Vilniaus universitetas

Duomenų Mokslo ir Skaitmeninių Technologijų Institutas

Informatikos inžinerijos krypties doktorantų konferencija

Veiklos ataskaita už 2024 m. kovo 29d. – 2024 m. spalio 3d.

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]		Dalyvavimas konferencijose [2]				Publikacijos [3]					
			Tarptautinėse		Nacionalinėse		Su citav. rodikliu			Be citav. rodiklio		
	Planas	Įvykdyta	Planas	Įvykdyta	Planas	Įvykdyta	Planas	Įvykdyta	Būklė [4]	Planas	Įvykdyta	Būklė [4]
I (2021/2022)	3	3			1	1						
II (2022/2023)	1	1	1	1		1		1	Publikuota	1	1	Publikuota
III (2023/2024)			1	1		1	1					
IV (2024/2025)							1					
Iš viso:	4	4	2	2	1	3	2	1		1	1	

2024 – II pusmetis

Vilniaus universitetas

Dalyvavimas Konferencijose 2023/2024 (I pusmetis)		
Planas	Įvykdyta	Konferencijos tipas
13th annual Counter Fraud, Cybercrime and Forensic Accounting Conference 2024.06.12/13 , Portsmouth, UK.	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.	Tarptautinė

Doktorantūros studijų pasiekimai

Vilniaus universitetas

Dalyvavimas tarptautinėse konferencijose	
	Aprašas
	D.Breskuvienė, G.Dzemyda „Clustering-based optimization in fraud detection classifier training“ EURO2022 2022.07.03 /06 ESPOO, FINLAND
	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
	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

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.	2021 m. gruodžio mėn. – 2022 m. vasario mėn.	Parengta mokslinės literatūros apžvalga. Esant poreikiui ji toliau yra pildoma.
	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.	<u>Mokslinio tyrimo vykdymas:</u>		Atlikta
	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:	2022 m. lapkričio mėn. – 2023 m. spalio mėn.	Atliktas teorinis tyrimas. Esant poreikiui toliau pildomas
	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.	2022 m. lapkričio mėn. – 2023 m. spalio mėn.	
	2.3. Empirinis tyrimas:		
	2.3.1. Sudarytų metodų pritaikymas praktinių uždavinių sprendimui.		Atliekamas empirinis tyrimas. Esant poreikiui toliau pildomas
	2.3.2. Gautų duomenų analizė, rezultatų apibendrinimas, išvadų parengimas.	2024 m. kovo mėn. – 2024 m. spalio mėn. 2024 m. spalio mėn. –2025 m. vasario mėn.	

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; 3.2. Analitinės disertacijos dalies parengimas; 3.3. Teorinės disertacijos dalies parengimas; 3.4. Eksperimentinės disertacijos dalies parengimas; 3.5. Bendrųjų išvadų formulavimas.	2024 m. spalio mėn. – 2025 m. vasario mėn. 2024 m. kovo mėn. – 2025 m. rugpjūčio mėn. 2024 m. kovo mėn. – 2025 m. rugpjūčio mėn. 2024 m. kovo mėn. – 2025 m. rugpjūčio mėn. 2024 m. kovo mėn. – 2025 m. rugpjūčio mėn.	Rengiama analitinė disertacijos dalis
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

- **Preliminari disertacijos tema:**

Klasifikatoriaus (nesubalansuotos) mokymo aibės optimizavimas, siekiant geresnės klasifikavimo kokybės.

- **Tyrimo objektai:**

Nesubalansuotų duomenų klasifikavimas finansiniams nusikaltimams identifikuoti.

- **Tikslas:**

Sukurti arba patobulinti jau egzistuojantį mašininio mokymosi algoritmą, siekiant pagerinti klasifikavimo rezultatus nesubalansuotiems duomenims.

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.

