

Justyna Wilk

Akademia Ekonomiczna we Wrocławiu

MIARY ODLEGŁOŚCI OBIEKTÓW OPISANYCH ZMIENNYMI SYMBOLICZNYMI Z WAGAMI

1. Wstęp

Wśród zmiennych symbolicznych (*symbolic variables*) szczególną rolę odgrywają zmienne symboliczne z wagami (*multi-valued modal variables*). Ich specyficzny charakter i nietypowa struktura uniemożliwiają zastosowanie klasycznych miar odległości, a także miar symbolicznych przeznaczonych do analizy podobieństwa obiektów opisanych pozostałymi typami zmiennych symbolicznych. Artykuł ma na celu omówienie (na przykładach empirycznych) specyfiki tego rodzaju zmiennych, zaprezentowanie miar odległości obiektów opisanych zmiennymi symbolicznymi z wagami (*probabilistic symbolic objects*), prezentację przykładu empirycznego wyznaczania odległości obiektów tego typu oraz przeprowadzenie klasyfikacji obiektów na podstawie uzyskanej macierzy odległości.

2. Specyfika zmiennych symbolicznych z wagami

Zmienne symboliczne z wagami charakteryzują się dwiema podstawowymi własnościami: pojedynczemu obiektowi jest przypisany więcej niż jeden wariant zmiennej, a wariantom przypisane są wagi, prawdopodobieństwa lub częstości [Brito, de Carvalho 2001, s. 13].

Zmienne te występują w praktyce ekonomicznej, a także mogą być efektem procesu agregacji obiektów symbolicznych pierwszego rzędu, opisanych zmiennymi „klasycznymi”¹ (zob. [Billard, Diday 2002, s. 8-14; Verde 2004, s. 304-305; Stephen, Hebrail, Lechevallier 2000, s. 78-105]). Mogą one przyjmować postać:

¹ „Zmienna klasyczna” to zmienna mierzona na mocnych lub słabych skalach pomiaru. Jest ona szczególnym przypadkiem zmiennej symbolicznej z wagami, gdzie obiekt ma jeden wariant zmiennej, o wadze równej 1.

- **przedziałów klasowych** (*interval-valued*), rozłącznych (*disjoint interval*) lub nierozłącznych (*non-disjoint interval*), zob. przykład w tab. 1.

Tabela 1. Rozkłady wag w poszczególnych grupach wiekowych kobiet

Nr	Wiek	Waga (kg)
1	20-29	[44,5-53,2) 8%; [53,2-58,6) 24%; [58,6-64,0) 30%; [64,0-69,4) 24%; [69,4-74,8) 6%; [74,8-82,0] 8%
2	30-39	[45,0-48,6) 2%; [48,6-52,2) 6%; [52,2-55,8) 40%; [55,8-59,4) 24%; [59,4-63,0) 24%; [63-67,5] 4%
3	40-49	[49,5-56,25) 4%; [56,25-60,75) 14%; [60,75-65,25) 20%; [65,25-69,75) 42%; [69,75-74,25) 14%; [74,25-83,25) 6%
4	50-59	[45,0-51,3) 4%; [51,3-56,7) 6%; [56,7-62,1) 20%; [62,1-67,5) 26%; [67,5-72,9) 28%; [72,9-85,5) 16%

Źródło: na podstawie [Billard, Diday 2003, s. 478].

Na przykład 8% kobiet w wieku 20-29 lat waży w granicach 44,5-53,2 kg.

- **list kategorii** (*set of categories*), o wariantach nominalnych (wielostanowych lub binarnych) lub porządkowych (zob. tab. 2),

Tabela 2. Struktura kadry naukowej uczelni X

Nr OS	Wydział	Kadra naukowa
1	A	profesor (13%), doktor hab. (30%), doktor (47%), doktorant (10%)
2	B	profesor (10%), doktor hab. (20%), doktor (55%), doktorant (8%), magister (7%)
3	C	profesor (16%), doktor hab. (30%), doktor (44%), magister (10%)

Źródło: opracowanie własne.

- **list wartości** (*set of values*), o wariantach ilorazowych, interwałowych, porządkowych.

Formalnie zmienna symboliczna z wagami Y_k (k – numer zmiennej, $k = 1, \dots, p$) ze zbioru Y (gdzie V_k zbiór realizacji zmiennej Y_k), jest odwzorowaniem $Y_k = (y_f, p(y_f))$ [Brito, Polaillon 2005, s. 79; Hardy, Lallemand 2004, s. 326], gdzie y_f wariant zmiennej Y_k ($f = 1, \dots, z$), a $p(y_f)$ i $q(y_f)$ to wagi wariantów zmiennej Y_k , odpowiednio dla obiektów symbolicznych o_i oraz o_j ($i, j = 1, \dots, n$),

gdzie $p(y_f) \in [0; 1]$, a $\sum_{f=1}^z p(y_f) = 1$. P i Q oznaczają rozkłady prawdopodobieństwa zmiennej Y_k , odpowiednio dla o_i oraz o_j . Na przykład, jeśli obiekt o_1 to kobiety w wieku 20-29 lat opisane zmienną Y_1 (waga w kg), to $P = \{[44,5-53,2) 8\%; [53,2-58,6) 24\%; [58,6-64,0) 30\%; [64,0-69,4) 24\%; [69,4-74,8) 6\%; [74,8-82,0] 8\%\}$ (zob. tab. 1). $d(P, Q)$ oznacza odległość obiektów o_i, o_j względem zmiennej Y_k .

