VILNIAUS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS FAKULTETAS
INFORMATIKOS KATEDRA

Baigiamasis bakalauro darbas

Mašinų mokymo metodų taikymas nesubalansuotiems duomenims

(Application of Machine Learning Methods for Imbalanced Data)

Atliko: 4 kurso 1 grupės studentas
Eligijus Silkartas (parašas)

Darbo vadovas:
Dr. Rimantas Kybartas (parašas)

Recenzentas:
Dr. Linas Laibinis (parašas)

Vilnius – 2019

Turinys

Įvadas....3
1. Klasifikatorių vertinimo metodika....4
1.1. Klasifikavimo matrica....4
1.2. Sprendimus priimančiojo ypatybių kreivė (ROC)....6
1.3. Preciziškumo-jautrumo kreivė (PRC)....7
2. Duomenų balansavimo metodika....9
2.1. Atsitiktinis duomenų išrinkinėjimas ir papildymas....9
2.2. Sintetinių duomenų generavimas....9
3. Atsitiktinio miško algoritmas duomenų klasifikavimui....11
3.1. Sprendimų medis....11
4. Tyrimas su realiais nesubalansuotais duomenimis....13
4.1. Tyrimo duomenys....13
4.2. Tyrimo rezultatai....15
4.3. Tyrimo su realiais duomenimis rezultatų apibendrinimas....16
5. Tyrimai su dirbtiniais duomenimis....19
5.1. Tyrimas, kai balansavimo metodai neturi įtakos....20
5.2. Tyrimas, kai krenta SMOTE kokybė....21
5.3. Tyrimas, kai SMOTE kokybė geriausia....23
6. Išvados....25
Šaltiniai....26

Įvadas

Realiame pasaulyje susiduriama su įvairiausiais duomenimis, kuriuose gali pasitaikyti įvairiausių dėsningumų, dėl kurių gali būti sunku išspręsti uždavinius, susijusiais su tais duomenimis. Vienas iš tokių atvejų yra, kai turimi duomenys yra nesubalansuoti.

Yra du skirtingi nesubalansuotų duomenų variantai [HE09]. Pirmasis yra kai disbalansas matomas tarp skirtingų klasių, pavyzdžiui, kai santykis tarp dviejų klasių duomenų yra 1:10, 1:100, 1:1000 ir panašiai. Čia minimas disbalansas tarp dviejų klasių, tačiau ši problema gali iškilti ir esant duomenims su daugeliu klasių. Antrasis nesubalansuotų duomenų variantas yra, kai disbalansas atsiranda klasės viduje, kai tos pačios klasės duomenų dalis gali ženkliai skirtis nuo kitos dalies. Nesubalansuotų duomenų problema dažnai pasitaiko tokiuose uždaviniuose kaip medicininių ligų diagnozė, kompiuterinių tinklų įsilaužimo aptikimas ar sukčiavimo atvejų aptikimas.

Įvairiuose šaltiniuose apie mašinų mokymo metodus ir nesubalansuotus duomenis yra akcentuojama, kad susidūrus su nesubalansuotais duomenimis greičiausiai reikia taikyti kokią nors balansavimo strategiją. Taip yra todėl, kad tradiciniai mašinų mokymo metodai pritaikyti darbą atlikti su subalansuotais duomenimis, o su nesubalansuotais duomenimis gali atsirasti įvairių netikslumų [HE09]. Visgi yra pastebėta, kad pasitaiko atvejų, kai klasifikatoriai, gauti treniruojant su nesubalansuotais duomenimis, išmoksta atpažinti duomenis geriau nei klasifikatoriai, kurie mokėsi taikant kurią nors vieną iš daugelio balansavimo strategijų [HE09].

Šis darbas koncentruojasi į klasifikavimo uždavinių su nesubalansuotais duomenimis sprendimus. Skyriuje nr. 1 skyriuje pateikiama klasifikatorių vertinimo kokybės vertinimo metodika, skyriuje nr. 2 pateikiami balansavimo metodai, kurie naudoti atliekant šio darbo tyrimus. Skyriuje nr. 3 nagrinėjamas mokymo algoritmas, kuris buvo naudojamas treniruojant klasifikavimo modelius šiame darbe.

Šio darbo tikslas yra palyginti duomenų balansavimo strategijų daroma įtaką klasifikatorių mokymui su klasifikatoriaus mokymu netaikant jokios strategijos. Taip pat siekiama parodyti, kada yra veiksminga taikyti balansavimo metodus, o kada ne. Buvo atliktas vienas tyrimas su realiais bankinių transakcijų duomenimis, aprašytas skyriuje nr. 4, ir trys tyrimai su dirbtiniais generuotais duomenimis, kurie aprašyti skyriuje nr. 5.

Gauti tyrimų rezultatai parodo, kad balansavimo metodai gali ne tik pagerinti klasifikavimo kokybę, bet ir pabloginti.

1. Klasifikatorių vertinimo metodika

Labiausiai paplitusios klasifikatorių vertinimo metrikos yra tikslumo (angl. Accuracy) ir klaidų dažnio įverčiai. Šios metrikos padeda nustatyti kaip gerai klasifikatorius geba atpažinti klases bei pagal šias metrikas galima palyginti skirtingus klasifikatorius to paties uždavinio sprendimui. Tikslumas skaičiuojamas sudedant visus teisingai atpažintus duomenis ir padalijamas iš visų duomenų skaičiaus. Klaidų dažnis skaičiuojamas iš vieneto atimant tikslumo įvertį.

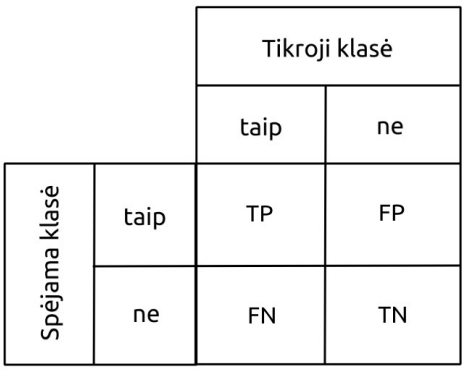
$$Accuracy = \frac{TP + TN}{P_C + N_C}; \qquad ErrorRate = 1 - Accuracy \tag{1}$$

Čia TP reiškia teisingai atpažintų teigiamos klasės duomenų skaičių, TN – teisingai atpažintų neigiamos klasės duomenų skaičių, P_C – teigiamos klasės visų duomenų skaičių, N_C – neigiamos klasės visų duomenų skaičių.

Šios vertinimo metrikos veikia pakankamai tikslai kai dirbama su subalansuotais duomenimis. Tačiau, jeigu duomenys yra nesubalansuoti, šios metrikos gali būti netikslios ir sudaryti apgaulingą įspūdį apie klasifikatoriaus gebėjimą atpažinti duomenis. Kad įsitikintume tuo, sakykime, jog dirbame su duomenų aibę, kurią sudaro 99 % neigiamos klasės duomenų ir 1 % teigiamos klasės duomenų. Pasirinkę klasifikatorių, kuris visada spėja, kad duomenys priklauso neigiamai klasei, gautume 99 % tikslumą ir tik 1 % klaidų dažnį. Nors šie įverčiai atrodo labai geri, mes žinome, kad toks klasifikatorius niekada neatpažintų teigiamos klasės duomenų. Todėl reikia naudotis kitokiomis vertinimo metrikomis dirbant su nesubalansuotais duomenimis.

1.1. Klasifikavimo matrica

Geresnis būdas vertinti klasifikatoriaus tikslumą dirbant nesubalansuotais duomenimis yra naudoti reikšmes iš klasifikavimo matricos. Klasifikavimo matrica – tai tokia matrica, kuri yra sudaryta iš N stulpelių ir N eilučių, kai duomenų aibėje yra N klasių. Dirbant su duomenimis, kur yra 2 klasės, tokia matrica turėtų du stulpelius ir dvi eilutes. Klasifikavimo matricos pavyzdys pateiktas paveikslėlyje nr. 1.



Pav. 1 Klasifikavimo matrica

Klasifikavimo matricos langeliai atspindi šias reikšmes:

- Tikrai teigiamas (TP) (angl. True Positive) – teigiamos klasės duomenų vektorius priskirtas teigiamai klasei.
- Tikrai neigiamas (TN) (angl. True Negative) – neigiamos klasės duomenų vektorius priskirtas neigiamai klasei.
- Klaidingai teigiamas (FP) (angl. False Positive) – neigiamos klasės duomenų vektorius priskirtas teigiamai klasei.
- Klaidingai neigiamas (FN) (angl. False Negative) – teigiamos klasės duomenų vektorius priskirtas neigiamai klasei.