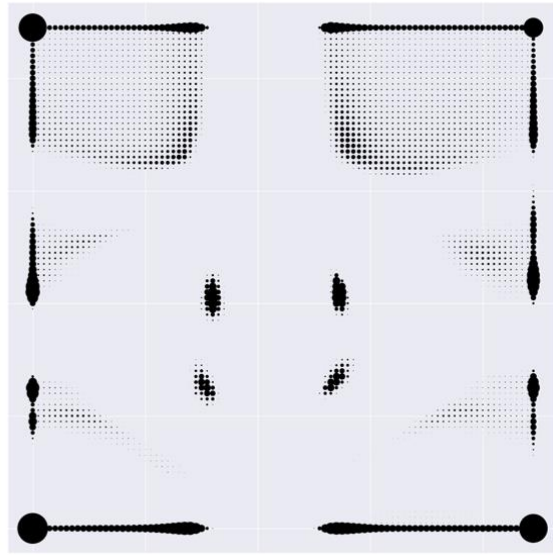
Atlikti darbai

- Pasiūlytas metodas sprendžiantis požymių atrinkimo uždavinį scenarijuose, kuriems būdingi labai nesubalansuoti duomenys
- Įdentifikuotos esamos nesubalansuotų duomenų požymių atrankos spragos
- Siūlomame metode integruotas SOM žemėlapis, kuris gali apdoroti požymių triukšmą ir nustatyti dėsningumus nesubalansuotuose duomenyse, kad pagerintų klasifikavimo uždavinius.
- Įvardintos problemos publikacijose susijusiose su finansinių nusikaltimų aptikimu, fokusuojantis labiausiai į sukčiavimus susijusius su kredito kortelėmis.

Požymių erdvės transformacija

Vilniaus universitetas

Current Age	Gender	City	State	FICO Score	Num Credit Cards	Card	Card Brand
36	-0.10	-0.31	-0.17	719	4	3	-0.14
44	-0.10	-0.50	-0.11	737	2	0	-0.14
47	0.10	-1.01	-0.4?	683	3	0	0.08
56	-0.10	-0.94	-0.11	782	5	2	-0.14
46	0.10	-0.79	-0.01	688	3	2	-0.14



	x	y	weight_vector
0	0	0	[0.41111601223857935, 0.5736991624061853, 0.13...
1	0	1	[0.4108185113360495, 0.5740438555705382, 0.131...
2	0	2	[0.41051656729556957, 0.5743938827703574, 0.13...
3	0	3	[0.4102081158436499, 0.5747521063450655, 0.131...
4	0	4	[0.40988994024247694, 0.5751225539807168, 0.13...

Map 90x90
Number of neurons 8 100

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**

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

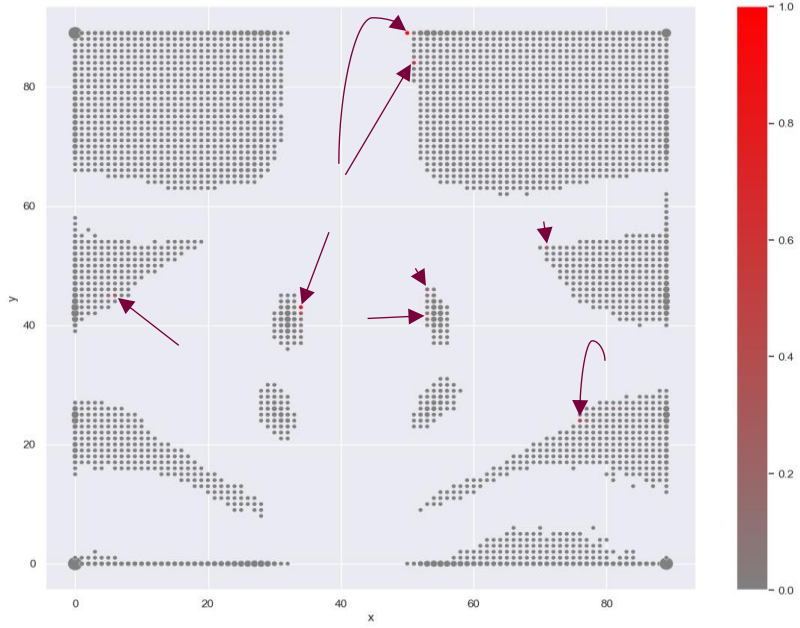
*FIDSOM with XGB classifier selecting 23 features. Data split is done by selecting 70% of the first data points for training and 30% remaining data points for testing. For SOM training, 500 iterations were used with a sigma of 15 and a learning rate of 1.8.

