PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1006 2003
Zastosowania statystyki i matematyki w ekonomii

Marek Walesiak

Akademia Ekonomiczna we Wrocławiu

MIARA ODLEGŁOŚCI OBIEKTÓW OPISANYCH ZMIENNYMI MIERZONYMI NA RÓŻNYCH SKALACH POMIARU

1. Wstęp

Miarę odległości d_{ik} między obiektami A_i, A_k ($i, k = 1, ..., n$) opisanymi zbiorem zmiennych o różnych skalach ich pomiaru zaproponował Gower [7] (zob. np. [5, s. 43-44; 6, s. 21-22; 9, s. 35-36]):

$$d_{ik} = \frac{\sum_{j=1}^{m} \delta_{ik}^{(j)} d_{ik}^{(j)}}{\sum_{j=1}^{m} \delta_{ik}^{(j)}}, \tag{1}$$

gdzie: $j = 1, ..., m$ – numer zmiennej,

$\delta_{ik}^{(j)} = 1$, gdy pomiaru na zmiennej j możemy dokonać dla obydwu obiektów A_i, A_k; w innych sytuacjach $\delta_{ik}^{(j)} = 0$.

Formuła Gowera uśrednia odległości wyznaczone dla poszczególnych zmiennych.

Dla zmiennej o numerze j zmierzonej na skali nominalnej (w tym binarnych) wielkość:

$$d_{ik}^{(j)} = \begin{cases} 0, \text{ gdy między obiektami dla wyników pomiaru na zmiennej } j\text{-tej zachodzi relacja równości,} \\ 1, \text{ gdy między obiektami dla wyników pomiaru na zmiennej } j\text{-tej zachodzi relacja różności.} \end{cases} \tag{2}$$

Jeśli w zbiorze znajdują się tylko zmienne nominalne wielostanowe, formuła (1) z podstawieniem (2) przyjmuje postać współczynnika Sokala i Michenera (por. [9, s. 28]):

ISSN 0324-8445

$$d_{ik} = \frac{\sum_{j=1}^{m} d_{ik}^{(j)}}{m} = \frac{m_r}{m}, \tag{3}$$

gdzie: m_r – liczba zmiennych, dla których między obiektami A_i, A_k zachodzi relacja różności,

m – liczba zmiennych.

Z kolei tylko dla zmiennych binarnych otrzymuje się formułę Sokala i Michenera (zob. [14, s. 28]):

$$d_{ik} = 1 - \frac{a+d}{a+b+c+d}, \tag{4}$$

gdzie: a (d) – liczba zmiennych, dla których obiekty A_i, A_k mają zgodne wartości występowania (braku występowania) odpowiedniego wariantu zmiennej – odpowiednio (+, +) i (–, –);

b (c) – liczba zmiennych, dla których obiekty A_i, A_k mają niezgodne wartości zmiennej – odpowiednio (+, –) i (–, +).

Etapem wstępnym konstrukcji miary (4) jest tab. 1.

Tabela 1. Sposób kodowania dla zmiennych nominalnych binarnych

Zmienna X_j		a_j	b_j	c_j	d_j
Obiekt A_i	Obiekt A_k				
+	+	1	0	0	0
+	–	0	1	0	0
–	+	0	0	1	0
–	–	0	0	0	1

„+" oznacza występuje; „–" oznacza nie występuje,

$\sum_{j=1}^{m} a_j = a$, $\sum_{j=1}^{m} b_j = b$, $\sum_{j=1}^{m} c_j = c$, $\sum_{j=1}^{m} d_j = d$.

Źródło: opracowanie własne.

Dla zmiennych o numerze j zmierzonych na skali interwałowej lub ilorazowej $d_{ik}^{(j)}$ jest zdefiniowane wzorem:

$$d_{ik}^{(j)} = \frac{|x_{ij} - x_{kj}|}{r_j}, \tag{5}$$

gdzie: r_j – rozstęp wyznaczony na podstawie wartości j-tej zmiennej.

