranje klastera volatilnosti, odnosno pojave da velike promjene pozitivnog predznaka prate velike promjene negativnog predznaka i obrnuto, što je najviše izraženo kod S&P500 indeksa u skupu za treniranje. Ovu pojavu moguće je jasno vidjeti na korelogramu apsolutnih vrijednosti logaritma prinosa (slika 47.) koji pokazuju značajnu autokorelaciju na svim prikazanim

pomacima, dok u skupu za testiranje ona nije prisutna.



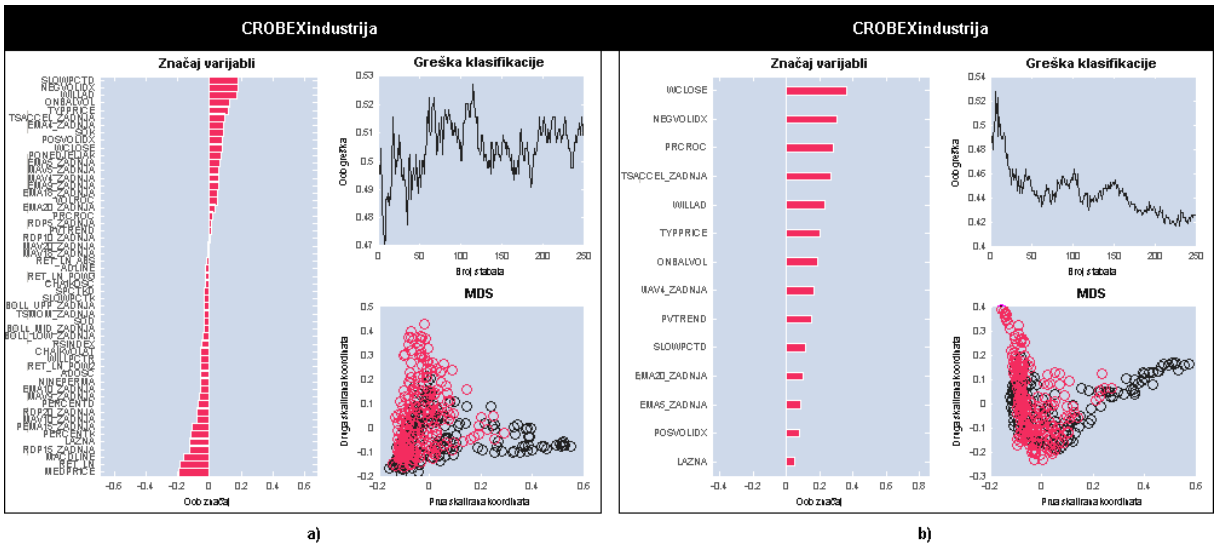
Slika 47. *Apsolutna vrijednost logaritma prinosa, ACF, PACF za indeks S&P500*

Izvor: izradila autorica

6.4. Odabir značajki i rezultati

Odabir značajki temeljio se na svim raspoloživim tehničkim indikatorima navedenima u poglavlju 5.7.4. Sam odabir vršio se kroz nekoliko iteracija na način da su u prvome koraku sudjelovale sve varijable, a nakon njihovog rangiranja, u sljedeći bi krug ušle samo one s pozitivnim značajem i značajem većim od lažne varijable. U svakoj se iteraciji kreirala datoteka za kasniji proces učenja, te datoteka za testiranje predviđanja. Postavke koje su korištene:

- broja stabala: 250,
- broj varijabli u svakoj podjeli: $\sqrt{n}$, gdje je n ukupan broj varijabli,
- veličina lista: 1.



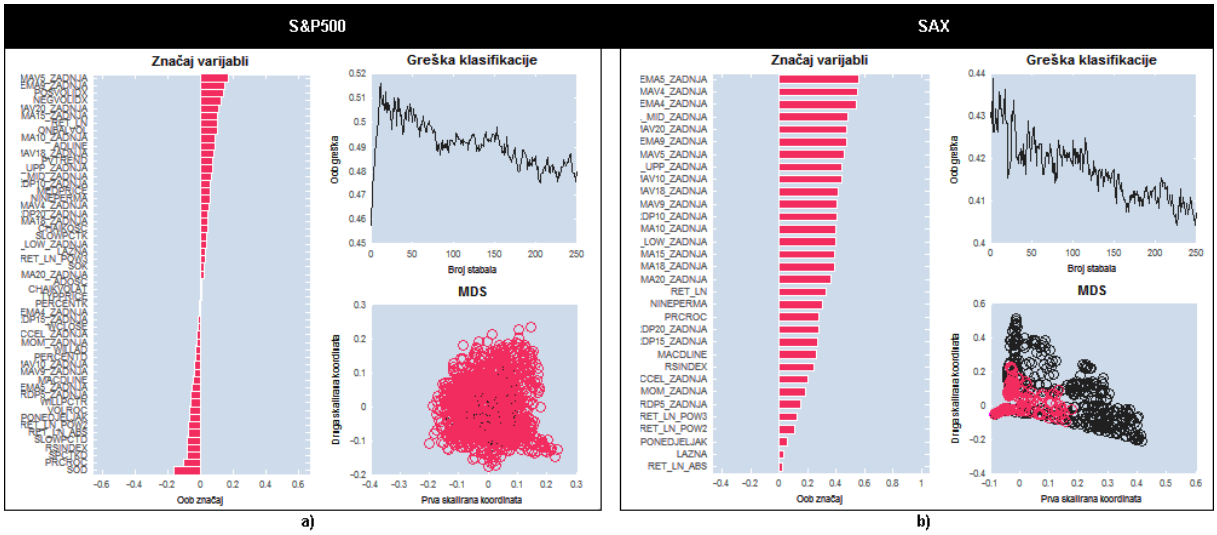
Slika 48. *Rangiranje varijabli po značaju za indeks CROBEXindustrija: a) rezultat nakon prve iteracije, b) rezultat nakon četvrte iteracije*

Izvor: izradila autorica

Slika 48. a) prikazuje rezultate rangiranja varijabli nakon prvog kruga, a slika 48. b) nakon četvrtoga za indeks CROBEXindustrija. Primjetan je pad i stabilizacija greške klasifikacije nakon izbacivanja varijabli s negativnim značajem. Međutim, ona i dalje ostaje prilično visoka.

Zanimljivo je da se prinos, čiji se predznak predviđa, uglavnom nije pokazao kao značajna varijabla, čak niti u samostalnoj usporedbi s lažnom. Vrijednosti prinosa bez ikakvih transformacija vjerojatno sadrže preveliki šum što ih čini neprikladnima za predviđanje.

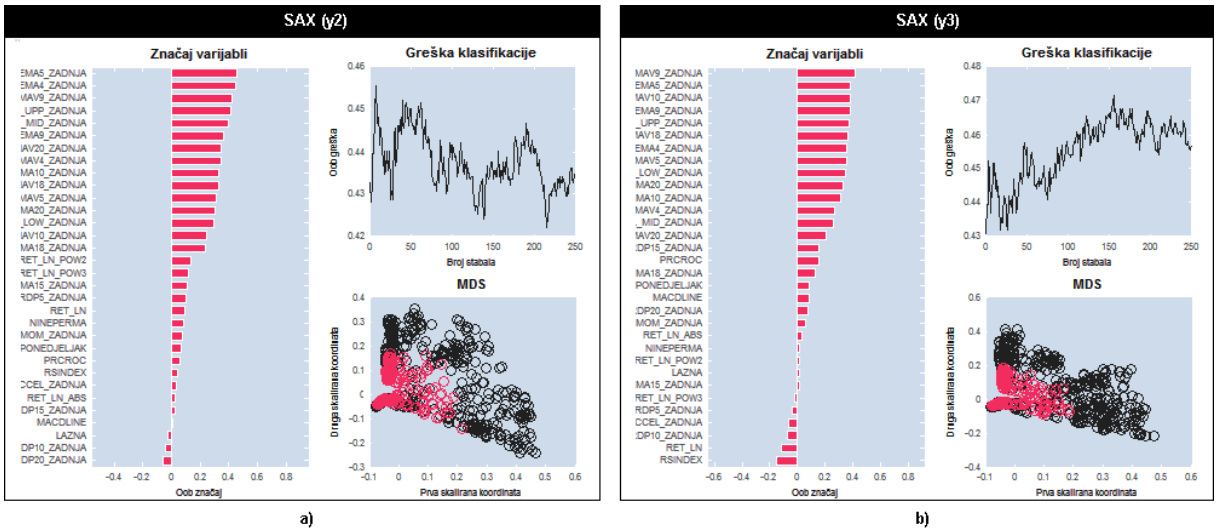
Slika 49. prikazuje rezultate rangiranja varijabli za indekse S&P500 i SAX kod predviđanja predznaka prinosa jedan dan unaprijed.



Slika 49. *Rezultat rangiranja varijabli po značaju nakon prve iteracije za predviđanje jedan dan unaprijed: a) indeks S&P500, b) indeks SAX*

Izvor:izradila autorica

Ono što je još zamjetno jest to da su samo kod indeksa SAX sve varijable već u prvom krugu odabira imale pozitivan značaj, dok se kod ostalih indeksa javlja veliki broj varijabli s negativnim značajem, što znači da se njihovim dodavanjem šteti rezultatima klasifikacije. Međutim, kod odabira varijabli za predviđanje dva ili tri dana unaprijed niti indeks SAX nije bio imun na pojavu varijabli s negativnim značajem (slika 50.), što ukazuje na povećanje poteškoća predviđanja što se dalje odmiče u budućnost.



Slika 50. *Rezultat rangiranja varijabli po značaju za indeks SAX kod predviđanja: a) dva dana unaprijed, b) tri dana unaprijed*

Izvor: izradila autorica

Najčešće visoko rangirane varijable kod indeksa SAX obuhvaćaju jednostavne i eksponencijalne pomične prosjeke i to pretežno za kraće periode (4 ili 5 dana), dok su se među najlošijima najčešće našle transformacije logaritma prinosa i varijabla "ponedjeljak".

Također, niti kod indeksa S&P500 transformacije logaritma prinosa nisu pronađene među značajnim varijablama, dok je sam logaritam prinosa bio visoko rangiran. Uz njega, među najbolje rangiranima našle su se i varijable koje stavljaju u vezu cijenu i volumen (što za indeks SAX nije bilo moguće izračunati) i to indeksi pozitivnog i negativnog volumena (eng. *Positive volume index***,** *Negative volume index*)**,** te ravnotežni volumen (eng. *On balance volume*). S obzirom na empirijski potvrđenu pozitivnu korelaciju volumena i promjene cijena (Karpoff, 1987.) te prediktivne moći neuobičajenih promjena volumena u predviđanju cijena (Sun, 2003.), to se moglo i očekivati. Kod indeksa CROBEXindustrija njima je pridodan i spori stohastički oscilator %D. Neki od inače u praksi često korištenih indikatora nisu se pokazali značajnima (npr. MACD ili RSI).

6.5. Rezultati treniranja i testiranja

Treniranje i testiranje provodilo se u dva dijela. U prvome je naglasak stavljen na ispitivanje utjecaja različitih kombinacija značajki na rezultat klasifikacije, pri čemu se pretraživanje parametara modela vršilo u širokom rasponu mogućih vrijednosti, dok se u drugome dijelu ispitivao utjecaj parametara modela na rezultat klasifikacije za kombinaciju

značajki koja se u prvome dijelu pokazala najboljom.

6.5.1. Rezultati odabira modela i treniranja za različite kombinacije značajki

Prije treniranja klasifikatora potrebno je odabrati kernel funkciju te optimalne parametre. S obzirom na rezultate i preporuke poznate iz literature (Kim (2003.) za problem klasifikacije ili Tay i Cao (2001.) za problem regresije), te općenito samih autora LibSVM biblioteke (Chang i Lin, 2001.) koji sugeriraju da, ako se uz odabir RBF kernela provodi i sustavan odabir modela, linearni kernel nije niti potrebno uzimati u obzir, dok polinomijalni, pogotovo onaj visokoga stupnja, može prouzročiti numeričke poteškoće. Sigmoidalni kernel također može biti problematičan budući da ne zadovoljava Mercerov uvjet[59], te za određene podatke problem kvadratnog programiranja možda neće imati rješenje. Ipak, Schoelkopf i Smola (2002.) smatraju da unatoč tome može dati dobre rezultate, ali uz određeni oprez. Međutim, budući da RBF kernel, uz dovoljno malu vrijednost σ^2 odnosno dovoljno velik γ ($\gamma = -1/2\sigma^2$), može ispravno klasificirati proizvoljno veliki broj primjera (Burges, 1998.) (iako uz rizik *overfittinga)*, ali i iz razloga što u preliminarnim testiranjima polinomijalni i sigmoidalni kernel nisu davali bolje rezultate, za provođenje eksperimenta odabran je RBF kernel.

Odabir modela, odnosno parametra γ i konstante C, vršio se *grid-serach* pretraživanjem (u nastavku GS), dok je za metodu evaluacije odabrana 5-struka unakrsna validacija koja kao evaluacijsku mjeru koristi točnost. Unatoč određenom nedostatku u primjeni s vremenskim nizovima[60], odabrana je iz razloga što se pokazala bržom od ostalih implementiranih metoda, a uz to je i statistički provjerena, dok su ostale samo prijedlozi koje bi trebalo detaljnije ispitati i utvrditi njihovu pouzdanost. S obzirom da je za testiranje izdvojen poseban skup podataka, važna je samo kao vodič pri odabiru parametara. Osim toga, kod problema ovakvoga tipa, jedina se prava ocjena dobiva stavljanjem sustava u rad odnosno njegovim testiranjem pri pokušaju da se ostvari zarada.

Za svaki od indeksa trenirano je više modela i to na temelju prethodno provedenoga odabira varijabli i različitih kombinacija parametara. U nastavku se daje pregled rezultata pretraživanja parametara i treniranja.

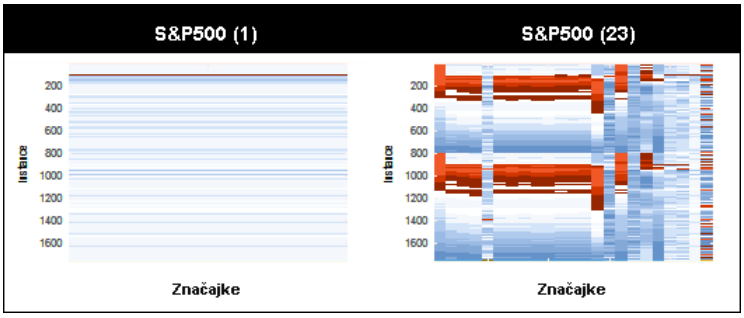
59 Vidi poglavlje 3.9.2.
60 Vidi poglavlje 5.8.2.

S&P500

U podacima za treniranje postoji manja neravnoteža u korist pozitivne klase (55,07%) zbog čega je isprobana verzija pretraživanja sa i bez kompenzacije neravnoteže**.** Trenirani su modeli s dvije kombinacije varijabli:

- samo logaritam prinosa,
- prema odabiru RF algoritma.

Slika 51. grafički prikazuje vrijednosti varijabli i to u gornjem dijelu negativne klase (koja u ovom slučaju označava pad prinosa ili situaciju bez promjene) a u donjem dijelu pozitivne klase (koja označava rast prinosa). Može se primijetiti da veći dio varijabli ima približno iste vrijednosti što čini upitan njihov doprinos razdvajanju klasa.



Slika 51. *Podaci indeksa S&P500 namijenjeni treniranju: podaci s jednom (lijevo) i podaci s 23 značajke (desno)*

Izvor: izradila autorica

Tablica 8. *Rezultati treniranja za indeks S&P500*

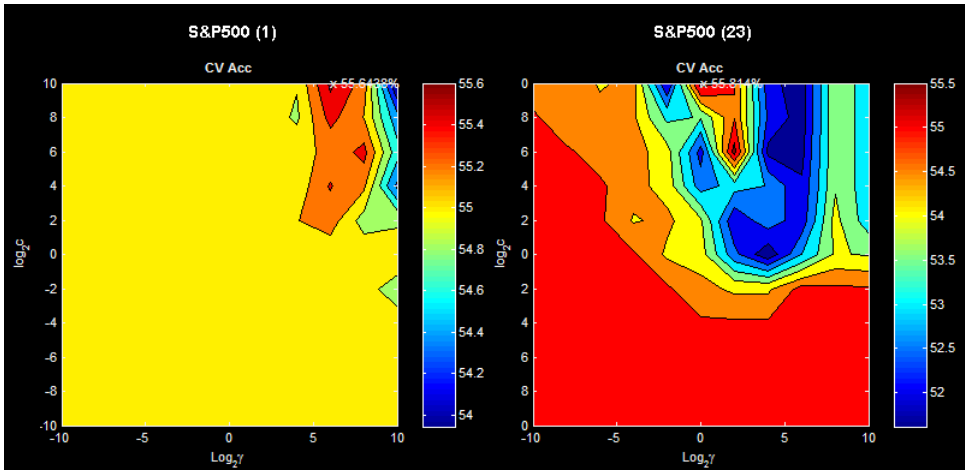
Broj značajki	Broj primjera	Pozitivni/ negativni	GS	C	g	Točnost	Broj potpornih vektora	Broj ograničenih potpornih vektora
1	1763	971/792	Da	1024	64	55,64	1582	1546
			proizvoljno	1024	1	55,19	1587	1575
23	1763	971/792	Da	1024	1	55,81	1404	982
			proizvoljno	64	4	55,70	1399	857
			proizvoljno	64	1	52,30	1539	1343
			Da + w	1024	4	51,13	1161	201

Proizvoljno – parametri odabrani bez GS, w – korištena je kompenzacije neravnoteže.

Izvor: izradila autorica

U tablici 8. može se primijetiti smanjenje točnosti u slučaju korištenja kompenzacije neravnoteže. No, to još ipak ne mora značiti da će model koji ima veću točnost prilikom treniranja ujedno postići i bolji rezultat na testiranju, pogotovo uzevši u obzir da postignuta

najveća točnost samo marginalno prelazi rezultat koji bi se dobio trivijalnim klasifikatorom odnosno takvim koji bi uvijek predviđao samo brojniju klasu.

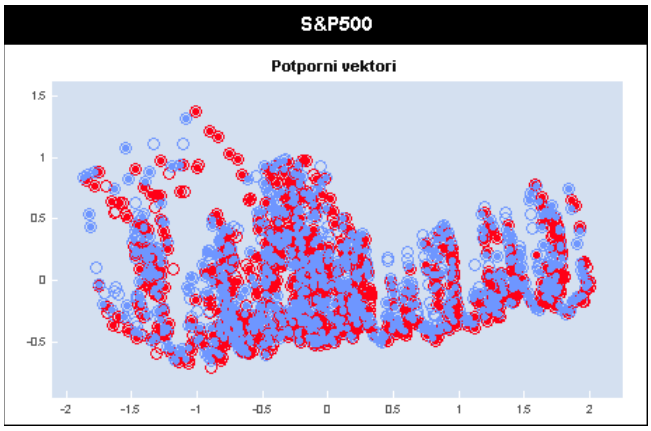


Slika 52. *Postignuta točnost za različite kombinacije parametara: podaci s jednom i 23 značajke*

Izvor: izradila autorica

Slika 52. prikazuje prostor kombinacija parametara i rezultirajuću vrijednost točnosti u slučaju podataka s jednom ili 23 značajke. Može se primijetiti da područje najveće točnosti zauzima samo malu površinu koja se nalazi u neposrednoj blizini područja najmanje točnosti, ali i prisutnost velike površine koja predstavlja područje jednake, ali samo nešto manje, točnosti za različite kombinacije parametara, što ukazuje na nestabilnost rješenja. Zbog toga je isprobana i proizvoljno odabrana kombinacija parametara.

