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

INFORMATIKA (09 P)

ERDVĖS-LAIKO DUOMENŲ KLASIFIKAVIMAS NAUDOJANT DISKRIMINANTINES FUNKCIJAS

Marta Karaliutė

2018 m. spalis

Mokslinė ataskaita DMSTI-DS-09P-18-1

VU Duomenų mokslo ir skaitmeninių technologijų institutas, Akademijos g. 4,
Vilnius LT-08663
www.mii.lt

Santrauka

Šiame darbe apžvelgiami statistiniai erdvės-laiko modeliai ir nagrinėjami erdvės-laiko duomenų kontekstinio (su apmokymu) klasifikavimo, naudojant diskriminantines funkcijas, uždaviniai. Pateikiamos klasifikavimo klaidos tikimybių ir aktualių klaidų formulės, naudojant diskriminantines funkcijas, pagristos marginaliniais ir sąlyginiais klasifikuojamo stebinio skirstiniais. Šios formulės bus taikomos atliekant skaičiavimus tiek su dirbtiniais, tiek su realiais duomenimis.

Reikšminiai žodžiai: erdvinė koreliacija, krigingas, klasifikavimo klaidos tikimybė, diskriminantinė funkcija.

Turinys

1 Įvadas

Norint spręsti statistinius uždavinius pirmiausia reikia surinkti duomenis. Dažnai erdvinių duomenų rinkiniai yra gana nedideli, o taškai, kuriuose atliekami stebėjimai, pasklidę netaisyklingai. Renkant duomenis tam tikrą laiko periodą (paprastai vienodais laiko intervalais), jų gali būti net ir labai daug. Pavyzdžiui, atliekant oro užterštumo tyrimus, monitoringo sistemą gali sudaryti mažiau nei šimtas taškų, tačiau kiekviename taške duomenys renkami kas valandą. Sprendžiant „erdvinį“ uždavinį, paprastai siekiama interpoliuoti arba įvertinti erdvinį vidurkį. Duomenys, rinkti tam tikrą laiko tarpą, dažniausiai naudojami ateities reikšmėms prognozuoti arba sezoniškumams tirti. Tuo tarpu erdvės-laiko uždaviniai jungia abu uždavinių tipus. Vienas iš akivaizdžių sprendimo būdų yra analizuoti erdvėje rinktus duomenis kiekvienu atskiru laiko momentu, t. y. ignoruoti reiškinio kitimą laike. Kita vertus, galima dirbti su laiko eilutėmis skirtinguose taškuose (daugiamatės laiko eilutės). Tačiau tada negalėtume modeliuoti, prognozuoti ar įvertinti reikšmių taškuose, nesančiuose imtyje. Bendru atveju reikia atsižvelgti į koreliacijas ir erdvėje, ir laike bei nustatyti ryšius tarp jų. (Dučinskas, Šaltytė-Benth 2003). Klasifikuojant stebėjimus svarbu įvertinti, kaip tiksliai tai atliekama. Dažnai taikomos tokios charakteristikos, kaip bendras klasifikavimo tikslumas bei klaidingo klasifikavimo tikimybių įverčiai. Klasifikavimo kokybę nusako klaidingo klasifikavimo tikimybių įverčiai, kurie parodo, kokia yra tikimybė suklysti klasifikavimo metu kiekvienai iš klasių (Čekanavičius ir Murauskas 2008). Trečiame skyriuje nagrinėjami Gausinių erdvės-laiko duomenų klasifikavimo metodai, naudojant įvairią erdvinę informaciją (geometrinę ir statistinę).

2 Erdvės ir laiko modeliai

Erdvės-laiko modelis paprastai užrašomas taip:

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

kur s – erdvės, o t – laiko koordinatės, kur $D \subset \mathbb{R}^d$ yra erdvinių indeksų aibė.

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

$$Z(s;t) = \mu(s;t) + \varepsilon(s;t),\ s \in D, t \in T, \quad (1)$$

čia $\mu(s;t)$ yra vidurkio funkcija; $\varepsilon(s;t)$ – nulinio vidurkio atsitiktinis laukas.

