IZZIVI STROJNEGA UČENJA Z NEURAVNOTEŽENIMI RAZREDI NA PRIMERU DETEKCIJE PREVAR PRI KASKO ZAVAROVANJU

David Šenica, Martin Jurkovič in Tadej Justin
Medius d.o.o., Tehnološki park 21, 1000 Ljubljana, Slovenia
david.senica@medius.si, martin.jurkovic@medius.si, tadej.justin@medius.si

Povzetek

Pri razvoju inteligentnih sistemov za pomoč pri odločanju, razvitih na osnovi strojnega učenja, se v praksi pogosto srečujemo s pomanjkanjem uporabnih podatkov, ki bi pripomogli k razvoju kvalitetnih modelov. Tako smo razvijalci velikokrat podvrženi majhnemu številu označenih vzorcev, ki so hkrati tudi neenakomerno zastopani v razredih razvrščanja. S tako zastopanostjo označenih vzorcev v razredih razvrščanja smo se srečali tudi pri razvoju razvrščevalnika sumljivih in regularnih škodnih dogodkov pri prijavi škode iz naslova avtomobilskega kasko zavarovanja. V prispevku opišemo izzive in nekatere rešitve pri razvoju razvrščevalnikov na podlagi neuravnoteženih označenih podatkov v razredih razvrščanja. Osredotočimo se na primer razvoja detektorja sumljivih škodnih prijav s pomočjo strojnega učenja in ovrednotimo predlagane rešitve. Posvetimo se interpretaciji in primerjavi evalvacij razvrščevalnikov, razvitih z nadzorovanim učenjem ter pridobljenih na podlagi dodatne manipulacije označenih vzorcev v učnih množici. Manipulacija vzorcev temelji na podlagi umetnega napihovanja minorno zastopanega razreda in na podlagi postopkov kleščenja vzorcev v razredu razvrščanja z večjim številom vzorcev pri nadzorovanem učenju. Dodatno izvedemo tudi nastavljanje optimalnih parametrov razvrščevalnika na podlagi algoritma Grid search in predstavimo rezultate.

Abstract

IMBALANCED DATA SETS CHALLENGES IN MACHINE LEARNING WITH A CASE STUDY IN CASCO CAR INSURANCE CLAIMS FRAUD DETECTION

When developing intelligent decision support systems based on machine learning we often encounter a lack of useful data that would help to develop quality models. Thus, developers are often subjected to a small number of labelled samples with their target class being unevenly distributed. We also encountered such a representation of imbalanced labelled samples while developing a classifier of suspicious and regular damage claims from casco car insurance. In this paper we describe the challenges and some solutions in the development of classifiers based on imbalanced labelled data in the target classes. We focus on the development of a suspicious claims detector with the help of machine learning and evaluate the proposed solutions. We shed light on the interpretation and comparison of evaluations of classifiers that were developed with supervised learning based on the manipulation of labelled samples in the learning set. Sample manipulation is achieved with the artificial inflation of samples with the minority class and through the procedure of squashing samples with the majority class. Additionally with the Grid search algorithm, we estimate the optimal classifier's hyper parameters, and present results.

Ključne besede

Avtomobilsko zavarovanje, nadzorovano strojno učenje, optimizacija razvrščevalnikov, prevare

Keywords

Car insurance, supervised machine learning, classifier optimization, fraud detection

UVOD

Prevaro v zavarovalništvu lahko opišemo kot dogodek, ko posameznik sebi ali drugemu priskrbi protipravno premoženjsko korist iz naslova zavarovalne odškodnine. Zavarovalnice poskušajo na različne načine preprečiti prevare. S podrobnim preučevanjem izvajanja prevar se ukvarja kriminalna psihologija, kjer so znani psihologi poskušali razumeti in opisati, kaj človeka žene k prevari in kako pojem prevare podrobneje opredeliti.

Odvračanje od prevar je aktivnost, ki se ukvarja z odstranitvijo razlogov oziroma predpogojev za pojav prevar. V tem kontekstu je naloga zavarovalnice, da onemogoči oz. oslabi vsaj enega izmed treh oglišč t.i. trikotnika prevar po Cressey-ju.

Slika 1:Trikotnik prevar po Cressey-ju.