Na slici 53. mogu se vidjeti podaci reprezentirani s 23 značajke u prikazu smanjene dimenzionalnosti. Ispunjeni kružići predstavljaju potporne vektore. Vidljivo je izrazito preklapanje klasa što čini podatke teško razdvojivima.



Slika 53. *Podaci indeksa S&P500 reprezentirani s 23 značajke u prikazu smanjene dimenzionalnosti: ispunjeni kružići označavaju potporne vektore*

Izvor: izradila autorica

CROBEXindustrija

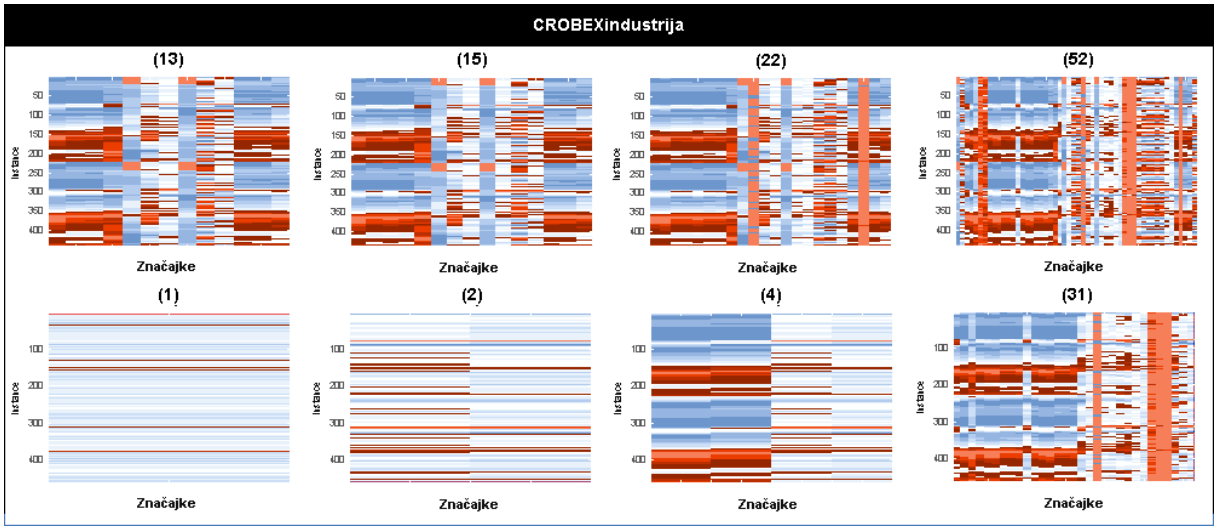
S obzirom da je ovaj niz kraći, te za njega čitav postupak znatno manje traje, provedeno je opsežnije ispitivanje. GS se provodio u tri koraka na način da se, nakon pronalaska optimalnih parametara u prvome, u sljedećem koraku pretraživanje vršilo u njihovoj okolini. Ispitano je više kombinacija varijabli:

- nekoliko kombinacija proizašlih iz odabira na temelju svih kreiranih indikatora,
- samo logaritam prinosa,
- nekoliko kombinacija proizašlih iz odabira među indikatorima koji se izračunavaju samo na temelju zaključne cijene.

Time se nastojalo utvrditi doprinose li dodatne informacije o ostalim cijenama i volumenu boljim rezultatima predviđanja.

S obzirom da su podaci uravnoteženi (50,23% u korist negativne klase), nije bilo potrebe za kompenzacijom neravnoteže.

Na slici 54. može se vidjeti da i u ovom slučaju neke značajke imaju jako slične vrijednosti, ali ipak pokazuju veću varijabilnost nego kod indeksa S&P500 (uspoređujući vrijednosti kombinacija od 22 i 23 značajke) što ukazuje da je moguće očekivati i njihov veći doprinos razdvajanju klasa.



Slika 54. *Podaci indeksa CROBEXindustrija namijenjeni treniranju: podaci s 13, 15, 22, 52 značajke bazirane na svim cijenama i volumenu (gornji red), podaci s 1, 2, 4, 31 značajkom baziranom samo na zaključnoj cijeni (donji red)*

Izvor: izradila autorica

Tablica 9. *Rezultati treniranja za indeks CROBEXindustrija*

Broj značajki	Broj primjera	Pozitivni/ negativni	GS	C	g	Točnost	Broj potpornih vektora	Broj ograničenih potpornih vektora
52	442	220/222	1.krug	1024	0.0625	52.94	336	190
			3.krug	4096	0.088388	55,43	298	60
22	442	220/222	1.krug	256	1	53,39	306	87
			3.krug	1024	0.70711	55,20	299	67
15	442	220/222	1.krug	256	4	56,79	292	49
			3.krug	2048	1,4142	60.18	272	67
13	442	220/222	1.krug	1024	1	57,01	309	172
			3.krug	8192	1,4142	59,05	265	37
1	460	232/228	1.krug	256	64	52,61	442	428
			3.krug	256	64	52,61	442	428
31 *	460	232/228	1.krug	4	0,0625	55,22	436	414
			3.krug	8	0,125	56,30	425	401
4 *	460	232/228	1.krug	256	1	54,13	419	400
			3.krug	256	1	54,13	419	400
2 *	460	232/228	1.krug	16	16	57,83	430	421
			3.krug	16	16	57,83	430	421

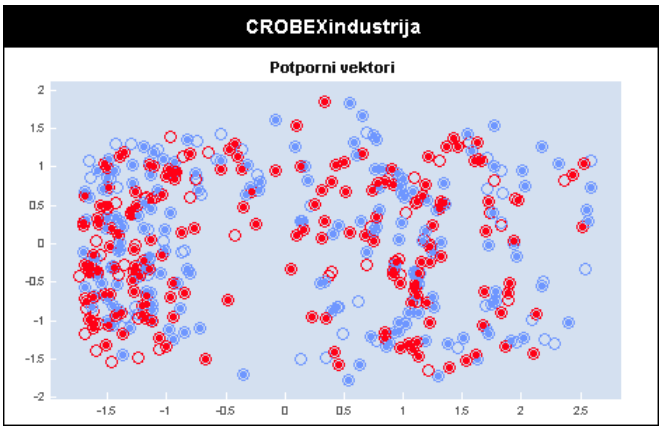
* – ulazne varijable birane samo među tehničkim indikatorima izračunatima na temelju zadnje cijene indeksa.

Izvor: izradila autorica

Iz tablice 9. vidljivo je da treći korak pretraživanja uglavnom doprinosi povećanju točnosti, ali s druge strane i naznaci da je došlo do *overfittinga* - izrazito velike vrijednosti parametra C i ponegdje γ, a što dodatno postaje vidljivo iz odnosa ukupnog broja potpornih

vektora i ograničenih potpornih vektora. Ograničeni potporni vektori su oni za koje vrijedi $\alpha_i = C$, odnosno oni koji krše marginu. Uz veći C, odnosno veću kaznu za kršenje margine, biti će i manje ograničenih potpornih vektora.

Ako se promatra utjecaj broja značajki na rezultate točnosti, može se vidjeti da je najmanja točnost dobivena korištenjem samo jedne značajke (logaritma prinosa). Također je vidljivo da se točnost ponešto smanjuje u slučaju biranja varijabli iz ograničenog skupa, ali se zato povećava broj potpornih vektora – kako njihov ukupan broj tako i broj ograničenih budući da je u tom slučaju pretraživanjem uglavnom biran manji C, ali i veći γ. Podaci očito ne sadrže dovoljno pravilnosti, odnosno ovakav odabir značajki nema dovoljnu diskriminatornu moć.



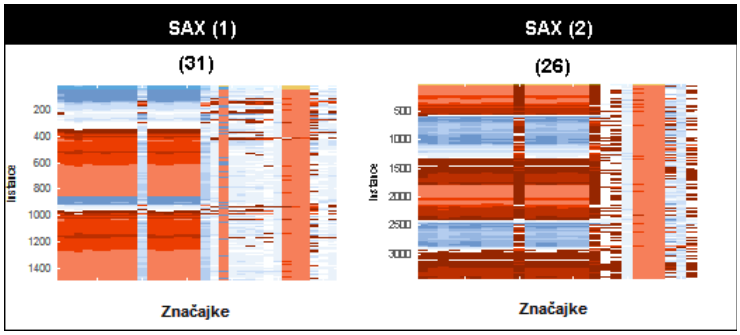
Slika 55. *Podaci indeksa CROBEXindustrija reprezentirani s 52 značajke u prikazu smanjene dimenzionalnosti: ispunjeni kružići označavaju potporne vektore*

Izvor: izradila autorica

Slika 55. prikazuje podatke s 52 varijable smanjene dimenzionalnosti. Ovdje je, iako i dalje prisutno, zamjetno mnogo manje preklapanje klasa nego kod indeksa S&P500.

SAX

Od ovog se indeksa najviše očekivalo, barem s obzirom na rezultate preliminarnih testiranja predvidljivosti. Među podacima kraćega niza prisutna je manja neravnoteža u korist negativne klase (57,32%) tako da je testirana kombinacija sa i bez kompenzacije neravnoteže. Kod dužega niza to nije bilo potrebno budući da su podaci prilično uravnoteženi (52,4 % u korist negativne klase). Trenirani su modeli za predviđanje jedan, dva i tri dana unaprijed.



Slika 56. *Podaci indeksa SAX namijenjeni treniranju: podaci kraćega niza s 31 značajkom (lijevo) i podaci dužega niza s 26 značajki (desno)*

Izvor: izradila autorica

Slika 56. prikazuje usporedbu značajki kraćega niza (s 31 značajkom) i dužega (s 26 značajki). U usporedbi s prethodnima primjetna je manja varijabilnost, ali treba i uzeti u obzir da su sve značajke dobivene samo na temelju zaključne cijene.

Tablica 10. Rezultati treniranja za indeks SAX

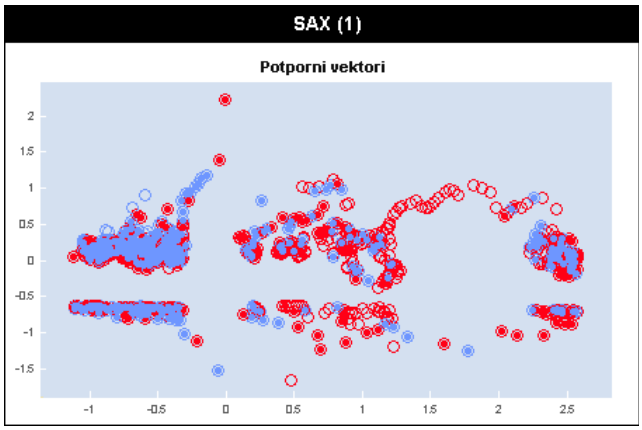
Broj značajki	Broj instanci	Pozitivni/ negativni	GS	C	g	Točnost	Broj potpornih vektora	Broj ograničenih potpornih vektora
Y1								
31	1495	638/857	Da	16	0,0625	61,67	1256	1238
30	1495	638/857	Da	16	0,0625	61,74	1235	
			Da + w	0,353	5,6569	59,93	1271	
5	1495	638/857	Da	64	256	60,60	1165	
			Da + w	0,25	1024	59,13	1301	
Y2								
31	1495	639/856	Da	0,25	4	59,33	1328	1162
29	1495	639/856	Da	0,25	4	59,26	1326	1166
26	1495	639/856	Da	1024	0,01563	59,06	1266	
Y3								
31	1495	638/857	Da	256	0,0625	59,06	1244	1181
24	1495	638/857	Da	1	1	59,00	1264	1185
17	1495	638/857	Da	256	0,125	59,80	1253	
Duži niz								
31 (y1)	3435	1635/1800	Da	256	0,25	56,36	3016	2856
26 (y2)	3435	1635/1800	Da	16	1	55,49	3070	2975
17 (y3)	3435	1635/1800	Da	64	0,25	54,99	3160	3137

Y1, Y2, Y3 – predviđanje jedan, dva ili tri dana unaprijed, w – korištena je kompenzacija neravnoteže

Izvor: izradila autorica

U tablici 10. može se uočiti da su birane manje vrijednosti konstante C, te da je kod dužeg niza postizana manja točnost. Slika 57. prikazuje prisutnost područja velikog

preklapanja klasa i područja manjeg preklapanja što ukazuje da je moguće očekivati ponešto bolji rezultat nego kod indeksa S&P500, ali i lošiji od onog postignutog kod indeksa CROBEXindustrija.



Slika 57. *Podaci indeksa SAX (kraći niz) reprezentirani s 31 značajkom u prikazu smanjene dimenzionalnosti: ispunjeni kružići označavaju potporne vektore*

Izvor: izradila autorica

Općenito se prilikom pretraživanja i treniranja može primijetiti da se točnost jedva ponešto mijenja s različitim kombinacijama varijabli ili parametara, ali i sklonost prema izrazito velikim vrijednostima konstante C, a ponekad i γ što ukazuje na *overfitting*. Keerthi i Lin (2003.) navode da se jaki *overfitting* događa kad je γ veliki jer je tada fleksibilnost klasifikatora velika. Dio sklonosti prema većim vrijednostima parametara dolazi i zbog načina implementacije jer se u slučaju jednakih vrijednosti dviju kombinacija parametara uzima ona posljednje odabrana.

Uz to se kod svih indeksa može primijetiti izrazito veliki broj potpornih vektora, čiji se ukupan broj znatno ne mijenja s promjenama vrijednosti konstante C, te se kreće oko 80% kod indeksa S&P500, nešto manje od 70% za indeks CROBEXindustrija, ali i 90% u slučaju značajki baziranih samo na zaključnoj cijeni, te preko 80% za indeks SAX. Manji C znači šire margine, što bi trebalo omogućiti i bolju sposobnost generalizacije, ali broj potpornih vektora kod nelinearnog modela ne ovisi samo o vrijednosti C nego i o kompleksnosti modela za čije je podešavanje potreban njihov veći broj. Tay i Cao (2001.) navode da gotovo svi podaci postaju potporni vektori u slučaju *overfittinga* ili *undefittinga*. Dodatno, broj potpornih vektora ovisi i o broju instanci, a povećava se i u slučaju kad u podacima nema dovoljno pravilnosti. Mana velikog broja potpornih vektora je i ta što doprinosi produljenju vremena treniranja i klasifikacije s obzirom da se svaka instanca uspoređuje sa svakim od potpornih

vektora.

Najveća postignuta točnost treniranja kreće se oko 60% za indekse CROBEXindustrija i SAX, dok je kod S&P500 ona jedva na razini omjera klasa. No bez testiranja na nepoznatom skupu ovaj podatak nema poseban značaj, iako bi već i ovakav rezultat bio jako poželjan.

6.5.2. Rezultati testiranja

Testiranje se provodilo na nekoliko načina:

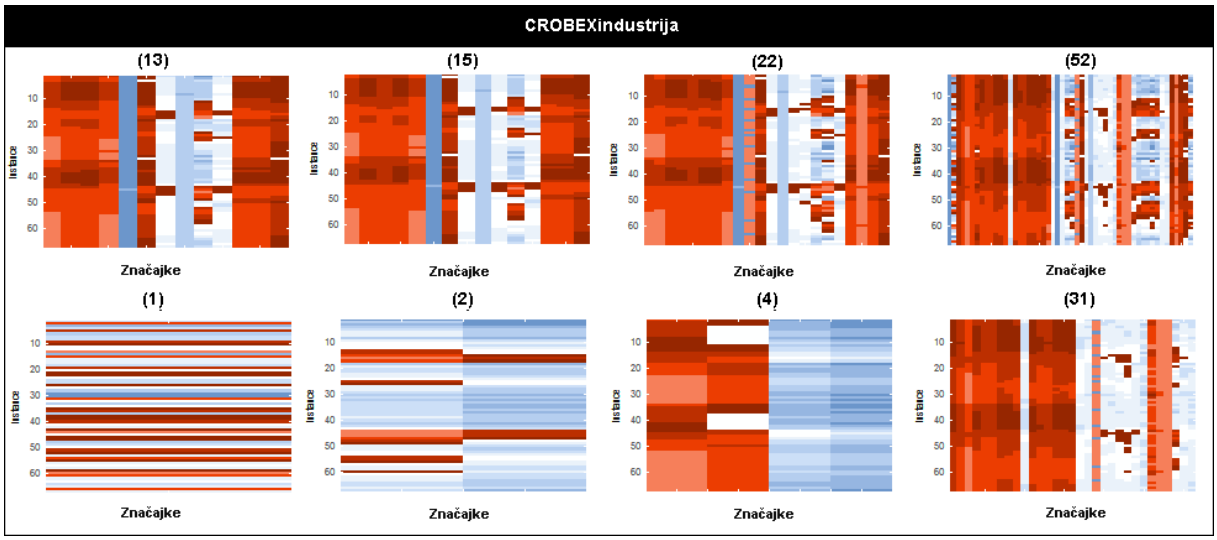
1. s odgovarajućim stvarnim podacima
2. sa slučajno generiranim podacima koji slijede Gaussovu distribuciju uz zadržavanje stvarnih oznaka klasa
3. s permutiranim oznakama klasa uz zadržavanje originalnih podataka čime je trebalo utvrditi je li klasifikator doista pronašao stvarnu vezu između podataka i oznaka klasa.

U nastavku se prikazuje dio karakterističnih razultata.

CROBEXindustrija

Kod ovog su indeksa podaci uravnoteženi i u skupu za testiranje (50,75% u korist pozitivnih primjera), te su prikazani na slici 58. Rezultati testiranja prema odabranim mjerama prikazani su u tablici 11, a radi veće preglednosti i u obliku konfuzijske matrice na slici 59.

Može se primijetiti da modeli trenirani s parametrima odabranima u trećem krugu GS postižu manju točnost prilikom testiranja od modela treniranih s parametrima odabranima u prvome krugu. S obzirom da je prilikom treniranja to bio obrnut slučaj, potvrđena je prethodna sumnja na *overfitting* i štetnost pretjerane optimizacije parametara.