Idealiu atveju, FP ir FN reikšmės būtų lygios nuliui. Toks klasifikatorius gebėtų puikiai atpažinti tiek teigiamas, tiek neigiamas klases.

Gavus šias reikšmes iš testavimo duomenų, galima apskaičiuoti įvairias metrikas, kurios mums leidžia geriau pažinti klasifikatoriaus galimybes ir įvertinti jo tikslumą.

Keletas metrikų, kurias galime apskaičiuoti, yra:

- Tikrai teigiamų dažnis arba jautrumas (angl. True Positive Rate (TPR), Recall) – matuoja kokia visų teigiamos klasės duomenų dalis buvo pažymėta kaip teigiama.

$$Recall = TPR = \frac{TP}{TP + FN} \tag{2}$$

- Klaidingai teigiamų dažnis (angl. False Positive Rate (FPR)) – matuoja kokia visų neigiamos klasės duomenų dalis buvo kaip teigiama.

$$FPR = \frac{FP}{TN + FP} \tag{3}$$

- Teigiamų spėjimų įvertis arba preciziškumas (angl. Positive Predictive Value (PPV), Precision) – matuoja kokia dalis duomenų, priskirtų teigiamai klasei, iš tikrųjų buvo teigiama.

$$Precision = PPV = \frac{TP}{TP + FP} \tag{4}$$

- F įvertis (angl. F-Measure) – matuoja klasifikavimo tikslumą pasinaudojant Recall ir Precision reikšmėmis. Koeficientas β nurodo Recall reikšmės svarbumą lyginant su Precision reikšme (dažniausiai $\beta = 1$).

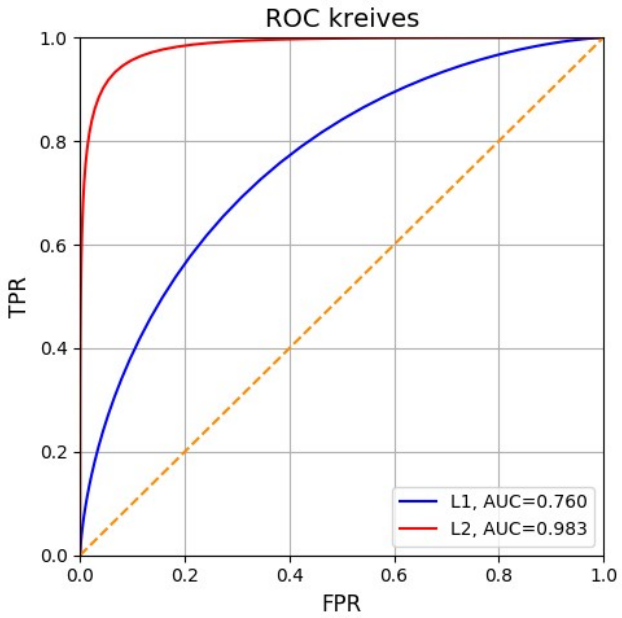
$$F\text{ - }Measure = \frac{(1+\beta)^2 \cdot Recall \cdot Precision}{\beta^2 \cdot Recall + Precision} \tag{5}$$

Kuo šių įverčių reikšmė didesnė, tuo geresnis klasifikatorius. Nors šios metrikos labiau tinka klasifikatorių vertinimui nei tikslumo ir klaidų dažnio (1), jų nepakanka norint susidaryti pilną vaizdą apie klasifikatoriaus kokybę. Pavyzdžiui, nėra aišku kaip palyginti skirtingus klasifikatorius tų pačių duomenų atpažinimui.

1.2. Sprendimus priimančiojo ypatybių kreivė (ROC)

Norint palyginti skirtingų klasifikatorių, dirbančių su tais pačiais duomenimis, kokybę, galima remtis sprendimus priimančiojo ypatybių kreive (angl. Receiver Operating Characteristic (ROC) Curve). Tai yra grafinis klasifikatorių kokybės pavaizdavimas koordinačių plokštumoje. Kreivė grafike vaizduoja sąryšį tarp anksčiau minėtų TPR (2) ir FPR (3) metrikų.

Dvejetainio klasifikavimo problemoje, klasifikatorius bandydamas atspėti ar duomenys priklauso teigiamai klasei apskaičiuoja tikimybę, kad priklauso tai klasei. Toks klasifikatorius taip pat turi ribą – fiksuota reikšmė, pagal kurią priima sprendimą kuriai klasei priklauso duomenys. Jei apskaičiuota tikimybė yra mažesnė negu toji riba, duomenys priskiriami prie neigiamos klasės, o jeigu tikimybė didesnė – prie teigiamos klasės. ROC kreivė vaizduoja sąryšį tarp FPR (x ašyje) ir TPR (y ašyje) reikšmių. Pagal šį grafiką galima parinkti klasifikavimo ribos reikšmę siekiant gauti norimus rezultatus.



Pav. 2: Grafikas su ROC kreivėmis.

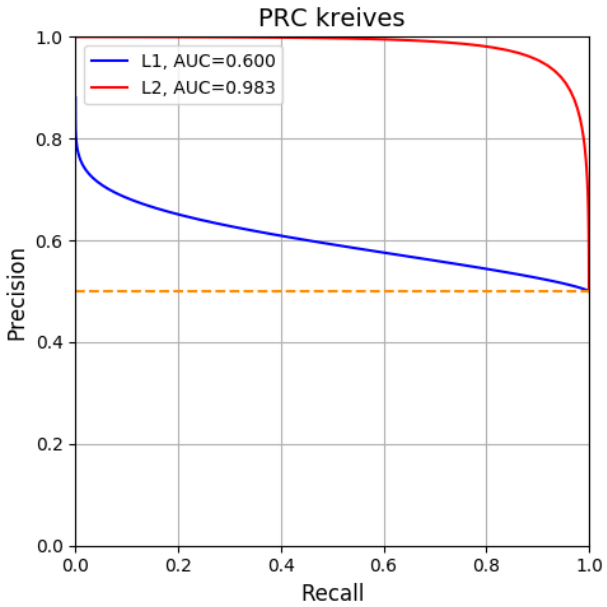
Paveikslėlyje nr. 2 matomas ROC kreivių grafikas. Remiantis šiuo grafiku galima palyginti du skirtingus klasifikatorius. Pirmojo klasifikatoriaus ROC kreivė pažymėta L1, antrojo – L2. Grafike esanti įstriža punktyrinė linija skirta pavaizduoti klasifikatoriaus, kuris sprendimus priima visiškai atsitiktinai, TPR ir FPR sąryšį. Kuo kreivė aukščiau virš įstrižainės linijos, tuo klasifikatoriaus tikslumas geresnis. Idealiu atvejų, kreivė turėtų eiti per tašką (0, 1) – tai reikštų, kad klasifikatorius teisingai atpažįsta visas teigiamas klases ir nei karto neklysta priskirdamas duomenys prie neigiamų klasių. Apibendrintą klasifikatoriaus kokybę galime įvertinti pagal ploto po kreive įvertį (angl. Area Under Curve (AUC)). Kuo AUC reikšmė didesnė, tuo geresnis klasifikatorius. Idealiu atveju AUC bus lygus 1, blogiausiu atveju bus lygus 0,5. Jei AUC reikšmė mažesnė nei 0,5 – tai nėra blogai, tiesiog reikėtų apkeisti klasifikatoriaus priskiriamų klasių reikšmes ir gautume AUC reikšmę didesne nei 0,5.

Iš paveikslėlio nr. 2 galime padaryti išvadą, kad klasifikatorius, kurio kreivė L2, dirba geriau nei klasifikatorius, kurio kreivė L1.

1.3. Preciziškumo-jautrumo kreivė (PRC)

Nors remiantis ROC kreivėmis galima daug ką sužinoti apie klasifikatoriaus kokybę, yra pastebėta, kad kai duomenys yra smarkiai nesubalansuoti, gali atsirasti netikslumų – tokiais atvejais klasifikatoriams vertinti labiau tinka preciziškumo-jautrumo kreivė (angl. Precision-Recall Curve (PRC)) [DG06].

PRC kreivė taip pat yra grafiškas klasifikatoriaus kokybės vaizdavimo būdas bei tinka palyginimui su kitais klasifikatoriais, dirbančiais su tais pačiais duomenimis. PRC kreivė vaizduoja sąryšį tarp Precision (PPV) (4) ir Recall (TPR) (2) reikšmių.