Gdy w zbiorze występują tylko zmienne mierzone na skali interwałowej i (lub) ilorazowej, formuła (1) z podstawieniem (5) to odległość miejska (pod warunkiem, że wcześniej przeprowadzono normalizację zmiennych z wykorzystaniem formuły przekształcenia ilorazowego z podstawą normalizacji równą rozstępowi j-tej zmiennej – zob. [14, s. 19]).

Miara odległości (1) przyjmuje wartości z przedziału [0; 1]. Kaufman i Rousseeuw [9, s. 35-36] zaproponowali ponadto, aby na podstawie wzoru (5) wyliczać odległość dla zmiennych mierzonych na skali porządkowej (po uprzednim porangowaniu wariantów zmiennej porządkowej). Propozycja ta jest nie do przyjęcia z punktu widzenia teorii pomiaru, ponieważ dla wyników pomiaru na skali porządkowej jedyną dopuszczalną operacją empiryczną jest zliczanie zdarzeń (tzn. ile można określić relacji mniejszości, większości i równości na wartościach tej skali).

Miara odległości Gowera uwzględniająca zróżnicowane wagi zmiennych przyjmuje postać (zob. [4, s. 103]):

$$d_{ik} = \frac{\sum_{j=1}^{m} w_{ik}^{(j)} d_{ik}^{(j)}}{\sum_{j=1}^{m} w_{ik}^{(j)}}, \tag{6}$$

gdzie: $d_{ik}^{(j)}$ określone wzorami (2) i (5),

$w_{ik}^{(j)}$ – wagi spełniające warunki: $w_{ik}^{(j)} \in [0; m]$, $\sum_{j=1}^{m} w_{ik}^{(j)} = m$.

Waga $w_{ik}^{(j)} = 0$, gdy pomiaru na zmiennej j-tej nie można dokonać dla obydwu obiektów A_i, A_k.

Propozycja odległości Gowera o postaci (1) i jej modyfikacja (6), choć zachęcająca z empirycznego punktu widzenia, budzi jednak wątpliwości:

- wprawdzie odległość ta zapisana jest za pomocą jednego wzoru, ale jest to faktycznie zabieg sztuczny, bowiem dla skali nominalnej, interwałowej i ilorazowej wykorzystuje się inne wzory (odpowiednio o numerach (2) i (5)),
- propozycja ta stosuje niedopuszczalną, z punktu widzenia teorii pomiaru, formułę (5) dla zmiennych mierzonych na skali porządkowej.

2. Uogólniona miara odległości a skale pomiaru zmiennych

W pracy Walesiaka [13] zaproponowano uogólnioną miarę odległości GDM (*Generalised Distance Measure*), w konstrukcji której wykorzystano ideę uogólnionego współczynnika korelacji obejmującego współczynnik korelacji liniowej Pearsona i współczynnik korelacji tau Kendalla (zob. [10, s. 19; 11, s. 266]):

$$d_{ik} = (1 - s_{ik})/2 = \frac{1}{2} - \frac{\sum_{j=1}^{m} w_j a_{ikj} b_{kij} + \sum_{j=1}^{m} \sum_{\substack{l=1 \\ l \neq i,k}}^{n} w_j a_{ilj} b_{klj}}{2\left[\left(\sum_{j=1}^{m} w_j a_{ikj}^2 + \sum_{j=1}^{m} \sum_{\substack{l=1 \\ l \neq i,k}}^{n} w_j a_{ilj}^2\right)\left(\sum_{j=1}^{m} w_j b_{kij}^2 + \sum_{j=1}^{m} \sum_{\substack{l=1 \\ l \neq i,k}}^{n} w_j b_{klj}^2\right)\right]^{\frac{1}{2}}}, \quad (7)$$

gdzie: $d_{ik}(s_{ik})$ – miara odległości (podobieństwa) między obiektami A_i, A_k,

w_j – waga j-tej zmiennej spełniająca warunki: $w_j \in (0; m)$, $\sum_{j=1}^{m} w_j = m$,

$i, k, l = 1, ..., n$ – numer obiektu,

$j = 1, ..., m$ – numer zmiennej.

