Klasterizavimo algoritmai didelės apimties medicinos duomenims

Ataskaita už III-ąjį doktorantūros kursą (2021/2022 metų I pusmetis)

Doktorantas: Roma Puronaitė

Vadovas: prof. dr. Audronė Jakaitienė

Doktorantūros pradžios ir pabaigos metai: 2017 - 2023

2022 m. kovas

Tyrimo objektas

- Klasterizavimo algoritmai
- Didelės apimties medicinos duomenys

Tyrimo tikslas

- Pasiūlyti metodą didelės apimties medicinos duomenims klasterizuoti, atsižvelgiant į duomenų dinamikos laike savybes.

Uždaviniai

1. Ištirti didelės apimties medicinos duomenų klasterizavimui dažniausiai taikomus klasterizavimo metodus.
2. Pasiūlyti klasterizavimo algoritmą ar esamo metodo patobulinimą, kuris atsižvelgtų į dinamiką laike.
3. Pritaikyti atrinktus algoritmus ir pasiūlytą sprendimą realiems medicinos duomenų rinkiniams.
4. Pasiūlyti algoritmo integravimo į sveikatos priežiūros įstaigos informacinę sistemą modelį.

Doktorantūros planas

Studijų metai	Egzaminai [1]		Dalyvavimas konferencijose [2]		Publikacijos [3]		
	Planas	Įvykdyta	Planas	Įvykdyta	Planas	Įvykdyta	Būklė [4]
I (2017/2018)	2	2		1		1	1 publikuota
II (2018/2019)	2	0	1	5		2	2 publikuota
III (2021/2022)	**0**	**1 (skola iš II metų) + Antrasis bus laikomas 2022-03-30**	**1**	**1 + (1)****	**1**	**2**	**2 publikuota**
IV (2022/2023)	0	0			*1*		
Iš viso:	4	2	2	6	2	3+1	

*Iki 2021 m. rugsėjo 30 d. akademinės atostogos

pav.: Visų studijų planas

Ataskaita už III mokslo metus

1 Egzaminai:

Egzaminai		
Planas	**Įvykdyta**	**Būklė**
Netiesiniai statistikos modeliai masinių duomenų analizėje	Netiesiniai statistikos modeliai masinių duomenų analizėje, 2022 m. kovo 18 d.	Išlaikytas
Daugiamačių duomenų vizualizavimo metodai	Daugiamačių duomenų vizualizavimo metodai, 2022 m. kovo 30 d.	Nelaikytas

2 Konferencijos:

Dalyvavimas konferencijose		
13th International Conference BIOMDLORE 2021, 2022 m. spalio mėn. 21-23 d, Vilnius, Lietuva	Bliudzius, Antanas, **Puronaite, Roma**, Trinkunas, Justas, Jakaitiene, Audrone, Kasiulevicius, Vytautas, „Research on physical activity variability and changes of metabolic profile in patients with prediabetes using Fitbit activity trackers data“, *13th International Conference BIOMDLORE 2021*, 2022 m. spalio mėn. 21-23 d, Vilnius, Lietuva Note: [1] Antanas Bliudzius and Roma Puronaite contributed equally to this work.	Tarptautinė
31st International Biometric Conference, 2022 m. liepos 10-15 d., Ryga, Latvija Priimtas žodinis pranešimas.	**Puronaite, Roma**, Ramanauskaitė, Dovilė, Burneikaitė, Greta, Švaikevičienė, Kristina, Šavareikaitė, Alicija, Vaitkute, Samanta, Jakaitienė, Audronė, Dambrauskas, Laimis, Jurevičienė, Elena, Trinkūnas, Justas, Kasiulevičius, Vytautas, Kazėnaitė, Edita, „Challenges of modeling depression and anxiety risk using data from large healthcare databases: systematic review and situation analysis“, *31st International Biometric Conference*, 2022 m. liepos 10-15 d., Ryga, Latvija	Tarptautinė

