



**Vilnius
universitetas**

Informatikos inžinerijos krypties doktorantų konferencija
Veiklos ataskaita už 2025 m. kovo 29d. – 2025 spalio 3d.

Feature Conversion for a Better Imbalanced Data Classification: A Financial Fraud Detection Case

(Požymių konversija siekiant gerinti nesubalansuotų duomenų klasifikavimą: finansinio sukčiavimo atvejis)

Dalia BRESKUVIENĖ– Informatikos inžinerija T 007 doktorantė

Darbo vadovas – prof. habil. dr. Gintautas DZEMYDA

Doktorantūros pradžios ir pabaigos metai: 2021.12.01 – 2025.11.30



2021–2025m.

Vilniaus
universitetas

STUDIJŲ PLANAS IR JO VYKDYMO SUVESTINĖ

Studijų metai	Egzaminai ¹	
	Planas	Įvykdyta
I (2021/2022)	3	3
II (2022/2023)	1	1
III (2023/2024)		
IV (2024/2025)		
Iš viso:	4	4

Studijų metai	Dalyvavimas konferencijose				Publikacijos					
	Tarptautinėse		Nacionalinėse		Su citav. rodikliu			Be citav. rodiklio		
	Planas	Įvykdyta	Planas	Įvykdyta	Planas	Įvykdyta	Būklė ⁴	Planas	Įvykdyta	Būklė ⁴
I (2021/2022)			1	1						
II (2022/2023)	1	1		1		1	Publikuota	1	1	Publikuota
III (2023/2024)	1	1		1	1					
IV (2024/2025)				1	1	1	Publikuota			
Iš viso:	2	2	1	4	2	2		1	1	

Doktorantūros studijų pasiekimai

Vilniaus
universitetas

Dalyvavimas tarptautinėse konferencijose

1.	D.Breskuvienė, G.Dzemyda „Clustering-based optimization in fraud detection classifier training“ EURO2022 2022.07.03 /06 ESPOO, FINLAND
2	D.Breskuvienė, G.Dzemyda „Adapt or fall behind: A deep dive into machine learning techniques for detection of the evolving fraud in the financial realm.“ 13th annual Counter Fraud, Cybercrime and Forensic Accounting Conference 2024.06.12/13 , Portsmouth, UK.

Publikacijos (tik su citavimo rodikliu)

	Bibliografinis aprašas	Būklė
1	D. Breskuvienė, G. Dzemyda, “Categorical Feature Encoding Techniques for Improved Classifier Performance when Dealing with Imbalanced Data of Fraudulent Transactions”, INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL , vol. 18, no. 3, Art. no. 3, May 2023, doi: 10.15837/ijccc.2023.3.5433.	Publikuota 2023.05.09
2	D. Breskuvienė, G. Dzemyda. “Enhancing credit card fraud detection: highly imbalanced data case” Journal of Big Data. ISSN 2196-1115. 2024, vol. 11, iss. 1, sp. [182]. doi: 10.1186/s40537-024-01059-5.	Publikuota 2024.12.28

Mokslinių tyrimų ir disertacijos rengimo planas:

	Darbo pavadinimas	Atlikimo terminai	
1.	<u>Mokslinių tyrimų disertacijos tema apžvalga ir analizė (Lietuvoje ir užsienyje):</u> 1.1. Disertacijos tyrimo objekto detalizavimas. 1.2. Atlikti būdų klasifikatorių veikimo optimizavimui analitinę apžvalgą. 1.3. Nustatyti (identifikuoti) mokslines problemas, kylančias uždaviniuose, susijusiuose su klasifikavimo kokybės optimizavimu, o taip pat ir naudojant giliuosius neuroninius tinklus. 1.4. Tyrimo tikslo suformavimas.	2021 m. gruodžio mėn. – 2022 m. vasario mėn. 2021 m. gruodžio mėn. – 2022 m. spalio mėn. 2022 m. kovo mėn. – 2022 m. spalio mėn. 2022 m. kovo mėn. – 2022 m. spalio mėn.	Parengta mokslinės literatūros apžvalga.
2.	<u>Mokslinio tyrimo vykdymas:</u> 2.1. Tyrimo metodikos sudarymas: 2.1.1. Tyrimo metodikos iškeltiems uždaviniams spręsti parinkimas; 2.1.2. Teorinio ir empirinio tyrimų suplanavimas pagal pasirinktą metodiką. 2.2. Teorinis tyrimas: 2.2.1. Klasifikatorių efektyvumo galimybių tyrimas optimizuojant mokymo aibės taškų parinkimą. 2.2.2. Giliųjų neuroninių tinklų panaudojimo galimybių optimaliai mokymo aibei rasti tyrimas. 2.3. Empirinis tyrimas: 2.3.1. Sudarytų metodų pritaikymas praktinių uždavinių sprendimui. 2.3.2. Gautų duomenų analizė, rezultatų apibendrinimas, išvadų parengimas.	2022 m. kovo mėn. – 2022 m. spalio mėn. 2022 m. kovo mėn. – 2022 m. spalio mėn. 2022 m. lapkričio mėn. – 2023 m. spalio mėn. 2022 m. lapkričio mėn. – 2023 m. spalio mėn. 2024 m. kovo mėn. – 2024 m. spalio mėn. 2024 m. spalio mėn. – 2025 m. vasario mėn.	Atlikta Atlikta Parengtas teorinis tyrimas. Parengtas empirinis tyrimas.

