PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 47 2009
TAKSONOMIA 16
Klasyfikacja i analiza danych – teoria i zastosowania

Andrzej Dudek
Uniwersytet Ekonomiczny we Wrocławiu

KONSTRUKCJA MACIERZY BURTA DLA OBIEKTÓW SYMBOLICZNYCH

1. Wstęp

Jednym z najważniejszych narzędzi wielowymiarowej analizy korespondencji jest macierz Burta. Sposób jej konstruowania dla danych mierzonych na skali nominalnej jest dobrze znany i opisany w literaturze przedmiotu. W przypadku danych symbolicznych w postaci listy wartości, w postaci listy wartości z wagami i interwałowych konstruowanie macierzy Burta polegało na zamianie zmiennych symbolicznych na zmienne nominalne. Podejście to jest związane z pewną utratą informacji, a niekiedy jest niemożliwe do realizacji (np.wymaga, aby wszystkie przedziały określone w zmiennej interwałowej były rozłączne).

W artykule zostanie zaproponowany sposób konstrukcji macierzy Burta wykorzystujący koncepcję kodowania rozmytego. W pierwszej części tekstu opisane zostaną pojęcia obiektu i zmiennej symbolicznej, w drugiej zaś przedstawiona będzie odległość Ichino-Yaguchiego dla zmiennych symbolicznych oraz zaproponowany lemat dotyczący tej odległości. W części trzeciej przedstawiona zostanie propozycja kodowania zmiennych symbolicznych w postaci listy wartości, w postaci listy wartości z wagami i interwałowych. Część czwarta zawiera przykład wykorzystania zaproponowanej techniki do przeprowadzenia wielowymiarowej analizy korespondencji.

2. Obiekty i zmienne symboliczne

Koncepcja obiektu symbolicznego zakłada, że do pełnej reprezentacji danych nie wystarczą tylko wartości liczbowe. Dlatego obiekt symboliczny może, oprócz pojedynczych wartości liczbowych, zawierać:
– łańcuchy tekstowe,
– zmienne symboliczne interwałowe (np. dochód – <2000,3000>),
– zmienne w postaci listy wartości (np. zalety {skromny, niepalący, pracowity}),

– zmienne w postaci listy wartości z wagami (wady {20% niegospodarna, 30% hałaśliwa, 50% zrzęda}).

Niezależnie od typu zmienne symboliczne mogą też mieć zdefiniowaną strukturę wewnętrzną lub określone wzajemne powiązania w postaci:
– zmiennych reprezentujących strukturę hierarchiczną,
– zmiennych zależnych hierarchicznie,
– zmiennych logicznie zależnych.

Zastosowanie obiektów symbolicznych w wielowymiarowej analizie statystycznej można podzielić na dwie grupy:

1. Adaptacja „klasycznych" metod, tak aby obiekty symboliczne mogły służyć jako dane wejściowe do algorytmów wielowymiarowej analizy statystycznej. Dotyczy to zwłaszcza obiektów zawierających zmienne symboliczne interwałowe i zmienne w postaci listy wartości.
2. Tworzenie nowych metod analizy danych przeznaczonych tylko dla obiektów symbolicznych (klasyfikacja metodą piramid, wizualizacja metodą *zoom star* [*Analysis*... 2000]).

Metody analizy korespondencji nie były do tej pory adaptowane dla danych symbolicznych, celem niniejszego artykułu było więc wypełnienie luki występującej w literaturze przedmiotu.

3. Miara odległości Ichino-Yaguchiego dla zmiennych symbolicznych

Struktura danych reprezentowanych w obiektach symbolicznych implikuje fakt, iż do pomiaru ich podobieństwa nie można stosować klasycznych miar, takich jak odległość miejska, euklidesowa, Canberra czy Clarka. Zamiast tego stosowane są inne miary, Bock i Diday [*Analysis*... 2000] wśród najważniejszych z nich wymieniają:

- Gowda i Krishna − miarę wzajemnego sąsiedztwa (*mutual neighborhood*);
- Hausdorffa − miarę odległości między zbiorami;
- Ichino i Yaguchiego − miarę opartą na pojęciach kartezjańskiego połączenia (*Cartesian join*) i kartezjańskiego przekroju (*Cartesian meet*) będących rozszerzeniami operatorów $\cup$ i $\cap$ na wszystkie typy danych reprezentowanych w obiektach symbolicznych;
- De Carvalho – rozszerzenie miar Ichino i Yaguchiego opierające się na pojęciach funkcji porównującej (*CF – comparison function*) i funkcji agregującej (*AF – aggregation function*) oraz na pojęciu potencjału opisowego obiektu symbolicznego.