Slika 58. *Podaci indeksa CROBEXindustrija namijenjeni testiranju: podaci s 13, 15, 22, 52 značajke bazirane na svim cijenama i volumenu (gornji red), podaci s 1, 2, 4, 31 značajkom baziranom samo na zaključnoj cijeni (donji red)*

Izvor: izradila autorica

Tablica 11. *Rezultati testiranja za indeks CROBEXindustrija*

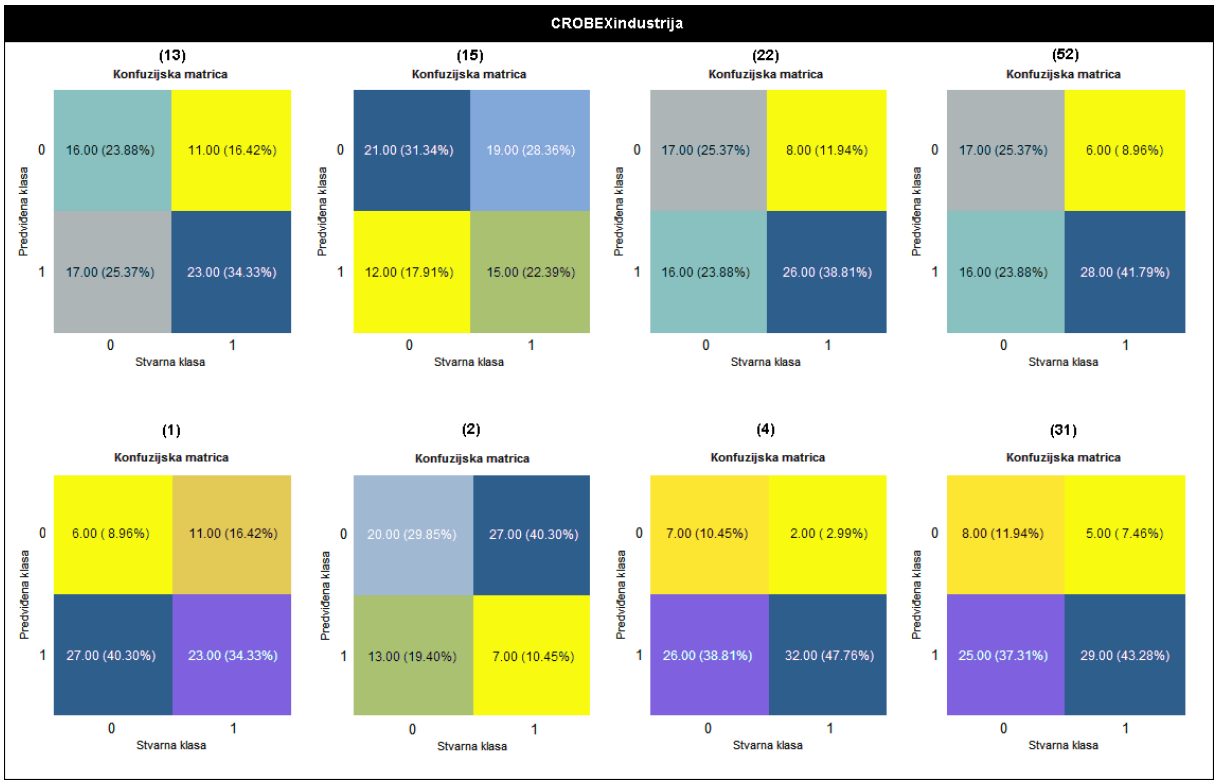
Broj značajki	GS	Točnost	AUC	TPR	TNR	FPR	FNR	PPV	NPV	G1	G2	F
52	1.krug	67,16	69,70	82,35	51,52	48,48	17,65	63,64	73,91	65,13	72,39	71,79
22	1.krug	64,18	59,02	76,47	51,52	48,48	23,53	61,90	68,00	62,76	68,80	68,42
	3.krug	58,21	55,30	64,71	51,52	48,48	35,29	57,89	58,62	57,74	61,21	61,11
15	1.krug	53,73	61,23	44,12	63,64	33,36	55,88	55,56	52,50	52,99	49,51	49,18
	3.krug	49,25	54,19	26,47	72,73	27,27	73,53	50,00	48,98	43,88	36,38	34,62
13	1.krug	58,21	57,31	67,65	48,48	51,52	32,35	57,50	59,26	57,27	62,37	62,16
	3.krug	53,73	56,33	44,12	63,64	36,36	55,88	55,56	52,50	52,99	49,51	49,18
1	1.krug	43,28	36,54	67,65	18,18	81,82	32,35	46,00	35,29	35,07	55,78	54,76
31*	1.krug	55,22	51,25	85,29	24,24	75,76	14,71	53,70	61,54	45,47	67,68	65,91
	3.krug	55,22	58,11	85,29	24,24	75,76	14,71	53,70	61,54	45,47	67,68	65,91
4*	1.krug	58,21	55,70	94,12	21,21	78,79	5,88	55,17	77,78	44,68	72,06	69,57
2*	1.krug	40,30	46,88	20,59	60,61	39,39	79,41	35,00	42,55	35,32	26,84	25,00

* – ulazne varijable birane samo među tehničkim indikatorima izračunatima na temelju zadnje cijene indeksa.

Izvor: izradila autorica

Međutim, kao što je ranije navedeno, sama točnost nije dovoljna mjera za ocjenu rezultata. Ovisno o preferencijama i sklonosti prema riziku, ali i općem trendu na tržištu, mijenjati će se i zainteresiranost za točnost predviđanja pozitivnih ili negativnih primjera. U vrijeme općeg pozitivnog tržišnog trenda postojat će veća zainteresiranost za izbjegavanje onih rjeđih situacija koje bi vodile gubitku s obzirom da u takvim uvjetima već i pasivna

strategija (*buy & hold*[61]) donosi dobit. S druge strane, u vrijeme negativnog trenda zainteresiranost će biti usmjerena na iskorištavanje onih ipak rijetkih prilika za ostvarenje profita budući da se u takvoj situaciji od gubitka najbolje čuva suzdržavanjem od ulaganja. Naravno, moguć je i drugačiji tijek razmišljanja i preferencija ulaganja stoga je potrebno proanalizirati i ostale mjere.



Slika 59. *Konfuzijska matrica za indeks CROBEXindustrija*

Izvor: izradila autorica

Gledajući konfuzijsku matricu, ako želimo iskoristiti prilike za ulaganje, potrebo je točno predvidjeti pozitivnu klasu (TP). Ako želimo izbjeći gubitak, trebamo minimizirati pogrešno predviđanje negativnih primjera (FP). FN označava propuštene prilike za zaradu, ali ne ostvaruje se stvarni gubitak, dok TN vodi suzdržavanju od ulaganja čime se izbjegava gubitak.

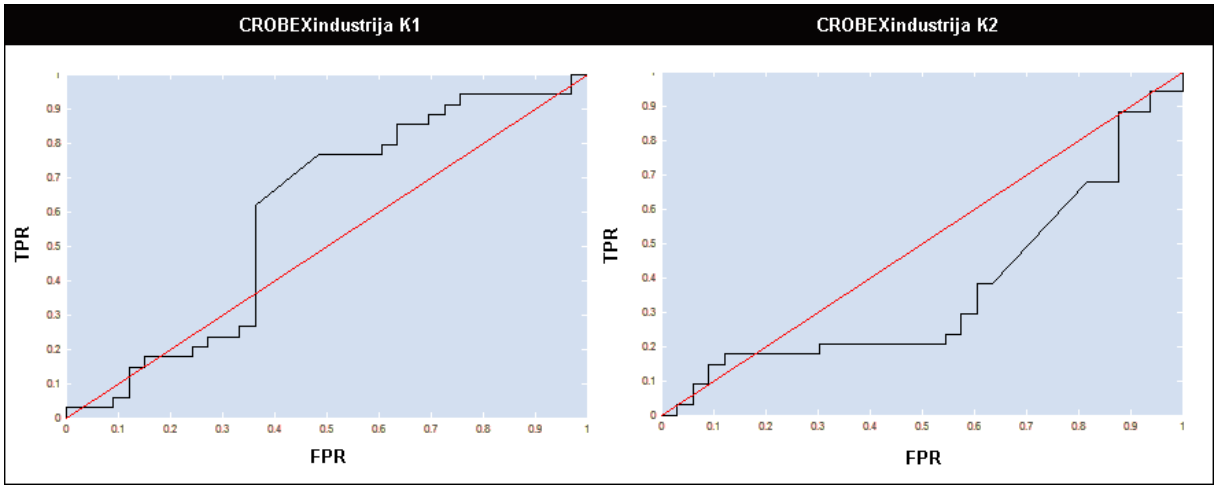
U slučaju indeksa CROBEXindustrija općenito se primjećuje postizanje boljeg rezultata na pozitivnim primjerima nego na negativnima. Ako se usporedi s kretanjem cijena (slika 45.) vidljiv je pad cijene u testnom razdoblju (o kojem prilikom treniranja ne znamo ništa),

61 *Buy & hold* ("kupi i drži") – investicijska strategija kod koje se dionica kupuje s namjerom dugoročnog držanja neovisno o fluktuacijama cijena čime se ujedno izbjegavaju transakcijski troškovi povezani sa špekulativnim kratkoročnim ulaganjima.

ali koji je započeo krajem razdoblja obuhvaćenoga podacima za treniranje. Uz pretpostavku da bi se taj trend mogao nastaviti, mjera koja bi mogla biti posebno od interesa je preciznost (PPV), a uz nju geometrijska sredina preciznosti i odziva (G2 u tablici 11.) odnosno njihova harmonijska sredina (F-mjera).

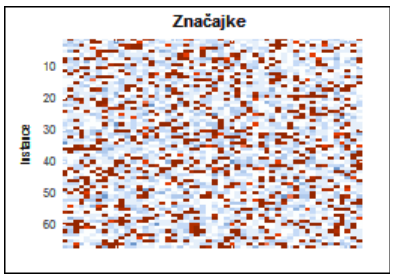
Zanimljivo je da se klasifikator izgrađen korištenjem svih ponuđenih varijabli pokazao najboljim i to prema gotovo svim mjerama, što ukazuje na doprinos dodatnih informacija (ostalih cijena, osim zaključne, i volumena) uspješnosti klasifikatora.

Slika 60. prikazuje usporedni prikaz ROC krivulja klasifikatora (K1) koji je po rezultatu odmah iza najboljega (drugi u tablici 11.) i onog izgrađenog korištenjem samo jedne varijable – logaritma prinosa (K2). Obje krivulje odstupaju od slučajnog klasifikatora i to najviše u srednjem dijelu, s tim da izgledaju gotovo kao zrcalna slika. K2 daje rezultat koji je ispod razine slučaja, ali biti će zanimljiva njegova provjera prilikom simulacije s obrnutim rezultatima predviđanja – hoće li možda tada biti na razini najboljega.



Slika 60. *Usporedba ROC krivulja klasifikatora temeljenog na podacima s 13 značajki (prikaz lijevo) i onog izgrađenog samo na temelju podataka o prinosu (prikaz desno) za indeks CROBEXindustrija*

Izvor: izradila autorica



Slika 61. *Slučajno generirani podaci*

Izvor: izradila autorica

Slika 61. prikazuje podatke sa slučajno generiranim vrijednostima, dok tablica 12. prikazuje usporedne rezultate pokušaja predviđanja najboljeg klasifikatora (prvi u tablici 11.) i to redom na originalnim podacima, na podacima s permutiranim oznakama klasa i na podacima sa slučajno generiranim vrijednostima .

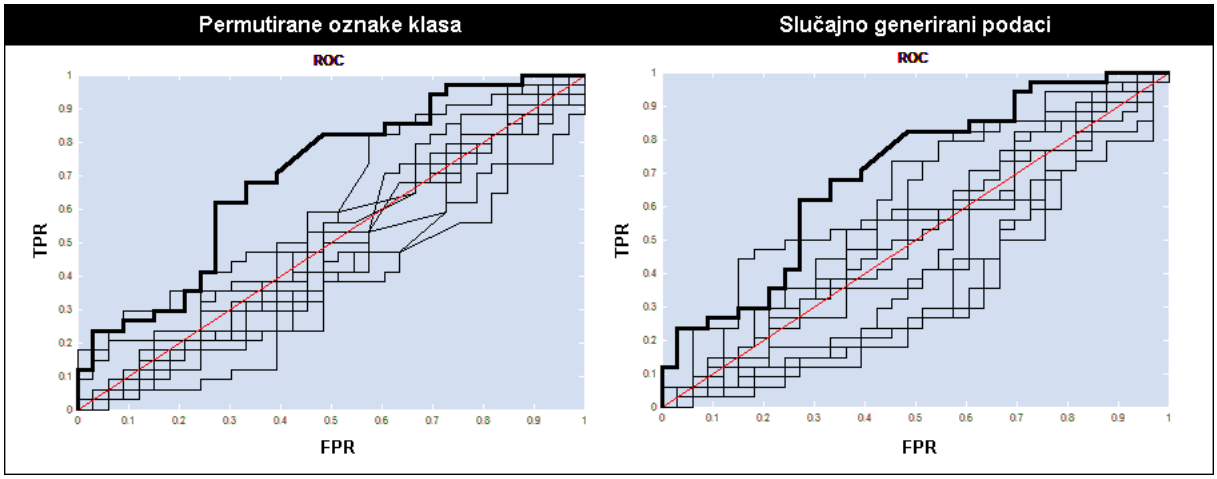
Tablica 12. *Rezultati testiranja najboljeg klasifikatora na različitim vrstama podataka*

	Točnost	AUC	TPR	TNR	FPR	FNR	PPV	NPV	G1	G2	F
Originalni podaci	67,16	69,70	82,35	51,52	48,48	17,65	63,64	73,91	65,13	72,39	71,79
Permutirane oznake klasa*	48,26	48,03	63,22	32,42	67,58	36,18	49,32	46,52	45,45	56,11	55,64
Slučajno generirani podaci*	49,25	42,60	0,00	100,00	0,00	100,00	0,00	49,25	0,00	0,00	0,00

* - prosječne vrijednosti za 10 ponavljanja

Izvor: izradila autorica

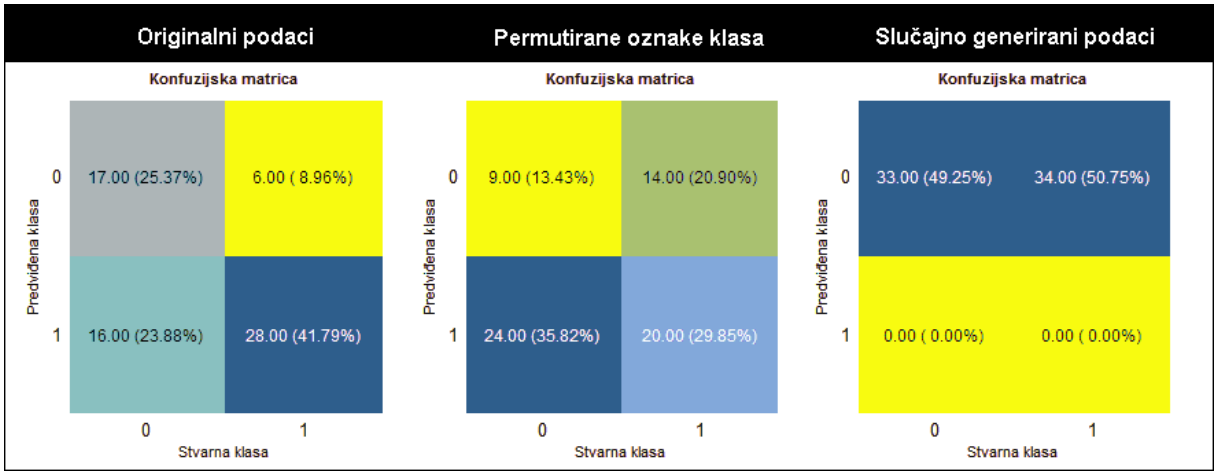
Rezultat predviđanja na slučajno generiranim podacima u svim je pokušajima bio ujednačen – klasifikator je uvijek predviđao istu klasu. Jedina varijacija vidljiva je u prikazu ROC krivulja (slika 62.), odnosno proizlazi iz toga koliko je klasifikator bio "uvjeren" u pripadnost određene instance jednoj od klasa. Radi lakše usporedbe, ROC krivulja dobivena s originalnim podacima na slici je podebljana. Može se vidjeti da se ona, od svih prikazanih, ipak najviše udaljava od slučajnog klasifikatora.



Slika 62. *ROC krivulje za indeks CROBEXindustrija: rezultat predviđanja na podacima s permutiranim oznakama klasa (lijevo) i slučajno generiranim podacima (desno). ROC krivulja originalnih podataka je podebljana.*

Izvor. izradila autorica

Također je i kod podataka s permutiranim oznakama klasa primjetna veća sklonost jednoj klasi, jasno vidljive iz konfuzijske matrice (slika 63.), ali ipak znatno manje izražene (u tablici 12. prikazane su srednje vrijednosti za 10 ponavljanja). Međutim, ukupan je rezultat prema svim mjerama lošiji od rezultata dobivenih na originalnim podacima. Moglo bi se reći da je u ovom slučaju doista došlo do učenja i da je klasifikator uspio prepoznati uzorke u podacima. Preostaje još njegova provjera na simulatoru trgovanja.



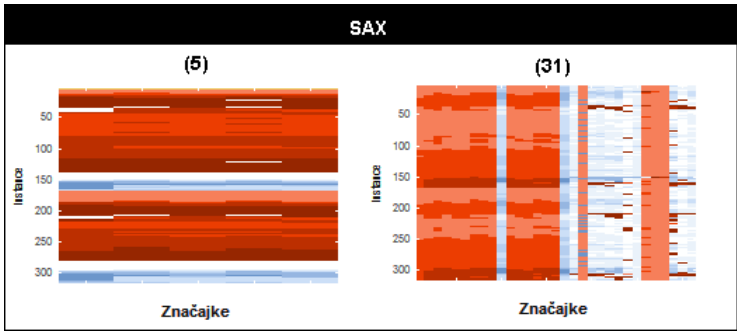
Slika 63. *Konfuzijska matrica za indeks CROBEXindustrija: originalni podaci, podaci s permutiranim oznakama klasa, slučajno generirani podaci.*

Izvor: izradila autorica

SAX

Prikazuju se rezultati testiranja klasifikatora izgrađenih za predviđanje predznaka promjene prinosa za jedan, dva i tri dana unaprijed. U podacima postoji manja neravnoteža u korist negativne klase i to redom: 53,16% (jedan dan unaprijed). 53,01% (dva dana unaprijed), 52,87% (tri dana unaprijed).

Rezultati testiranja dani su u tablici 13., a dio rezultata prikazan je i u obliku konfuzijske matrice (slika 65.). Slika 64. prikazuje podatke za testiranje predviđanja jedan dan unaprijed s pet i 31 značajkom.