3. Miary odległości zmiennych symbolicznych z wagami

Podobieństwo obiektów o_i oraz o_j względem zmiennej symbolicznej z wagami Y_k mierzy się przez porównanie rozkładów prawdopodobieństwa zmiennej dla poszczególnych obiektów, odpowiednio P i Q [Malerba, Esposito, Monopoli 2002, s. 33]. Dwa obiekty są więc tym bardziej podobne, im mniej różnią się co do wartości wag poszczególnych wariantów, tzn. im bardziej proporcjonalne są ich rozkłady prawdopodobieństwa P i Q .

Funkcja $d(P, Q)$ jest nazywana miarą odległości obiektów opisanych zmienną symboliczną z wagami wtedy, gdy spełnia podstawowe warunki (opracowanie z wykorzystaniem [Gatnar, Walesiak 2004, s. 39; Malerba, Esposito, Monopoli 2002]):

- 1) nieujemności: $d(P, Q) \geq 0$;
- 2) zwrotności: $d(P, Q) = 0 \Leftrightarrow P = Q$;²
- 3) symetryczności: $d(P, Q) = d(Q, P)$;

jeśli ponadto spełniony jest warunek:

- 4) nierówności trójkąta: $d(P, Q) \leq d(P, Z) + d(Q, Z)$,

to miara odległości zwana jest metryką.

Prezentowane miary odległości zmiennych (*comparison functions*) są w większości adaptacją miar sformułowanych dla rozkładów prawdopodobieństwa przez Csiszar [1967, s. 299-318] oraz Ali i Silvey [1966, s. 131-142] oraz klasycznych miar podobieństwa obiektów, a także miar wykorzystywanych w segmentacji obrazów (*image segmentation*) (zob. [Malerba, Esposito, Monopoli 2002, s. 33-35; Bock 2000, s. 153-165; ASSO 2004, s. 122-123; Rubner, Puzicha, Tomasi, Buhmann 2001; Bacelar-Nicolau 2000, s. 161]). Są to następujące miary:

- **wskaźnik KL** (*Kullback-Leibler divergence, I-coefficient, relative- lub cross entropy, information distance*) przyjmuje postać:

$$d_{KL}(P, Q) := \sum_{y \in V_k} q(y_f) \cdot \log \frac{q(y_f)}{p(y_f)}. \quad (1)$$

Jest to miara niesymetryczna, $d_{KL} \in [0; 1]$, nie istnieje dla $p(y_f) = 0$, tzn. nie może być wariantu o wadze równej 0. Rozwinięciem miary KL , spełniającym podstawowe założenia funkcji odległości jest **wskaźnik J** (*J-coefficient, Jeffrey divergence*). Przybiera on postać:

$$d_J(P, Q) := d_{KL}(P, Q) + d_{KL}(Q, P) \text{ lub} \quad (2)$$

$$d_J(P, Q) := \frac{d_{KL}(P, Q) + d_{KL}(Q, P)}{2}. \quad (3)$$

² W przypadku funkcji bliskości s , $P = Q$ dla $s(P, Q) = 1$.

Y. Rubner, J. Puzicha, C. Tomasi i J.N. Buchmann proponują operowanie średnią arytmetyczną zamiast samymi wagami wariantów i definiują wskaźnik (wzór 2) jako **wskaźnik J'**:

$$d_{J'}(P, Q) := \sum_{y \in V_k} q(y_f) \cdot \log \frac{q(y_f)}{\hat{e}(y_f)} + p(y_f) \cdot \log \frac{p(y_f)}{\hat{e}(y_f)}, \quad (4)$$

gdzie: $\hat{e}(y_f) = \frac{p(y_f) + q(y_f)}{2}$ to średnia arytmetyczna prawdopodobieństw wariantu y_f dla obu obiektów,

- **wskaźnik χ^2 (χ^2 divergence)** przybiera postać:

$$d_{\chi^2}(P, Q) := \sum_{y \in V_k} \frac{|p(y_f) - q(y_f)|^2}{p(y_f)}. \quad (5)$$

Wskaźnik χ^2 przyjmuje dodatnie wartości, tzn. $\chi^2 \geq 0$. Podobnie jak wskaźnik KL jest miarą niesymetryczną i nie istnieje dla $p(y_f) = 0$. Y. Rubner, J. Puzicha, C. Tomasi i J.N. Buchmann podają zmodyfikowaną wersję tego współczynnika, która redukuje ograniczenia χ^2 w postaci **miary χ_s^2** :

$$d_{\chi_s^2}(P, Q) = \sum_{y \in V_k} \frac{(p(y_f) - \hat{e}(y_f))^2}{\hat{e}(y_f)}, \quad (6)$$

gdzie: $\hat{e}(y_f) = \frac{p(y_f) + q(y_f)}{2}$ to średnia arytmetyczna prawdopodobieństw wariantu y_f dla obu obiektów.

Miara χ_s^2 spełnia podstawowe założenia funkcji odległości, przyjmuje wartości w przedziale $\chi_s^2 \in [0; 1]$ oraz istnieje dla $p(y_f) = 0$.

- **wskaźnik Hellingera (*Hellinger coefficient*)**:

$$d_H^{(s)}(P, Q) := \sum_{y \in V_k} p^{(1-s)}(y_f) \cdot q^s(y_f), \quad (7)$$

gdzie: $s \in (0; 1)$.