Mokslinių tyrimų ir disertacijos rengimo planas:

Darbo pavadinimas		Atlikimo terminai	
3.	<u>Atskirų daktaro disertacijos dalių (tyrimo metodikos, rezultatų, ginamų teiginių, išvadų, ir kt.) parengimas:</u>		
	3.1. Tikslų, uždavinių, tyrimo metodikos, ginamųjų teiginių patikslinimas;	2024 m. spalio mėn. – 2025 m. vasario mėn.	Atlikta
	3.2. Analitinės disertacijos dalies parengimas;	2024 m. kovo mėn. – 2025 m. rugpjūčio mėn.	Atlikta
	3.3. Teorinės disertacijos dalies parengimas;	2024 m. kovo mėn. – 2025 m. rugpjūčio mėn.	Atlikta
	3.4. Eksperimentinės disertacijos dalies parengimas;	2024 m. kovo mėn. – 2025 m. rugpjūčio mėn.	Atlikta
	3.5. Bendrųjų išvadų formulavimas.	2024 m. kovo mėn. – 2025 m. rugpjūčio mėn.	Atlikta
4.	Daktaro disertacijos parengimas ir svarstymas padalinyje	2025 m. rugsėjo mėn.	Daktaro disertacija parengta
5	Daktaro disertacijos gynimas	2025 m. lapkričio mėn.	

Disertacijos tyrimo objektai ir tikslas

Tyrimo objektas:

Šio tyrimo objektas – finansinio sukčiavimo aptikimo procesas, ypatingą dėmesį skiriant požymių atrankos metodams, juos konvertuojant į kitą skaitmeninę erdvę. Tyrime analizuojama, kaip įvairūs požymių atrankos metodai gali padėti efektyviau parinkti informatyvius požymius, siekiant pagerinti nesubalansuotų duomenų klasifikavimo tikslumą taikant mašininio mokymosi algoritmus kreditinių kortelių sukčiavimo aptikimui.

Tikslas:

Šio tyrimo tikslas - sukurti metodą, racionaliai sumažinti esamą požymių rinkinį, kad būtų pagerintas nesubalansuotų duomenų klasifikavimo tikslumas, siekiant geriau aptikti sukčiavimo atvejus.

Tyrimo uždaviniai

- Pasiūlyti nuoseklią mašininio mokymosi strategiją, grindžiamą duomenų klasterizavimu į prasmingas grupes, šių grupių balansavimu bei individualiai pritaikytų klasifikavimo metodų taikymu, siekiant pagerinti sukčiavimo atvejų atpažinimą.
- Įvertinti tikslinių (angl. target-based) ir į tikslinį kintamąjį neatsižvelgiančių (angl. target-agnostic) kodavimo metodikų įtaką nesubalansuotiems duomenims klasifikuoti.
- Įvertinti saviorganizuojančio žemėlapių galimybes ir panaudojimą kaip tarpinę požymių atrinkimo proceso dalį: transakcijoms klasterizuoti ir požymiams konvertuoti į naują skaitinę erdvę siekiant geriausio požymių parinkimo tiriamame duomenų rinkinyje.
- Sukurti į nesubalansuotus duomenis orientuotą metodą esamam požymių rinkiniui pagerinti ir sumažinti, siekiant geresnių klasifikavimo rezultatų.
- Pasiūlyti eksperimentinę aplinką, kuri užtikrina modelių gebėjimą prisitaikyti prie besikeičiančių sukčiavimo elgsenos tendencijų taikant laiku grindžiamus transakcijų duomenų aibės padalijimus (mokoma ankstesniais duomenimis nei testuojama) ir atlikti eksperimentus su viešai prieinamais duomenimis, siekiant įrodyti siūlomo sprendimo efektyvumą.

