UNIVERZA NA PRIMORSKEM
FAKULTETA ZA MATEMATIKO, NARAVOSLOVJE IN
INFORMACIJSKE TEHNOLOGIJE

Magistrsko delo

Modeliranje časovnih vrst satelitskih meritev

(Time-Series modeling on the satellite data)

Ime in priimek: Duško Topić
Študijski program: Računalništvo in informatika, 2. stopnja
Mentor: izr. prof. dr. Janez Žibert

Koper, september 2017

Ključna dokumentacijska informacija

Ime in PRIIMEK: Duško TOPIĆ

Naslov magistrskega dela: Modeliranje časovnih vrst satelitskih meritev

Kraj: Koper

Leto: 2017

število listov: 49 število slik: 13 število tabel: 18

število referenc: 33

Mentor: izr. prof. dr. Janez Žibert

UDK: 004.414.23(043.2)

Ključne besede: Analiza podatkovne zbirke, modeliranje podatkov, metoda podpornih vektorjev, časovne vrste, izpeljava značilk, RapidMiner, Matlab.

Izvleček: V magistrskem delu bosta predstavljena dva možna pristopa modeliranja časovnih vrst. Prvi pristop je neposredno modeliranje, drugi pristop pa modeliranje nad izpeljanimi značilkami. Pri slednjem je v fazi predpriprave podatkov potrebno iz podatkovne zbirke izpeljati ustrezne značilke, ki bodo predstavljale novo množico za modeliranje - v tem primeru so to statistični funkcionali. V prvem delu bo podrobneje analizirana podatkovna zbirka satelitskih podatkov vegetacije določenega področja v različnih časovnih obdobjih, nato bo predstavljen proces izgradnje modelov za oba pristopa. V drugem delu bodo predstavljeni rezultati modeliranja detekcije vegetacije in tudi binarne detekcije specifičnega tipa vegetacije, poudarek bo na primerjavi natančnosti obeh pristopov.

Key words documentation

Name and SURNAME: Duško TOPIĆ

Title of final project paper: Time-Series modeling on the satellite data

Place: Koper

Year: 2017

Number of pages: 49 Number of figures: 13 Number of tables: 18

Number of references: 33

Mentor: Assoc. Prof. Janez Žibert, PhD

UDC: 004.414.23(043.2)

Keywords: Database analysis, data modeling, Support Vector Machine, time-series data modeling, feature extraction, RapidMiner, Matlab.

Abstract: In this master's thesis two possible approaches of modeling time-series will be presented. The first approach is direct modeling, the second one is modeling on extracted features. The difference in the latter case is in the data preparation process, where new features must be extracted from the initial database, that will present the new dataset for modeling - in this case statistical functionals will be used. In the first part of this thesis the database of satellite data of vegetation in the specific area in various time periods will be analyzed, then the process of building models will be presented. for both approaches. In the second part the modeling results of vegetation detenction and binary detection of specific vegetation type will be presented, whilst the focus will be on comparison of the accuracy of both approaches.

Kazalo vsebine

Kazalo tabel

Kazalo slik

Seznam kratic

AUC	ploščina pod ROC krivuljo
EKG	elektrokardiogram
HMM	prikriti Markov model
ROC	karakteristika delovanja sprejemnika
SVM	metoda podpornih vektorjev
VAR	vektorski avtoregresijski model
ZDA	Združene države Amerike

Zahvala

Posebej bi se zahvalil mentorju izr. prof. dr. Janezu Žibertu za dragocene nasvete pri izgradnji modelov ter za usmeritve in popravke pri pripravi magistrskega dela.

Posebna zahvala gre tudi ženi Vidi za vso podporo in potrpežljivost skozi celoten študij, pa tudi otrokoma Nikoli in Filipu, ki sta mi vlila dodatno motivacijo za dokončanje študija.

Zahvaljujem se tudi staršem, bratu in vsem ostalim prijateljem, ki so me ves čas vzpodbujali in podpirali.

1 UVOD

Podatki, ki vsebujejo meritve v časovnih vrstah, so prisotni na marsikaterem področju, med drugim v financah pri analiziranju borznega trga [26], v meteorologiji za potrebe analize in napovedovanja vremenskih pojavov, v seizmologiji za analize in napovedovanje potresov [21] [1], v medicini pri analizi EKG signalov [22], v elektrotehniki pri razpoznavanju govora [11] ali na področju telekomunikacije pri napovedi obremenitev brezžičnega omrežja [13], pa tudi v agronomiji pri detekcijah vegetacije, kot se bo izkazalo v nadaljevanju. Časovna vrsta je niz določenega parametra, pridobljenega oz. izmerjenega v določenem časovnem zaporedju. Časovni interval in razpon meritev sta lahko poljubna (odvisno od tipa podatkov), pomembno je le, da oprema in pogoji za meritve parametrov ostanejo nespremenjeni, saj se s tem izognemo morebitnem popačenju rezultatov.

Za boljšo predstavo si oglejmo primer časovne vrste. V tem magistrskem delu razpolagamo s podatkovno zbirko s področja vegetacije poljščin z določenim številom vhodnih parametrov, ki so bili na enak način večkrat izmerjeni na enakih geografskih področjih v razponu 18-ih mesecev, in sicer ob neenakomernih časovnih intervalih - kar v konkretnem primeru ne igra bistvene vloge, saj zadostuje že po ena meritev iz posameznega dela sezone rastja. V tem primeru posamezna časovna vrsta predstavlja meritev posameznega vhodnega parametra ob vseh datumih meritev. Tabela 1 prikazuje konkreten primer časovne vrste za parameter rdeče valovne dolžine. V prvi vrstici so zapisani datumi meritev (prikazanih je prvih 6 meritev), druga vrstica pa prikazuje izmerjeno vrednost parametra v določenem datumu. Skupek vseh meritev istega parametra tvori niz, imenovan časovna vrsta.

Tabela 1: Primer časovne vrste parametra rdeče valovne dolžine.

Datum	11.4.2013	18.5.2013	15.6.2013	2.7.2013	29.7.2013	18.8.2013
Vrednost	0.411	0.734	0.237	0.144	0.720	1.382

V tem magistrskem delu smo uporabili dva orodja, in sicer za proces predpriprave podatkovne zbirke ter izračun statističnih funkcionalov (predstavljeni bodo v razdelku 3.2) smo izvedli z orodjem Matlab [31] s pomočjo razširitvenega modula »Statistics

and Machine Learning Toolbox≪. Pripravljeno podatkovno zbirko smo nato uvozili v RapidMiner [32], s katerim smo izpeljali različne tehnike modeliranja in tudi primerjali pridobljene rezultate.

V razdelku 1.1 se nahaja pregled drugih raziskovalnih del, ki so povezani s področjem magistrskega dela. V poglavju 2 je predstavljen namen magistrskega dela, podrobneje so tudi opisane delovne hipoteze s predvidenimi rezultati. V poglavju 3 je najprej opisana podatkovna zbirka, zatem pa metode in pristopi, ki so uporabljeni pri modeliranju časovnih vrst, vključno z evalvacijo postopkov. V poglavju 4 se nahaja interpretacija ter primerjava rezultatov obeh pristopov modeliranja, v poglavju 5 pa so povzete zaključne ugotovitve, vključno z možnimi nadaljnjimi raziskavami na tem področju.

1.1 Pregled področja

V nadaljevanju se nahaja kratek pregled nekaterih primerljivih raziskovalnih del na področju modeliranja časovnih vrst. V splošnem poznamo dva pristopa:

- neposredno modeliranje, kjer podatkovno množico v obliki časovne vrste neposredno modeliramo z ustreznimi statističnimi modeli,
- posredno modeliranje, kjer časovno vrsto predstavimo na drug, enakovreden način (npr. s statističnimi funkcionali), pridobljene značilke nato modeliramo z ustreznim modelom.

Oglejmo si nekaj raziskav, ki so uporabile pristop neposrednega modeliranja. V raziskavi ti. spremljanja zdravja infrastrukture (ang. *Structural Health Monitoring*) [18] je na podlagi pridobljenih meritev potrebno pripraviti model, ki bo napovedal prihodnje vrednosti meritev posameznega senzorja. V tem primeru torej ne gre za klasifikacijo, temveč za napovedovanje prihodnjih vrednosti v časovni vrsti (ang. *Forecasting*). Vhodno podatkovno zbirko predstavljajo meritve senzorjev, ki med drugim periodično beležijo temperaturo infrastrukture ter pritiske oz. silo na določen del infrastrukture, senzorje gibanja itd. Ker senzorji pridobivajo meritve v določenem časovnem intervalu, je v podatkih prisotna časovna vrsta. Avtor je zgradil dva modela, kot prvi je uporabil model ARIMA, ki vsebuje avtoregresijske člene (ang. *Auto Regressive*) in vrednosti napak preteklih napovedi (ang. *Moving Average*) ter kot drugi model okenske transformacije (ang. *Windowing*), kjer za želeno časovno vrsto izberemo velikost okna (število preteklih meritev) in poljubno združevalno funkcijo za napoved novih vrednosti.

V naslednji raziskavi [24] je predstavljena problematika neposrednega modeliranja časovnih vrst, ki pride do izraza predvsem v veliki količini meritev v podatkovni

množici, in sicer časovna kompleksnost modeliranja. Kot primer je naveden postopek klasifikacije z metodo prileganja z ukrivljanjem časovne osi (ang. *Dynamic Time Warping*), ki ima kvadratično časovno kompleksnost izračuna. Kot rešitev predstavlja metodo za klasificiranje, ti. BOSS VS (ang. *Bag-Of-Symbolic-Fourier-Approximation-Symbols in Vector Space*), ki je časovno manj zahteven (linearna časovna kompleksnost). Testiranje metode na več podatkovnih množicah je pokazalo, da natančnost ostane nespremenjena, kvečjemu se v nekaterih primerih tudi izboljša.

Sledi raziskovalno delo [29], v katerem je analizirana izpostavljenost pacientovega tveganja, da se okuži s ti. C.*diff* (lat. *Clostridium difficile*) bakterijo v času hospitalizacije. Podatkovno zbirko predstavljajo parametri bolnikov, ki so ostali hospitalizirani v bolnišnici dlje od 7 dni (saj je verjetnost okužbe v prvem tednu praktično nična), vrednosti posameznih parametrov pa se beležijo za vsak bolnišnični dan, torej imamo časovno vrsto. Cilj raziskovalnega dela je pripraviti model, ki bo znal predvideti, ali bo bolnik v naslednjem bolnišničnem dnevu močno izpostavljen okužbi z omenjeno bakterijo. Avtor je primerjal natančnost treh različnih metod. Prva je metoda prikritih Markovih modelov oz. HMM (ang. Hidden Markov Model), kjer se predpostavi, da obstaja zaporedje skritih stanj, ki določajo ugotovitve z neko verjetnostjo. V raziskavi sicer razpolagamo z dvema stanjema, kjer eno stanje predstavlja nizko stopnjo tveganja, drugo pa visoko stopnjo tveganja okužbe. Drugi predlagan postopek je uporaba mere podobnosti (ang. *similarity measure*), ki kot rezultat vrne stopnjo ujemanja dveh objektov. Kot algoritem je uporabljen SVM, kot jedro pa funkcija RBF (ang. *Radial Basis Function*). Kot že omenjeno, se je v obeh primerih izvajal pristop neposrednega modeliranja.

V isti raziskavi je kot tretji postopek avtor uporabil drugi pristop, in sicer izpeljavo časovnih značilk iz neposrednih meritev. To pomeni, da je avtor izpeljal 17 statističnih funkcionalov, ki predstavljajo nove značilke. Večina je specifičnih glede na tip podatkov, kot primer lahko navedemo značilko, kjer so poznejši dnevi bolj obteženi, ali pa značilka, ki predstavlja vsoto tveganj v zadnjih 3 dneh. Z izpeljanimi značilkami je ustvarjena nova podatkovna množica, nad katero se nadaljuje z modeliranjem na običajen način - v tem primeru je avtor uporabil algoritem SVM. Izkaže se, da je tretji postopek modeliranja s pomočjo izpeljave značilk za odtenek bolj natančno v primerjavi s preostalima dvema predstavljenima modeloma.

Oba pristopa je prav tako možno zaslediti v raziskavah nad meritvami, kjer so prisotna nenadna občutna odstopanja, kot je v primeru napovedi gibanja menjalniških tečajev. Pristop z neposrednim modeliranjem je uporabljen v napovedi [25] gibanja tečaja ka-

zakstanske valute napram ameriškem dolarju, pa tudi v napovedi [12] ameriškega dolarja napram evru. V obeh primerih je uporabljen model ARIMA, medtem ko avtor v raziskavi [4] predlaga alternativen pristop, in sicer uporabo kombinacije nevronskih mrež z izpeljavo statističnih značilk za modeliranje napovedi tečaja srbskega dinarja napram evru. Primerjava obeh pristopov [16] je pokazala, da se z uporabo nevronskih mrež doseže večja natančnost v primerjavi z ARIMA modelom.

Sledi še nekaj primerov posrednega modeliranja. V splošnem je pri vsakem primeru možnih več načinov implementacij, odvisno od tipa podatkov, s katerim razpolagamo. V primeru velikega števila časovnih meritev, posebej če se vrednosti ponavljajo v periodični obliki, je praviloma ustrezna uporaba Furier-ovega transforma [20]. Podobna je tudi naslednja raziskava [14], katere cilj je klasificirati stanje pljuč (normalno ali abnormalno) na podlagi vhodne meritve, in sicer zvoka, posnetega z električnim stetoskopom. Iz vhodnih signalov so izpeljane nove značilke - statistični funkcionali. Nad slednjimi se je nato za nadaljnji proces modeliranja uporabil postopek valovne transformacije (ang. *Wavelet Transform*).