Jest to miara niesymetryczna, $d_H^{(s)}(P, Q) \in [0; 1]$. Jeśli $d_H^{(s)}(P, Q) = 0$, to obiekty są całkowicie niepodobne, natomiast gdy $d_H^{(s)}(P, Q) = 1 \Leftrightarrow P = Q$, obiekty są identyczne (proporcjonalne). Szczególnym przypadkiem miary Hellingera dla $s = \frac{1}{2}$ jest **miara Bhattacharyya (*Bhattacharyya coefficient*)**:

$$d_B^{1/2}(P, Q) := \sum_{y \in V_k} \sqrt{p(y_f) \cdot q(y_f)}. \quad (8)$$

Podobnie jak miara Hellingera $d_B^{1/2}(P, Q) = 1 \Leftrightarrow P = Q$, $d_B^{1/2}(P, Q) \in < 0; 1 >$. W przeciwieństwie do miernika Hellingera jest symetryczna.

Rozwinięciem miary Hellingera jest niesymetryczna miara odległości zmiennych – **wskaźnik Chernoffa** (*Chernoff's distance*):

$$d_{CH}^{(s)}(P, Q) := -\log d_H^{(s)}(P, Q). \quad (9)$$

Jest on mniej wrażliwy na wartości odbiegające (odstające) od pozostałych niż miara *KL*. Zwykle przyjmuje niższe wartości niż *KL*. Obiekty są identyczne, jeśli $d_B^{1/2}(P, Q) = 0$. Wskaźnik $d_C^{(s)}(P, Q)$ nie istnieje, gdy każdy z analizowanych obiektów ma wyłącznie jeden wariant zmiennej (o wadze równej 1), np. gdy $P = \{\text{zielony} (1,00), \text{niebieski} (0,0)\}$, a $Q = \{\text{zielony} (0,0), \text{niebieski} (1,0)\}$. Dla $s \leq 0$ lub $s \geq 1$ wskaźnik przyjmuje wartości ujemne.

H. Bacelar-Nicolau do pomiaru odległości proponuje sformułowany przez K. Matusita [1951, s. 1-30; 1955, s. 631-640] **wskaźnik affinity** (*affinity coefficient*).

Jest on rozwinięciem miary Bhattacharyya $d_B^{1/2}(P, Q)^2 = 2 \cdot (1 - d_A(P, Q))$:

$$d_A(P, Q) = 1 - \frac{\left(d_B^{1/2}(P, Q)\right)^2}{2}. \quad (10)$$

Jest symetryczny, zawiera się w przedziale $d_A(P, Q) \in [0; 1]$, $d_A(P, Q) = 1 \Leftrightarrow P = Q$, natomiast gdy $d_A(P, Q) = 0 \Leftrightarrow P \neq Q$. Dla zmiennych binarnych wskaźnik Affinity jest nazywany miarą Ochiai. Może być wykorzystywany, gdy w zbiorze, oprócz zmiennych symbolicznych z wagami, znajdują się zmienne innego typu, np. zmienne binarne. Kolejnym rozwinięciem miary Hellingera jest asymetryczny **wskaźnik Renyi** (*Renyi's divergence* lub *information gain*), dla którego $d_R^{(s)}(P, Q) = 0$ oznacza $P = Q$:

$$d_R^{(s)}(P, Q) := \frac{-\log d_H^{(s)}(P, Q)}{s - 1}, \quad (11)$$

- **odległość Minkowskiego L_w :**

$$d_w(P, Q) := \sqrt[w]{\sum_{y \in V_K} |p(y_f) - q(y_f)|^w}, \quad (12)$$

gdzie $w \geq 1$.

Odległość Minkowskiego jest metryką. Dla $w = 1$ odległość Minkowskiego jest nazywana odległością miejską (*Manhattan distance* lub *variation distance*), natomiast w przypadku zmiennych binarnych – odległością Hamminga. Gdy $w = 2$, L_w nazywana jest odległością Euklidesową (*Euclidean distance*).

- **współczynnik K (*K divergence*):**

$$d_K(P, Q) := \sum_{y \in V_K} q(y_f) \cdot \log \frac{q(y_f)}{\frac{1}{2} p(y_f) + \frac{1}{2} q(y_f)}. \quad (13)$$

Jest on miarą niesymetryczną, przyjmującą wartości większe od zera lub równe zero. Między współczynnikiem K a miarą KL zachodzi zależność $d_K(P, Q) := d_{KL}(P, \frac{1}{2}P + \frac{1}{2}Q)$, a między metryką Minkowskiego (dla $w=1$) $d_K(P, Q) \leq d_w(P, Q) \leq 2$. Nie istnieje, gdy $q(y) = 0$.

Zestawienie własności omówionych miar zawiera tab. 3.

Tabela 3. Charakterystyka omówionych miar odległości

Miara \ Własność	KL	J	J'	χ^2	χ_s^2	H	B	CH	A	R	L_w	K
Nieujemność	[0;1]	[0;1]	[0;1]	[0; ∞]	[0;1]	[0;1]	[0;1]	[0;1]	[0;1]	[0; ∞]	[0;1]	[0; ∞]
Symetryczność	–	+	+	–	+	–	+	–	+	–	+	–
Zwrotność	+	+	+	+	+	+	+	+	+	+	+	+
$P=Q$, gdy $d(P, Q) = \dots$	0	0	0	0	0	1	1	0	1	0	0	0
Funkcja odległości	+	+	+	+	+	–	–	+	–	+	+	+
Nierówność trójkąta	–	–	–	–	–	–	–	–	–	–	+	–
Analiza wielowymiarowa	+	+	+	+	+	+	+	+	+	+	+	+
nie istnieje, gdy ...	$p(y_i)=0$	$p(y_i)=0$	–	$p(y_i)=0$	–	–	–	zob. tekst	–	–	–	$q(y_i)=0$
Sodas v. 2.0 ³	–	+	–	+	–	–	–	+	–	+	+	–

Źródło: opracowanie własne z wykorzystaniem [Rubner, Puzicha, Tomasi, Buhmann 2001, s. 29-33; Malerba, Esposito, Monopoli 2002, s. 33-35; Bock 2000, s. 153-165; Bacelar-Nicolau 2000, s.161-162].