Ginamieji teiginiai

- Siūloma klasterizavimu grįsta mašininio mokymo strategija, apimanti individualų klasterių balansavimą ir klasifikavimą, leidžia reikšmingai pagerinti sukčiavimo atvejų atpažinimo tikslumą bei sumažinti klaidingai priskirtų sukčiavimo atvejų skaičių teisėtoms operacijoms.
- Tikslinės kodavimo metodikos, lyginant su tradicinėmis, į tikslinį kintamąjį neatsižvelgiančiomis kodavimo schemomis, pasižymi pranašumu klasifikavimo užduotyse, kuriose egzistuoja didelis klasių disbalansas ir aukštas kategorinių požymių kardinalumas.
- Siūlomas FID-SOM (Feature Selection for Imbalanced Data Using SOM) metodas, pagrįstas konkurencinio mokymosi principais, taikant SOM kaip požymių konversijos mechanizmą, kuriame BMU svorio vektorių dispersija naudojama požymių svarbai įvertinti, pagerina požymių atrankos procesą užtikrindamas tikslesnį sukčiavimo aptikimą.
- Siekiant įvertinti modelių gebėjimą prisitaikyti prie besikeičiančių sukčiavimo elgsenos tendencijų, reikia taikyti laiku grindžiamus transakcijų duomenų aibės padalijimus (mokoma ankstesniais duomenimis nei testuojama), kadangi tai leidžia vertinti modelio adaptyvumą prognozuojant ateities įvykius finansinio sukčiavimo aptikime.

Pagrindinės metrikos klasifikavimo gerumui įvertinti

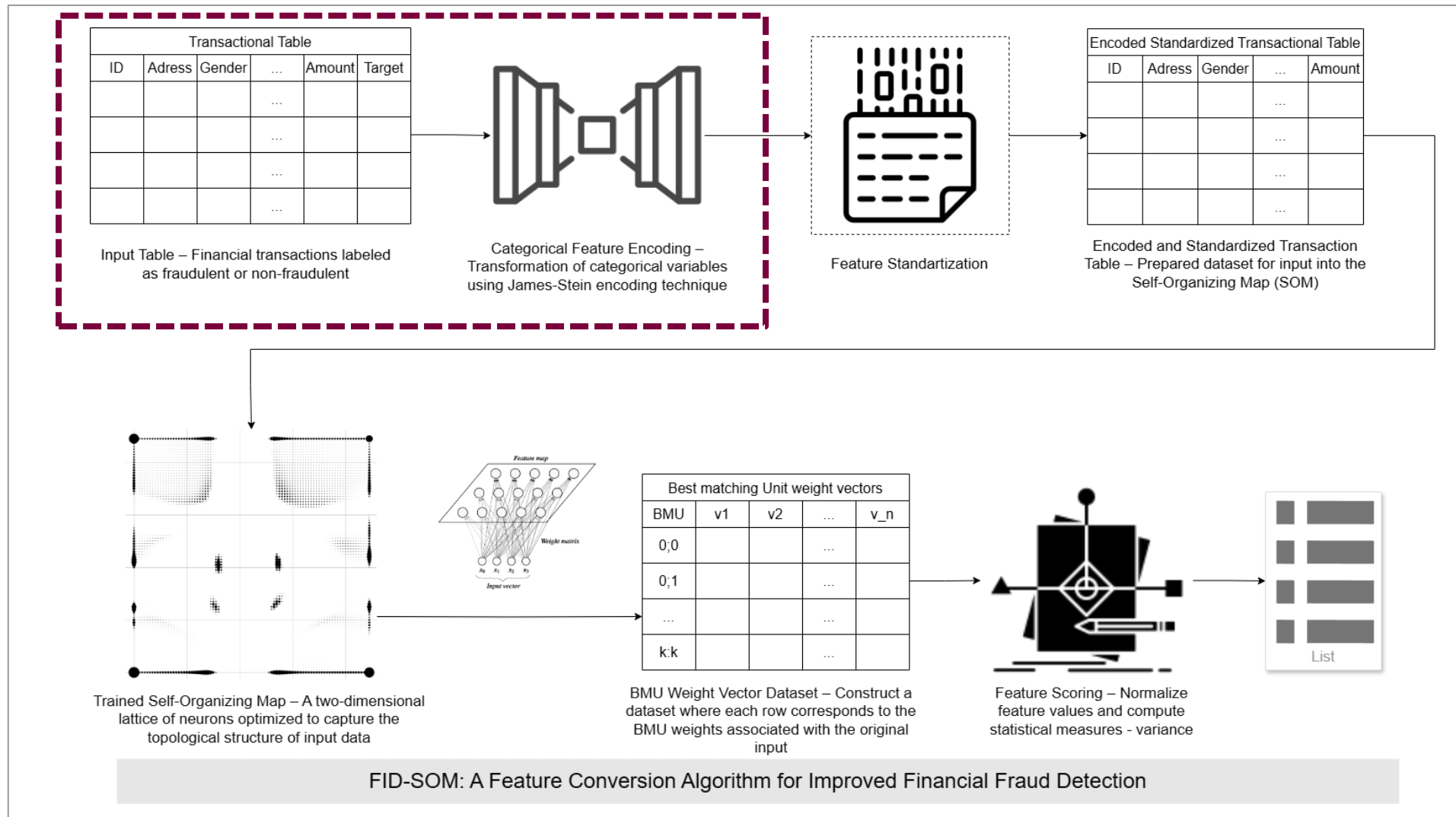
		Predicted	
Actual		<i>TN</i> - regular transactions labeled as regular	<i>FP</i> - regular transactions labeled as fraudulent
		<i>FN</i> - fraudulent transactions labeled as regular	<i>TP</i> - fraudulent transactions labeled as fraudulent