Naslednji primer uporabe pri izpeljavi statističnih funkcionalov je raziskava za razpoznavo in detekcijo govora ob uporabi odprtokodne rešitve OpenEAR [8]. Avtor nad izvorno podatkovno množico z neposrednimi meritvami izpelje 35 statističnih funkcionalov. Zatem nadaljuje s klasičnim postopkom modeliranja nad novo podatkovno množico, v konkretnem primeru je izbran model k-NN (ang. *k nearest neighbours*) oz. k najbližjih sosedov, kjer za vsak primer poišče v dani množici k najbližjih oz. najbolj podobnih primerov in hkrati oceni verjetnostno porazdelitev k-tih primerov po razredih.

Sledita dve raziskavi, pri obeh so iz vhodnih signalov izpeljane nove značilke, konkretno statistični funkcionali. Pri prvi raziskavi [23] je avtor primerjal metodo SVM z metodo najbližjih sosedov nad več različnimi podatkovnimi množicami - izkaže se, da je bila metoda SVM bolj natančna v 5-ih primerih, medtem ko metoda najbližjih sosedov v preostalih 3-eh primerih. Pri drugi raziskavi [15] je avtor modeliral nad podatkovno množico meritev srčnih utripov. Tokrat se je avtor osredotočil le na uporabo SVM, a je primerjal natančnost klasifikacije z različnim številom značilk. Pričakovano se izkaže, da se z večanjem števila značilk izboljša natančnost modela.

2 NAMEN

Kot že omenjeno, bosta v magistrskem delu podrobneje predstavljena dva možna pristopa modeliranja časovnih vrst. Običajno se modeliranje izvaja nad nespremenjenimi podatki oz. parametri, torej z neposrednim modeliranjem podatkovne zbirke. Tak pristop je sicer lahko učinkovit, a podatkovna zbirka se z dodajanjem novih parametrov, ali še pomembneje, novih časovnih obdobij meritev lahko občutno poveča, s tem pa se poveča tudi časovna kompleksnost izračuna. Poleg tega je težava še v pripravi modelov, saj je v primeru dodajanja nove časovne meritve potrebno ponoviti celoten postopek modeliranja. Ravno zaradi slednjega razloga je problematika še vedno aktualna, podatkovne zbirke se v sedanjosti na dnevni bazi povečujejo v nepreglednih količinah. Poglavitni razlogi so v porasti števila senzorjev, ki periodično izvajajo meritve, zmožnost hitre pretočnosti podatkov, medtem ko diskovne kapacitete za shranjevanje ne predstavljajo večjega problema. Tudi algoritmi v različnih orodjih so že optimizirani do te mere, da v realnosti bistvenih izboljšav ni za pričakovati, zato se je potrebno osredotočiti na optimalno pripravo podatkov, da se bo model učil le na najbolj informativnih parametrih.

Namen magistrskega dela je na realnih podatkih, kjer so prisotne časovne vrste, predstaviti različne možne pristope k reševanju problema. Za doseganje uspešne realizacije namena bodo v nadaljevanju definirani cilji tega dela.

Raziskovalno vprašanje, s katerim se ukvarjamo v magistrskem delu, je, ali obstaja razlika v klasifikaciji podatkov pri modeliranju časovnih vrst z dvema metodama: z metodo, kjer značilke predstavljajo neposredne časovne meritve in metodo z izpeljanimi časovnimi značilkami. Iz tega sledi delovna hipoteza, ki jo preverejamo: Model z izpeljanimi časovnimi značilkami je vsaj enakovreden neposredenemu modeliranju časovnih vrst. Delovno hipotezo preverjamo na satelitskih meritvah za detekcijo različnih vegetacijskih tipov.

Cilji magistrskega dela so naslednji:

- pregled ostalih raziskav in metod z enako tematiko,
- zgraditi optimiziran model, ki bo potrdil delovno hipotezo,

- identificirati najbolj informativne parametre - značilke - za modeliranje časovnih vrst,
- izvesti primerjavo med različnimi načini modeliranja časovnih vrst.

3 METODE

V nadaljevanju bo predstavljena metodologija modeliranja na realni podatkovni zbirki. Po podrobnejšem pregledu le-te bo večji poudarek na predstavitvi dveh metod za modeliranje časovnih vrst.

3.1 Podatkovna zbirka satelitskih podatkov

Podatkovna zbirka, ki smo jo uporabili v primeru modeliranja časovnih vrst, je namenjena detekciji vegetacije iz satelitskih podatkov, zgrajena je iz 619644 poligonov, kjer vsak poligon predstavlja točno določeno geografsko območje in hkrati je za vsak poligon pridobljenih 11 satelitskih meritev v letih 2012 in 2013. Definiranih je 9 različnih tipov poljščin:

- ječmen,
- pšenica,
- koruza,
- oljna buča,
- oljna ogrščica,
- tritikala,
- neznana poljščina,
- druga poljščina,
- ne-poljščina.

Prvih 8 tipov hkrati spada pod kategorijo poljščina, medtem ko pod zadnji tip spadajo vsi preostali poligoni, ki niso klasificirani kot poljščina. Detektirali bomo, ali se na izbranem poligonu nahaja katera izmed 6-ih znanih poljščin ali pa le-ta sploh ni poljščina. Tipa neznana poljščina ter druga poljščina ne bomo poizkušali detektirati, saj sta preveč splošna, v katerih se nahaja več različnih poljščin, bodisi znanih ali neznanih, kar pomeni, da iz take množice poligonov ne moremo pridobiti uporabnega

vzorca. Uporabili bomo satelitske meritve iz sezone 2013 – teh je 9, in sicer med aprilom in oktobrom, meritve so zajete približno na vsake 3-4 tedne.

Satelitske meritve so zajete s strani satelita Sentinel 2, s katerim upravlja evropska vesoljska agencija ESA (ang. *European Space Agency*). V raziskovalne namene je satelit zajemal slike v slovenski regiji v več različnih časovnih intervalih. Satelit vsebuje senzorje za zajem ti. osnovnih meritev, kot so vrednosti v različnih valovnih dolžinah (rdeča, zelena, modra valovna dolžina in NIR ter RedEdge valovni dolžini). Na podlagi pridobljenih osnovnih meritev se naknadno izpelje še nekaj relevantnih indeksov (npr. ARI-1, BAI, PSRI-NIR itd.), ki ustvarijo še dodatne informacije in bodo predstavljeni v nadaljevanju.

Posamezna meritev vsebuje 11 parametrov, vsi so navedeni v tabeli 2. ARI-1 (ang. *Anthocyanin Reflectance Index* 1) beleži vrednost antocianina – gre za vodotopen pigment, veliko ga je v novih in odmrlih listih. Višje vrednosti parametra je opaziti v pomladnih in jesenskih mesecih. BAI (ang. *Burn Area Index*) beleži indeks pogorelosti zemlje. Višje vrednosti je opaziti v poletnih mesecih, kar je za pripisati večjim sušnim obdobjem. Sledijo parametri, ki beležijo prisotnost rdeče, zelene in modre valovne dolžine ter parametri, ki beležijo vrednosti v različnih valovnih dolžinah, in sicer NIR (ang. *Near Infrared*), ki beleži na valovni dolžini $700 - 1200nm$ ter RedEdge, ki beleži na valovni dolžini $690 - 730nm$. Rdeča valovna dolžina beleži vrednosti med $600 - 700nm$ valovne dolžine.

Naslednji parameter je PSRI-NIR (ang. *Plant Senescence Reflectance Index*), gre za indeks odbojnosti senescence rastlin v ti. NIR območju. ChlRedEdge beleži razmerje med vrednostjo klorofila v dveh valovnih dolžinah, in sicer rdeča ter RedEdge. Parameter NDVI (ang. *Normalized Difference Vegetation Index*) je vegetacijski indeks, ki beleži stopnjo vegetacije. Je najbolj znan in tudi najbolj razširjen vegetacijski indeks, saj je enostaven za izračun, hkrati pa določi vrednost, iz katere se jasno razbere tip rastja [33]. Razpon vrednosti indeksa je med 0 in 1. Negativne vrednosti predstavljajo vodnato področje (npr. morje, jezero), vrednosti okrog ničle predstavljajo gola, nerodovitna področja (npr. kamen, puščava, sneg), medtem ko pozitivne vrednosti predstavljajo znake rastja, od travnatih površin pri nizkih vrednostih, pa vse do tropskega pragozda (bogato rastje) pri vrednosti 1. Zadnji parameter je NDVI-Green, ki je produkt indeksa NDVI ter zelene valovne dolžine. Za konec je potrebno omeniti, da so vsi parametri zveznega tipa, gre za realna števila. Parametrom sledi devet razrednih atributov, ki predstavljajo numerična binarna števila, ki določajo pripadnost poligona k posamezni poljščini (0-ne pripada, 1-pripada). Posamezen poligon lahko pripada le

Tabela 2: Vhodni parametri s pripadajočo enačbo. Parametri, ki nimajo zapisane enačbe, so pridobljeni na podlagi neposrednih meritev senzorjev.

Parameter	**Izračun**
ARI-1	$\frac{1}{green} - \frac{1}{rededge}$
BAI	$\frac{1}{(0.1-red)^2+(0.06-nir)^2}$
Blue	/
Red	/
Green	/
NIR	/
RedEdge	/
PSRI-NIR	$\frac{red-blue}{nir}$
ChlRedEdge	$\left(\frac{nir}{rededge}\right)^{-1}$
NDVI	$\frac{nir-red}{nir+red}$
NDVI-Green	$ndvi \cdot green$

enemu tipu poljščine hkrati.

Poleg vhodnih parametrov smo naknadno dodali še dodaten parameter Δ_t (v nadaljevanju: delta v času t), ki beleži spremembo vrednosti parametra med dvema zaporednima poligonoma v določenem času. V bistvu gre za funkcijo odvoda, ki je definirana na naslednji način:

$$\Delta_t = \frac{c_{t+1} - c_{t-1}}{2}$$

Parameter delta nam veliko pove o dinamiki v podatkih, npr. koliko se vrednosti parametrov spremenijo med dvema časovnima obdobjema, oziroma nakazujejo trend spreminjanja časovne vrste ob času t. Na primer, če predpostavimo, da imamo vse vrednosti določenega parametra enake skozi vsa časovna obdobja, bi bile vrednosti delte povsod enake 0, saj bi bila razlika med meritvami prav tako enaka 0.

V tem trenutku se je število parametrov iz začetnih 11 podvojilo na 22 parametrov za posamezen datum meritve, saj je delta posebej pridobljen nad vsakim izvornim parametrom. Za lažje razumevanje nadaljnje interpretacije je v tabeli 3 definirana notacija atributov. Najprej je zapisanih 11 atributov v zaporedju datum_parameter, kjer se poleg določenega datuma zapišejo vsi izmerjeni parametri (z izjemo delte). Temu

sledi naslednjih 11 atributov v zaporedju datum_parameter_delta, kjer so za posamezen datum in parameter zapisane še vrednosti parametra delta. Tako imamo za posamezen datum skupno 22 atributov. Nato se postopek ponovi z naslednjimi datumi. Ker je skupno 9 različnih datumov, je končno število atributov v podatkovni zbirki 198.

Tabela 3: Prikaz notacije zapisanih atributov v podatkovni zbirki.

Ime atributa
2013-04-11_ari-1
2013-04-11_bai
...
2013-04-11_ari-1_delta
2013-04-11_bai_delta
...
2013-05-18_ari-1
...
2013-10-26_rededge_delta

Na koncu smo podatkovni zbirki dodali še 8 binarnih razrednih atributov, po enega za vsak tip poljščine. Vsak poligon lahko ima le na enem izmed teh vrednost enako 1 (na tistem tipu, kateremu dejansko pripada), na vseh ostalih pa vrednost enako 0.

Podatkovno zbirko smo razdelili na dva dela, učno (75% delež) in testno množico (25% delež). Na tako obsežni količini podatkov zaradi pomanjkanja resursov ni bilo možno izpeljati postopkov modeliranja, zato smo izdelali novo, razvojno podatkovno množico, ki predstavlja 10% delež naključno izbranih vzorcev izvorne podatkovne zbirke, s tem ohranimo značilnosti izvorne zbirke (razmerje med poljščinami, razmerje med vzorci znotraj posamezne poljščine itd.). Tako imamo 44.935 vzorcev v učni in 14.980 v testni množici. Tabela 4 prikazuje razporeditev podatkov v razvojni množici. Kot je razvidno, je razmerje podatkov med učno in testno množico dokaj uravnoteženo.

Omejitve, s katerimi smo se srečali, so nezmožnost modeliranja celotne podatkovne zbirke zaradi preobsežne količine zajetih podatkov. Omejitev smo odpravili tako, da smo ustvarili manjšo, razvojno podatkovno zbirko, nad katero je možno izvajati postopke modeliranja v sprejemljivem času. Druga omejitev so prisotnost manjkajočh podatkov. Pristop k reševanju problema je tak, da smo v primeru več kot 3-eh manjkajočih podatkov meritev istega parametra, parameter enostavno odstranili iz podatkovne zbirke.

