Vilniaus universitetas

Duomenų Mokslo ir Skaitmeninių Technologijų Institutas

Informatikos inžinerijos krypties doktorantų konferencija

Veiklos ataskaita už 2021 m. gruodžio 1d. – 2022 m. spalio 1d.

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

Disertacijos tema, tyrimo objektai ir tikslas

- **Preliminari disertacijos tema:**
- Klasifikatoriaus (nesubalansuotos) mokymo aibės optimizavimas, siekiant geresnės klasifikavimo kokybės.

- **Tyrimo objektai:**
- Finansinio sukčiavimo duomenų optimizavimas, siekiant tikslesnių mašininio mokymosi metodų rezultatų.

- **Tikslas:**
- Sukurti arba patobulinti nesubalansuotų duomenų optimizavimo metodą, skirtą anomalinių įvykių identifikavimui ir jų prevencijai.

Tyrimo uždaviniai

- Identifikuoti tinkamus metodus nesubalansuotos mokymo aibės optimizavimui;
- Identifikuoti aktualias mokslines problemas, kylančias uždaviniuose, susijusiuose su finansinio sukčiavimo aptikimu;
- Sukurti arba patobulinti algoritmą nesubalansuotos mokymo aibės optimizavimui atsižvelgiant į naujo taško klasifikavimą;
- Pritaikyti sukurtą arba patobulintą metodą nesubalansuotiems duomenims ir atlikti gautų duomenų analizę, rezultatų apibendrinimą, išvadų parengimą.

2021–2022m.

Vilniaus universitetas

STUDIJŲ PLANAS IR JO VYKDYMO SUVESTINĖ

Studijų metai	Egzaminai [1]		Dalyvavimas konferencijose [2]		Publikacijos [3]		
	Planas	Įvykdyta	Planas	Įvykdyta	Planas	Įvykdyta	Būklė [4]
I (2021/2022)	3	3	1	2			Priimta publikavimui
II (2022/2023)	1		1		1		
III (2023/2024)			1		1		
IV (2024/2025)					1		
Iš viso:	**4**	**3**	**3**	**2**	**3**		

2021–2022m.

Vilniaus universitetas

ATASKAITINIŲ METŲ DARBO PLANAS IR JO SUVESTINĖ

Egzaminai		
Mašininis mokymasis	Išlaikyta – pirmas pusmetis	Išlaikyta. Įvertinimas 10
Netiesiniai statistikos modeliai masinių duomenų analizėje	Išlaikyta – antras pusmetis	Išlaikyta. Įvertinimas 10
Informatikos ir informatikos inžinerijos tyrimo metodai ir metodika	Išlaikyta – antras pusmetis	Išlaikyta. Įvertinimas 9

2021–2022m.

Vilniaus universitetas

ATASKAITINIŲ METŲ DARBO PLANAS IR JO SUVESTINĖ

Konferencijos pavadinimas	Pranešimo pavadinimas	Konferencijos tipas
12th Conference: Data Analysis Methods for Software Systems, 2021.12.02/04 , Druskininkai	„Forbearance prediction using XGBoost and LightGBM Models“	Nacionalinė konferencija
EURO 2022 2022.07.03 /06 ESPOO, FINLAND	„Clustering-based optimization in fraud detection classifier training“	Tarptautinė konferencija

Publikacijos tipas	Publikacijos pavadinimas	Literatūros šaltinis
PROCEEDINGS	„Forbearance prediction using XGBoost and LightGBM Models“	Bernatavičienė J. (2021) “Data Analysis Methods for Software Systems”, Vilnius University Proceedings, 17, pp. 1-82. doi: 10.15388/DAMSS.12.2021
Chapter in the book “Data Science in Application	„Imbalanced Data Classification Approach Based on Clustered Training Set“	Priimta publikavimui

Mokslinių tyrimų ir disertacijos rengimo planas:

