KVAZIKANONIČKA DISKRIMINATIVNA ANALIZA U PROSTORU S UNIVERZALNOM METRIKOM

Zdenko Milonja, Vesna Dobrić i Konstantin Momirović
Sveučilište u Zagrebu

Predložena je nova metoda diskriminativne analize kojom se postiže zadovoljenje jednog od kriterija uvjetne optimalnosti. Taj se kriterij može definirati kao maksimalna separacija centroida grupa u prostoru koji je maksimalno reprezentativan za neki univerzum varijabli. Metoda se sastoji u maksimiziranju kovarijanci između linearnih kompozita varijabli reparametriziranih na Harrisovu univerzalnu metriku, i projekcija tih kompozita u prostor razapet vektorima indikatorske matrice koja definira pripadanje objekata prirodnim ili eksperimentalno formiranim subpopulacijama.

QUASICANONICAL DISCRIMINANT ANALYSIS IN UNIVERSAL METRIC SPACE

Zdenko Milonja, Vesna Dobrić and Konstantin Momirović
University of Zagreb

A new model for discriminant analysis is proposed so that a criterion of conditional optimality is satisfied. This criterion can be defined as maximal separation between group centroids in the space maximally representative for a universe of variates. The proposed method is defined as maximization of covariances between linear composites of variates reparametrized to Harris universal metrics, and projections of these composites into the space spanned by vectors of an indicator matrix, defining the position of each object in one and only one of natural or experimentally generated subpopulations.

1 UVOD

Kvazikanonička diskriminativna analiza u standardnoj metrici (Štalec i Momirović, 1984; Dobrić i Momirović, 1984; Dobrić, 1986) separira centroide na skupu ne nužno ortogonalnih latentnih dimenzija tako da maksimizira njihovu udaljenost bez obzira na intragrupnu konfiguraciju vektorâ varijabli. Zbog slabe osjetljivosti na objekte koji zauzimaju ekstremne pozicije, te neosjetljivosti na singularnost totalne i ma koje intragrupne matrice kovarijanci, ova je metoda vrlo pogodna za rutinske statističke analize i za grubu separaciju agregata formiranih pri različitim statističkim istraživanjima. Međutim, kvazikanoničke diskriminativne funkcije definirane su prije svega sposobnošću varijabli da separiraju grupe, a tek sekundarno (iako u osjetno većoj mjeri od standardnih kanoničkih diskriminativnih funkcija) njihovom reprezentativnošću za neki univerzum varijabli. Kako je reperezentativnost diskriminativnih funkcija, u mnogim slučajevima, gotovo jednako važna kao i njihova diskriminativna moć, predložena je jedna nova procedura kojom se formiraju kvazikanoničke diskriminativne varijable iz skupa varijabli koji je, reparametrizacijom na Harrisovu univerzalnu metriku, učinjen maksimalno reprezentativnim za univerzum varijabli iz koga je izvučen.

2 METODA

Neka je $E \subset P$ neki skup od n objekata definiran kao uzorak iz neke nehomogene populacije P, neka je $V \subset U$ neki skup linearno nezavisnih varijabli izabran iz nekog homogenog univerzuma varijabli U, i neka je G neka nominalna varijabla kojom je definirano g subpopulacija iz P.

Definirajmo

$$S = E \otimes G$$

kao indikatorsku matricu koja opisuje pripadanje objekata iz E subpopulacijama iz P, i

$$B = E \otimes V$$

kao matricu podataka koja opisuje objekte iz E na skupu varijabli V.

