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

INFORMATIKA (N009)

KONTEKSTINIS DUOMENŲ, APRAŠOMŲ ERDVINIAIS APIBENDRINTAIS TIESINAIS MODELIAIS, KLASIFIKAVIMAS

Eglė Zikarienė

2019 m. spalis

Mokslinė ataskaita DMSTI-DS-N009-19-4

VU Duomenų mokslo ir skaitmeninių technologijų institutas, Akademijos g. 4,
Vilnius LT-08412
www.mii.lt

Santrauka

Šis darbas apima teorinę disertacijos dalį, kurioje pateikiama kontekstinio klasifikavimo klaidų tikimybių skaičiavimo ir vertinimo metodika. Siūlomi įvairūs aktualių klasifikavimo klaidų (ang. actual error rates) įvertiniai, kurių reikšmės panaudojamos nustatant klasifikatorių efektyvumą (ang. performance of classifiers) ir palyginant erdvinių duomenų modelius. Pradėti tirti duomenys apie Šakotųjų banguolių padengimą Baltijos jūroje, analizei taikomas erdvinės Beta regresijos modelis. Trumpai aprašomi parašyti straipsniai šia tema ir koks numatomas darbų praplėtimas disertacijos tema.

Reikšminiai žodžiai: Erdvinė analizė, Bajeso diskriminantinė funkcija, Tikroji klaidos tikimybė.

Turinys

1 Įvadas

Parengta erdvinių duomenų, aprašomų apibendrintais tiesiniais modeliais, kontekstinio klasifikavimo klaidų tikimybių skaičiavimo ir vertinimo metodika. Pagrindinis disertacijos darbo tikslas – atlikti erdvinių duomenų klasifikavimą taikant Bajeso diskriminantinės funkcijos procedūrą. Diskriminantinės funkcijos modeliams paprastai naudojamas tik Gauso skirstinys, todėl šioje disertacijoje planuojamas šių modelių išplėtimas dviem kryptimis: eksponentinių skirstinių šeimos nariais tiek tolydžiu, tiek diskrečiu atveju ir eliptinių skirstinių šeimos nariais.

Orientuojamasi į klasifikavimo metodus, pagrįstus sąlyginiais klasifikuojamo erdvės taško (ang. focal location) skirstiniais, priklausančiais eksponentinių skirstinių šeimai. Siūlomi įvairūs tikrųjų klasifikavimo klaidų (ang. actual error rates) įvertiniai, kurių reikšmės panaudojamos nustatant klasifikatorių efektyvumą (ang. performance of classifiers) ir palyginant erdvinių duomenų modelius. Siūlomi tikros klasifikavimo klaidos vertinimo metodai: R-metodas, aposteriorinės tikimybės metodas, U-metodas.

2 Klasifikavimo procedūra

Darbe nagrinėsime atsitiktinio lauko stebinius $\left\{Z(s): s \in D \subset R^p\right\}$. Stebėjimo $Z(s)$ modelis populiacijoje Ω_l yra

$$Z(s) = x'(s)\beta_l + \varepsilon(s),$$

čia $x(s)$ yra $q \times 1$ neatsitiktinių regresorių vektorius, ir β_l yra $q \times 1$ parametrų vektorius, $l = 1,...,L$. Klaidos narys atsitiktinis laukas $\left\{\varepsilon(s): s \in D\right\}$ generuojamas stacionarus nulinio vidurkio laukas su kovariacine funkcija, apibrėžiama pagal modelį visiems $s,u \in D$

$$\operatorname{cov}\left\{\varepsilon(s), \varepsilon(u)\right\} = \sigma^2 r(s-u),$$

čia σ^2 yra skalės parametras, o $r(\cdot)$ - erdvinės koreliacijos funkcija. (Dučinskas *et al.* 2015)

Tarkime, nagrinėjamas $Z_0 = Z(s_0)$ klasifikavimo uždavinys su mokymo imtimi T į vieną iš aukščiau aprašytų populiacijų.