Cressey je že v petdesetih letih 20. stoletja na podlagi empiričnih raziskav postavil hipotezo, da so za izvedbo prevare potrebni trije ključni predpogoji, ki morajo biti izpolnjeni sočasno in v pravem medsebojnem razmerju [1]. Čeprav najdemo v sodobni literaturi tudi razširitve in drugačne napotke za preprečevanje prevar [2], se zavarovalnice vseeno še danes poslužujejo prav zgoraj opisanega principa.

V tem prispevku se bomo osredotočili na detekcijo prevare in tako skušali preprečiti njeno realizacijo s pomočjo strojnega učenja. Na okrnjenih in psevdonimiziranih podatkih škodnih spisov bomo izpostavili realne probleme, s katerimi smo se srečali pri razvoju takega sistema in izpostavili nekaj rezultatov, ki smo jih pridobili s preizkusi različnih pristopov pri učenju modela za detekcijo suma prevar pri neenakomerno uteženih razredih razvrščanja.

METODE

Avtomatska detekcija prevar je v zavarovalništvu zelo zaželena, saj vsako prizadevanje na tem področju lahko razbremeni preiskovalce prevar in jim omogoča večjo učinkovitost. Zavarovalniška panoga velja za informacijsko dobro podkrepljeno in v veliki meri digitalizirano okolje, vseeno pa se najde prostor za izboljšave in prilagoditve.

Za uspešno detekcijo prevar so škodni spisi ključen vir informacij. Dandanes so v veliki meri digitalizirani, vendar žal ponekod tudi pomanjkljivi za digitalno avtomatsko obravnavo. Njihovi ključni deli so zapisani kot nestrukturirani podatki, čeprav bi uvedba šifrantov lahko bistveno pripomogla pri preglednosti in analizi škodnih spisov. Tako se v škodnih spisih na primer pojavlja opis škodnega primera kot tekstovni opis npr. “Viden udarec na levem delu

zadnjega odbijača" ali "Razbita desna sprednja utripalka" ipd. Šifrant tovrstnih opisov poškodovanih delov vozila bi lahko bistveno izboljšal možnost enostavnejše analize podatkov in posledično tudi pripomogel k boljšemu modeliranju detekcije prevar s pomočjo strojnega učenja.

Slovensko zavarovalno združenje, GIZ, nam je priskrbelo podatkovno zbirko psevdonimiziranih podatkov o škodnih spisih zavarovalnic, ki so članice združenja. Te podatke, ki so temeljili na opisu ključnih dogodkov v škodnih spisih, smo strukturirali in jih uporabili za enostavno modeliranje značilk. Zaradi specifike in občutljivosti podatkov jih žal v tem prispevku ne moremo podrobneje opisati. Iz pridobljene zbirke smo izpeljali 20 značilk. Z redukcijo dimenzije z algoritmom Principal Component Analysis, PCA [3], smo za nadaljnje delo izbrali 18 komponent PCA. Tako smo v smislu modeliranja prevar s strojnim učenjem vsakemu vzorcu – škodnemu spisu – pripisali 18-razsežni vektor, ki opisuje škodni spis oz. prijavo škodnega dogodka. Iz zbirke smo izbrali označen material in pridobili 5463 označenih vzorcev škodnih dogodkov. Kot prevara je bilo označenih 305, kot regularni spis pa 5158 vzorcev. Sklop označenih podatkov so označile zavarovalnice oz. njihovi preiskovalci prevar. Označen material v zbirki škodnih spisov predstavlja ključne podatke, s katerimi s strojnim učenjem gradimo modele.

Kljub temu, da lahko tudi pri nenadzorovanem učenju pridobimo relativno dobre rezultate, pa je potrebno omeniti, da večina algoritmov za pridobivanje modelov s strojnim učenjem pri nenadzorovanem učenju deluje na principu rojenja (ang. clustering) [6], kjer se z določeno mero izračuna oddaljenost posameznega vzorca od rojev vzorcev in se jih uredi glede na največjo oddaljenost. Pri tem pristopu govorimo o iskanju osamelcev. V primeru nadzorovanega učenja pa izrabimo preteklo znanje označevalcev na že označenem materialu. V postopku učenja skušamo v razvrščevalniku določiti razrede razvrščanja na različne matematično kompleksne načine [4]. Velikokrat v tem postopku rečemo, da modeliramo preteklo znanje ekspertov področja razvrščanja za dotičen problem.