W uproszczonej postaci formułę odległości (7) można zapisać jako:

$$d_{ik} = (1 - s_{ik})/2 = \frac{1}{2} - \frac{\sum_{j=1}^{m} w_j a_{ikj} b_{kij} + \sum_{j=1}^{m} \sum_{\substack{l=1 \\ l \neq i,k}}^{n} w_j a_{ilj} b_{klj}}{2\left[\sum_{j=1}^{m} \sum_{l=1}^{n} w_j a_{ilj}^2 \cdot \sum_{j=1}^{m} \sum_{l=1}^{n} w_j b_{klj}^2\right]^{\frac{1}{2}}}. \quad (8)$$

Stosowanie konkretnych konstrukcji miar odległości (8) jest uzależnione od skal pomiaru zmiennych. Dla zmiennych mierzonych na skali ilorazowej i (lub) interwałowej w formule (8) stosowane jest podstawienie:

$$\begin{aligned} a_{ipj} &= x_{ij} - x_{pj} \quad \text{dla} \quad p = k, l \\ b_{krj} &= x_{kj} - x_{rj} \quad \text{dla} \quad r = i, l \end{aligned}, \quad (9)$$

gdzie: x_{ij} (x_{kj}, x_{lj}) – i-ta (k-ta, l-ta) obserwacja na j-tej zmiennej.

Zasób informacji skali porządkowej jest nieporównanie mniejszy. Jedyną dopuszczalną operacją empiryczną na skali porządkowej jest zliczanie zdarzeń (tzn. wyznaczanie liczby relacji większości, mniejszości i równości). W związku z tym w konstrukcji miernika odległości musi być wykorzystana informacja o relacjach, w jakich pozostają porównywane obiekty w stosunku do pozostałych obiektów ze zbioru A. Dla zmiennych mierzonych na skali porządkowej w formule (8) stosuje się podstawienie [12, s. 44-45]:

$$a_{ipj}(b_{krj})=\begin{cases} 1 & \text{jeżeli } x_{ij} > x_{pj} \ (x_{kj} > x_{rj}) \\ 0 & \text{jeżeli } x_{ij} = x_{pj} \ (x_{kj} = x_{rj}) \\ -1 & \text{jeżeli } x_{ij} < x_{pj} \ (x_{kj} < x_{rj}) \end{cases}, \text{ dla } p = k,\ l;\ r = i,\ l. \tag{10}$$

Wtedy w mianowniku wzoru (8) pierwszy czynnik oznacza liczbę relacji większości i mniejszości określoną dla obiektu i, czynnik drugi zaś liczbę relacji większości i mniejszości określoną dla obiektu k.

Zasób informacji skali nominalnej zezwala na zliczanie zdarzeń, tzn. wyznaczanie liczby relacji równości i różności. W związku z tym w konstrukcji miernika odległości musi być wykorzystana tego typu informacja. W mianowniku wzoru (8) czynniki iloczynu oznaczają liczbę relacji równości i różności określoną dla obiektu i oraz k, zatem $\sum_{j=1}^{m}\sum_{l=1}^{n} w_j a_{ilj}^2 = \sum_{j=1}^{m}\sum_{l=1}^{n} w_j b_{klj}^2 = m(n-1)$.