Pristatant klasifikavimo procedūrą paprastumo dėlei pereisime prie modelio dviejų klasių atveju, tai yra $l = 1,2$. Tuomet Bajeso diskriminantinė funkcija (BDF), minimizuojanti klaidingo klasifikavimo tikimybę, yra suformuota pagal sąlyginio tankio santykio logaritmą (Fukunaga, 1990). Jos bendrąjį pavidalą galima užrašyti taip:

$$W\left(Z_0 \middle| T\right) = \ln\left(\frac{p\left(Z_0 \middle| T, \Omega_1\right)P(\Omega_1)}{p\left(Z_0 \middle| T, \Omega_2\right)P(\Omega_2)}\right) = \ln\left(\frac{p\left(Z_0 \middle| T, \Omega_1\right)}{p\left(Z_0 \middle| T, \Omega_2\right)}\right) + \ln\left(\frac{P(\Omega_1)}{P(\Omega_2)}\right)$$

čia Z_0 - klasifikuojamas stebinys, T - mokymo imtis, Ω_l - objekto klasė, $l = 1,2$ - klasės numeris, $p\left(Z_0 \middle| T, \Omega_l\right)$ - sąlyginis tankis, $P(\Omega_l)$ - apriorinė klasės tikimybė. Sudarant BDF naudojamos sąlyginio tankio funkcijos. Praktikoje mokymo imtis ir klasifikuojamas taškas yra tame pačiame erdviniame koreliuotame lauke, todėl į aptartas diskriminantines funkcijas įvedama klasifikuojamo taško erdvinė priklausomybė su mokymo imtimi (Dučinskas 2009), (Stabingienė *et al*. 2010).

Eksponentinių skirstinių šeimai bendrąją sąlyginio tankio išraišką galima užrašyti taip:

$$p\left(Z_0|T;\theta\right)=\exp\left\{A_0\left(T;\theta\right)T_0\left(Z_0\right)-B_0\left(T;\theta\right)+C_0\left(Z_0\right)\right\}$$

čia $T_0\left(Z_0\right)$ - minimali pakankama statistika, $C_0\left(\cdot\right)$ yra funkcija nuo Z_0, $A_0\left(\cdot\right)$ ir $B_0\left(\cdot\right)$ yra funkcijos nuo T ir nežinomų parametrų θ (Furukawa 2004). Tinkamas $A_0\left(\cdot\right)$ pasirinkimas apibrėžia priklausomybės tipą tarp kaimyninių reikšmių, o $B_0\left(\cdot\right)$ tuomet yra tinkama normalizavimo funkcija.

Šiame darbe nagrinėjant eksponentinių skirstinių šeimą laikomasi prielaidos, kad tarp erdvinių duomenų galioja porinis (*ang. pairwise*) ryšys, kai tankio funkciją galima užrašyti tokiu pavidalu:

$$p\left(z\right)=f\left(\alpha+\sum_{1\le i\le n}z_i\beta_i+\sum_{1\le i<j\le n}\sum z_iz_j\beta_{ij}\right)$$

$\alpha,\beta_i,\beta_{ij}$ - funkcijos koeficientai.

Tokiu atveju eksponentinių šeimos skirstinių modeliams, aprašomiems aukščiau pateikta sąlyginio tankio funkcija su galiojančia prielaida, įvedamas terminas *auto-modeliai* (Besag 1974). Gautos BDF išraiškos atsitiktinių laukų auto modeliams.

Auto-Gamma modelis

Įveskime pažymėjimus: $S^0=\left(s_0,S\right)$ - bendra klasifikuojamo taško ir mokymo imties vietų aibė, kai $S=\left(s_1,...,s_n\right)$ – mokymo vietų aibė. $T'_{0l}=\left(z_0^l,T'\right)$ – klasifikuojamo taško reikšmės ir mokymo imties aibė, čia $T'=\left(z_1,...,z_n\right)$ - mokymo imties aibė. Sąlyginį Gamma skirstinį galima užrašyti taip:

$$Z_0^l|T\sim\Gamma\left(\alpha^l,\gamma_0\left(T\right)\right)$$

kai α^l - mastelio parametras kiekvienai klasei *l*, $\gamma_0\left(T\right)$ - formos parametras, kurio išraiška $\gamma_0\left(T\right)=\theta_0+\sum_{t=1}^{n}\theta_{0t}z_t$ .Tuomet sąlyginį Gammą skirstinio tankį galima užrašyti taip:

$$p\left(Z_0^l=z_0|T\right)=\exp\left\{\begin{array}{l}-\frac{1}{\alpha^l}z_0+\left(\theta_0+\sum_{t=1}^{n}\theta_{0t}z_t\right)\ln z_0+\left(\theta_0+\sum_{t=1}^{n}\theta_{0t}z_t\right)\ln\left(\frac{1}{\alpha^l}\right)-\\-\ln\left(\Gamma\left(\theta_0+\sum_{t=1}^{n}\theta_{0t}z_t\right)\right)-\ln z_0\end{array}\right\}.$$

Sąlyginio Gamma skirstinio atveju gaunama BDF išraiška:

$$W\left(Z_0|T\right)=\ln\left(\frac{p_1\left(Z_0|T,\Omega_1\right)}{p_2\left(Z_0|T,\Omega_2\right)}\right)+\ln\left(\frac{P\left(\Omega_1\right)}{P\left(\Omega_2\right)}\right)=\left(\frac{\alpha^1-\alpha^2}{\alpha^1\alpha^2}\right)z_0+\left(\theta_0+\sum_{t=1}^{n}\theta_{0t}z_t\right)\ln\left(\frac{\alpha^2}{\alpha^1}\right)+\gamma$$

,

kai $\gamma=\ln\left(\frac{P\left(\Omega_1\right)}{P\left(\Omega_2\right)}\right)$.

Auto-Beta modelis

Sąlyginį Betą skirstinį galima užrašyti taip:

$$Z_0^l|T\sim Beta\left(p^l,q\right),$$

čia p,q- mastelio parametrai, p - skirtingi kiekvienai klasei l, $p^l(T)=\theta_0^l+\sum_{t=1}^{n}\theta_{0t}z_t$. Tuomet Beta skirstinio tankį galima užrašyti taip:

$$p\left(Z_0^l=z_0\left|T\right.\right)=\exp\left\{\begin{array}{l}\left(\left(\theta_0^l+\sum_{t=1}^{n}\theta_{0t}z_t\right)-1\right)\ln z_0+(q-1)\ln(1-z_0)+\\ +\ln\left(\Gamma\left(\theta_0^l+\sum_{t=1}^{n}\theta_{0t}z_t+q\right)\Big/\Gamma\left(\theta_0^l+\sum_{t=1}^{n}\theta_{0t}z_t\right)\Gamma(q)\right)\end{array}\right\}$$

Sąlyginio Beta skirstinio atveju gaunama BDF išraiška:

$$W\left(Z_0\left|T\right.\right)=\ln\left(\frac{p_1\left(Z_0\left|T,\Omega_1\right.\right)}{p_2\left(Z_0\left|T,\Omega_2\right.\right)}\right)+\ln\left(\frac{P(\Omega_1)}{P(\Omega_2)}\right)=\left(p^1-p^2\right)\ln z_0+\left(\frac{\Gamma\left(p^1+q\right)\Gamma\left(p^2\right)}{\Gamma\left(p^1\right)\Gamma\left(p^2+q\right)}\right)+\ln\left(\frac{P(\Omega_1)}{P(\Omega_2)}\right)=$$

$$=\left(\theta_0^1-\theta_0^2\right)\ln z_0+\ln\left(\frac{\Gamma\left(\theta_0^1+\sum_{t=1}^{n}\theta_{0t}z_t+q\right)\Gamma\left(\theta_0^2+\sum_{t=1}^{n}\theta_{0t}z_t\right)}{\Gamma\left(\theta_0^1+\sum_{t=1}^{n}\theta_{0t}z_t\right)\Gamma\left(\theta_0^2+\sum_{t=1}^{n}\theta_{0t}z_t+q\right)}\right)+\gamma$$

,

kai $\gamma=\ln\left(\frac{P(\Omega_1)}{P(\Omega_2)}\right)$.