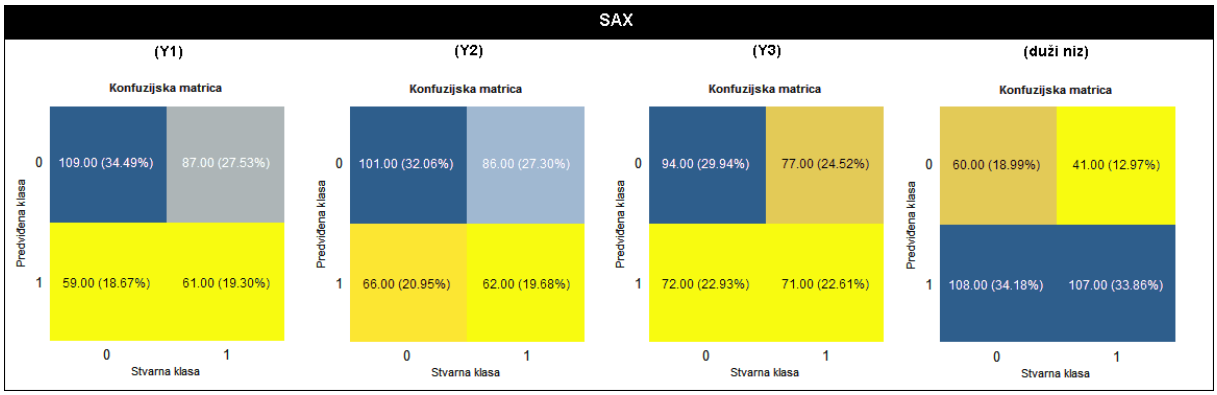
Slika 64. *Podaci indeksa SAX namijenjeni testiranju predviđanja promjene predznaka prinosa jedan dan unaprijed: podaci s pet i 31 značajkom*

Izvor: izradila autorica

Tablica 13. *Rezultati testiranja za indeks SAX*

Broj značajki	Točnost	AUC	TPR	TNR	FPR	FNR	PPV	NPV	G1	G2	F
31	53,80	51,70	41,22	64,88	35,11	58,78	50,83	55,61	51,71	45,77	45,52
30	52,22	52,07	40,54	62,50	37,50	59,46	48,78	54,40	50,34	44,47	44,28
5	53,48	56,17	27,70	76,19	23,81	72,30	50,62	54,47	45,94	37,45	35,81
y2											
31	51,75	47,43	41,89	60,48	39,52	58,11	48,44	54,01	50,33	45,05	44,93
29	52,06	48,11	43,24	59,98	40,12	56,76	48,85	54,35	50,89	45,96	45,88
26	48,89	50,75	66,89	32,93	67,07	33,11	46,92	52,88	46,94	56,02	55,15
y3											
31	52,55	50,18	47,97	56,63	43,37	52,03	49,65	54,97	52,12	48,80	48,80
24	51,91	52,97	59,46	45,18	54,82	40,54	49,16	55,56	51,83	54,07	53,82
17	49,68	46,28	38,51	59,64	40,36	61,49	45,97	52,11	47,93	42,08	41,91
Duži niz											
31 (y1)	52,85	56,31	72,30	35,71	64,29	27,70	49,77	59,41	50,81	59,98	58,95
26 (y2)	47,30	51,94	77,03	20,96	79,04	22,97	46,34	50,72	40,18	59,75	57,87
17 (y3)	47,13	48,39	100,00	0,00	100,00	0,00	47,13	0,00	0,00	68,65	64,07

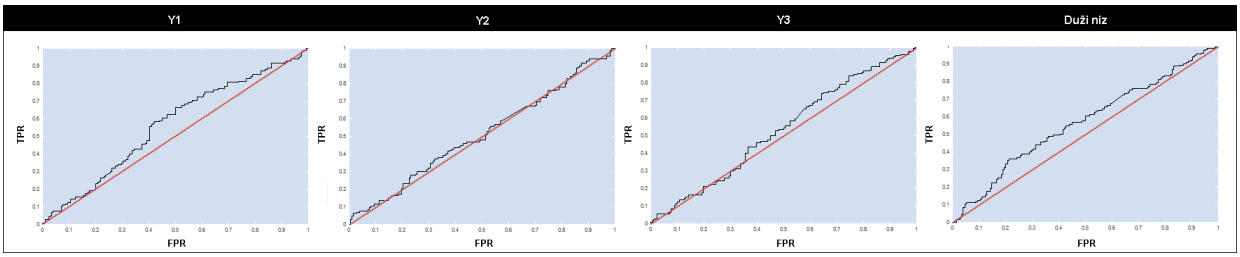
Izvor: izradila autorica



Slika 65. *Konfuzijska matrica za indeks SAX*

Izvor: izradila autorica

U ovom slučaju situacija je prilično nejasna i teško je razlučiti koji je klasifikator bolji, ako uopće ikoji. Samo u slučaju predviđanja jedan dan unaprijed postiže se točnost na razini omjera klasa. Općenito se može primijetiti postizanje ponešto boljeg rezultata u predviđanju negativne klase od one pozitivne što je, zapravo, uzrokovano neravnotežom u podacima. Međutim, uspoređujući ROC krivulje (slika 66.) i mjeru AUC (tablica 13.) , neosjetljivih na neravnoteže u podacima, jasno je vidljivo da su klasifikatori jako blizu razine slučaja.



Slika 66. *ROC krivulje indeksa SAX za najveći postignuti AUC: u predviđanju jedan, dva, tri dana unaprijed te dužeg niza u predviđanju promjene predznaka jedan dan unaprijed*

Izvor: izradila autorica

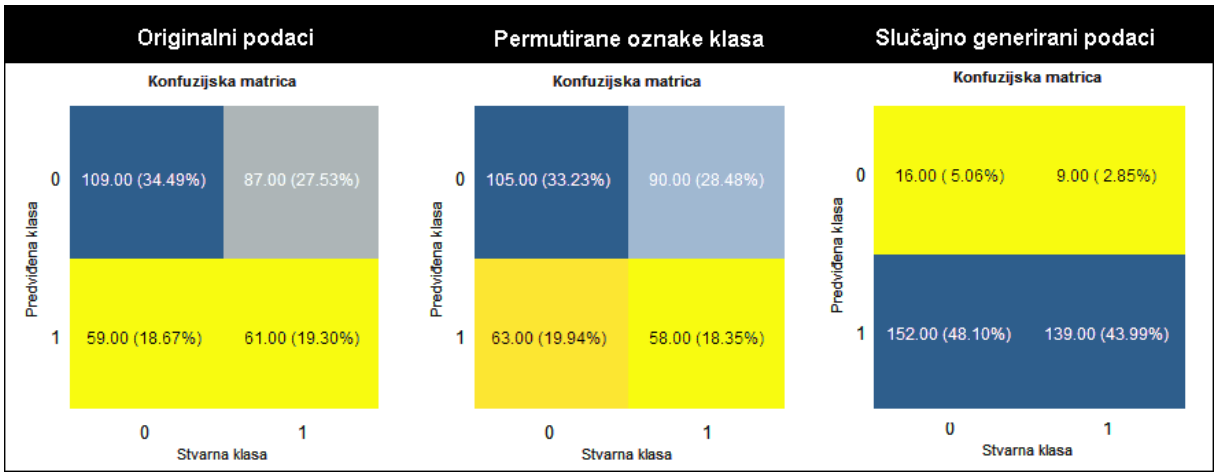
Za provjeru s permutiranim i slučajnim vrijednostima kraćega niza odabran je prvi klasifikator iz tablice 13. koji je, unatoč manjem AUC, ipak postigao ponešto bolje vrijednosti nekoliko drugih mjera u odnosu na preostale iz grupe. Rezultati su prikazani u tablici 14.

Tablica 14. *Rezultat predviđanja na različitim vrstama podataka za indeks SAX (kraći niz)*

	Točnost	AUC	TPR	TNR	FPR	FNR	PPV	NPV	G1	G2	F
Originalni podaci	53,80	51,70	41,22	64,88	35,11	58,78	50,83	55,61	51,71	45,77	45,52
Permutirane oznake klasa*	50,53	50,20	39,19	62,50	37,50	60,81	47,93	53,85	49,49	43,34	43,12
Slučajno generirani podaci*	48,62	49,42	93,11	8,87	91,13	6,89	47,38	58,95	28,56	66,41	62,80

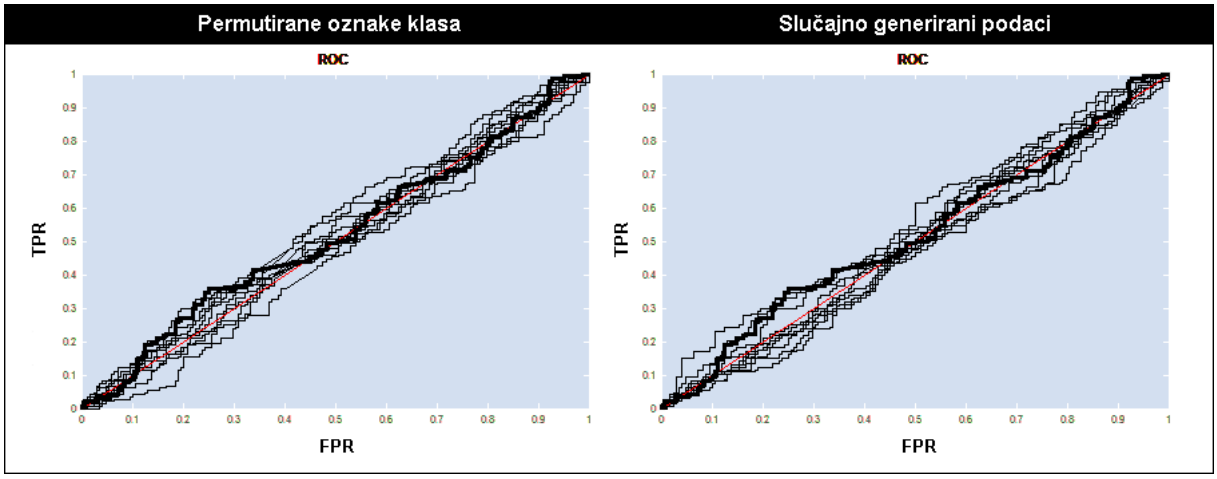
* - prosječne vrijednosti za 10 ponavljanja

Izvor: izradila autorica



Slika 67. *Konfuzijska matrica za indeks SAX (kraći niz): originalni podaci, podaci s permutiranim oznakama klasa, slučajno generirani podaci.*

Izvor: izradila autorica



Slika 68. *ROC krivulje za indeks SAX (kraći niz): rezultat predviđanja na podacima s permutiranim oznakama klasa (lijevo) i slučajno generiranim podacima (desno). ROC krivulja originalnih podataka je podebljana.*

Izvor: izradila autorica

I kod ovog je indeksa rezultat predviđanja na slučajno generiranim podacima gotovo ujednačen – uz minimalnu varijaciju predviđa uvijek istu klasu što je jasno vidljivo iz konfuzijske matrice (slika 67.). Međutim, za razliku od indeksa CROBEXindustrija, kod predviđanja podataka s permutiranim ozanakama rezultat se gotovo ne razlikuje od rezultata predviđanja na originalnim podacima. Ako se pogleda usporedni prikaz ROC krivulja (slika 68.) jasno je da se klasifikator ne razlikuje od slučajnoga. Može se reći da u ovom slučaju nije došlo do učenja, odnosno klasifikator ne prepoznaje vezu između podataka i oznaka klasa.

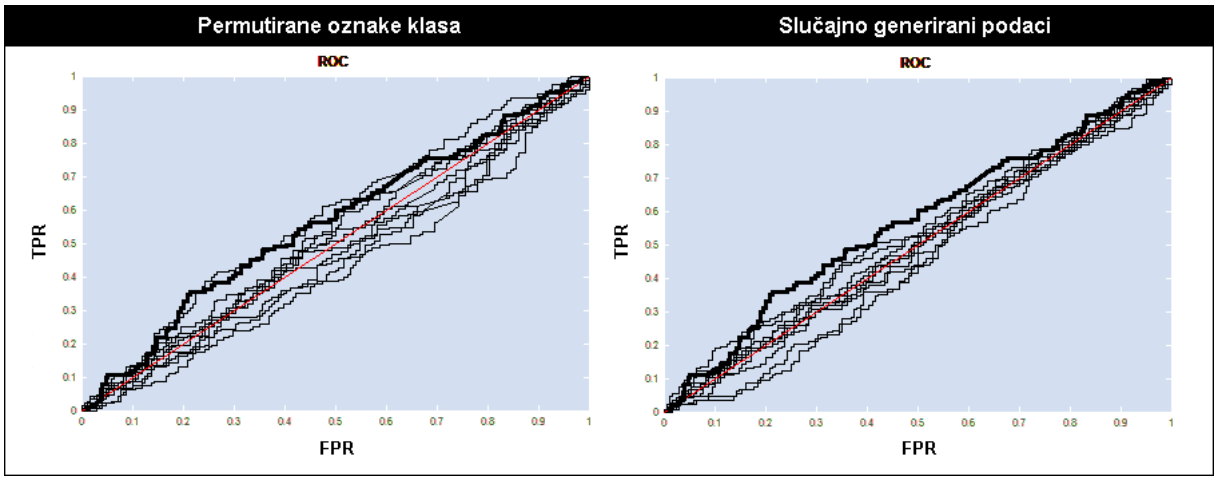
Tablica 15. *Rezultat predviđanja na različitim vrstama podataka za indeks SAX (duži niz)*

	Točnost	AUC	TPR	TNR	FPR	FNR	PPV	NPV	G1	G2	F
Originalni podaci	52,85	56,31	72,30	35,71	64,29	27,70	49,77	59,41	50,81	59,98	58,95
Permutirane oznake klasa*	47,15	49,18	66,22	30,36	69,64	33,79	45,58	50,50	44,82	54,88	53,99
Slučajno generirani podaci*	53,16	49,43	0,00	100,00	0,00	100,00	0,00	53,16	0,00	0,00	0,00

* - prosječne vrijednosti za 10 ponavljanja

Izvor: izradila autorica

Sličan je rezultat, iako ponešto bolji, dobiven klasifikatorom izgrađenog korištenjem duljeg niza podataka (tablica 15.). Iako su rezultati na originalnim podacima tek ponešto bolji od onih na podacima s permutiranim oznakama klasa, na slici 69. je vidljivo da se ROC krivulje permutiranih i slučajnih podataka ipak uglavnom nalaze ispod krivulje originalnih podataka.



Slika 69. *ROC krivulje za indeks SAX (duži niz): rezultat predviđanja na podacima s permutiranim oznakama klasa (lijevo) i slučajno generiranim podacima (desno). ROC krivulja originalnih podataka je podebljana.*

Izvor: izradila autorica

S&P500

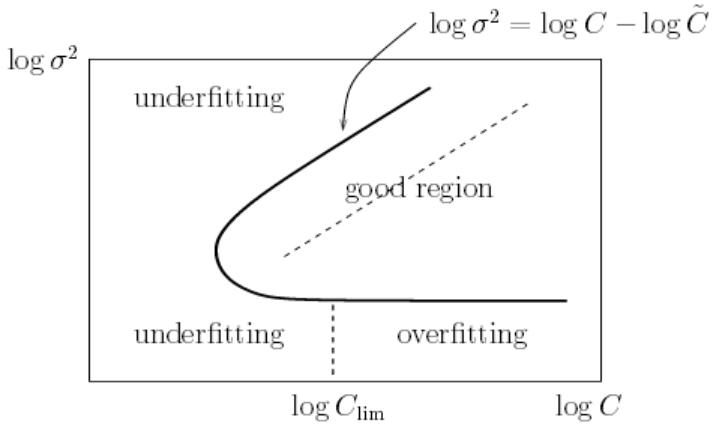
Rezultati testiranja ovog indeksa ne prikazuju se jer niti jedan od izgrađenih klasifikatora nije pokazao uspješnost u predviđanju promjene predznaka prinosa jedan dan unaprijed, a pogotovo ne dva ili tri dana unaprijed. Klasifikatori su uvijek predviđali istu klasu – onu brojniju. S obzirom na umjerenu neravnotežu u podacima, isprobana je kompenzacija neravnoteže, koja se, unatoč tome što je doprinijela pojavi određene varijacije, pokazala nedostatnom za poboljšanje rezultata. Keerthi i Lin (2003.) ovakvo ponašanje

povezuju s jakim *underfittingom,* odnosno slučaja nedovoljne prilagođenosti podacima zbog korištenja modela nedovoljne kompleksnosti, ali navodi da se to događa za $C \to 0$ uz fiksirani γ ili za $\gamma \to \infty$ i fiksirani C na dovoljno malu vrijednost ili $\gamma \to 0$ uz fiksirani C, što ovdje nije bio slučaj. Međutim, ako se usporedi s rezultatima testiranja predvidljivosti, može se vidjeti da se za podatke namijenjene testiranju klasifikatora nije mogla odbaciti hipoteza o slučajnom hodu.

6.5.3. Ispitivanje utjecaja različitih kombinacija parametara na rezultat

Drugi dio treniranja i testiranja provodi se kako bi se ispitao odnos kombinacija parametara i rezultata dobivenih na osnovu različitih evaluacijskih mjera što bi trebalo pridonijeti boljem razumijevanju i tumačenju rezultata dobivenih u prvome dijelu treniranja i testiranja.

Keerthi i Lin (2003.) su istražujući asimptotsko ponašanje greške generalizacije u slučaju ekstremnih vrijednosti parametara γ i C utvrdili odnose prikazane na slici 70. iz koje je vidljivo da do pojave *underfittinga* dolazi općenito u slučaju malih vrijednosti C, a do *overfittinga* u slučaju velikih vrijednosti C i malih vrijednosti σ^2, odnosno velikih vrijednosti γ.



Slika 70. *Prostor underfittinga, overfittinga i optimalnih kombinacija parametara σ^2 i C*
Izvor: Keerthi i Lin (2003.)

Tay i Cao (2001.) proveli su istraživanje primjene SVM algoritma u predviđanju financijskih vremenskih nizova pri čemu su utvrdili konkretne vrijednosti (prikazane u tablici 16.) za koje se može uočiti prethodno opisano ponašanje.

Tablica 16. *Utjecaj parametara na pojavu underfittinga ili overfittinga*

	γ	C
Underfitting	< 0,005	<10
Optimalno	0,005 – 0,5	10 – 100
Overfitting	> 0,5.	>100

Izvor: izradila autorica prema (Tay i Cao, 2001.)

Ovdje se ispitivanje vršilo za kombinacije parametara sljedećih vrijednosti:

$C = \{0.1, 1,10,100,1000\}$, $\gamma = \{0.01, 0.1, 1, 10\}$.

CROBEXindustrija

Upotrijebljeni su podaci s 52 značajke i 442 instance. Rezultati treniranja prikazani su u tablici 17., a testiranja u tablici 18.

