PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 950 2002
Ekonometria 10

Marek Walesiak

POMIAR PODOBIEŃSTWA OBIEKTÓW W ŚWIETLE SKAL POMIARU I WAG ZMIENNYCH[1]

1. Wstęp

Wykorzystanie metod klasyfikacji (metody podziału, metody klasyfikacji hierarchicznej, metody wizualizacji – skalowanie wielowymiarowe, analiza korespondencji) i metod porządkowania liniowego opartego na wzorcu rozwoju wymaga sformalizowania pojęcia „podobieństwo obiektów". Stopień podobieństwa obiektów kwantyfikuje się za pomocą miar podobieństwa, wśród których wyróżnia się miary odległości oraz bliskości (por. Dąbrowski i Laus-Mączyńska [1978], s. 49-51; Gatnar [1998], s. 27; Walesiak [1985]).

Funkcja $d: A \times A \to R$ (zbiór liczb rzeczywistych) będzie nazywana miarą odległości wtedy i tylko wtedy, gdy spełnione są warunki $(d(A_i, A_k) = d_{ik}$, dla $i, k = 1, 2, \ldots, n)$: nieujemności: $d_{ik} \geq 0$, zwrotności: $d_{ik} = 0 \Leftrightarrow i = k$ i symetryczności: $d_{ik} = d_{ki}$. Jeśli ponadto spełniony jest warunek nierówności trójkąta $d_{ik} \leq d_{il} + d_{kl}$, to miara odległości jest metryką.

Funkcja $g: A \times A \to R$ będzie nazywana miarą bliskości wtedy i tylko wtedy, gdy spełnione będą warunki $(g(A_i, A_k) = g_{ik}$, $i, k = 1, 2, \ldots, n)$: nieujemności: $0 \leq g_{ik} < 1$ dla $i \neq k$, zwrotności: $g_{ik} = 1 \Leftrightarrow i = k$, symetryczności: $g_{ik} = g_{ki}$.

Sposoby transformacji funkcji bliskości na funkcję odległości wyrażają m.in. formuły (por. Zakrzewska [1987], s. 212): $d_{ik} = 1 - g_{ik}$, $d_{ik} = \sqrt{1 - g_{ik}}$, $d_{ik} = -\log g_{ik}$.

Wszystkie miary podobieństwa mają analogiczną interpretację (chociaż ze względu na odmienne konstrukcje przybierają na ogół różne wartości liczbowe). Dwa obiekty są tym bardziej podobne, im mniej różnią się co do wartości zmiennych.

[1] Pracę wykonano częściowo w ramach projektu badawczego nr 5 H02B 030 21, finansowanego przez Komitet Badań Naukowych w latach 2001-2003.

ISSN 0324-8445
ISSN 1507-3866

2. Charakterystyka miar odległości obiektów z punktu widzenia skal pomiaru i wag zmiennych

Stosowanie konkretnych konstrukcji miar podobieństwa jest uzależnione od skal pomiaru zmiennych[2]. Problem stosowania różnych miar podobieństwa w zasadzie nie występuje wtedy, gdy wszystkie zmienne opisujące badane obiekty są mierzone na skali jednego typu. W literaturze wypracowano wiele propozycji miar podobieństwa mających zastosowanie do zmiennych mierzonych na skali: ilorazowej, przedziałowej i (lub) ilorazowej, nominalnej (w tym dla zmiennych binarnych).

Bardzo dobry przegląd różnych typów miar podobieństwa przedstawiono m.in. w pracach: Cormack [1971]; Cox i Cox [1994], s. 10; Gordon [1999], s. 20-21; Gower [1971]; Anderberg [1973], s. 98-130; Kaufman i Rousseeuw [1990], s. 4-37.

Tabela 1 zawiera zestawienie podstawowych miar odległości dla zmiennych mierzonych na **skali ilorazowej i (lub) przedziałowej**. Podstawową miarą odległości obiektów A_i, A_k, opisanych za pomocą zmiennych mierzonych na skali przedziałowej i (lub) ilorazowej, jest metryka Minkowskiego. Szczególnymi jej przypadkami są odległości: miejska, euklidesowa i Czebyszewa. Cenną zaletą tych trzech miar odległości jest to, że mają interpretację geometryczną. W praktyce wykorzystuje się dwie pierwsze miary, tzn. odległości miejską i euklidesową.

W konstrukcji miar odległości z wagami zróżnicowanymi (1) przyjęto założenie, że ważeniu podlegają wartości zmiennych. Zatem macierz ważonych obserwacji na zmiennych przybiera postać:

$$\left[w_j \cdot z_{ij}\right] = \begin{bmatrix} w_1 z_{11} & w_2 z_{12} & \cdots & w_m z_{1m} \\ w_1 z_{21} & w_2 z_{22} & \cdots & w_m z_{2m} \\ \vdots & \vdots & & \vdots \\ w_1 z_{n1} & w_2 z_{n2} & \cdots & w_m z_{nm} \end{bmatrix}, \tag{1}$$

gdzie z_{ij} – znormalizowana wartość j-tej zmiennej zaobserwowana w i-tym obiekcie.

