Janez Jug

RAZVRSCANJE BINARNIH ENOT S PODPROGRAMOM CLUSTER V SPSS/PC+

Podprogram CLUSTER v SPSS/PC+ praviloma omogoča hierarhično združevanje realnih enot. Nekatere mere podobnosti in različnosti, ki so definirane za realne enote, imajo binarne ekvivalente. Neposredno lahko izračunamo binarno evklidsko razdaljo, kvadrat binarne evklidske razdalje in Ochiaijevo mero podobnosti, z nekaj programiranja pa lahko s kvadratom binarne evklidske razdalje in številom spremenljivk izračunamo pet mer podobnosti in eno mero različnosti za binarne enote.

CLUSTERING OF BINARY CASES WITH PROCEDURE CLUSTER FROM SPSS/PC+ The procedure CLUSTER enables the agglomerative hierarchical clustering of numerical data. Because some similarity measures for numerical data have binary equivalents, it is also possible to cluster binary ones. We can directly compute binary Euclidean distance, binary squared Euclidean distance and Ochiai similarity measure and on the basis of the programming five similarity and one dissimilarity measures.

GRUPIGADO DE BINARAJ UNUOJ PER PROCEDO CLUSTER EL SPSS/PC+ - La procedo CLUSTER ebligas la hierarkian kunigan grupigadon de realaj unuoj (objektoj). Car kelkaj mezuroj de proksimeco por realaj unuoj havas binarajn ekvivalentojn, estas eble grupigi ankaŭ binarajn unuojn. Rekte oni povas kalkuli binaran Euklidan distancon, kvadrigitan binaran Euklidan distancon kaj similecmezuron de Ochiai kaj per iom da programado ankoraŭ kvin mezurojn de malsimileco kaj unu mezuron de simileco.

1. Uvod

Podprogram CLUSTER v programskem paketu SPSS/PC+ je namenjen hierarhičnemu razvrščanju enot, opisanih s številskimi spremenljivkami, vendar pa z njim lahko razvrščamo tudi enote, opisane z binarnimi (dihotomnimi) spremenljivkami. Za številske enote podprogram CLUSTER omogoča izračun kvadrata evklidske razdalje, evklidske razdalje, razdalje Manhattan, razdalje Čebiševa,

kosinusa kota med vektorjema enot in razdalj v absolutni potenčni metriki[1]. Nekatere od teh mer različnosti in podobnost imajo binarne ekvivalente, zato jih lahko uporabimo za računanje različnosti ali podobnosti med enotami, opisanimi z binarnimi spremenljivkami. Poleg tega pa lahko izračunamo Sokal-Michenerjevo in še nekatere druge mere podobnosti.

2. Binarni ekvivalenti mer različnosti in podobnosti v podprogramu CLUSTER

Mere podobnosti in različnosti za binarne enote so določene s frekvencami v asociacijski tabeli (kontingenčni tabeli 2 x 2) za enoti, med katerima merimo podobnost ali različnost. Asociacijsko tabelo za enoti, opisani s spremenljivkami, ki imajo vrednosti 1 (prisotnost značilnosti) ali 0 (odsotnost značilnosti), prikazuje tabela 1. Frekvenca *a* pove, na koliko spremenljivkah imata enoti hkrati pozitiven odgovor, frekvenca *d* pa, na koliko spremenljivkah imata enoti hkrati negativen odgovor. Frekvenci *b* in *c* štejeta, na koliko spremenljivkah se enoti ne ujemata. Vsota vseh štirih frekvenc (a + b + c + d) je enaka številu spremenljivk *m*.

Tabela 1: Asociacijska tabela za enoti X in Y.

		Enota Y	
		1	0
Enota X	1	a	b
	0	c	d

---

1. Razdalja med enotama je r-ti koren vsote absolutnih razlik med vrednostima na vsaki spremenljivki na p-to potenco.

$$d_{p,r}(X,Y) = \left(\sum_{k=1}^{m} |x_i - y_i|^p \right)^{1/r}$$

Z ustrezno izbiro parametrov p in r dobimo evklidsko razdaljo, kvadrat evklidske razdalje, razdalje Minkowskega, razdaljo Manhattan, minimum, maksimum in več drugih razdalj (Norušis, 1986, C-3).

