

Andrzej Dudek

MIARY PODOBIENSTWA OBIEKTÓW SYMBOLICZNYCH. MIARY DE CARVALHO

1. Wstęp

W wielu współczesnych zagadnieniach klasyfikacyjnych tradycyjne podejście do reprezentowania danych w postaci macierzy lub wielowymiarowej tablicy liczb przestaje być wystarczające, a coraz więcej autorów postuluje uwzględnianie w badaniach również danych nie mających reprezentacji numerycznej. Jedną z dziedzin taksonomii, w których tak się postępuje, jest symboliczna analiza danych będąca rozszerzeniem wielowymiarowej analizy statystycznej na dane reprezentowane w postaci tablic obiektów symbolicznych. Każda komórka w obiekcie symbolicznym (zmienna symboliczna) może zawierać dane w postaci pojedynczych wartości liczbowych, przedziałów liczbowych, zbiorów danych, zbiorów danych wraz z wagami lub wartości tekstowych [1].

Do mierzenia odległości między danymi symbolicznymi, ze względu na ich postać, nie mogą być użyte miary stosowane w wielowymiarowej analizie statystycznej, takie jak odległość Minkowskiego, euklidesowa, Mahalanobisa, Canberra, Clarka czy odległość GDM (por. [12, s. 24]). Zamiast tego stosowane są inne miary, z których nie wszystkie spełniają postulaty metryki, dlatego w przypadku obiektów symbolicznych zamiast pojęcia *miary odległości* częściej używa się pojęcia *miary podobieństwa* (lub analogicznie *miary niepodobieństwa*).

Ze względu na złożoność tematu mierzenia podobieństwa obiektów symbolicznych próba jego przedstawienia w ramach jednego artykułu mogłaby być nieskuteczna z powodu braku miejsca. Dlatego opis tematu został podzielony na dwie części zgodnie z podziałem znanym z literatury przedmiotu [10, s. 480; 1, s. 220; 13]. W pierwszej części [5] przedstawiono najważniejsze problemy związane z mierzeniem podobieństwa, najważniejsze grupy miar podobieństwa obiektów

symbolicznych oraz zaprezentowano miary Ichino i Yaguchiego [9] wraz z przykładami dla różnych typów zmiennych symbolicznych.

Artykuł niniejszy stanowi drugą część prezentacji miar podobieństwa obiektów symbolicznych, poświęconą grupie miar zaproponowanych przez de Carvalho. Przedstawione zostaną pojęcia operatora połączenia, operatora przekroju, funkcji porównującej, funkcji agregującej i potencjału obiektu wraz z przykładami, przedyskutowane własności metryczne miar de Carvalho oraz określone najważniejsze zastosowania tych miar wraz z przykładem klasyfikacji danych symbolicznych metodą Warda z wykorzystaniem miar de Carvalho.

2. Obiekt symboliczny i zmienna symboliczna

Przedmiotem symbolicznej analizy danych są obiekty symboliczne, czyli te obiekty, w których poszczególne składowe (zienne symboliczne) mogą być reprezentowane w postaci [1]:

- danych numerycznych,
- łańcuchów tekstowych (danych jakościowych),
- przedziałów liczbowych,
- danych w postaci listy wartości,
- danych w postaci listy wartości z wagami.

Zmienne w obiektach symbolicznych mogą ponadto [3]:

- reprezentować strukturę hierarchiczną (*taxonomic variable*),
- być hierarchicznie zależne – jeśli jedna ze zmiennych jest określona tylko wtedy, gdy inna zmienna przyjmuje określoną wartość,
- być związane zależnościami logicznymi – jeśli między wartościami zmiennych muszą zachodzić określone związki, aby dane były poprawne.

Ponadto każda ze zmiennych symbolicznych może przybierać wartość *NA* oznaczającą, że wartość tej zmiennej nie jest określona.

Struktura danych reprezentowanych w obiektach symbolicznych implikuje fakt, iż do pomiaru ich podobieństwa nie można stosować klasycznych miar odległości. Zamiast tego stosowane są inne miary [1]:

- Gowdy i Krishny – miara wzajemnego sąsiedztwa (*mutual neighborhood value*);
- Ichino i Yaguchiego – miara oparta na pojęciach operatorów połączenia (*Cartesian join*) i przekroju (*Cartesian meet*), będącymi rozszerzeniami operatorów $\cup$ i $\cap$ na wszystkie typy danych w obiektach symbolicznych;
- de Carvalho – rozszerzenie miar Ichino i Yaguchiego oparte na pojęciach funkcji porównującej (CF – *comparison function*) funkcji agregującej (AF – *aggregation function*) oraz potencjału obiektu (*description potential*).

3. Miary de Carvalho oparte na pojęciach funkcji porównującej i funkcji agregującej

Pojęcie operatora połączenia (*Cartesian join*) i operatora przekroju (*Cartesian meet*) wprowadzone zostało przez Ichino i Yaguchiego [9]. Operatory te można traktować jako rozszerzenie znanych z rachunku zbiorów operatorów sumy zbiorów $\cup$ i przekroju zbiorów $\cap$ na wszystkie typy danych określonych dla zmiennych symbolicznych.