Pav. 3: Grafikas su preciziškumo-jautrumo kreivėmis.

Paveikslėlyje nr. 3 matome pavaizduotas dviejų klasifikatorių PRC kreives – L1 ir L2. Punktyrinė linija, einanti lygiagrečiai su x ašimi, vaizduoja atsitiktinai spėliojančio klasifikatoriaus kokybę. PRC kreivės nuo ROC kreivių skiriasi tuo, kad klasifikatoriaus kokybė yra geresnė kreivei esant arčiau viršutinio dešiniojo kampo. Plotas po PRC kreive, kaip ir ROC kreivėse, rodo bendrą klasifikatoriaus kokybę. Iš paveiksliuko galime teigti, kad klasifikatorius, kurio kreivė yra L2, duomenis klasifikuoja tiksliau nei klasifikatorius, kurio kreivė L1.

2. Duomenų balansavimo metodika

Literatūroje yra nemažai pavyzdžių, įrodančių, kad taikant tradicinius mašinų mokymo metodus nesubalansuotiems duomenims gali įsivelti įvairių skaičiavimo netikslumų, kurie blogina klasifikatoriaus kokybę [HG09]. Todėl ir šiame darbe aptarsime keletą pasirinktų duomenų balansavimo metodų, kurių veiksmingumas yra įrodytas.

2.1. Atsitiktinis duomenų išrinkinėjimas ir papildymas

Paprasčiausi metodai spręsti nesubalansuotų duomenų problemai – atsitiktinai išrinkinėti duomenis iš daugumos klasės (angl. Random Under-Sampling) arba atsitiktinai papildyti mažumos klasę (angl. Random Over-Sampling). Atsitiktinai išrinkinėjant duomenis siekiama, kad daugumos klasės duomenų kiekis sumažėtų iki mažumos klasės duomenų kiekio. Atsitiktinis mažumos klasės pildymas veikia paimant atsitiktinį duomenų įrašą, padarant jo kopiją ir pridedant prie mažumos klasės, visa tai tęsiama tol, kol mažumos klasės duomenų kiekis sutampa su daugumos klasės duomenų kiekiu. Nors šie metodai atrodo gana paprasti, kartais jie gali veikti ne ką prasčiau nei tie metodai, kurie protingiau atlieka balansavimą [BPM04]. Pagrindinis atsitiktinio duomenų išrinkinėjimo trūkumas yra tas, kad išrenkant duomenis yra prarandama informacija, tai reiškia, kad klasifikatorius gali neišmokti atpažinti kai kurių svarbių daugumos klasės duomenų bruožų [HAP89]. Su atsitiktiniu papildymu yra kita problema – padarius daugybę vienodų duomenų kopijų klasifikatorius gali persimokyti – pradėti atpažinti tik mokymo duomenis, o kitus realybėje sutinkamus duomenis imti priskirti netinkamai klasei [MWB07].

2.2. Sintetinių duomenų generavimas

Kitas metodas skirtas spręsti nesubalansuotų duomenų problemai – sintetinė perteklinio mažumos klasės duomenų papildymo technika (angl. Synthetic Minority Oversampling Technique(SMOTE)). SMOTE metodas gana plačiai taikomas, kadangi yra įrodyta, kad jis gali padėti pasiekti geresnių rezultatų [CBH+02].

$$x_{new} = x_i + (x_{ik} - x_i) \cdot \delta \tag{6}$$

SMOTE pasirinktam duomenų vektoriui iš mažumos klasės atranda K-artimiausių kaimynų (skaičiuojami Euklido erdvės atstumai tarp duomenų atributų). Norint sukurti naują sintetinį duomenų įrašą, atsitiktinai parenkamas vienas iš rastų K kaimynų, tada skaičiuojamas pirmiau pasirinkto duomenų įrašo ir atrinkto kaimyno atributų vektorių skirtumas. Gautas vektorius padauginamas iš atsitiktinio skaičiaus intervale [0, 1] ir galiausiai sudedamas su

pradinio duomenų vektoriaus atributais. Galutinis rezultatas ir bus naujo įrašo atributų vektorius. Visas šis procesas užrašomas formule (6). Veiksmai kartojami tol, kol mažumos klasės duomenų kiekis sutampa su daugumos klasės duomenų kiekiu.

Formulėje (6) x_{new} yra naujas įrašas, x_i – pradinis pasirinktas įrašas, x_{ik} – pradiniam įrašui atsitiktinai atrinktas K-artimiausias kaimynas, δ – atsitiktinis skaičius intervale [0, 1].

SMOTE metodas gali pagerinti klasifikavimo kokybę, tačiau jis turi ir trūkumų, dėl kurių yra galimybė atsirasti klaidoms klasifikavime. Taip nutinka dėl to, kad SMOTE gali per didelę svarbą suteikti kai kurių atributų reikšmėms arba kita priežastis – susidariusi didelė duomenų dispersija [WJ04].

3. Atsitiktinio miško algoritmas duomenų klasifikavimui

Šiame darbe duomenų klasifikavimui pasirinktas naudoti atsitiktinio miško metodas, kadangi jis tinka klasifikavimo uždaviniams, taip pat jis turi nedaug hiperparametrų (angl. Hyperparameter), o tai palengvina jo taikymą [Don18]. Atsitiktinis miškas yra sudarytas iš sprendimų medžių. Kiekvienam medžiui perduodamas duomenų vektorius, kurį apdoroję grąžina atsakymą. Pagal dažniausiai pasitaikantį medžių atsakymą priimamas galutinis sprendimas.

3.1. Sprendimų medis

Kad suprastume kaip veikia atsitiktinio miško metodas, turime suprasti kaip veikia jo komponentai – sprendimų medžiai. Sprendimų medis priima duomenų vektorių kaip įvestį ir tam tikrus patikrinimus priima sprendimą, kuriai klasei priklauso duomenų vektorius. Sprendimų medžio pavyzdys pateiktas paveikslėlyje nr. 4.

Sprendimų medžio veikimo principas gana aiškus. Dabar galime panagrinėti kaip konstruojamas toks medis. Duomenų klases galima atskirti pagal jų atributų reikšmes, taigi reikia sugalvoti sąlygą, pagal kurią medis galėtų lengviau priimti sprendimą apie duomenų klasę. Kad rastume optimalų sprendimą medžiui, turime optimizuoti informacijos gavimo reikšmę, kuri parodo kaip pasikeičiau duomenų entropija.

$$H(S) = -\sum_i p_i \log_2 p_i \tag{7}$$

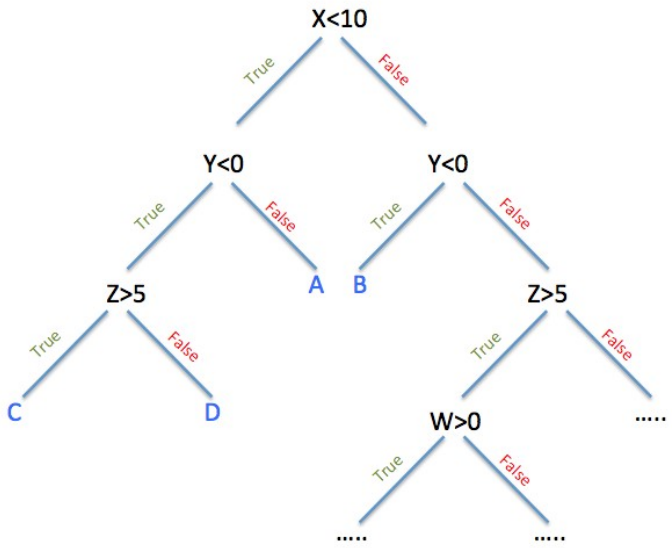
$$IG(S, a) = H(S) - H(S \mid \mathrm{a}) = -\sum_i p_i \log_2 p_i - \sum_{v \in Values(a)} \frac{|S_v|}{|S|} H(S_v) \tag{8}$$

Entropija skaičiuojama pagal formule (7), joje S žymi nagrinėjamą sistemą, p_i – i-tojo sistemos elemento kiekio santykį su likusių sistemos elementų kiekiu. Entropija matuojamas sistemos netvarkingumas. Netvarkinga sistema yra tokia sistema, apie kurią negalima pateikti jokių tikslių išvadų. Pavyzdžiui, turint maišą, kuriame yra 3 balti ir 3 juodi kamuoliukai, negalėtume garantuoti, kad labiau tikėtina ištraukti juodą kamuoliuką nei baltą ar atvirkščiai – tokia sistema yra netvarkinga. Kada entropija būna lygi 0, tai reiškia, kad sistema yra tvarkinga. Traukiant kamuoliuką iš maišo su 3 baltais kamuoliukais nekiltų abejonių, kad ištrauktumėme baltą kamuoliuką – tokia sistema tvarkinga.