Ataskaita už III mokslo metus

1 Publikacijos:

Publikacijos			
Planas	**Įvykdyta**	**Būklė**	**Publikacijos tipas**
Technology and health care	Bliūdžius, A., **Puronaitė, R.**, Trinkūnas, J., Jakaitienė, A., & Kasiulevičius, V. (2022). Research on physical activity variability and changes of metabolic profile in patients with prediabetes using Fitbit activity trackers data. Technology and health care, 30(1), 231-242. doi:10.3233/THC-219006 [DB: Scopus; Social Sciences Citation Index (Web of Science)] [Indėlis: 0,268] [Citav. rod.: 1.285 (2020, SCIE)] [M.kr.: M 001,N 009] *Note: Antanas Bliudzius and Roma Puronaite contributed equally to this work.*	Publikuota	IF 1.285
International journal of environmental research and public health	Jurevičienė, E., Burneikaitė, G., Dambrauskas, L., Kasiulevičius, V., Kazėnaitė, E., Navickas, R., **Puronaitė, R.**, Smailytė, G., Visockienė, Ž., & Danila, E. (2022). Epidemiology of chronic obstructive pulmonary disease (COPD) comorbidities in Lithuanian national database: a cluster analysis. International journal of environmental research and public health, 19(2), 1-14. doi:10.3390/ijerph19020970 [DB: MEDLINE; Embase; Scopus; Social Sciences Citation Index (Web of Science); Science Citation Index Expanded (Web of Science)] [Indėlis: 0,067] [Citav. rod.: 3.390 (2020, SCIE); 3.390 (2020, SSCI)] [M.kr.: M 001]	Publikuota	IF 3.390

Ataskaita už III mokslo metus

1 Planas:

	Darbo pavadinimas	Atlikimo terminai	Pastabos
1	Mokslinių tyrimų disertacijos tema apžvalga ir analizė (Lietuvoje ir užsienyje): Anotuotos bibliografijos sudarymas. Mokslinės literatūros apžvalga. Egzistuojančių metodų taikymo medicinos duomenims analizavimas.	2017 m. spalio mėn. – 2018 m. birželio mėn.	Naudojant sisteminės literatūros apžvalgos metodą, atlikta literatūros analizė šiomis temomis: Sisteminė literatūros apžvalga – poliligotumo modelių klasterizavimo metodai (išnagrinėta 151 publikacija, identifikuoti šioje tematikoje naudojami klasterizavimo metodai) ir **Sisteminė literatūros apžvalga – netiesiniai statistikos modeliai: poliligotumo analizės tyrimai naudojant didelės apimties medicinos duomenų masyvus (išnagrinėta 206 publikacijos, identifikuoti šioje tematikoje naudojami klasterizavimo metodai)** **Apžvelgtos antriniu tikslu naudojamos administracinės duomenų bazės analizės strategijos.**
2	Mokslinio tyrimo vykdymas: 2.1. Tyrimo metodikos sudarymas: 1. Disertacijos tikslo formulavimas. 2. Disertacijos uždavinių formulavimas.	2018 m. birželio mėn. – 2018 m. lapkričio mėn.	Suformuluotas tikslas ir uždaviniai
	2.2. Teorinis tyrimas: Matematinio modelio sudarymas. Algoritmų konstravimas ir tobulinimas.	2018 m. lapkričio mėn. – 2019 m. birželio mėn.	Nagrinėjami ir išbandomi nagrinėtoje literatūroje pasiūlyti metodai atsižvelgiantys į duomenų dinamiką. Empiriniui tyrimui atrinkta: • Skerspjūvio tipo analizė • Temporalinio fenotipavimo metodai (PARAFAC2, neneigiamas matricos faktorizavimas) • Laiko dimensijos transformavimas (siūloma transformacijai naudoti Poincare grafikus (indeksus), laiko eilučių išlyginimą, netiesinius statistinius modelius) • Kiti: gilusis mokymas, gilieji rekurentiniai neuroniniai tinklai, arba giliojo mokymo teksto analizės metodai.