Dla miar odległości z wagami zróżnicowanymi (2) przyjęto założenie, że ważeniu podlegają odległości cząstkowe wyznaczone dla j-tej zmiennej (por. Gordon [1999], s. 30). Zastosowanie wag w_j pozwala wyznaczyć średnią ważoną odległość między obiektami A_i i A_k.

W literaturze można spotkać trzy sposoby ustalania wag zmiennych. Wagi ustala się albo metodą ekspertów (metoda *a priori*), albo z użyciem algorytmów obliczeniowych opierających się na informacjach zawartych w danych pierwotnych (surowych). Można też wykorzystać metodę opartą na obu tych ujęciach. Szerzej o zagadnieniu ważenia zmiennych napisano w pracach: Bąk [1999], s. 44-47; Borys

---

[2] Skale pomiaru zmiennych omówiono m.in. w pracach: Stevens [1959], Walesiak [1993; 1996].

Tabela 1. Podstawowe miary odległości dla zmiennych mierzonych na skali ilorazowej i (lub) przedziałowej

Nazwa miary	Odległość d_{ik}		
	wagi jednakowe	wagi zróżnicowane (1)	wagi zróżnicowane (2)
Minkowskiego (dla $p \geq 1$)	$\sqrt[p]{\sum_{j=1}^{m} \lvert z_{ij} - z_{kj} \rvert^p}$	$\sqrt[p]{\sum_{j=1}^{m} w_j^p \lvert z_{ij} - z_{kj} \rvert^p}$	$\sqrt[p]{\sum_{j=1}^{m} w_j \lvert z_{ij} - z_{kj} \rvert^p}$
– miejska (dla $p = 1$)	$\sum_{j=1}^{m} \lvert z_{ij} - z_{kj} \rvert$	$\sum_{j=1}^{m} w_j \lvert z_{ij} - z_{kj} \rvert$	
– euklidesowa (dla $p = 2$)	$\sqrt{\sum_{j=1}^{m} \lvert z_{ij} - z_{kj} \rvert^2}$	$\sqrt{\sum_{j=1}^{m} w_j^2 \lvert z_{ij} - z_{kj} \rvert^2}$	$\sqrt{\sum_{j=1}^{m} w_j \lvert z_{ij} - z_{kj} \rvert^2}$
– Czebyszewa (dla $p \to \infty$)	$\max_j \lvert z_{ij} - z_{kj} \rvert$	$\max_j w_j \lvert z_{ij} - z_{kj} \rvert$	
Canberra	$\sum_{j=1}^{m} \frac{\lvert z_{ij} - z_{kj} \rvert}{(z_{ij} + z_{kj})}$		$\sum_{j=1}^{m} w_j \frac{\lvert z_{ij} - z_{kj} \rvert}{(z_{ij} + z_{kj})}$
Braya-Curtisa	$\frac{\sum_{j=1}^{m} \lvert z_{ij} - z_{kj} \rvert}{\sum_{j=1}^{m} (z_{ij} + z_{kj})}$	$\frac{\sum_{j=1}^{m} w_j \lvert z_{ij} - z_{kj} \rvert}{\sum_{j=1}^{m} w_j (z_{ij} + z_{kj})}$	
Clarka	$\sqrt{\frac{1}{m} \sum_{j=1}^{m} \left(\frac{z_{ij} - z_{kj}}{z_{ij} + z_{kj}} \right)^2}$		$\sqrt{\frac{1}{m} \sum_{j=1}^{m} w_j \left(\frac{z_{ij} - z_{kj}}{z_{ij} + z_{kj}} \right)^2}$
Jeffreysa-Matusita	$\sum_{j=1}^{m} \left(\sqrt{z_{ij}} - \sqrt{z_{kj}} \right)^2$	$\sum_{j=1}^{m} w_j \left(\sqrt{z_{ij}} - \sqrt{z_{kj}} \right)^2$	

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

(1) – ważeniu podlegają wartości zmiennych (wagi liniowe),

(2) – ważeniu podlegają odległości cząstkowe wyznaczone dla j-tej zmiennej,

z_{ij} (z_{kj}) – znormalizowana wartość j-tej zmiennej dla i-tego (k-tego) obiektu.

Źródło: opracowanie własne na podstawie prac: Bąk [1999], s. 19-22, 62-63; Gordon [1981], s. 21-22; Gordon [1999], s. 20-21; Wedel i Kamakura [1998], s. 47; Zaborski [2001], s. 44; Zeliaś i in. [2000], s. 83-85.

[1984], s. 318-325; Abrahamowicz i Zając [1986]; Grabiński [1984], s. 25-30; Milligan [1989]. Problem „ważenia" zmiennych nie został dotychczas zadowalająco rozwiązany. Williams stwierdza nawet, że ważenie zmiennych jest manipulowaniem wartościami zmiennych (por. Aldenderfer i Blashfield [1984], s. 21). Z tego

względu często w badaniach empirycznych zakłada się, że zmienne są jednakowo ważne z punktu widzenia badanego problemu (takie stanowisko przyjmują m.in. Sneath i Sokal [1973][3]).