Informacijos gavimas skaičiuojamas pagal (8) formulę. Informacijos gavimo reikšmė parodo, kokią informacijos vertę gautume, jeigu nagrinėjamą sistemą skaidytume pagal

atributą *a*, kaip pažymėta formulėje. Taip yra todėl, kad iš pradinės sistemos entropijos atimame naujai gautų entropijų sumą.

Pagal informacijos gavimo formulę galime atrinkti tokį atributą *a*, pagal kurį skaidydami sistemą sumažintume entropiją labiausiai. Pagal šį atributą ir konstruojama sąlyga. Šis procesas kartojamas tol, kol gaunamų naujų sistemų entropijos netampa lygios 0.



Pav. 4: Sprendimų medžio pavyzdys.

4. Tyrimas su realiais nesubalansuotais duomenimis

Išnagrinėjome keletą duomenų balansavimo metodų, klasifikatoriaus mokymą pagal atsitiktinio miško algoritmą bei kaip vertinti klasifikatoriaus, dirbančio su nesubalansuotais duomenimis, kokybę. Dabar galime aptarti patį tyrimą.

Tyrimo tikslas yra pamatyti kokią įtaką nesubalansuotiems duomenims padaro jų balansavimas naudojant minėtus balansavimo metodus ir ar visada yra naudinga atlikti duomenų balansavimą.

4.1. Tyrimo duomenys

Tyrimui atlikti buvo panaudoti bankinių transakcijų duomenys iš Kaggle.com internetinio puslapio. Kiekviena transakciją turi 31 atributą, iš kurių 30 yra skaičiaus pavidalo ir 1 kategorinis atributas. Detalesnė informacija apie transakcijos atributus pateikta lentelėje nr. 1.

Lentelė 1. Transakcijų atributų lentelė su paaiškinimais.

Pavadinimas	**Paaiškinimas**
Laikas (Time)	Laiko atributas yra realus skaičius didesnis už arba lygus 0. Jis pasako, kiek sekundžių praėjo nuo pirmosios transakcijos šioje duomenų aibėje užfiksavimo laiko iki nagrinėjamos transakcijos užfiksavimo laiko.
Kiekis (Amount)	Kiekio atributas pasako koks pinigų kiekis buvo pervestas per šią transakciją. Tai yra realus skaičius didesnis už arba lygus 0.
Klasė (Class)	Klasės atributas yra vienintelis kategorinis atributas šioje duomenų aibėje. Visoje duomenų aibėje pasitaiko dvi skirtingos klasės – autentiška arba apgaulinga. Autentiškos klasės žymimos „0“, apgaulingos – „1“.
V1, …, V28	Dėl konfidencialumo priežasčių šie atributai „užmaskuoti“ – jų reikšmės buvo normalizuotos ir transformuotos pagal principinių komponentų analizės (angl. Principle Component Analysis (PCA)) metodą. Gautos reikšmės po transformacijos yra tarpusavyje koreliacijų neturintys realūs skaičiai.

Transakcijų duomenų aibę sudaro 284807 transakcijos, tarp kurių yra 284315 (99,828 %) autentiškos (žymimos „0“) ir 492 (0,172 %) apgaulingos (žymimos „1“). Pastebime, kad duomenys yra smarkiai nesubalansuoti. Transakcijų pavyzdžiai pateikti lentelėje nr. 2.

Lentelė 2: Transakcijų duomenų pavyzdžiai.

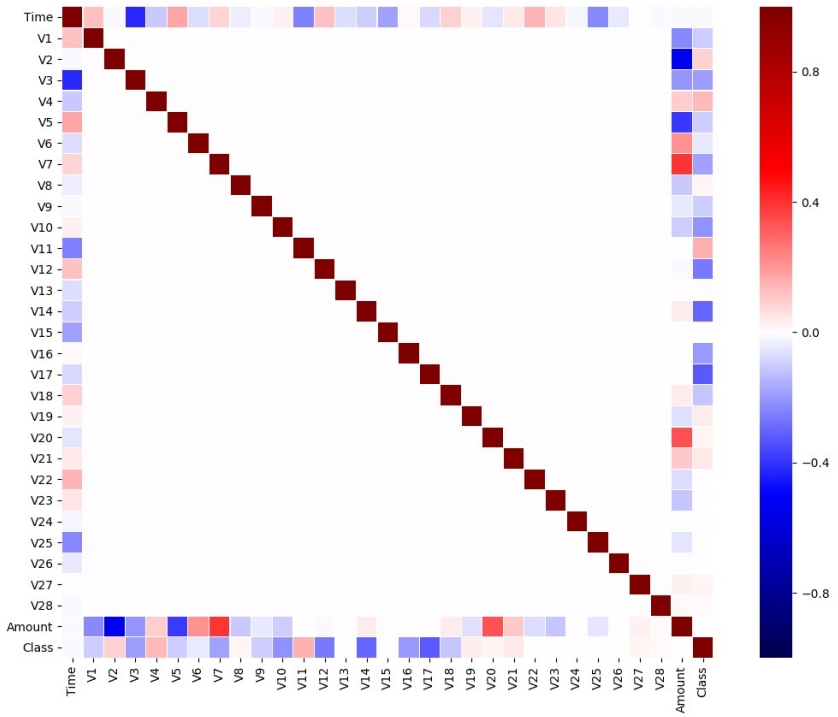
#	Time	V1	V2	V3	...	V27	V28	Amount	Class
0	0.0	-1,36	-0,072	2,536	...	0,134	-0,021	149,62	0
1	0.0	1,19	0,266	0,166	...	-0,009	0,014	2,69	0
2	1.0	-1,36	-1,340	1,773	...	-0,055	-0,059	378,66	0
3	1.0	-0,966	-0,185	1,793	...	0,062	0,061	123,50	0
4	2.0	-1,159	0,877	1,549	...	0,219	0,215	69,99	0

Žinome, kad siekiant užtikrinti originalių duomenų konfidencialumą, atributai V1, …, V28 „užmaskuoti“ – originalūs atributai buvo normalizuoti ir transformuoti pagal principinių komponentų analizės (PCA) metodą, po kurios buvo gauti principiniai komponentai (angl. Principal Component), kurie tarpusavyje neturi koreliacijų. Šių atributų koreliacijos nebuvimas pavaizduotas koreliacijų matricoje paveikslėlyje nr. 5.

Koreliacijų matricos langeliai įgija reikšmes intervale [-1, 1], bet dėl aiškesnio atvaizdavimo jie pažymėti spalvomis:

- Raudona – reiškia teigiamą koreliaciją, kuo tamsesnė raudona spalva, tuo koreliacija stipresnė.
- Mėlyna – reiškia neigiamą koreliaciją, kuo tamsesnė mėlyna spalva, tuo koreliacija stipresnė.
- Balta – reiškia koreliacijos nebuvimą.

Kadangi atributai V1, …, V28 neturi koreliacijų tarpusavyje, šiems transakcijų duomenims nebūtų naudinga taikyti mokymo metodus, kurie naudojasi koreliacija, pavyzdžiui Bajeso tinklus.



Pav. 5: Transakcijų atributų koreliacijų matrica.