Ataskaita už III mokslo metus

1 Planas:

	1. Siūlomų algoritmų pritaikymas medicinos duomenims. 2. Siūlomų algoritmų tobulinimas, atsižvelgiant į gautus rezultatus.	**2018 m. birželio mėn. – 2022 m. vasario mėn.***	• Diabetas • Lėtinė obstrukcinė plaučių liga • Depresija ir nerimas Kiekvieno paciento atveju buvo sudaryti dvinariai kintamieji atitinkantys ligos (išskirta 30 lėtinių ligų) buvimą arba nebuvimą atitinkamo paciento ligos istorijoje per tyrimo laikotarpį. Šiuose pogrupiuose pritaikius hierarchinį klasterizavimą (angl. HCA) gauti ligų klasteriai, kurie buvo vertinti gydytojų specialistų, dalis šių klasterių įvertinti kaip turintys klinikinį paaiškinimą. Taip pat pritaikyta tiriamoji faktorinė analizė (angl. EFA). Koreliacijos matrica sudaryta naudojant tetrachorinę koreliaciją tarp dvinarių kintamųjų porų. Atlikta vizuali daugiamačių poliligotumo duomenų analizė Pradėta: • Temporalinio fenotipavimo metodai (PARAFAC2, neneigiamas matricos faktorizavimas) Numatoma: • Laiko dimensijos transformavimas (siūloma transformacijai naudoti Poincare grafikus (indeksus), laiko eilučių išlyginimą, netiesinius statistinius modelius) • Kiti: gilusis mokymas, gilieji rekurentiniai neuroniniai tinklai, arba giliojo mokymo teksto analizės metodai.
	2.4. Gautų duomenų analizė, apibendrinimas, išvadų parengimas: 1. Algoritmų tikslumo įvertinimas, palyginimas su kitų autorių metodais, atrinktais remiantis išanalizuota moksline literatūra. 2. Gautų rezultatų apibendrinimas. 3. Išvadų parengimas.	2020 m. vasario mėn. – 2020 m. liepos mėn. **2022 m. vasario mėn. – 2022 m. liepos mėn.***	**2021/2022 metais I pusmetis** Planuojamos publikacijos (pradėta rengti): Depresijos ir nerimo prognozavimas remiantis administracinio pobūdžio duomenimis, pritaikant Poincare plot metodą ir apskaičiuotus indeksus (Health Informatics tematikos žurnale) Pranešimas konferencijoje (piimta): **Puronaite, Roma**, Ramanauskaitė, Dovilė, Burneikaitė, Greta, Švaikevičienė, Kristina, Šavareikaitė, Alicija, Vaitkute, Samanta, Jakaitienė, Audronė, Dambrauskas, Laimis, Jurevičienė, Elena, Trinkūnas, Justas, Kasiulevičius, Vytautas, Kazėnaitė, Edita, „Challenges of modeling depression and anxiety risk using data from large healthcare databases: systematic review and situation analysis“, *31st International Biometric Conference*, 2022 m. liepos 10-15 d., Ryga, Latvija Pranešimas pristatomas 2021/2022 metais II pusmetis

Ataskaita už III mokslo metus

1 Planas:

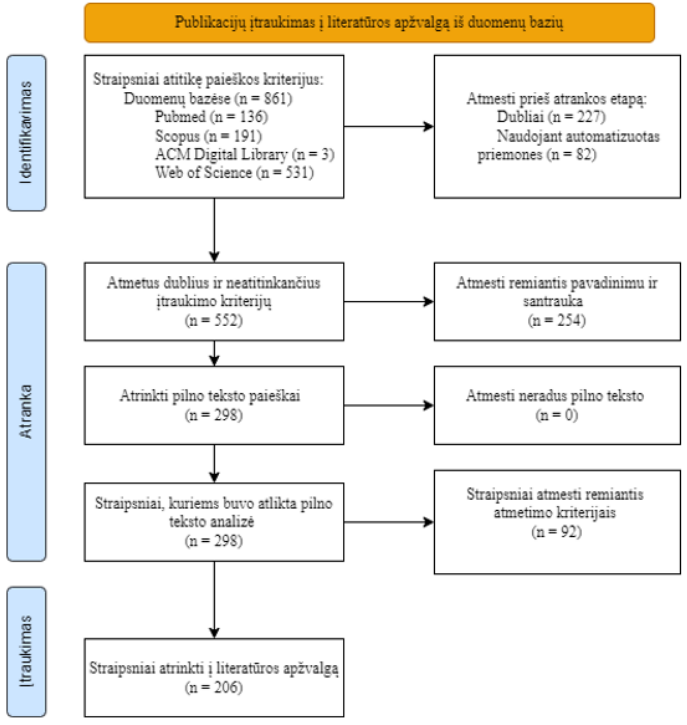
			Planuojama: Pranešimas tarptautinėje statistikos konferencijoje: ISCB2023 (Rezultatų apibendrinimas) 2022/2023 metais II pusmetis
3	Atskirų daktaro disertacijos dalių (tyrimo metodikos, rezultatų, ginamų teiginių, išvadų, ir kt.) parengimas: 1. Tyrimų apžvalga ir analizė. 2. Tyrimo metodikos sudarymas. 3. Teorinis tyrimas. 4. Empirinis tyrimas. 5. Gautų duomenų analizė, apibendrinimas. 6. Išvados, įvadas, literatūros sąrašas.	2020 m. rugsėjo mėn. – 2021 m. gegužės mėn. **2022 m. rugsėjo mėn. – 2023 m. gegužės mėn.***	2021/2022 metais I pusmetis Pradėtos rengti dalys: mokslinės literatūros apžvalga, tyrimo metodika, duomenų analizė.
4	Daktaro disertacijos parengimas ir svarstymas padalinyje	2021 m. birželio mėn. **2023 m. birželio mėn.***	
5	Daktaro disertacijos gynimas	2021 m. rugsėjo mėn. **2023 m. rugsėjo mėn.***	

