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

DOKTORANTŪROS METINĖ ATASKAITA

2017 m. spalio mėn. 1 d. – 2018 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:**

  KU Jūros tyrimų instituto duomenys

- **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).

2017/2018 m. m. darbo planas

- Išlaikyti 2 egzaminus („*Atpažinimo teorija*“, „*Daugiamačių duomenų vizualizavimo metodai*“).

- Mokslinių tyrimų disertacijos tema apžvalga ir analizė (Lietuvoje ir užsienyje):

  1. Erdvės-laiko duomenų analizės metodų apžvalga.

  2. Erdvės-laiko duomenų kontekstinio klasifikavimo metodų analizė.

Ataskaita už 2017/2018 mokslo metus:

- Išlaikyti egzaminai: *„Atpažinimo teorija“*, *„Daugiamačių duomenų vizualizavimo metodai“*.

- Mokslinių tyrimų disertacijos tema apžvalga ir analizė (Lietuvoje ir užsienyje).

Erdvės ir laiko modeliai

$$\{Z(s;t): s \in D, t \in T\}$$

Bendras erdvės-laiko duomenų modelis gali būti užrašytas tokia forma:

$$Z(s;t) = \mu(s;t) + \varepsilon(s;t),\ s \in D, t \in T.$$

Erdviniams duomenims modelis parenkamas iš žinomų parametrinių modelių rinkinio. Yra du būdai, kaip tai galima apibendrinti erdvės-laiko duomenims.

1. Sudaryti metriką erdvės-laiko atžvilgiu ir tuomet taikyti izotropinius modelius. Atstumas tarp dviejų erdvės-laiko taškų $(s_1; t_1)$ ir $(s_2; t_2)$ yra:

$$|(s_1; t_1) - (s_2; t_2)| = |(s_1 - s_2; t_1 - t_2)| =$$

$$= \left(a(x_1 - x_2)^2 + b(y_1 - y_2)^2 + c(t_1 - t_2)\right)^{\frac{1}{2}}.$$

2. „atskirti“ erdvės ir laiko priklausomybę:

- *Adityvus erdvės-laiko modelis*:

$$C_{DT}(h_s, h_t) = C_D(h_s) + C_T(h_t)$$

- *Multiplikatyvus erdvės-laiko modelis (atskiriama (separabili) kovariacinė funkcija)*:

$$C_{DT}(h_s, h_t) = C_D(h_s) \cdot C_T(h_t)$$

Erdvės-laiko stebinių kontekstinis (su apmokymu) klasifikavimas

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 sąlyginė Bajeso diskriminantinė funkcija (SBDF), minimizuojanti klaidingo klasifikavimo tikimybę, yra suformuota pagal sąlyginio tankio santykio logaritmą

$$W_m(Z_0) = ln\frac{\pi_1 P_1(Z_0|M)}{\pi_2 P_2(Z_0|M)} =$$
$$= \left(Z_0 - (m - XB)'\alpha_0 - B'H'x_0/2\right)'\Sigma^{-1}B'G'x_0/\rho + \gamma$$

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.

Įterptinė (*ang. plug-in*) sąlyginė Bajeso diskriminantinė funkcija (PSBDF):

$$W_M(Z_0;\widehat{\Psi}) =$$

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

Tada aktualioji klasifikavimo klaida pagal PSBDF lygi

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

2018/2019 m. m. darbo planas:

- Išlaikyti likusius 2 egzaminus (*„Optimizacijos teorija, algoritmų sudėtingumas“, „Skaitiniai metodai“*).

- Straipsnio „Expected error regret in linear discrimination of balanced spatial Gaussian time series” rengimas ir dalyvavimas konferencijoje “Duomenų analizės metodai programų sistemoms“ (DAMSS).

Mokslinio tyrimo vykdymas:

- 1. Tyrimo metodikos sudarymas:

1.1.Tinkamos tyrimo metodikos parinkimas;

1.2. Teorinio ir empirinio tyrimo planavimas.

- 2. Teorinis tyrimas:

2.1. Erdvės-laiko duomenų (ELD) klasifikavimo klaidų tikimybių ir jų įvertinių analitinių formulių išvedimas bei savybių tyrimas, taikant ML ir Bajeso parametrų įvertinius;

2.2. ELD duomenų vidutinės klasifikavimo į dvi klases rizikos aproksimacijos išvedimas;

2.3. ELD duomenų vidutinės klasifikavimo klaidos aproksimacijos išvedimas daugiaklasių atveju.

Ačiū už dėmesį