Tabela 4: Razporeditev meritev v razvojni podatkovni množici.

Tip poljščine	**Učna množica**	**Testna množica**
Poljščina	21661 (48.2%)	7232 (48.3%)
Ne-poljščina	23274 (51.8%)	7748 (51.7%)
Skupaj	44935 (100%)	14980 (100%)

Za lažjo predstavo, kako je v praksi videti podatkovno zbirko v obliki časovne vrste v podatkovni množici, si oglejmo naslednjo matriko.. V primeru vektorja parametrov $p = p_1, p_2, \ldots, p_n$, kjer n predstavlja število vhodnih parametrov, ter meritev $t = t_1, t_2, \ldots, t_m$, kjer m predstavlja število časovnih obdobij, dobimo matriko P dimenzije $n \cdot m$, kjer n predstavlja število vrstic, m pa število stolpcev:

$$P = \begin{bmatrix} p_1^{t_1} & p_1^{t_2} & \ldots & p_1^{t_m} \\ p_2^{t_1} & p_2^{t_2} & \ldots & p_2^{t_m} \\ & \vdots & & \\ p_n^{t_1} & p_n^{t_2} & \ldots & p_n^{t_m} \end{bmatrix}$$

Hitro opazimo, da se z dodajanjem časovnih obdobij sorazmerno poveča tudi velikost matrike. Postopek je do neke točke sicer še možno obvladovati, a nimamo modela, ki bi uspešno deloval nad poljubnim številom časovnih meritev, saj se ob vsaki nadaljnji časovni meritvi poveča število atributov. Poleg tega se lahko pojavi problematika, da pri vsakem parametru iz različnih razlogov (npr. manjkajoči podatki) nimamo vsakič m časovnih meritev, kar lahko privede do popačenja rezultatov ali celo nezmožnosti izvedbe algoritmov za modeliranje. V takih primerih je potrebno izpeljati postopek interpolacije oziroma ekstrapolacije. Razlika med postopkoma je v tem, da pri interpolaciji imamo znani vrednosti parametra v času t_{i-1} ter t_{i+1}, mi pa želimo oceniti manjkajočo vmesno vrednost parametra v času t_i. Ekstrapolacija pa je metoda, ki počne ravno obratno - na podlagi vrednosti parametrov do zadnjega znanega časovnega obdobja, recimo jim $t_1, t_2, \ldots, t_j$, ocenimo vrednost parametra ob vseh nadaljnjih časih $t_{j+1}, t_{j+2}, \ldots, t_m$, dokler ne dosežemo m časovnih obdobij. Slednja metoda pride v poštev v primerih, ko imamo na razpolago nižjo količino časovnega razpona meritev, posledično je potrebno ročno dodati nove, primerljive vrednosti.

Potrebno je še omeniti, da so vrednosti vseh parametrov normalizirane, kot metoda normalizacije je uporabljena ti. statistična standardizacija podatkov oz. Z-transformacija.

Cilj postopka je doseči porazdelitev vrednosti meritev s povprečjem enakim 0 ter varianco enako 1. Formula za izračun standardizacije:

$$Z = \frac{X - u}{s},$$

kjer X predstavlja vektor vrednosti parametra, parameter u predstavlja povprečno vrednost vhodnega parametra, parameter s pa standardni odklon.

V našem primeru sicer ni bilo potrebe po uporabi postopkov interpolacije ali ekstrapolacije, saj smo razpolagali z znanimi vrednostmi meritev, prav tako smo za vsak parameter imeli enako število časovnih obdobij. Možnost bi bila, da bi uporabili postopek interpolacije na način, da med dvema obdobjema z znanimi meritvami vrinemo novo, vmesno obdobje z interpoliranimi vrednostmi. To možnost smo opustili, saj je 9 meritev v eni sezoni predstavljalo ustrezen razpon meritev.

3.2 Izpeljava časovnih značilk

Drugi pristop modeliranja bo nad izpeljanimi značilkami, kar pomeni, da bo v fazi predpriprave podatkov iz izvorne podatkovne zbirke potrebno izpeljati ustrezne značilke, ki bodo predstavljale novo podatkovno množico za modeliranje. Cilj postopka je imeti nespremenjeno število atributov, ne glede na količino časovnih meritev. S tem dosežemo, da ni potrebno bistveno spreminjati oz. prilagajati modela v primeru dodajanja meritev novega časovnega obdobja. Izpeljane značilke bodo v našem primeru statistični funkcionali, navedeni v tabeli 5. V prvem stolpcu je naveden opis značilke, v drugem stolpcu je zapisana pripadajoča formula za izpeljavo, v tretjem stolpcu pa je zapisano ime, ki se bo nahajalo v atributu v podatkovni zbirki.

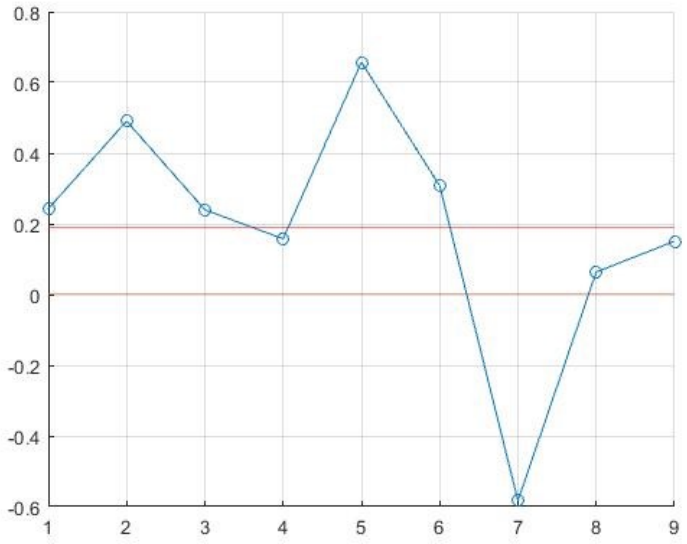
Prvi štirje statistični funkcionali predstavljajo najvišjo ter najnižjo vrednost znotraj vektorja, ter pripadajoči indeks, kjer se vrednost nahaja. Naslednji statistični funkcional predstavlja razpon (ang. *range*) v meritvah, torej razliko med najvišjo in najnižno vrednostjo vektorja. Nato sledita povprečna vrednost vektorja ter povprečna absolutna sprememba – to je razlika med izbrano vrednostjo v primerjavi z vrednostjo meritve v naslednjem časovnem obdobju. Sledi statistična metoda, ki meri razpršenost podatkov, in sicer standardni odklon. Asimetrija (ang. *skewness*) in sploščenost (ang. *kurtosis*) ugotavljata, koliko so podatki simetrični glede na povprečje in sploščeni v primerjavi z normalno porazdelitvijo. Visoka vrednost asimetrije pomeni, da podatki niso simetrični, visoka vrednost sploščenosti pa pomeni, da je prisotno veliko število ti. osamelcev (ang. *outlier*). V primeru nizkih vrednosti velja ravno nasprotno, torej pri

Tabela 5: Statistični funkcionali, izpeljani nad vhodnim vektorjem $x = [x_1, x_2, \ldots, x_n]$, kjer n predstavlja število časovnih obdobij, x_i pa meritev v i-tem časovnem obdobju.

Opis	**Izračun**	**Ime**
Najvišja vrednost s pripadajočim indeksom	$\max x_i$	max; max-index
Najnižja vrednost s pripadajočim indeksom	$\min x_i$	min; min-index
Razpon	$\max x_i - \min x_i$	range
Povprečje	$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$	mean
Povprečna absolutna sprememba	$\frac{1}{N}\sum_{i=1}^{N-1} \lvert x_i - x_{i+1}\rvert$	aac
Standardni odklon	$\sqrt{\frac{\sum_{i=1}^{N}(x_i-\bar{x})^2}{N-1}}$	stdev
Asimetrija	$\frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^3}{\left(\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2}\right)^3}$	skewness
Sploščenost	$\frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^4}{\left(\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2\right)^2}$	kurtosis
Križno povprečje	*št. križanj od povprečja*	crossmean
25-percentil	$n = \lceil \frac{25}{100} \cdot N \rceil$	25prctile
50-percentil	$n = \lceil \frac{50}{100} \cdot N \rceil$	50prctile
75-percentil	$n = \lceil \frac{75}{100} \cdot N \rceil$	75prctile

asimetriji so podatki simetrični, pri sploščenosti pa nimamo osamelcev.

Naslednji statistični funkcional je križno povprečje. Kot primer si vzemimo sliko 1, kjer je prikazan vektor meritev BAI indeksa na določenem poligonu, kjer x os predstavlja časovno obdobje (9 meritev, zato 9 točk), medtem ko y os predstavlja vrednost meritve v posameznem časovnem obdobju. Križno povprečje beleži, kolikokrat signal preseka povprečno vrednost, ki je v tem primeru enaka 0.2 (rdeča premica na sliki). V konkretnem primeru na sliki lahko razberemo, da je vrednost križnega povprečja enaka 3, saj signal prestopi povprečno vrednost med 3. in 4., med 4. in 5. ter med 6. in 7. časovnim obdobjem.



Slika 1: Primer vektorja točk, nad katerim se izmeri križno povprečje.

Zadnji trije funkcionali predstavljajo 25-percentil, mediano (oz. 50-percentil) ter 75-percentil. Skupno imamo izpeljanih 14 statističnih funkcionalov. Ker smo vseh 14 statističnih funkcionalov aplicirali nad vsakim parametrom, vključno z deltami, skupno torej nad 22-imi parametri, imamo skupno 308 atributov v podatkovni zbirki. Za lažje razumevanje nadaljnje interpretacije je v tabeli 6 definirana notacija atributov za novo podatkovno zbirko. Najprej je zapisanih 11 atributov v zaporedju funkcional_parameter, kjer se poleg določenega statističnega funkcionala zapišejo vsi izmerjeni parametri (z izjemo delte). Temu sledi naslednjih 11 atributov v zaporedju funkcional_parameter_delta, kjer so za posamezen statistični funkcional in parameter zapisane še vrednosti parametra delta. Tako imamo za posamezen statistični funkcional skupno 22 atributov. Nato se postopek ponovi s preostalimi statističnimi funkcionali.

3.3 Neposredno modeliranje časovnih vrst

Pri tem pristopu izpeljemo klasičen postopek modeliranja nad pridobljenimi signali - to je množica parametrov, pridobljenih iz meritev na posamezni točki. Vsakemu parametru dodamo še dinamično značilko, imenovano delta, ki beleži razliko posameznega parametra med dvema sosednjima meritvama. Ker imamo 11 vhodnih parametrov, s tem pridobimo 22 značilk. Vsaki značilki je potrebno dodati še meritve iz 9-ih časovnih

Tabela 6: Prikaz notacije zapisanih atributov v podatkovni zbirki izpeljanih časovnih značilk.

Ime atributa
max_ari-1
max_bai
...
max_ari-1_delta
max_bai_delta
...
max-index_ari-1
...
75-prctile_rededge_delta

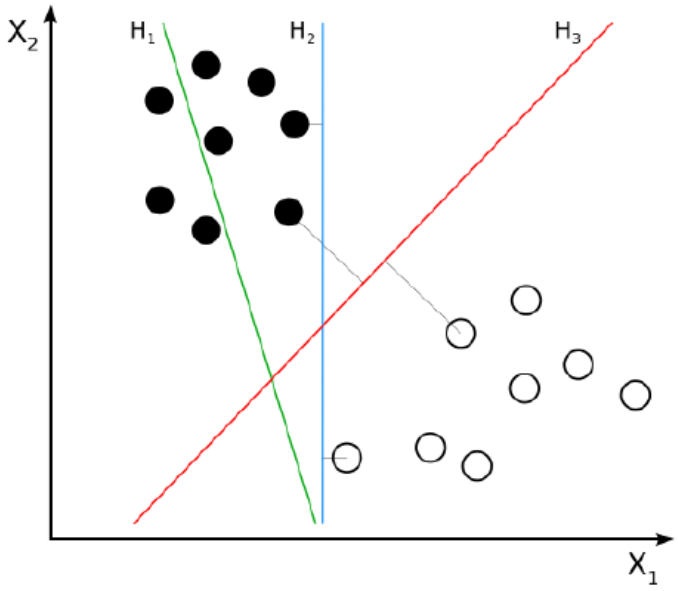
obdobij, tako da imamo skupno 198 značilk.

3.4 Metoda podpornih vektorjev

Modeliranje smo izvedli z algoritmom SVM (ang. *Support Vector Machine*) oziroma metodo podpornih vektorjev [30]. Gre za algoritem, ki določi tolikšno mejo med dvema razredoma, da dosežemo maksimalni razmik med vzorci [27]. Izhodišče za metodo SVM je učna množica meritev, za katere vemo, kateremu razredu (v našem primeru poljščini) pripadajo. Vsako meritev predstavimo z vektorjem v vektorskem prostoru, z metodo SVM nato poiščemo v tem večdimenzionalnem prostoru hiperravnino, ki ločuje meritve iz različnih razredov. Razdaljo vektorjev, ki se nahajajo najbližje hiperravnini, poskušamo maksimirati, namreč večja kot je razdalja praznega območja med razredi, toliko bolj natančno deluje klasifikacija novih meritev [28]. Algoritem hkrati za posamezen atribut določi še utež w, tako da skupek uteži v bistvu definira vektor na hiperravnini. Višja kot je vrednost uteži, bolj je atribut pomemben za model, na ta način pridobimo še dodatno informacijo o pomembnosti ter vplivu posameznih atributov.