Miary odległości dla zmiennych mierzonych na skali ilorazowej i (lub) przedziałowej zamieszczone w tab. 1 wykorzystują w obliczeniach znormalizowane wartości zmiennych. Wyznaczanie odległości z wykorzystaniem pierwotnych wartości zmiennych x_{ij} jest możliwe za pomocą odległości Mahalanobisa (por. Jajuga [1990], s. 22):

$$d_{ik} = \left[(\mathbf{x}_i - \mathbf{x}_k)^{\mathrm{T}} \mathbf{S}^{-1} (\mathbf{x}_i - \mathbf{x}_k)\right]^{0,5} \tag{2}$$

lub w zapisie skalarnym:

$$d_{ik} = \left[\sum_{j=1}^{m}\sum_{l=1}^{m} s_{jl}\left(x_{ij} - x_{kj}\right)\left(x_{il} - x_{kl}\right)\right]^{0,5}, \tag{3}$$

gdzie: s_{jl} – element macierzy odwrotnej do macierzy kowariancji,

$\mathbf{S}$ – macierz kowariancji zbioru obserwacji,

$\mathbf{x}_i$ – wielowymiarowa obserwacja określona wzorem $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{im}]^{\mathrm{T}}$.

Macierz kowariancji zbioru obserwacji $\mathbf{S}$ wyznacza się ze wzoru:

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^{\mathrm{T}}, \tag{4}$$

gdzie $\bar{\mathbf{x}}$ – wektor średnich zbioru obserwacji.

Przy obliczaniu odległości Mahalanobisa brana jest pod uwagę macierz kowariancji zbioru obserwacji, zatem następuje ujednolicenie wartości zmiennych pod względem jednostki miary i rzędu wielkości (zob. Jajuga [1993], s. 58).

Jeśli normalizacji zbioru obserwacji dokona się z wykorzystaniem przekształcenia Mahalanobisa, to odległość euklidesowa równa się odległości Mahalanobisa wyznaczonej z wykorzystaniem pierwotnych wartości zmiennych (por. Jajuga [1993], s. 59).

Przekształcenie Mahalanobisa pozwala przeprowadzić normalizację łącznie dla wszystkich zmiennych (zob. Jajuga [1993], s. 58; Jajuga i Walesiak [2000], s. 110):

$$\mathbf{z}_i = \mathbf{S}^{-0,5}(\mathbf{x}_i - \bar{\mathbf{x}}). \tag{5}$$

---

[3] Zob. Aldenderfer i Blashfield [1984], s. 21.

Macierz $\mathbf{S}^{-0,5}$ wyznacza się ze wzoru (por. Jajuga [1993], s. 58):

$$\mathbf{S}^{-0,5} = \left(\mathbf{G}\mathbf{L}^{0,5}\mathbf{G}^{\mathrm{T}}\right)^{-1}, \tag{6}$$

gdzie: $\mathbf{L}^{0,5}$ – macierz diagonalna o wymiarach $m \times m$ (na głównej przekątnej tej macierzy znajdują się pierwiastki kwadratowe wartości własnych macierzy **S** uporządkowane malejąco);

G – macierz ortogonalna o wymiarach m x m, której kolumny są unormowanymi wektorami własnymi odpowiadającymi uporządkowanym malejąco wartościom własnym macierzy **S**.

Miarę odległości obiektów, którą można stosować w sytuacji, gdy w zbiorze są zmienne mierzone na **skali porządkowej**, zaproponowano w pracy Walesiaka [2000]. Uogólniona miara odległości GDM przybiera postać:

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

gdzie: d_{ik} (s_{ik}) – miara odległości (bliskości),

$i, k, l = 1, 2, \ldots, n$ – numer obiektu,

$j = 1, 2, \ldots, m$ – numer zmiennej,

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

Stosowanie konkretnych konstrukcji miary odległości (7) jest uzależnione od skal pomiaru zmiennych. Dla zmiennych mierzonych na skali ilorazowej i (lub) przedziałowej w formule (7) stosowane jest podstawienie:

$$\begin{aligned} a_{ipj} &= x_{ij} - x_{pj} \quad \text{dla } p = k, l, \\ b_{krj} &= x_{kj} - x_{rj} \quad \text{dla } r = i, l. \end{aligned} \tag{8}$$

Zastosowanie formuły (7) dla zmiennych mierzonych na skali ilorazowej i (lub) przedziałowej wymaga wcześniejszej normalizacji zmiennych. Normalizację zmiennych przeprowadza się celem ich sprowadzenia do porównywalności. Po normalizacji dla podstawienia (8) w miejsce symbolu x wystąpi symbol z. Niezależnie od tego jednak, czy przeprowadzi się normalizację, czy też nie, wartości miary (7) z podstawieniem (8) zawierają się w przedziale [0; 1].

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 porównywane obiekty pozostają w stosunku do pozostałych obiektów ze zbioru A. Dla zmiennych mierzonych na skali porządkowej w formule (7) stosuje się podstawienie (Walesiak [1993a], s. 44-45):

$$a_{ipj}\,(b_{krj}) = \begin{cases} 1 & \text{jeżeli } x_{ij} > x_{pj}\ (x_{kj} > x_{rj}) \\ 0 & \text{jeżeli } x_{ij} = x_{pj}\ (x_{kj} = x_{rj}), \text{ dla } p = k, l;\ r = i, l. \\ -1 & \text{jeżeli } x_{ij} < x_{pj}\ (x_{kj} < x_{rj}) \end{cases} \tag{9}$$

Wtedy w mianowniku wzoru (7) pierwszy czynnik oznacza liczbę relacji większości i mniejszości określoną dla obiektu i, czynnik drugi zaś liczbę relacji większości i mniejszości określoną dla obiektu k.

