Vilniaus universitetas
Duomenų mokslo ir skaitmeninių technologijų institutas
L I E T U V A

DOKTORANTŪROS METINĖ ATASKAITA

2018 m. spalio mėn. 1 d. – 2019 m. rugsėjo mėn. 30 d.

INFORMATIKOS STUDIJŲ PROGRAMOS

DOKTORANTĖ MARTA KARALIUTĖ

Disertacijos pavadinimas: Erdvės-laiko duomenų klasifikavimas naudojant diskriminantines funkcijas

Doktorantūros laikotarpis: 2017 – 2021

Vadovas: prof. dr. Kęstutis Dučinskas

Konsultantas: prof. habil. dr. Gintautas Dzemyda

- **Tyrimo objektas:**

  Erdvės-laiko duomenų klasifikavimo metodai

- **Tyrimo tikslas:**

  Atlikti erdvės-laiko duomenų statistinį klasifikavimą naudojant diskriminavimo funkciją bei klasifikavimo klaidų tikimybes. Išvestų formulių pagrindu sukurti algoritmus ir juos pritaikyti realių duomenų analizei.

Tyrimo uždaviniai:

- Erdvės-laiko duomenų (ELD) klasifikavimo klaidų tikimybių ir jų įvertinių analitinių formulių išvedimas bei savybių tyrimas, taikant ML ir Bajeso parametrų įvertinius;
- ELD duomenų vidutinės klasifikavimo į dvi klases rizikos aproksimacijos išvedimas;
- ELD duomenų vidutinės klasifikavimo klaidos aproksimacijos išvedimas daugiaklasių atveju.

Planuojami rezultatai:

- Siūlomų klasifikavimo metodų realizavimas, įvairių parametrų įtakos klasifikavimo rizikai tyrimas naudojant dirbtinius (generuotus) ir realius duomenis bei specializuotą programinę įrangą (R-INLA).

2018/2019 m. m. darbo planas

- Išlaikyti 2 egzaminus (*„Optimizacijos teorija, algoritmų sudėtingumas“, „Skaitiniai metodai“*).
- Dalyvavimas tarptautinėje konferencijoje.
- Straipsnis tarptautiniame mokslo žurnale.
- Mokslinio tyrimo vykdymas:

  1. Tyrimo metodikos sudarymas
  2. Teorinis tyrimas

Ataskaita už 2018/2019 mokslo metus:

- Išlaikyti egzaminai: *„Optimizacijos teorija, algoritmų sudėtingumas“, „Skaitiniai metodai“*.

- Dalyvauta doktorantų bendrųjų gebėjimų mokymuose:
  - Latex (1,25 ECTS kredito)
  - Atvirosios prieigos kompetencijų tobulinimas, taikant žaidimo metodą (0,5 ECTS kredito)

Ataskaita už 2018/2019 mokslo metus:

- Dalyvauta konferencijose:
  - Duomenų analizės metodai programų sistemoms (DAMSS), Druskininkai 2018 (posteris).
  - Spatial Statistics 2019: Towards Spatial Data Science, Sitges, Ispanija (posteris).
  - XII International Conference COMPUTER DATA ANALYSIS & MODELING 2019 Stochastics & Data Science, Minskas, Baltarusija (mokslinis pranešimas).

Ataskaita už 2018/2019 mokslo metus:

- Publikacijos:
  - Marta Karaliutė, Kęstutis Dučinskas, Laura Šaltytė-Vaisiauskė. Expected Error Rate in Linear Discrimination of Balanced Spatial Gaussian Time Series. XII International Conference COMPUTER DATA ANALYSIS & MODELING 2019 Stochastics & Data Science, Minskas, Baltarusija. 2019, 172-175.
  - Marta Karaliutė, Laura Šaltytė-Vaisiauskė, Kęstutis Dučinskas. Linear Discriminant Analysis of Spatio-temporal Unemployment Rate Data in Lithuania. Spatial Statistics. 2019 (teikiamas).