Niektóre zaprezentowane współczynniki, m.in. KL czy χ^2 , to miary bliskości. Sposób transformacji funkcji bliskości na funkcję odległości wyrażają formuły $d'(P, Q) := 1 - d(P, Q)$, $d'(P, Q) := \sqrt{1 - d(P, Q)}$, $d'(P, Q) := -\log d(P, Q)$ [Gatnar, Walesiak 2004, s. 39]. Nie wszystkie wymienione miary spełniają warunek symetryczności, m.in. miara Chernoff'a czy χ^2 . Można temu zaradzić poprzez założenie $d''(P, Q) = d(P, Q) + d(Q, P)$. Wyrażenie $d''(P, Q)$ spełnia wszystkie własności funkcji odległości (z wyjątkiem nierówności trójkąta) [Malerba, Esposito, Monopoli 2002, s. 35].

4. Miary odległości obiektów

Funkcja odległości d_{ij} obiektów symbolicznych o_i oraz o_j (*aggregation function*) jest zagregowaną wartością odległości zmiennych $d(P, Q)$, np. ważona odległość Minkowskiego [Malerba, Esposito, Monopoli 2002, s. 36]:

³ Program Sodas v. 2.0 można pobrać ze strony ASSO Project <http://www.info.fundp.ac.be/asso/sodaslink.htm>

$$d_{ij}^w = \sqrt[w]{\sum_{k=1}^p (\omega_k \cdot d(P, Q))^w}, \quad (14)$$

gdzie: ω_k to waga zmiennej Y_k ($k = 1, \dots, p$), $\omega_k \in (0; 1)$, $\sum_{k=1}^p \omega_k = 1$.

Modyfikacją miary d_{ij}^w jest normalizowana miara odległości:

$$d_{ij}^w = 1 - \frac{\prod_{f=1}^z \left[\sqrt[w]{2} - \sqrt[w]{\sum_{y_f} |p(y_f) - q(y_f)|^w} \right]}{(\sqrt[w]{2})^z} = 1 - \frac{\prod_{f=1}^z (\sqrt[w]{2} - \sqrt[w]{L_w})}{(\sqrt[w]{2})^z}. \quad (15)$$

5. Przykład empiryczny

Stopa bezrobocia⁴ w końcu lipca br. w Polsce kształtowała się na poziomie 17,9% (2.809,0 tys. osób⁵). Wielkość bezrobocia w poszczególnych województwach waha się pomiędzy 13,9 a 27,5%. Najwyższa stopa bezrobocia występowała w województwach warmińsko-mazurskim, zachodniopomorskim, lubuskim i kujawsko-pomorskim (22,5-27,5%), najniższa w małopolskim, mazowieckim, podlaskim i wielkopolskim (13,9-15,1%).

Obiektami poddanymi klasyfikacji będą województwa ze względu na wybrane zmienne. Zmienne poddane analizie zawiera tab. 4. Charakterystykę województw zawiera załącznik 1.

Tabela 4. Zmienne symboliczne z wagami i ich charakterystyka

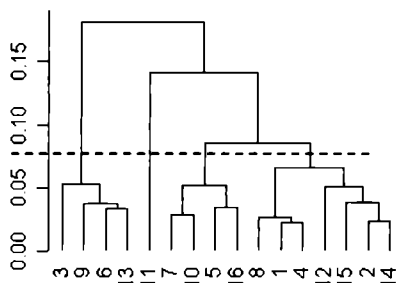
Nr	Zmienna	Warianty zmiennej z udziałem procentowym	Typ wariantów
Y ₁	Płeć	kobieta (50,68-58,17%), mężczyzna (41,83-49,32%)	Listy kategorii binarne
Y ₂	Grupy wiekowe (w latach)	[18-24] (19,72-26,83%), [25-34] (26,08-31,89%), [35-44] (19,00-23,19%), [45-54] (17,66-27,59%), [55-59] (2,29-5,00%), [60-64] (0,33-1,04%)	Przedziały klasowe rozłączne
Y ₃	Wykształcenie	wyższe (3,25-7,00%), policealne i średnie zawodowe (17,69-25,83%), średnie ogólnokształcące (6,54-8,84%), zasadnicze zawodowe (29,44-35,99%), gimnazjalne i poniżej (24,77-39,53%)	Listy kategorii nominalne
Y ₄	Staż pracy (w latach)	do 1 roku (10,72-18,10%), <1-5] (15,82-22,05%), <5-10] (10,84-15,56%), <10-20] (13,74-19,08%), <20-30] (7,55-14,21%), 30 i więcej (0,75-2,44%), bez stażu (18,28-33,86%)	Przedziały klasowe rozłączne
Y ₅	Czas pozostawania bez pracy (w miesiącach)	do 1 miesiąca (6,82-8,98%), <1-3] (8,30-10,94%), <3-6] (10,81-14,08%), <6-12] (15,90-18,42%), <12-24] (15,18-17,84%), 24 i więcej (30,48-40,79%)	Przedziały klasowe rozłączne

Źródło: opracowanie na podstawie danych [Statystyki...].