Ataskaita už III mokslo metus

Literatūros apžvalga papildyta nauja dalimi - Netiesiniai statistikos modeliai: poliligotumo analizės tyrimai naudojant didelės apimties medicinos duomenų masyvus

Sisteminės apžvalgos pagrindinės komponentės:

1. Poliligotumas – kelios vienu metu egzistuojančios lėtinės ligos.
2. Netiesiniai modeliai.
3. Antrinis (pakartotinis) sveikatos duomenų naudojimas.



Ataskaita už III mokslo metus

1. Dažniausiai naudojami pakartotinai naudojamų (ne pagal pirminę paskirtį) duomenų šaltiniai, kai analizei buvo naudoti netiesiniai statistiniai modeliai: administraciniai, elektroniniai sveikatos įrašai, sveikatos tyrimų (statistinių) duomenys.
2. Dažniausiai naudoti netiesiniai statistikos modeliai: logistinė regresija, Puasono, neigiama binominė regresijos.
3. Nėra vieningo poliligotumo apibrėžimo: poliligotumas vertinamas pagal ligų skaičių, pagal ligų grupes, naudojant poliligotumo indeksus.
4. Dažniausios temos: poliligotumas ir sveikatos priežiūros resursų naudojimas (išlaidos, paslaugų dažnis) - 15 proc., poliligotumo tyrimai (poliligotumo struktūra, klasteriai) - 10,7 proc., poliligotumas ir socialiniai netolygumai (pagal gyvenamąją vietą, socialinį statusą) - 9,2 proc.