Dla zmiennych mierzonych na skali nominalnej w formule (8) stosuje się podstawienia:

a) dla porównywanych obiektów i, k

$$a_{ikj} \cdot b_{kij} = \begin{cases} 1 & \text{dla } x_{ij} = x_{kj} \\ -1 & \text{dla } x_{ij} \neq x_{kj} \end{cases}, \tag{11a}$$

b) dla pozostałych obiektów $(l = 1, ..., n; l \neq i, k)$

$$a_{ilj} \cdot b_{klj} = \begin{cases} 1 & \text{dla } x_{ij} = x_{kj} \wedge (x_{ij}, x_{kj} = x_{lj} \vee x_{ij}, x_{kj} \neq x_{lj}) \\ -1 & \text{dla } x_{ij} \neq x_{kj} \wedge (x_{ij}, x_{kj} \neq x_{lj} \vee x_{ij} \neq x_{lj};\ x_{kj} = x_{lj} \vee x_{ij} = x_{lj};\ x_{kj} \neq x_{lj}) \end{cases}. \tag{11b}$$

Jeśli w zbiorze znajdują się tylko zmienne nominalne wielostanowe, formuła (8) z podstawieniem (11a) i (11b) przyjmuje postać [14, s. 27]:

$$d_{ik} = \frac{\sum_{j=1}^{m} w_j d_{ik}^{(j)}}{\sum_{j=1}^{m} w_j} = \frac{\sum_{j=1}^{m} w_j d_{ik}^{(j)}}{m}, \tag{12}$$

gdzie: $d_{ik}^{(j)}$ określone wzorem (2),

w_j – waga j-tej zmiennej spełniająca warunki: $w_j \in (0; m)$, $\sum_{j=1}^{m} w_j = m$.

We wzorze (12) ważeniu podlega *de facto* relacja równości i różności. Nie jest istotny rozkład wag dla zmiennych, dla których między obiektami A_i, A_k zachodzi relacja różności. Niezależnie bowiem od rozkładu wag dla poszczególnych zmiennych $\sum_{j=1}^{m} w_j d_{ik}^{(j)}$ jest stała.

3. Konstrukcja miary odległości umożliwiająca pomiar podobieństwa obiektów opisanych zmiennymi mierzonymi na różnych skalach pomiaru

Konstrukcja miary odległości d_{ik}, która umożliwia uwzględnienie w badaniach zmiennych mierzonych na skali ilorazowej i (lub) interwałowej (I), porządkowej (P), nominalnej (N), bazuje na propozycji zawartej w pracy [2, s. 152]:

$$d_{ik} = \frac{w_1 d_{ik}^N + w_2 d_{ik}^P + w_3 d_{ik}^I}{w_1 + w_2 + w_3}, \tag{13}$$

gdzie: $N(P, I)$ – podzbiór zmiennych nominalnych (porządkowych, interwałowych i ilorazowych),

$w_1(w_2, w_3)$ – wagi przypisane odległościom wyznaczonym na podstawie zmiennych nominalnych (porządkowych, interwałowych i ilorazowych),

$w_1, w_2, w_3 \in (0, m)$;

$w_1 + w_2 + w_3 = m$ (liczba zmiennych).

Wagi w_1, w_2, w_3 mogą oznaczać liczbę zmiennych w poszczególnych podzbiorach lub merytoryczną ważność poszczególnych podzbiorów zmiennych w wyznaczeniu miary odległości d_{ik} o postaci (13).

Formuła o postaci (13) uśrednia odległości cząstkowe wyznaczone na podstawie poszczególnych podzbiorów zmiennych (nominalnych, porządkowych, interwałowych i ilorazowych).