Paprastai klasifikuojant objektus labai dažnai populiacijos parametrai yra nežinomi. Tokiu atveju yra naudojami jų įvertiniai, o pačios funkcijos yra vadinamos įterptosiomis Bajeso diskriminantinėmis funkcijomis (*ang. Plug-in Bayesian dickriminant function (PBDF)*) (Dučinskas 2009). Įterptoji Bajeso diskriminantinė funkcija, užrašoma taip:

$$W\left(Z_0\left|T,\hat{\theta}\right.\right)=\ln\left(\frac{p\left(Z_0\left|T,\Omega_1,\hat{\theta}_1\right.\right)P(\Omega_1)}{p\left(Z_0\left|T,\Omega_2\hat{\theta}_2\right.\right)P(\Omega_2)}\right)=\ln\left(\frac{p\left(Z_0\left|T,\Omega_1,\hat{\theta}_1\right.\right)}{p\left(Z_0\left|T,\Omega_2,\hat{\theta}_2\right.\right)}\right)+\ln\left(\frac{P(\Omega_1)}{P(\Omega_2)}\right),$$

čia $\hat{\theta}_l$ - nežinomų parametrų vertinimai klasėje l. Nežinomi parametrų įverčiai pasirenkami: ML, REML, Bajeso parametrų įvertinimai.

Klasifikuojant stebinius svarbu įvertinti, kaip tiksliai tai atliekama. Dažnai taikomos tokios charakteristikos, kaip bendras klasifikavimo tikslumas bei klaidingo klasifikavimo tikimybių įverčiai.

3 Klasifikavimo klaidos vertinimo kriterijai

Klasifikavimo kokybę nusako klaidingo klasifikavimo tikimybių įverčiai, kurie parodo, kokia yra tikimybė suklysti klasifikavimo metu kiekvienai iš klasių (Čekanavičius ir Murauskas 2008).

Tikroji (*angl. or actual error rate or conditional error rate*) – tai klasifikatoriaus, klaidingai klasifikuojančio atsitiktinai parinktą objektą, tikėtina tikimybė. Tai klaidos tikimybė su be galo didele bandymo aibe, sudaryta iš to paties skirstinio, kaip mokymo duomenų.

Tikroji klaidos tikimybė *(ang. actual error rate (AER)*) Bajeso diskriminantinei funkcijai $W\left(Z_0\left|T,\hat{\theta}\right.\right)$, yra apibrėžiama taip (Dučinskas 2009):

$$P\left(\hat{\theta}\right)=\sum_{l}\pi_{l}\hat{P}_{0l}$$

Čia, kai $l=1,2$,

$$\hat{P}_{0l}\left(t\right)=P_{0T}\left(\left(-1\right)^{l}W\left(Z_{0}\left|T,\hat{\theta}\right.\right)>0\left|\Omega_{l}\right.\right)$$

P_{0l} yra tikimybė stebinio reikšmę Z_0 priskirti ne tai klasei (kai diskriminantinės funkcijos $W\left(Z_{0}\left|T,\hat{\theta}\right.\right)$ reikšmė yra didesnė už 0, priskiriame stebinį Z_0 vienai klasei, kai $W\left(Z_{0}\left|T,\hat{\theta}\right.\right)$ - kitai klasei). Toliau pateikiami tikrųjų klasifikavimo klaidų (ang. actual error rates) vertinimo metodai, kurie bus naudojami nustatant klasifikatorių efektyvumą (ang. performance of classifiers).

R-metodas

R-metodas (ang. Resubstitution estimator), tai stebinių santykis mokymo imtyje iš klasės Ω_l, tai yra klaidingo klasifikavimo santykis naudojant pasirinktą klasifikavimo taisyklę. Taikant šį metodą mokymo imtis taip pat panaudojama įvertinti diskriminantinę funkciją. Tai yra, jei n_1 ir n_2 yra imtys atitinkamai iš klasės Ω_1 ir Ω_2, tuomet n_1 ir n_2 naudojamos apskaičiuoti diskriminantinę funkciją. Jei klaidingai suklasifikuotų imties taškų skaičius klasėse Ω_1 ir Ω_2 yra atitinkamai m_1 ir m_2, tai klaidų tikimybių P_1 ir P_2 įvertiniai atitinkamai yra $\frac{m_1}{n_1}$ ir $\frac{m_2}{n_2}$, tuomet R-metodo klaidos tikimybės įvertinimas apibrėžiamas taip (Ikechukwu 2016):