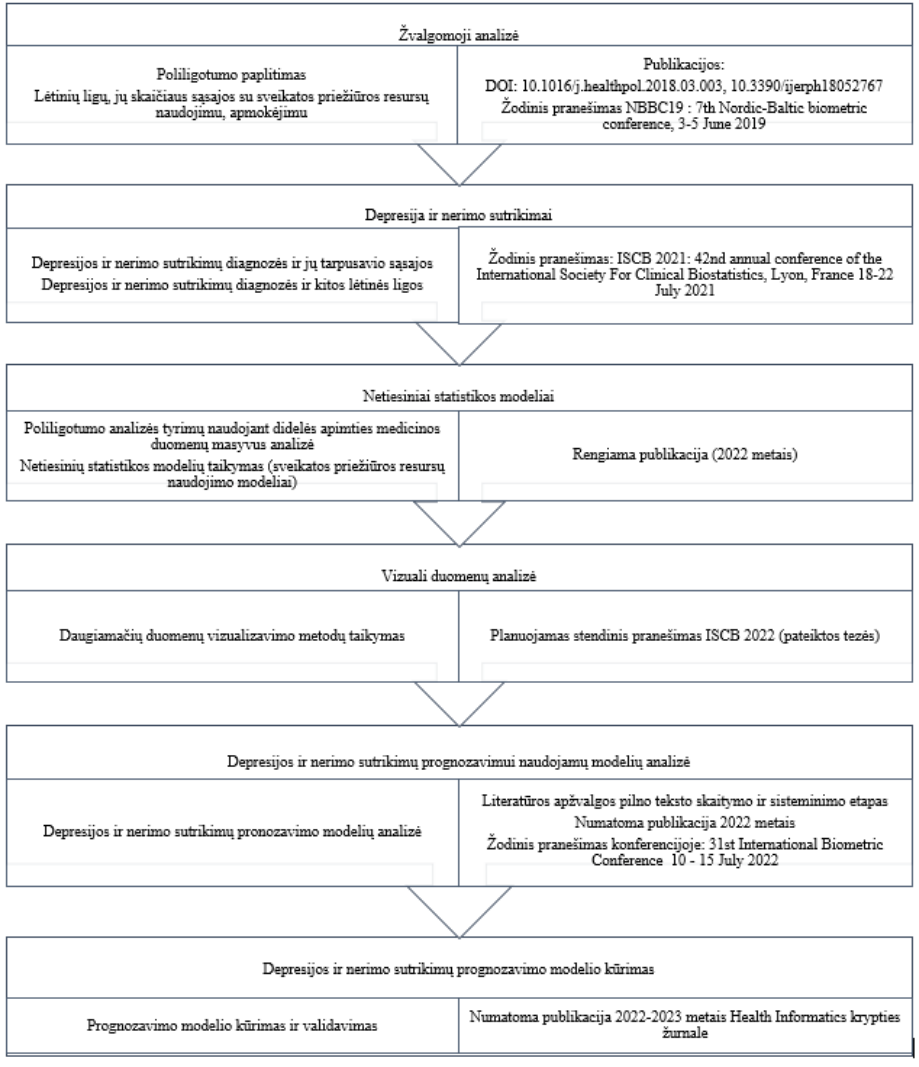
Ataskaita už III mokslo metus

1 Duomenų parengimas analizei ir klasterizavimo algoritmai

Skerspjūvio tipo	Temporalinis fenotipavimas (fenotipavimas laike)	Laiko dimensijos transformacija	Kiti būdai

• Būsena kintamiesiems įvertinama apimant visą laikotarpį, pvz. pacientas turėjo x diagnozę tarp 2012-2014 metų.
• Pirminis objektas matrica:
  • Pacientai x Diagnozės
• Pvz. hierarchinis klasterizavimas, tiriamoji faktorinė analizė.
	• Būsena gali kisti, atsižvelgiama į laiko žymę, eilės tvarką, pvz. pacientas turėjo x diagnozę 2012, 2014 metais, bet neturėjo 2013.
	• Pirminis objektas tenzorius:
	  • Pacientai x Diagnozės x Vizitai (laiko žymė)
	• Pvz. tenzorių faktorizacija (PARAFAC2 algoritmas ir kt.)
		• Būsenos kaita įvertinama atliekant transformaciją
		• Pirminis objektas tenzorius:
		  • Pacientai
		  • Diagnozės
		  • Vizitas (laiko žymė)
		• Po transformacijos klasterizavimas atliekamas matricoms
		  • Matrica:
		  • pvz. Pacientai x Diagnozės
		• Laiko eilučių išlyginimo ir panašumo vertinimas
		• Pvz. dinaminis laiko skalės kraipymas-> klasterizavimo algoritmai (HCA, EFA)
			• Gilieji neuroniai tinklai
			• Gilusis mokymas
			• Teksto analizė
			• Kryptiniai grafai
			• Jungtinis klasterizavimas

1 Mokslinio tyrimas "Lėtinių neinfekcinių ligų paplitimo, tarpusavio sąveikos, sveikatos priežiūros paslaugų bei vaistų vartojimo bei klinikinių baigčių vertinimas Lietuvoje".



III (2021/2022) mokslo metų II pusmečio planas

1. Egzaminas: Daugiamačių duomenų vizualizavimo metodai (2022 m. kovo 30 d.)
2. Publikacija sisteminės apžvalgos "Netiesiniai statistikos modeliai: poliligotumo analizės tyrimai naudojant didelės apimties medicinos duomenų masyvus"pagrindu
3. Publikacija (Darbinis pavadinimas: "Multimorbidity, depression and anxiety patterns: temporal phenotyping approach"- statusas parengta)