LMD konf. darbai, 217–221
1997, Vilnius

Klasterių skaičiaus nustatymas panaudojant neparametrines statistikas

D. Šimoliūnas (VDU)

1. ĮVADAS

Šiame darbe nagrinėjamas Gauso skirstinių mišinio klasifikavimo uždavinys. Tarkime, turime q skirtingų klasifikavimo objektų klasių, ir kiekvieną iš jų atitinka atsitiktinis d-matis požymių vektorius Y_i su pasiskirstymo funkcija Φ_i, $i = 1, \dots, q$, kurios skirtingoms klasėms nesutampa. Stebimas atsitiktinis dydis (a.d.) $X = Y_\nu$, kur ν – diskretus a.d., įgyjantis reikšmes $1, \dots, q$ su apriorinėmis tikimybėmis $P\{\nu = i\} = p_i$, $\sum_{i=1}^q p_i = 1$. Tuomet a.d. X turės pasiskirstymo funkciją

$$F(x) = \sum_{i=1}^q p_i \Phi_i(x). \tag{1}$$

Gauso mišinių modelyje pasiskirstymo funkcijos $\Phi_i(x)$ yra Gauso pasiskirstymo funkcijos su vidurkių vektoriais M_i ir kovariacinėmis matricomis R_i. Pasižymėję φ_i atitinkamą pasiskirstymo tankį, turime

$$\varphi(x) = \varphi(x \mid M_i, R_i) = \frac{1}{(\sqrt{2\pi})^d (\det R_i)^{1/2}} \exp\left\{ -\frac{1}{2}(x - M_i)^T R_i^{-1} (x - M_i) \right\}.$$

Pagrindinis klasifikavimo uždavinio tikslas yra stebėjimo rezultatų $X_1, \dots, X_n$ pagrindu nustatyti objekto su požymių vektoriumi X priklausomybės i-jai klasei, $i = 1, \dots, d$, tikimybę. Pažymėję šias tikimybes $\pi(i, x)$ turime

$$\pi(i, x) = P\{\nu = i \mid X = x\} = \frac{p_i \varphi_i(x)}{f(x)}, \qquad i = 1, \dots, q, \tag{2}$$

kur $f(x)$ – a.d. X pasiskirstymo tankis. Taigi, šios tikimybės pilnai nusakomos x reikšme ir parametrų vektoriumi $\Theta = (p_i, M_i, R_i, i = 1, \dots, q)$.

Vienas iš būdų įvertinti šias tikimybes yra nežinomų Θ komponenčių įverčių radimas maksimalaus tikėtinumo metodu (MTM) ir taikymas lygybės (2). Praktikoje MTM įverčio radimui dažniausia taikomas EM algoritmas ar įvairios jo modifikacijos. Tačiau esant didelei duomenų dimensijai, šių įverčių radimas yra sudėtingas uždavinys. Be to, jei parametrų skaičius nedaug skiriasi nuo imties tūrio, tai įverčiai nebus pakankamai tikslūs. Antra, yra žinoma, kad EM algoritmu gaunami įverčiai, kai tikėtinumo funkcija sudėtinga, konverguoja į MTM įvertį, jei jo pradinis priartėjimas $\Theta^{(0)}$ parinktas netoli. Tačiau, kai d didelis, taip parinkti $\Theta^{(0)}$ praktikoje yra sudėtinga (plačiau žr. [1]).

Praktikoje dažnai pasitaiko atvejų, kai duomenų projektavimas į mažesnio matavimo poerdvį $H \subset R^d$ išsaugo visą statistinę informaciją apie duomenų klasterinę struktūrą. Paprastumo dėlei, priimsime prielaidą, kad $\mathbf{E}X = 0$ ir $\operatorname{cov}(X, X) = I_d$, kur I_d žymi vienetinę $d \times d$ matricą. Pažymėkime x_H vektoriaus x projekciją į H.

Apibrėžimas. Tiesinis poerdvis $H \subset R^d$, tenkinantis sąlygą

$$\pi(i, x) = P\{\nu = i \mid X = x\} = P\{\nu = i \mid X_H = x_H\}, \quad \forall x \in R^d \tag{3}$$

ir turintis mažiausią dimensiją iš tokių poerdvių klasės vadinamas diskriminantine erdve.

Tegul $k = \dim H$. Jei $k \ll d$, vertinamų parametrų skaičius gali žymiai sumažėti. Be to, jei visų klasių kovariacinės matricos lygios

$$R_1 = R_2 = \ldots = R_q, \tag{4}$$

tai $H = \operatorname{Span}\{M_1, \ldots, M_q\}$. Kadangi $p_1 M_1 + \ldots, p_q M_q = 0$, gauname nelygybę $k \leqslant q - 1$ ir daugeliu praktinių atvejų $k = q - 1$. Todėl šį faktą galima naudoti nežinomam klasių skaičiui q vertinti.