Miara o postaci (7) z podstawieniem (8) jest stosowana jako miara odległości dla zmiennych mierzonych na skali przedziałowej i (lub) ilorazowej. Wprowadzenie do wzoru (7) podstawienia (9) oznacza, że jest to miara odległości dla zmiennych mierzonych na skali porządkowej. Płynie stąd wniosek, że miary (7) nie można stosować bezpośrednio, gdy zmienne są mierzone jednocześnie na różnych skalach. Zastosowanie miary (7) z podstawieniem (9) rozwiązuje częściowo ten problem, ale wtedy zostaje osłabiona skala pomiaru dla grupy zmiennych mierzonych na skali przedziałowej i (lub) ilorazowej (zostają one przekształcone w zmienne porządkowe, ponieważ w obliczeniach uwzględniane są tylko relacje większości, mniejszości i równości).

Uogólniona postać miary odległości, w której uwzględnia się wagi zmiennych, określona jest wzorem (por. Walesiak [1999], s. 170):

$$d_{ik} = \frac{1}{2} - \frac{\sum_{j=1}^{m} w_j\, a_{ikj}\, b_{kij} + \sum_{j=1}^{m} \sum_{\substack{l=1 \\ l \neq i,\, k}}^{n} w_j\, a_{ilj}\, b_{klj}}{2\left[\sum_{j=1}^{m}\sum_{l=1}^{n} w_j\, a_{ilj}^2 \cdot \sum_{j=1}^{m}\sum_{l=1}^{n} w_j\, b_{klj}^2\right]^{\frac{1}{2}}}, \tag{10}$$

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

Miara odległości d_{ik} o postaci (10) (zob. Walesiak [1999], s. 171):

- może być stosowana w sytuacji, gdy obiekty opisane są zmiennymi mierzonymi na skali ilorazowej, przedziałowej lub porządkowej;
- przybiera wartości z przedziału [0; 1]; wartość 0 oznacza, że dla porównywanych obiektów i, k między odpowiadającymi sobie obserwacjami na zmiennych za-

chodzą tylko relacje równości. W razie podstawienia (9) wartość 1 oznacza, że gdy dla porównywanych obiektów i, k między odpowiadającymi sobie obserwacjami na zmiennych porządkowych zachodzą tylko relacje większości (mniejszości) lub relacje większości (mniejszości) oraz relacje równości, jeżeli relacje te są zachowane w stosunku do pozostałych obiektów (a więc obiektów o numerach $l = 1, 2, \ldots, n$; gdzie $l \neq i, k$);

- spełnia warunki: nieujemności $d_{ik} \geq 0$, zwrotności $d_{ii} = 0$, symetryczności $d_{ik} = d_{ki}$ (dla wszystkich $i, k = 1, 2, \ldots, n$);
- nie zawsze spełnia warunek nierówności trójkąta (potwierdziły ten wniosek przeprowadzone analizy symulacyjne);
- 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});
- nie zmienia wartości w wyniku transformacji wartości zmiennych za pomocą dozwolonego na danej skali przekształcenia matematycznego (na skali porządkowej – dowolna ściśle monotonicznie rosnąca funkcja; na skali przedziałowej – funkcja liniowa; na skali ilorazowej – funkcja liniowa jednorodna[4]).

W literaturze z zakresu statystycznej analizy wielowymiarowej nie zaproponowano dotychczas innych miar odległości dla **zmiennych porządkowych**. Miara odległości Kendalla ([1966], s. 181) o postaci (11) nie jest typową miarą dla zmiennych porządkowych:

$$d_{ik} = \sum_{j=1}^{m} \frac{\left(x_{ij} - x_{kj}\right)^2}{s_j}, \tag{11}$$

gdzie s_j – odchylenie standardowe dla j-tej zmiennej.

Zastosowanie tej miary odległości wymaga uprzedniego porangowania obserwacji. Formuła ta jest w rzeczywistości kwadratem odległości euklidesowej (po uprzedniej normalizacji zmiennych polegającej na podzieleniu wszystkich obserwacji przez ich wariancję). Kwadrat odległości euklidesowej wolno, z punktu widzenia teorii pomiaru, stosować tylko dla zmiennych ze skali przedziałowej i (lub) ilorazowej. Miara odległości Kendalla nie jest typową miarą dla zmiennych mierzonych na skali porządkowej, stosując ją bowiem, zakłada się, że odległości pomiędzy sąsiednimi wartościami na skali porządkowej są sobie równe (na skali porządkowej odległości między dowolnymi dwiema wartościami nie są znane). Takich propozycji jak powyższa w literaturze jest więcej (zob. np. Kaufman i Rousseeuw [1990], s. 30). Przyjmuje się wtedy upraszczające założenie, że rangi są mierzone co najmniej na skali przedziałowej (wtedy dopuszcza się wyznaczanie różnic między wartościami skali).

[4] Zob. Cegiełka, Stachowski i Szymański [2000], s. 79.