Erdvės ir laiko modeliai

Norime klasifikuoti erdvės-laiko stebėjimus Gauso atsitiktiniame lauke $\{Z(s;t): s \in D \subset \mathbb{R}^2, t \in [0,\infty)\}$.

Stebėjimo $Z(s;t)$ modelis populiacijoje Ω_l yra

$$Z(s;t) = \mu_l(s;t) + \varepsilon(s;t).$$

Turime $S_n = \{s_i \in D; i = 1,\dots,n\}$ vietų, kuriose paimti mokymo stebėjimai.

Suformuosime T-mačius stebinių vektorius kiekvienam erdvės taškui $Z_i = \big(Z(s_i, 1), \dots, Z(s_i, T)\big)', i = 0, \dots, n.$

Tada mokymo imtis bus $n \times T$ matrica $M = \begin{pmatrix} M_1 \\ M_2 \end{pmatrix}$,

kur $M'_1 = (Z_1, \dots, Z_{n1})$, $M'_2 = (Z_{n1+1}, \dots, Z_n)$.

Tuomet $Z_i \sim N_T(m_i, \Sigma)$.

Nagrinėjama $Z_0 = \big(Z(s_0, 1), \dots, Z(s_0, T)\big)$ klasifikavimo problema, kai duota mokymo imtis M.

Kai populiacijos apriorinės tikimybės yra žinomos $\pi_1(s)$ ir $\pi_2(s)$. Tada Bajeso diskriminantinė funkcija (BDF), minimizuojanti klaidingo klasifikavimo tikimybę, yra suformuota pagal sąlyginio tankio santykio logaritmą

$$W_m(Z_0;\Psi)=\left(Z_0-(m-XB)'\alpha_0-\frac{B'H'x_0}{2}\right)'\Sigma^{-1}B'G'x_0/\rho+\gamma$$

Kai $W_m(Z_0)>0$, taškas s_0 priskiriamas 1 klasei,

o kai $W_m(Z_0)<0$, taškas s_0 priskiriamas 2 klasei.

Kai populiacijų parametrai nežinomi, yra naudojami jų įvertiniai, o pačios funkcijos yra vadinamos įterptosiomis Bajeso diskriminantinėmis funkcijomis (PBDF) (*ang. plug-in Bayesian discriminant function)*:

$$W_M\left(Z_0;\widehat{\Psi}\right)=\left(Z_0-\left(M-X\hat{B}\right)'\alpha_0-\hat{B}'H'x_0/2\right)'\hat{\Sigma}^{-1}\hat{B}'G'x_0/\rho+\gamma.$$

Nagrinėjamos šios erdvės-laiko kovariacinės struktūros:

Erdvės:

1. remiantis sferiniu semivariogramos modeliu, t. y. $R = (r_{ij})$

kur $r_{ij} = r(|s_i - s_j|) = \begin{cases} 1 - \frac{3}{2}\frac{|s_i - s_j|}{\lambda} + \frac{1}{2}\left(\frac{|s_i - s_j|}{\lambda}\right)^3, & |s_i - s_j| \le \lambda; \\ 0, & |s_i - s_j| > \lambda \end{cases}$

2. Markovo modelis, t. y. $R = (I - \alpha_R W)^{-1}$.

Laiko kovariacijos matrica Σ apskaičiuojama momentų metodu (empirinis įvertis) arba pagal Yule-Walker lygtis AR(2) modeliui.