$$R=\frac{m_1+m_2}{n_1+n_2}.$$

U-metodas

Naudojant U-metodą arba „vieno neįtraukimo“ (ang. leave-one-out) vertinimo metodą visi išskyrus vieną stebinį yra naudojami apibrėžiant klasifikavimo taisyklę, o ši taisyklė yra naudojama klasifikuojant neįtrauktąjį stebinį. Ši procedūra kartojama kiekvienam stebiniui, todėl *N* dydžio imtyje kiekvienas stebinys yra klasifikuojamas naudojant funkciją su $N-1$ stebinių. Bendru atveju, galima sakyti, kad tai m – lypė kryžminė patikra, naudojama klaidos tikimybės įvertinimui, $R\left(cv\right)$, pagal (Lachenbruch, 1967), dviejų klasių atveju $R\left(cv\right)=\sum_{i=1}^{2}\sum_{j=1}^{2}n_{ij}/n_{i}$. Daugelio autorių tyrimai, įskaitant ir (Efron, 1983), parodė, kad *m*-lypė kryžminė patikra turi didelę dispersiją. Todėl, nors $R\left(cv\right)$ gali būti nepaslinktas įvertinimas, pasitikėjimas, su kuriuo galima tikėtis, kad $R\left(cv\right)$priartės prie $R\left(T\right)$, nėra didelis. Pagrindinis šio metodo privalumas yra tai, kad jis gauna nepaslinktą klasifikavimo vidutinės tikrosios klaidos tikimybę su mokymo imtimis n_1-1 ir n_2 iš klasių Ω_l. Tačiau negalima daryti prielaidos, kad U-metodas turi nedidelį poslinkį lyginant su sąlygine tikrosios klaidos tikimybe. Vienas šio įvertinimo trūkumų yra tai, kad jam reikia daugiau skaičiavimų, nei R-metodui. Kitas trūkumas yra didelė dispersija. Pagrindinis dėmesys lyginant šiuos metodus skiriamas klasifikavimo vidutinės tikrosios klaidos tikimybės vertinimo poslinkiui, bet dispersija taip pat labai svarbus faktorius.

Aposteriorinės tikimybės metodas

Šis vertinimo metodas aprašytas (Moore, 1973). Kai daroma prielaida, jog apriorinės klasių tikimybės yra lygios, jei parametrai θ yra žinomi ir turima diskriminantinė funkcija W, klaidingo klasifikavimo aposteriorinė tikimybė yra:

$$\left[\min\left\{f\left(z,\theta_1\right), f\left(z,\theta_2\right)\right\}\right] \Big/ \left[f\left(z,\theta_1\right) + f\left(z,\theta_2\right)\right],$$

kai θ yra nežinomi, klaidingo klasifikavimo aposteriorinės tikimybės, paremtos diskriminantine funkcija W išraiška, įvertinamos taip:

$$\left[\min\left\{f\left(z,\hat{\theta}_1\right), f\left(z,\hat{\theta}_2\right)\right\}\right] \Big/ \left[f\left(z,\hat{\theta}_1\right) + f\left(z,\hat{\theta}_2\right)\right].$$

4 Erdvinės beta regresijos modelių pritaikymas dumblių padengimo analizei

Pagrindinis dėmesys skiriamas apibendrintiems tiesiniams ir adityviems (GAM) modeliams, su tolydžiu priklausomu kintamuoju, pasiskirsčiusiu pagal Beta dėsnį. Pradėti tirti duomenys apie Šakotųjų banguolių padengimą Baltijos jūroje, analizei taikomas erdvinės Beta regresijos modelis. Sprendžiamas modelio, besisiskiriančio erdvinės informacijos įtraukimo lygmeniu, parinkimo ir parametrų vertinimo uždavinys. Atliekamas modelių palyginimas pagal AIC kriterijų. Modelių parametrų vertinimo algoritmai realizuoti statistinių skaičiavimų programinėje aplinkoje R. Rezultatai pristatyti konferencijose:

Zikarienė E., Dučinskas K. Application of spatial beta regression for modelling of the algae concentration index. Spatial statistics 2019, Sitges, Ispanija 2019 m. liepos 9 - 13 d.
Zikarienė E., Dučinskas K. Implementation of generalized additive models for spatial beta regression. Computer data analysis and modeling: stochastics and data science. Minskas, Baltarusija 2019 m. rugsėjo 18-22 d.