⁴ Stopa bezrobocia to procentowy udział liczby bezrobotnych w liczbie ludności aktywnej zawodowo.

⁵ Dane obejmują osoby zarejestrowane w urzędach pracy jako bezrobotne, zgodnie z definicją zawartą w Ustawie z 20 kwietnia 2004 r. o promocji zatrudnienia i instytucjach rynku pracy (DzU nr 99, poz. 1001 z późn. zm.).

Podobieństwo zmiennych i obiektów określono za pomocą odległości Euklidesowej (wzór 12), a następnie sklasyfikowano obiekty hierarchiczną aglomeracyjną metodą Warda [zob. Gordon 1999, s. 79, 81-82]. Zatrzymanie procesu aglomeracji na odległości wiązania równej 0,07 pozwoliło wyodrębnić cztery klasy⁶ obiektów (rys. 1).



Lp.	Województwo	Nr klasy	Lp.	Województwo	Nr klasy
1	Dolnośląskie	4	9	Podkarpackie	1
2	Kujawsko-Pomorskie	4	10	Podlaskie	3
3	Lubelskie	1	11	Pomorskie	2
4	Lubuskie	4	12	Śląskie	4
5	Łódzkie	3	13	Świętokrzyskie	1
6	Małopolskie	1	14	Warmińsko-Mazurskie	4
7	Mazowieckie	3	15	Wielkopolskie	4
8	Opolskie	4	16	Zachodniopomorskie	3

Rys. 1. Klasyfikacja obiektów metodą Warda na podstawie uzyskanej macierzy odległości (Załącznik 2)
Źródło: opracowanie własne w programie R 2.0.1, na podstawie danych z Załącznika 2.

W skupieniu 1 jest najwięcej osób, na tle innych klas, w wieku od 18 do 24 lat (nie mniej niż 23,91% bezrobotnych poszczególnego województwa), od 25 do 34 lat (minimum 28,5%), osób z wykształceniem wyższym (min. 5,45%) oraz policealnym i średnim zawodowym (min. 24,4%) i osób bez odbytego stażu pracy (min. 25,6%). **Klasa 2** notuje większość najwyższych wartości analizowanych wskaźników, tj. bezrobotnych kobiet (min. 55,11%), osób w wieku od 35 do 44 lat (min. 20,96%), od 45 do 54 (min. 23,9%), badanych z wykształceniem średnim ogólnokształcącym (min. 7,69%), zawodowym (min. 34,18%), gimnazjalnym i niżej (min. 34,18%), ze stażem pracy powyżej jednego roku i najdłuższym okresem pozostawania bez pracy od 1 do 24 miesięcy. **W skupieniu 3** 50% to województwa o najniższej stopie bezrobocia. Jest tu największy odsetek bezrobotnych mężczyzn (min. 46,79%), osób w wieku od 55 do 59 lat (min. 4,44%), od 60 do 64 lat (0,6%), osób ze stażem pracy nie przekraczającym jednego roku (co najmniej 14,9%) oraz pozostającym bez pracy powyżej 24 miesięcy (min. 35,8%). Do **klasy 4** należą województwa o największej stopie bezrobocia. Ta klasa skupia obiekty o relatywnie niskich wartościach rozpatrywanych kategorii. Wyprzedza inne klasy jedynie pod względem czasu pozostawania bez pracy do jednego miesiąca – min. 7,64%.

⁶ Sposoby interpretacji wyników klasyfikacji obiektów symbolicznych zob. [Brito 2004, s. 209, 211-213].

6. Podsumowanie

Przy wyznaczaniu odległości obiektów opisanych zmiennymi symbolicznymi z wagami należy zastosować miary odpowiednie dla rozkładów prawdopodobieństwa. Innym sposobem jest zamiana na zmienne „klasyczne”⁷ i zastosowanie „klasycznych” lub symbolicznych miar odległości [Bock, Diday 2000, s. 165-176].

Ze względu na sposób wyznaczania odległości, gdzie poszczególny przedział wartości jest traktowany jako kategoria zmiennej, problemem jest wyznaczenie podobieństwa obiektów opisanych zmiennymi w postaci nierozłącznych przedziałów klasowych. Rozwiązań można szukać w metodyce algebry przedziałowej.

Literatura

- Ali S.M., Silvey S.D. (1966), *A General Class of Coefficient of Divergence of One Distribution from Another*, „Journal of the Royal Statistical Society, Ser. B”, vol. 2.
- ASSO Developers Partners (2004), *Information Society Technologies Programme. User Manual for SODAS 2 Software*.
- Bacelar-Nicolau H. (2000), *The Affinity Coefficient*, [w:] H.H. Bock, E. Diday (Eds.), *Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data*, Springer-Verlag, Berlin, Heidelberg, s. 160-165.
- Billard L., Diday E. (2003), *From the Statistics of Data to the Statistic of Knowledge: Symbolic Data Analysis*, „Journal of the American Statistical Association”, June 2003, vol. 98, nr 470-487.
- Billard L., Diday E. (2002), *Symbolic Data Analysis: Definitions and Examples*, Ceremade, Dauphine.
- Bock H.H. (2000), *Dissimilarity Measures for Probability Distributions*, [w:] H.H. Bock, E. Diday (Eds.), *Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data*, Springer-Verlag, Berlin, Heidelberg, s. 153-165.
- Bock H.H., Diday E. (2000), *Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data*, Springer-Verlag, Berlin, Heidelberg.
- Brito P. (2004), *Clustering Interpretation. Interpreting Clusters by Using the Module CLINT*, [w:] ASSO Developers Partners, *Information Society Technologies Programme. User Manual for Sodas 2 Software*, s. 207-214.