Aby móc zdefiniować miarę Ichino-Yaguchiego, należy wprowadzić pojęcia operatora połączenia (*Cartesian join*) i operatora przekroju (*Cartesian meet*).

Operator połączenia jest oznaczany symbolem $\oplus$, a jego definicja jest zależna od typu danych, których dotyczy.

1. Dane numeryczne. Dla danych numerycznych (liczbowych) operator połączenia jest zdefiniowany zgodnie ze wzorem (1):

$$A \oplus B = \begin{cases} (A,B) \Leftrightarrow A < B \\ (B,A) \Leftrightarrow A \geq B \end{cases}, \tag{1}$$

gdzie: A, B – zmienne numeryczne, (A,B) – przedział liczbowy.

2. Przedziały liczbowe. Dla danych w postaci przedziału liczbowego operator połączenia jest zdefiniowany zgodnie ze wzorem (2):

$$(a_1,a_2) \oplus (b_1,b_2) = \left(\min(a_1,b_1), \max(a_2,b_2)\right), \tag{2}$$

gdzie: $(a_1,a_2),(b_1,b_2)$ − przedziały liczbowe.

3. Inne typy danych. Dla pozostałych typów danych operator połączenia jest zdefiniowany zgodnie ze wzorem (3):

$$A \oplus B = A \cup B\,, \tag{3}$$

gdzie: A, B − zmienne w postaci zbioru wartości numerycznych lub nienumerycznych.

Operator przekroju jest oznaczony symbolem ⊗. Niezależnie od typu danych, których dotyczy operator przekroju (*Cartesian meet*), jest on zawsze zdefiniowany zgodnie ze wzorem (4):

$$A \otimes B = A \cap B\,, \tag{4}$$

gdzie: A, B − zmienne dowolnego typu.

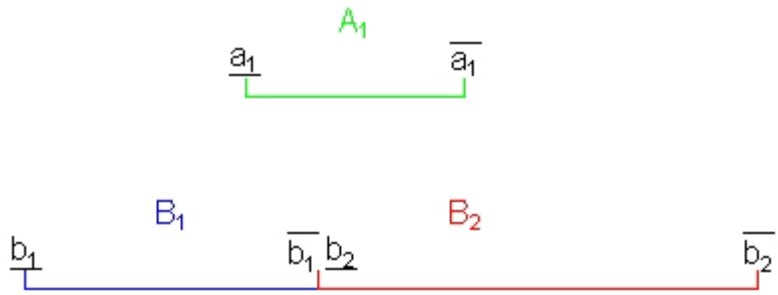
Miara niepodobieństwa zmiennych. Operatory ⊕, ⊗ służą do zdefiniowana miary niepodobieństwa (*dissimilarity measure*) dla zmiennych dowolnego typu. Miara Ichino-Yaguchiego jest zdefiniowana zgodnie ze wzorem (5):

$$\phi(A,B) = \left|A \oplus B\right| - \left|A \otimes B\right| + \gamma\left(2 \cdot \left|A \oplus B\right| - \left|A\right| - \left|B\right|\right), \tag{5}$$

gdzie: $\phi(A,B)$ − miara niepodobieństwa zmiennych symbolicznych, A, B − zmienne dowolnego typu, ⊕ − operator połączenia, ⊗ − operator przekroju, $|\ |$ − symbol oznaczający dla danych w postaci przedziału liczbowego jego długość, a dla pozostałych danych − liczbę elementów zbioru, γ − parametr z przedziału <0, ½ >.

Lemat I. Niech:

$A_1 = \left\langle \underline{a_1}, \overline{a_1} \right\rangle$ − przedział liczbowy, $B_1 = \left\langle \underline{b_1}, \overline{b_1} \right\rangle, B_2 = \left\langle \underline{b_2}, \overline{b_2} \right\rangle$ − przedziały liczbowe takie, że $\underline{b_1} \leq \underline{a_1} \leq \overline{b_1}, \underline{b_2} \leq \overline{a_1} \leq \overline{b_2}, \overline{b_1} = \underline{b_2}$ (czyli przedziały B_1, B_2 są przedziałami sąsiadującymi, a A_1 ma niepustą część wspólną z każdym z nich, a równocześnie jest podzbiorem ich sumy − sytuację tę graficznie przedstawia rys. 1).