Apibrėžimas. $K(s,t;u,r)$ yra vadinama stacionaria erdvės-laiko kovariacine funkcija, jeigu $K(s,t;u,r) = C(s-u;t-r)$, kur $s,u \in \mathbb{R}^d$, $t,r \in \mathbb{R}$, $C(\cdot,\cdot)$ – bet kokia funkcija, tenkinanti atitinkamas kovariacinės funkcijos savybes.

Jeigu atsitiktinis procesas $Z(s;t)$ turi pastovų vidurkį μ ir stacionarią kovariacinę funkciją $C(h;\tau)$, jis yra vadinamas antros eilės (arba silpnai) stacionariu. Stiprus $Z(s;t)$ stacionarumas reikalauja, kad visos tikimybinės atsitiktinių procesų $Z(s;t)$ ir $Z(s+h;t+\tau)$ charakteristikos, visiems $h \in \mathbb{R}^d$ ir $\tau \in \mathbb{R}$, sutaptų.

Remiantis erdvės-laiko kovariacinės funkcijos išraiška, stacionari erdvės-laiko koreliacinė funkcija užrašoma:

$R(h;\tau) \equiv \frac{C(h;\tau)}{C(0;0)}$, $h \in \mathbb{R}^d$ ir $\tau \in \mathbb{R}$.

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

Galima sudaryti metriką erdvės-laiko atžvilgiu ir tuomet taikyti izotropinius modelius, t. y. modelius, skirtus erdvinių duomenų analizei. Atstumas tarp dviejų erdvės-laiko taškų $(s_1;t_1)$ ir $(s_2;t_2)$ yra:

$$|(s_1;t_1)-(s_2;t_2)| = |(s_1-s_2;t_1-t_2)| =$$
$$= \left(a(x_1-x_2)^2 + b(y_1-y_2)^2 + c(t_1-t_2)\right)^{\frac{1}{2}}, \qquad (2)$$

kur a, b ir c yra teigiami skaičiai, taško $s \in \mathbb{R}^2$ koordinatės yra (x, y) ir t – laikas. Koeficientai a ir b nusako, ar erdvei būdinga geometrinė anizotropija, o koeficientas c nusako, ar būdinga anizotropija tarp erdvės ir laiko. Iš tikrųjų, atstumas erdvėje ir atstumas laike yra ne tas pats. Erdvei nėra būdingas "sutvarkymas", tačiau laikui yra būdinga tėkmė (praeitis-ateitis). Nėra akivaizdaus erdvės-laiko metrikos (atstumo) parinkimo būdo, nes, tarkime, nėra akivaizdaus tarpusavio ryšio tarp laiko ir erdvės „atstumo" matavimo vienetų. Metrika (2) yra sudaroma tariant, kad erdvė-laikas yra tiesiog aukštesnio matavimo Euklido erdvė.

Kitas būdas yra tam tikra prasme „atskirti" erdvės ir laiko priklausomybę.

Adityvus erdvės-laiko modelis:

$C_{DT}(h_s, h_t) = C_D(h_s) + C_T(h_t)$, kur $C_D(h_s)$, $h_s = s_1 - s_2$, $s_1, s_2 \in D$ ir $C_T(h_t)$, $h_t = |t_1 - t_2|$, $t_1, t_2 \in T$, yra kovariacinės funkcijos, apibrėžtos atitinkamai erdvėje ir laike.

Tačiau šis modelis esant tam tikram taškų išsibarstymui neatitiks kovariacinėms funkcijoms būdingų savybių. Panaši problema gali iškilti ir taikant erdvės-laiko semivariogramos adityvų modelį.

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

Dviejų kovariacinių funkcijų sandauga $C_{DT}(h_s, h_t) = C_D(h_s) \cdot C_T(h_t)$ yra erdvės-laiko modelis, atitinkantis kovariacinėms funkcijoms būdingas savybes (Dučinskas, Šaltytė-Benth 2003).