Oglejmo si primer na sliki 2. Vzorce imamo razporejene v dva razreda, črni ter beli krogi. Tri premice, H_1, H_2 ter H_3, predstavljajo možne primere hiperravnine, s tem da premica H_1 ne ločuje razredov, premica H_2 sicer ločuje razrede, a razdalja med razredi ni optimalna (povzroči slabšo natančnost pri klasifikaciji), medtem ko H_3 predstavlja optimalno ločitev razredov, saj je postavljena tako, da ima največjo možno razdaljo

med vzorcema različnih razredov.



Slika 2: Primer delovanja SVM algoritma [30].

V praksi se v veliki večini primerov pojavi problem, da vseh podatkov ni možno ločiti z linearno mejo. Rešitev predvideva definicijo ti. "mehke meje", s katero ločimo večino točk v prostoru, hkrati pa sprejmemo nek vnaprej dogovorjen delež napačno klasificiranih primerov. Zato izrazu pri metodi SVM dodamo novo spremljivko z določeno utežjo, ki počne ravno to - to utež imenujemo parameter C.

Včasih pa niti linearna mehka meja ne zadostuje. Rešitev takega problema predvideva uporabo nelinearnih transformacij, kjer si pomagamo z uporabo različnih funkcij, s katerimi preslikamo točke iz prostora razvrščanja v drug prostor, kjer je točke možno linearno ločevati, hkrati pa se še vedno ohranja enak skalarni produkt. Definicija preslikave:

$$K(x, y) = \langle \Phi(x), \Phi(y) \rangle ,$$

kjer je $K(x, y)$ točka, ki s pomočjo funkcije Φ izpelje nelinearno transformacijo. Takim funkcijam pravimo jedra. Čeprav točke preslikamo v drug prostor, v tem prostoru še vedno uporabljamo skalarni produkt, zato lahko uporabimo enako metodologijo za izračun hiperravnine in podpornih vektorjev. Kot primer jedra je smiselno omeniti funkcijo RBF (ang. *Radial Basis Function*) [19].

V našem primeru smo izbrali pristop modeliranja z linearno mehko mejo. Obstaja verjetnost, da bi z uporabo drugih funkcij oz. algoritmov morda dosegli boljše rezultate.

Med drugimi možnimi metodami lahko izpostavimo linearno regresijo [2] ter GMM oz. model mešanice Gaussovih porazdelitev (ang. *Gaussian mixture model*) [7]. A potrebno je poudariti, da cilj celotnega magistrskega dela ni iskati algoritem, ki bo najbolj natančen, temveč primerjava med obema pristopoma, torej med neposrednim modeliranjem in modeliranjem nad izpeljanimi značilkami.

SVM ponuja nastavitev številnih parametrov, ki vplivajo na delovanje. Priporočljivih vrednosti, ki bi enako dobro delovale na vseh modelih ni, zato je potrebno poiskati optimalne vrednosti, s katerimi bomo pridobili najboljše rezultate. Orodje RapidMiner omogoča ravno tak postopek optimizacije parametrov, operator se imenuje "Optimize Parameters", ki za definiran podproces (v našem primeru je to model za klasifikacijo posamezne poljščine) in razpon parametrov pridobi natančnosti klasifikacij modelov za vse možne kombinacije vrednosti izbranih parametrov. Taka metoda optimizacije se imenuje GridSearch. Recimo, da izberemo 3 parametre ter 10 različnih vrednosti za posamezen parameter. Ker proces preizkusi natančnosti za vse možne kombinacije, bi v tem primeru imeli skupno $10 \times 10 \times 10 = 1000$ iteracij. Rezultat procesa je tabela parametrov ter natančnost modela z izbranim kriterijem, v našem primeru smo se omejili na optimizacijo parametra C pri algoritmu SVM, medtem ko smo kot kriterij natančnosti izbrali AUC (kriterij bo podrobneje opisan v sekciji 3.7). Rezultat procesa optimizacije pri modelu detekcije koruze je prikazan v tabeli 7. Zaradi obsežnih podatkov in hitrejše obdelave smo se omejili zgolj na optimizacijo parametra C ter izvedli 11 iteracij z razponom vrednosti parametra med 0.001 in 100 [32] - omenjene vrednosti priporoča RapidMiner. Videti je, da se natančnost modela rahlo izboljšuje z višanjem parametra do vrednosti 0.1, z uporabo višjih vrednosti pa natančnost močno upade. Podobne rezultate smo pridobili tudi pri enakem postopku optimizacije parametrov za detekcijo drugih poljščin, zato smo prišli do zaključka, da je najbolj optimalno uporabiti vrednost parametra $C = 0.1$.

3.5 Priprava učne množice za modeliranje

V fazi predobdelave podatkov smo zaznali občutno neuravnoteženost med pozitivnimi ter negativnimi klasifikacijami znotraj posamezne poljščine. Tabela 8 prikazuje razporeditev meritev tipov poljščin. V drugem stolpcu je zapisana dejanska porazdelitev v učni množici. Razmerje pri koruzi in pšenici je še v mejah sprejemljivega, pri preostalih pa je neuravnoteženost bolj opazna, posebej pri tritikali (razmerje 1 : 130). Obstaja velika verjetnost, da bo zaradi tega dosežena slabša kvaliteta modela, kot če bi bilo razmerje bolj uravnoteženo [6]. Iz tega razloga smo za vsak tip poljščine pripravili po dve prilagojeni učni množici na način, da smo na prvi učni množici dosegli

Tabela 7: Rezultati procesa optimizacije parametrov pri detekciji koruze.

Iteracija	**C parameter**	**AUC**
1	0.001	0.819
2	0.003	0.827
3	0.010	0.833
4	0.032	0.834
5	0.100	0.895
6	0.316	0.551
7	0.999	0.744
8	3.162	0.412
9	10.00	0.412
10	31.62	0.412
11	100.0	0.412

razmerje 1 : 3 (3 negativne meritve na eno pozitivno), na drugi pa 1 : 1 (število pozitivnih ter negativnih meritev je enako) - količine so razvidne v tretjem in četrtem stolpcu.

Omenjena razmerja učnih množic smo dosegli tako, da smo zajeli ustrezen delež meritev iz izvorne podatkovne zbirke (619644 poligonov). V nekaterih primerih (npr. tritikala) niti ta način ni zadostoval za dosego razmerja 1 : 1, saj je bilo enostavno premalo pozitivnih meritev tega tipa. Težavo smo odpravili z uporabo metode bootstrapping, ki naključno podvaja obstoječe pozitivne meritve, dokler ne dosežemo želenega razmerja.

Vsebino testne množice smo namenoma pustili nespremenjeno ne glede na tip poljščine (razmerje je enako učni množici z dejansko porazdelitvijo signalov), samo na ta način se lahko prepričamo, ali bi model, naučen na modificirani učni množici, učinkovito deloval tudi na realnih podatkih.

Poglavitna prednost neposrednega modeliranja časovnih vrst je v tem, da je proces dokaj enostaven in zahteva uporabo osnovnih tehnik modeliranja, učinkovit je v primeru zmernega števila značilk in časovnih meritev. Ravno slednje pa lahko predstavlja pomembno pomanjkljivost, saj se v primeru večjega števila časovnih meritev občutno poveča tudi število značilk. Konkretno, v našem primeru smo iz 11-ih parametrov in 9-ih časovnih meritev izpeljali 198 značilk. Hipotetično, če bi želeli zajeti meritve iz treh sezon, bi lahko imeli 27 časovnih meritev, število izpeljanih značilk pa bi zraslo na 594, kar je že težko obvladljivo, poleg tega model po vsej verjetnosti ne bi zmogel

Tabela 8: Razporeditev meritev tipov poljščin v razvojni podatkovni množici v zaporedju false / true. V oklepaju pri dejanskih porazdelitvah je zapisano razmerje vrednosti razrednega atributa.

Tip poljščine	**Učna množica dejanska porazdelitev**	**Učna množica razmerje** 1:3	**Učna množica razmerje** 1:1
Ječmen	42.732 / 2.203 (1 : 19)	29.928 / 9.808	21.377 / 19.617
Koruza	38724 / 6211 (1 : 6)	30223 / 9900	21311 / 20419
Oljna buča	43878 / 1057 (1 : 41)	30696 / 10826	19733 / 20569
Oljna ogrščica	44022 / 913 (1 : 48)	33043 / 8778	22028 / 18434
Pšenica	40406 / 4529 (1 : 9)	32232 / 9288	22160 / 18576
Tritikala	44593 / 342 (1 : 130)	31227 / 9735	22305 / 19470
Tip poljščine	**Testna množica**		
Ječmen	14289 / 691 (1 : 20)		
Koruza	12942 / 2038 (1 : 6)		
Oljna buča	14626 / 354 (1 : 41)		
Oljna ogrščica	14682 / 298 (1 : 49)		
Pšenica	13432 / 1548 (1 : 9)		
Tritikala	14865 / 115 (1 : 129)		

zaključiti izračunov v časovno sprejemljivem okvirju. Poleg tega predstavlja dodaten problem sam proces modeliranja, namreč pripravljen model deluje le na fiksno število značilk. V primeru dodajanja novih značilk (npr. meritev v novem časovnem obdobju), je potrebno pripraviti popolnoma nov model z novim procesom optimiziranja parametrov itd. Rešitev za omenjen problem bo predstavljena v nadaljevanju.

3.6 Modeliranje časovnih vrst s pomočjo časovnih značilk

Druga tehnika modeliranja predstavlja drugačen pristop. Namesto, da se izvedejo algoritmi za učenje nad neposrednimi značilkami, kot je opisano v prejšnji sekciji, se na podlagi izvornih meritev 11-ih parametrov v 9-ih časovnih obdobjih izpelje nove značilke.

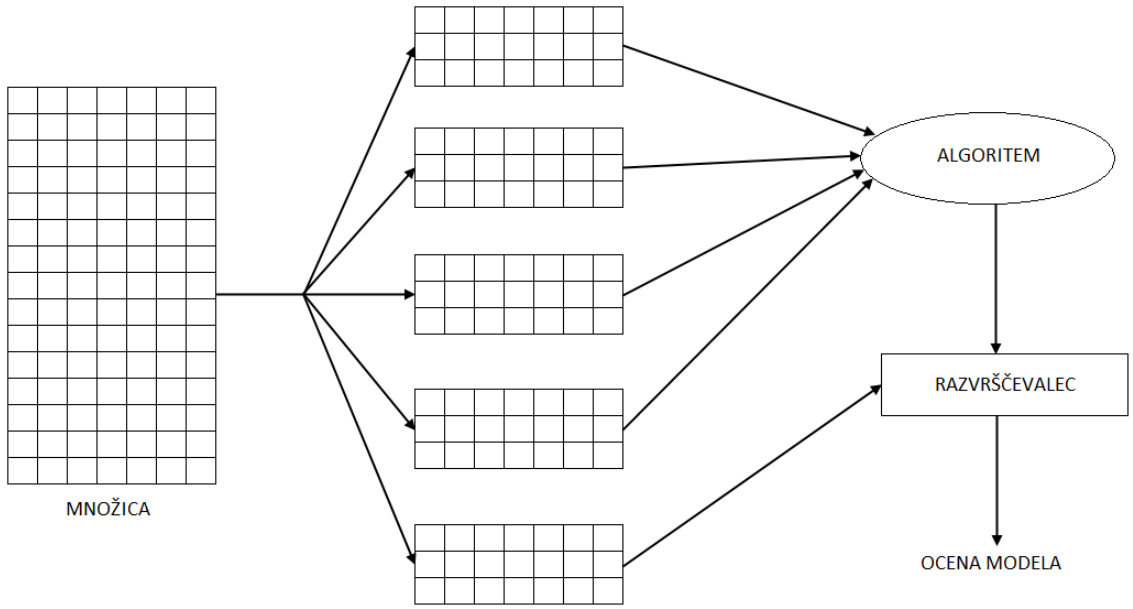
Postopek modeliranja nad izpeljanimi časovnimi značilkami je enak, kot je predstavljeno pri neposrednem modeliranju značilk, razlikuje se le podatkovna množica. Sedaj imamo še vedno 22 vhodnih parametrov (11 parametrov s pripadajočimi deltami), to-

krat pa smo za vsak parameter izračunali 14 statističnih funkcionalov, naštetih v tabeli 5. Skupno imamo torej 308 atributov. Tudi tokrat smo uporabili metodo SVM, glede optimizacije parametrov pa se izkaže, da je uporaba vrednosti parametra $C = 0.1$ najbolj optimalna izbira. Prav tako smo izpeljali modeliranje nad tremi različnimi učnimi množicami, saj smo tudi v tem primeru razpolagali z neuravnoteženimi podatki.

3.7 Evalvacija postopkov modeliranja

Validacijo učne množice smo izvedli z metodo 10-kratnega križnega preverjanja (ang. *10-fold cross validation*) [5], evalvacijo pa smo izvedli nad testno množico – gre za ločeno podatkovno množico, ki ni bila uporabljena v procesu učenja.