Pravilno interpretacijo pridobljenih rezultatov v veliko primerih otežujejo neuravnoteženi razredi razvrščanja. Prepoznavanje prevar v škodnih spisih zavarovalnic je tipičen predstavnik tovrstnega izziva, saj imamo pri dveh razredih razvrščanja število vzorcev v prid regularnim spisom skoraj v razmerju 17:1. Pri tovrstnih problemih se v literaturi priporoča uporaba algoritmov nenadzorovanega učenja, saj se minorno zastopan razred razvrščanja lahko obravnava kot osamelce. Priporočena je torej uporaba algoritmov, kot so Variational Autoencoders (VAE) [7], Isolation Forest (IF) [8] ali One class support vector machine [9]. Sami smo preizkusili delovanje VAE in IF, vendar o rezultatih v tem prispevku ne poročamo, saj skušamo izpostaviti izzive, s katerimi se srečamo pri nadzorovanjem učenju z neuravnoteženimi podatki.

Detekcija osamelcev velja v splošnem za manj natančen način detekcije potencialnih prevar, kot pa to lahko dosežemo z udejanjenim razvrščevalnikom na podlagi nadzorovanega učenja. Pri razvrščanju sumljivih in regularnih spisov uporabimo znanje, ki so ga preiskovalci na danih podatkih v spisih že odkrili, in posledično lahko govorimo, da skušamo modelirati do sedaj odkrite tipe prevar, ki jih bo razvrščevalnik sumljivih in regularnih spisov lahko razvrstil. Obstaja kar nekaj primernih učnih algoritmov, ki so primerni za tovrsten problem razvrščanja [4]. Vseeno smo razvili več razvrščevalnikov z 10-kratnim navzkrižnim preverjanjem in rezultate primerjali med seboj, na podlagi evalvacije pa izbrali najbolj primernega. Najboljše rezultate je dosegel algoritem, ki ga uvrščamo v skupino odločitvenih dreves – Random Forest [10]. Ker pa razpolagamo z neuravnoteženimi razredi razvrščanja, lahko hitro pridobimo

rezultate, ki so po vrednostih osnovnih mer vrednotenja razvrščevalnika precej uspešni, vendar je ta uspešnost zavajujoča, saj k uspešnemu rezultatu prispevajo le pravilno razvrščeni vzorci regularnega razreda, medtem ko ni nobenega pravilno razvrščenega razreda z manjšim številom vzorcev. Razlog za tovrstno delovanje lahko iščemo v tem, da je odločilna meja med razredoma razvrščanja postavljena precej v prid normalnemu razredu (regularnih škodnih spisov), kar posledično pomeni, da smo pri razvoju tovrstnih sistemov primorani poiskati optimalne hiperparametre učnega algoritma, ki omogočajo pridobitev boljših rezultatov. Alternativno rešitev pa lahko iščemo pri manipulaciji učne množice na tak način, da poskusimo uravnotežiti vzorce v razredih razvrščanja. Manipulacijo nad številom učnih vzorcev lahko štejemo med optimizacijo učnega algoritma, saj učnemu algoritmu omogočimo razviti optimiziran razvrščevalnik.

V literaturi [11] se srečujemo z dvema pristopoma k tovrstni manipulaciji števila vzorcev učne množice. Pri obeh dosežemo, da razpolagamo s približno uravnoteženim številom vzorcev v razredih razvrščanja. Prvi temelji na tem, da vzorcem minornega razreda razvrščanja dodajamo umetno tvorjene vzorce, ki so bili na različne načine umetno generirani na podlagi obstoječih vzorcev v dotočnem razredu. Drugi pristop pa temelji na tem, da odstranimo vzorce iz razredov in njihovo število izenačimo z manj zastopanim razredom. Tudi tu se uporablja več načinov izločanja vzorcev. V našem prispevku smo se odločili preizkusiti po enega izmed obeh načinov. Za generiranje umetnih vzorcev v minorno zastopanem razredu sumljivih spisov smo uporabili algoritem ADASyn [12]. Pri odstranjevanju vzorcev pa smo razvili lastni pristop, ki temelji na naključnem izboru vzorcev iz desetih rojev. Vzorce, ki so pripadniki razreda regularnih spisov, smo s pomočjo rojenja z algoritmom KMeans [13] rojili v deset razredov. Iz teh rojev smo nato naključno, vendar čim bolj enakomerno po rojih izbrali približno enako število vzorcev, kot pa jih je vseboval razred razvrščanja sumljivih škodnih spisov.