Miara odległości d_{ik} o postaci (13):

- może być stosowana w sytuacji, gdy obiekty opisane są zmiennymi mierzonymi na skali ilorazowej i (lub) interwałowej, porządkowej oraz nominalnej,
- przybiera wartości z przedziału [0; 1]; wartość 0 oznacza, że dla porównywanych obiektów i, k między odpowiadającymi sobie obserwacjami na zmiennych zachodzą tylko relacje równości,
- spełnia warunki: nieujemności, zwrotności, symetryczności (dla wszystkich $i, k = 1, ..., n$),
- istnieje przynajmniej jedna para obiektów w zbiorze badanych obiektów A, dla której obserwacje na zmiennych nie są identyczne (dla uniknięcia zera w mianowniku d_{ik}^P i d_{ik}^I),
- nie zmienia wartości w wyniku transformacji wartości zmiennych za pomocą dozwolonego na danej skali przekształcenia matematycznego (na skali nominalnej: funkcja wzajemnie jednoznaczna; na skali porządkowej: dowolna ściśle mo-

notonicznie rosnąca funkcja; na skali interwałowej: funkcja liniowa; na skali ilorazowej: funkcja liniowa jednorodna, zob. [3, s. 79]).

4. Podsumowanie

W pracy Walesiaka [13] zaproponowano uogólnioną miarę odległości GDM, która umożliwia uwzględnienie w badaniach zmiennych mierzonych na skali: a) ilorazowej i (lub) interwałowej, b) porządkowej. W artykule zaproponowano wersję miary odległości GDM uwzględniającą zmienne mierzone na skali nominalnej. Ponadto zaproponowano konstrukcję miary odległości umożliwiającą pomiar podobieństwa obiektów opisanych zmiennymi mierzonymi na różnych skalach pomiaru. Formuła ta uśrednia odległości cząstkowe wyznaczone na podstawie poszczególnych podzbiorów zmiennych (odpowiednio nominalnych, porządkowych, interwałowych i ilorazowych).

Literatura

[1] Arabie P., Hubert L.J., De Soete G., *Clustering and Classification*, World Scientific, Singapore 1996.
[2] Bock H.H., Diday E. (red.), *Analysis of Symbolic Data*, Springer-Verlag, Berlin, Heidelberg 2000.
[3] Cegiełka K., Stachowski E., Szymański K. (red.), *Matematyka. Encyklopedia dla wszystkich*, WNT, Warszawa 2000.
[4] Cox T.F., Cox M.A.A., *A General Weighted Two-way Dissimilarity Coefficient*, „Journal of Classification" 2000 Vol. 17, s. 101-121.
[5] Everitt B.S., Landau S., Leese M., *Cluster Analysis*, Edward Arnold, London 2001.
[6] Gordon A.D., *Classification*, Chapman and Hall/CRC, London 1999.
[7] Gower J.C., *A General Coefficient of Similarity and Some of its Properties*, „Biometrics" 1971 (27), s. 857-874.
[8] Jajuga K., Walesiak M., Bąk A., *On the General Distance Measure*, [w:] M. Schwaiger and O. Opitz (red.), *Exploratory data analysis in empirical research*, Springer-Verlag, Berlin, Heidelberg 2003, s. 104-109.
[9] Kaufman L., Rousseeuw P.J., *Finding Groups in Data: an Introduction to Cluster Analysis*, Wiley, New York 1990.
[10] Kendall M.G., *Rank Correlation Methods*, Griffin, London 1955.
[11] Kendall M.G., Buckland W.R., *Słownik terminów statystycznych*, PWE, Warszawa 1986.
[12] Walesiak M., *Statystyczna analiza wielowymiarowa w badaniach marketingowych*, Prace Naukowe AE we Wrocławiu nr 654, Seria: Monografie i Opracowania nr 101, AE, Wrocław 1993.
[13] Walesiak M., *Propozycja uogólnionej miary odległości w statystycznej analizie wielowymiarowej*, [w:] red. J. Paradysz, *Statystyka regionalna w służbie samorządu lokalnego i biznesu*, Internetowa Oficyna Wydawnicza, Centrum Statystyki Regionalnej, Akademia Ekonomiczna w Poznaniu, Poznań 2002, s. 115-121.
[14] Walesiak M., *Uogólniona miara odległości w statystycznej analizie wielowymiarowej*, AE, Wrocław 2002.