Miarę podobieństwa obiektów A_i, A_k wykorzystywaną w sytuacji, gdy są one opisane za pomocą **zmiennych nominalnych wielostanowych**, zaproponowali Sokal i Michener (por. Kaufman i Rousseeuw [1990], s. 28):

$$d_{ik} = \frac{\sum_{j=1}^{m}\left(1 - g_{ik}^{(j)}\right)}{m} = \frac{m - m_r}{m}, \tag{12}$$

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

m – liczba zmiennych,

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

Miara odległości obiektów opisanych zmiennymi nominalnymi wielostanowymi uwzględniająca zróżnicowane wagi zmiennych przybiera postać:

$$d_{ik} = \frac{\sum_{j=1}^{m} w_j\left(1 - g_{ik}^{(j)}\right)}{\sum_{j=1}^{m} w_j} = \frac{m - \sum_{j=1}^{m} w_j g_{ik}^{(j)}}{m}. \tag{13}$$

We wzorze (13) *de facto* ważeniu podlega relacja równości i różności. Nie jest istotny rozkład wag dla zmiennych, dla których między obiektami A_i, A_k zachodzi relacja równości. Niezależnie bowiem od rozkładu wag dla poszczególnych zmiennych $\sum_{j=1}^{m} w_j g_{ik}^{(j)}$ jest stała.

W literaturze dotyczącej wielowymiarowej analizy statystycznej wypracowano bardzo dużo miar podobieństwa obiektów opisanych za pomocą tylko **zmiennych nominalnych binarnych**. Etapem wstępnym konstrukcji tych miar jest tab. 2.

Niech $\sum_{j=1}^{m} a_j = a$, $\sum_{j=1}^{m} b_j = b$, $\sum_{j=1}^{m} c_j = c$, $\sum_{j=1}^{m} d_j = d$, gdzie a (d) oznacza liczbę zmiennych, dla których obiekty A_i, A_k mają zgodne wartości występowania (braku występowania) odpowiedniego wariantu zmiennej – odpowiednio (+, +) i (–, –); b (c) – liczbę zmiennych, dla których obiekty A_i, A_k mają niezgodne wartości zmiennej – odpowiednio (+, –) i (–, +).

Zestawienie wybranych miar odległości obiektów będących funkcją a, b, c i d dla zmiennych nominalnych binarnych przedstawia tab. 3.

Tabela 2. Sposób kodowania 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".
Źródło: opracowanie własne.

Tabela 3. Zestawienie wybranych miar odległości obiektów dla zmiennych nominalnych binarnych

Nazwa miary	Odległość d_{ik}	Wagi przypisane parom zgodnym i niezgodnym
Sokala i Michenera	$\frac{b+c}{a+b+c+d} = 1 - \frac{a+d}{a+b+c+d}$	jednakowe
Jaccarda	$\frac{b+c}{a+b+c} = 1 - \frac{a}{a+b+c}$	jednakowe
Czekanowskiego	$\frac{b+c}{2a+b+c} = 1 - \frac{2a}{2a+b+c}$	zróżnicowane (podwójna waga przypisana parom zgodnym)
Rogersa i Tanimoto	$\frac{2(b+c)}{a+d+2(b+c)} = 1 - \frac{a+d}{a+d+2(b+c)}$	zróżnicowane (podwójna waga przypisana parom niezgodnym)

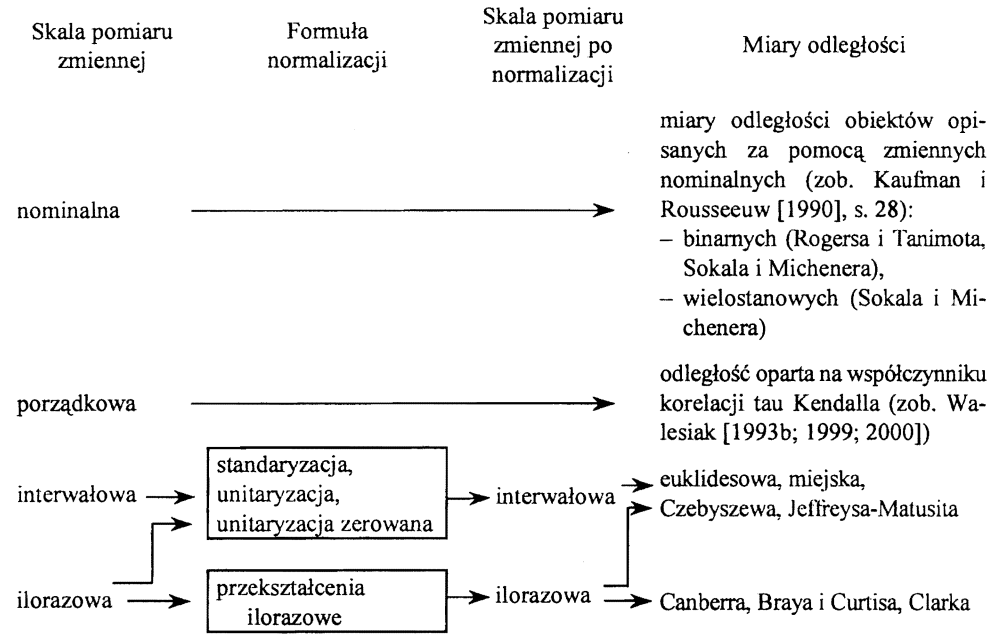
Źródło: opracowanie własne na podstawie pracy: Anderberg [1973], s. 89, 113-117.