4.2. Tyrimo rezultatai

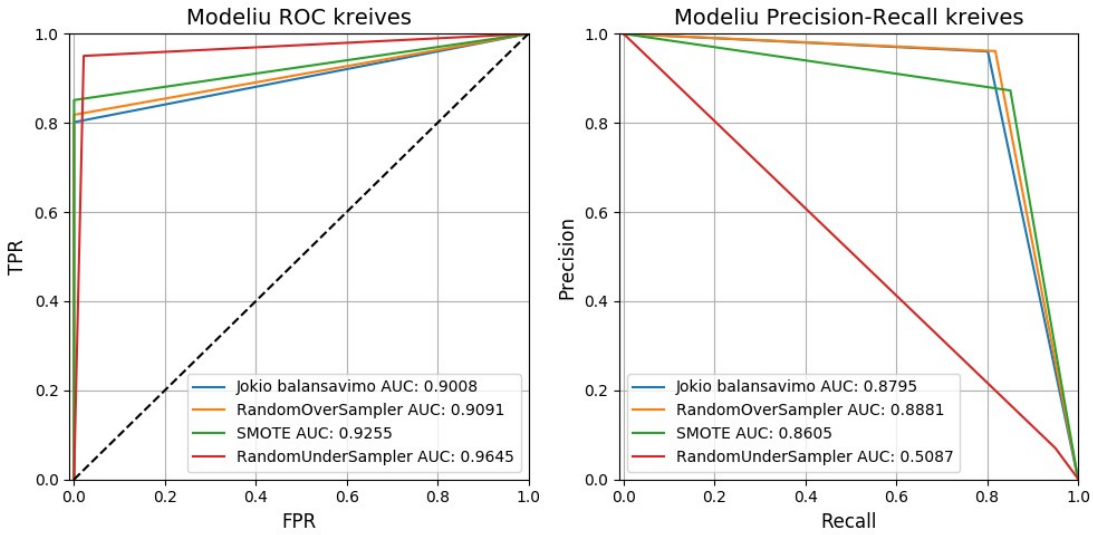
Tyrimas atliktas naudojantis atsitiktinio miško mokymo metodu. Atsitiktinis miškas generuojamas su 10 sprendimų medžių, kurie generuojami pagal informacijos gavimo formulę (8). Pateikiami keturių skirtingų gautų modelių rezultatai: mokantis be duomenų balansavimo, mokantis su subalansuotais duomenims pritaikius atsitiktinio išrinkimo, atsitiktinio papildymo ir SMOTE balansavimo metodus. Transakcijų duomenys paskirstyti į mokymo aibę ir testavimo aibę. Mokymo aibė sudaryta iš 75% (213605 vnt.) pradinių duomenų, testavimo – iš likusių 25% (71202 vnt.). Pateikiami išmokytų modelių spėjimų rezultatai (lentelė nr. 3) bei pateikiami modelių vertinimo metrikų įverčiai (lentelė nr. 4), aprašyti modelių vertinimo skyriuje, taip pat pateikiami ROC ir PR kreivių grafikai (pav. nr. 6).

Lentelė 3: Modelių rezultatai su transakcijų duomenimis.

Modelis	TP	FP	TN	FN
Jokio balansavimo	97	4	71077	24
Random Over-Sampling	99	4	71077	22
SMOTE	103	15	71066	18
Random Under-Sampling	115	1522	69599	6

Lentelė 4: Modelių įverčiai. Geriausi stulpelių rezultatai paryškinti.

Modelis	TPR	FPR	Precision	F1	AU ROC	AU PRC
Jokio balansavimo	0,802	**5,6e-05**	0,960	0,873	0,901	0,880
Random Over-Sampling	0,818	**5,6e-05**	**0,961**	**0,884**	0,909	**0,888**
SMOTE	0,851	2,1e-04	0,873	0,862	0,923	0,861
Random Under-Sampling	**0,950**	0,021	0,070	0,131	**0,965**	0,509



Pav. 6: Modelių ROC ir PR kreivių grafikai.

4.3. Tyrimo su realiais duomenimis rezultatų apibendrinimas

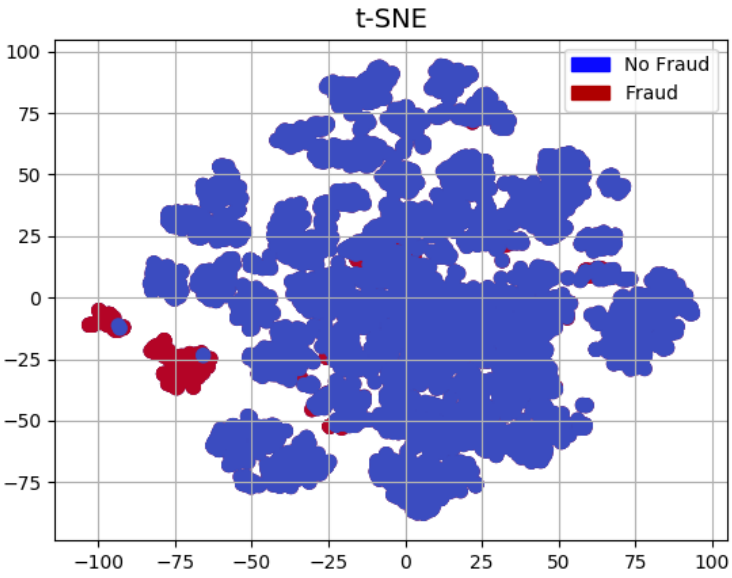
Pažvelgę į gautus modelių rezultatus pastebėsime, kad modelis mokytas netaikant balansavimo metodo transakcijas atpažįsta ne ką prasčiau nei kiti modeliai. Jeigu remtumėmės vien tik ROC kreivių AUC įverčiu, galėtume pagalvoti, kad Random Under-Sampling ir SMOTE modeliai yra pranašesni už kitus du. Tačiau PR kreivių grafikas rodo kiek kitokį vaizdą – Random Under-Sampling modelio rezultatas yra ženkliai prastesnis nei kitų modelių,

o SMOTE trečioje vietoje. Tai yra priežastis, kodėl geriau naudoti PR kreivių grafiką nei ROC kreivių grafiką dirbant su nesubalansuotais duomenimis – ROC kreivės pateikia per daug optimistiškus rezultatus. Rezultatų lentelėje (nr. 3) matome, kad Random Under-Sampling modelis net 1522 transakcijas palaikė apgaulingomis, o tai gana prastas rezultatas. SMOTE ir Random Over-Sampling modeliai veikia gana panašiai kartu su modeliu, kuris mokėsi su nesubalansuotais duomenimis.

Iš rezultatų galime įsitikinti, kad balansavimo metodai nesuteikė jokio ženklaus pranašumo prieš mokymąsi su nesubalansuotais duomenimis, nors literatūroje ganėtinai dažnai akcentuojama, kad susidūrus su nesubalansuotų duomenų problema, būtina subalansuoti duomenų klases. Tokie atvejai nėra dažni [HE09], tačiau susidūrus nesubalansuotais duomenimis visai neprasta mintis pasižiūrėti ar tikrai būtina balansuoti duomenis.

Yra pastebėta, kad mokymosi algoritmas su nesubalansuotais duomenimis gali veikti taip pat gerai ar net geriau nei taikant balansavimo metodus tam tikru atveju – kai tarp duomenų klasių nėra arba yra minimalus persidengimas bei esant tokių pavienių duomenų mažumos klasėje, kurie labiau panašūs į daugumos klasės duomenis [BPM04]. Iš tikrųjų galime patikrinti ar šis atvejis pasitaiko ir su tyrime naudotais transakcijų duomenimis. Paprasčiausia tuo įsitikinti būtų grafiškai atvaizduojant duomenų aibes.

Kaip žinome, transakcijų duomenis sudaro 31 atributas, tai norint atvaizduoti šiuos duomenis reikėtų piešti 31-matės erdvės grafiką, o tai nebūtų nei paprasta, nei neaišku kokį piešinį gautume. Tačiau šiai problemai spręsti yra kitas būdas – galima duomenis klasterizuoti, t. y. pirmiausia redukuoti duomenų dimensiją iki 2-matės erdvės ir nesudėtingai atvaizduoti duomenų taškus plokštumoje. Klasterizaciją atlikti galima pasinaudojant t-skirstinio stochastiniu kaimynų įterpimo metodu (angl. t-distributed Stochastic Neighbour Embedding (t-SNE)). Šiame darbe šis metodas pasirinktas todėl, kad jis sugeba labai aiškiai klasterizuoti duomenis lyginant su kitais daugiamačių duomenų vizualizavimo metodais [MH08]. Vienintelis t-SNE trūkumas, kad jis atlieka labai daug skaičiavimų ir ilgai užtrunka, todėl šis grafikas buvo generuotas tik su 10000 (3,5%) daugumos klasės duomenų. Tačiau to tiek duomenų užtenka, kad galėtume pastebėti svarbiausias duomenų savybes.



Pav. 7: Transakcijų duomenys klasterizuoti pagal t-SNE metodą.