Tablica 17. *CROBEXindustrija: rezultati treniranja za različite kombinacije parametara*

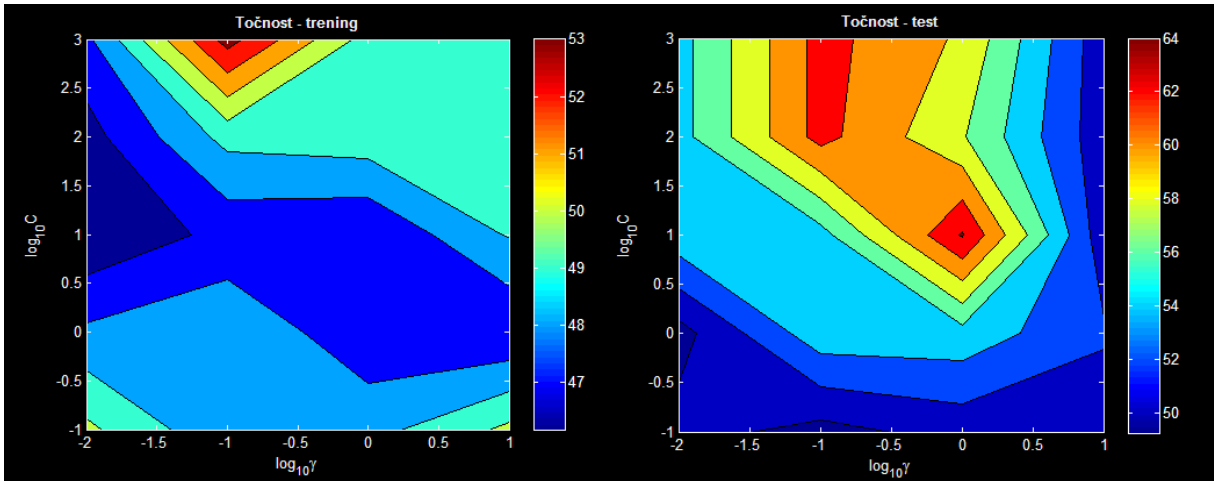
C	g	Točnost	Brroj potpornih vektora	Broj ograničenih potpornih vektora
0,1	0,01	50,2262	440	440
1	0,01	48,19	436	433
10	0,01	46,1538	422	414
100	0,01	46,6063	406	380
1000	0,01	47,7376	393	343
0,1	0,1	48,19	440	440
1	0,1	48,8326	426	411
10	0,1	47,2851	407	357
100	0,1	49,3213	376	274
1000	0,1	53.3937	322	127
0,1	1	48,6425	441	437
1	1	47,2851	418	353
10	1	47,0588	372	141
100	1	49,5475	329	7
1000	1	49,5475	327	0
0,1	10	50,2262	442	434
1	10	47,0588	441	233
10	10	49,0950	439	0
100	10	49,0950	439	0
1000	10	49,0950	439	0

Izvor. izradila autorica

Tablica 18. *CROBEXindustrija: rezultati testiranja za različite kombinacije parametara*

C	g	Točnost	AUC	TPR	TNR	FPR	FNR	PPV	NPV	G1	G2	F
0,1	0,01	50,75	33,69	100,00	0,00	100,00	0,00	50,75	0,00	0,00	71,24	67,33
1	0,01	49,25	33,24	0,00	100,00	0,00	100,00	0,00	49,25	0,00	0,00	0,00
10	0,01	55,22	65,78	85,29	24,24	75,76	14,71	53,70	61,54	45,47	67,68	65,91
100	0,01	55,22	57,22	85,29	24,24	75,76	14,71	53,70	61,54	45,47	67,68	65,91
1000	0,01	55,22	50,76	75,53	36,36	63,64	26,47	54,35	57,14	51,71	63,22	62,50
0,1	0,1	49,25	48,84	0,00	100,00	0,00	100,00	0,00	49,25	0,00	0,00	0,00
1	0,1	55,22	47,06	85,29	24,24	75,75	14,71	53,70	61,54	45,47	67,68	65,91
10	0,1	55,22	58,78	85,29	24,24	75,76	14,71	53,70	61,54	45,47	67,68	65,91
100	0,1	62,69	66,76	85,29	39,39	60,61	14,71	59,18	72,22	57,97	71,05	69,88
1000	0,1	62,69	66,62	85,29	39,39	60,61	14,71	59,18	72,22	57,97	71,05	69,88
0,1	1	50,75	54,99	26,17	75,76	24,24	73,53	52,94	50,00	44,78	37,44	35,29
1	1	55,22	54,01	97,06	12,12	87,88	2,94	53,23	80,00	34,30	71,88	68,75
10	1	64,18	56,42	91,18	36,36	63,64	8,82	59,62	80,00	57,58	73,73	72,09
100	1	58,21	57,35	79,41	36,36	63,64	20,59	56,25	63,16	53,74	66,83	65,85
1000	1	59,70	57,71	82,35	36,36	63,64	17,65	57,14	66,67	54,72	68,60	67,47
0,1	10	50,75	43,49	100,00	0,00	100,00	0,00	50,75	0,00	0,00	71,24	67,33
1	10	52,24	51,52	94,12	9,09	90,91	5,88	51,61	60,00	29,25	69,70	66,67
10	10	50,75	50,00	100,00	0,00	100,00	0,00	50,75	0,00	0,00	71,24	67,33
100	10	50,75	50,00	100,00	0,00	100,00	0,00	50,75	0,00	0,00	71,24	67,33
1000	10	50,75	50,00	100,00	0,00	100,00	0,00	50,75	0,00	0,00	71,24	67,33

Izvor: izradila autorica

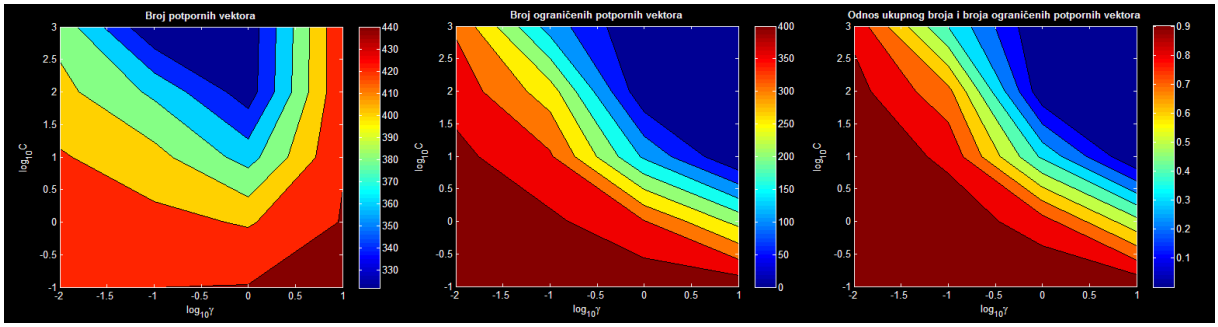


Slika 71. *Usporedba točnosti treniranja (lijevo) i testiranja (desno) za indeks CROBEXindustrija*

Izvor: izradila autorica

Slika 71. prikazuje usporedbu točnosti dobivenu na podacima za treniranje i testiranje. Može se primijetiti da se područja najveće točnosti jednim dijelom preklapaju te da se kod skupa za testiranje nakon $\gamma = 1$ i za $C > 10$ točnost mijenja uglavnom samo s promjenom γ

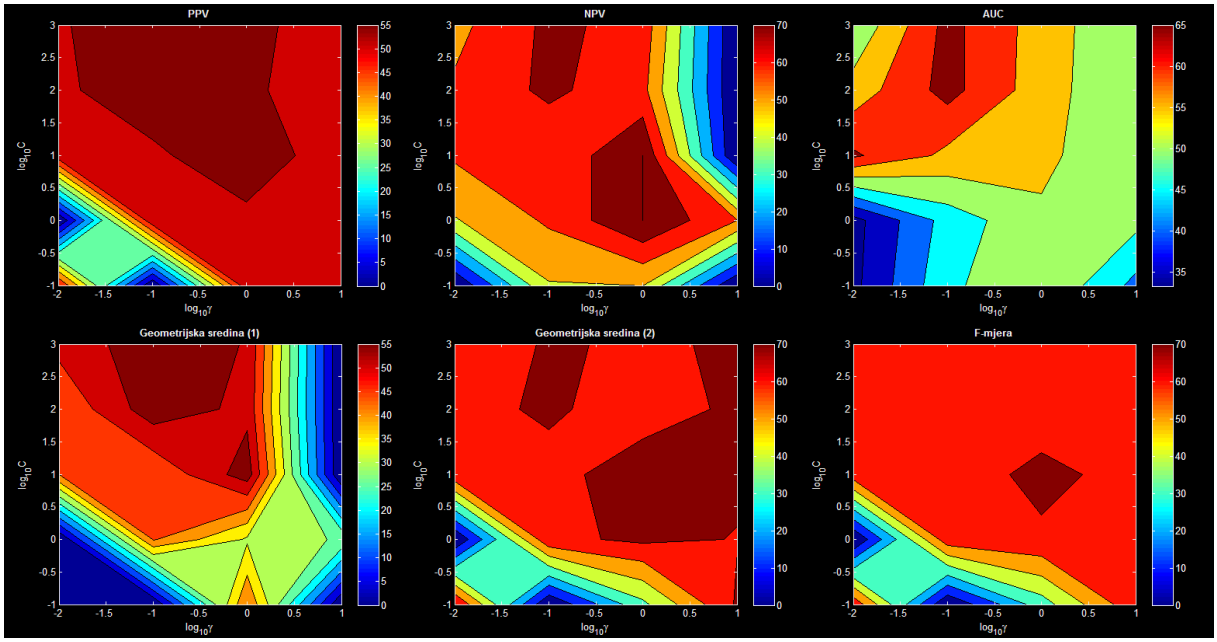
(pojava gotovo paralelnih linija s osi koja prikazuje vrijednosti C), odnosno smanjuje se s povećanjem γ.



Slika 72. *Broj potpornih vektora indeksa CROBEXindustrija: ukupan broj (prikaz lijevo), broj ograničenih (prikaz u sredini), njihov međusobni odnos (prikaz desno)*

Izvor: izradila autorica

Slika 72. prikazuje broj potpornih vektora za različite kombinacije parametara. Može se vidjeti da broj ograničenih potpornih vektora podjednako ovisi o C kao i o γ (smanjuje se kako se povećavaju C i γ), dok njihov ukupan broj do određene vrijednosti γ (u ovom slučaju $\gamma = 1$) ovisi otprilike podjednako o oba parametra, ali iza toga (za $\gamma > 1$) ovisi samo o γ, odnosno povećava se s povećanjem vrijednosti γ. Iz tablice 18. vidljivo je dostizanje kritične vrijednosti parametara C i γ iza kojih više nema ograničenih potpornih vektora čime se izjednačava s klasifikatorom tvrdih margina.



Slika 73. *Odnos različitih kombinacija parametara i evaluacijskih mjera za indeks CROBEXindustrija: PPV, NPV, AUC (gornji red), G1, G2, F-mjera (donji red)*

Izvor: izradila autorica

Uspoređujući ostale mjere, može se primijetiti uglavnom poprilično slaganje s obzirom na područja njihove najbolje ili najlošije vrijednosti. Općenito se najbolji rezultat dobiva za vrijednosti parametara koje su blizu $\gamma = 1$ i $C = 10$, dok se najlošiji rezultat dobiva za istovremeno male vrijednosti C i γ ili za velike vrijednosti γ neovisno o C, što je i u skladu s rezultatima (Tay i Cao, 2001.) i (Keerthi i Lin, 2003.) koji također napominju da u slučaju velikog γ i kad je C veći od određene kritične vrijednosti, SVM klasifikator ne ovisi više o C.

SAX

Korišteni su podaci kraćega niza za predviđanje jedan dan unaprijed s 1495 instanci i 31 značajkom. Rezultati treniranja prikazani su u tablici 19. a testiranja u tablici 20.

Tablica 19. *SAX: rezultati treniranja za različite kombinacije parametara*

C	g	Točnost	Broj potpornih vektora	Broj ograničenih potpornih vektora
0,1	0,01	57,3244	1289	1272
1	0,01	57,3244	1284	1271
10	0,01	57,3244	1281	1272
100	0,01	60,5368	1255	1240
1000	0,01	60,7358	1233	1199
0,1	0,1	57,3244	1281	1269
1	0,1	60,4682	1266	1250
10	0,1	61,806	1232	1188
100	0,1	61,0033	1217	1143
1000	0,1	58,7291	1206	1088
0,1	1	58,2609	1287	1218
1	1	60,8696	1258	1134
10	1	59,2642	1217	1059
100	1	57,9933	1184	930
1000	1	57,3913	1124	738
0,1	10	58,0602	1385	1157
1	10	60,1338	1323	1015
10	10	57,9933	1226	665
100	10	57,2575	1114	288
1000	10	56,6555	993	72

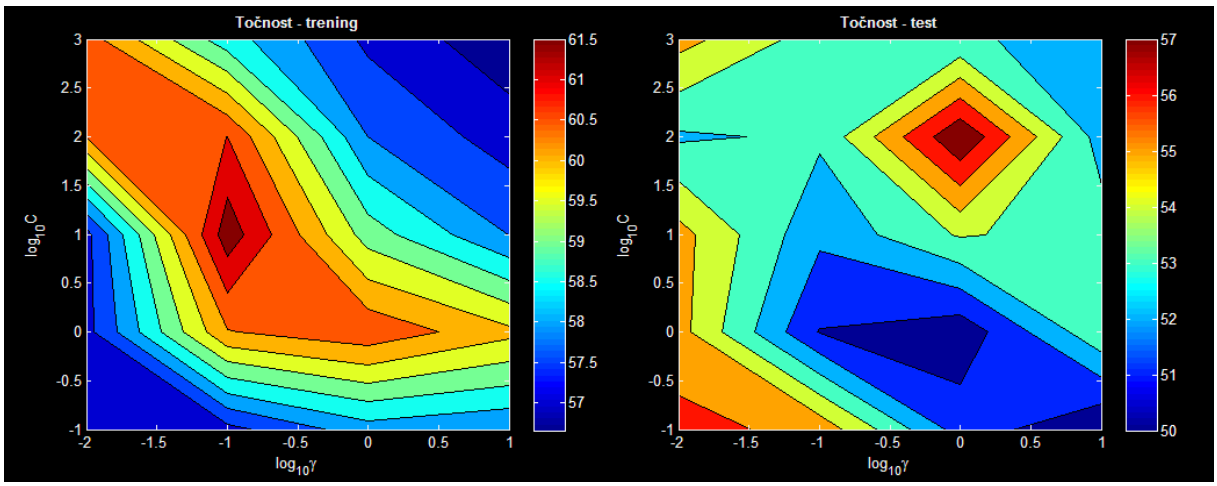
Izvor: izradila autorica

Tablica 20. *SAX: rezultati testiranja za različite kombinacije parametara*

C	g	Točnost	AUC	TPR	TNR	FPR	FNR	PPV	NPV	G1	G2	F
0,1	0,01	56,33	48,11	14,86	92,86	7,14	85,14	64,71	55,32	37,15	31,01	24,18
1	0,01	55,38	47,79	12,84	92,86	7,14	87,16	61,29	54,71	34,53	28,05	21,23

C	g	Točnost	AUC	TPR	TNR	FPR	FNR	PPV	NPV	G1	G2	F
10	0,01	55,38	47,87	12,84	92,86	7,14	87,16	61,29	54,74	34,53	28,05	21,23
100	0,01	52,85	50,56	35,14	68,45	31,55	64,86	49,52	54,50	49,04	41,71	41,11
1000	0,01	55,38	53,25	39,86	69,05	30,95	60,14	53,15	56,59	52,46	46,03	45,56
0,1	0,1	55,70	47,13	20,27	86,90	13,10	79,73	57,69	55,30	41,97	34,20	30,00
1	0,1	50,95	48,65	43,24	57,74	42,26	56,76	47,41	53,59	49,97	45,28	45,23
10	0,1	52,22	51,77	41,22	61,90	38,10	58,78	48,80	54,45	50,51	44,85	44,69
100	0,1	53,16	55,77	47,30	58,33	41,67	52,70	50,00	55,68	52,53	48,63	48,61
1000	0,1	53,48	56,45	48,65	57,74	42,26	51,35	50,35	56,07	53,00	49,49	49,48
0,1	1	51,58	47,54	49,32	53,57	46,43	50,68	48,34	54,55	51,40	48,83	48,83
1	1	50,32	51,45	43,92	55,95	44,05	56,08	46,76	53,11	49,57	45,32	45,30
10	1	54,11	57,10	52,70	55,36	44,64	47,30	50,98	57,06	54,01	51,83	51,83
100	1	57,91	58,16	33,11	79,76	20,24	66,89	59,04	57,51	51,39	44,21	42,42
1000	1	53,16	53,82	7,43	93,45	6,55	92,57	50,00	53,40	26,35	19,28	12,94
0,1	10	50,00	50,14	37,16	61,31	38,69	62,84	45,83	52,55	47,73	41,27	41,04
1	10	53,80	55,01	31,76	73,21	26,79	68,24	51,09	54,91	48,22	40,28	39,17
10	10	53,48	48,62	16,22	86,31	13,69	83,78	51,06	53,90	37,41	28,78	24,62
100	10	52,53	48,19	12,84	87,50	12,50	87,16	47,50	53,26	33,52	24,69	20,21
1000	10	52,53	48,97	13,51	86,90	13,10	86,49	47,62	53,28	34,27	25,37	21,05

Izvor: izradila autorica

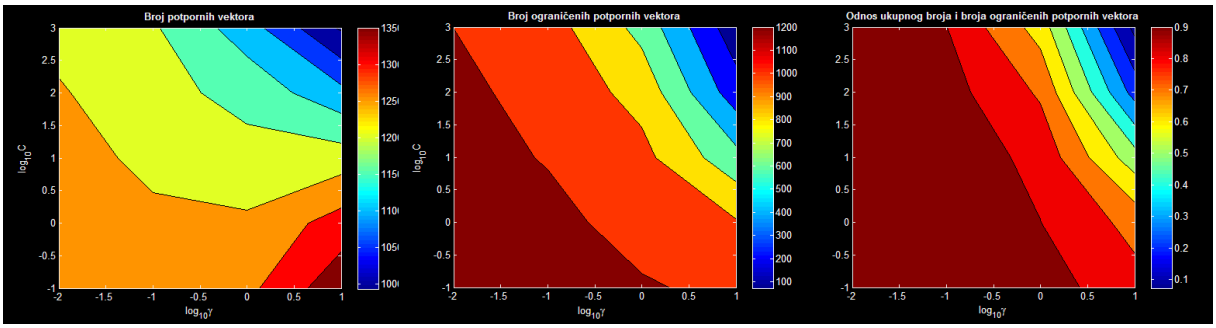


Slika 74. *Usporedba točnosti treniranja (lijevo) i testiranja (desno) za indeks SAX*

Izvor: izradila autorica

Slika 74. otkriva razlog neuspjeha prethodno izgrađenih klasifikatora. Nema preklapanja područja najveće točnosti podataka za treniranje i testiranje. Ako se usporedi s grafikonom kretanja cijena (slika 45.), rezultat upućuje da je došlo do određene značajnije promjene. Podaci za treniranje obuhvaćaju razdoblje globalne financijske krize u kojem je ostvaren veliki pad cijena, dok je razdoblje obuhvaćeno podacima za testiranje mirnije. Naznaka oporavka primjetna je neposredno pred kraj razdoblja za treniranje stoga je jasno da