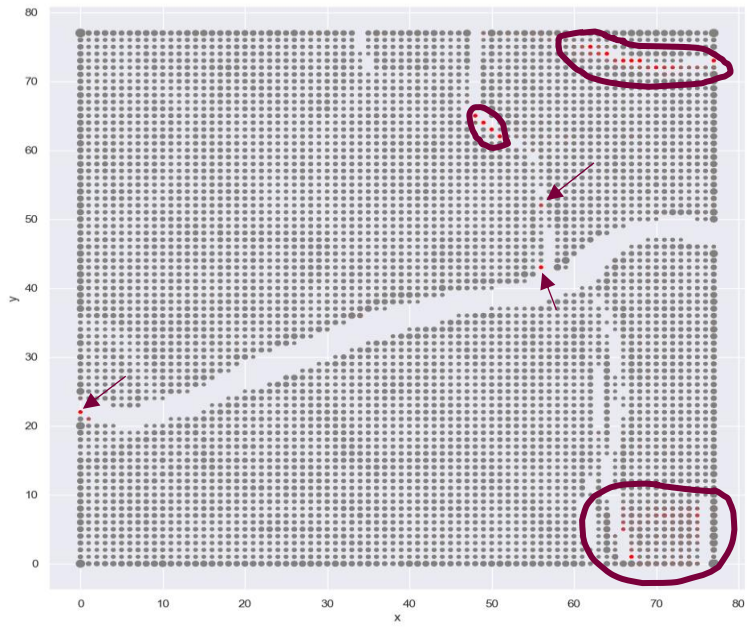
[50] Fiore, U., De Santis, A., Perla, F., Zanetti, P., Palmieri, F.: Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. Information Sciences **479**, 448–455 (2019)

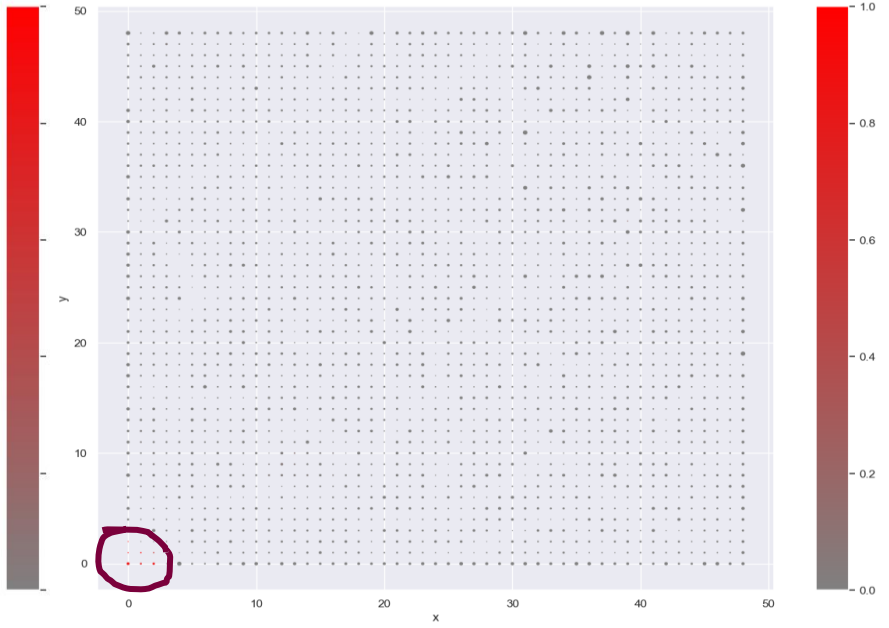
[51] Fanai, H., Abbasimehr, H.: A novel combined approach based on deep autoencoder and deep classifiers for credit card fraud detection. Expert Systems with Applications **217**, 119562 (2023)