W zagadnieniu klasyfikacji oraz skalowania wielowymiarowego w zbiorze mogą być zmienne mierzone na różnych skalach pomiaru; z kolei zagadnienie porządkowania liniowego wymaga, aby w zbiorze były zmienne mierzone przynajmniej na skali porządkowej (ze względu na to, że porządkowanie obiektów staje się możliwe, gdy dopuszczalne jest określenie na wartościach zmiennych relacji większości i mniejszości).

Problem stosowania konkretnych konstrukcji miar podobieństwa w zagadnieniu klasyfikacji i skalowania wielowymiarowego nie występuje w zasadzie wtedy, gdy wszystkie zmienne są mierzone na skali pomiaru jednego typu. Dla zmiennych mierzonych na skali jednego typu istnieją rozmaite konstrukcje miar podobieństwa. Z kolei w zagadnieniu porządkowania liniowego wypracowano wiele konstrukcji syntetycznych mierników rozwoju w przypadku, gdy w zbiorze znajdują się zmienne mierzone tylko na skali przedziałowej i (lub) ilorazowej. Różne konstrukcje mierników odnoszących się do tych grup zmiennych omówił m.in. Walesiak [1990].

Przy wyborze miar odległości obiektów opisanych zmiennymi mierzonymi na skali przedziałowej i (lub) ilorazowej należy wziąć pod uwagę zastosowaną formu-

łę normalizacji wartości zmiennych. Klasyfikację formuł normalizacyjnych oraz miar podobieństwa obiektów z punktu widzenia skal pomiaru zmiennych przedstawia rys. 1.



Rys. 1. Klasyfikacja formuł normalizacyjnych oraz miar odległości obiektów z punktu widzenia skal pomiaru zmiennych

Źródło: zob. Jajuga i Walesiak [2000], s. 109.

Sytuacja komplikuje się wtedy, gdy w zbiorze znajdują się zmienne mierzone na skalach różnych rodzajów. Na podstawie literatury przedmiotu (por. Kaufman i Rousseeuw [1990], s. 32-37; Kolonko [1979]; Gordon [1981], s. 25-27; Jajuga [1989]; Walesiak [1993b]) do rozwiązania tego problemu można wykorzystać następujące sposoby:

a. Przeprowadzić klasyfikację, skalowanie wielowymiarowe i porządkowanie liniowe zbioru obiektów osobno dla każdej grupy zmiennych. Gdy tak otrzymane rezultaty są w miarę zgodne, problem można uznać za rozwiązany. Sytuacja komplikuje się wtedy, gdy wyniki te znacznie odbiegają od siebie.

b. Wykorzystać w analizie tylko zmienne jednego ustalonego typu (dominującego w zbiorze zmiennych) z odrzuceniem zmiennych innego typu. Wyniki uzyskane na podstawie zbioru zmiennych uzyskanego w taki sposób są na ogół bardzo zniekształcone (wskutek tego, że należy zrezygnować z części informacji, które niosą odrzucone zmienne).

c. W praktyce zaniedbać to, że zmienne są mierzone na skalach różnych typów, i stosować metody właściwe w odniesieniu do zmiennych jednego typu. Zmienne nominalne i porządkowe traktuje się zazwyczaj tak jak przedziałowe i ilorazowe: stosuje się więc do nich techniki właściwe tym skalom. Sposób ten, choć atrakcyjny z aplikacyjnego punktu widzenia, jest nie do przyjęcia ze względów metodologicznych (następuje tu bowiem sztuczne wzmocnienie skali pomiaru).

d. Dokonać transformacji zmiennych tak, by sprowadzić je do skali jednego typu. Podstawowa reguła teorii pomiaru mówi, że jedynie rezultaty pomiaru w skali mocniejszej mogą być transformowane na liczby należące do skali słabszej. Wynika z tego, że wszystkie obserwacje na zmiennych należy przekodować na pomiary na skali najsłabszej. Tej operacji towarzyszy jednak utrata informacji. Proponowane są również w tym względzie procedury wzmacniania skal pomiaru (por. Anderberg [1973], s. 53-69; Pociecha [1986]). Są to aproksymacyjne metody przekształcania skal słabszych w silniejsze, opierające się na pewnych dodatkowych informacjach. Z punktu widzenia teorii pomiaru wzmacnianie skal jest jednak niemożliwe, ponieważ z mniejszej ilości informacji nie można uzyskać większej ilości informacji.

e. Posłużyć się metodami (miarami podobieństwa, konstrukcjami syntetycznych mierników rozwoju) dopuszczającymi wykorzystanie zmiennych mierzonych na różnych skalach.

Miarę odległości między obiektami opisanymi zbiorem zmiennych o różnych skalach ich pomiaru zaproponował Gower [1971]:

$$d_{ik} = \frac{\sum_{j=1}^{m} \delta_{ik}^{(j)} d_{ik}^{(j)}}{\sum_{j=1}^{m} \delta_{ik}^{(j)}}. \tag{14}$$

Czynnik $\delta_{ik}^{(j)}$ przybiera wartość 1, gdy pomiaru na zmiennej j można dokonać dla obydwu obiektów i, k. W innych sytuacjach przybiera wartość 0.