Drugi pristop k optimizaciji razvrščevalnika ne spreminja učne množice, ampak optimizira parametre uporabljenega klasifikatorja. Pri tem pristopu se je vredno najprej vprašati, kakšne so zahteve sistema, ki ga razvijamo. V našem primeru želimo razviti detektor sumljivih škodnih spisov, ki bo uporabljen kot dodatni sistem za pomoč pri odločitvi, kateri škodni spis je vredno ročno preučiti. Razviti sistem bo tako preiskovalcem ponujal nabor škodnih spisov, ki jih bo označil za sumljive. Pri takem sistemu težko dosežemo, da bo sistem ponudil v pregled vse škodne spise, ki bodo rezultirali kot dejanske prevare. Pričakujemo lahko določen odstotek napačno razvrščenih vzorcev. Tak tip napake lahko minimiziramo z optimizacijo razvrščanja, pri čemer se osredotočimo na optimizacijo preciznosti in priklica v razredu razvrščanja sumljivih škodnih spisov. Želimo si, da bi obe meri dosegli kar se da visoko vrednost. Na ta način naš sistem nekaj škodnih spisov, ki bodo rezultirali v dejanske prevare, ne bo detektiral, vendar bo v pregled ponudil tudi manj takih, ki predstavljajo dejansko regularne škodne spise.

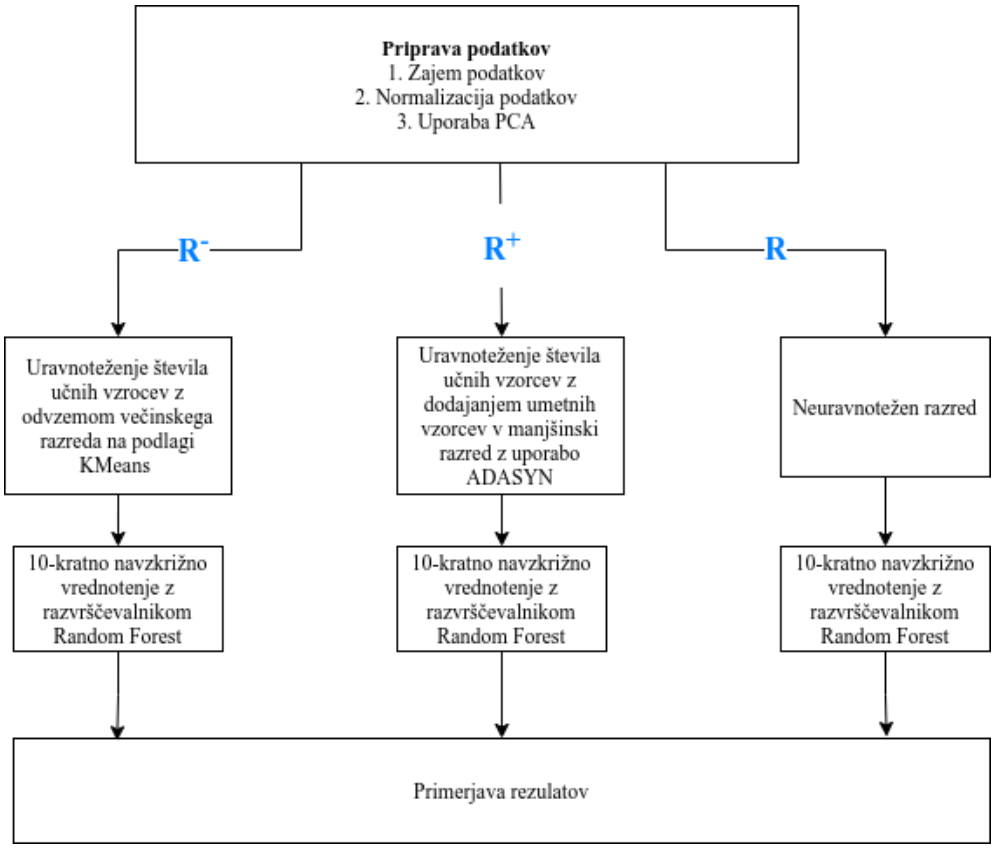
Za doseganje optimalnega delovanja razvrščevalnika smo uporabili algoritem Grid search [11], ki omogoča avtomatsko evalvacijo razvitega razvrščevalnika na enaki učni/testni množici pri različni konfiguraciji hiperparametrov učnega algoritma. Obenem pa nam že enostavna analiza vrednosti priklica in preciznosti pri spreminjajočem se pragu odločanja precej izboljša zahteve detekcije sumljivih škodnih spisov, kar predstavimo kot enostavnejšo alternativno rešitev iskanju optimalne točke delovanja razvrščevalnika.