Metoda k-kratnega križnega preverjanja deluje tako, da podatkovno množico z N instancami razdelimo na k enakih podmnožic, kjer k običajno ni večji od 10. Na sliki 3 se nahaja primer za $k = 5$. Prvih $k - 1$ podmnožic uporabimo kot učno množico, nad katero izvajamo algoritem za učenje (v našem primeru SVM), k-to podmnožico pa kot testno množico, na kateri razvrstimo instance na podlagi prej naučenega modela. V zadnjem koraku ocenimo natančnost modela z izbranim kriterijem. Postopek ponovimo k-krat, s tem da v vsakem koraku zamenjamo testno množico. Končno stopnjo napake dobimo s povprečenjem vseh iteracij.

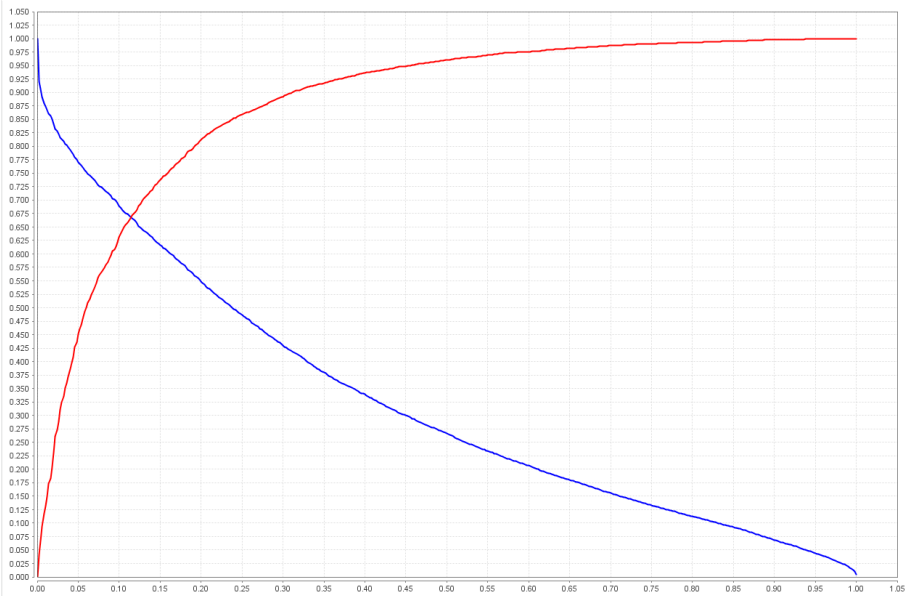


Slika 3: Delovanje križnega preverjanja v i-tem koraku, kjer $1 \leq i \leq k$.

Kot kriterij kakovosti modelov smo izbrali ploščino AUC (ang. *Area Under the Curve*)

pod ROC krivuljo (ang. *Receiver Operating Characteristics).* Gre za dvodimenzionalen graf, kjer x os predstavlja delež lažnih pozitivnih primerov (ang. *False Positive Rate*), y os pa delež resničnih pozitivnih primerov (ang. *True Positive Rate*) [9,10]. Na sliki 4 je z rdečo barvo označena ROC krivulja, z modro barvo pa krivulja, ki predstavlja prag (ang. *threshold*), ki določa mejno vrednost napovedi verjetnosti pozitivnega razreda. Če se napoved nahaja nad pragom, jo označimo kot pozitivno, sicer negativno [3]. Ordinatna os predstavlja senzitivnost, ki določi delež pravilno razvrščenih pozitivnih napovedi, medtem ko abscisna os predstavlja $1-$specifičnost, kjer specifičnost pomeni delež pravilnih negativnih napovedi.

Težava je v tem, da v primeru dveh ROC krivulj ni možno kvantitativno primerjati razlike med njima, temveč zgolj vizualno. Da bi lahko pridobili neko natančno definirano, številsko razliko, je dodatno vpeljan parameter AUC, ki izračuna ploščino pod ROC krivuljo. Vrednost AUC bo vedno v razponu med 0 in 1 (v nadaljevanju bomo označevali razpon med $0-100\%$ površino), saj računa ploščino na enotskem kvadratu. V praksi pomeni, da mora vrednost AUC stremeti k 100% pokritosti površine, vsekakor pa mora biti vsaj strogo večja od 50%, saj gre v nasprotnem primeru za neustrezen model, saj bi v takem primeru dosegli enako natančnost, kot če bi naključno določali pripadnost razredu.



Slika 4: Primer ROC krivulje.

3.8 Primer modela

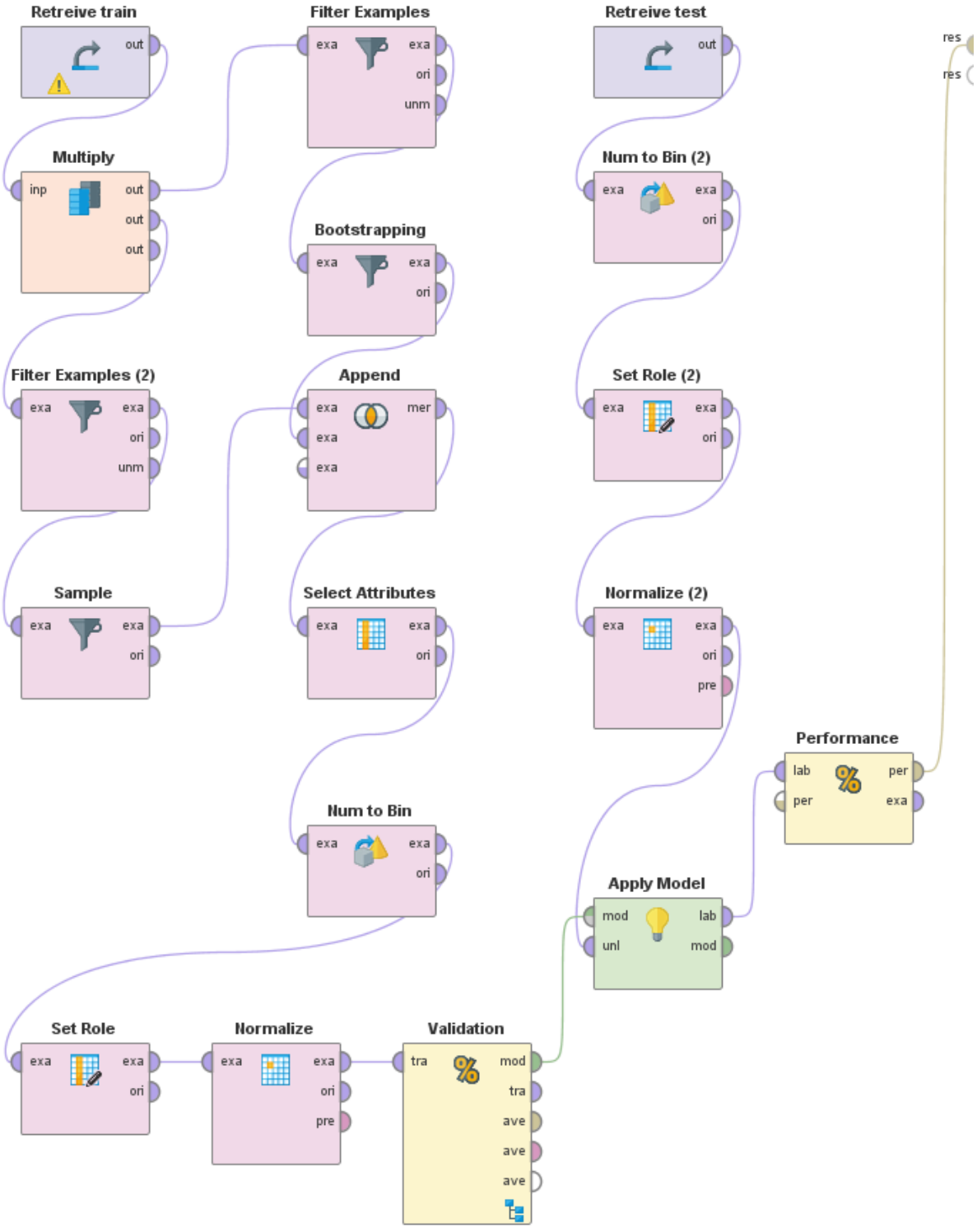
Za lažje razumevanje celotnega procesa modeliranja, si oglejmo primer na sliki 5, ki prikazuje model, zgrajen v RapidMiner. Konkretno, gre za primer detekcije tritikale. Enak model se aplicira tako pri neposrednem kot posrednem načinu modeliranja, razlikujeta se le vhodna učna ter testna množica znotraj operatorjev Retreive train ter Retreive test.

Proces poteka od leve proti desni, v prvem koraku uvozi učno množico, katero kasneje podvoji (operator multiply), da lahko ločimo pozitivne ter negativne meritve (operator filter examples). Cilj tega koraka je pridobiti razvojno učno množico z razmerjem 1:3, kot navedeno v tabeli 8. Za negativne meritve zadostuje zgolj zajem manjšega deleža naključno izbranih negativnih meritev (operator sample). Pri pozitivnih meritvah se pojavi težava, ker je tudi v učni množici premalo zajetih pozitivnih meritev, zato je potrebno uporabiti operator bootstrapping, ki naključno podvaja obstoječe pozitivne meritve do želenega razmerja.

Sledi operator append, ki združi pozitivne ter negativne meritve v eno podatkovno množico. Zatem izločimo razredne atribute (operator select attributes), katerih ne obravnavamo, kot so ječmen, pšenica itd. Tako nam preostane razredni atribut tritikala, katerega je potrebno pretvoriti iz numeričnega v binomski zapis (operator num to bin). Z operatorjem set role eksplicitno definiramo atribut tritikala kot razredni atribut. Za zaključek predpriprave učne množice je potrebno vrednosti še normalizirati (operator normalize).

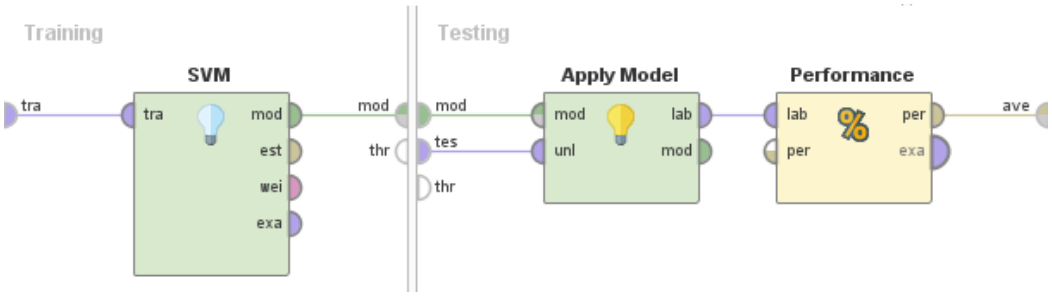
Nad pripravljeno učno množico izpeljemo postopek validacije, in sicer 10-kratno navzkrižno preverjanje (operator validation). Operator vsebuje ti. podproces, ki je definiran na sliki 6 (podproces bo opisan v nadaljevanju). Po zaključenem postopku imamo pripravljen model, kateremu je potrebno ovrednotiti natančnost. V naslednjem koraku preberemo testno množico (operator retreive test) ter jo pripravimo za evalvacijo (izbira razrednega atributa ter sprememba v binarni zapis, normalizacija vrednosti). Z operatorjem apply model, naučen model apliciramo na novih, neznanih meritvah (testna množica) ter izmerimo natančnost s kriterijem AUC (operator performance).

Oglejmo si še sliko 6, ki predstavlja podproces validacije 10-kratnega navzkrižnega preverjanja. Leva polovica podprocesa je namenjena pripravi procesa nad učno množico, desna polovica pa nad testno množico. V našem primeru se v levi polovici nahaja zgolj operator SVM, ki izpelje modeliranje z istoimenskim algoritmom nad učno množico.



Slika 5: Primer izgradnje modela v RapidMiner.

Model se nato prenese v desno polovico, kjer je prostor za pripravo procesa nad testno množico. V našem primeru smo model aplicirali na testni množici ter izmerili natančnost z AUC. Opisan podproces se ponovi k-krat (parameter k se določi pred



Slika 6: Validacija v RapidMiner.

izvedbo procesa), v našem primeru se je ponovil 10-krat.

4 REZULTATI IN RAZPRAVA

Kot že prej omenjeno, smo vsa modeliranja izvedli z metodo SVM, z 10-kratnim križnim preverjanjem. Evalvacijo modela smo izvedli nad testno množico. Kot mero natančnosti smo izbrali AUC ploščino pod ROC krivuljo. Rezultati AUC za posamezno učno množico pri neposrednem modeliranju časovnih vrst se nahajajo v tabeli 9, medtem ko se rezultati AUC pri modeliranju izpeljanih značilk nahajajo v tabeli 10. Pri obeh primerih se rezultati nanašajo na testno množico. V splošnem se izkaže, da je model, zgrajen nad učno množico v razmerju 1:3 (3 negativni vzorci na 1 pozitivnega) najbolj natančen, podobna natančnost se doseže tudi z razmerjem 1:1. Model z učno množico z dejansko porazdelitvijo se izkaže za najmanj natančnega, razlog je pripisati premajhnemu številu pozitivnih vzorcev, nad katerimi bi se model lahko naučil (razmerja dejanske porazdelitve so zapisana v tabeli 8). Izjema je model za klasifikacijo ne-poljščine – ta je dosegel sprejemljivo natančnost z dejansko porazdelitvijo učne množice, saj je le-ta že zgrajena v razmerju 1:1.

Tabela 9: Prikaz natančnosti klasifikacije neposrednega modeliranja časovnih značilk na testni množici, izražene z AUC.