- Recall (atsako rodiklis) yra kritiškas, nes jis parodo, kiek visų tikrųjų sukčiavimo atvejų modelis sugeba aptikti.
- F1 subalansuoja precision ir recall. Vien tik aukštas recall gali reikšti daug klaidingų aliarmų (false positives), kurie apkrauna specialistus ir kelia klientų nepasitenkinimą. F1 įvertina kompromisą – kiek sėkmingai modelis pagauna sukčiavimus neprigeneruodamas per daug klaidingų signalų.

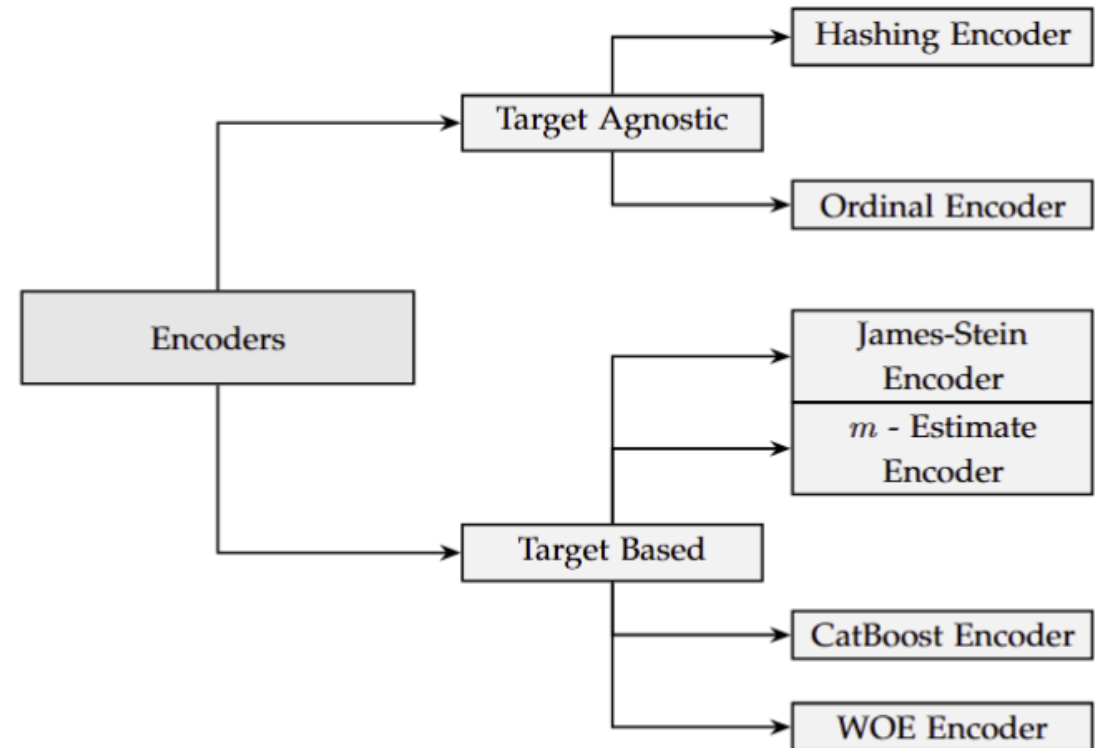
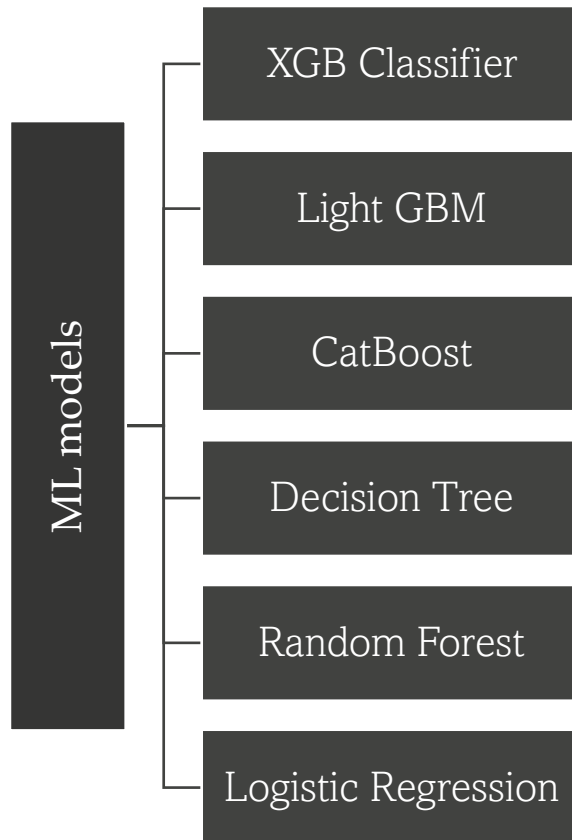
$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