REZULTATI

Predstavljeni rezultati se osredotočajo na optimizacijo razvrščevalnika pri neenakomerno zastopanih razredih razvrščanja. V našem primeru si želimo, da bi razvrščevalnik čim manj

sumljivih spisov, ki so dejansko prevare, označil kot regularne spise in čim manj regularnih spisov označil kot sumljive škodne spise. Ko rezultate predstavimo s konfuzijsko matriko, hitro razberemo, da skušamo doseči čim večji priklic pri čim večji preciznosti. Zato vse evalvacijske mere predstavljamo v smislu osnovnih evalvacijskih mer razvrščevalnika [4], to so preciznost, priklic in mera F1 vsakega izmed razredov razvrščanja, dodamo pa tudi izračun uteženega povprečja obravnavanih mer.

Rezultati se nanašajo na uporabo algoritma za nadzorovano učenje Random Forest iz skupine odločitvenih dreves. Ta je pri medsebojni primerjavi večjega števila algoritmov za učenje razvrščevalnikov dosegel najboljše rezultate. Rezultate poročamo glede na vnaprej izbrano testno množico. S tem zagotovimo, da so rezultati med seboj primerljivi. Čeprav v nekaterih primerih manipuliramo učno množico, pri teh postopkih vedno pazimo, da vzorci, ki so del testne množice, nikdar niso prisotni v učni množici.



Slika 2: Diagram uporabljenih metod, način evalvacije in primerjava

Najprej predstavimo vrednotenje razvitih razvrščevalnikov, kjer smo manipulirali z učnimi vzorci učne množice tako, da smo odvzeli vzorce regularnih spisov na način, ki je opisan v poglavju 2. V tabeli 1 smo tak ravrščevalnik označili z R-. Razvrščevalnik, razvit s pomočjo učne množice, ki smo ji umetno dodali vzorce z algoritmom ADASyn, pa z R+. T črko R, pa smo označili razvrščevalnik, ki je bil razvit nad celotnim označenim materialom. Diagram na Sliki 2 za lažje razumevanje uporabljenih algoritmov za manipulacijo učnih vzorcev dodatno predstavi uporabljene metode. Tabela 1 prikazuje, da oba postopka za uravnoteženje razredov pripomoreta k doseganju boljših rezultatov. Lastna implementacija postopka izbora regularnih spisov doseže celo boljše rezultate kot pa če manj zastopane vzorce umetno generiramo z

algoritmom ADASyn. Vidimo, da razpoznavalnik R- bistveno popravi rezultat priklica pri razvrščanju sumljivih škodnih spisov, pri relativno majhnem poslabšanju preciznosti. Na podlagi Tabele 1 smo se odločili, da bomo vse nadaljnje optimizacije preizkusili na podlagi učne množice, ki smo jo uporabili pri učenju razvrščevalnika R-.

Sistem optimizacije	**Razred**	**Preciznost**	**Priklic**	**F1**
R	Sumljiv	0,867	0,043	0,081
	Regularen	0,509	0,993	0,673
	Povprečen utežen	**0,688**	**0,518**	**0,377**
R^-	Sumljiv	0,643	0,849	0,732
	Regularen	0,778	0,528	0,629
	Povprečen utežen	**0,710**	**0,689**	**0,680**
R^+	Sumljiv	0,718	0,426	0,535
	Regularen	0,592	0,833	0,692
	Povprečen utežen	**0,655**	**0,630**	**0,614**