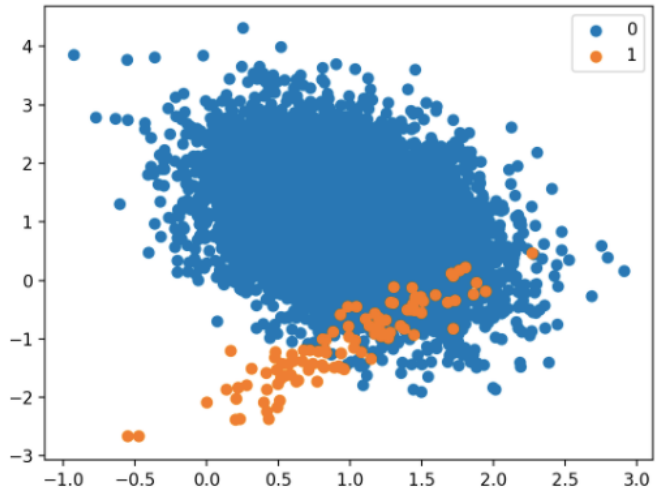
	Darbo pavadinimas	Atlikimo terminai
1.	**Mokslinių tyrimų disertacijos tema apžvalga ir analizė (Lietuvoje ir užsienyje):**	
	1.1. Disertacijos tyrimo objekto detalizavimas.	2021 m. gruodžio mėn. – 2022 m. vasario mėn.
	1.2. Atlikti būdų klasifikatorių veikimo optimizavimui analitinę apžvalgą.	**2021 m. gruodžio mėn. – 2022 m. spalio mėn.**
	1.3. Nustatyti (identifikuoti) mokslines problemas, kylančias uždaviniuose, susijusiuose su klasifikavimo kokybės optimizavimu, o taip pat ir naudojant giliuosius neuroninius tinklus.	**2022 m. kovo mėn. – 2022 m. spalio mėn.**
	1.4. Tyrimo tikslo suformavimas.	**2022 m. kovo mėn. – 2022 m. spalio mėn.**
2.	**Mokslinio tyrimo vykdymas:**	
	2.1. Tyrimo metodikos sudarymas:	
	2.1.1. Tyrimo metodikos iškeltiems uždaviniams spręsti parinkimas;	**2022 m. kovo mėn. – 2022 m. spalio mėn.**
	2.1.2. Teorinio ir empirinio tyrimų suplanavimas pagal pasirinktą metodiką.	**2022 m. kovo mėn. – 2022 m. spalio mėn.**
	2.2. Teorinis tyrimas:	
	2.2.1. Klasifikatorių efektyvumo galimybių tyrimas optimizuojant mokymo aibės taškų parinkimą.	2022 m. lapkričio mėn. – 2023 m. spalio mėn.
	2.2.2. Giliųjų neuroninių tinklų panaudojimo galimybių optimaliai mokymo aibei rasti tyrimas.	2022 m. lapkričio mėn. – 2023 m. spalio mėn.
	2.3. Empirinis tyrimas:	
	2.3.1. Sudarytų metodų pritaikymas praktinių uždavinių sprendimui.	2024 m. kovo mėn. – 2024 m. spalio mėn.
	2.3.2. Gautų duomenų analizė, rezultatų apibendrinimas, išvadų parengimas.	2024 m. spalio mėn. –2025 m. vasario mėn.
3.	**Atskirų daktaro disertacijos dalių (tyrimo metodikos, rezultatų, ginamų teiginių, išvadų, ir kt.) parengimas:**	
	3.1. Tikslų, uždavinių, tyrimo metodikos, ginamųjų teiginių patikslinimas;	2024 m. spalio mėn. – 2025 m. vasario mėn.
	3.2. Analitinės disertacijos dalies parengimas;	2024 m. kovo mėn. – 2025 m. rugpjūčio mėn.
	3.3. Teorinės disertacijos dalies parengimas;	2024 m. kovo mėn. – 2025 m. rugpjūčio mėn.
	3.4. Eksperimentinės disertacijos dalies parengimas;	2024 m. kovo mėn. – 2025 m. rugpjūčio mėn.
	3.5. Bendrųjų išvadų formulavimas.	2024 m. kovo mėn. – 2025 m. rugpjūčio mėn.
4.	Daktaro disertacijos parengimas ir svarstymas padalinyje	2025 m. rugsėjo mėn.
5	Daktaro disertacijos gynimas	2025 m. lapkričio mėn.