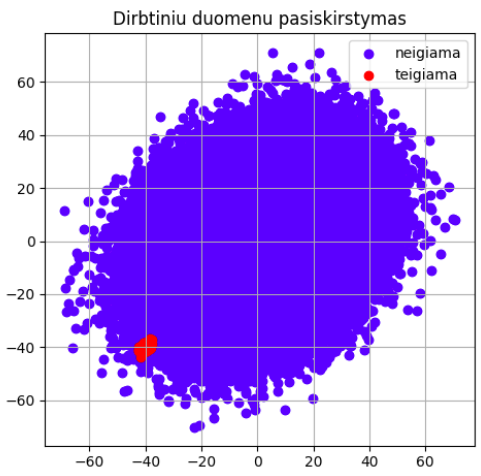
Paveikslėlyje nr. 7 pateiktas transakcijų duomenų klasterizavimo grafikas pasitelkus t-SNE. Grafike raudonai pažymėti taškai reiškia apgaulingas transakcijas, mėlynai – autentiškas. Vienas iš keleto dalykų, kuriuos galime pastebėti grafike, yra tai, kad duomenys triukšmingi – grafike nesimato jokių koreliacijų tarp taškų koordinačių. Ir tai nėra keista, kadangi kaip matėme koreliacijų matricoje (pav. nr. 5) didžioji dauguma duomenų iš tikrųjų tarpusavyje neturi jokių koreliacijų. Kitas dalykas yra tas, kad skirtingų klasių duomenys beveik nepersidengia – apgaulingos transakcijos yra aiškiai nutolusios nuo autentiškų, neskaitant kelių kraštutinių atvejų. Kaip jau minėjome, esant tokiai savybei, galima sukurti modelį, kuris gerai gebėtų atpažinti duomenys net ir nepritaikius jokių duomenų balansavimo metodų.

5. Tyrimai su dirbtiniais duomenimis

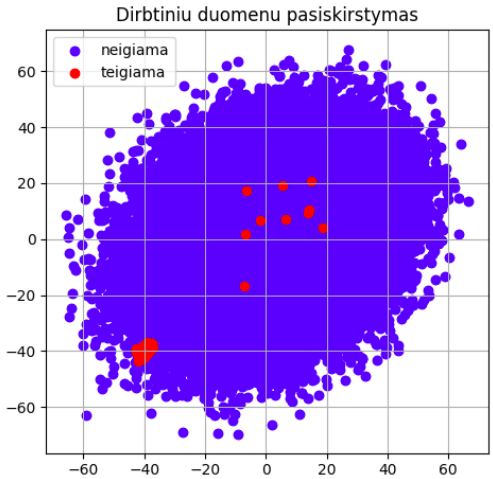
Po atlikti tyrimo su realiais duomenimis, galime įsitikinti, kad pasitaiko atvejų, kada balansavimo metodai nepagerina klasifikavimo kokybės ar net padaro prastesne. Siekiant geriau suprasti, kada tokie atvejai pasitaiko, atliktas tyrimas su sugeneruotais dirbtiniais duomenimis. Tyrime su realiais duomenimis pamatėme, kad klasifikatorius mažumos klasę gerai, kai jos duomenys nepersidengia arba mažai persidengia su daugumos klase.

Šiame skyriuje išnagrinėsime tris skirtingus atvejus su dirbtiniais duomenimis, kai mažumos ir daugumos klasių duomenys persidengia. Duomenų aibę iš viso sudaro 200000 dvimačių duomenų, o santykis tarp mažumos ir daugumos klasių duomenų kiekio yra 1:1000. Pirmajame atvejyje (pav. nr. 8) duomenų klasės generuotos pagal Gauso skirstinį, o mažumos klasė yra viename duomenų klasteryje. Antrajame atvejyje (pav. nr. 9) duomenys generuoti kaip ir pirmuoju atveju, tik šį kartą papildomai atsiranda išsibarsčiusių mažumos klasės duomenų po daugumos klasę. Trečiajame atvejyje (pav. nr. 10) mažumos klasė yra pasidalinusi į skirtingus duomenų klasterius.

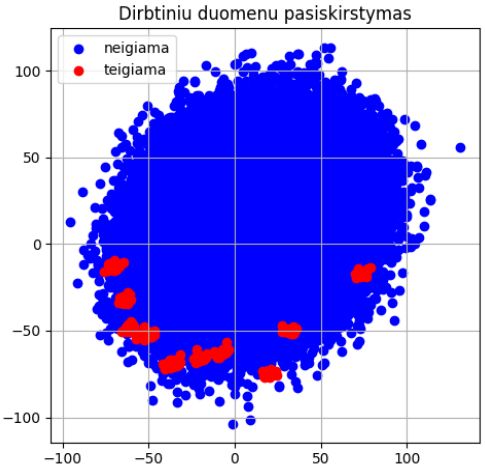
Kiekvienam tyrimui naudojami tie patys balansavimo metodai: atsitiktinis išrinkimas (Random Under-Sampling), atsitiktinis perrinkimas (Random Over-Sampling) bei SMOTE. Taip pat sukurtas modelis, kuris mokėsi su duomenimis netaikant balansavimo metodo. Mokymo algoritmas yra atsitiktinio miško metodas, kuriame yra 10 skirtingų sprendimų medžių, kurių struktūra buvo generuota taikant informacijos gavimo formulę. Pateikiami tyrimų modelių rezultatai, jų įverčiai bei ROC ir PR kreivių grafikai. Apibendrinami tyrimų rezultatai bei aprašomi pastebėjimai.



Pav. 8: Mažumos klasė viename klasteryje.



Pav. 9: Mažumos klasės klasteris ir išsibarstę taškai.



Pav. 10: Mažumos klasės klasteriai.

5.1. Tyrimas, kai balansavimo metodai neturi įtakos

Šiuo tyrimu buvo siekta parodyti, kad modelis mokydamasis su nesubalansuotais duomenimis, kai mažumos klasės duomenys persidengia su daugumos klasės duomenimis, gali gerai išmokti atpažinti šiuos duomenis ir gali veikti taip pat gerai ar geriau nei modeliai, kurie mokėsi atpažinti balansavimo metodų apdorotus duomenis.

Modelių įverčių lentelėje (nr. 6) matome, kad Random Under-Sampling modelis atpažino visus mažumos klasės duomenis, tačiau kaip matome modelių rezultatų lentelėje (nr. 5), jis net 521 daugumos klasės duomenų neteisingai priskyrė prie mažumos klasės. Tai labai didelis klaidų skaičius, todėl galime teigti, jog šis modelis pasirodė prasčiausiai. Taip nutiko todėl, kad mokymo procese buvo išmesti dideli kiekiai duomenų apie daugumos klasę, todėl modelis nesugebėjo tinkamai išmokti atpažinti jos duomenų.

Modelis, mokydamasis dirbti su duomenimis, neapdorotais jokių balansavimo metodų, atpažino mažiausiai mažumos klasės duomenų, tačiau jis tik vieną kartą daugumos klasės duomenų įrašą priskyrė prie mažumos klasės. Pasirodo, kad net ir persidengiant klasėms, duomenų balansavimas ne visada yra būtinas norint gauti gerus rezultatus. Taip nutinka, kai visi arba dauguma mažumos klasės duomenų patenka į vieną duomenų klasterį.

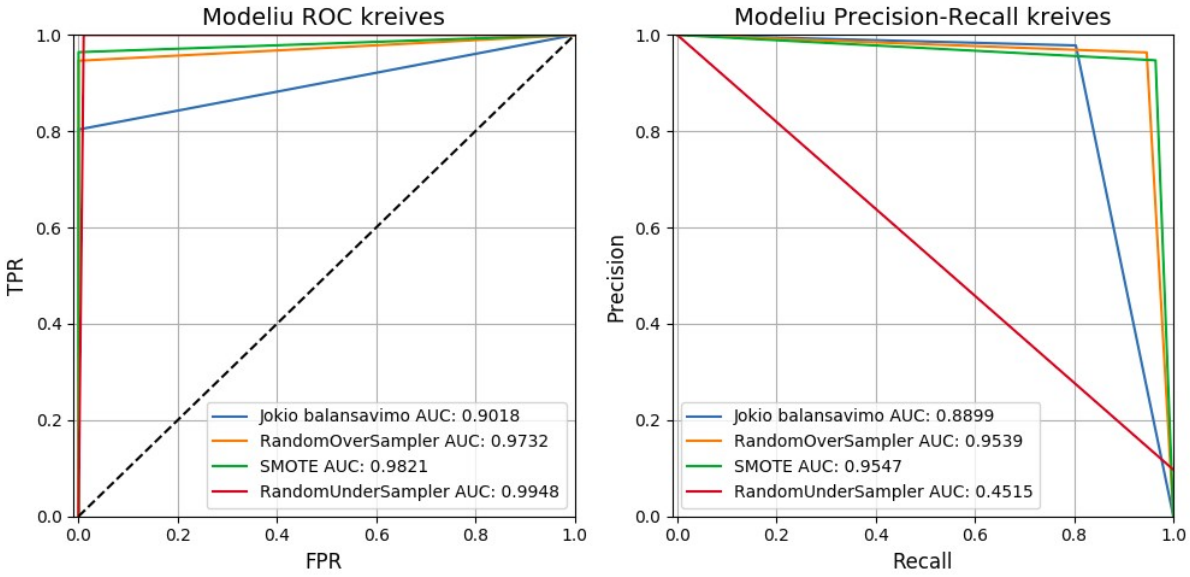
Random Over-Sampling ir SMOTE modeliai pasirodė geriausiai šiame tyrime, jie abu suklydo tik 5 kartus. Tai reiškia, kad balansavimo metodai turėjo teigiamą įtaką, tačiau jų nepakako, kad duomenys būtų idealiai atpažinti.