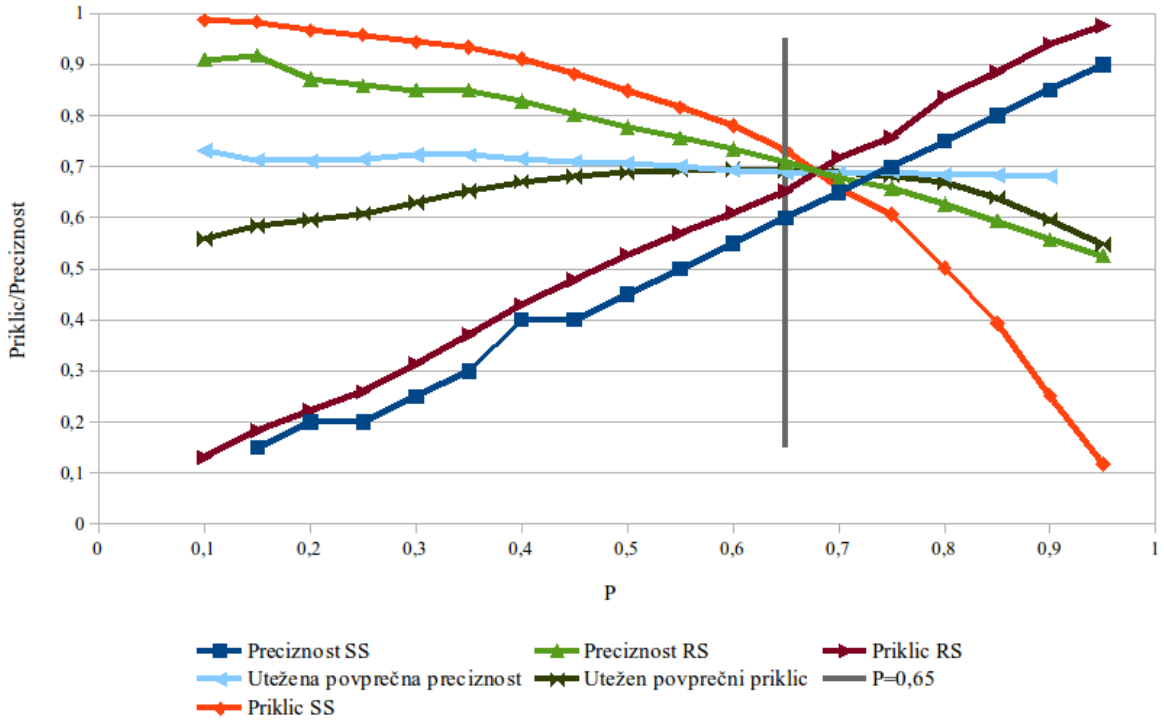
Tabela 1: Primerjava vrednotenja razvrščevalnikov z 10-kratnim navzkrižnim preverjanjem pri manipulaciji števila vzorcev v učni množici R- (odvzemanje regularnih spisov), R+ (napihovanje sumljivih spisov z ADASyn algoritmom) in R (razpoznavalnik pri neuravnoteženi učni množici)

Nekateri algoritmi, ki omogočajo udejanjanje razvrščevalnikov, delujejo tako, da vsakemu testnemu vzorcu pripišejo verjetnost pripadnosti (P) enemu izmed razredov razvrščanja. Tudi razvrščevalnik, naučen na podlagi algoritma Random Forest, omogoča pripis takšne vrednosti. Uravnotežen razvrščevalnik predpostavi, da je meja med binarnim razredom razvrščanja 0,5. Če pa želimo pridobiti optimizirane rezultate za doseganje konkretnega cilja razvrščevalnika, pa moramo ta prag nastaviti sami.

Najbolj enostaven princip za doseganje maksimalnega priklica pri maksimalni preciznosti razreda razvrščanja sumljivih škodnih spisov predstavlja iskanje preseka med krivuljo preciznosti v odvisnosti od vrednosti P in vrednostjo priklica v odvisnosti od vrednosti P. Slika

3 prikazuje določanje take meje, ki smo jo pri izvedenem preizkusu določili kot nov prag, ki omogoča razvrščanje z doseganjem večje preciznosti pri sicer manjšem priklicu. Nov prag smo postavili pri vrednosti P=0,65. Pri tem smo zagotovili, da nismo optimizirali samo razvrščanja vzorcev v razred sumljivih spisov, pač pa tudi v razred regularnih škodnih spisov.

Tabela 2 prikazuje primerjavo med evalvacijskimi merami pri vrednosti praga odločanja P=0,5 in P=0,65. Vidimo, da lahko pri pragu 0,65 preciznost razvrščanja sumljivih spisov povečamo za skoraj 13 odstotkov, vendar se to zgodi na račun priklica, saj se ta poslabša za skoraj 32 odstotkov.