Neka je e vektor od n jedinica, $\mathbf{C} = \mathbf{e}(\mathbf{e}^T\mathbf{e})^{-1}\mathbf{e}^T$ centroidni projektor, i

$$\mathbf{V}^{-2} = diag\,((\mathbf{B}^T\mathbf{B} - \mathbf{B}^T\mathbf{C}\mathbf{B})\frac{1}{n})^{-1}.$$

Matricu

$$\mathbf{Z} = (\mathbf{I} - \mathbf{C})\mathbf{B}\mathbf{V}^{-1}$$

definirat ćemo kao matricu podataka u Harrisovoj univerzalnoj metrici sa matricom kovarijanci

$$\mathbf{R} = \mathbf{Z}^T\mathbf{Z}\frac{1}{n}.$$

Kvazikanonička diskriminativna analiza u prostoru s univerzalnom metrikom može se definirati kao rješenje problema

$$\mathbf{Z}\,\mathrm{x}_p = \mathrm{k}_p \qquad \alpha_p = \mathrm{k}_p^T\mathrm{l}_p\frac{1}{n} = max$$

$$\mathbf{S}(\mathbf{S}^T\mathbf{S})^{-1}\mathbf{S}^T\mathbf{Z}\,\mathrm{x}_p = \mathrm{l}_p \qquad \alpha_p \geq \alpha_{p+1}$$

$$p = 1, \ldots, q$$

$$q = min\,(m,\ g - 1)$$

$$\mathrm{x}_p^T\mathrm{x}_q = \delta_{pq}$$

odnosno kao ekstremizacija funkcija

$$\begin{aligned}\alpha_p = \mathrm{k}_p^T\mathrm{l}_p\frac{1}{n} = \mathrm{l}_p^T\mathrm{l}_p\frac{1}{n} &= \delta_p^2 \\ &= \mathrm{x}_p^T\mathbf{Z}^T\mathbf{S}(\mathbf{S}^T\mathbf{S})^{-1}\mathbf{S}^T\mathbf{Z}\,\mathrm{x}_p\frac{1}{n} \\ &= \mathrm{x}_p^T\mathbf{A}\,\mathrm{x}_p\end{aligned}$$

gdje je $\mathbf{A}$ matrica kovarijanci varijabli reskaliranih na Harrisovu metriku i njihovih projekcija u prostor koji razapinju vektori selektorske matrice, te istovremeno matrica intergrupnih kovarijanci Harrisovih varijabli.

Prema tome, $\alpha_p = \delta_p^2$, pa su korelacije između varijabli K_p i L_p

$$\eta_p = \delta_p / \sigma_p,$$

$$p = 1, \ldots, q$$

gdje su

$$\sigma_p^2 = \mathrm{x}_p^T \mathbf{R}\, \mathrm{x}_p,$$

$$p = 1, \ldots, q$$

varijance varijabli k_p. Vrijednosti η_p nazvat ćemo kvazikanoničkim koeficijentima diskriminacije u prostoru s univerzalnom metrikom, a varijable

$$\mathrm{y}_p = \mathrm{k}_p\, \delta_p^{-1},$$

$$p = 1, \ldots, q$$

kvazikanoničkim diskriminativnim varijablama u tom prostoru.

Definirajmo

$$\mathbf{U}^2 = diag\, \mathbf{R}$$

kao matricu varijanci Harrisovih varijabli. Standardizirana struktura varijabli $\mathbf{Y} = (\mathrm{y}_p)$ bit će definirana matricom

$$\mathbf{F} = \mathbf{U}^{-1}\, \mathbf{Z}^T\, \mathbf{Y}\, \frac{1}{n} = \mathbf{U}^{-1}\, \mathbf{R}\, \mathbf{X}\, \Delta^{-1}$$

gdje je $\mathbf{X} = (\mathrm{x}_p)$ a $\Delta^2 = (\delta_p^2)$, $p = 1, \ldots, q$.