Aktualioji klaidos tikimybė *(ang. actual error rate (AER))* pagal PBDF $W_M\left(Z_0;\widehat{\Psi}\right)$ yra apibrėžiama:

$$P\left(\widehat{\Psi}\right)=\sum_{l=1}^{2}\pi_l P\left((-1)^l W_M\left(Z_0;\widehat{\Psi}\right)>0|M\right).$$

Lema. *Aktualioji klaidos tikimybė pagal PBDF yra*

$$P\left(\widehat{\Psi}\right)=\sum_{l=1}^{2}\pi_l \Phi\left(\hat{Q}_l\right),$$

kur $\hat{Q}_l=(-1)^l\dfrac{\left(x_0'(B_l-H\hat{B}/2)+\alpha_0' X(\Delta\hat{B})\right)\widehat{\Sigma}^{-1}\hat{B}\prime G\prime x_0/\rho+\gamma}{\sqrt{x_0' G\hat{B}\widehat{\Sigma}^{-1}\Sigma\widehat{\Sigma}^{-1}\hat{B}\prime G\prime x_0/\rho}}$,

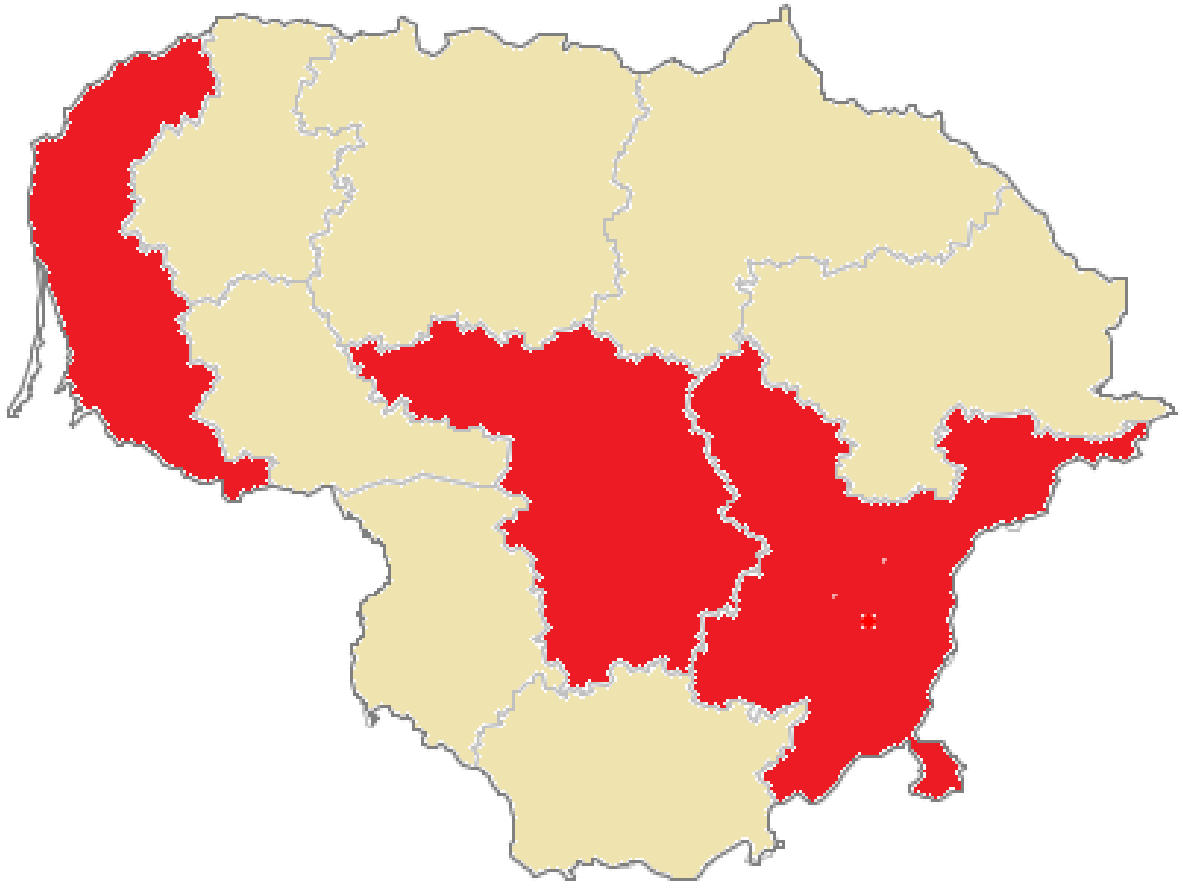
$\Delta\hat{B}=\hat{B}$-*B*.