⁷ Zamiana na zmienne „klasyczne” oznacza potraktowanie wariantów zmiennej pierwotnej jako „nowe” samodzielne zmienne, a ich wagi jako „nowe” warianty. obiektowi odpowiada jeden wariant „nowej” zmiennej.

- Brito P., de Carvalho F.A.T. (2001), *Symbolic Clustering of Constrained Probabilistic Data*, [w:] O. Opitz, M. Schwaiger (Eds.), *Exploratory Data Analysis in Empirical Research*, „Studies in Classification, Data Analysis and Knowledge Organization”, Springer-Verlag, Heidelberg, s. 12-22.
- Brito P., Polaillon G. (2005), *Structuring Probabilistic Data by Galois Lattices*, „Mathematics and Social Sciences” vol. 43, nr 169, s. 77-104.
- Csiszar I. (1967), *Information-type Measures of Difference of Probability Distributions and Indirect Observations*, „Studia Scient. Math. Hung.” vol. 2.
- Gatnar E., Walesiak M. (red.) (2004), *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*, AE, Wrocław.
- Gordon A.D. (1999), *Classification*, Chapman & Hall/CRC, London.
- Hardy A., Lallemand P. (2004), *Clustering of Symbolic Objects Described by Multi-Valued and Modal Variables*, [w:] D. Banks, L. House, F.R. McMorris, P. Arabie, W. Gaul (Eds.), *Clustering and Data Mining Applications*, „Studies in Classification, Data Analysis and Knowledge Organization”, Springer-Verlag, Berlin-Heidelberg, s. 325-332.
- Malerba D., Esposito F., Monopoli M. (2002), *Comparing Dissimilarity Measures for Probabilistic Symbolic Objects*, [w:] A. Zanasi, C.A. Brebbia, N.F.F. Ebecken, P. Melli (Eds.) *Data Mining III*, „Series Management Information Systems”, WIT Press, Southampton, UK, vol. 6, s. 31-40.
- Matusita K. (1951), *Decision Rules Based on Distance for Problems of Fit, Two Samples and Estimation*, „Ann. Math. Stat.” 3, s. 1-30.
- Matusita K. (1955), *On the Theory of Statistical Decision Functions*, „Ann. Math. Stat.” 26, s. 631-640.
- Rubner Y., Puzicha J., Tomasi C., Buhmann J.M. (2001), *Empirical Evaluation of Dissimilarity Measures for Color and Texture*, „Computer Vision and Image Understanding” 84, s. 25-43.
- Stephan V., Hebrail G., Lechevallier Y. (2000), *Generation of Symbolic Objects from Relational Databases*, [w:] H.H. Bock, E. Diday (Eds.), *Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data*, Springer-Verlag, Berlin, Heidelberg, s. 78-105.
- Statystyki rynku pracy, www.praca.gov.pl.
- Verde R. (2004), *Clustering Methods in Symbolic Data Analysis*, [w:] D. Banks, L. House, F.R. McMorris, P. Arabie, W. Gaul (Eds.), *Clustering and Data Mining Applications*, „Studies in Classification, Data Analysis and Knowledge Organization”, Springer-Verlag, Berlin-Heidelberg, s. 299-317.