Slika 3: Določitev optimalnega praga P=0,65 razvrščevalnika pri izrisu priklica in preciznosti regularnih spisov, sumljivih spisov in uteženega povprečja

	P=0,5			**P=0,65**		
Razred	**Preciznost**	**Priklic**	**F1**	**Preciznost**	**Priklic**	**F1**
Sumljiv (SS)	0,643	0,849	0,732	0,778	0,518	0,622
Regularen (RS)	0,778	0,528	0,629	0,639	0,852	0,730
Povprečen utežen	**0,710**	**0,689**	**0,680**	**0,709**	**0,685**	**0,676**

Tabela 2: Primerjava rezultatov 10 kratnega navzkrižnega preverjanja pri različnem pragu P razvrščevalnika

Bolj napredno optimizacijo iskanja parametrov učnega algoritma pa predstavlja algoritem Grid search. Algoritem izvaja učenje razvrščevalnikov v območju vrednosti parametrov, ki jih določi razvijalec, sam pa na podlagi evalvacije razvrščevalnikov in podane cenilke skuša poiskati optimalne hiperparametre razvrščevalnika. Tabela 3 prikazuje primerjavo rezultatov med optimiziranimi razvrščevalniki ter katere parametre smo določili iterativno s pomočjo algoritma Grid search. Določili smo dva različna optimizacijska kriterija. Najprej smo skušali optimizirati priklic, nato tudi preciznost. Tabela 3 prikazuje optimalno izbrane parametre pri maksimizaciji priklica in preciznosti.

Parameter učnega algortima Random Forest	**Optimizacija priklica**	**Optimizacija preciznosti**
Maksimalna globina drevesa	5	25
Maksimalni delež obravnavanih značilk pri vejitvi	1	1
Minimalno število vzorcev za delitev veje	3	2
Minimalno število vzorcev za določitev lista	1	1
Število dreves v gozdu	200	100

Tabela 3: Parametri razvrščevalnikov pri optimizaciji priklica in preciznosti

Primerjava rezultatov evalvacije s tako razvitimi parametri je prikazana v Tabeli 4. Vidimo, da smo z optimiranjem parametrov razvrščevalnika pri razredu razvrščanja sumljivih spisov v smislu preciznosti pridobili približno 5 odstotkov, na račun zmanjšanja priklica za približno 10 odstotkov. Tako razvrščevalnik razvit glede na optimizacijo priklica, kot optimizacijo preciznosti dosegata zelo podobne rezultate, je pa razvrščevalnik, s parametri ki smo ga razvili s pomočjo optimizacije preciznosti bistveno bolj kompleksen od tistega, ki smo ga razvili na podlagi optimizacije priklica.

Sistem optimizacije	**Razred**	**Preciznost**	**Priklic**	**F1**
Privzeti hiper parametri	Sumljiv	0,643	0,849	0,732
	Regularen	0,778	0,528	0,629

	Povprečen utežen	**0,710**	**0,689**	**0,680**
Optimizacija priklica	Sumljiv	0,696	0,757	0,725
	Regularen	0,734	0,669	0,700
	Povprečen utežen	**0,715**	**0,713**	**0,713**
Optimizacija preciznosti	Sumljiv	0,698	0,734	0,716
	Regularen	0,720	0,682	0,700
	Povprečen utežen	**0,709**	**0,708**	**0,708**

Tabela 4: Primerjava rezultatov po optimizaciji algoritma grid-search - najboljši rezultati.

ZAKLJUČEK

V prispevku smo opisali postopek pridobivanja optimalnega razvrščevalnika škodnih spisov za namen detekcije sumljivih škodnih spisov. Namen tovrstnega razvrščevalnika je zavarovalniškim preiskovalcem olajšati delovanje in jim v predogled ponuditi sumljive škodne spise, ki bi lahko bili obravnavani v postopku nadaljnjega raziskovanja morebitne prevare. Cilj tovrstnega sistema je raziskovalcem ponuditi sumljive škodne spise s čim večjo preciznostjo razvrščanja ob največjem možnem priklicu razvrščanja.