Neuronų kategorizavimas pagal transakcijų pobūdžio intensyvumą

Vilniaus universitetas



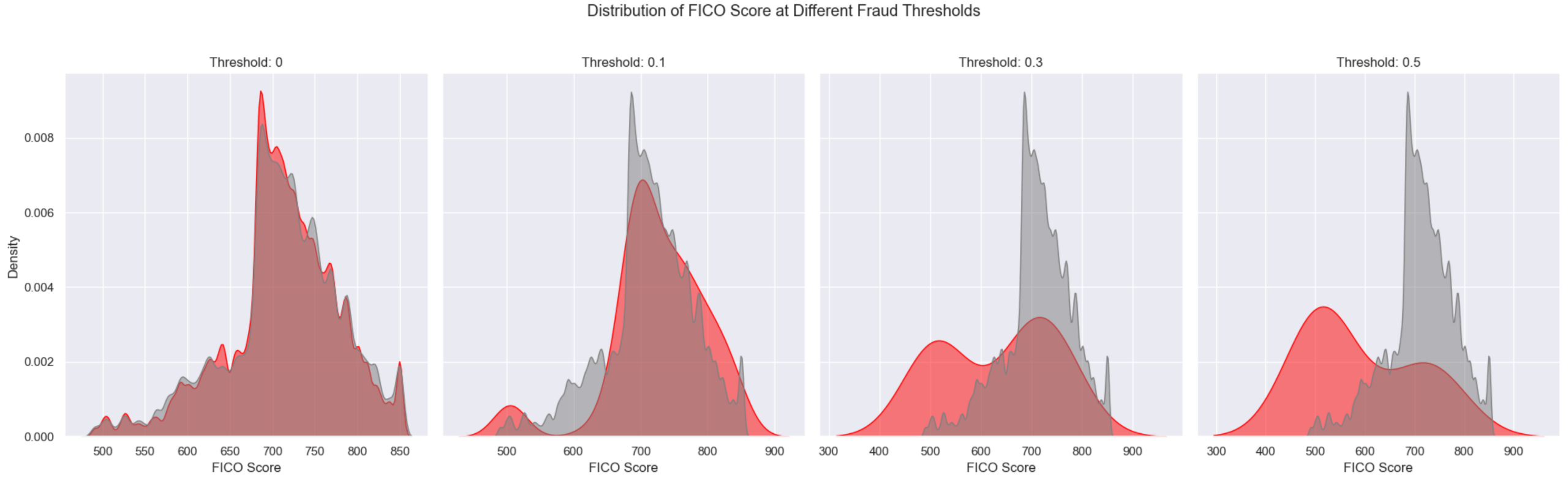




Pasiūlyto algoritmo patobulinimas

Vilniaus universitetas

- BMU pažymimas kaip neuronas turinti finansinio sukčiavimo transakcijas, jei jame yra daugiau nei x% sukčiavimo atvejų



Informacijos prieaugio ir KL – divergencijos panaudojimas

Vilniaus universitetas

Informacijos prieaugis (Information Gain) yra matas, naudojamas duomenų moksle, siekiant nustatyti, kiek informacijos tam tikras požymis prisideda prie klasifikacijos proceso. Jis matuoja, kiek sumažėja entropija, kai duomenys suskirstomi pagal konkretų požymį.

$$Entropy(D) = -\sum_{i=1}^{c} p(i) \log_2 p(i),$$

$$IG = Entropy(D) - \sum_{j=1}^{k} \frac{n_j}{n} Entropy(D_j),$$

KL divergencija (Kullback-Leibler divergencija) yra skirtumo matas tarp dviejų tikimybių pasiskirstymų. Ji nurodo, kiek papildomos informacijos reikia, kad būtų aprašytas vienas pasiskirstymas naudojant kitą.

$$D_{KL}(P||Q) = \sum_{i} P(i) \log \frac{P(i)}{Q(i)},$$

KITO PUSMEČIO DARBO PLANAS

- Pabaigti eksperimentus susijusius su naujomis hipotezėmis
- Apibendrinti rezultatus
- Pristatyti rezultatus DAMSS konferencijoje
- Ruošti disertacijos tekstą.

Vilniaus universitetas

Ačiū už dėmesį!

Klausimai?