FID-SOM (Feature Selection for Imbalanced Data Using SOM)



Kategorinių požymių kodavimo eksperimentai: nesubalansuotų duomenų atvejis



Kategorinių požymių kodavimo metodų F1 rezultatai

Encoder	Boosting			Ensemble			Non-linear		
	Max	Avg.	Std. Dev.	Max	Avg.	Std. Dev.	Max	Avg.	Std. Dev.
jse	0.9107	<u>0.8049</u>	0.0899	0.8708	0.7595	0.1004	0.8257	<u>0.7604</u>	0.0498
woe	0.9119	0.7861	0.1076	0.8674	<u>0.7651</u>	0.0955	0.8246	0.7529	0.0544
mee	0.9118	0.7875	0.1253	0.8685	0.7584	0.1016	0.8235	0.7607	0.0503
cbe	0.8830	0.6221	0.2145	0.8518	0.6619	0.1577	0.7909	0.5244	0.2377
oe	0.9092	0.7706	0.1487	0.8699	0.7074	0.1528	0.8197	0.7449	0.0578
he	0.7800	0.4786	0.2541	0.7185	0.4356	0.2750	0.6340	0.4136	0.2054

Tyrimas pavadinimu “Categorical Feature Encoding for Improved Classifier Performance when Dealing with Imbalanced Data of Fraudulent Transactions” buvo publikuotas žurnale INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL



INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
Online ISSN 1841-9844, ISSN-L 1841-9836, Volume: 18, Issue: 3, Month: June, Year: 2023
Article Number: 5433, <https://doi.org/10.15837/ijcce.2023.3.5433>



CCC Publications



Categorical Feature Encoding Techniques for Improved Classifier Performance when Dealing with Imbalanced Data of Fraudulent Transactions

D. Breskuvienė, G. Dzemyda

Dalia Breskuvienė*

Data Science and Digital Technologies Institute
Vilnius University, Lithuania
Akademijos g. 4, Vilnius 08412, Lithuania
dalia.breskuviene@mif.vu.lt

*Corresponding author: dalia.breskuviene@mif.vu.lt

Gintautas Dzemyda

Data Science and Digital Technologies Institute
Vilnius University, Lithuania
Akademijos g. 4, Vilnius 08412, Lithuania
gintautas.dzemyda@mif.vu.lt

Abstract

Fraudulent transaction data tend to have several categorical features with high cardinality. It makes data preprocessing complicated if categories in such features do not have an order or meaningful mapping to numerical values. Even though many encoding techniques exist, their impact on highly imbalanced massive data sets is not thoroughly evaluated.

Two transaction datasets with an imbalance lower than 1% of frauds have been used in our study. Six encoding methods were employed, which belong to either target-agnostic or target-based groups. The experimental procedure has involved the use of several machine-learning techniques, such as ensemble learning, along with both linear and non-linear learning approaches.