Kadar imamo namesto številskih spremenljivk binarne spremenljivke, ki imajo vrednosti 0 in 1, sta v enačbi za razdaljo Manhattan

$$d_1(X,Y) = \sum_{i=1}^{m} |x_i - y_i|$$

in v enačbi za kvadrat evklidske razdalje

$$d_2^2(X,Y) = \sum_{i=1}^{m} (x_i - y_i)^2$$

količini

$$|x_i - y_i| \quad \text{in} \quad (x_i - y_i)^2$$

nič, če se enoti X in Y ujemata na i-ti spremenljivki, in ena, če se ne ujemata, zato sta razdalja Manhattan in kvadrat evklidske razdalje med enotama X in Y enaki vsoti frekvenc *b* in *c*, ki pove, na koliko spremenljivkah se enoti X in Y ne ujemata (Anderberg, 1973, 113).

Če imajo binarne spremenljivke vrednosti 0 in 1, so v enačbi za izračun kosinusa kota med vektorjema

$$s_{cos}(X,Y) = (\sum_{i=1}^{m} X_i Y_i)/((\sum_{i=1}^{m} X_i^2)(\sum_{i=1}^{m} Y_i^2))^{1/2}$$

količine

$$\sum_{i=1}^{m} X_i Y_i = a \qquad \sum_{i=1}^{m} X_i^2 = a + b \qquad \sum_{i=1}^{m} Y_i^2 = a + c$$

zato je kosinus kota med binarnima vektorjema enot

$$s_{cos}(X,Y) = \frac{a}{((a+b)(a+c))^{1/2}} = \left(\frac{a}{a+b} \cdot \frac{a}{a+c}\right)^{1/2}$$

oziroma Ochiaijeva mera podobnosti (Anderberg, 1973, 72).

Od velikega števila znanih mer podobnosti in različnosti za binarne enote[2] lahko v podprogramu CLUSTER zahtevamo izračun

---

2. Mere podobnosti in različnosti za binarne enote in spremenljivke so opisane v: Anderberg (1973, pogl. 4 in 5), Baroni-Urbani in Buser (1976), Cheetham in Hazel (1969), Clifford in Stephenson (1975, 49-82), Doran in Hodson (1975, pogl. 6), Everitt, 1974, 49-59), Ferligoj (1989, 37-42), Gordon (1981, 13-32), Legendre in Legendre (1983, pogl. 6), Lorr (1983, 22-44), Romesburg (1984, 141-158), Sneath in Sokal (1973, 129-134), Spath (1980, 24-30), Steinhausen in Langer (1977, 53-56).

binarne evklidske razdalje, kvadrata binarne evklidske razdalje in Ochiaijeve mere podobnosti[3].

3. Računanje drugih mer podobnosti za binarne enote

Za razvrščanje binarnih enot navadno uporabljamo Sokal-Michenerjevo ali Jaccardovo mero podobnosti. Ker so tudi binarni ekvivalenti mer različnosti in podobnosti v podprogramu CLUSTER definirani s frekvencami iz asociacijske tabele, lahko z njimi izračunamo nekatere mere podobnosti in različnosti za binarne enote.

Sokal-Michenerjevo mero podobnosti $s_{SM}(X,Y)$ lahko izračunamo, če poznamo vsoto neujemanj $b + c$, tj. kvadrat binarne evklidske razdalje, in število spremenljivk m. To gre takole:

$$s_{SM}(X,Y) = \frac{a + d}{a + b + c + d} = \frac{m - d_2^2(X,Y)}{m} = 1 - \frac{d_2^2(X,Y)}{m}$$

Na ta način lahko iz kvadrata binarne evklidske razdalje (oz. binarne razdalje Manhattan) in števila spremenljivk izračunamo pet mer podobnosti in eno mero različnosti, ki so v tabeli 2.

Matriko mer podobnosti ali različnosti izračunamo tako, da najprej s podprogramom CLUSTER izračunamo matriko kvadratov evklidskih razdalj med binarnimi enotami in jo izpišemo v posebno datoteko. Nato privzamemo, da je to matrika z m spremenljivkami v stolpcih, ki opisujejo m enot. Iz vrednosti vsake spremenljivke izračunamo novo spremenljivko, katere vrednosti sestavljajo en stolpec matrike izbrane mere podobnosti ali različnosti med binarnimi enotami. Novo spremenljivko izračunamo z ukazom COMPUTE po obrazcu iz tabele 2. Ker imamo m spremenljivk, uporabimo m stavkov COMPUTE, da iz spremenljivk S1 do Sm izračunamo nove spremenljivke D1 do Dm. Primer programa je v prilogi.