Apibrėžimas. Atsitiktinis procesas $Z(s;t)$ turi separabilią (atskiriamą) erdvės-laiko kovariacinę funkciją, jeigu visiems $s, u \in \mathbb{R}^d$ ir $t, r \in \mathbb{R}$, galima užrašyti:

$$cov\big(Z(s;t), Z(u;r)\big) = C^{(s)}(s,u) \cdot C^{(t)}(t,r), \quad (3)$$

kur $C^{(s)}$ ir $C^{(t)}$ yra atitinkamai erdvės ir laiko kovariacinės funkcijos (Genton 2007). Stacionariuoju atveju $C_{DT}(h_s, h_t) = C^{(s)}(h_s) \cdot C^{(t)}(h_t)$ (4)

Atsitiktinio proceso $Z(s;t)$ erdvės-laiko semivariograma γ apibrėžiama taip:

$$var\big(Z(s;t) - Z(u;r)\big) \equiv \gamma(s,u;t,r), \quad (5)$$

stacionarumo atveju $2\gamma(h;\tau)$; $h \in \mathbb{R}^d$, $\tau \in \mathbb{R}$.

Prognozei taške (s_0, t_0) naudojamos tiesinės prognozės (krigingo) metodai. Priklausomai nuo parametrinio apibrėžtumo laipsnio naudojamas paprastas, ordinarus arba universalus krigingas (Cressie 1993, Cressie & Wikle 2015). Universalus krigingas dar buvo nagrinėtas darbe Lesauskienė ir Dučinskas (2003).

3 Erdvės-laiko stebinių kontekstinis (su apmokymu) klasifikavimas

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

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

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

kur $\mu_l(s;t)$ – determinuotas erdvės-laiko trendas.

Pažymėjus $S_n = \{s_i \in D; i = 1, \dots, n\}$ vietų, kuriose paimti mokymo stebėjimai, galima pavadinti ją mokymo vietų aibe (*ang. set of training locations (STL)*). S_n yra padalintas į dviejų nesusikertančių sąjungų poaibius, t. y. $S_n = S^{(1)} \cup S^{(2)}$, kur $S^{(l)}$ yra S_n poaibis, kuriame ir paimti $Z(\cdot)$ stebėjimai iš Ω_l, $l = 1, 2$, $n = n_1 + n_2$. Nagrinėsime tiktai subalansuotus laike stebinius, t. y. $t = 1, \dots, \mathrm{T}$ visiems $s_i \in D$, $i = 0, \dots, n$. 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 $Z_i \sim N_T(m_i, \Sigma)$,

kur $m_i = \begin{cases} m_i^{(1)} = (\mu_1(s_i,1) \quad \dots \quad \mu_1(s_i,T))', & i = 1,\dots,n_1 \\ m_i^{(2)} = (\mu_2(s_i,1) \quad \dots \quad \mu_2(s_i,T))', & i = n_1+1,\dots,n \end{cases}$.

O kovariacinės $T \times T$ matricos Σ elementai apibrėžiami tokiu būdu $\sigma_{tr} = C^{(t)}(t,r)$, $t,r = 1,\dots,T$.

Bus nagrinėjamas tiesinis regresinis vidurkio modelis

$\mu_l(s;t) = \beta_l'(t)x(s)$,

kur $x(s) = \big(x_1(s),\dots,x_q(s)\big)'$ – regresorių vektorius, $\beta_l(t)$ – regresijos koeficiento vektorius.

Įveskime pažymėjimus $x(s) = x(s_i)$, $i = 0,\dots,n$ ir $B_l = (\beta_l(1),\dots,\beta_l(T))$, $B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix}$, $X_1 = \big(x_1,\dots,x_{n_1}\big)'$, $X_2 = \big(x_{n_1+1},\dots,x_{n_1}\big)'$ ir $X = X_1 \oplus X_2$.

Tada mokymo imties M modelis yra

$$M = XB + E,$$

kur E yra $n \times T$ atsitiktinių klaidų matrica, kuri turi matricinį normalųjį skirstinį

$$E \sim N_{n\times p}(0, R \otimes \Sigma).$$

Čia $R = \big(r_{ij}; i,j = 1,\dots,n\big)$ – erdvinės koreliacijos matrica tarp stebėjimų STL. AR(p) atveju, Σ yra $T \times T$ Tioplico (Toeplitz) matrica.

