NEKE MJERE UDALJENOSTI I SLIČNOSTI OBJEKATA OPISANIH NA SKUPU NOMINALNIH VARIJABLI U MAHALANOBISOVOM PROSTORU

KONSTANTIN MOMIROVIĆ
Sveučilište u Zagrebu

VALERIJ B. KUDRJAVCOV
Državni Univerzitet u Moskvi

Predložena je kvantifikacija jednog skupa nominalnih varijabli izvedena transformacijom binarno kodiranih rezultata u kategorijama u parcijalni Mahalanobisov oblik. Nad tako kvantificiranim varijablama predložena je jedna nova klasa mjera udaljenosti, definirana udaljenostima Minkowskog, i jedna nova klasa mjera sličnosti, definirana skalarnim produktima vektora entiteta.

KLJUČNE RIJEČI

prostor Mahalanobisa / prostori Minkowskog / klasifikacija / nenumeričke varijable / mjere udaljenosti / mjere sličnosti

SOME MEASURES OF DISTANCE AND SIMILARITY OF OBJECTS DESCRIBED ON THE SET OF CATEGORICAL VARIABLES IN MAHALANOBIS SPACE.

Quantification of a set of binary coded categorical variables defined by transformation to partial Mahalanobis space is proposed. On the set of so transformed variables a set of distance measures in Minkowski space is defined, as well as some measures of similarity, defined as scalar products of vectors of entities.

KEY WORDS

Mahalanobis space / Minkowski space / classification / categorical variables / distance measures / similarity measures

1. UVOD

Za objekte opisane na skupu nominalnih varijabli predložen je do sada niz mjera udaljenosti ili sličnosti. Medju njima se najviše upotrebljavaju, ili bar najčešće navode u literaturi o metodama klasifikacije[1)], indeksi Jackarda, Chekanowskog, Sokala i Michenera, Russela i Raoa, dva indeksa koje je predložio Kulchinski, zatim mjere Sokala i Sneatha, Ochaia, Rodgersa i Tanimota, Benzecrieovo rastojanje izmedju profila, te obična i ponderirana Euklidska udaljenost. Izvorno, sve su ove mjere definirane na standardno kodiranim nominalnim varijablama; lako je pokazati da su mnoge medju njima ekvivalentne ili bukvalno identične, i da se gotovo sve svode na normiranje broja karakteristika koje su zajedničke za oba objekta (Jambu, 1988).

Činjenica da je za slučaj kada su objekti opisani na nominalnim varijablama predloženo tako mnogo mjera prilično je jak argument da ni jedna od njih nije osobita. Prema tome, proširivanje skupa tih mjera ne može proizvesti veće štete. U ovom radu predložena je jedna nova klasa mjera udaljenosti, definirana udaljenostima Minkowskog u Mahalanobisovom prostoru; prije toga je, naravno, taj prostor definiran za nominalne varijable, reprezentirane konkatenacijom indikatorskih matrica. U tom su prostoru definirane i dve klase mjera sličnosti; prva na osnovu skalarnih produkata vektora objekata, a druga na osnovu mjera udaljenosti Minkowskog. Nadjeno je da transformacija nominalnih varijabli u Mahalanobisov oblik ima mnoga pogodna svojstva, koja se prenose i na mjere udaljenosti, i na mjere sličnosti; ono što je medju tim svojstvima najvažnije je da se udaljenost objekata opisanih na nominalnim varijablama može tretirati vrlo slično udaljenosti u prostoru kvantitativnih, eliptično ili čak normalno distribuiranih varijabli, što znači i da je moguće odrediti relativan udio svake varijable u mjerama udaljenosti ili sličnosti objekata.