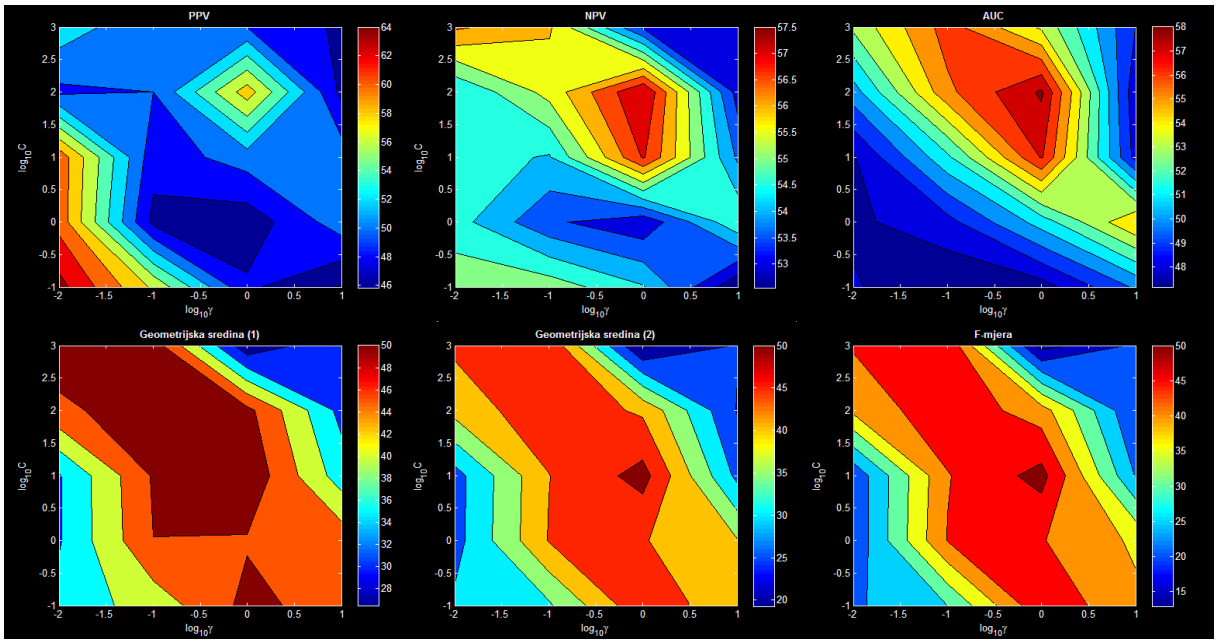
u takvoj situaciji klasifikator ne uspijeva identificirati uzorke kojih u podacima više nema. Optimalni parametri odabrani prilikom treniranja vode izrazito lošim rezultatima tijekom testiranja.



Slika 75. *Broj potpornih vektora indeksa SAX: ukupan broj (prikaz lijevo), broj ograničenih (prikaz u sredini), njihov međusobni odnos (prikaz desno)*

Izvor: izradila autorica

Broj ograničenih potpornih vektora te odnos broja ograničenih i ukupnog broja potpornih vektora pokazuje slično kretanje kao i kod indeksa CROBEXindustrija. Međutim, ukupan broj pokazuje donekle obrnuto kretanje – iza $\gamma = 1$ počinje ovisiti sve više o vrijednosti konstante C, a sve manje o vrijednosti γ.



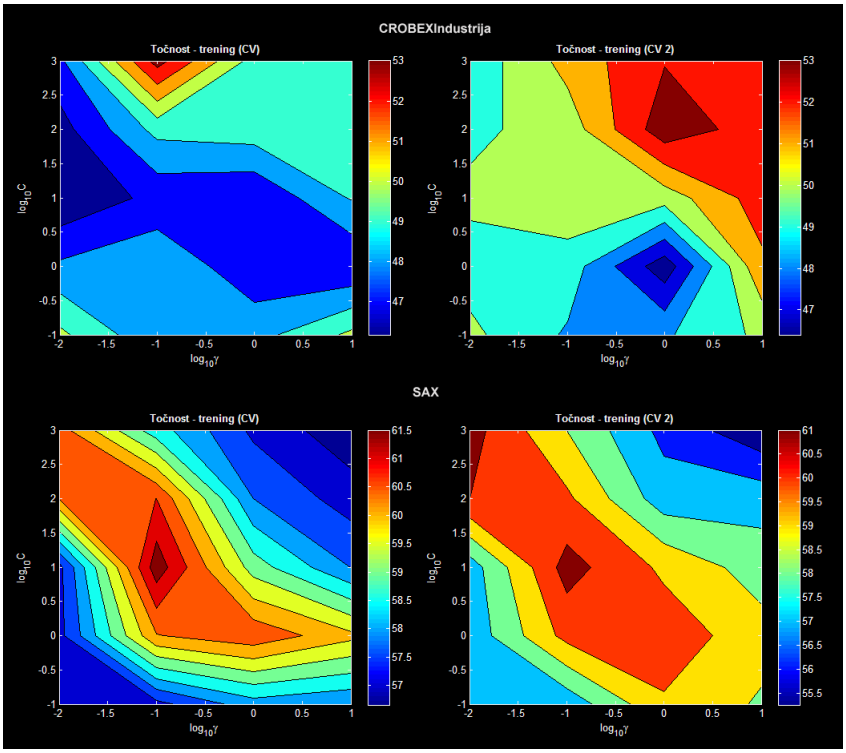
Slika 76. *Odnos različitih kombinacija parametara i evaluacijskih mjera za indeks SAX: PPV, NPV, AUC (gornji red), G1, G2, F-mjera (donji red)*

Izvor: izradila autorica

Uspoređujući rezultate različitih mjera evaluacije vidljivo je znatno manje područje njihovog podudaranja, ali koje je ipak u skladu s označenim "dobrim" područjem na slici 70. Kao najbolja kombinacija parametara pokazala se $\gamma = 1$ i $C = 100$. Međutim, i uz takvu kombinaciju maksimalna točnost iznosi 57,91% uz AUC = 58,16% (omjer klasa je 53,16%)

...

Među rezultatima indeksa CROBEXindustija moguće je primijetiti postizanje veće točnosti testiranja od one prilikom treniranja. Zbog korištenja 5-struke unakrsne validacije prilikom procjene točnosti treniranja na relativno malom broju instanci, pri njezinom uzastopnom ponavljanju rezultati variraju i do nekoliko postotnih poena. Slika 77. prikazuje promjenu optimalne kombinacije parametara pri uzastopnom ponavljanju unakrsne validacije do koje dolazi kod indeksa CROEBXindustrija. Kod indeksa SAX to nije izraženo zbog korištenja većeg skupa podataka kako kod treniranja tako i testiranja.



***Slika** 77.Utjecaj ponovljenog provođenja unakrsne validacije na odabir optimalnih parametara: indeks CROBEXindustrija i mali broj instanci (gornji red), indeks SAX i veći broj instanci (donji red)*

Izvor: izradila autorica

6.6. Simulacija trgovanja

Simulacija trgovanja trebala bi dati konačan odgovor na pitanje uspješnosti klasifikatora i korisnosti rezultata predviđanja prilikom njihove upotrebe unutar zadanih ograničenja. Međutim, u tome se od stvarnosti odstupa u nekoliko bitnih segmenata:

- indeksima se ne trguje[62] već se trguje *futures* ugovorima[63] vezanima za određeni indeks. Simulator je prilagođen trgovanju dionicama, ali radi jednostavnosti pretpostavljen je jednaki mehanizam trgovanja. S obzirom da je svrha provjera rezultata predviđanja, a ne testiranje konkretnog indeksa ili samog procesa trgovanja, u tom smislu ovakvo odstupanje može biti prihvatljivo jer su umjesto indeksa, koristeći potpuno jednaku proceduru dolaska do rezultata, mogle biti uzete u obzir dionice.
- transakcijski troškovi nisu uzeti u obzir.

Simulacija se provodi za vremenski period obuhvaćen testiranjem. Uobičajeni pristup u literaturi je usporedba rezultata s *buy & hold* strategijom, ali kako Chen i Navet (2007.) napominju, česti su tada zaključci da testirani algoritam daje bolje rezultate od navedene strategije u situacijama kad tržište pada, a lošije kad tržište raste. Što nimalo ne iznenađuje s obzirom da je, kad tržište kontinuirano pada, *buy & hold* najgora strategija, a kada kontinuirano raste, vjerojatno najbolja jer donosi dobit bez transakcijskih troškova i uz smanjeni rizik, čime je, ujedno, postavljeno i ograničenje na njezinu prikladnost za ocjenu rezultata. Zbog toga se usporedba vrši i s rezultatima dobivenima na temelju slučajno generiranih signala za kupnju ili prodaju te obrtanjem generiranog signala dobivenoga na temelju rezultata predviđanja.

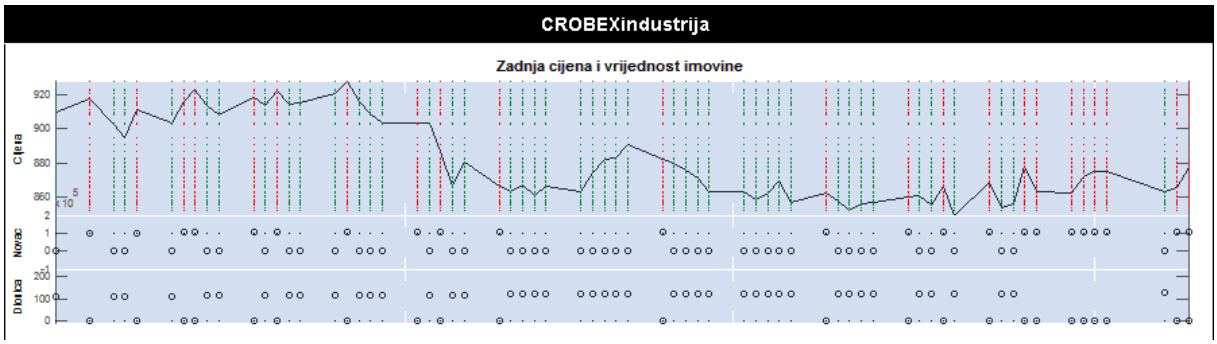
6.6.1. CROBEXindustrija

Iz tablice 21. može se vidjeti da je najbolji rezultat (prinos od 9,79% na početno ulaganje od 100.000) postignut korištenjem rezultata predviđanja najboljega klasifikatora, što je prikazano i na slici 78. S obzirom na negativan tržišni trend, to se svakako može smatrati

62 Doduše, efekt kupnje indeksa mogao bi se postići kupnjom svih dionica u njegovom sastavu i to u onim omjerima u kojima su i sadržane. To je izvedivo u slučaju indeksa koji u svom sastavu imaju mali broj dionica, poput indeksa CORBEX10, dok je prilično nerealistično kod indeksa kao što je S&P500.

63 Futures ugovor je standardizirani ugovor o kupnji ili prodaji predmetne imovine definiranog datuma u budućnosti po tržišno određenoj cijeni.

dobrim rezultatom, pogotovo uzevši u obzir da je *buy & hold* strategija ostvarila gubitak od -3,49%. Slučajno generirani signal također ostvaruje prosječni gubitak od -1,61%, a gubitak (-12,07%) se ostvaruje i obrtanjem signala dobivenoga na temelju rezultata predviđanja najboljega. Time postignuti rezultat dobiva na uvjerljivosti.



Slika 78. *Rezultati simulacije trgovanja korištenjem rezultata najboljeg klasifikatora za indeks CROBEXindustrija: zelene vertikalne linije označavaju signal za kupnju, crvene za prodaju.*

Izvor: izradila autorica

Tablica 21. *Rezultati simulacije za indeks CROBEXindustrija*

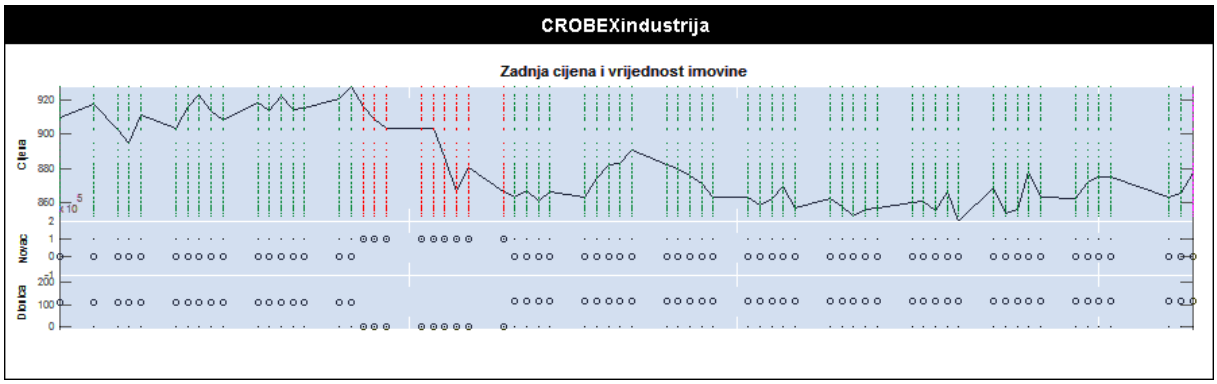
Broj značajki	Točnost testiranja %	Strategija	Prinos %
52	67,16	kupi/prodaj	9,79
		obrnuta	-12,07
22	64,18	kupi/prodaj	7,31
		obrnuta	-10,03
15	53,73	kupi/prodaj	6,34
		obrnuta	-9,23
13	58,21	kupi/prodaj	5,97
		obrnuta	-8,90
1	43,28	kupi/prodaj	-6,65
		obrnuta	3,41
31*	55,22	kupi/prodaj	-2,59
		obrnuta	-0,90
4*	58,21	kupi/prodaj	2,35
		obrnuta	-5,69
2*	40,30	kupi/prodaj	-7,62
		obrnuta	4,45
/	/	slučajno generirani signal**	-1,61
/	/	*buy & hold*	-3,49
2 klasifikatora			
52 / 22	67,16 / 64,18	kupi/prodaj/drži	5,59
		obrnuta	-9,38

* - samo zadnja cijena, ** - prosjek 15 simulacija (kupi/prodaj) uz standardnu devijaciju 2,47

Izvor: izradila autorica

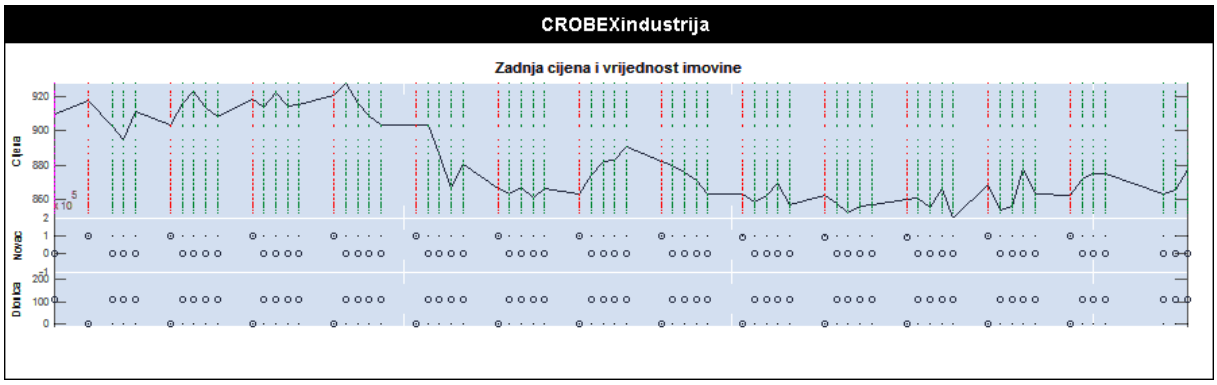
Međutim, ako se uzme u obzir da su u tom razdoblju izvršene 32 transakcije (generirana su 43 signala "kupi" i 23 signala "prodaj"), transakcijski troškovi i ostali stvarni tržišni uvjeti mogli bi poništiti dobar rezultat. Također, unatoč usporedbi sa slučajno generiranim signalom, teško je razdvojiti koliko su ostvarenju rezultata doprinijela sama predviđanja, a koliko ograničenja. Ako je pogrešno predviđen rast i u skladu s time generiran signal "kupi", a zbog nedostatne količine novca to nije učinjeno, pukim smo slučajem spašeni od pogrešnog ulaganja i gubitka. Ali s druge strane, ograničenja sprječavaju i da se realizira svaka povoljna prilika.

I ostali su klasifikatori, čije su ulazne varijable birane među svim mogućima, ostvarili pozitivne rezultate, dok su oni bazirani samo na zaključnoj cijeni ostvarili osjetno lošije rezultate. Slika 79. prikazuje kako se jedan takav gotovo pretvorio u *buy & hold* strategiju budući da su u čitavom periodu izvršene samo tri transakcije, dok se drugi (slika 80.) može nazvati "*prodaj ponedjeljkom*" budući da se signal za prodaju generira samo ponedjeljkom. Zanimljivo je da je to bio jedan od rijetkih slučajeva u kojem se varijabla "ponedjeljak" visoko rangirala s obzirom na značaj prilikom odabira RF algoritmom. Međutim, oba su klasifikatora u potpunosti beskorisna.



Slika 79. *Klasifikator "buy & hold"*

Izvor: izradila autorica



Slika 80. *Klasifikator "prodaj ponedjeljkom"*

Izvor: izradila autorica

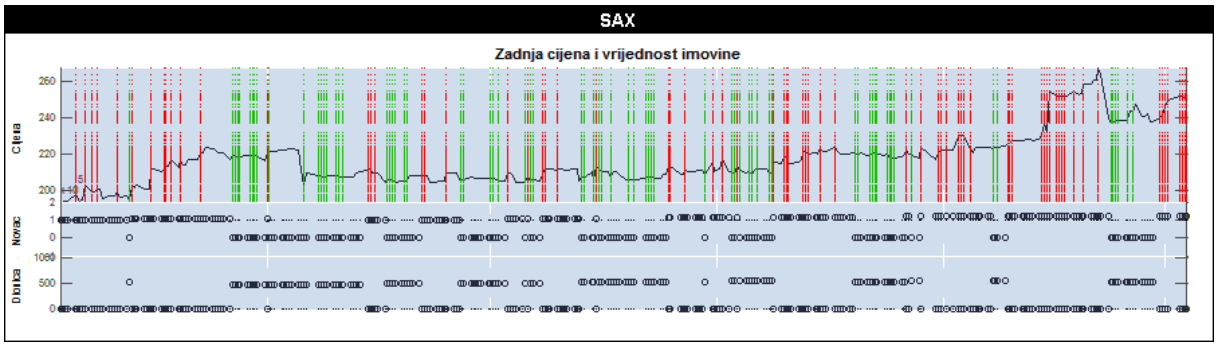
Klasifikator temeljen samo na prinosu kao ulaznoj varijabli ostvario je gubitak od -6,65%. Obrtanjem njegovog predviđanja dobiven je pozitivan rezultat (3,41%), ali ipak nije dostignut rezultat onog najboljega.