Lentelė 5: Modelių rezultatai, kai mažumos klasė viename klasteryje.

Modelis	TP	FP	TN	FN
Jokio balansavimo	45	1	49943	11
Random Over-Sampling	53	2	49942	3
SMOTE	54	3	49941	2
Random Under-Sampling	56	521	49423	0

Lentelė 6: Modelių įverčiai, kai mažumos klasė viename klasteryje.

Modelis	TPR	FPR	Precision	F1	AU ROC	AU PRC
Jokio balansavimo	0,803	**2e-05**	**0,978**	0,882	0,901	0,890
Random Over-Sampling	0,946	4e-05	0,964	0,955	0,973	0,954
SMOTE	0,964	6e-05	0,947	**0,956**	0,982	**0,955**
Random Under-Sampling	**1,000**	0,010	0,097	0,177	**0,995**	0,452



Pav. 11: Modelių ROC ir PR kreivių grafikai.

5.2. Tyrimas, kai krenta SMOTE kokybė

Dabar panagrinėsime atvejį, kai tarp mažumos klasės yra išsibarsčiusių pavienių duomenų, kurie persidengia su daugumos klase ir kokia įtaką tai turi mokymo modeliams.

Daugiausiai mažumos klasės duomenų atpažino Random Under-Sampling modelis kaip matome iš lentelių nr. 7 ir nr. 8, tačiau jo atotrūkis nuo kitų modelių yra minimalus pagal šį vertinimą. Tiek Random Under-Sampling, tiek SMOTE modeliai suklydo daugybę kartų neteisingai priskirdami daugumos klasės duomenis prie mažumos klasės – 1132 ir 495 kartus

atitinkamai. Random Under-Sampling modelis klydo dėl tos pačios priežasties kaip ir pirmajame tyrime su dirbtiniais duomenimis. SMOTE kokybė, priešingai nei pirmajame tyrime, krito ženkliai – taip nutiko todėl, kad balansavimo metodas sugeneravo daug naujų duomenų tarp išsibarsčiusių taškų grupės, kur realiai jų neturėtų būti visai. Taigi, taikant SMOTE metodą turėtume būti atsargūs, kad nepasitaikytų tokių duomenų išsibarstymų. Jeigu jų pasitaiko, vienas būdas su tuo kovoti yra juos pašalinti visiškai, taip pagerinant mokymo modelio kokybę.

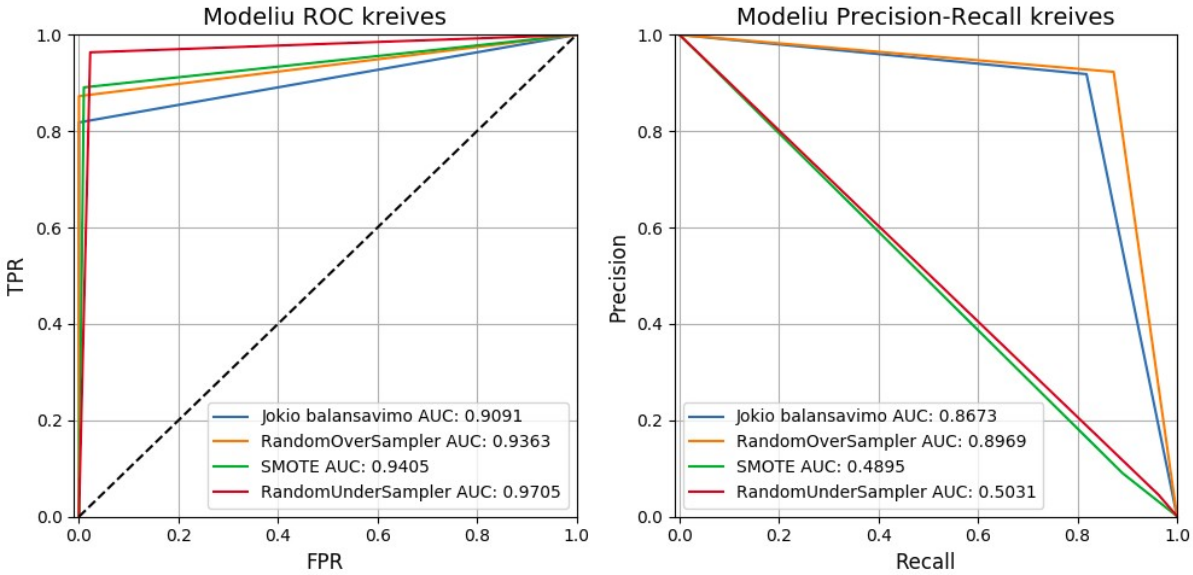
Random Over-Sampling modelis bei modelis be jokio balansavimo parodė geriausius rezultatus – abu klydo mažai kartų ir atpažino didžiąją dalį mažumos klasės duomenų. Jie abu klydo daugiau kartų nei pirmajame tyrime, kadangi nesugebėjo atpažinti išsibarsčiusių mažumos klasės duomenų. Tačiau likusius duomenis iš mažumos klasės klasterio išmoko atpažinti ganėtinai gerai.

Lentelė 7: Modelių rezultatai, kai mažumos klasė viename klasteryje bei turi išsibarsčiusių pavienių taškų.

Modelis	**TP**	**FP**	**TN**	**FN**
Jokio balansavimo	45	4	49941	10
Random Over-Sampling	48	4	49941	7
SMOTE	49	495	49450	6
Random Under-Sampling	53	1132	48813	2

Lentelė 8: Modelių įverčiai, kai mažumos klasė viename klasteryje bei turi išsibarsčiusių pavienių taškų.

Modelis	**TPR**	**FPR**	**Precision**	**F1**	**AU ROC**	**AU PRC**
Jokio balansavimo	0,818	**8e-05**	0,918	0,865	0,909	0,867
Random Over-Sampling	0,873	**8e-05**	**0,923**	**0,897**	0,936	**0,900**
SMOTE	0,891	0,010	0,090	0,164	0,941	0,490
Random Under-Sampling	**0,964**	0,023	0,045	0,085	**0,971**	0,503



Pav. 12: Modelių ROC ir PR kreivės.

5.3. Tyrimas, kai SMOTE kokybė geriausia

Tiek tyrime su realiais bankinių transakcijų duomenimis, tiek pirmajame bei antrajame tyrime su dirbtiniais duomenimis, nematėme atvejo, kad balansavimo metodai turėtų teigiamą poveikį mokymuisi. Todėl šiuo trečiuoju tyrimu su dirbtiniais duomenimis siekta pamatyti, kada yra naudinga balansuoti duomenis.

Pažvelgę į rezultatų lenteles nr. 9 ir nr. 10, pamatysime kiek kitokius rezultatus nei ankstesniuose tyrimuose. Visi modeliai, išskyrus Random Under-Sampling modelį, padarė gerokai daugiau klaidų atpažįstant duomenis. Esant mažumos klasės duomenų pasiskirstymui į daug skirtingų klasterių, modeliams sunkiau aptikti panašumus, leidžiančius geriau atpažinti mažumos klasę.

Random Under-Sampling sugebėjo atpažinti visus mažumos klasės duomenis, tačiau neteisingai atpažintų daugumos klasės duomenų kiekis ir vėl didelis, dėl priežasties aprašytos pirmojo tyrimo su dirbtiniais duomenimis skyrelyje.