[1)] *Vidi, na primjer, Anderberg (1973), Jambu (1988), Hartigan (1975), Jardine and Sibson (1971), Ferligoj (1989), Diday (1979) itd.*

2. PARCIJALNI MAHALANOBISOV PROSTOR

Neka je E neki skup od n objekata, i neka je V neki skup od m kvantitativnih, linearno nezavisnih varijabli. Neka je $\mathbf{E}$ sumacioni vektor reda n, tako da je $\mathbf{P} = \mathbf{E}(\mathbf{E}^t\mathbf{E})^{-1}\mathbf{E}^t$ centroidni projektor. Neka je

$$\mathbf{D} = E \times V$$

matrica podataka reda (n, m); centrirana matrica podataka biće sada

$$\mathbf{B} = (\mathbf{I}-\mathbf{P})\mathbf{D}$$

gdje je $\mathbf{I}$ matrica identiteta reda n. Neka je

$$\mathbf{B} = \mathbf{Y}\Lambda\mathbf{X}^t$$

bazična struktura matrice $\mathbf{B}$, sa lijevim vektorima $\mathbf{Y}$, $\mathbf{Y}^t\mathbf{Y} = \mathbf{I}$, desnim vektorima $\mathbf{X}$, $\mathbf{X}^t\mathbf{X} = \mathbf{X}\mathbf{X}^t = \mathbf{I}$ i dijagonalnom matricom singularnih vrijednosti Λ. Neka je

$$\mathbf{C} = \mathbf{B}^t\mathbf{B}$$

matrica disperzija varijabli iz V na skupu E. Tada ako $|\mathbf{C}| \neq 0$, puni Mahalanobisov prostor definiran je sa

$$\begin{aligned}\mathbf{M} &= \mathbf{B}\mathbf{C}^{-1/2} \\ &= \mathbf{Y}\mathbf{X}^t.\end{aligned}$$

Neka su Λ_q, $\mathbf{Y}_q$, $\mathbf{X}_q$ matrice prvih q, $q < m$ singularnih vrijednosti, lijevih i desnih svojstvenih vektora matrice $\mathbf{B}$. Parcijalni Mahalanobisov prostor ranga q biće

$$\mathbf{M}_q = \mathbf{Y}_q\mathbf{X}_q^t.$$

Puni Mahalanobisov prostor ima ova, dobro poznata svojstva:

(1) Metričku invarijantnost, jer je

$$\begin{aligned} \mathbf{M}^* &= \mathbf{B}^*\mathbf{C}^{*-1/2} \\ &= \mathbf{BW}(\mathbf{WCW})^{-1/2} \\ &= \mathbf{YX}^t \\ &= \mathbf{M} \end{aligned}$$

za svaku regularnu dijagonalnu matricu $\mathbf{W}$

(2) Ortonormalnost, jer je, očito,

$$\mathbf{M}^t\mathbf{M} = \mathbf{I}$$

(3) Optimalnu, pod kriterijem najmanjih kvadrata, aproksimaciju izvornih podataka, jer je

$$\text{trace } ((\mathbf{B} - \mathbf{M})^t(\mathbf{B} - \mathbf{M})) = \min$$

u odnosu na bilo koju drugu ortogonalizaciju vektora iz $\mathbf{B}$ (Mahalanobis, 1936; Wilks, 1962; Kendall and Stuart, 1973).

3. MAHALANOBISOV PROSTOR NOMINALNIH VARIJABLI

Neka je $E = \{e_i ; i = 1,...,n\}$ skup objekata, i neka su V_j; $j = 1,...,m$ nominalne varijable definirane skupovima kategorija $V_j = \{v_{jk} ; k = 1,...,q_j\}$ tako da je $v_{jk} \cap v_{jl} = 0 \mid k \neq l \; \forall \; V_j$. Neka su

$$\mathbf{S}_j = (s_{jik}) \qquad \begin{matrix} j = 1,...,m \\ i = 1,...,n \\ k = 1,...,q \end{matrix}$$

indikatorske matrice definirane funkcijom

$$\begin{cases} s_{jik} = 1 \; \forall \; e_i \in v_{jk} \\ s_{jik} = 0 \; \forall \; e_i \notin v_{jk} \end{cases}$$

i neka je

$$\mathbf{S} = (\mathbf{S}_1 ... \mathbf{S}_j ... \mathbf{S}_m)$$

matrica reda (n,q), $q = \sum_{j=1}^{m} q_j$, dobijena konkatenacijom matrica $\mathbf{S}_j$.