Tip poljščine	**Učna množica dejanska porazdelitev**	**Učna množica razmerje** 1:3	**Učna množica razmerje** 1:1
Ječmen	73.0%	85.9%	53.9%
Koruza	88.9%	89.5%	89.3%
Oljna buča	82.3%	90.1%	90.1%
Oljna ogrščica	85.7%	92.7%	92.6%
Pšenica	88.5%	89.6%	90.1%
Tritikala	69.4%	89.3%	56.4%
Ne-poljščina	88.0%	/	/

Tudi na podlagi tabele 10 se izkaže, da je v tem primeru najbolj ustrezen model, zgrajen nad učno množico z razmerjem 1:3, medtem ko dejanska porazdelitev povzroči slabšo natančnost, a predstavlja bolj podobno razmerje v primerjavi z realnim stanjem podatkovne zbirke. Iz tega razloga smo se odločili, da bomo pri vseh nadaljnjih postopkih

Tabela 10: Prikaz natančnosti klasifikacije izpeljanih časovnih značilk na testni množici, izražene z AUC.

Tip poljščine	**Učna množica dejanska porazdelitev**	**Učna množica razmerje** 1:3	**Učna množica razmerje** 1:1
Ječmen	66.5%	83.6%	63.5%
Koruza	85.0%	86.4%	87.3%
Oljna buča	70.6%	86.6%	86.6%
Oljna ogrščica	82.9%	89.4%	90.2%
Pšenica	80.6%	83.9%	85.1%
Tritikala	64.3%	84.4%	83.6%
Ne-poljščina	87.3%	/	/

modeliranja za učno množico uporabili množico z razmerjem 1:3.

Tabela 11 prikazuje razliko v natančnosti AUC ploščine med neposrednimi meritvami in izpeljanimi značilkami. Videti je, da so razlike v nekaterih primerih minimalne, kot lahko vidimo na primerih detekcije ne-poljščine, ječmena ali oljne ogrščice. Pri preostalih primerih se rezultati razlikujejo do 5% v korist modeliranja z neposrednimi meritvami. Čeprav se je model z izpeljanimi značilkami v splošnem izkazal za slabšega, še vedno dosega sprejemljivo napovedno moč.

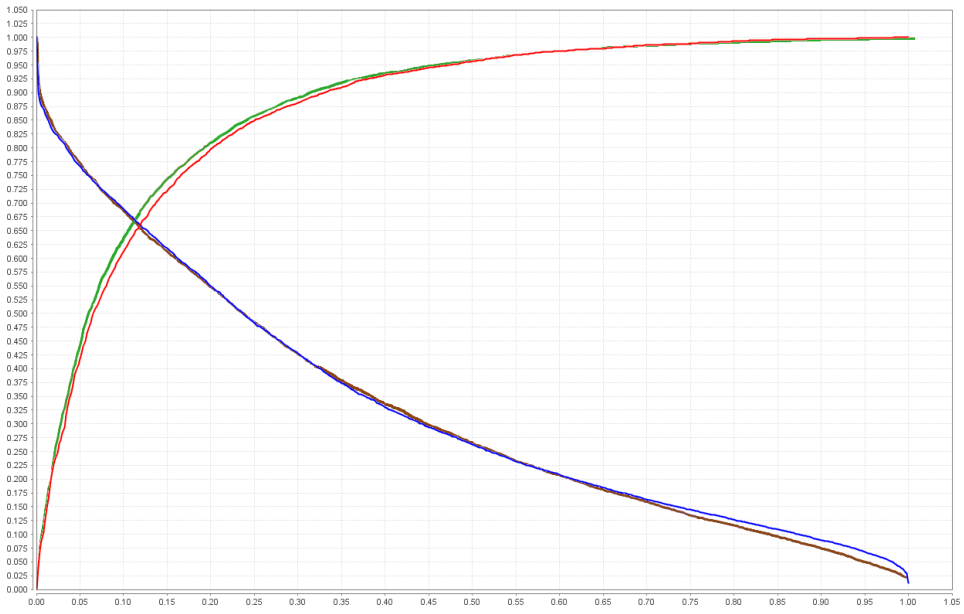
Tabela 11: Primerjave natančnosti klasifikacij posamezne poljščine pri obeh pristopih modeliranja.

Tip poljščine	**Neposredne meritve AUC**	**Izpeljane značilke AUC**	**Razlika**
Ječmen	85.9%	83.6%	2.3%
Koruza	89.5%	83.9%	5.6%
Oljna buča	90.1%	86.6%	3.5%
Oljna ogrščica	92.7%	89.4%	3.3%
Pšenica	89.7%	83.9%	5.8%
Tritikala	89.4%	84.4%	5%
Ne-poljščina	88.1%	87.3%	0.8%

4.1 Rezultati detekcije poljščin

V nadaljevanju bodo analizirani rezultati modelov detekcije posameznih poljščin, katere značilke prispevajo največji delež oz. imajo največjo utež pri rezultatu, poskusili bomo tudi raziskati razlog zakaj je temu tako. Vsi navedeni postopki so predstavljeni v prejšnjem poglavju, sedaj bodo analizirani zgolj rezultati.

Prvi model je namenjen detekciji ne-poljščine. Natančnost modela je v obeh primerih praktično enaka, in sicer pri neposrednem modeliranju je ploščina AUC enaka 88.1%, medtem ko je pri modeliranju izpeljanih značilk AUC enak 87.3%. V obeh primerih se rezultati nanašajo na testno množico. ROC krivulje s pripadajočimi pragovi se nahajajo na sliki 7. Z rdečo barvo je označena krivulja, ki predstavlja ROC za model z izpeljanimi značilkami, z modro barvo pa krivulja, ki določa prag oz. mejno vrednost napovedi verjetnosti pozitivnega razreda. Zelena barva je v tem primeru sicer manj vidna, ker se v večini prekriva z rdečo krivuljo, predstavlja pa ROC izpeljanih značilk. Podobno velja za krivuljo rjave barve, ki določa prag pozitivnega razreda za izpeljane značilke, hkrati pa se v veliki večini prekriva z modro krivuljo. Na podlagi grafa je videti, da sta si modela praktično enakovredna.



Slika 7: Primerjava ROC krivulj obeh pristopov za detekcijo ne-poljščine.

Da bi bolje razumeli pridobljene rezultate, si oglejmo še uteži atributov, ki jih je določil SVM algoritem glede na razredni atribut na učni množici. Tabela 12 prikazuje uteži prvih 5 najbolj informativnih značilk. Pri modelu z neposrednimi meritvami je višje

vrednosti opaziti pri parametrih BAI ter PSRI-NIR, navedenih na levi polovici tabele. Najbolj informativen parameter z najvišjo utežjo je BAI v mesecu juliju. Pri modelu z izpeljanimi značilkami sta najbolj informativna parametra PSRI-NIR in NDVI, kot je razvidno iz desne polovice tabele. Iz celotnega seznama uteži pri obeh modelih gre razbrati, da ima parameter PSRI-NIR velik vpliv na detekcijo ne-poljščine, posebej pri modelu z neposrednimi meritvami. Naslednji parametri se med seboj izmenjujejo po pomembnosti, nekoliko izstopata parametra NDVI ter zelena valovna dolžina v različnih mesecih. Pri izpeljanih značilkah ima poleg PSRI-NIR prav tako pomembno vlogo parameter NDVI, kar je tudi razumljivo, saj parameter predstavlja stopnjo vegetacije.

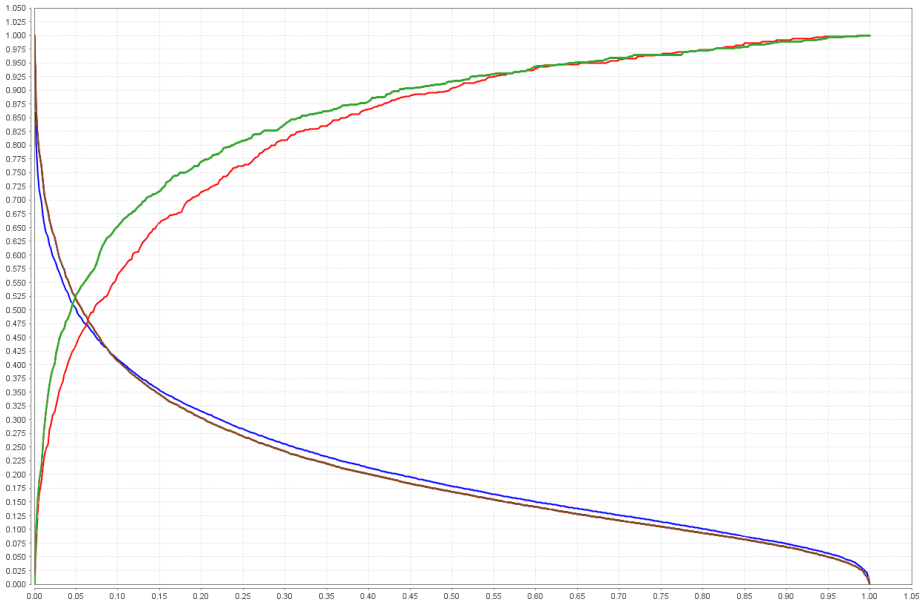
Tabela 12: Pregled uteži obeh pristopov za detekcijo ne-poljščine na učni množici.

Neposredne meritve		**Izpeljane značilke**	
2013-07-29_bai	0.457	75prctile_psri_nir	0.304
2013-07-29_psri-nir	0.451	mean_psri-nir	0.287
2013-05-18_green	0.417	25prctile_ndvi_delta	0.252
2013-10-08_psri-nir	0.345	stdev_ndvi_delta	0.251
2013-05-18_psri-nir	0.344	stdev_ndvi	0.238

Drugi model izvaja detekcijo ječmena. Tako kot pri vseh ostalih modelih, gre tudi v tem primeru za binarno klasifikacijo, kar pomeni, da model za vsak poligon zgolj detektira, ali gre za določen tip poljščine (v tem primeru ječmen) ali ne. AUC je pri neposrednem modeliranju enak 85.9%, pri modeliranju izpeljanih značilk pa je enak 83.6%. V obeh primerih se rezultati nanašajo na testno množico. Slika 8 prikazuje primerjavo ROC krivulj obeh tipov modeliranja. Tokrat je razlika bolj opazna, posledično je tudi razlika med AUC nekoliko višja, medtem ko je določen prag praktično identičen.

Vizualna razlika med ROC krivuljama je v tem, da je zelena krivulja (neposredne meritve) bolj strma v prvi polovici v primerjavi z rdečo krivuljo (izpeljane značilke). To pomeni, da model z neposrednimi meritvami doseže višji delež pravilno klasificiranih pozitivnih primerov ob nižjem deležu napačno klasificiranih pozitivnih primerov.

Oglejmo si, katere značilke so najbolj prispevale k pridobljenemu rezultatu. Tabela 13 prikazuje 5 značilk z najvišjo utežjo na učni množici. Pri neposrednih meritvah najbolj izstopata sprememba zelene ter modre valovne dolžine v mesecu juniju, kar gre pripisati dejstvu, da v tem obdobju ječmen običajno dozori [17], torej se barvna struktura



Slika 8: Primerjava ROC krivulj obeh pristopov za detekcijo ječmena.

poljščine najbolj spremeni.

Tabela 13: Pregled uteži obeh pristopov za detekcijo ječmena na učni množici.

Neposredne meritve		**Izpeljane značilke**	
2013-06-15_green_delta	0.541	aac_psri-nir_delta	0.325
2013-06-15_blue_delta	0.436	mean_ndvi-green_delta	0.291
2013-05-18_green	0.353	min_blue	0.276
2013-05-18_blue	0.319	mean_red_delta	0.206
2013-04-11_blue	0.305	range_ndvi	0.203

Pri izpeljanih značilkah se, razumljivo, informativnost parametrov drugače razporedi, saj podatkovna množica ne vsebuje neposrednih meritev v določenem mesecu, temveč le določen statistični funkcional za celotno obdobje meritev. Tako sta najbolj informativna povprečna absolutna sprememba PSRI-NIR parametra ter povprečna sprememba parametra NDVI-GREEN.

Tretji model izvaja detekcijo koruze. AUC je pri neposrednem modeliranju enak 89.5%, pri modeliranju izpeljanih značilk pa je enak 83.9%. V obeh primerih se rezultati nanašajo na testno množico. Slika 9 prikazuje ROC krivulji obeh modelov, kjer je tudi razvidna razlika. V primerjavi z modelom za detekcijo ječmena, je v tem primeru raz-

lika med krivuljama nekoliko višja v drugi polovici krivulje (pri ječmenu je bila razlika bolj izrazita v prvi polovici), kar pomeni, da je model za boljšo pravilno klasifikacijo pozitivnih primerov, porabil nižji delež napačno klasificiranih pozitivnih napovedi. Posledično je tudi AUC pri detekciji koruze višji za približno 3%, torej je model toliko bolj natančen.

Slika 9: Primerjava ROC krivulj obeh pristopov za detekcijo koruze.