Podprogram CLUSTER lahko razvršča na podlagi matrike različnosti in matrike podobnosti. Matriko podobnosti transformira v različnost samo pri Wardovi metodi, pri vseh drugih metodah pa razvršča na podlagi matrike podobnosti[4]. Če hočemo binarne enote hierarhično razvrščati v skupine na podlagi matrike različnosti,

---

3. S podprogramom CORRELATION pa lahko izračunamo binarno obliko Pearsonovega korelacijskega koeficienta. ϕ je mera podobnosti med spremenljivkami in ima vrednosti na intervalu [0,1].

$$\phi(X,Y) = \frac{ad - bc}{((a + b)(a + c)(b + d)(c + d))^{1/2}}$$

4. Kadar uporabimo metode CENTROID, MEDIAN in WARD, opozori, da je pri teh metodah združevanja priporočljiva uporaba kvadrata evklidske razdalje.

Tabela 2: Mere podobnosti in različnosti, ki jih izračunamo s kvadratom evklidske razdalje in številom spremenljivk

Mera podobnosti		Izračun
Sokal-Michener	$\frac{a+d}{a+b+c+d}$	$s = 1 - \frac{d_2^2(X,Y)}{m}$
Sokal in Sneath 1	$\frac{2(a+d)}{2(a+d)+b+c}$	$s = \frac{2(m - d_2^2(X,Y))}{2m - d_2^2(X,Y)}$
Rogers in Tanimoto	$\frac{a+d}{a+d+2(b+c)}$	$s = \frac{m - d_2^2(X,Y)}{m + d_2^2(X,Y)}$
Sokal in Sneath 3	$\frac{a+d}{b+c}$	$s = \frac{m}{d_2^2(X,Y)} - 1$
Hamann	$\frac{(a+d)-(b+c)}{a+b+c+d}$	$s = 1 - \frac{2d_2^2(X,Y)}{m}$
Mera različnosti		
Variančna različnost[5]	$\frac{b+c}{4(a+b+c+d)}$	$d = \frac{d_2^2(X,Y)}{4m}$

izračunamo namesto mer podobnosti iz tabele 2 transformacije teh mer v različnosti z vrednostmi na intervalu od 0 do 1.

Sokal-Michenerjevo mero podobnosti transformiramo v različnost s transformacijo $d = 1 - s$, to pa izrazimo s kvadratom binarne evklidske razdalje ter številom spremenljivk takole:

$$d_{SM}(X,Y) = 1 - s_{SM}(X,Y) = 1 - \frac{a+d}{a+b+c+d} =$$

$$= \frac{(a+b+c+d)-(a+d)}{a+b+c+d} = \frac{b+c}{a+b+c+d} = \frac{d_2^2(X,Y)}{m}$$

5. Ta mera različnosti je v podprogramu PROXIMITIES v SPSS-X (SPSS, 1988, 832).

Podobno izpeljemo tudi transformacije drugih mer podobnosti iz tabele 2 v mere različnosti. Vrsto transformacije in način izračuna prikazuje tabela 3.

Tabela 3: Transformacije mer podobnosti v različnosti in izračun

Ime	Transformacija	Izračun različnosti
Sokal-Michener	$d = 1 - s$	$d = \frac{d_2^2(X,Y)}{m}$
Sokal in Sneath 1	$d = 1 - s$	$d = \frac{d_2^2(X,Y)}{2m - d_2^2(X,Y)}$
Rogers in Tanimoto	$d = 1 - s$	$d = \frac{2d_2^2(X,Y)}{m + d_2^2(X,Y)}$
Hamann	$d = 1 - \lvert s \rvert$	$d = 1 - \left\lvert 1 - \frac{2d_2^2(X,Y)}{m} \right\rvert$

S transformacijo dveh mer podobnosti v različnosti dobimo komplement Sokal-Michenerjeve mere podobnosti. Če mero podobnosti Sokala in Sneatha 3, ki ima vrednosti na intervalu I0, ∞I, transformiramo v različnost s transformacijo d = 1/(1 + s), dobimo komplement Sokal-Michenerjeve mere podobnosti. Tudi transformacija Hamannove mere podobnosti, ki je definirana na območju I-1,1I, v različnost tako, da bo različnost najmanjša pri meri podobnosti +1 in največja pri meri podobnosti -1 (transformacija d = (1 - s)/2), dobimo ta komplement.