Na področju integracije strojnega učenja v informacijski poslovni proces se srečujemo z različnimi tipi podatkov, zelo pogosto pa so podatki, ki so del poslovnega podatkovnega toka, neenakomerno zastopani v razredih razvrščanja. Ta prispevek ponuja nekaj praktičnega vpogleda v reševanje tovrstnih izzivov na primeru razvrščanja sumljivih in regularnih škodnih spisov, kjer smo z odstranjevanjem vzorcev iz učne množice razreda razvrščanja regularnih škodnih spisov razvili bolj optimalno razvrščanje. Nadalje smo z uporabo algoritma Grid search izvedli dodatno optimizacijo parametrov algoritma Random Forest in s tem še dodatno izboljšali rezultat razvrščanja.

Predstavljeni rezultati dosegajo F1 mero med 0,63 in 0,73 v razredu razvrščanja sumljivih spisov. Mera odraža tako uspešnost priklica kot tudi preciznosti, zato jo v zaključku izpostavljamo kot končno mero uspešnosti. Ocenjujemo, da bi se rezultat lahko še bistveno izboljšal z večjim številom označenih prevar v danem materialu in z uporabo fuzije tako nenadzorovanih kot nadzorovanih postopkov pri razvrščanju škodnih spisov med sumljive in regularne. Prav tako pa je še nekaj prostora pri izpeljavi značilk, ki bi še bolj enoznačno odražale škodni dogodek.

Optimizacija parametrov algoritmov strojnega učenja je s pomočjo sodobnih algoritmov relativno trivialna in lahko dodatno pripomore pri izboljšanju razvrščanja, vendar je močno odvisna od zastavljenih ciljev, ki jih skušamo z razvitim razvrščevalnikom doseči. Zato je zaželeno, pri vpeljavi strojnega učenja v poslovni proces, se čim bolj približati zastavljenim ciljem, ki so v različnih panogah in predvidenih optimizacijah zelo različni. Ker so tovrstni sistemi močno odvisni od učnih podatkov je potrebno posebno pozornosti posvetiti tudi razlagam uspešnosti razvrščanja oziroma uspešnosti vpeljave strojnega učenja v informacijski sistem. Temu pa hiše za razvoj programske opreme malokrat posvetijo dovolj pozornosti.

VIRI IN LITERATURA

[1] CRESSEY, D. R.: Other people's money: a study in the social psychology of embezzlement Glencoe, IL: The Free Press, 1953

[2] KASSEM, R. in HIGSON, A.: The new fraud triangle model. Journal of emerging trends in economics and management sciences, 3(3), 2012, str. 191-195.

[3] SVANTE, Wold, ESBENSEN, Kim in GELADI Paul: "Principal component analysis." Chemometrics and intelligent laboratory systems 2.1-3, 1987, str. 37-52.

[4] BISHOP, Christopher M.: "Pattern recognition." Machine learning 128.9., 2006.

[5] WONG, Tzu-Tsung: "Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation." Pattern Recognition 48.9, 2015, str. 2839-2846.

[6] PAVEŠIĆ, Nikola: Razpoznavanje vzorcev: Uvod v analizo in razumevanje vidnih in slušnih signalov, Fakulteta za elektrotehniko, Ljubljana, 2000

[7] IVANOV, Oleg, FIGURNOV, Michael in VETROV, Dmitry.: "Variational autoencoder with arbitrary conditioning." arXiv preprint arXiv:1806.02382, 2018

[8] LIU Fei, Tony, Kai Ming, TING, in Zhi-Hua, ZHOU.: "Isolation forest." 2008 eighth ieee international conference on data mining. IEEE, 2008.

[9] Yin, Shen, Xiangping Zhu, and Chen Jing: Fault detection based on a robust one class support vector machine., Neurocomputing 145, 2014, str. 263-268.

[10] MAHESH, Pal: Random forest classifier for remote sensing classification., International journal of remote sensing 26.1, 2005, str. 217-222.

[11] AJINKYA, More: Survey of resampling techniques for improving classification performance in unbalanced datasets., arXiv preprint arXiv:1608.06048, 2016.

[12] He, Haibo, et al. "ADASYN: Adaptive synthetic sampling approach for imbalanced learning." 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence). IEEE, 2008.

[13] Likas, Aristidis, Nikos Vlassis, and Jakob J. Verbeek. "The global k-means clustering algorithm." Pattern recognition 36.2 (2003): 451-461.

[14] LIASHCHYNSKYI, Petro, in LIASHCHYNSKYI, Pavlo: Grid search, random search, genetic algorithm: a big comparison for NAS., arXiv preprint arXiv:1912.06059 (2019).