Załącznik 1. Dane symboliczne z wagami, charakteryzujące województwa

	plec	wiek	wyższ
Dolnośl.	k (0.54), m (0.46)	18-24 (0.20), 25-34 (0.26), 35-44 (0.20), 45-54 (0.28), 55-59 (0.06), 60-64 (0.01)	wyższ (0.04), polc (0.22), sred. (0.07), zas. (0.32), gimn. (0.35)
Kuj.-Pom.	k (0.56), m (0.44)	18-24 (0.25), 25-34 (0.28), 35-44 (0.21), 45-54 (0.22), 55-59 (0.03), 60-64 (0.00)	wyższ (0.03), polc (0.18), sred. (0.07), zas. (0.35), gimn. (0.36)
Lubel.	k (0.51), m (0.49)	18-24 (0.26), 25-34 (0.32), 35-44 (0.19), 45-54 (0.19), 55-59 (0.03), 60-64 (0.00)	wyższ (0.07), polc (0.26), sred. (0.08), zas. (0.31), gimn. (0.28)
Lubus.	k (0.53), m (0.47)	18-24 (0.21), 25-34 (0.27), 35-44 (0.21), 45-54 (0.26), 55-59 (0.04), 60-64 (0.00)	wyższ (0.04), polc (0.21), sred. (0.07), zas. (0.34), gimn. (0.34)
Łódz.	k (0.51), m (0.49)	18-24 (0.21), 25-34 (0.27), 35-44 (0.20), 45-54 (0.26), 55-59 (0.05), 60-64 (0.01)	wyższ (0.05), polc (0.21), sred. (0.08), zas. (0.29), gimn. (0.36)
Małopol.	k (0.55), m (0.45)	18-24 (0.27), 25-34 (0.26), 35-44 (0.22), 45-54 (0.20), 55-59 (0.03), 60-64 (0.00)	wyższ (0.05), polc (0.26), sred. (0.08), zas. (0.36), gimn. (0.25)
Mazow.	k (0.51), m (0.49)	18-24 (0.21), 25-34 (0.26), 35-44 (0.20), 45-54 (0.25), 55-59 (0.05), 60-64 (0.01)	wyższ (0.06), polc (0.23), sred. (0.08), zas. (0.32), gimn. (0.32)
Opol.	k (0.56), m (0.44)	18-24 (0.21), 25-34 (0.26), 35-44 (0.22), 45-54 (0.25), 55-59 (0.05), 60-64 (0.01)	wyższ (0.04), polc (0.21), sred. (0.06), zas. (0.33), gimn. (0.36)
Podkarp.	k (0.54), m (0.46)	18-24 (0.25), 25-34 (0.32), 35-44 (0.23), 45-54 (0.18), 55-59 (0.02), 60-64 (0.00)	wyższ (0.06), polc (0.24), sred. (0.08), zas. (0.35), gimn. (0.27)
Podlas.	k (0.51), m (0.49)	18-24 (0.24), 25-34 (0.26), 35-44 (0.21), 45-54 (0.23), 55-59 (0.04), 60-64 (0.01)	wyższ (0.06), polc (0.24), sred. (0.09), zas. (0.30), gimn. (0.32)
Pomor.	k (0.55), m (0.45)	18-24 (0.23), 25-34 (0.26), 35-44 (0.21), 45-54 (0.24), 55-59 (0.04), 60-64 (0.00)	wyższ (0.42), polc (0.20), sred. (0.08), zas. (0.34), gimn. (0.34)
Ślask.	k (0.57), m (0.43)	18-24 (0.22), 25-34 (0.27), 35-44 (0.21), 45-54 (0.24), 55-59 (0.04), 60-64 (0.01)	wyższ (0.05), polc (0.22), sred. (0.07), zas. (0.33), gimn. (0.33)
Świętokrz.	k (0.51), m (0.49)	18-24 (0.24), 25-34 (0.30), 35-44 (0.20), 45-54 (0.22), 55-59 (0.04), 60-64 (0.00)	wyższ (0.07), polc (0.25), sred. (0.08), zas. (0.34), gimn. (0.26)
Warm.-Maz.	k (0.54), m (0.46)	18-24 (0.23), 25-34 (0.27), 35-44 (0.22), 45-54 (0.24), 55-59 (0.04), 60-64 (0.00)	wyższ (0.04), polc (0.20), sred. (0.07), zas. (0.31), gimn. (0.38)
Wielkop.	k (0.58), m (0.42)	18-24 (0.26), 25-34 (0.26), 35-44 (0.20), 45-54 (0.22), 55-59 (0.04), 60-64 (0.00)	wyższ (0.04), polc (0.23), sred. (0.07), zas. (0.35), gimn. (0.31)
Zachodniop.	k (0.53), m (0.47)	18-24 (0.20), 25-34 (0.27), 35-44 (0.21), 45-54 (0.27), 55-59 (0.05), 60-64 (0.01)	wyższ (0.04), polc (0.18), sred. (0.08), zas. (0.31), gimn. (0.40)

	staz	bez_pracy
Dolnośl.	do 1 (0.14), 1-5 (0.18), 5-10 (0.13), 10-20 (0.18), 20-30 (0.14), 30 i (0.02), bez s (0.21)	do 1 (0.08), 1-3 (0.10), 3-6 (0.13), 6-12 (0.18), 12-24 (0.17), pow.2 (0.34)
Kuj.-Pom.	do 1 (0.14), 1-5 (0.22), 5-10 (0.15), 10-20 (0.16), 20-30 (0.10), 30 i (0.01), bez s (0.22)	do 1 (0.08), 1-3 (0.09), 3-6 (0.13), 6-12 (0.18), 12-24 (0.17), pow.2 (0.36)
Lubel.	do 1 (0.16), 1-5 (0.17), 5-10 (0.11), 10-20 (0.14), 20-30 (0.08), 30 i (0.01), bez s (0.34)	do 1 (0.07), 1-3 (0.09), 3-6 (0.12), 6-12 (0.17), 12-24 (0.16), pow.2 (0.39)
Lubus.	do 1 (0.13), 1-5 (0.19), 5-10 (0.14), 10-20 (0.19), 20-30 (0.13), 30 i (0.02), bez s (0.20)	do 1 (0.08), 1-3 (0.11), 3-6 (0.14), 6-12 (0.18), 12-24 (0.17), pow.2 (0.31)
Łódz.	do 1 (0.15), 1-5 (0.19), 5-10 (0.14), 10-20 (0.17), 20-30 (0.13), 30 i (0.02), bez s (0.20)	do 1 (0.07), 1-3 (0.09), 3-6 (0.12), 6-12 (0.17), 12-24 (0.17), pow.2 (0.37)
Małopol.	do 1 (0.14), 1-5 (0.19), 5-10 (0.14), 10-20 (0.17), 20-30 (0.10), 30 i (0.01), bez s (0.26)	do 1 (0.08), 1-3 (0.09), 3-6 (0.13), 6-12 (0.18), 12-24 (0.17), pow.2 (0.35)
Mazow.	do 1 (0.16), 1-5 (0.18), 5-10 (0.13), 10-20 (0.15), 20-30 (0.10), 30 i (0.02), bez s (0.26)	do 1 (0.07), 1-3 (0.09), 3-6 (0.12), 6-12 (0.16), 12-24 (0.17), pow.2 (0.39)
Opol.	do 1 (0.11), 1-5 (0.20), 5-10 (0.14), 10-20 (0.19), 20-30 (0.14), 30 i (0.02), bez s (0.19)	do 1 (0.08), 1-3 (0.10), 3-6 (0.13), 6-12 (0.17), 12-24 (0.17), pow.2 (0.34)
Podkarp.	do 1 (0.14), 1-5 (0.20), 5-10 (0.13), 10-20 (0.16), 20-30 (0.08), 30 i (0.01), bez s (0.29)	do 1 (0.08), 1-3 (0.08), 3-6 (0.12), 6-12 (0.16), 12-24 (0.15), pow.2 (0.41)
Podlas.	do 1 (0.16), 1-5 (0.19), 5-10 (0.12), 10-20 (0.15), 20-30 (0.10), 30 i (0.01), bez s (0.27)	do 1 (0.08), 1-3 (0.10), 3-6 (0.13), 6-12 (0.17), 12-24 (0.15), pow.2 (0.36)
Pomor.	do 1 (0.12), 1-5 (0.21), 5-10 (0.16), 10-20 (0.19), 20-30 (0.12), 30 i (0.02), bez s (0.18)	do 1 (0.07), 1-3 (0.09), 3-6 (0.13), 6-12 (0.17), 12-24 (0.18), pow.2 (0.35)
Ślask.	do 1 (0.14), 1-5 (0.16), 5-10 (0.13), 10-20 (0.17), 20-30 (0.12), 30 i (0.02), bez s (0.27)	do 1 (0.01), 1-3 (0.10), 3-6 (0.14), 6-12 (0.17), 12-24 (0.17), pow.2 (0.34)
Świętokrz.	do 1 (0.13), 1-5 (0.19), 5-10 (0.13), 10-20 (0.15), 20-30 (0.10), 30 i (0.01), bez s (0.29)	do 1 (0.07), 1-3 (0.08), 3-6 (0.11), 6-12 (0.18), 12-24 (0.18), pow.2 (0.38)
Warm.-Maz.	do 1 (0.14), 1-5 (0.20), 5-10 (0.15), 10-20 (0.16), 20-30 (0.11), 30 i (0.01), bez s (0.23)	do 1 (0.08), 1-3 (0.10), 3-6 (0.12), 6-12 (0.18), 12-24 (0.17), pow.2 (0.36)
Wielkop.	do 1 (0.14), 1-5 (0.21), 5-10 (0.14), 10-20 (0.17), 20-30 (0.11), 30 i (0.02), bez s (0.21)	do 1 (0.08), 1-3 (0.10), 3-6 (0.14), 6-12 (0.18), 12-24 (0.17), pow.2 (0.32)
Zachodniop.	do 1 (0.18), 1-5 (0.19), 5-10 (0.13), 10-20 (0.15), 20-30 (0.12), 30 i (0.01), bez s (0.23)	do 1 (0.07), 1-3 (0.09), 3-6 (0.12), 6-12 (0.18), 12-24 (0.18), pow.2 (0.36)