Taip pat parašyta mokslinė publikacija:
Zikarienė E., Dučinskas K. 2019. Implementation of generalized additive models for spatial beta regression. Proceedings of the XII International Conference. Computer data analysis and modeling: stochastics and data science. p. 341-343.

5 Išvados

Išvestos Bajeso diskriminantinės funkcijos išraiškos eksponentinių skirtinių šeimos nariams tolydžiu atveju: Gama ir Beta skirstiniams. Išanalizuoti tikslios klasifikavimo klaidos vertinimo kriterijai, kuriuos planuojama panaudoti nustatant klasifikatorių efektyvumą (ang. performance of classifiers) ir palyginant erdvinių duomenų modelius. Atlikta realių duomenų apie Šakotųjų banguolių padengimą Baltijos jūroje, analizė, kai taikomas erdvinės Beta regresijos modelis. Sudaryti modeliai su skirtingu erdvinės informacijos įtraukimu.

6 Literatūra

Besag J. 1974. Spatial Interaction and the Statistical Analysis of Lattice Systems. Journal of the Rotal Statistical Society, vol. 36, p. 192-236.

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

Efron B. 1983. Estimating the error rate of the error rate of a prediction rule: improvement on cross validation. Journal of the American Statistical Association, vol. 78, p. 316-331.

Fukunaga K. 1990. Introduction to Statistical Pattern recognition. Second edition. New York: Academic press.

Furman E. 2008. On a multivariate gamma distribution. Statistics & probability letters, vol. 78, p. 2353-2360.

Furukawa K. 2004. Development of Markov random field models based on exponential family conditional distributions. Retrospective Theses and Dissertations, vol. 939.

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.

Dučinskas K., Dreižienė L. 2011. Supervised classification of the scalar Gaussian random field observations under a deterministic spatial sampling design. Austrian Journal of Statistics, vol. 40 (1-2), p. 25-36.

Dučinskas K., Dreižienė L. Zikarienė E. 2015. Multiclass classification of the scalar Gaussian random field observation with known spatial correlation function. Statistics & probability letters, vol. 98, p. 107-114. ISSN: 0167-7152. Prieiga per internetą: <http://www.sciencedirect.com/science/article/pii/S0167715214004118#>.

Dučinskas K., Zikarienė E. 2015. Actual error rates in classification of the T-distributed random field observation based on plug-in linear discriminant function. Informatica, vol. 26, no. 4, p. 557-568. ISSN: 0868-4952. Prieiga per internetą: <http://www.mii.lt/Informatica/htm/INFO1061.htm>.

Dučinskas K., Zikarienė E., Dreižienė L. 2014. Comparison of Performances of Plug-in Spatial Classification Rules Based on Bayesian and ML Estimators. Proceedings of the 3rd International Conference on Pattern Recognition Applications and Methods, p. 161-166.

Ikechukwu E. 2016. Evaluation of Error Rate Estimators in Discriminant Analysis with Multivariate Binary Variables. American Journal of Theorical and Applied Statistics, vol. 5, p. 173-179.

Lachenbruch P. A. 1967. An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis. Biometrics, vol. 23, p. 639-645.

Moore D. H. 1973, Evoluation of five Discriminant procedures for binary variables. Journal of American Statistical Association, vol. 68, p. 399-404.

Stabingienė L., Stabingis G., Dučinskas K. 2010. Comparison of linear discriminant functions in image classification. Lietuvos matematikos rinkinys. LMD darbai, vol. 51, p. 227-231. ISSN 0132-2818.

Zikarienė E., Dučinskas K. 2019. Implementation of generalized additive models for spatial beta regression. Proceedings of the XII International Conference. p. 341-343.