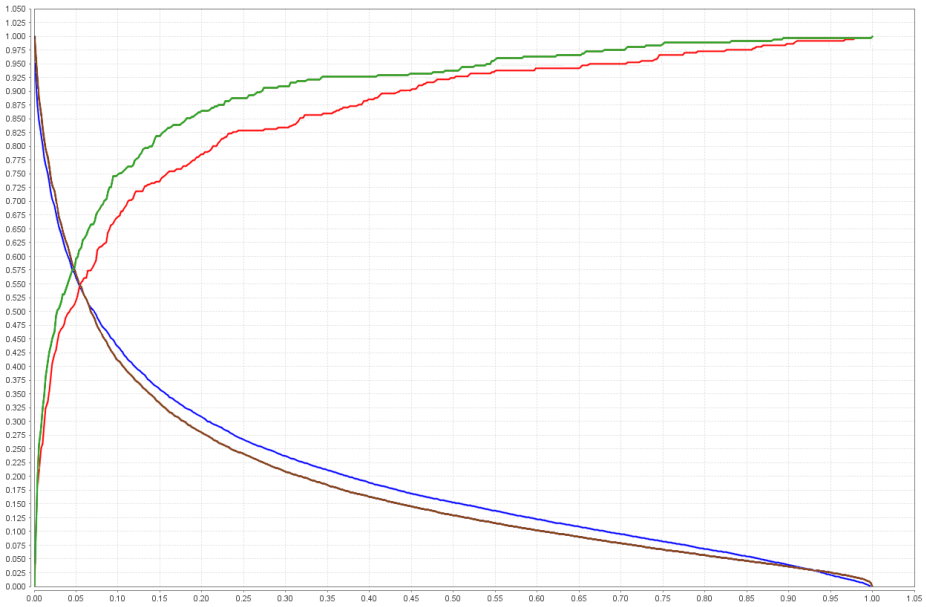
Najbolj informativne značilke tega modela so predstavljene v tabeli 14, uteži se nanašajo na učno množico. Pri neposrednih meritvah izrazito prevladuje parameter chlrededge ter sprememba vrednosti le-tega v poletnih mesecih. Pri izpeljanih značilkah je rezultat podoben, chlrededge spada med bolj informativne parametre. Iz pridobljenih uteži lahko sklepamo, da je razmerje vrednosti klorofila med rdečo in RedEdge valovno dolžino, najverjetneje v listih koruze, parameter z največjo utežjo.

Četrti model izvaja detekcijo oljne buče. AUC je pri neposrednem modeliranju enak 90.1%, pri modeliranju izpeljanih značilk pa je enak 86.6%. V obeh primerih se rezultati nanašajo na testno množico. Slika 10 prikazuje ROC krivulji obeh modelov. Tokrat je graf nekoliko bolj stopničaste oblike, razlog je v nižjem deležu pozitivnih primerov v testni množici, na katero namenoma nismo vplivali. Če opazujemo pragove na vseh grafih, opazimo, da ni bistvenih odstopanj pri obeh načinih modeliranja, medtem ko je razlika v natančnosti nekoliko bolj opazna.

Podobno kot pri modelu za detekcijo koruze, je tudi v primeru detekcije oljne buče

Tabela 14: Pregled uteži obeh pristopov za detekcijo koruze na učni množici, .

Neposredne meritve		**Izpeljane značilke**	
2013-07-02_chlrededge	0.512	25prctile_rededge	0.243
2013-07-29_chlrededge_delta	0.378	aac_nir_delta	0.243
2013-06-15_chlrededge_delta	0.368	min_chlrededge	0.229
2013-07-29_bai	0.337	mean_rededge	0.201
2013-07-29_ari-1	0.284	25prctile_chlrededge	0.199



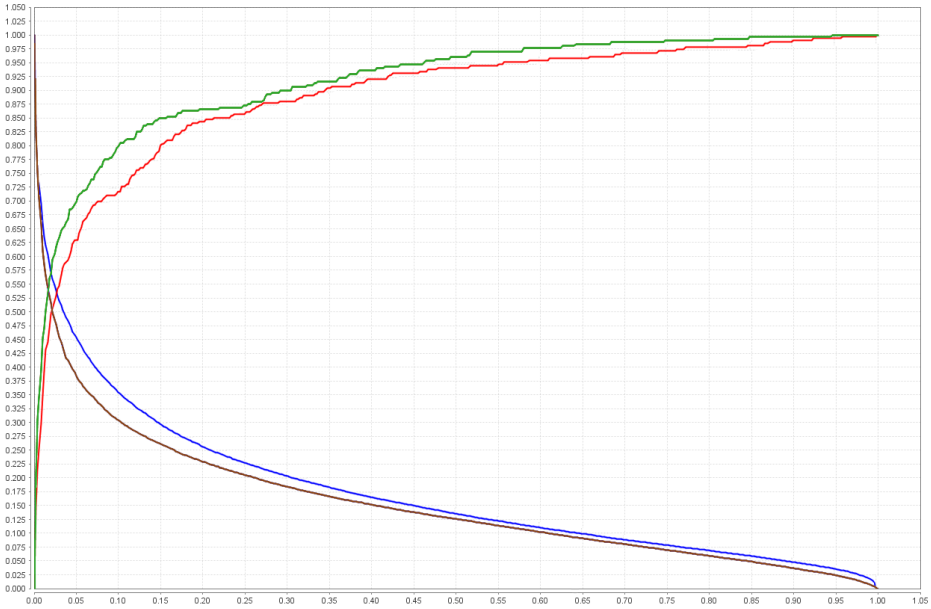
Slika 10: Primerjava ROC krivulj obeh pristopov za detekcijo oljne buče.

parameter chlrededge najbolj informativen na učni množici, saj pri neposrednih meritvah predstavlja prvo in tretjo značilko z najvišjo utežjo, pri izpeljanih značilkah pa kar prve tri, kot je razvidno iz tabele 15.

Peti model izvaja detekcijo oljne ogrščice. AUC je pri neposrednem modeliranju enak 92.7%, pri modeliranju izpeljanih značilk pa je enak 89.4%. V obeh primerih se rezultati nanašajo na testno množico. Oglejmo si sliko 11, ki prikazuje ROC krivulji obeh modelov. Razlika med grafoma je minimalna, saj oba modela že vsebujeta visoko vrednost AUC na učni množici, posledično se ROC krivulja izredno strmo pomika od leve proti desni, kar pomeni, da model doseže visoko natančnost klasifikacije ob minimalnem deležu napačno klasificiranih pozitivnih primerov.

Tabela 15: Pregled uteži obeh pristopov za detekcijo oljne buče na učni množici.

Neposredne meritve		**Izpeljane značilke**	
2013-07-29_chlrededge	0.473	min_chlrededge	0.405
2013-06-15_rededge_delta	0.415	mean_chlrededge	0.320
2013-07-02_chlrededge_delta	0.402	25prctile_chlrededge	0.310
2013-07-02_rededge	0.337	min_bai	0.280
2013-07-29_red	0.322	min_ndvi-green	0.266



Slika 11: Primerjava ROC krivulj obeh pristopov za detekcijo oljne ogrščice.

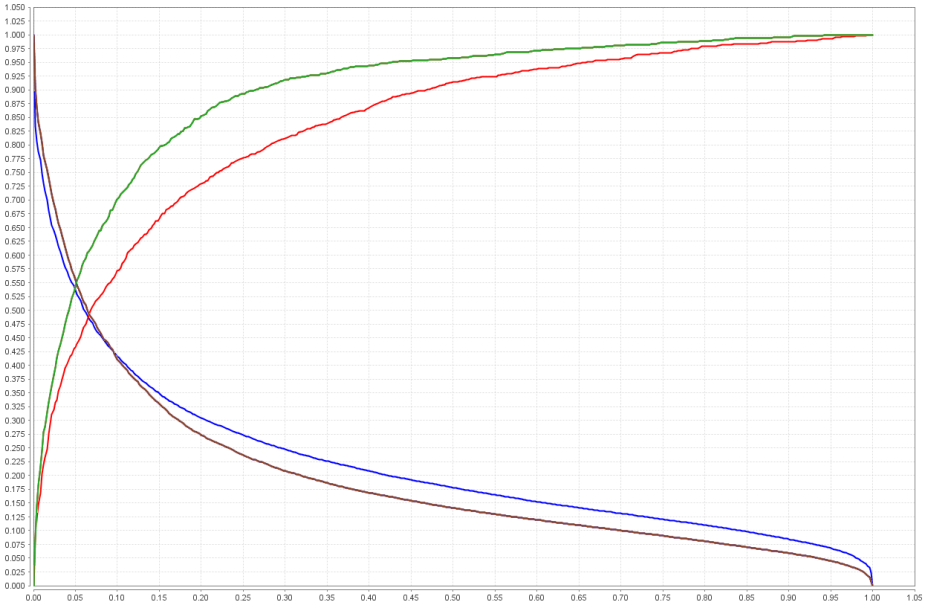
Ob pogledu na tabelo 16 (rezultati se nanašajo na učno množico) je opaziti ponavljajoči se trend glede najvišjega deleža uteži, torej parameter chlrededge je najbolj informativen pri obeh modelih. Pri neposrednih meritvah nekoliko izstopa še parameter NDVI v mesecu juniju, medtem ko vsi preostali parametri pri obeh modelih predstavljajo nižji delež, tako da posebnega vzorca ni bilo moč zaznati oz. razbrati.

Šesti model izvaja detekcijo pšenice. AUC je pri neposrednem modeliranju enak 89.7%, pri modeliranju izpeljanih značilk pa je enak 83.9%, kar predstavlja najvišjo razliko med vsemi znanimi tipi poljščin. V obeh primerih se rezultati nanašajo na testno množico. Kot sta prikazani ROC krivulji na sliki 12, je razlika v natančnosti očitna že na prosti pogled, saj je model z neposrednimi meritvami v bistveno manjšem deležu napačno klasificiral pozitivne primere, da bi dosegel pravilno klasifikacijo dejansko pozitivnih

Tabela 16: Pregled uteži obeh pristopov za detekcijo oljne ogrščice na učni množici.

Neposredne meritve		**Izpeljane značilke**	
2013-06-15_chlrededge	0.500	aac_chlrededge_delta	0.330
2013-06-15_ndvi	0.419	25prctile_rededge	0.287
2013-06-15_red	0.339	min_ari-1	0.280
2013-05-18_rededge	0.332	mean_red_delta	0.276
2013-07-29_nir_delta	0.330	75prctile_ndvi	0.268

primerov.



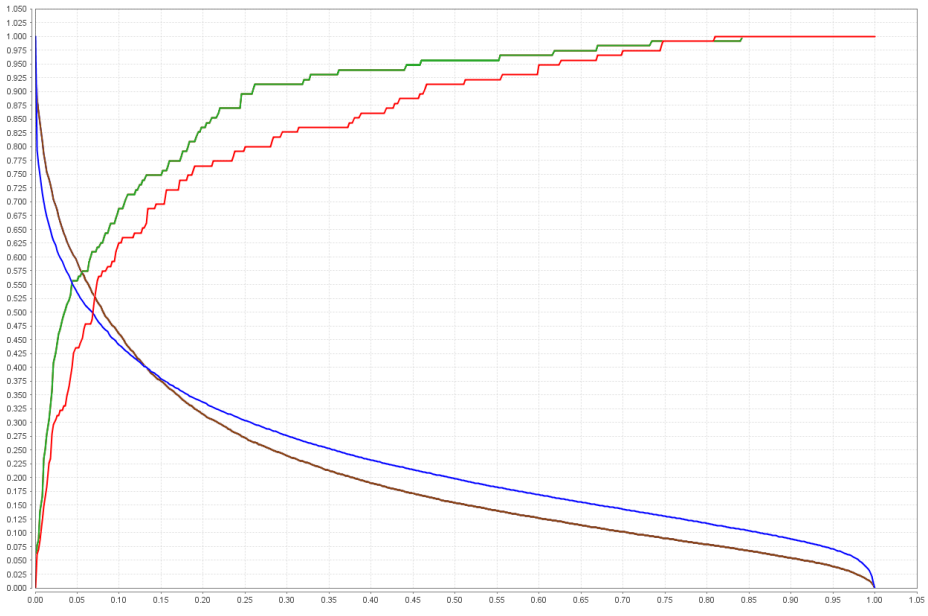
Slika 12: Primerjava ROC krivulj obeh pristopov za detekcijo pšenice.

Odgovor na vprašanje, kaj je prispevalo k tolikšni razliki v natančnosti modela, se nahaja v tabeli 17. Če si ogledamo vrednosti uteži, je pri izpeljanih značilkah možno razbrati, da ni nobenega parametra, ki bi posebej izstopal - sicer je PSRI-NIR najbolj informativen, a ne v tolikšni meri, saj recimo 75-percentil rdeče valovne dolžine (ki predstavlja peti najbolj informativen parameter) dosega primerljivo vrednost uteži. Pri neposrednih meritvah pa nekoliko bolj izstopa parameter chlrededge, delno tudi rdeča valovna dolžina. To pomeni, da omenjena parametra največ prispevata k doseganju optimalne natančnosti modela in predvsem razliki v primerjavi z modelom z izpeljanimi značilkami. V obeh primerih se uteži nanašajo na učno množico.

Tabela 17: Pregled uteži obeh pristopov za detekcijo pšenice na učni množici.