Literatūros apžvalga ir naudojami įrankiai:

- Web of Science
- Mendeley
- Straipsnių dokomentacija excel formate:

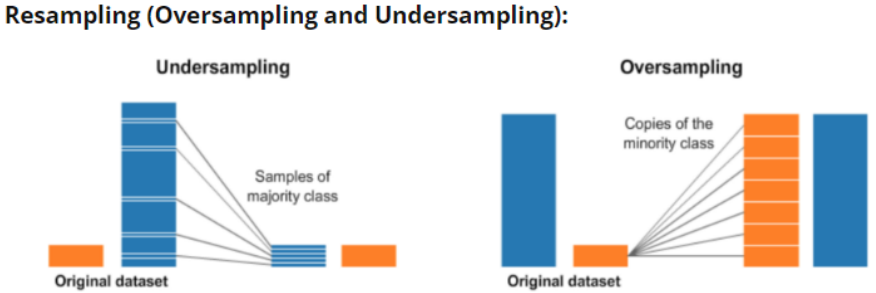
	A	B	C	D	E	F	G	H	I	J	K	L	M	
1	**Topic**	**Nr.**	**Pavadinimas**	**Metai**	**Santrauka**	**Duomenų rinkiniai**	**Metrikc**	**Rezulta**	**Algoritmai**	**Pastabos**	**Citavim**	**WoS SCIE ša**	**Leidinys**	**Kategorija (1 kategorija –publikacija vi bus panaudoti tyrimuose. 2 kategorija – publikacija a**
2	ML algorythms	1	**Performance analysis of boosting classifiers in recognizing activities of daily living**	2020	Physical activity is essential for physical and mental health, and its absence is highly associated with severe health conditions and disorders. Therefore, tracking activities of daily living can help promote quality of life. Wearable sensors in this regard can provide a reliable and economical means of tracking such activities, and such sensors are readily available in smartphones and watches. This study is the first of its kind to develop a wearable sensor-based physical activity classification system using a special class of supervised machine learning approaches called boosting algorithms. The study presents the performance analysis of several boosting algorithms (extreme gradient boosting—XGB, light gradient boosting machine—LGBM, gradient boosting— GB, cat boosting—CB and AdaBoost) in a fair and unbiased performance way using uniform dataset, feature set, feature selection method, performance metric and cross-validation techniques. The study utilizes the Smartphone-based dataset of thirty individuals. The results showed that the proposed method could accurately classify the activities of daily living with very high performance (above 90%). These findings suggest the strength of the proposed system in classifying activity of daily living using only the smartphone sensor's data and can assist in	The dataset used in this study is a publically available dataset developed using thewaist-mounted smartphone: Anguita, D.; Ghio, A.; Oneto, L.; Parra, X.; Reyes-Ortiz, J.L. A Public Domain Dataset for Human ActivityRecognition Using Smartphones. In Proceedings of the European Symposium on Artificial Neural Networks,Bruges, Belgium, 24-26 April 2013; pp. 24-26.	F1		extreme gradient boosting (XGB), light gradient boosting machine (LGBM), gradient boosting (GB), cat boosting (CB), AdaBoost.	Short description of used machine learning algorithms. Feature selection based on correlation	27	Q2	International Journal of Environmental Research and Public Health	3 kategorija
3	Imbalanced data training and optimisation	2	**Learning from imbalanced data**	2009	With the continuous expansion of data availability in many large-scale, complex, and networked systems, such as surveillance, security, Internet, and finance, it becomes critical to advance the fundamental understanding of knowledge discovery and analysis from raw data to support decision-making processes. Although existing knowledge discovery and data engineering techniques have shown great success in many real-world applications, the problem of learning from imbalanced data (the imbalanced learning problem) is a relatively new challenge that has attracted growing attention from both academia and industry. The imbalanced learning problem is concerned with the performance of learning algorithms in the presence of underrepresented data and severe class distribution skews. Due to the inherent complex characteristics of imbalanced data sets, learning from such data requires new understandings, principles, algorithms, and tools to transform vast amounts of raw data efficiently into information and knowledge representation. In this paper, we provide a comprehensive review of the development of research in learning from imbalanced data. Our focus is to provide a critical review of the nature of the problem, the state-of-the-art technologies, and the current assessment metrics used to evaluate learning performance under the imbalanced learning scenario. Furthermore, in order to					Issues when learning from imbalanced data	3,523	Q1	IEEE Transactions on Knowledge and Data Engineering	1kategorija

1.2. Atlikti būdų klasifikatorių veikimo optimizavimui analitinę apžvalgą



- Nesubalansuotos mokymų aibės problemos mašininio mokymosi kontekste:
  - Standartiniai mašininio mokymosi algoritmai yra dažniausiai nepritaikyti nesubalansuotoms duomenų aibėms
  - Nesubalansuotos duomenų aibės dydis dažniausiai kelia papildomų problemų
  - Triukšmas duomenyse sukvalifikuojamas kaip viena iš klasių

- Dažniausiai siūlomi sprendimo būdai:
  - Didesnės klasės narių išretinimas mokymų aibėje (undersampling)
  - Mažesnės klasės narių generavimas/kopijavimas (oversampling)