Źródło: opracowanie na podstawie danych: [Statystyki ...].

Załącznik 2. Macierz odległości obiektów, na podstawie danych z Załącznika 1

	Dolno- śląskie	Kujaw- pomor.	Lubel- skie	Lubuskie	Łódzkie	Mato- polskie	Ma- zow.	Opol- skie	Pod- kar- packie	Podla- skie	Pomor- skie	Śląskie	Święto- krzyskie	Warm- mazur.	Wielko- polskie	Zachod- nopomo- rskie
Dolnośląskie	0															
Kujawsko-pomorskie	0,0451	0														
Lubelskie	0,0858	0,0830	0													
Lubuskie	0,0225	0,0407	0,0951	0												
Łódzkie	0,0289	0,0448	0,0791	0,0374	0											
Małopolskie	0,0620	0,0494	0,0551	0,0581	0,0653	0										
Mazowieckie	0,0423	0,0509	0,0535	0,0547	0,0342	0,0508	0									
Opolskie	0,0250	0,0415	0,1031	0,0259	0,0361	0,0663	0,0586	0								
Podkarpackie	0,0781	0,0544	0,0434	0,0786	0,0752	0,0374	0,0549	0,0795	0							
Podlaskie	0,0493	0,0552	0,0475	0,0590	0,0425	0,0472	0,0286	0,0642	0,0521	0						
Pomorskie	0,0939	0,0900	0,1351	0,0869	0,0959	0,1158	0,1016	0,0877	0,1179	0,1111	0					
Śląskie	0,0379	0,0467	0,0726	0,0455	0,0491	0,0492	0,0460	0,0481	0,0666	0,0471	0,0970	0				
Świętokrzyskie	0,0633	0,0574	0,0424	0,0640	0,0591	0,0331	0,0372	0,0708	0,0343	0,0392	0,1098	0,0565	0			
Warmińsko-mazurskie	0,0364	0,0232	0,0764	0,0394	0,0342	0,0593	0,0403	0,0389	0,0615	0,0449	0,0955	0,0450	0,0578	0		
Wielkopolskie	0,0426	0,0274	0,0789	0,0392	0,0546	0,0365	0,0543	0,0433	0,0571	0,0527	0,0966	0,0406	0,0550	0,0417	0	
Zachodniopomorskie	0,0370	0,0450	0,0813	0,0477	0,0338	0,0753	0,0401	0,0509	0,0821	0,0497	0,1071	0,0527	0,0689	0,0304	0,0598	0

Źródło: opracowanie własne w programie Sodas v. 2.0.

DISSIMILARITY MEASURES FOR PROBABILISTIC SYMBOLIC OBJECTS

Summary

The paper presents the way how to measure distance between probabilistic symbolic objects (objects described by modal variables). Firstly there was presented an idea and the types of modal variables (with examples) and then symbolic dissimilarity measures suggested to analyze. Finally the example of matching distance between probabilistic symbolic objects was illustrated and then cluster analysis was applied.