AER įvertinimas apskaičiuojamas pagal Leave-One-Out įvertį:

$$\mathrm{LO} = \left(\sum_{\mathrm{i}=1}^{\mathrm{n}_1} \mathrm{H}\left(-\mathrm{W}_\mathrm{M}\left(\mathrm{Z}_i, \widehat{\Psi}_{(-i)}\right)\right) + \sum_{\mathrm{i}=\mathrm{n}_1+1}^{\mathrm{n}} \mathrm{H}\left(\mathrm{W}_\mathrm{M}\left(\mathrm{Z}_i, \widehat{\Psi}_{(-i)}\right)\right) \right) \Big/ n$$

kur $\mathrm{H}(\cdot)$ yra Heaviside funkcija, $\widehat{\Psi}_{(-i)}$ - Ψ įvertis iš Z, išskyrus Z_i.

Lietuvos nedarbo lygio diskriminantinė analizė



Klasifikavimo rezultatai

			Laiko modeliai					
			Empirinis įvertis			AR(2) (Yule – Walker)		
			Klaidingai suklasifikuotos apskritys	LO	Apskritys	Klaidingai suklasifikuotos apskritys	LO	Apskritys
Erdvės modeliai	**Sferinė koreliacija**		3	0.3	V, K, Kl	1	0.1	Kl
	Markovo modeliai	**α_R=0.1**	2	0.2	Kl, U	1	0.1	Kl
		α_R=0.2	2	0.2	Kl, U	1	0.1	Kl
		α_R=0.3	2	0.2	Kl, U	1	0.1	Kl
		α_R=0.4	2	0.2	Kl, U	1	0.1	Kl
		α_R=0.5	2	0.2	Kl, U	1	0.1	Kl
		α_R=0.6	2	0.2	Kl, U	1	0.1	Kl
		α_R=0.7	2	0.2	Kl, U	1	0.1	Kl
		α_R=0.8	2	0.2	Kl, U	1	0.1	Kl
		α_R=0.9	2	0.2	Kl, U	1	0.1	Kl
		α_R=1.0	6	0.6	K, Kl, A, M, T, U	3	0.3	KL, P, T

Išvados

1. PBDF buvo įvertintas atsižvelgiant į skirtingas apriorinių tikimybių π_i ir erdvinės koreliacijos reikšmes. Pagal gautus rezultatus tinkamesnis klasifikavimo atvejis yra AR(2) modelis. Abiejų erdvinių struktūrų tikslumas yra vienodas.

2. Klaipėdos apskritis buvo klasifikuojama neteisingai. Galima matyti, kad Klaipėdoje nedarbo lygis yra didesnis nei Vilniaus ir Kauno apskrityse. Taigi Klaipėdą galima laikyti apskritimi tarp aukštesnio ir žemesnio nedarbo lygio.

3. Mažas LO visiems erdviniams atvejams ir AR(2) laiko kovariacijos modelis rodo, kad siūloma erdvės-laiko diskriminantinė analizė yra gana tiksli.

2019/2020 m. m. darbo planas:

- Moksliniai tyrimai

  3. Empirinis tyrimas:

  Siūlomų klasifikavimo metodų realizavimas, įvairių parametrų įtakos klasifikavimo rizikai tyrimas naudojant dirbtinius (generuotus) ir realius duomenis bei specializuotą programinę įrangą (R-INLA).

- Dalyvavimas tarptautinėje konferencijoje.
- Straipsnis tarptautiniame mokslo žurnale.

Ačiū už dėmesį