3.1. Operator połączenia

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

Dla danych **numerycznych** (liczbowych) operator połączenia jest zdefiniowany zgodnie ze wzorem

$$A_j \oplus B_j = \begin{cases} (A_j, B_j) \Leftrightarrow A_j < B_j, \\ (B_j, A_j) \Leftrightarrow A_j \geq B_j, \end{cases} \quad (1)$$

gdzie A_j, B_j oznaczają zmienne numeryczne, (A_j, B_j) – przedział liczbowy.

Dla danych w postaci **przedziału liczbowego** operator połączenia jest zdefiniowany zgodnie ze wzorem

$$(a_1, a_2) \oplus (b_1, b_2) = (\min(a_1, b_1), \max(a_2, b_2)), \quad (2)$$

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

Dla danych o **strukturze hierarchicznej** operator połączenia jest zdefiniowany zgodnie ze wzorem

$$A_j \oplus B_j = \begin{cases} A_j \cup B_j \Leftrightarrow N(A_j) = N(B_j), \\ L(N(B_j \cup A_j)) \Leftrightarrow N(A_j) \neq N(B_j), \end{cases} \quad (3)$$

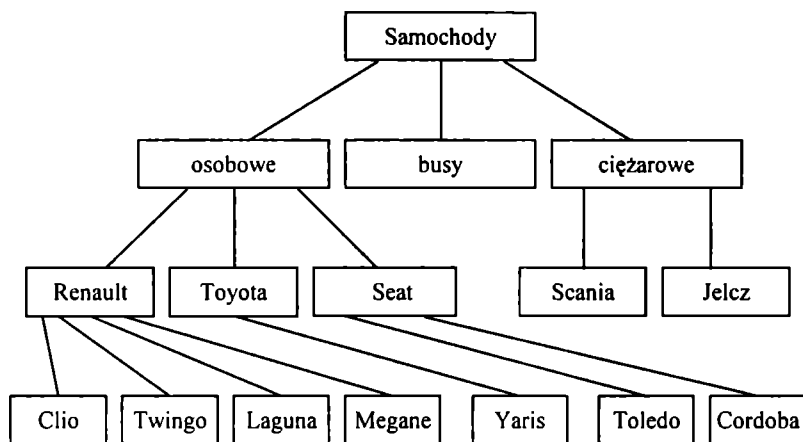
gdzie A_j, B_j oznaczają zmienne o strukturze hierarchicznej, $N(.)$ – najniżej położony element drzewa, który jest rodzicem dla wszystkich elementów zmiennej, $L(.)$ – zbiór wszystkich liści drzewa położonych poniżej węzła.

Przykład 1

Na rys. 1 przedstawiona jest struktura hierarchiczna związana ze zmienną *marka_samochodu*.

Niech: zbiór $A = \{\text{Twingo, Clio}\}$; zbiór $B = \{\text{Megane, Laguna}\}$.

Ponieważ $N(A) = N(B)$, więc zgodnie ze wzorem 3 operator połączenia A i B jest równy: $A \oplus B = \{\text{Twingo, Clio, Megane, Laguna}\}$.



Rys. 1. Struktura hierarchiczna zmiennej obiektu symbolicznego

Źródło: opracowanie własne.

Przykład 2

Niech: zbiór $C = \{\text{Twingo, Clio}\}$; zbiór $D = \{\text{Megane, Toledo}\}$.

Podobnie jak w przykładzie 1 strukturę hierarchiczną związaną ze zmienną *marka_samochodu* przedstawiono na rys. 1.

Ponieważ $N(C)$ jest różny od $N(D)$, a $N(C \cup D) = \{\text{osobowe}\}$, więc zgodnie ze wzorem (3) wartością połączenia zbiorów C i D jest: $C \oplus D = \{\text{Twingo, Clio, Megane, Laguna, Yaris, Toledo, Cordoba}\}$.

Dla pozostałych typów danych operator połączenia jest zdefiniowany zgodnie ze wzorem

$$A \oplus B = A \cup B, \quad (4)$$

gdzie A, B oznaczają zmienne w postaci zbioru wartości numerycznych lub nienu-
merycznych.

3.2. Operator przekroju

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

$$A_j \otimes B_j = A_j \cap B_j, \quad (5)$$

gdzie A_j, B_j oznaczają zmienne symboliczne dowolnego typu.

De Carvalho [2], opierając się na pojęciach operatora połączenia i operatora przekroju, wprowadza pojęcia funkcji porównującej (CF – *comparison function*) i funkcji agregującej (AF – *aggregation function*).

3.3. Funkcja porównująca

Funkcja porównująca mierzy podobieństwo zmiennych symbolicznych. De Carvalho definiuje pięć typów funkcji porównującej, a do ich zdefiniowania wykorzystuje wskaźniki zgodności (niezgodności) zmiennych zgodnie z tab. 1.