Rys. 1. Przedziały $B_1 = \langle \underline{b_1}, \overline{b_1} \rangle, B_2 = \langle \underline{b_2}, \overline{b_2} \rangle$ i przedział $A_1 = \langle \underline{a_1}, \overline{a_1} \rangle$ takie, że $\underline{b_1} \le \underline{a_1} \le \overline{b_1}, \underline{b_2} \le \overline{a_1} \le \overline{b_2}, \overline{b_1} = \underline{b_2}$ – ilustracja graficzna

Źródło: opracowanie własne.

Teza. Dla tak zdefiniowanych przedziałów A_1, B_1, B_2, dla $\gamma = \frac{1}{2}$ zachodzi:

$$\varphi(A_1, B_1) + \varphi(A_1, B_2) = \varphi(B_1, B_2), \tag{6}$$

gdzie : $\varphi(Z_1, Z_z)$ – odległość Ichino-Yaguchiego dla zmiennych symbolicznych.

Dowód:

$$\varphi(A_1, B_1) + \varphi(A_1, B_2) = \varphi(B_1, B_2)$$

$$\overline{a_1} - \underline{b_1} - \left(\overline{b_1} - \underline{a_1}\right) + \frac{1}{2}\left(2\left(\overline{a_1} - \underline{b_1}\right) - \left(\overline{a_1} - \underline{a_1}\right) - \left(\overline{b_1} - \underline{b_1}\right)\right) + \overline{b_2} - \underline{a_1} - \left(\overline{a_1} - \underline{b_2}\right) +$$

$$+ \frac{1}{2}\left(2\left(\overline{b_2} - \underline{a_1}\right) - \left(\overline{a_1} - \underline{a_1}\right) - \left(\overline{b_2} - \underline{b_2}\right)\right) = \overline{b_2} - \underline{b_1} + \frac{1}{2}\left(2\left(\overline{b_2} - \underline{b_1}\right) - \left(\overline{b_1} - \underline{b_1}\right) - \left(\overline{b_2} - \underline{b_2}\right)\right)$$

$$\overline{a_1} - \underline{b_1} - \overline{b_1} + \underline{a_1} + \overline{a_1} - \underline{b_1} - \frac{1}{2}\overline{a_1} + \frac{1}{2}\underline{a_1} - \frac{1}{2}\overline{b_1} + \frac{1}{2}\underline{b_1} + \overline{b_2} - \underline{a_1} - a_1 + \underline{b_2} + \overline{b_2} - \underline{a_1} - \frac{1}{2}\overline{a_1} + \frac{1}{2}\underline{a_1} - \frac{1}{2}\overline{b_2} + \frac{1}{2}\underline{b_2} =$$

$$\frac{3}{2}\overline{b_2} - \frac{3}{2}\underline{a_1} + \frac{1}{2}\underline{b_2} - \frac{1}{2}\overline{a_1}$$

$$\frac{3}{2}\overline{b_2} - \frac{3}{2}\underline{a_1} + \frac{1}{2}\underline{b_2} - \frac{1}{2}\overline{a_1} = \frac{3}{2}\overline{b_2} - \frac{3}{2}\underline{a_1} + \frac{1}{2}\underline{b_2} - \frac{1}{2}\overline{a_1}$$

Analogicznie można udowodnić, że w przypadku *k* sąsiadujących ze sobą przedziałów $B_1, B_2, ..., B_k$ takich, że każdy ma niepustą część wspólną z przedziałem A_1, a równocześnie A_1 jest zawarty w ich sumie, zachodzi (7):

$$\varphi(A_1, B_1) + \varphi(A_1, B_2) + \ldots + \varphi(A_1, B_k) = \varphi(B_1, B_2) + \varphi(B_1, B_3) + \ldots + \varphi(B_1, B_k). \tag{7}$$

Zależności 6 i 7 posłużą do sformułowania zasad kodowania przedziałów liczbowych w macierzy kodów.