Neposredne meritve		**Izpeljane značilke**	
2013-06-15_chlrededge_delta	0.412	25prctile_psri-nir	0.294
2013-06-15_red_delta	0.379	25prctile_rededge	0.266
2013-07-02_red	0.311	mean_psri-nir	0.240
2013-05-18_rededge	0.305	75prctile_psri-nir	0.236
2013-07-29_bai	0.298	75prctile_red	0.236

Sedmi in hkrati tudi zadnji model izvaja detekcijo tritikale. AUC je pri neposrednem modeliranju enak 89.4%, pri modeliranju izpeljanih značilk pa je enak 87.8% V obeh primerih se rezultati nanašajo na testno množico. ROC krivulji na sliki 13 sta še bolj stopničasti iz enakega razloga kot pri detekciji oljne buče, torej manjše število pozitivnih primerov v testni podatkovni množici. Ne glede na to pa je rezultat enak, kot v vseh ostalih primerih, torej tudi v tem primeru je model z neposredno meritvijo bolj natančen v primerjavi z modelom z izpeljanimi značilkami, saj se ROC krivulja neposrednih meritev po celotnem grafu nahaja nad ROC krivuljo izpeljanih značilk.



Slika 13: Primerjava ROC krivulj obeh pristopov za detekcijo tritikale.

Pregled uteži v tabeli 18 prikaže podoben razlog za tolikšno odstopanje - pri izpeljanih značilkah ni parametra, ki bi bistveno odstopal z utežjo, torej vsi parametri so med seboj podobno informativni. Pri neposrednih meritvah tokrat najbolj izstopa parameter

NIR v mesecu juliju.

Tabela 18: Pregled uteži obeh pristopov za detekcijo tritikale.

Neposredne meritve		**Izpeljane značilke**	
2013-07-29_nir	0.364	stdev_red	0.249
2013-07-29_ari-1	0.298	stdev_blue	0.247
2013-07-29_chlrededge	0.297	aac_green_delta	0.233
2013-10-08_red	0.278	75prctile_chlrededge_delta	0.211
2013-06-15_chlrededge_delta	0.268	mean_chlrededge_delta	0.205

Če povzamemo pridobljene uteži značilk, sta pri detekciji ne-poljščine vsekakor najbolj informativna parametra NDVI ter PSRI-NIR, kar je razumljivo, saj indeks NDVI s svojo vrednostjo definira tip rastja - torej zgolj s tem parametrom je možno zavreči tisti delež poligonov, kjer rastje ni prisotno (npr. voda, skalnata območja itd.).

Pri detekciji preostalih 6-ih znanih poljščin prav tako pogosto najdemo parametra NDVI ter PSRI-NIR, sicer v manjši meri (nižje vrednosti uteži), saj v veliki večini prevladuje parameter ChlRedEdge z najvišjimi vrednostmi uteži. Razlog je predvsem v koncentraciji klorofila v posamezni poljščini [33], namreč večja kot je koncentracija klorofila, toliko večja je tudi vpojnost v rdeči valovni dolžini, kar povzroči slabši odboj. Zaradi vse večje vpojnosti se prisotnost klorofila razširi tudi v druge valovne dolžine proti NIR področju, kar posledično povzroči povečanje odboja, saj je odboj klorofila v NIR področju občutno večji.

Na podlagi pridobljenih uteži je smiselno analizirati, katere značilke so najbolj koristne in obratno, torej ali je morda katero izmed značilk bolj smiselno izločiti iz postopkov modeliranja. Opaziti je, da so pri praktično vseh modelih pri izpeljanih značilkah med najbolj koristnimi prisotne značilke, ki beležijo percentile posameznega parametra. Zato je take značilke vsekakor smiselno obdržati tudi pri nadaljnjih postopkih. Podobno velja za značilke, ki beležijo povprečje, standardni odklon, povprečno absolutno spremembo ter najnižjo vrednost posameznega parametra.

Na drugi strani je med najmanj koristnimi značilkami opaziti indekse najvišjih ter najnižjih vrednosti ter križno povprečje. Razloge gre pripisati prenizkem razponu vrednosti, poleg tega so vse vrednosti cela števila, kar pomeni, da so vrednosti v veliki večini primerov enake. Ravno iz tega razloga ni moč pričakovati, da bi omenjene

značilke bistveno vplivale na kakovost modela, zato bi bilo smiselno razmisliti o izločitvi omenjenih značilk pri nadaljnjih postopkih modeliranja.

Pri neposrednih meritvah je stanje podobno, le da tokrat prevladujejo značilke poletnih mesecev, medtem ko pri značilkah v mesecih april ter oktober ni zaznati večjih uteži. Iz tega lahko sklepamo, da bi bilo smiselno preveriti, ali bi izločitev omenjenih značilk bistveno vplivala na kvaliteto modela pri nadaljnjih postopkih modeliranja.

5 ZAKLJUČEK

Na podlagi pridobljenih rezultatov lahko zaključimo s trditvijo, da je model z neposrednimi značilkami bolj primeren za detekcijo vegetacije, če zajamemo podatke iz ene sezone. V primeru morebitne analize več sezon pa bi bilo potrebno razmisliti o smiselnosti uporabe modela z izpeljanimi značilkami, saj je tudi ta model dosegal sprejemljivo natančnost pri detekcijah vseh 7 tipov vegetacij, je pa model ustrezneje zgrajen v primeru povečanega števila značilk, tudi v smislu časovne zahtevnosti izračuna, saj za novo časovno obdobje ni potrebno ustvariti novih značilk, temveč se zgolj ponovi izračun statističnih funkcionalov za izgradnjo nove podatkovne množice z izpeljanimi značilkami.

Poleg tega ne smemo pozabiti, da poleg detekcije poljščine, omenjena podatkovna zbirka vsebuje še delitev vegetacije na 6 različnih znanih poljščin. Klasifikacija znanih poljščin je zagotovo eden od možnih predlogov za nadaljevanje raziskovalnega dela. Specifika pri pristopu s klasifikacijo je v tem, da so podatki nesorazmerno uravnoteženi, torej v celotni podatkovni množici se nahaja le nizek delež takih podatkov (lahko tudi 5% delež), ki so klasificirani kot specifična poljščina. V takih primerih je potrebno biti previden pri ustrezni interpretaciji rezultatov (npr. 99% natančnost še ne pomeni, da smo zgradili kvaliteten model), predvsem pri uporabi objektivnih mer evalvacije (npr. AUC). Pri tem pristopu je razumljivo pričakovati nižjo stopnjo natančnosti, kot pri binarni detekciji. Prednost je seveda v tem, da imamo enoten model za klasifikacijo nerazvrščenih poligonov, namesto obstoječih 7, ki binarno odločajo, ali poligon pripada določenemu tipu poljščine.

V splošnem lahko trdimo, da je modeliranje časovnih vrst specifičen postopek, odvisen od tipa podatkov, dolžine časovne vrste, in seveda količine podatkov. Izpeljava novih značilk pa je uporabna alternativna rešitev, ki je lahko najmanj enakovredna, v določenih primerih tudi bolj učinkovita v primerjavi z ostalimi pristopi [29]. S to trditvijo lahko tudi potrdimo začetno hipotezo tega raziskovalnega dela, da je model z izpeljanimi značilkami vsaj enakovreden neposrednem modeliranju časovnih vrst.

Za konec si oglejmo še, ali smo poleg hipoteze izpolnili še vse zastavljene cilje razisko-

valnega dela:

- Pregledali in povzeli smo nekaj primerljivih raziskovalnih del na temo obdelave časovnih vrst na realnih podatkih,
- zgradili in optimizirali smo model z izpeljanimi značilkami, ki je vsaj enakovreden neposrednem modeliranju časovnih vrst,
- izpostavili smo najbolj informativne značilke s pomočjo najbolj obteženih značilk pri metodi SVM za posamezen model,
- izvedli smo primerjavo med dvema načinoma modeliranja časovnih vrst ter analizirali razlike v natančnosti.

Literatura

[1] M. Alam, Time series modeling for forecasting the earthquake behavior in Indonesia, *Proc. of Water and Geoscience*, 2015, 174–179.

[2] E. Baidoo, *An Analysis of Accuracy using Logistic Regression and Time Series*, Grey Literature from PhD Candidates, Kennesaw State University, 2016.

[3] M. Biček, *Grafični gradnik za merjenje kvalitete klasifikatorja s pomočjo krivulj*, Diplomsko delo, Univerza v Ljubljani, 2009.

[4] J. Božić in Đ. Babić, EUR/RSD Exchange Rate Forecasting Using Hybid Wavelet-Neural Model, *Computer Science and Information Systems 12* 2 (2015), 487–508.

[5] M. Bramer, *Principles of Data Mining*, Springer, 2007.

[6] N.V. Chawla, Data Mining for Imbalanced Datasets: An Overview: O. Maimon, L. Rokach, *Data Mining and Knowledge Discovery Handbook*, Springer US, 2005, 853–867.

[7] E. Eirola, in A. Lendasse, Gaussian Mixture Models for Time Series Modelling, Forecasting and Interpolation, *Advances in Intelligent Data Analysis XII, London*, 2013. 162–173

[8] F. Eyben, M. Wöllmer in B. Schuller, OpenEAR - Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit, *Proceedings 3rd InternationalConference on Affective Computing and Intelligent Interaction and Workshops*, 2009. 576–581

[9] J. Fan, S. Upadhye in A. Worster, Understanding receiver operating characteristics (ROC) curves, *Canadian Journal of Emergency Medicine* 8 (2006), 19–20.

[10] T. Fawcett, An introduction to ROC analysis, *Pattern Recognition Letters - Special issue: ROC analysis in pattern recognition* 27 (2006), 861–874.

[11] S. Fong, K. Lan in R. Wong, Classifying Human Voices by Using Hybrid SFX Time-Series Preprocessing and Ensemble Feature Selection, *BioMed Research International* 2013 (2013), 1–27.

[12] L. Ghalayini, Modeling and Forecasting the US Dollar/Euro Exchange Rate, *International Journal of Economics and Finance* 6 (2014), 194–207.

[13] S. Gowrishankar, A Time Series Modeling and prediction of wireless Network Traffic, *Georgian Electronic Scientific Journal: Computer Science and Telecommunications* 2 (2008), 40–52.

[14] A. Hashemi, H. Arabalibiek in K. Agin, Classification of Wheeze Sounds Using Wavelets and Neural Networks, *International Conference on Biomedical Engineering and Technology*, 2011. 127–131

[15] A. Kampouraki, G. Manis in C. Nikou, Heartbeat Time Series Classification With Support Vector Machines, *IEEE Transactions on Information Technology in Biomedicine* 13 (2009), 512–518.

[16] J. Kamruzzaman in R.A. Sarker, Comparing ANN Based Models with ARIMA for Prediction of Forex Rates, *ASOR Bulletin* 22 (2003), 2–11.

[17] D. A.Kocjan, Ječmen, *Naša žena* 2 (1999), 19–20.

[18] H. Kosorus, J. Hönigl in J. Küng, Using R, Weka and RapidMiner in Time Series Analysis of Sensor Data for Structural Health Monitoring, *DEXA '11 Proceedings of the International Workshop on Database and Expert Systems Applications*, 2011. 306–310

[19] Q. Liu, C. Chen, Y. Zhang in Z. Hu, Feature Selection for Support Vector Machines with RBF Kernel, *Artificial Intelligence Review* 2011 (2011), 99–115.

[20] T.W. Liao, Clustering of time series data – a survey, *The Journal of the Pattern Recognition Society* 38 (2005), 1857–1874.

[21] A.H. Nury, M. Koch in M.J.B. Alam, Time Series Analysis and Forecasting of Temperatures in the Sylhet Division of Bangladesh, *Proceedings of 4th International Conference on Environmental Aspects of Bangladesh*, 2013. 267–271

[22] M. Perc, Nonlinear time series analysis of the human electrocardiogram, *European journal of physics* 26 (2005), 757–768.

[23] J. Rodriguez, in L. Kuncheva, Time series classification: Decision forests and SVM on interval and DTW features, *Proceedings of the Workshop on Time Series Classification, 13th International Conference on Knowledge Discovery and Data Mining*, 2007. 162–171

[24] P. Schäfer, Scalable time series classification, *Data Mining and Knowledge Discovery* 5 (2016), 1273–1298.

[25] D. Tlegenova, *Forecasting Exchange Rates Using Time Series Analysis: The sample of the currency of Kazakhstan*, Cornell University Library, e-arhiv, 2015.

[26] K.C. Tseng, O. Kwon in L.C. Tjung, Time series and neural network forecast of daily stock prices, *Investment Management and Financial Innovations* 9 (2012), 32–54.

[27] V.N. Vapnik, An Overview of Statistical Learning Theory, *The Journal of IEEE Transactions on Neural Networks* 10 (1999), 988–999.

[28] Ž. Vlahušić, *Primerjalna analiza odprtokodnih programov za podatkovno rudarjenje*, Diplomsko delo, Univerza v Ljubljani, 2013.

[29] J. Wiens, J.V. Guttag in E. Horvitz, Patient Risk Stratification for Hospital-Associated C.diff as a Time-Series Classification Task, *NIPS'12 Proc. 25th International Conference on Neural Information Processing Systems*, 2013. 467–475

[30] I.H. Witten in E. Frank, *Data Mining: Practical machine learning tools and techniques, Second Edition*, Morgan Kaufmann, 2005.

[31] MATLAB - MathWorks.
URL: https://www.mathworks.com/products/matlab.html (13.8.2017)

[32] RapidMiner - Open Source Data Science Platform.
URL: https://rapidminer.com/ (13.8.2017)

[33] SEOS - Remote Sensing and GIS in Agriculture.
URL: http://www.seos-project.eu/modules/agriculture/agriculture-c01-s02.html (13.8.2017)