Kako su korelacije varijabli iz $\mathbf{Y}$ elementi matrice

$$\mathbf{M} = \mathbf{Y}^T\, \mathbf{Y}\, \frac{1}{n} = \Delta^{-1}\, \mathbf{X}^T\, \mathbf{R}\, \mathbf{X}\, \Delta^{-1},$$

sklop ovako definiranih diskriminativnih faktora bit će definiran matricom

$$\mathbf{A} = \mathbf{F}\,\mathbf{M}^{-1} = \mathbf{U}^{-1}\,\mathbf{X}\,\Delta\ .$$

Uočimo da η_p, koeficijenti kvazikanoničke diskriminacije, nisu maksimizirani. Zbog toga su aproksimativni testovi značajnosti tih koeficijenata

$$f_p = \eta_p^2\,((n\ -\ 2)\,/\,(1\ -\ \eta_p^2))$$

koji, pod hipotezom $H_0 : \eta_p = 0 \Leftrightarrow \alpha_p = 0$ imaju Fisher-Snedeckorovu distribuciju sa 1 i $(n\ -\ 2)$ stupnjeva slobod

Neka je $\mathbf{L} = (\mathrm{l}_p) = \mathbf{S}\,(\mathbf{S}^T\,\mathbf{S})^{-1}\,\mathbf{S}^T\,\mathbf{Z}\,\mathbf{X}$. Očito,

$$\mathbf{L}^T\,\mathbf{L}\,\frac{1}{n} = \mathbf{X}^T\,\mathbf{A}\,\mathbf{X} = \Delta^2$$

pa su varijable l_p ortogonalne. Kako je i

$$\mathbf{K}^T\,\mathbf{L}\,\frac{1}{n} = \Delta^2\,,$$

varijable iz $\mathbf{L}$ i $\mathbf{K}$ tvore semibiortogonalni sustav, jer

$$\mathbf{K}^T\,\mathbf{K}\,\frac{1}{n} = \mathbf{X}^T\,\mathbf{R}\,\mathbf{X}$$

nije, u općem slučaju, dijagonalna matrica.

Međutim, kako su

$$\mathbf{H} = \mathbf{L}\,\Delta^{-1}$$

standardizirane glavne komponente matrice $\mathbf{S}\,(\mathbf{S}^T\,\mathbf{S})^{-1}\,\mathbf{S}^T\,\mathbf{Z}$. relacije između tih komponenata i standardiziranih kvazikanoničkih diskriminativnih varijabli $\mathbf{Y} = \mathbf{Z}\,\mathbf{X}\,\Sigma^{-1}$, gdje je $\Sigma^2 = diag\,(\mathbf{X}^T\,\mathbf{R}\,\mathbf{X})$, su ortogonalne za različito indeksirane varijable, jer je

$$\mathbf{Y}^T \mathbf{H} \frac{\mathbf{1}}{n} = \Sigma^{-1} \mathbf{X}^T \mathbf{A} \mathbf{X} \Delta^{-1} = \eta$$

dijagonalna matrica.

Bitno svojstvo kvazikanoničke diskriminativne analize u Harrisovom prostoru je da, osim što rezultat ne zavisi od metrike izvornih varijabli, varijable koje su bolji reprezentanti skupa V više sudjeluju u određivanju diskriminativnog prostora, jer je $\mathbf{U}^2$ matrica Guttmanovih procjena unikviteta varijabli iz V. Prema tome, kvazikanonička diskriminativna analiza u Harrisovom prostoru generira diskriminativne funkcije koje zadovoljavaju uvjetni optimum definiran maksimalnom separacijom centroida grupa u prostoru koji je maksimalno reprezentativan za univerzum varijabli U.

LITERATURA

[1] Dobrić, V. (1986). *On a class of robust method for multivariate data analysis.* COMPSTAT 1988, Proc. on Computational Statistics, Physica Verlag, Heidelberg, 211-216.

[2] Dobrić, V. and K. Momirović (1984). ***An algorithm and program for robust discriminant analysis.*** Proceedings of 8th Bosnian Symposium in Informatics "Jahorina 84", Sarajevo, 213:1-5.

[3] Štalec, J. and K. Momirović (1982). ***On a very simple model for robust discriminant analysis.*** Proc. 6th International Symposium "Computer at the University", 515:1-16.