Iš PR kreivių grafiko matome (pav. nr. 13) matome, kad šįkart geriausiai pasirodė SMOTE modelis. Priešingai nei antrame tyrime su dirbtiniais duomenimis, kad mažumos klasės duomenys buvo išsibarstę pavieniui, šįkart mažumos klasė buvo skirtinguose klasteriuose. Dėl šios priežasties, SMOTE generuodamas dirbtinius duomenis, nesukūrė jų ten, kur jų neturėtų būti. Šis modelis neteisingai atpažino 39 daugumos klasės duomenis – taip įvyko todėl, kad jie pateko į mažumos klasės klasterius. Vienas būdas spręsti šią problemą būtų pašalinti

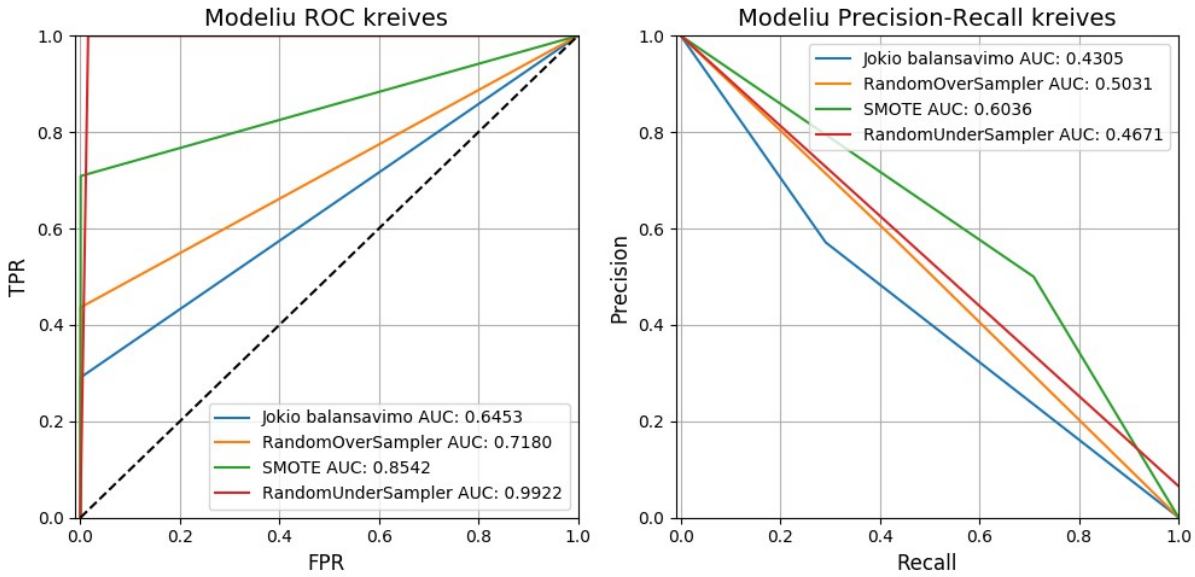
daugumos klasės duomenis, kurie patenka tarp mažumos klasės klasterių, tačiau tai galėtų pakenkti daugumos klasės duomenų atpažinimo kokybei.

Lentelė 9: Modelių rezultatai, kai mažumos klasė skirtinguose klasteriuose.

Modelis	TP	FP	TN	FN
Jokio balansavimo	16	12	49933	39
Random Over-Sampling	24	18	49927	31
SMOTE	39	39	49906	16
Random Under-Sampling	55	781	49164	0

Lentelė 10: Modelių įverčiai, kai mažumos klasė skirtinguose klasteriuose.

Modelis	TPR	FPR	Precision	F1	AU ROC	AU PRC
Jokio balansavimo	0,291	**2,2e-04**	**0,571**	0,386	0,645	0,431
Random Over-Sampling	0,436	3,6e-04	**0,571**	0,495	0,718	0,503
SMOTE	0,710	7,8e-04	0,500	**0,586**	0,854	**0,604**
Random Under-Sampling	**0,984**	0,016	0,066	0,123	**0,992**	0,467



Pav. 13: Modelių ROC ir PR kreivės.

6. Išvados

Skaitant įvairius šaltinius apie mašinų mokymo metodus ir nesubalansuotus duomenis, dažnai yra akcentuojama, jog klasių duomenų disbalanso gali įsivelti daugybė klaidų treniruojant modelį ir tai gali kenkti klasifikavimo kokybei. Tačiau iš šiame darbe aprašytų tyrimų galime įsitikinti, jog tai ne visada yra tiesa. Į klausimą ar reikia balansuoti duomenis, ar nereikia, nėra vieno paprasto atsakymo – tai priklauso nuo pačių duomenų. Taigi, rekomendacija ateityje susidūrus su nesubalansuotais duomenimis būtų pirmiausia juos atidžiai ištirti ir apgalvoti ar tikrai verta juos balansuoti, galbūt yra geresnių strategijų kaip pasiekti geresnius klasifikavimo tikslumus.

Darbe aprašyti tyrimai parodė, kad atsitiktinio išrinkimo (angl. Random Under-Sampling) metodas gali labai pakenkti klasifikavimo tikslumui. Tačiau tai neįrodo, jog šis metodas yra netinkamas naudoti. Metodo veiksmingumas priklauso nuo pačių duomenų ir yra atliktą tyrimų, kurie parodo, kad ir atsitiktinai išrinkinėjant duomenis iš daugumos klasės galima pasiekti geresnių rezultatų, negu visai netaikant balansavimo metodo [WP01].

Kitas dalykas, pastebėtas atliktuose tyrimuose, yra tai, kad atsitiktinio perrinkimo (angl. Random Over-Sampling) metodas kiekvieno tyrimo atveju veikė taip pat gerai arba geriau nei klasifikavimo modelis gautas treniruojant su nesubalansuotais duomenimis. Taigi atsitiktinio perrinkimo strategija dažnai yra geresnė strategija nei mokymas be jokio balansavimo metodo. Visgi, šio metodo taikymas nėra garantija, kad gauti rezultatai bus geresni, kadangi įmanomi atvejai, kai vienas ar keli mažumos duomenų klasės duomenų įrašai gali įsivelti į daugumos klasės duomenų klasterį. Jei ten atsirastų daug papildomų mažumos klasės kopijų, tai galėtų turėti neigiamą įtaką klasifikavimo modelio rezultatams.

Šia tema galima būtų atlikti daugiau tyrimų, panagrinėti daugiau atvejų kada balansavimas yra būtinas ir kokiais atvejais kuris balansavimo metodas yra pranašiausias ar atvirkščiai. Taip pat būtų galima atlikti tyrimus kombinuojant įvairias duomenų apdorojimo strategijas, siekiant pašalint įvairias anomalijas duomenyse. Be viso to, yra ir daugiau duomenų balansavimo strategijų, kurias taip pat galima būtų išnagrinėti siekiant pastebėti kiekvienos jų pranašumus ir trūkumus.

Šaltiniai

[DG06] J. Davis and M. Goadrich, “The Relationship between Precision-Recall and ROC Curves,” Proc. Int’l Conf. Machine Learning, pp. 233-240, 2006.

[MWB07] D. Mease, A.J. Wyner, and A. Buja, “Boosted Classification Trees and Class Probability/Quantile Estimation,” J. Machine Learning Research, vol. 8, pp. 409-439, 2007.

[HAP89] R.C. Holte, L. Acker, and B.W. Porter, “Concept Learning and the Problem of Small Disjuncts,” Proc. Int’l J. Conf. Artificial Intelligence, pp. 813-818, 1989.

[CBH+02] N.V. Chawla, K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer, “SMOTE: Synthetic Minority Over-Sampling Technique,” J. Artificial Intelligence Research, vol. 16, pp. 321-357, 2002.

[WJ04] B.X. Wang and N. Japkowicz, “Imbalanced Data Set Learning with Synthetic Samples,” Proc. IRIS Machine Learning Workshop, 2004.

[Don18] Donges, N. T*he Random Forest Algorithm*. (2018) https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd

[BPM04] G.E.A.P.A. Batista, R.C. Prati, and M.C. Monard, “A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data,” ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, pp. 20-29, 2004.

[MH08] van der Maaten, L.J.P.; Hinton, G.E. (Nov 2008). *"Visualizing Data Using t-SNE"*. Journal of Machine Learning Research. **9**: 2579–2605.

[WP01] G.M. Weiss and F. Provost, “*The Effect of Class Distribution on Classifier Learning: An Empirical Study,”* Technical Report ML-TR-43, Dept. of Computer Science, Rutgers Univ., 2001