Isprobana je i kombinacija dvaju pojedinačno najboljih klasifikatora pri čemu su signali za kupnju ili prodaju generirani samo u slučaju njihovog slaganja. Ostvareni prinos je ipak manji nego što su ga ostvarili svaki pojedinačno (5,59% u odnosu na 9,79% i 7,31%), ali je također generirano i manje signala te realizirano manje transakcija (15 u odnosu na 32 i 18) što bi možda i moglo kompenzirati razliku u prinosu.

Također, potrebno je uzeti u obzir da su klasifikatori binarni i da ne prepoznaju magnitudu promjene zbog čega su moguće situacije da mali broj grešaka na većim promjenama cijene u konačnici dovede do velike razlike u ostvarenom prinosu.

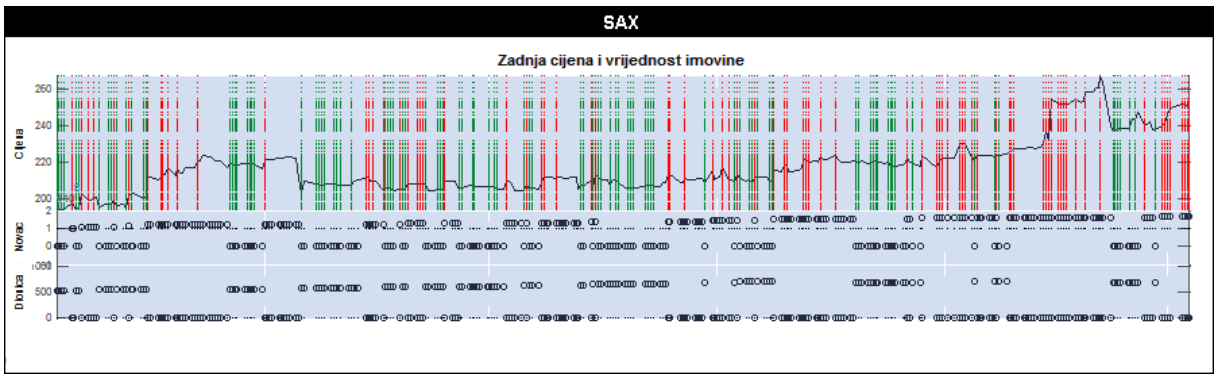
6.6.2. SAX

Ovdje situacija na prvi pogled djeluje prilično iznenađujuće. Najbolji rezultat dobiven je kombinacijom dvaju klasifikatora (63,64%) pri čemu, premda tržište kontinuirano raste, *buy & hold* ostvaruje "samo" 29,32% prinosa. Kombinacija od tri klasifikatora ipak nije uspjela ostvariti bolje od toga. Dodavanje trećega znatno je pogoršalo rezultat, dok je dodavanje drugoga omogućilo postizanje boljeg rezultata nego što bi to ostvario jedan samostalno (tablica 22.). Slika 81. prikazuje najbolji postignut rezultat kombinacijom triju, a slika 82. ukupno najbolje postignut rezultat iz koje je vidljivo da je u potonjem slučaju generiran ponešto veći broj signala a ujedno i izvršen veći broj transakcija. Izgledno je da je upravo povećana dinamika trgovanja doprinijela boljem rezultatu.



Slika 81. *Rezultat simulacije postignut s tri klasifikatora za indeks SAX*

Izvor: izradila autorica



Slika 82. *Najbolji rezultat simulacije za indeks SAX*

Izvor: izradila autorica

Međutim, ono što posebno iznenađuje jest da su svi klasifikatori, čak i pojedinačno, ostvarili rezultat koji poprilično nadmašuje *buy & hold* strategiju. Teško bi bilo povjerovati da je to rezultat dobrog učenja s obzirom da u prethodno provedenom testiranju nije bilo moguće razlikovati rezultat od slučaja iako i usporedba sa slučajno generiranim signalom također ukazuje na bolji rezultat dobiven na temelju predviđanja.

Klasifikator izgrađen korištenjem većeg skupa podataka za učenje i namijenjen predviđanju jedan dan unaprijed postigao je pojedinačno ponešto bolji rezultat od ostalih temeljenih samo na kraćem nizu podataka. Dodavanje drugoga, koji se u testiranju pokazao znatno lošijim, uspjelo je poboljšati rezultat, ali tek za nekoliko postotnih poena.

Za usporedbu je prikazan i rezultat koji bi bio ostvaren da je tijekom treniranja prepoznata optimalna kombinacija parametara (kao što je to utvrđeno u poglavlju 6.5.3), a koja, osim u testiranju, daje i najbolji pojedinačni rezultat tijekom simulacije.

Tablica 22. *Rezultati simulacije za indeks SAX*

Broj dana unaprijed	Broj značajki	Točnost testiranja %	Strategija	Prinos %
3 klasifikatora				
1	30	52,22	3+/3-	24,28
2	26	48,89		
3	17	49,68		
			obrnuta	1,93
1	5	53,48	3+/3-	11,43
2	31	51,75		
3	24	51,91		
2 klasifikatora				
1	30	52,22	2+/2-	47,38
2	26	48,89		
			obrnuta	-12,28
1	31	53,80	2+/2-	63,64
2	26	48,89		
			obrnuta	-20,29
1*	31	52,85	2+/2-	55,40
2*	26	47,30		
1 klasifikator				
1	31	53,80	kupi/prodaj	42,03
1	30	52,22	kupi/prodaj	40,93
1	5	53,48	kupi/prodaj	2,45
1*	31	52,85	kupi/prodaj	48,53
1**	31	57,91	kupi/prodaj	55,42
/	/	/	slučajno generirani signal***	12,11
/	/	/	*buy& hold*	29,32

* - duži niz, ** - s optimalnim parametrima, *** - prosjek za 15 simulacija (kupi/prodaj/drži) uz standardnu devijaciju 13,99

Izvor: izradila autorica

Ukupno uzevši, u ovom je slučaju situacija pomalo nejasna. S jedne strane rezultati testiranja ukazuju da nije došlo do učenja, dok simulacija ukazuje na odstupanje od slučaja. Stoga se nameće zaključak kako je, zahvaljujući spletu određenih okolnosti i ograničenja, moguće privremeno ostvarivati iznimno pozitivne rezultate i to većim dijelom zahvaljujući pukoj sreći i možda pokojem dobro pogođenom predviđanju, ali i to da se kombinacijom klasifikatora od kojih svaki uči svoj zadatak, a koji samostalno možda nisu dovoljno uspješni, može doprinijeti poboljšanju rezultata. Međutim, za čvršće zaključke nužno bi bilo provesti puno opsežnije ispitivanje, što zbog ograničenja korištenog računala nije bilo izvedivo u razumnom vremenu.

7. ZAKLJUČAK

Stroj s potpornim vektorima (SVM) pripada danas samom vrhu klasifikacijskih algoritama strojnog učenja koji bilježi uspješnost u rješavanju najraznovrsnijih problema. Stoga je za potrebe ovoga rada istraženo njegovo ponašanje u rješavanju zahtjevnog zadatka predviđanja smjera kretanja cijena na tržištima vrijednosnica.

Kako bi se doprinijelo boljem razumijevanju samoga algoritma, u teorijskom dijelu rada izložene su osnovne postavke teorije učenja. Ukazano je da se model učenja sastoji od skupa hipoteza koje učeći stroj može implementirati i algoritma za učenje koji iz danog skupa kandidata hipoteza bira onu koja najbolje predviđa odgovor ciljne funkcije pri čemu se odabir vrši korištenjem raspoloživog skupa primjera za učenje. Je li učenje bilo uspješno određuje se na temelju uspješnosti generalizacije, odnosno sposobnosti predviđanja na novim primjerima. Međutim, za postizanje dobrih rezultata predviđanja, nameće se potreba ograničavanja kapaciteta klase funkcija iz koje se bira konačna hipoteza. Mjeru kapaciteta predstavlja VC dimenzija, a u slučaju kad je ona beskonačna, ne može se očekivati nikakva generalizacija.

Kao prvi algoritam proistekao iz statističke teorije učenja, SVM svoju izvrsnu sposobnost generalizacije zahvaljuje upravo implementaciji principa strukturne minimizacije rizika, temeljenog na simultanoj minimizaciji empirijskog rizika i kapaciteta, što mu, za razliku od ostalih algoritama, omogućava da dobro generalizira čak i u slučaju malog broja primjera za učenje. Učenje SVM-a svodi se na rješavanje optimizacijskog problema pronalaženja hiperravnine s maksimalnom marginom. Razlog za korištenje širokih margina je njihov doprinos većoj otpornosti na šum, a ujedno i smanjenju kapaciteta, čime je omogućeno korištenje nelinearnih transformacija ulaznih podataka mapiranih u visokodimenzionalni prostor značajki, s ciljem smanjenja empirijskog rizika. Primjenom kernela, koji se mogu shvatiti kao mjere sličnosti, takvo se mapiranje vrši samo implicitno te se omogućava konstruiranje nelinearnih decizijskih funkcija u ulaznome prostoru ekvivalentnima linearnim decizijskim funkcijama u prostoru značajki, a time i efikasnija primjena u rješavanju brojnih stvarnih problema.

U eksperimentalnom dijelu rada predviđalo se kretanje burzovnih indeksa pri čemu su

kao ulazne varijable korišteni tehnički indikatori, dok je izlaznu varijablu predstavljala promjena predznaka prinosa na određeni dan u budućnosti. Iako su indeksi odabrani na temelju rezultata testiranja predvidljivosti vremenskog niza, samo je u jednom slučaju postignut zadovoljavajući rezultat kada se moglo doista i zaključiti da su utvrđene veze između podataka i pripadajućih oznaka klasa. Glavni uzrok tome može se smatrati nedovoljna prediktivna moć odabranih tehničkih indikatora, što je i očekivano s obzirom da se time ne odbacuje hipoteza efikasnog tržišta. No, da li bi odabir drugačije vrste inputa mogao značajnije poboljšati rezultate, tek bi trebalo istražiti. Stoga bi neki budući rad mogao biti usmjeren na:

- korištenje drugačije vrste inputa (npr. dobivenih na temelju analize sentimenta),
- implementaciju drugačije metode za odabir značajki, s obzirom da se implementirana metoda bazirana na Random forest algoritmu nije pokazala dovoljno uspješnom (premda je u ovome slučaju teško razlučiti koliko su tome doprinijele inicijalno loše odabrane varijable, a koliko sama metoda),
- daljnju razradu ideje kombiniranja klasifikatora od kojih svaki rješava svoj zadatak ili njezino proširenje na izgradnju hibridnog modela kombinacijom više algoritama,
- a ostajući u kontekstu SVM-a, bila bi zanimljiva mogućnost pronalaženja kernela koji bi iskorištavao specifičnosti financijskih vremenskih nizova.

LITERATURA

Knjige

1. ABU-MOSTAFA, Y. S., MAGDON-ISMAIL, M. i LIN, H.-T. (2012.) *Learning from data.* AMLbook.com + dodatna poglavlja [Online] Dostupno na: http://www.amlbook.com/support.html [Pristupljeno: 25. travnja 2015.]
2. ACHELIS, S. B. (2001.) *Technical Analysis from A to Z*. New York: McGraw Hill.
3. BAHOVEC, V. i ERJAVEC, N. (2009.) *Uvod u ekonometrijsku analizu*. Zagreb: Element.
4. GARCIA, S., LUENGO, J. i HERRERA, F. (2015.) *Data Preprocessing in Data Mining.* New York: Springer.
5. HAMEL, L. (2009.) *Knowledge Discovery with Support Vector Machines.* New Yersey: JohnWiley&Sons, Inc.
6. HYNDMAN, R. i ATHANASOPOULOS, G. (2013.) *Forecasting: principles and practice.* OTexts. [Online] Dostupno na: https://www.otexts.org/book/fpp. [Pristupljeno: 27. travnja 2015.]
7. JAPKOWITZ, N. i SHAH M. (2011.) *Evaluatong learning algorithms:A Classification Perspective.* Cambridge: Cambridge University Press.
8. KECMAN V. (2001.) *Learning and Soft Computing: Support Vector Machines, Neural Networks, and Fuzzy Logic Models*. London: MIT Press.
9. MITCHELL, T. (1997.) *Machine Learning.* New York:McGraw Hill.
10. MURPHY, J. J. (1999.) *Technical analysis of the financial markets: A comprehensive guide to trading methods and applications*. Paramus: NewYork Institute of Finance.
11. NEGNEVITSKY M. (2005.) *Artificial Intelligence: A Guide to Intelligent Systems*, Second Edition, Pearson Education, Essex
12. SCHOELKOPF, B. i SMOLA, A. J. (2002.) *Learning with Kernels:Support Vector Machines, Regularization, Optimization, and Beyond.* London: MIT Press.

13. SHARP, H., ROGERS, Y. i PREECE, J. (2002.) Interaction Design: Beyond Human-Computer Interaction. West Sussex: John Wiley & Sons Ltd.
14. STANCZYK, U. i JAIN, L. C. (ur.) (2015.) *Feature Selection for data and Pattern Recognition.* New York: Springer.
15. TORGO, L. (2010.) *Data mining with R: learning with case studies*. Chapman & Hall/CRC.
16. VIDUČIĆ, LJ. (2006.) *Financijski menadžment, V. dopunjeno i izmijenjeno izdanje.* Zagreb: RRiF plus

Članci

1. ABU-MOSTAFA, Y. S. i ATTYA, A. (1996.) Introduction to Financial Forecasting. *Applied Intelligence.* 6 (3) str. 205–213
2. AKBANI, R., KWEK, S. i JAPKOWICZ, N. (2004.) Applying support vector machines to imbalanced datasets. Na: *Machine Learning: ECML 2004.* str. 39-50. Springer Berlin Heidelberg.
3. ATSALAKIS, G.S. i VALAVANIS, K.P. (2009.) Surveying stock market forecasting techniques – Part II: Soft computing methods. *Expert Systems with Applications* 36 (3) str. 5932-5941.
4. BACIU, O. A. (2014.) Ranking Capital Markets Efficiency: The Case of Twenty European Stock Markets. *Journal of Applied Quantitative Methods.*9(3).str. 24-33.
5. BARBIĆ, T (2010.a) Pregled razvoja hipoteze efikasnog tržišta.*Privredna kretanja i ekonomska politika.* 20(124). str. 29-62.
6. BARBIĆ, T. (2010.b) Testiranje slabog oblika hipoteze efikasnog tržišta na hrvatskom tržištu kapitala. *Zbornik Ekonomskog fakulteta u Zagreb.* 8(1). str.155-172.
7. BOSER, B.E., GUYON, I.M. i VAPNIK, V.N. (1992.) A Training Algorithm for Optimal Margin Classifiers, *Proceedings Fifth ACM Workshop on Computational Learning Theory, COLT 1992.* Pittsburgh, PA, SAD, 27.-29.07.1992. Pittsburgh: *s*tr. 144–152.
8. BOULESTEIX, A. L., JANITZA S., KRUPPA, J., KOENIG, I. R. (2012.) Overview

of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery.* 2(6). str. 493-507.

9. BREIMAN, L. (2001.) Random forests. *Machine learning. 45*(1). str. 5-32.
10. BURGES, C. J. C.(1998.) A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery.* 2. str.121-164.
11. CAO, L. i TAY, F.E.H.(2001.) Financial Forecasting Using Support Vector Machines. *Neural Computing & Applications.*10(2) str.184-192.
12. CHANG, C.-C. i LIN, C.-J. (2001.) LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology.* 2:27:1-27:27. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.
13. CHEN, S.-H. i NAVET, N. (2007.) Failure of genetic-programming induced trading strategies: Distinguishing between efficient markets and inefficient algorithms. U: *Computational Intelligence in Economics and Finance, Volume II.* Springer-Verlag: Berlin.str. 169-182.
14. CHEN, Y. W., i LIN, C. J. (2006.) Combining SVMs with various feature selection strategies. U: *Feature Extraction Foundations and Applications* (str. 315-324). Springer Berlin Heidelberg.
15. CONT, R. (1999.) Statistical properties of financial time series. Na: *Symposium on Mathematical Finance*, Fundan University, Shangai, 10.-24.08.1999.
16. CONT, R. (2001.) Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance.* 1. str. 223-236.
17. CORTES, C. i VAPNIK, V. (1995.) Support-Vector Networks, *MachineLearning.* 20. Str. 273-297.
18. FAMA, E. F. (1970.) Efficient capital markets: A review of theory and empirical work. *The journal of Finance.* 25(2). str. 383-417.
19. FAMA, E. F. (1991.) Efficient capital markets: II. *The journal of finance*,. 46(5). str. 1575-1617.
20. GUYON, I., i ELISSEEFF, A. (2003.) An introduction to variable and feature

selection. *The Journal of Machine Learning Research. 3. str.* 1157-1182.

21. HELLSTROEM, T. i HOLMSTROEM, K. (1998.) *Predicting the Stock Market* [Online] Dostupno na: http://www.technicalanalysis.org.uk/general/HeHo98.pdf. [Pristupljeno: 23. svibnja 2015.]

22. HUANG, W., NAKAMORI, Y. i WANG, S.-Y. (2005.) Forecasting stock market movement direction with support vector machine. *Computers & Operations research*. 32(10) str. 2513-2522

23. KARA, Y., BOYACIOGLU, M.A. i BAYAKAN, Ö.K. (2011.) 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.* 38(5) str. 5311-5319.

24. KARAGIANNOPOULOS , M., ANYFANTIS, D., KOTSIANTIS, S. B., PINTELAS, P. E. (2007.) Feature selection for regression problems. *Proceedings of HERCMA'07*.

25. KARPOFF, J. M. (1987.) The Relation Between Price Changes and Trading Volume: A Survey. *The Journal of Financial and Quantitative Analysis.* 22(1). str. 109-126.

26. KEERTHI, S. S. i LIN, C.-J. (2003.) Asymptotic Behaviors of Support Vector Machines with Gaussian Kernel. *Neural Computation.* 15(7). str. 1667-1689.

27. KIM, K. (2003.) Financial time series forecasting using support vector machines. *Neurocomputing.* 55 (1-2) str. 307-319.

28. KRISTOUFEK, L i VOSVRDA, M. (2013.) Measuring capital market efficiency: Global and local correlations structure. *Physica A: Statistical Mechanics and its Applications.*392(1). str. 184-193.

29. LEONARDELLI, J. (2012.) Playing with history can affect your future: how handling missing dana can impact parameter estimation and risk measure [Online] Dostupno na: http://www.frgrisk.com/system/resources/14/document/Playing%20with%20history%20can%20effect%20your%20future_original.pdf [Pristupljeno: 25. veljače 2015.]

30. LO, A. W. (2007.) Efficient markets hypothesis. U:*The New Palgrave Dictionary of Economics, Second Edition.* New York: Palgrave Macmillan.