Mokslines problemas, kylančios uždaviniuose, susijusiuose su klasifikavimo kokybės optimizavimu

Vilniaus universitetas

- Oversampling:
  - Pridėjus papildomų tos pačios eilutės kopijų modelis gali persimokyti (overfitting) [Random oversampling]
  - Gali atsirasti klasių persidengimas [SMOTE]
  - Reikalauja pakankamai daug atminties laikymui ir skaičiavimams
- Undersampling:
  - Išmetus įrašus iš duomenų aibės gali būti prarasta svarbi informacija apie daugumos klasę. [Random undersampling]
- Unsupervised learning:
  - Klasterizavimas: tinkamas klasterių kiekio parinkimas ir kintamieji optimaliai skirstantys duomenis į klasterius

Tyrimo metodikos iškeltiems uždaviniams spręsti parinkimas

01

Kuriamas metodas

02

Parenkami duomenų rinkiniai kuriamam metodui įvertinti

03

Atliekami eksperimentai

04

Grįžtamasis ryšys: metodo idėjos vystymas atsižvelgiant į rezultatus

Teorinio ir empirinio tyrimų suplanavimas pagal pasirinktą metodiką.

Publikacija knygoje „Data Science in Application“
Leidykla „Springer“

Imbalanced Data Classification Approach Based on Clustered Training Set

Dalia Breskuvienė, Gintautas Dzemyda

Institute of Data Science and Digital Technologies, Vilnius University
Akademijos str. 4, Vilnius LT-08412, Lithuania
dalia.breskuviene@gmail.com,gintautas.dzemyda@mif.vu.lt
http://www.mii.lt

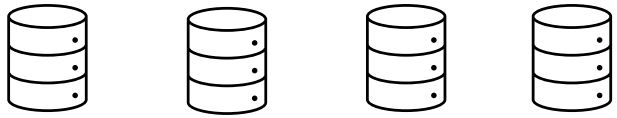
Abstract. Fraud detection is a system that prevents criminals from obtaining financial assets. The research aims to increase machine learning prediction quality on fraudulent cases as well as decrease false positive and false negative cases in prediction. Fraudulent data like credit card transactions are usually imbalanced data, and standard machine learning techniques cannot achieve the desired quality levels in this scenario. This paper proposes a clustering-based classification method to improve the *recall*. For the experimental evaluation, we use a credit card transaction database. Firstly we suggest finding the optimal features and number of clusters to create smaller, more homogeneous training sets, which we train on separate machine learning models. The second step is to find relevant percentages to undersample each cluster to compensate for sharply imbalanced data. Our baseline *recall* is 0.845. By applying the proposed method, we improved the *recall* to 0.867. Moreover, classification of fraudulent cases that were labeled as regular decreased from 323 to 278, i.e. by 13.9%. The statistical test has shown that decrease is significant.

Keywords: Imbalanced data, *k*-means, Undersampling, *recall*, Classification, Fraud detection

Straipsnio grafinė santrauka

Vilniaus universitetas

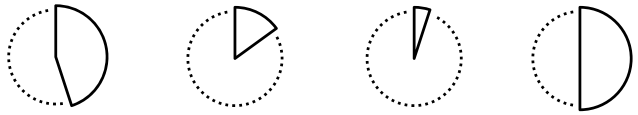
Training set Clustering



Experimental search for the best feature subset and number of clusters

Measure – Silhouette score

Undersampling each cluster



$$F1 = 2 * \frac{precision * recall}{precision + recall}$$

$$precision = \frac{TP}{TP + FP} \qquad recall = \frac{TP}{TP + FN}$$

Train the models

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

Unseen data classification

Shortest Euclidian distance to a cluster center

KITO PUSMEČIO DARBO PLANAS

Vilniaus universitetas

Egzaminai:

- Fundamentalieji informatikos ir informatikos inžinerijos metodai

Publikacijos:

- Publikacija Springerio leidinyje

Konferencijos:

- 13th Conference „Data Analysis Methods for Software Systems“,

Pranešimai:

- Lietuvos Aktuarų Asociacijoje (2022.10.20)

Teorinis tyrimas:

- Klasifikatorių efektyvumo galimybių tyrimas optimizuojant mokymo aibės taškų parinkimą.
- Giliųjų neuroninių tinklų panaudojimo galimybių optimaliai mokymo aibei rasti tyrimas.

Vilniaus universitetas

Ačiū už dėmesį!

Klausimai?