



**Vilnius
universitetas**

Duomenų Mokslo ir Skaitmeninių Technologijų Institutas

Informatikos inžinerijos krypties doktorantų konferencija
Veiklos ataskaita už 2024 m. spalio 3d. – 2025 m. kovo 28d.

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	

2024/2025 – I pusmetis

Publikacijos 2024/2025 I pusmetis			
Planas	Ivykdyta	Būklė	Publikacijos tipas
Journal of Big Data	Breskuvienė, Dalia; Dzemyda, Gintautas. 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	CA WoS Impact Factor 12,4

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. Esant poreikiui ji toliau yra pildoma.
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 Parengtas teorinis tyrimas. Esant poreikiui toliau pildomas Parengtas empirinis tyrimas. Esant poreikiui toliau pildomas

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. Esant poreikiui tikslinama.
	3.2. Analitinės disertacijos dalies parengimas;	2024 m. kovo mėn. – 2025 m. rugpjūčio mėn.	Rengiama
	3.3. Teorinės disertacijos dalies parengimas;	2024 m. kovo mėn. – 2025 m. rugpjūčio mėn.	Rengiama
	3.4. Eksperimentinės disertacijos dalies parengimas;	2024 m. kovo mėn. – 2025 m. rugpjūčio mėn.	Rengiama
	3.5. Bendrųjų išvadų formulavimas.	2024 m. kovo mėn. – 2025 m. rugpjūčio mėn.	Rengiama
4.	Daktaro disertacijos parengimas ir svarstymas padalinyje	2025 m. rugsėjo mėn.	
5	Daktaro disertacijos gynimas	2025 m. lapkričio mėn.	

Disertacijos tema, tyrimo objektai ir tikslas

- Disertacijos tema:

Požymių konversija siekiant gerinti finansinio sukčiavimo aptikimą

EN: Feature conversion for better financial fraud detection

Tyrimo objektai:

Finansinio sukčiavimo aptikimas specializuojantis į kredito kortelių sukčiavimo atvejus.

EN: The object of this research is financial fraud detection and its special case - credit card fraud detection.

Tikslas:

Sukurti metodą, leidžiantį patobulinti ir sumažinti esamą požymių rinkinį, kad būtų pagerintas nesubalansuotų duomenų klasifikavimo tikslumas, siekiant užtikrinti geresnį sukčiavimo aptikimą.

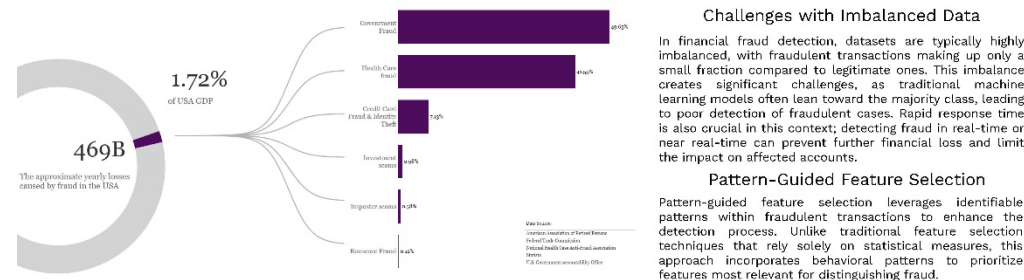
EN: Develop a method to enhance and reduce an existing feature set to improve the classification performance on imbalanced data for better fraud detection.

Tyrimo uždaviniai

- Nustatyti atitinkamus mokslinius iššūkius, kylančius sprendžiant su finansinio sukčiavimo nustatymu susijusias užduotis.
- Nustatyti tinkamus metodus nesubalansuotiems mokymo duomenų rinkiniams optimizuoti.
- Įvertinti saviorganizuojančio žemėlapių galimybes ir panaudojimą transakcijoms agreguoti ir toliau kurti naują požymių sistemą.
- Sukurti į nesubalansuotų duomenų rinkinį orientuotą metodą esamam požymių rinkiniui pagerinti ir sumažinti, siekiant geresnių klasifikavimo rezultatų.
- Sukurti eksperimentinę aplinką, pagrįstą etaloniniais realaus pasaulio duomenimis, siekiant įrodyti siūlomo sprendimo efektyvumą.
- Apibendrinti rezultatus ir suformuluoti išvadas.