Our study emphasizes the significance of carefully selecting an appropriate encoding approach for imbalanced datasets and machine learning algorithms. Using target-based encoding techniques can enhance model performance significantly. Among the various encoding methods assessed, the James-Stein and Weight of Evidence (WOE) encoders were the most effective, whereas the CatBoost encoder may not be optimal for imbalanced datasets. Moreover, it is crucial to bear in mind the curse of dimensionality when employing encoding techniques like hashing and One-Hot encoding.

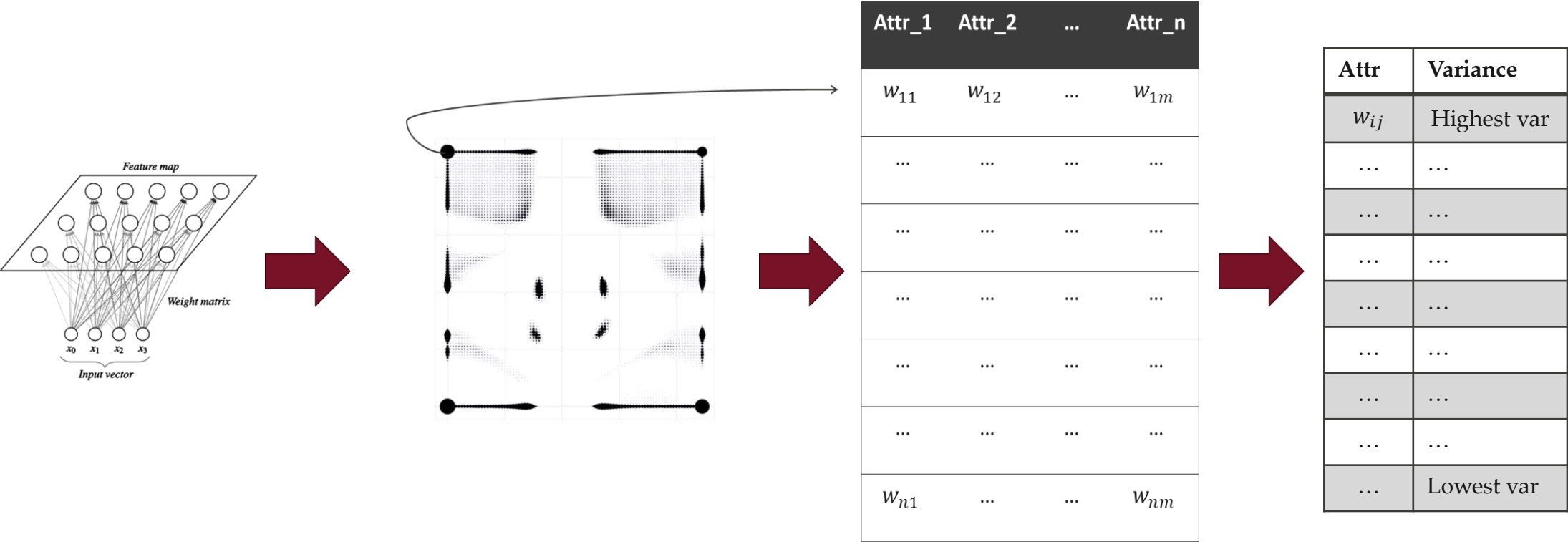
Keywords: imbalanced data, classifier, feature encoding, high-cardinality, fraud detection.

1 Introduction

Financial fraud is a significant issue for businesses and individuals alike, with losses amounting to billions of euros each year. The negative effect of fraud is felt across all sectors of every country and

Neuronų laimėtojų svorio vektorių duomenų rinkinys

Sudaromas duomenų rinkinys, kuriame kiekviena eilutė atitinka neuronų laimėtojų svorio vektorių, susijusį su pradiniais įvesties duomenimis.



Eksperimentinė aplinka

Metrikos

- F1
- MCC
- G-MEAN
- AUC ROC
- AUC PR

Klasifikatoriai

- XGBoost
- Random Forest
- CatBoost

Konkurentai – Požymių atrankos metodai

- Univariate FS method based on χ^2 -test
- Univariate FS method based on F-test
- Univariate FS method based on Mutual Information
- Recursive Feature Elimination
- XGB Importance

Sintetinis rinkinys 1

Ne sukčiavimas (%)	99,86%
Sukčiavimas (%)	0,14%
Transakcijų skaičius	24 386 900
Požymių skaičius	25