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

$$d_{ik}^{(j)} = \begin{cases} 0 & \text{gdy między obiektami dla wyników pomiaru na zmiennej } j\text{-tej} \\ & \text{zachodzi relacja równości,} \\ 1 & \text{gdy między obiektami dla wyników pomiaru na zmiennej } j\text{-tej} \\ & \text{zachodzi relacja różności.} \end{cases} \tag{15}$$

Jeśli w zbiorze znajdują się tylko zmienne nominalne, formuła (14) przybiera postać współczynnika Sokala i Michenera (12). Z kolei tylko dla zmiennych binarnych otrzymuje się formułę Sokala i Michenera, zaprezentowaną w tab. 3.

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

$$d_{ik}^{(j)} = \frac{|x_{ij} - x_{kj}|}{r_j}, \tag{16}$$

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

Gdy w zbiorze występują tylko zmienne mierzone na skali przedziałowej i (lub) ilorazowej, formuła (14) to odległość miejska (pod warunkiem, że wcześniej przeprowadzono normalizację zmiennych z wykorzystaniem formuły przekształcenia ilorazowego o postaci $z_{ij} = x_{ij} / r_j$).

Miara odległości (14) przybiera wartości z przedziału [0; 1]. Kaufman i Rousseeuw [1990], s. 35-36 zaproponowali ponadto, aby na podstawie wzoru (16) 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, albowiem 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 przybiera postać (zob. Cox i Cox [2000], s. 103):

$$d_{ik} = \frac{\sum_{j=1}^{m} w_{ik}^{(j)} d_{ik}^{(j)}}{\sum_{j=1}^{m} w_{ik}^{(j)}}, \tag{17}$$

gdzie $w_{ik}^{(j)}$ – 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: i, k.

Propozycja odległości Gowera o postaci (14) i (17), choć zachęcająca z empirycznego punktu widzenia, budzi jednak wątpliwości. Wprawdzie odległość ta jest zapisana za pomocą jednego wzoru, ale jest to faktycznie zabieg sztuczny, albowiem dla skal nominalnej i przedziałowej oraz ilorazowej wykorzystuje się inne wzory (odpowiednio o numerach (15) i (16)).

Dotychczas w empirycznych zastosowaniach zagadnienia klasyfikacji i porządkowania liniowego, gdy w zbiorze zmiennych występowały zmienne mierzone

co najmniej na skali porządkowej, wykorzystywano sposób *c*, w którym zmienne porządkowe traktowano jak zmienne przedziałowe lub ilorazowe. Zastosowanie formuły (10) z podstawieniem (9) pozwala wykorzystać zgodny z teorią pomiaru sposób d), w którym obserwacje na zmiennych przedziałowych i ilorazowych zostają przekodowane na pomiary na zmiennych porządkowych.

3. Uwagi końcowe

W artykule zaprezentowano przegląd miar odległości obiektów w świetle skal pomiaru i sposobów wprowadzania wag zmiennych. Artykuł jest rozszerzoną i uzupełnioną wersją artykułu Walesiaka [1994]. W nowej wersji artykuł rozszerzono i uzupełniono problematyką dotyczącą sposobów uwzględniania wag w formułach odległości, nowych formuł odległości, nowych publikacji z omawianego zakresu.

Literatura

Abrahamowicz M., Zając K. (1986): *Metoda ważenia zmiennych w taksonomii numerycznej i procedurach porządkowania liniowego*. W: Prace Naukowe AE we Wrocławiu nr 328, s. 5-17.

Aldenderfer M.S., Blashfield R.K. (1984): *Cluster Analysis*. Beverly Hills: Sage.

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

Bąk A. (1999): *Modelowanie symulacyjne wybranych algorytmów wielowymiarowej analizy porównawczej w języku C++*. Wrocław: Wyd. AE.

Borys T. (1984): *Kategoria jakości w statystycznej analizie porównawczej*. Prace Naukowe AE we Wrocławiu nr 284. Seria: Monografie i Opracowania nr 23.

Cegiełka K., Stachowski E., Szymański K. (red.) (2000): *Matematyka. Encyklopedia dla wszystkich*. Warszawa: WNT.

Cormack R.M. (1971): *A Review of Classification (with Discussion)*. „Journal of the Royal Statistical Society", Ser. A part 3, s. 321-367.

Cox T.F., Cox M.A.A. (1994): *Multidimensional Scaling*. London: Chapman & Hall.

Cox T.F., Cox M.A.A. (2000): *A General Weighted Two-way Dissimilarity Coefficient*. „Journal of Classification", vol. 17, s. 101-121.

Dąbrowski M., Laus-Mączyńska K. (1978): *Metody wyszukiwania i klasyfikacji informacji*. Warszawa: WNT.

Gatnar E. (1998): *Symboliczne metody klasyfikacji danych*. Warszawa: PWN.

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

Gordon A.D. (1999): *Classification*. 2nd Edition, London: Chapman and Hall/CRC.

Gower J.C. (1971): *A General Coefficient of Similarity and Some of its Properties*. „Biometrics" (27), s. 857-874.

Grabiński T. (1984): *Wielowymiarowa analiza porównawcza w badaniach dynamiki zjawisk ekonomicznych*. Zeszyty Naukowe AE w Krakowie, Seria specjalna: Monografie nr 61.