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

Tarkime kad r_0 – erdvinių koreliacijų vektorius tarp Z_0 ir stebinių aibėje S_n, t. y. $r_0 = (r_{01},\dots,r_{0n})$.

Tada populiacijoje Ω_l, sąlyginis Z_0 skirstinys, kai duota $M = m$ yra Gauso, t. y.

$(Z_0|M = m;\Omega_l) \sim N_T(\mu_{lm}^0, \Sigma_{0m})$,

kur sąlyginis vidurkis μ_{lm}^0 yra

$$\mu_{lm}^0 = E(Z_0|M = m;\Omega_l) = B_l'x_{i0} + \alpha_0'(m - XB)$$

ir sąlyginė kovariacija Σ_{0m} yra

$$\Sigma_{0m} = Var(Z_0|M = m;\Omega_l) = \rho\Sigma$$

ir $\alpha_0 = R^{-1}r_0$, $\rho = 1 - r_0'\alpha_0$.

Priminsime, kad populiacijoje Ω_l, marginalinis skirstinys yra taip pat Gauso,

$(Z_0;\Omega_l) \sim N_T(\mu_l^0, \Sigma_0)$,

kur mardinalinis vidurkis μ_l^0 yra

$$\mu_l^0 = E(Z_0;\Omega_l) = B_l'x_{i0}$$

ir mardinalinė kovariacija Σ_0 yra

$\Sigma_0 = Var(Z_0; \Omega_l) = \Sigma.$

Mahalanobio atstumo kvadratas tarp marginalinių skirstinių taške $s = s_0$ yra

$$\Delta^2 = (\mu_1^0 - \mu_2^0)'\Sigma^{-1}(\mu_1^0 - \mu_2^0),$$

kur $\mu_l = B_l' x_{0l}$, $l = 1, 2$.

o Mahalanobio atstumo kvadratas tarp sąlyginių skirstinių taške $s = s_0$ yra

$$\Delta_0^2 = (\mu_{1m}^0 - \mu_{2m}^0)'\Sigma_{0m}^{-1}(\mu_{1m}^0 - \mu_{2m}^0) = \Delta^2/\rho.$$

Tarkime, kad $H = (I_q, I_q)$ ir $G = (I_q, -I_q)$, kur I_q - vienetinė q-eilės matrica.

Tarkime, kad populiacijos apriorinės tikimybės yra žinomos $\pi_1(s)$ ir $\pi_2(s)$, $(\pi_1(s) + \pi_2(s) = 1)$. Tada sąlyginė Bajeso diskriminantinė funkcija (SBDF), minimizuojanti klaidingo klasifikavimo tikimybę, (Duda, Hart, Stork 2000) yra suformuota pagal sąlyginio tankio santykio logaritmą

$$W_m(Z_0) = ln\frac{\pi_1 P_1(Z_0|M)}{\pi_2 P_2(Z_0|M)} = (Z_0 - (m - XB)'\alpha_0 - B'H'x_0/2)'\,\Sigma^{-1}B'G'x_0/\rho + \gamma \quad (6)$$

kur $\gamma = ln(\pi_1(s_0)/\pi_2(s_0))$.

Apriorinės tikimybės, atsižvelgiančios į erdvinį kontekstą, dažnai skaičiuojamos pagal atvirkštinio atstumo (*ang. inverse distance*) metodą

$\pi_1(s_0) = \sum_{i=1}^{n_1} \frac{1}{d(s_0,s_i)} / \sum_{i=1}^{n} \frac{1}{d(s_0,s_i)}$ ir $\pi_2(s_0) = \sum_{j=n_1+1}^{n} \frac{1}{d(s_0,s_j)} / \sum_{i=1}^{n} \frac{1}{d(s_0,s_i)}$,

kur $d(\cdot,\cdot)$ - Euklido atstumo funkcija tarp taškų.