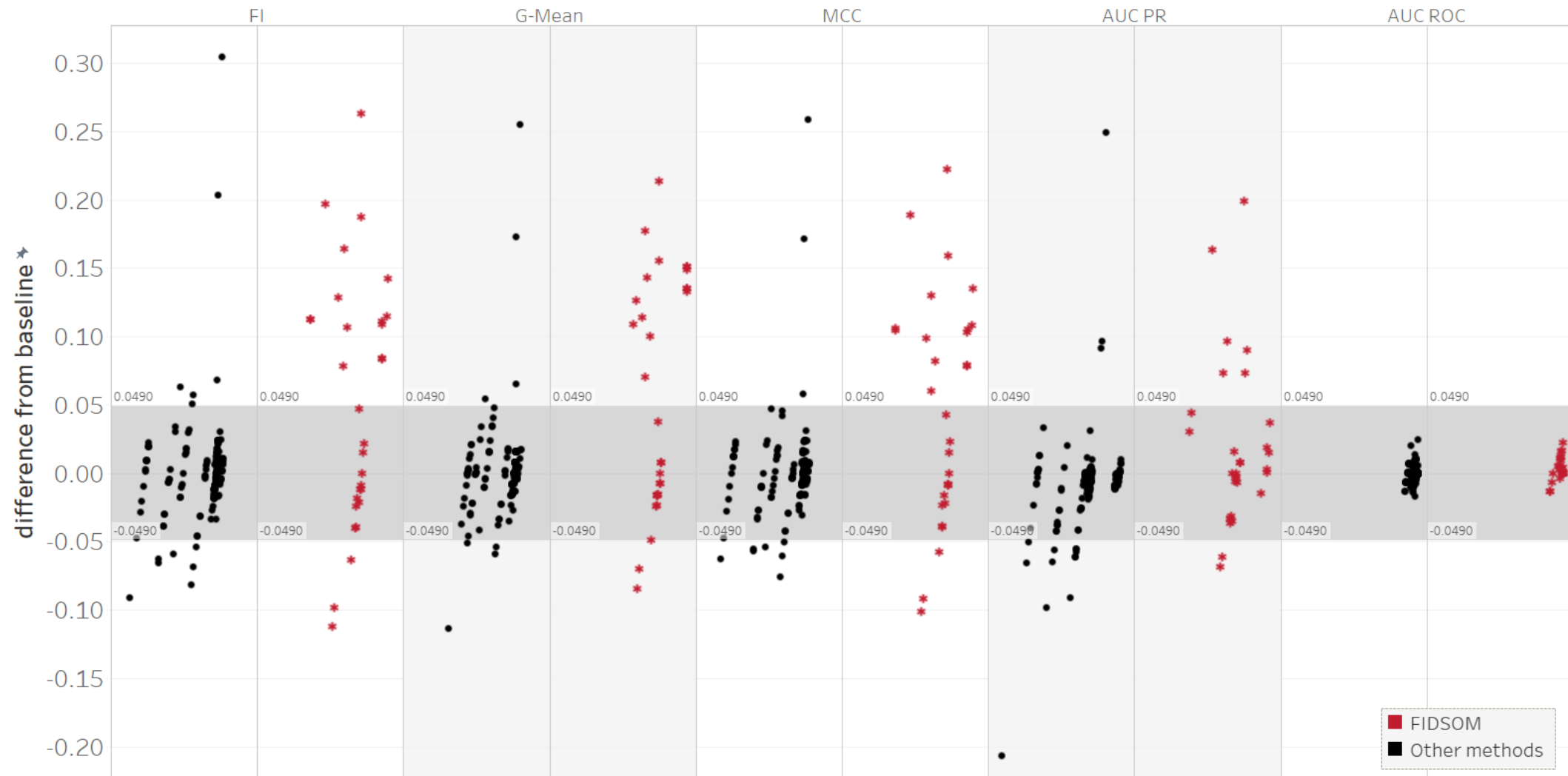
Sintetinis rinkinys 2

Ne sukčiavimas (%)	99,48%
Sukčiavimas (%)	0,52%
Transakcijų skaičius	1 852 394
Požymių skaičius	11

Realus rinkinys

Ne sukčiavimas (%)	99,83%
Sukčiavimas (%)	0,172%
Transakcijų skaičius	284 807
Požymių skaičius	30

Rezultatai



FID-SOM
palyginimas
su kitais
požymių
atrankos
metodais

Method	# of winnings	Total	Percentage
Baseline	2	45	4.44%
FID-SOM	32	45	71.11%
Univariate FS method based on 2-test	8	45	17.78%
Univariate FS method based on F-test	6	45	13.33%
Univariate FS method based on Mutual Information	8	45	17.78%
Recursive Feature Elimination	5	45	11.11%
XGB Importance	6	45	13.33%