Jajuga K. (1989): *Podstawowe metody analizy wielowymiarowej w przypadku występowania zmiennych mierzonych na różnych skalach*. Akademia Ekonomiczna we Wrocławiu. Praca wykonana w ramach CPBP 10.09 (materiał powielony).

Jajuga K. (1990): *Statystyczna teoria rozpoznawania obrazów*. Warszawa: PWN.
Jajuga K. (1993): *Statystyczna analiza wielowymiarowa*. Warszawa: PWN.
Jajuga K., Walesiak M. (2000): *Standardisation of Data Set under Different Measurement Scales*. W: *Classification and Information Processing at the Turn of the Millennium*. Red. R. Decker, W. Gaul. Springer-Verlag, Berlin, Heidelberg, s. 105-112.
Jajuga K., Walesiak M., Bąk A. (2001): *On the General Distance Measure*. Referat na 25 Konferencję Naukową Niemieckiego Towarzystwa Klasyfikacyjnego (Gesellschaft für Klassifikation e.V.), Uniwersytet w Monachium, 14-16 marca 2001.
Kaufman L., Rousseeuw P.J. (1990): *Finding Groups in Data: an Introduction to Cluster Analysis*. New York: Wiley.
Kendall M.G. (1966): *Discrimination and Classification*. W: *Multivariate Analysis I*. Red. P.R. Krishnaiah. New York, London: Academic Press, s. 165-185.
Kolonko J. (1979): *O wykorzystaniu w badaniach taksonomicznych danych pierwotnych mierzonych na skalach różnego typu. Materiały konferencyjne nt. Metody taksonomiczne i ich zastosowanie w badaniach ekonomicznych*. Szklarska Poręba, 25.10.1979 r. (materiał powielony).
Milligan G.W. (1989): *A Validation Study of a Variable Weighting Algorithm for Cluster Analysis*. „Journal of Classification" No. 1, s. 53-71.
Milligan G.W., Cooper M.C. (1988): *A Study of Standarization of Variables in Cluster Analysis*. „Journal of Classification" No. 2, s. 181-204.
Sneath P.H.A., Sokal R.R. (1973): *Numerical Taxonomy*. San Francisco: W.H. Freeman and Co.
Stevens S.S. (1959): *Measurement, Psychophysics and Utility*. W: *Measurement; Definitions and Theories*. C.W. Churchman, P. Ratoosh. New York: Wiley, s. 18-61.
Taksonomiczna analiza przestrzennego zróżnicowania poziomu życia w Polsce w ujęciu dynamicznym (2000). Red. A. Zeliaś. Kraków: Wyd. AE.
Walesiak M. (1985): *Metody klasyfikacji w badaniach strukturalnych*. Akademia Ekonomiczna we Wrocławiu (rozprawa doktorska).
Walesiak M. (1990): *Syntetyczne badania porównawcze w świetle teorii pomiaru*. „Przegląd Statystyczny" z. 1-2, s. 37-46.
Walesiak M. (1993a): *Statystyczna analiza wielowymiarowa w badaniach marketingowych*. Prace Naukowe AE we Wrocławiu nr 654. Seria: Monografie i Opracowania nr 101.
Walesiak M. (1993b): *Strategie postępowania w badaniach statystycznych w przypadku zbioru zmiennych mierzonych na skalach różnego typu*. „Badania Operacyjne i Decyzje" nr 1, s. 71-77.
Walesiak M. (1996): *Metody analizy danych marketingowych*. Warszawa: PWN.
Walesiak M. (1999): *Distance Measure for Ordinal Data*. „Argumenta Oeconomica" No. 2 (8), s. 167-173.
Walesiak M. (2000): *Propozycja uogólnionej miary odległości w statystycznej analizie wielowymiarowej*. Referat na Konferencję naukową nt. „Statystyka regionalna w służbie samorządu lokalnego i biznesu" (Kiekrz k. Poznania, 5-7 czerwca 2000 r.).
Walesiak M., Bąk A., Jajuga K. (2002): *Uogólniona miara odległości – badania symulacyjne*. W: *Klasyfikacja i analiza danych – teoria i zastosowania*. Red. K. Jajuga, M. Walesiak. Taksonomia 9. Prace Naukowe AE we Wrocławiu nr 942, s. 116-127.
Wedel M., Kamakura W.A. (1998): *Market Segmentation. Conceptual and Methodological Foundations*. Boston, Dordrecht, London: Kluwer Academic Publishers.
Zaborski A. (2001): *Skalowanie wielowymiarowe w badaniach marketingowych*. Wrocław: Wyd. AE we Wrocławiu.
Zakrzewska M. (1987): *O miarach podobieństwa obiektów i cech przydatnych w psychologicznych zastosowaniach analizy skupień* (rozdz. 7). W: *Wielozmienne modele statystyczne w badaniach psychologicznych*. Red. J. Brzeziński. Warszawa, Poznań: PWN.

SIMILARITY MEASURES FROM THE POINT OF VIEW SCALES OF MEASUREMENT AND VARIABLES WEIGHTS

Summary

The article contains widen version of earlier article of Walesiak [1994]. In the paper the following problems are added and discussed:

- the classification of distance measures in the light of measurement scales,
- the construction of distance measures with variable weights,
- the new distance measures.