4. Propozycja kodowania zmiennych symbolicznych i konstrukcji macierzy Burta

Greenacre [1984] zamiast kodowania w konwencji 0 lub 1 zaproponował w przypadku braku informacji o dokładnej przynależności do danej kategorii ustalenie prawdopodobieństwa przynależności do kategorii rozważanej cechy lub wartość jakiejś innej miary określającej rozmytą regułę przynależności do danej grupy, tak aby każdej kategorii przyporządkować ułamek, przy czym suma dla każdej zmiennej i każdej obserwacji ma dawać 1.

Zmienne w postaci listy wartości. W odpowiednie pola macierzy kodów wpisujemy $|A|^{-1}$. Tabele 1 i 2 przedstawiają wartości przykładowych zmiennych w postaci listy wartości oraz fragment macierzy znaczników dla tych zmiennych.

Tabela 1. Przykładowe zmienne w postaci listy wartości

Obiekt	Zmienna A	Zmienna B
1	{A,C}	{czerwony, żółty, biały}

Źródło: opracowanie własne.

Tabela 2. Fragment macierzy znaczników dla danych z tab. 1

Obiekt	Zmienna A			Zmienna B			
	A	B	C	czerwony	żółty	biały	zielony
1	0,5	0	0,5	0,33	0,33	0,33	0

Źródło: opracowanie własne.

Zmienne w postaci listy wartości z wagami. W odpowiednie pola macierzy kodów wpisujemy wagi.

Tabela 3. Przykładowe zmienne w postaci listy wartości z wagami

Obiekt	Zmienna A	Zmienna B
1	{20% TAK, 80% NIE}	{25% X, 35%Y, 40%Z}

Źródło: opracowanie własne.

Tabela 4. Fragment macierzy znaczników dla danych z tab. 3

Obiekt	Zmienna A			Zmienna B		
	TAK	NIE	NIE WIEM	X	Y	Z
1	0,2	0,8	0	0,25	0,35	0,40

Źródło: opracowanie własne.

Zmienne interwałowe. Ustalamy przedziały bazowe B_1, B_2,…, B_k (np. dzieląc równomiernie dziedzinę zmiennej na przedziały równej długości). W odpowiednie pola macierzy kodów wpisujemy: 0 − jeśli przedział A nie ma części wspólnej z B_i, 1 − jeśli przedział A jest równy B_i. W pozostałych wypadkach wyznaczamy zbiór B_m,…,B_n przedziałów bazowych posiadających część wspólną z A i w i-te pole macierzy kodów wpisujemy:

$$\left(1-\frac{\phi(A,B_i)}{\sum_{i=m+1}^{n-1}\phi(B_{i,}B_n)}\right)(m-n)^{-1}.$$

Tabela 5. Przykładowe wartości zmiennej interwałowej

Obiekt	Zmienna A
1	<0,2>
2	<1,3>

Źródło: opracowanie własne.

Tabela 6. Fragment macierzy znaczników dla danych z tab. 5

Obiekt	Zmienna A		
	<0,2>	<2,5>	**<5,7>**
1	1	0	0
2	0,6	0,4	0

Źródło: opracowanie własne.

Należy zaznaczyć, że na mocy lematu 1 suma kodów dla pojedynczej zmiennej pojedynczego obiektu będzie zawsze dawać 1.

Po utworzeniu macierzy kodów macierz Burta jest tworzona zgodnie ze wzorem (8):