31. LO, A. W. i MACKINLAY, A. C (1989.) The size and power of the variance ratio test

in finite samples: A Monte Carlo investigation. *Journal of Econometrics.* 40(2) str. 203-238.

32. LO, A. W., REPIN, D. V. i STEENBARGER, B. N. (2005.) Fear and greed in financial markets: A clinical study of day-traders. *Cognitive Neuroscientific Foundations of Economic Behavior.* 95(2). str. 352-359

33. MALKIEL B. G. (2003.a) *A random walk down Wall Street: the time-tested strategy for successful investing.* New York : W.W. Norton.

34. MALKIEL, B. G. (2003.b) The efficient market hypothesis and its critics. *Journal of economic perspectives.* 17(1). str. 59-82.

35. PREIS, T., MOAT, H. S., i STANLEY, H. E. (2013.) Quantifying trading behavior in financial markets using Google Trends. *Scientific reports. 3*.

36. SAJTER D. (2013.) Algoritamsko i visoko-frekventno trgovanje, *Ekonomska misao i praksa.* 22(1). str. 321-336.

37. SI, J., MUKHERJEE, A., LIU, B., LI, Q., LI, H., DENG, X. (2013.) Exploiting Topic based Twitter Sentiment for Stock Prediction. *ACL.* (2) str. 24-29.

38. SMOLA, A. J., i SCHOELKOPF, B. (2004.) A tutorial on support vector regression. *Statistics and computing.* 14(3). str. 199-222.

39. SUN, W. (2003.) Relationship between trading volume and security prices and returns. *Area Exam Report, Technical Report, MIT Laboratory for Information and Decision Systems.*

40. ŠKRINJARIĆ, T. (2012.) Kalendarski učinci u prinosima dionica. *Ekonomski pregled. 63*(11). str. 651-678.

41. TAY, F.E.H. i CAO, L. (2001.) Application of support vector machines in financial time series forecasting. *Omega.* 29 (4) str. 309-317

42. THE GOVERNMENT OFFICE FOR SCIENCE (2012.) *Foresight: The Future of Computer Trading in Financial Markets Final Project Report.* London. [Online]

Dostupno na:
https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/289431/

12-1086-future-of-computer-trading-in-financial-markets-report.pdf [Pristupljeno: 20.

lipnja 2014.]

43. TSAIH, R., HSU, Y. i LAI, C.C. (1998.) Forecasting S&P 500 stock index futures with a hybrid AI system. *Decision Support Systems*. 23(2). str.161-174.
44. VAPNIK, V. N. (1999.) An overview of statistical learning theory. *Neural Networks, IEEE Transactions on.* 10(5). str. 988-999.
45. ZEMKE, S. (2002.) On developing a financial prediction system: Pitfalls and possibilities. Na: *Proceedings of DMLL-2002 Workshop at ICML-2002*. Sydney, Australia.

Ostalo

1. MATLAB

   http://www.mathworks.com/products/matlab/

2. The Bratislava Stock Exchange.
   http://www.bsse.sk/bcpben/Trading/Indices/SAXIndex.aspx

3. Yahoo Finance

   http://finance.yahoo.com/

4. Zagrebačka burza

   http://www.zse.hr/

POPIS SLIKA

Slika 1. Komponente problema nadziranog učenja.....5
Slika 2. Mogućnost savršenog razdvajanja podataka: a) linearnom, b) nelinearnom granicom odlučivanja.....7
Slika 3. Overfitting: iako se kompleksnija funkcija (crvena linija) savršeno prilagođava podacima za učenje (označeni kružićima), njenim se odabirom postiže veća greška prilikom predviđanja na novim primjerima.....8
Slika 4. VC dimenzija: linearnom funkcijom u dvodimenzionalnom prostoru moguće je "razbiti" tri točke.....9
Slika 5. Nelinearnom funkcijom u dvodimenzionalnom prostoru moguće je "razbiti" tri kolinearne točke (lijevo) ili četiri točke (desno) što nije moguće postići linearnom funkcijom 10
Slika 6. Odnos empirijskog rizika i VC-pouzdanosti: empirijski rizik opada s povećanjem kompleksnosti skupa hipoteza iz kojeg se bira f, ali povećava se i kazna koja se plaća za dodatnu kompleksnot. Zbrajanje empirijskog rizika i VC-pouzdanosti rezultira granicom generalizacije.....11
Slika 7. Vektori.....15
Slika 8. Linearno separabilni podaci.....17
Slika 9. Razdvajajuća hiperravnina i margina na primjeru binarne klasifikacije.....18
Slika 10. Široka margina i šum: unatoč prisutnosti šuma (predstavljenog kružnicama) podaci će biti ispravno klasificirani.....18
Slika 11. Graf ciljne funkcije.....19
Slika 12. Graf Lagrangeove funkcije.....20
Slika 13. Samo potporni vektori određuju hiperravninu.....23
Slika 14. VC dimenzija i uske margine.....25
Slika 15. VC dimenzija i široke margine.....25
Slika 16. Neseparabilni podaci: ako se dopusti određeno toleriranje greške i dalje je moguće konstruirati linearnu granicu odlučivanja (lijevo), situacija u kojoj nije moguće konstruirati linearnu granicu odlučivanja (desno).....26
Slika 17. Meke margine: primjeri označeni brojevima 4 i 5 krše marginu, ali i dalje su ispravno klasificirani. Primjer označen brojem 6 pogrešno je klasificiran.....28
Slika 18. Utjecaj konstante C na širinu margine: mala vrijednost konstante C stvara široke

margine (lijevo), velika vrijednost stvara uske margine (desno).....29
Slika 19. Mapiranje u visoko dimenzionalni prostor značajki: korištenje nelinearnih transformacija ulaznih podataka omogućava stvaranje linearnih granica odlučivanja u prostoru značajki koje odgovaraju nelinearnim granicama odlučivanja u ulaznome prostoru.....31
Slika 20. Granice odlučivanja u ulaznom prostoru dobivene primjenom različitih kernela: a) polinomijalnim, b)RBF kernelom.....33
Slika 21. Shematski prikaz aplikacije.....47
Slika 22. Kotacija Zagrebačke burze.....48
Slika 23. Shema pripreme podataka za učenje.....49
Slika 24. Box-Cox transformacija: a) originalni podaci, b) transformirani podaci i širenje histograma kao posljedica transforamcije.....55
Slika 25. Pomični prosjeci: jednostavni (plava linija), eksponencijalni (crvena linija) za 40 dana. Eksponencijalni prosjek pokazuje nešto veću osjetljivost.....56
Slika 26. Oblici redukcije podataka: odabir značajki (gore), odabir instanci (u sredini), diskretizacija (dole).....62
Slika 27. Shema metode filtra za odabir značajki.....64
Slika 28. Shema metode omotača za odabir značajki.....65
Slika 29. Shema odabira modela i učenja.....69
Slika 30. Evaluacija metodom pomičnog prozora: slijedno pomicanje (lijevo), slučajni odabir početne točke (desno).....72
Slika 31. Neravnoteža u podacima i optimalna hiperravnina: naučena(lijevo), idealna (desno) 74
Slika 32. Shema evaluacije konačnog modela.....76
Slika 33. Krivulja učenja: s povećanjem količine primjera za učenje empirijski se rizik približava stvarnom riziku.....77
Slika 34. ROC prostor: stavlja u odnos stopu stvarno pozitivnih i stopu lažno pozitivnih primjera.....81
Slika 35. Sučelje za pregled i podjelu podataka.....86
Slika 36. Sučelje za testiranje efikasnosti.....87
Slika 37. Sučelje za generiranje značajki.....88
Slika 38. Sučelje za odabir značajki.....89
Slika 39. Grafički prikaz rezultata pretraživanja optimalnih parametara modela.....90

Slika 40. Sučelje za odabir modela i treniranje....91
Slika 41. Sučelje za testiranje modela....92
Slika 42. Konfuzijska matrica i dodatne evaluacijske mjere....93
Slika 43. Sučelje za simulaciju trgovanja....94
Slika 44. Glavni koraci u eksperimentu....95
Slika 45. Kretanja zaključne cijene odabranih indeksa s označenim podjelama podataka na skupove za učenje i testiranje....100
Slika 46. Logaritam prinosa, ACF i PACF: a) CROBEXindustrija, b) S&P500, c) SAX (1), d) SAX (2)....102
Slika 47. Apsolutna vrijednost logaritma prinosa, ACF, PACF za indeks S&P500....103
Slika 48. Rangiranje varijabli po značaju za indeks CROBEXindustrija: a) rezultat nakon prve iteracije, b) rezultat nakon četvrte iteracije....103
Slika 49. Rezultat rangiranja varijabli po značaju nakon prve iteracije za predviđanje jedan dan unaprijed: a) indeks S&P500, b) indeks SAX....104
Slika 50. Rezultat rangiranja varijabli po značaju za indeks SAX kod predviđanja: a) dva dana unaprijed, b) tri dana unaprijed....105
Slika 51. Podaci indeksa S&P500 namijenjeni treniranju: podaci s jednom (lijevo) i podaci s 23 značajke (desno)....107
Slika 52. Postignuta točnost za različite kombinacije parametara: podaci s jednom i 23 značajke....108
Slika 53. Podaci indeksa S&P500 reprezentirani s 23 značajke u prikazu smanjene dimenzionalnosti: ispunjeni kružići označavaju potporne vektore....109
Slika 54. Podaci indeksa CROBEXindustrija namijenjeni treniranju: podaci s 13, 15, 22, 52 značajke bazirane na svim cijenama i volumenu (gornji red), podaci s 1, 2, 4, 31 značajkom baziranom samo na zaključnoj cijeni (donji red)....110
Slika 55. Podaci indeksa CROBEXindustrija reprezentirani s 52 značajke u prikazu smanjene dimenzionalnosti: ispunjeni kružići označavaju potporne vektore....111
Slika 56. Podaci indeksa SAX namijenjeni treniranju: podaci kraćega niza s 31 značajkom (lijevo) i podaci dužega niza s 26 značajki (desno)....112
Slika 57. Podaci indeksa SAX (kraći niz) reprezentirani s 31 značajkom u prikazu smanjene dimenzionalnosti: ispunjeni kružići označavaju potporne vektore....113
Slika 58. Podaci indeksa CROBEXindustrija namijenjeni testiranju: podaci s 13, 15, 22, 52

značajke bazirane na svim cijenama i volumenu (gornji red), podaci s 1, 2, 4, 31 značajkom baziranom samo na zaključnoj cijeni (donji red).....115
Slika 59. Konfuzijska matrica za indeks CROBEXindustrija.....116
Slika 60. Usporedba ROC krivulja klasifikatora temeljenog na podacima s 13 značajki (prikaz lijevo) i onog izgrađenog samo na temelju podataka o prinosu (prikaz desno) za indeks CROBEXindustrija.....117
Slika 61. Slučajno generirani podaci.....118
Slika 62. ROC krivulje za indeks CROBEXindustrija: rezultat predviđanja na podacima s permutiranim oznakama klasa (lijevo) i slučajno generiranim podacima (desno). ROC krivulja originalnih podataka je podebljana.....119
Slika 63. Konfuzijska matrica za indeks CROBEXindustrija: originalni podaci, podaci s permutiranim oznakama klasa, slučajno generirani podaci.....119
Slika 64. Podaci indeksa SAX namijenjeni testiranju predviđanja promjene predznaka prinosa jedan dan unaprijed: podaci s pet i 31 značajkom.....120
Slika 65. Konfuzijska matrica za indeks SAX.....121
Slika 66. ROC krivulje indeksa SAX za najveći postignuti AUC: u predviđanju jedan, dva, tri dana unaprijed te dužeg niza u predviđanju promjene predznaka jedan dan unaprijed.....121
Slika 67. Konfuzijska matrica za indeks SAX (kraći niz): originalni podaci, podaci s permutiranim oznakama klasa, slučajno generirani podaci.....122
Slika 68. ROC krivulje za indeks SAX (kraći niz): rezultat predviđanja na podacima s permutiranim oznakama klasa (lijevo) i slučajno generiranim podacima (desno). ROC krivulja originalnih podataka je podebljana.....122
Slika 69. ROC krivulje za indeks SAX (duži niz): rezultat predviđanja na podacima s permutiranim oznakama klasa (lijevo) i slučajno generiranim podacima (desno). ROC krivulja originalnih podataka je podebljana.....123
Slika 70. Prostor underfittinga, overfittinga i optimalnih kombinacija parametara i.....124
Slika 71. Usporedba točnosti treniranja (lijevo) i testiranja (desno) za indeks CROBEXindustrija.....126
Slika 72. Broj potpornih vektora indeksa CROBEXindustrija: ukupan broj (prikaz lijevo), broj ograničenih (prikaz u sredini), njihov međusobni odnos (prikaz desno).....127
Slika 73. Odnos različitih kombinacija parametara i evaluacijskih mjera za indeks CROBEXindustrija: PPV, NPV, AUC (gornji red), G1, G2, F-mjera (donji red).....127

Slika 74. Usporedba točnosti treniranja (lijevo) i testiranja (desno) za indeks SAX.....129
Slika 75. Broj potpornih vektora indeksa SAX: ukupan broj (prikaz lijevo), broj ograničenih (prikaz u sredini), njihov međusobni odnos (prikaz desno).....130
Slika 76. Odnos različitih kombinacija parametara i evaluacijskih mjera za indeks SAX: PPV, NPV, AUC (gornji red), G1, G2, F-mjera (donji red).....130
Slika 77.Utjecaj ponovljenog provođenja unakrsne validacije na odabir optimalnih parametara: indeks CROBEXindustrija i mali broj instanci (gornji red), indeks SAX i veći broj instanci (donji red).....131
Slika 78. Rezultati simulacije trgovanja korištenjem rezultata najboljeg klasifikatora za indeks CROBEXindustrija: zelene vertikalne linije označavaju signal za kupnju, crvene za prodaju. 133
Slika 79. Klasifikator "buy & hold".....134
Slika 80. Klasifikator "prodaj ponedjeljkom".....135
Slika 81. Rezultat simulacije postignut s tri klasifikatora za indeks SAX.....136
Slika 82. Najbolji rezultat simulacije za indeks SAX.....136

POPIS TABLICA

Tablica 1. Testiranje varijanti hipoteze slučajnog hoda.....52
Tablica 2. Tehnički indikatori korišteni u aplikaciji.....57
Tablica 3. Primjer kombinacija parametara za grid search pretraživanje.....69
Tablica 4. Indeksi i razdoblja podvrgnuta testiranju efikasnosti tržišta.....97
Tablica 5. Rezultati testa omjera varijanci.....98
Tablica 6. Odabrani indeksi i podjele podataka na skupove za učenje i testiranje.....99
Tablica 7. Pregled deskriptivne statistike za logaritme prinosa.....100
Tablica 8. Rezultati treniranja za indeks S&P500.....107
Tablica 9. Rezultati treniranja za indeks CROBEXindustrija.....110
Tablica 10. Rezultati treniranja za indeks SAX.....112
Tablica 11. Rezultati testiranja za indeks CROBEXindustrija.....115
Tablica 12. Rezultati testiranja najboljeg klasifikatora na različitim vrstama podataka.....118
Tablica 13. Rezultati testiranja za indeks SAX.....120
Tablica 14. Rezultat predviđanja na različitim vrstama podataka za indeks SAX (kraći niz) 121
Tablica 15. Rezultat predviđanja na različitim vrstama podataka za indeks SAX (duži niz). 123
Tablica 16. Utjecaj parametara na pojavu underfittinga ili overfittinga.....125
Tablica 17. CROBEXindustrija: rezultati treniranja za različite kombinacije parametara.....125
Tablica 18. CROBEXindustrija: rezultati testiranja za različite kombinacije parametara.....126
Tablica 19. SAX: rezultati treniranja za različite kombinacije parametara.....128
Tablica 20. SAX: rezultati testiranja za različite kombinacije parametara.....128
Tablica 21. Rezultati simulacije za indeks CROBEXindustrija.....133
Tablica 22. Rezultati simulacije za indeks SAX.....137

SAŽETAK

Stroj s potpornim vektorima pripada danas samom vrhu klasifikacijskih algoritama strojnog učenja s uspješnom primjenom u rješavanju najraznovrsnijih problema. Kao prvi algoritam proistekao iz statističke teorije učenja, svoju izvrsnu sposobnost generalizacije zahvaljuje implementaciji principa strukturne minimizacije rizika, baziranog na simultanoj minimizaciji empirijskog rizika i VC dimenzije, odnosno kapaciteta klase funkcija koje učeći stroj implementira.

U radu se istražuje mogućnost predviđanja smjera kretanja cijena na tržištima vrijednosnica primjenom stroja s potpornim vektorima pri čemu su kao ulazne varijable korišteni tehnički indikatori, dok je izlaznu varijablu predstavljao predznak prinosa na određeni dan u budućnosti. S obzirom da, osim o samome algoritmu, uspješnost klasifikacije ovisi i o ostalim elementima sustava, ispitivan je utjecaj odabira značajki, različitih kombinacija parametara, neravnoteže u podacima te duljine niza na rezultate klasifikacije. Uspoređivane su različite evaluacijske mjere, a rezultati predviđanja testirani su i u simulatoru trgovanja gdje se pokazalo da se bolji rezultat može dobiti kombinacijom više klasifikatora na način da svaki od njih uči rješavati svoj zadatak.

Od tri burzovna indeksa, iako odabrana za eksperiment na temelju testova predvidljivosti vremenskog niza, samo su kod jednoga u konačnici postignuti zadovoljavajući rezultati s obzirom da se kao osnovna prepreka boljim rezultatima pokazala nedovoljna prediktivna moć odabranih tehničkih indikatora.

Ključne riječi: stroj s potpornim vektorima, burzovni indeks, klasifikacija.

SUMMARY

Support vector machine today belongs to the very top of the machine learning classification algorithm, with successful application in resolution of all kinds of problems. As the first algorithm was the result of the statistical learning theory, it owes its excellent generalization performance to the implementation of the structural risk minimalization principle, based on simultaneous minimalization of the empirical risk and VC-dimension, i.e. the capacity of the class of functions implemented by the learning machine.

In this paper, the possibility of prediction of the stock market price movement was researched by means of implementation of the support vector machine, where technical indicators were used as input variables, while the sign of the excess of returns on a specific date in the future represented the output variable. Given that, apart from the algorithm itself, the classification performance also depends on other system elements, the impact of the choice of features on the classification results, i.e. different parameter combinations, data imbalance, as well as the series length, were also examined. Different evaluation measures were compared and the prediction results were also tested in a market trading simulator, where it was shown that a better result could be obtained by the combination of multiple classifiers, in the manner that each one learns how to solve its own task.

Out of the three stock market indexes, although chosen for the experiment on the basis of the time series predictability tests, ultimately in only one of them satisfactory results were achieved, given that insufficient predictive power of selected technical indicators proved to represent the main obstacle in obtaining better results.

Keywords: support vector machine, stock market index, classification.