Kai $W_m(Z_0) > 0$, taškas s_0 priskiriamas 1 klasei, t. y. $S^{(1)}$ poaibiui, o kai $W_m(Z_0) < 0$, taškas s_0 priskiriamas 2 klasei, t. y. $S^{(2)}$ poaibiui.

Marginalinė Bajeso diskriminantinė funkcija (MBDF)

$$W(Z_0) = ln\frac{\pi_1 P_1(Z_0)}{\pi_2 P_2(Z_0)} = (Z_0 - B'H'x_0/2)'\Sigma^{-1}B'G'x_0 + \gamma. \quad (7)$$

Tada klasifikavimo klaidos tikimybė pagal SBDF lygi

$P_m = \sum_{l=1}^{2} \pi_l \Phi(Q_l)$,

kur $Q_l = -\Delta_0/2 + (-1)^l \gamma/\Delta_0$.

Tada klasifikavimo klaidos tikimybė pagal MBDF lygi

$P = \sum_{l=1}^{2} \pi_l \Phi(Q_l)$,

kur $Q_l = -\Delta/2 + (-1)^l \gamma/\Delta$.

Tačiau praktinėse situacijose dažnai populiacijų parametrai nežinomi. Nagrinėsime atvejį, kai B ir Σ nežinomi, o R žinomas. Naudosimės šių parametrų įvertiniais $\hat{B}$ ir $\hat{\Sigma}$ pagal mokymo imtį M. Pačios funkcijos yra vadinamos įterptosiomis

Bajeso diskriminantinėmis funkcijomis (*Bayesian dickriminant function (PBDF)*) (Dučinskas 2009). Toliau naudosimės pažymėjimais $\Psi = (\{B, \Sigma\})$ ir $\widehat{\Psi} = \left(\{\hat{B}, \widehat{\Sigma}\}\right)$.

Išstačius $\widehat{\Psi}$ vietoj Ψ į lygtį (6) gauname įterptinę (*ang. plug-in*) SBDF, kurią žymėsime PSBDF,

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

Tada aktualioji klasifikavimo klaida pagal PSBDF lygi

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

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

Išstačius $\widehat{\Psi}$ vietoj Ψ į lygtį (7) gauname įterptinę (*ang. plug-in*) MBDF, kurią žymėsime PMBDF

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

Tada aktualioji klasifikavimo klaida pagal PMBDF lygi

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

kur $\Delta\hat{B} = \hat{B} - B$ (Šaltytė-Benth, Dučinskas, 2005).

Išvestos klasifikavimo klaidų tikimybių ir aktualiųjų klasifikavimo klaidų formulės bus lyginamos ir skaitiškai iliustruojamos naudojant generuotus Gauso laukų duomenis. Bus naudojami statistinių programos R paketai geoR ir spatial.

4 Literatūra

1. Cressie N. (1993). Statistics for Spatial Data. Wiley & Sons, New York.

2. Cressie N., Wikle C. K. 2015, Statistics for spatio-temperal data. John Wiley & Sons.

3. Čekanavičius V., Murauskas G. 2009. Statistika ir jos taikymai III. Vilnius: TEV.

4. Dučinskas K. 2009. Approximation of the expected error rate in classification of the Gaussian random field observations. Statistics and Probability Letters, vol. 79, p. 138-144.

5. Dučinskas K., Šaltytė-Benth J. 2003. Erdvinė statistika. Klaipėda: Klaipėdos universiteto leidykla.

6. Duda Richard O., Hart Peter E., Stork David G. 2000. Pattern Classification.

7. Genton M. G. 2007. Separable approximations of space–time covariance matrices. Environmetrics, 18, pp. 681–695.

8. Lesauskiene E., Dučinskas K. 2003. Universal kriging for spatio-temporal data, Mathematical Modelling and Analysis, 8:4, 283-290.
9. Saltyte-Benth J., Ducinskas K. 2005. Linear Discriminant Analysis of Multivariate Spatial-Temporal Regressions. Scandinavian Journal of Statistics, 32, 281-294.