Ako je $n > q$, rang matrice $\mathbf{S}$ biće $r = q-m+1$ ako u skupu $V = \{v_{jk} ; \; j = 1,...,m; \; k = 1,...,q_j\}$ nema ni jednog para identičnih kategorija (Benzecrie, 1979; Momirović, 1988). Prema tome, bazična struktura ove matrice biće

$$\mathbf{S} = \mathbf{Y} \Lambda \mathbf{X}^t$$

gdje je $\Lambda = (\lambda_p)$; $p = 1,...,r$; $\lambda_p > 0 \; \forall \lambda_p$ dijagonalna matrica r singularnih vrijednosti matrice $\mathbf{S}$, $\mathbf{Y} = (\mathbf{y}_p)$; $p = 1,...,r$; $\mathbf{Y}^t\mathbf{Y} = \mathbf{I}$ matrica njenih lijevih svojstvenih vektora, a $\mathbf{X} = (\mathbf{x}_p)$; $p = 1,...,r$; $\mathbf{X}^t\mathbf{X} = \mathbf{I}$

matrica njenih desnih svojstvenih vektora pridruženih nenultim singularnim vrijednostima (Girko, 1988; Bertsekas, 1987). Očito, $\mathbf{Y}\mathbf{Y}^t = \mathbf{P}_Y$ je projektor reda n i ranga r, a $\mathbf{X}\mathbf{X}^t = \mathbf{P}_X$ je projektor reda q i ranga r.

Mahalanobisov prostor vektora matrice $\mathbf{S}$ definiran je vektorima matrice

$$\mathbf{M} = \mathbf{Y}\mathbf{X}^t$$

i ima ova očigledna svojstva:

(1) $$\mathbf{M}^t\mathbf{M} = \mathbf{X}\mathbf{X}^t = \mathbf{P}_X$$

(2) $$\mathbf{S}^t\mathbf{M} = \mathbf{X}\Lambda\mathbf{X}^t$$

(3) $$(\mathbf{S}-\mathbf{M})^t(\mathbf{S}-\mathbf{M}) = \mathbf{X}(\Lambda-\mathbf{I})^2\mathbf{X}^t$$

pri čemu je trace $(\mathbf{X}(\Lambda-\mathbf{I})^2\mathbf{X}^t) = \min$,

i

(4) $$f(\mathbf{m}_p) = N(\mu_p, \sigma^2_p) \qquad p = 1,\ldots,q$$

gdje su $\mathbf{m}_p$ vektori matrice $\mathbf{M}$, a N oznaka normalne raspodjele, jer je, ustvari, $\mathbf{M} = \mathbf{S}\mathbf{X}\Lambda^{-1}\mathbf{X}^t$, pa su vektori $\mathbf{m}_p$ linearna kombinacija identično distribuiranih vektora $\mathbf{s}_p$, $p = 1,\ldots,q$ matrice $\mathbf{S}$.

Prema tome, transformacija nominalnih varijabli reprezentiranih matricom $\mathbf{S}$ u Mahalanobisov oblik definiran matricom $\mathbf{M}$ generira kvantitativne, quaziortogonalne varijable koje su, pod kriterijem najmanjih kvadrata, najsličnije originalnim binarnim varijablama iz matrice $\mathbf{S}$.

4. MJERE UDALJENOSTI U PROSTORU MINKOWSKOG NOMINALNIH VARIJABLI TRANSFORMIRANIH U MAHALANOBISOV OBLIK

Neka je $M = (m_{ip})$; $i = 1,\ldots,n$; $p = 1,\ldots,q$ matrica podataka transformiranih u Mahalanobisov oblik. Mjera udaljenosti ma koja dva objekta (e_i, e_h) u prostoru Minkowskog reda l biti će

$$d_{ih}^{(l)} = \left(\sum_{p=1}^{q} (|m_{ip} - m_{hp}|)^l \right)^{1/l} \qquad l \neq 0$$

očito, za $l = 2$

$$d_{ih}^{(l)} = \left(\sum_{p=1}^{q} (m_{ip} - m_{hp})^2 \right)^{1/2}$$

ova će mjera, formalno definirana kao Euklidska udaljenost, biti, ustvari, Mahalanobisova udaljenost izmedju objekata e_i i e_h.