2. DISKRIMINANTINĖS ERDVĖS CHARAKTERIZACIJA

Nagrinėsime lygių kovariacinių matricų atvejį. Tegul H k-matė – diskriminantinė erdvė. Pažymėsime jos bazinius vektorius $u_1, \ldots, u_k \in R^d$, t.y. $H = \operatorname{Span}(u_1, \ldots, u_k)$. Tegul $H^\perp$ – tiesinis poerdvis, toks kad $H^\perp \oplus H = R^d$, t.y. $H^\perp$ yra ortogonalus H papildinys iki R^d. Pažymėkime $\xi(u) = u^T X$ – vienmatį a.d., o jo pasiskirstymo funkciją F_u. Tegul Φ_u žymi atitinkamą Gauso pasiskirstymo funkciją su vidurkiu 0 ir dispersija $\|u\|^2$. Kaip parodyta [2], jei galioja (4), tai $u \in H \Leftrightarrow F_u \neq \Phi_u$.

Praktikoje identifikuojant H, natūralu pradžioje rasti tas kryptis, kurias atitinkančios X projekcijos turi skirstinius, labiausiai nutolusius nuo normaliųjų. Tam įveskime atstumą ρ tarp dviejų vienamačių pasiskirstymo funkcijų F ir G, $\rho(F, G) > 0$, kai $F \neq G$ ir $\rho(F, F) = 0$. Papildomai reikalausime, kad ρ reikšmės nepriklausytų nuo F ir G poslinkio ir mastelio parametrų. Pvz.,

$$\rho(F, G) = \max_t |(F(t) - G(t))|.$$

Galima naudoti ir atstumą tarp atitinkamų tankio funkcijų, pvz., $\rho(f, g) = \|f - g\|_2^2 / \|f\|^2$.

Funkcionalą $Q(u) = \rho(F_u, \Phi_u)$ vadinsime projektavimo indeksu. Tuomet vektoriai $u_1, \ldots, u_k$, apibrėžti lygybėmis

$$\begin{aligned} u_1 &= \arg\max_{h \in R^d} Q(h) \\ u_i &= \arg \max_{h:(h, u_j)=0,\, j=1,\ldots,i-1} Q(h), \quad i = 2, \ldots, k \end{aligned} \tag{5}$$

sudarys diskriminantinės erdvės bazę. Be to, $\forall h \in H^{\perp}$: $Q(h) = 0$. Taigi, su projektavimo indeksų pagalba galima identifikuoti diskriminantinę erdvę (plačiau šis klausimas aptartas [2]).

3. DISKRIMINANTINĖS ERDVĖS STATISTINIS IDENTIFIKAVIMAS

Praktikoje diskriminantinę erdvę tenka įvertinti imties pagrindu. Tam apskaičiuojamas projektavimo indekso įvertis $\widehat{Q}(u)$. Kadangi nežinome tikrojo duomenų projekcijos skirstinio, tenka imti jo statistinį įvertį $\widehat{F}_u$, t.y. $\widehat{Q}(u) = \rho(\widehat{F}_u, \widehat{\Phi}_u)$ įvertina F_u skirstinio artumą Gauso skirstiniui.

Aišku, $\widehat{Q}(u) > 0$ bet kokiam $u \in H$. Tačiau ir kai $h \in H^{\perp}$, $\widehat{Q}(u)$ gali būti nelygus 0, bet artės į 0, kai $n \to \infty$. Todėl norint įvertinti diskriminantinę erdvę, pagal $\widehat{Q}(h)$ reikšmę tenka nuspręsti ar $Q(h) = 0$. Tam būtina žinoti $\max_{h \in H^{\perp}} \widehat{Q}(h)$ pasiskirstymo funkciją, kurią pažymėsime $G(y)$. $G(y)$ nepriklauso nuo parametrų Θ. Jei $v_1, \ldots, v_{d-k}$ yra ortonormuota bazė $H^{\perp}$, tai $\xi(v_1), \ldots, \xi(v_{d-n})$ yra nepriklausomi standartiniai Gauso dydžiai. Todėl $G(y)$ yra nusakoma pasirinktos atstumo funkcijos ρ, erdvės $H^{\perp}$ dimensijos $d - k$ ir imties tūrio n, t.y. konkrečiam atstumui ρ ir F_u įvertinimo būdui $\widehat{F}_u$ turime

$$G(y) = G_{n,d-k}(y).$$

Modeliavimo būdu galima kiek norima tiksliai įvertinti G. Kadangi tikrojo k nežinome, šį skirstinį tenka įvertinti įvairiems $k = 0, \ldots, d-1$. $\widehat{G}_{n,j}$ randame taip:
1) generuojame j-mačių Gauso vektorių n tūrio imtis;
2) kiekvienai iš jų skaičiuojame $\max_h \widehat{Q}(h)$;
3) pagal gautus rezultatus randame empirinę skirstinio funkciją $\widehat{G}_{n,j}$.