Jaccardove in njej enakovrednih mer podobnosti, ki v števcu ne upoštevajo ujemanja v negativnih vrednostih, na ta način ne moremo izračunati, zato moramo za razvrščanje binarnih enot na podlagi teh mer podobnosti uporabiti SPSS-X ali kak drug program za razvrščanje.

LITERATURA

Anderberg, M. R. (1973): *Cluster Analysis for Applications*. New York: Academic Press.

Baroni-Urbani, C., Buser, M. W. (1976): "Similarity of Binary Data". *Systematic Zoology*. 25:251-259.

Cheetham, A. H., Hazel, J. E. (1969): "Binary (presence-absence) Similarity Coefficients". *Journal of Paleontology*. 43:1130-1136.

Clifford, H. T., Stephenson W. (1975): *An Introduction to Numerical Classification*. New York: Academic Press.

Doran, J. E., Hodson, F. R. (1975): *Mathematics and Computers in Archaeology*. Cambridge, Mass.: Harvard University Press.

Everitt, B. (1974): *Cluster Analysis*. London: Heinemann Educational Books.

Ferligoj, A. (1989): *Razvrščanje v skupine*. Teorija in uporaba v družboslovju. Ljubljana: Raziskovalni institut FSPN.

Gordon, A. D. (1981): *Classification*. London: Chapman and Hall.

Legendre, L., Legendre, P. (1983): *Numerical Ecology*. Amsterdam: Elsevier Scientific Publishing Company.

Lorr, M. (1983): *Cluster Analysis for Social Scientists*. San Francisco: Jossey-Bass.

Norušis, M. J. (1986): *Advanced Statistics SPSS/PC+™ For the IBM PC/XT/AT*. Chicago: SPSS Inc.

Romesburg, H. C. (1984): *Cluster Analysis for Researchers*. Belmont, California: Lifetime Learning Publications.

Sneath, P. H. A., Sokal, R. R. (1973): *Numerical Taxonomy*. The Principles and Practice of Numerical Classification. San Francisco: W. H. Freeman and Company.

Spath, H. (1980): *Cluster Analysis Algorithms for Data Reduction and Classification of Objects*. Chichester: Ellis Horwood Ltd.

SPSS Inc. (1986): *SPSS^{x} User's Guide*. 2nd Edition. New York: McGraw-Hill.

SPSS Inc. (1988): *SPSS-X™ User's Guide*. 3rd Edition. Chicago: SPSS Inc.

Steinhausen, D., Langer, K. (1977): *Clusteranalyse*. Einfuhrung in Methoden und Verfahren der automatischen Klassifikation. Berlin, New York: Walter de Gruyter.

Priloga: Program za razvrščanje binarnih enot s programom SPSS/PC+

```
SET LISTING='bled1.lis'/RESULTS='razdalj.dat'.

* Vhodni podatki
DATA LIST
  /ID 2 S1 to S17 4-20.
BEGIN DATA.
 1 01011001100000100
 2 10100110000000000
 3 01111101011000000
 4 10100000000000000
 5 10011100000010100
 6 11111011011101011
 7 11110111110000110
 8 11111001000000101
 9 11001001001110001
END DATA.

* Računanje matrike kvadratov binarne evklidske razdalje.
CLUSTER S1 TO S17
  /WRITE
  /PRINT = NONE
  /PLOT = NONE.

* Računanje matrike komplementov Sokal-Michenerjevega koeficienta.
DATA LIST FILE='razdalj.dat' FREE / S1 TO S9.
COMPUTE D1 = S1/17.
COMPUTE D2 = S2/17.
COMPUTE D3 = S3/17.
COMPUTE D4 = S4/17.
COMPUTE D5 = S5/17.
COMPUTE D6 = S6/17.
COMPUTE D7 = S7/17.
COMPUTE D8 = S8/17.
COMPUTE D9 = S9/17.
FORMATS D1 TO D9 (F8.6).
SET RES='sokal.mat'.
WRITE VAR= D1 TO D9.

* Razvrščanje.
DATA LIST MATRIX FREE FILE='sokal.mat'
  /PROBLEMI POGLEDI KRT TELEKS DELO NOVAREV MLADINA GLEDISTA BORBA.
CLUSTER PROBLEMI TO BORBA
  /READ
  /METHOD = SINGLE COMPLETE WARD
  /PRINT = CLUSTER(2,6) SCHEDULE DISTANCE
  /PLOT = DENDROGRAM.
FINISH.
```