$$B = Z^T \times Z\,. \tag{8}$$

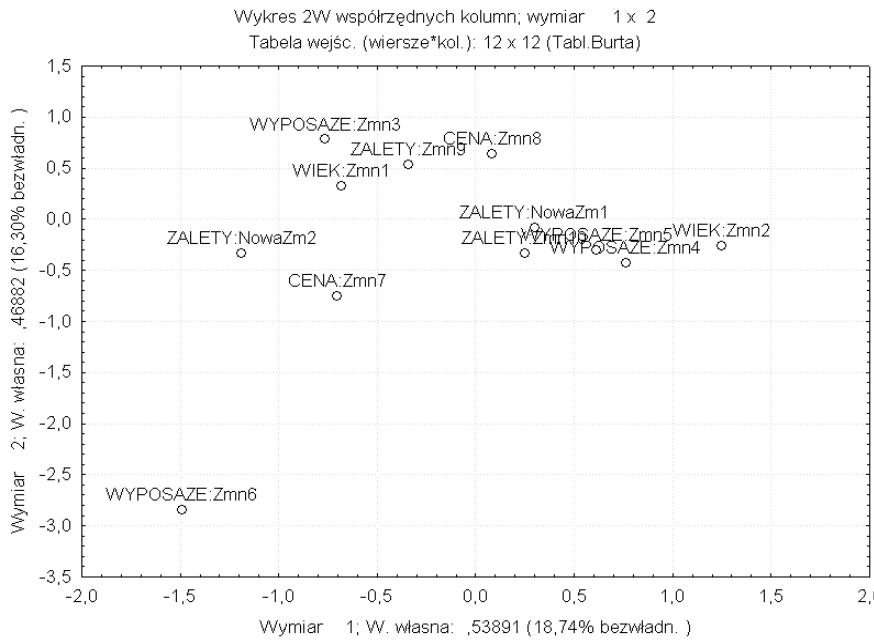
5. Przykład wielowymiarowej analizy korespondencji dla danych symbolicznych

Jako dane wejściowe do wielowymiarowej analizy korespondencji posłużyły dane symboliczne dotyczące samochodów osobowych. Zmienna *Spalanie* jest zmienną interwałową, zmienna *Wyposażenie* − zmienną w postaci listy wartości, zmienna *Wiek* − zmienną interwałową, zmienna *Zalety* − zmienną w postaci listy wartości. Tabela 7 przedstawia macierz Burta dla tych danych.

Tabela 7. Macierz Burta dla danych symbolicznych

Spalanie		Wyposażenie				Wiek		Zalety			
A	B	A	B	C	D	A	B	A	B	C	D
1,34	0,1771	1,1925	0,1925	0,1925	0,1925	0,693	1,077	0,8925	0,2925	0,2925	0,2925
0,1771	3,66	0,0575	0,5575	0,5575	0,0575	0,207	1,023	0,3575	0,3575	0,4575	0,0575
1,1925	0,0575	1,25	0,0625	0,0625	0,0625	0,225	1,025	0,7625	0,1625	0,1625	0,1625
0,1925	0,5575	0,0625	1,75	0,3125	0,0625	0,225	0,525	0,2125	0,2125	0,2625	0,0625
...	...	...	...	...	...	...	...	...	...	...	...

Źródło: opracowanie własne.



Rys. 2. Efekty wielowymiarowej analizy korespondencji dla danych z tab. 7
Źródło: opracowanie własne.

Rysunek 2 Przedstawia efekty analizy korespondencji dla tych danych.

6. Uwagi końcowe

W artykule zaproponowano sposób konstrukcji macierzy Burta dla danych symbolicznych. Może ona posłużyć do przeprowadzenia wielowymiarowej analizy korespondencji dla danych tego typu.

Zaproponowana metoda może znaleźć zastosowanie m.in. do (por. [*Statystyczna…* 2009]):

- segmentacji rynku,
- określenia pozycji produktu na rynku,
- monitorowania skuteczności kampanii reklamowej,
- rozpoznawania luk na rynku.

Literatura

Analysis of symbolic data. Explanatory methods for extracting statistical information from complex data (2000), red. H.H. Bock, E. Diday, Springer Verlag, Heidelberg.

Greenacre M. (1984), *Theory and applications of correspondence analysis*, Academic Press, New York.

*Statystyczna analiza wielowymiarowa z wykorzystaniem programu **R*** (2009), red. M. Walesiak, E. Gatnar, PWN, Warszawa.
Symbolic data analysis. Conceptual statistics and data mining (2006), red. L. Billard, E. Diday, Wiley, Chichester.

THE CONSTRUCTION OF BURT TABLE FOR SYMBOLIC OBJECTS

Summary

Burt table is one of the most important tools of multidimensional correspondence analysis. The algorithm of creation for categorical data is well-known and described in literature of subject.

In this paper, an extension of methods of correspondence analysis onto data represented in form of symbolic objects is proposed. Basing on fuzzy coding introduced by Greenacre (1984), the method of creation code matrix and Burt table is proposed for symbolic data in form of intervals, list of categories, and list of categories with weights.