Atlikti darbai

- Pakoreguotas straipsnis pagal recenzentų pastabas. Straipsnis publikuotas 2024.12.28
- Atliktas pristatymas DAMSS konferencijoje “Highly imbalanced data case: Pattern -Guided Feature Selection to Detect Financial Fraud Challenges with Imbalanced Data”
- Disertacijos teksto rašymas



Proposed Method

Algorithm FID-SOM (Feature selection for imbalanced Data Using SOM)

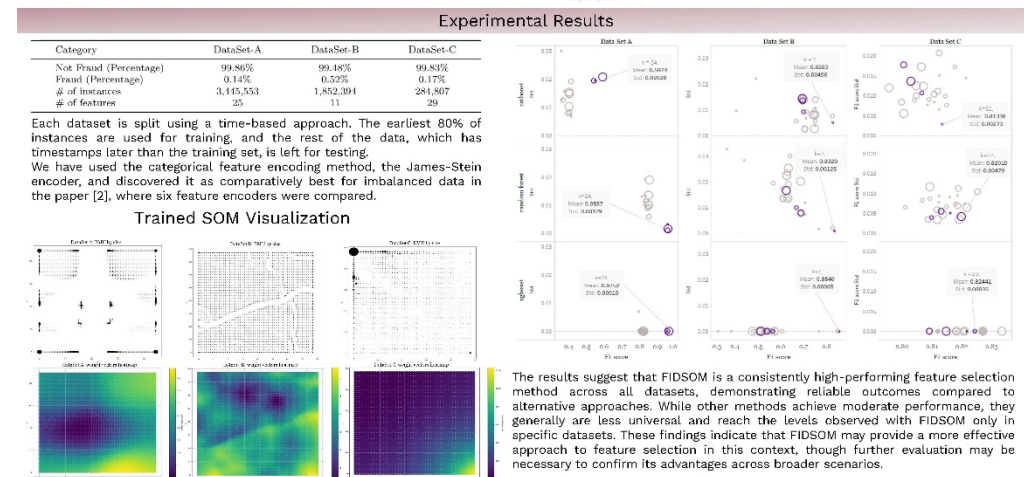
Require X: Dataset
Require params: SOM parameters
Require d: Desired number of features.
Ensure: Features subset ensuring high classifier performance

Procedure SELECTFEATURES:

1. train SOM using parameters *params* with dataset *X*
2. form a new dataset W_{SOM} containing n_{SOM} weight vectors of *m* attributes corresponding to *m* features of dataset *X*
3. normalize W_{SOM} dataset attributes to a scale of [0,1]
4. calculate the variance of each attribute
5. sort attributes based on variance in descending order
6. select *d* attributes from the top of the list
7. select features for dataset *X* corresponding to the kept attributes

We propose employing Self-Organizing Maps (SOM), which is a special type of artificial neural network, to select valuable features. The SOM algorithm is applied to the dataset to cluster similar transactions together and identify patterns of fraudulent behavior.

This method dynamically adapts to the data's inherent characteristics, ensuring an automatic and data-driven feature selection process. It significantly enhances the method's suitability for diverse scientific applications, where datasets often vary in dimensionality and complexity. FID-SOM was compared with univariate feature selection methods utilizing the F-test, χ^2 -test and mutual information, the Recursive Feature Elimination method, and the XGB Importance method. The goodness of the feature selection methods was evaluated using F1 score, MCC, G-Mean, AUC-PR, and AUC-ROC metrics when performing XGBoost, CatBoost, and Random algorithms on three datasets.

A diagram of a Self-Organizing Map (SOM) network. It shows an 'Input Space' at the bottom with several input nodes. These nodes are connected to a grid of 'Neurons' in the middle. Each connection is labeled with a 'Weights' value. The output of the neurons is shown as a 'Latent Space' at the top.

Publikuotų straipsnių statistika

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

Enhancing credit card fraud detection: highly imbalanced data case

CA WoS

Article ⓘ

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

5 Citations

35 References

Breskuvienė, D and Dzemyda, G

May 2023 | INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL 18 (3)

Article ⓘ

Enhancing credit card fraud detection: highly imbalanced data case

Breskuvienė, D and Dzemyda, G

Dec 28 2024 | JOURNAL OF BIG DATA 11 (1)

50 References

Google scholar

PAVADINIMAS ⓘ	CITUOTA	METAI
Categorical feature encoding techniques for improved classifier performance when dealing with imbalanced data of fraudulent transactions	15	2023
D Breskuvienė, G Dzemyda		
International Journal of Computers Communications & Control 18 (3)		

Enhancing credit card fraud detection: highly imbalanced data case

D Breskuvienė, G Dzemyda

Journal of Big Data 11 (1), 182

2024

Research Gate

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

Research Interest Score 16.5

Citations 11

Recommendations 1

Reads ⓘ 1,427

May 2023 · International Journal of Computers, Communications & Control 18(3)

DOI: 10.15837/ijccc.2023.3.5433

License · CC BY-NC 4.0

[Learn about stats on ResearchGate](#)

Enhancing credit card fraud detection: highly imbalanced data case

Research Interest Score 4.6

Citations 1

Recommendations 7

Reads ⓘ 56

December 2024 · Journal of Big Data 11(1)

DOI: 10.1186/s40537-024-01059-5

License · CC BY-NC-ND 4.0

[Learn about stats on ResearchGate](#)

Dalia Breskuvienė · Gintautas Dzemyda

FID-SOM algorithm

Algorithm 1 FID-SOM (Feature selection for Imbalanced Data Using SOM)

Require: X : Dataset

Require: $params$: SOM parameters

Require: d : Desired number of features.