Od posebnog interesa mogu biti još Hemmingova udaljenost u Mahalanobisovom prostoru jer, za $l = 1$

$$d_{ih}^{(1)} = \sum_{p=1}^{q} |m_{ip} - m_{hp}|$$

i, za $l = \infty$

$$d_{ih}^{(\infty)} = \max_{p} |m_{ip} - m_{hp}|.$$

5. MJERE SLIČNOSTI OBJEKATA OPISANIH NAD SKUPOM NOMINALNIH VARIJABLI TRANSFORMIRANIH U MAHALANOBISOV OBLIK

Mjere sličnosti objekata iz E najjednostavnije je definirati kao skalarne produkte njihovih vektora u Mahalanobisovom prostoru, dakle kao elemente matrice

$$\mathbf{M}\mathbf{M}^t = \mathbf{P}_Y .$$

Medjutim, ako ove mjere treba, kao što je običaj, omedjiti sa 1 , te se mjere mogu definirati kao elementi matrice

$$\mathbf{R} = \mathbf{V}^{-1}\mathbf{P}_Y\mathbf{V}^{-1}$$

gdje je $\mathbf{V}^2 = \text{diag } \mathbf{P}_Y$.

Mjere sličnosti se, naravno, mogu izvesti i iz mjera udaljenosti. Neka je

$$\mathbf{D}^{(l)} = (d_{ih}^{(l)}) \qquad i,h = 1,\ldots,n$$

matrica mjera udaljenosti dobijenih, nakon transformacije u Mahalanobisov oblik, u nekoj l metrici Minkowskog. Tada će mjere sličnosti biti elementi matrice

$$\mathbf{Q}^{(l)} = (q_{ih}^{(l)}) \qquad i,h = 1,\ldots,n$$

gdje su elementi

$$q_{ih}^{(l)} = 1.0 - d_{ih}^{(l)}/\hat{d}^{(l)} \qquad i,h = 1,\ldots,n$$

a $\hat{d}^{(l)} = \max\limits_{i,h} (d_{ih}^{(l)})$.

LITERATURA

1. Anderberg, M.R.:
   Cluster analysis for application.
   Academic Press, New York, 1973.

2. Benzecrie, J.P.:
   L' analyse des données. 1. La taxinomie.
   Dunod, Paris, 1979.

3. Bertsekas, D.:
   Uslovnaya optimizaciya i metody mnozhiteley Lagrangea.
   Radio i svyaz, Moskva, 1987.

4. Diday, E.:
   Optimisation et classification automatique.
   INRIA, Paris, 1979.

5. Ferligoj, A.:
   Razvrščanje v skupine.
   Metodološki zveski, 4, JUS, Ljubljana, 1989.

6. Girko, V.L.:
   Spektral'naya teoriya sluchaynyh matric.
   Nauka, Moskva, 1988.

7. Hartigan, J.A.:
   Clustering algorithms.
   Wiley, New York, 1975.

8. Jambu, M.:
   Ierarhicheskiy klaster-analiz i sootvetstviya. (perev. B.G. Mirkina).
   Finansy i statistika, Moskva, 1988.

9. Jardine, N. and R. Sibson:
   Mathematical taxonomy.
   Wiley, New York, 1971.

10. Kendall, M.G. and A. Stuart:
    The advanced theory of statistics.
    Hafner, New York, 1973.

11. Mahalanobis, P.C.:
    On the generalized distance in statistics.
    Proceeding of National Institute of Science of India, 2(1936), 49-55.

12. Momirović, K.:
    Uvod u analizu nominalnih varijabli.
    Metodološke sveske, 3, JUS, Ljubljana, 1988.

13. Wilks, S.:
    Mathematical statistics.
    Wiley, New York, 1962.