4. SPRENDIMŲ APIE DISKRIMINANTINĖS ERDVĖS DIMENSIJĄ PRIĖMIMAS

Norėdami statistiškai įvertinti diskriminantinę erdvę, turime įvertinti jos bazinius vektorius $u_1, \ldots, u_k$, maksimizuodami $\widehat{Q}(u)$. Kadangi tikrojo k nežinome, ieškome $\hat{u}_1, \ldots, \hat{u}_d$:

$$\begin{aligned} \hat{u}_1 &= \arg\max_h \widehat{Q}(h) \\ \hat{u}_i &= \arg \max_{h:(h,\hat{u}_j)=0,\, j=1,\ldots,i-1} \widehat{Q}(h), \quad i = 2, \ldots, d. \end{aligned} \tag{6}$$

Todėl $\widehat{Q}(\hat{u}_1) \geqslant \widehat{Q}(\hat{u}_2) \geqslant \ldots \geqslant \widehat{Q}(\hat{u}_d) \geqslant 0$.

Jei $\widehat{Q}(\hat{u}_{j+1}) \cong 0$ ir $\widehat{Q}(\hat{u}_j) > 0$, tai jį laikysime diskriminantinės erdvės dimensijos įverčiu $\hat{k}$. Aprašysime tai detaliau.

Šis uždavinys sprendžiamas, nuosekliai tikrinant hipotezes:

$$\begin{aligned} H_0 &: k = k_0 \\ H_1 &: k > k_0, k_0 = 0, \ldots, d-1. \end{aligned} \tag{7}$$

$\hat{k}$ laikysime mažiausią k_0, prie kurio nulinę hipotezę (7) priimame. Nulinę hipotezę priimame, jei prie užsiduoto reikšmingumo lygmens α gauname $\widehat{Q}(u_{k_0+1}) \leqslant \varepsilon$, kur ε apibrėžiamas lygybe $\widehat{G}_{n,d-k_0}(\varepsilon) = 1 - \alpha$. Čia $\widehat{G}_{n,j}$ yra modeliavimo būdu gautas empirinis $G_{n,j}$ analogas. Tuomet diskriminantinės erdvės įvertis:

$$\widetilde{H} = \mathrm{Span}(\hat{u}_1, \ldots, \hat{u}_k).$$

5. TYRIMŲ REZULTATAI

Praktinuiose tyrimuose buvo tiriamas šiame darbe aprašyto metodo pritaikymas klasterių skaičiaus nustatymui. Šis metodas buvo lyginamas su tradiciniu AIC (Akaikės informacinis kriterijus) kriterijumi. Projektavimo indeksas buvo apibrėžtas naudojant atstumą tarp neparametrinio tankio įverčio $\tilde{f}$, gauto k-kaimynų metodu, ir Gauso pasiskirstymo tankio metrikoje L_1:

$$Q(u) = \rho(f_u, \varphi_u) = \|f_u - \varphi_u\|_1,$$

$$\widehat{Q}(u) = \frac{2}{n}\sum_{t=1}^{n}\left[1 - \frac{\varphi_u(X_t)}{\tilde{f}_u(X_t)}\right]_+.$$

Plačiau buvo išnagrinėtas atvejis, kai klasių skaičius $q = 2$, t.y. $\dim H = 1$ ir $d = 2$ bei $d = 3$. Parametrus Θ buvo stengiamasi parinkti tokius, kad išryškėtų kurio nors metodo privalumai. Iš gautų rezultatų galima padaryti tokias išvadas.

1. Duomenų projektavimo metodo efektyvumas auga, kai $n \to \infty$. Buvo atliekami tyrimai su skirtingais imties tūriais $n = 200, 500, 1000$. Esant tam tikram Θ, prie $n = 200$, klasės nebuvo atskirtos, o prie $n = 1000$ įvertis buvo tikslus.
2. Duomenų projektavimo metodo įverčiai tikslesni, jei skirstinys nėra simetrinis, t.y. $p_1 \neq p_2 \neq 0{,}5$. Simetriniu atveju šio metodo pranašumai nežymūs.
3. Duomenų projektavimo metodas yra žymiai tikslesnis, kai $p_1 \ll p_2$. Daugeliu atvejų, kai AIC kriterijus nesugebėjo duoti tikslaus įverčio, metodas vis dar tiksliai vertino klasių skaičių.

Ligšioliniai tyrimai kol kas rodo, kad duomenų projektavimo metodas klasių skaičiui įvertinti gali būti efektyvus. Vienintelis trūkumas, kurį šiuo metu jau galima nurodyti, tai dideli skaičiavimo veiksmų kiekiai ir iš to sekantis gana ilgas skaičiavimo laikas.

Norėčiau padėkoti gerbiamam prof. R. Rudzkiui už visokeriopą pagalbą rašant šį straipsnį.

LITERATŪRA

[1] M. Radavičius, R. Rudzkis, Statisticai estimation of a mixture of Gaussian distrbutions, *Acta Applicandae Mathematicae*, **38** (1995), 37–54.

[2] Р. Рудзкис, М. Радавичюс, Целенаправленное проэцирование в моделях гауссовских распределений, сохраняющих информацию о кластерной структуре, *Liet. Mat. Rink.*, **37**(4) (1997).

Klasterių skaičiaus nustatymas panaudojant neparametrines statistikas

D. Šimoliūnas

A Gaussian mixture model is investigated in this paper. A method for estimation of unknown number of clusters in the case of equal covariance matrices is discribed. This method is based on the use of projection pursuit. The rezults of analysis of the method by simulation are shortly discussed.