Ensure: features subset ensuring high classifier performance

- 1: **procedure** SELECTFEATURES
 - 2: train SOM using parameters $params$ with dataset X
 - 3: form a new dataset W_{BMU} containing n_{BMU} weight vectors of m attributes corresponding to m features of dataset X
 - 4: normalize W_{BMU} dataset attributes to a scale of $[0,1]$
 - 5: calculate the variance of each attribute
 - 6: sort attributes based on variance in descending order
 - 7: select d attributes from the top of the list
 - 8: select features for dataset X corresponding to the kept attributes
 - 9: **end procedure**
-

Požymių erdvės transformacija

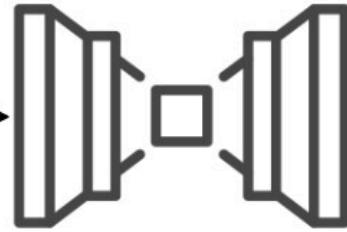
Vilniaus
universitetas

As an example:
3 000 000 transactions
with 30 features

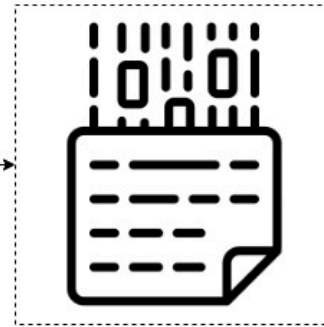
Categorical Feature Encoding Techniques for
Improved Classifier Performance when
Dealing with Imbalanced Data of Fraudulent
Transactions

Transactional Table				
ID	Adress	Gender	...	Amount
			...	
			...	
			...	
			...	

Input Table



Feature Encoding

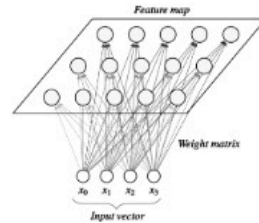
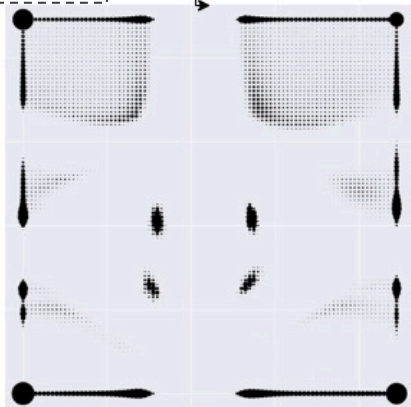


Feature Standardization

Encoded Standardized Transactional Table				
ID	Adress	Gender	...	Amount
			...	
			...	
			...	
			...	

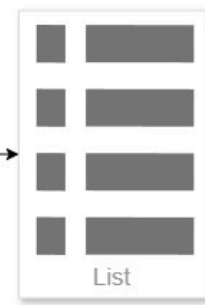
Input to Feature
Transformation Step

Map 42x42
With 30 length
weight vectors



≤ 1764 rows and 30 columns

Best matching Unit weight vectors				
BMU	v1	v2	...	v_n
0;0			...	
0;1			...	
...			...	
k:k			...	



List

Enhancing credit card fraud detection: highly imbalanced data case

Pasiūlyto metodo palyginimas su kitais klasikiniais metodais ir publikacijomis

Table 9: Comparison with other papers splitting data in a time based manner

Paper	Year	F1-Score	Recall	Precision
[50]	2019	0.82	0.73	0.93
[51]	2023	0.84	0.74	0.97
FIDSOM*	2024	0.85	0.76	0.97

*FIDSOM with XGB classifier selecting 23 features. Data split is done by selecting 70% of the first data points for training and 30% remaining data points for testing. For SOM training, 500 iterations were used with a sigma of 15 and a learning rate of 1.8.

[50] Fiore, U., De Santis, A., Perla, F., Zanetti, P., Palmieri, F.: Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. *Information Sciences* **479**, 448–455 (2019)

[51] Fanai, H., Abbasimehr, H.: A novel combined approach based on deep autoencoder and deep classifiers for credit card fraud detection. *Expert Systems with Applications* **217**, 119562 (2023)

KITO PUSMEČIO DARBO PLANAS

- Daktaro disertacijos parengimas ir svarstymas padalinyje
- Daktaro disertacijos gynimas



**Vilniaus
universitetas**

Ačiū už dėmesį!



Klausimai?