Duomenų rinkiniai: 3

Metrikos: 5

Mašininio mokymosi algoritmai: 3

- Siūloma klasterizavimu grįsta mokymosi strategija, kurioje kiekvienas klasteris subalansuojamas ir klasifikuojamas atskirai, leidžia reikšmingai pagerinti sukčiavimo atvejų klasifikavimo atsako rodiklį (angl. Recall) – nuo 0.845 iki 0.867. Eksperimentų rezultatai rodo, kad integravus klasterizaciją kaip išankstinį apdorojimo žingsnį, modelio gebėjimas teisingai identifikuoti apgaulingas transakcijas pagerėjo, o klaidingai legaliomis priskirtų sukčiavimo atvejų sumažėjo 13,9 %.

- Eksperimentiniai tyrimo rezultatai atskleidė, kad kategorinių požymių kodavimo strategijos, įtraukiančios tikslinio kintamojo informaciją – konkrečiai James-Stein ir Weight of Evidence (WOE) metodai – leidžia užtikrinti ženkliai aukštesnę modelių diskriminacinę gebą dirbant su labai nesubalansuotais ir aukštą kardinalumą turinčiais duomenimis. Taikant James-Stein kodavimą, vidutiniai F1 rodikliai siekė 0,8049 su stiprinimo tipo klasifikatoriais, 0,7595 su ansambliniais modeliais ir 0,7604 su nelinijiniais; WOE rezultatai atitinkamai buvo 0,7861, 0,7651 ir 0,7529. Tuo metu target-agnostic metodai, tokie kaip Label Encoding ar Hashing, reikšmingai atsiliko – kai kuriais atvejais F1 reikšmė nesiekė nė 0,5. *Pažymėtina, kad visų modelių hiperparametrai liko numatytieji – eksperimentų tikslas buvo ne optimizuoti galutinį rezultatą, bet objektyviai įvertinti skirtingų kodavimo strategijų poveikį.*
- Hashing kodavimo metodas, kaip ir dažnai naudojamas One-Hot kodavimas (nors šis metodas eksperimente nebuvo įtrauktas), prisideda prie „daugiamatiškumo prakeiksmo“ (angl. curse of dimensionality), ypač kai kategorinių požymių kardinalumas yra labai aukštas – pvz., „Merchant City“ turi 11 391 unikalias reikšmes. Tai ne tik blogina modelio klasifikuoti, bet ir didina skaičiavimų aptikimo sąnaudas bei riziką persimokyti.

Disertacijos išvados

- Eksperimentiniais tyrimais buvo parodyta, kad pasiūlytas naujas požymių atrankos metodas FID-SOM, pasitelkiantis saviorganizuojančius žemėlapius ir remiantis neuronų laimėtojų (BMU) svorių vektorių komponentų dispersija informatyviausių požymių identifikavimui, pasirodė esąs geriausias net 71,11 % visų testuotų konfigūracijų, t. y. daugumoje požymių rinkinių ir klasifikatorių kombinacijų jis pasiekė aukščiausią klasifikavimo rezultatą. Metodas pranoko tradicinius atrankos sprendimus, tokius kaip F testas, X^2 testas, tarpusavio informacija (angl. mutual information), rekursyvinė požymių eliminacija (angl. Recursive Feature Elimination) bei XGBoost svarbos analizė (angl. XGB Importance).
- Eksperimentiniai rezultatai parodė, kad taikant laiku pagrįstą transakcijų duomenų padalijimą, testuojant FID-SOM pasiekė F1 rodiklį 0,85, atsako rodiklį 0,76. Tuo tarpu naudojant atsitiktinį duomenų dalijimą, tie patys modeliai grąžino gerokai aukštesnius rodiklius (F1 rodiklis 0,88, atsako rodiklis 0,82), tačiau toks rezultatų padidėjimas laikytinas dirbtiniu ir siejamas su duomenų nutekėjimu tarp mokymo ir testavimo rinkinių. Šis skirtumas parodė, kad atsitiktinis skirstymas gali sudaryti klaidingą įspūdį apie modelio efektyvumą, o laiku grįstas padalijimas suteikia patikimesnę pagrindą vertinti modelių gebėjimą būti atspariam besikeičiančioms sukčiavimo tendencijoms.

Planai iki doktorantūros studijų pabaigos (2025-11-30)

- Redakatoriaus pastabų taisymas
- Disertacijos gynimas



**Vilniaus
universitetas**

Ačiū už dėmesį!

Klausimai?