Tabela 1. Wskaźniki zgodności i niezgodności de Carvalho

	Zgodność	Niezgodność	Razem
Zgodność	$\alpha = \mu(A_j \cap B_j)$	$\beta = \mu(A_j \cap c(B_j))$	$\mu(A_j)$
Niezgodność	$\chi = \mu(c(A_j) \cap B_j)$	$\delta = \mu(c(A_j) \cap c(B_j))$	$\mu(c(A_j))$
Razem	$\mu(B_j)$	$\mu(c(B_j))$	

$\mu(A_j) = \begin{cases} \text{długość przedziału} & \text{dla zmiennych w postaci przedziałów liczbowych,} \\ |A_j| & \text{dla pozostałych typów zmiennych,} \end{cases}$

$c(A_j), c(B_j)$ – dopełnienie zmiennej względem jej dziedziny.

Źródło: [1, s. 174].

Jako funkcja porównująca może być użyta jedna z funkcji określona wzorami:

$$S_1 = \frac{\alpha}{\alpha + \beta + \chi}, \quad (6)$$

$$S_2 = \frac{2\alpha}{2\alpha + \beta + \chi}, \quad (7)$$

$$S_3 = \frac{\alpha}{\alpha + 2(\beta + \chi)}, \quad (8)$$

$$S_4 = \frac{1}{2} \cdot \left[\frac{\alpha}{\alpha + \beta} + \frac{\alpha}{\alpha + \chi} \right], \quad (9)$$

$$S_5 = \frac{\alpha}{\sqrt{(\alpha + \beta)(\alpha + \chi)}}, \quad (10)$$

gdzie $\alpha, \beta, \chi, \delta$ oznaczają wskaźniki zgodności i niezgodności zmiennych.

Każdej funkcji porównującej S_i określającej podobieństwo zmiennych symbolicznych odpowiada miara niepodobieństwa zmiennych zgodnie ze wzorem

$$d_i = 1 - S_i, \text{ dla } i = 1, 2, 3, 4, 5. \quad (11)$$

3.4. Funkcja agregująca

Dla miar niepodobieństwa zmiennych zdefiniowanych w (11) funkcja agregująca określa miary niepodobieństwa obiektów. Wzór (12) definiuje miary niepodobieństwa obiektów O_1 i O_2 :

$$d_q^i(O_1, O_2) = \sqrt[q]{\sum_{j=1}^q [w_j d_i(A_j, B_j)]^q} \quad \text{dla } i = 1, 2, 3, 4, 5, \quad (12)$$

gdzie $O_1 = (A_1, A_2, \dots, A_p)$, $O_2 = (B_1, B_2, \dots, B_p)$ oznaczają obiekty symboliczne składające się z p zmiennych, $j = 1, 2, \dots, p$ – numer zmiennej, w_j – wagę związaną z j -tą zmienną symboliczną, $d_i(A_j, B_j)$ jest zdefiniowane jak w (11), q – liczbę całkowitą dodatnią.

4. Miary de Carvalho oparte na pojęciu potencjału obiektu

De Carvalho [2] proponuje również inny sposób mierzenia niepodobieństwa obiektów symbolicznych wykorzystujący pojęcie potencjału obiektu (*description potential*).

4.1. Potencjał obiektu

Dla obiektu symbolicznego zawierającego zmienne dowolnych typów jego potencjał jest zdefiniowany zgodnie ze wzorami:

$$\pi(O_1) = \prod_{j=1}^p \mu(A_j), \quad (13)$$

$$\pi(O_2) = \prod_{j=1}^p \mu(B_j), \quad (14)$$

gdzie $\pi(O_1)$, $\pi(O_2)$ oznaczają potencjał obiektu O_1 i O_2 , $O_1 = (A_1, A_2, \dots, A_p)$, $O_2 = (B_1, B_2, \dots, B_p)$ – obiekt symboliczny opisany p zmiennymi,

$$\mu(A_j) = \begin{cases} \text{długość przedziału} & \text{dla zmiennych w postaci przedziałów liczbowych} \\ |A_j| & \text{dla pozostałych typów zmiennych,} \end{cases}$$

$||$ – symbol oznaczający dla danych w postaci przedziału liczbowego jego długość, dla pozostałych danych liczba elementów zbioru.

Tabela 2. Obiekty symboliczne ze zmiennymi różnych typów

O_1	A_1	A_2	A_3
	$\{a, b\}$	10,2	$\langle 2, 5 \rangle$
O_2	B_1	B_2	B_3
	$\{b, c, d\}$	8,5	$\langle 4, 8 \rangle$

Źródło: opracowanie własne.

Przykład 3

W tabeli 2 przedstawione są przykładowe obiekty symboliczne. Wzory (15) i (16) przedstawiają ich potencjały.

$$\pi(O_1) = 2 \cdot 1 \cdot 3 = 6, \quad (15)$$

$$\pi(O_2) = 3 \cdot 1 \cdot 4 = 12. \quad (16)$$

4.2. Operator połączenia i operator przekroju dla obiektów

Działanie operatorów połączenia i przekroju zmiennych można rozszerzyć na obiekty symboliczne zgodnie z definicjami (17) i (18)

$$O_1 \oplus O_2 = (A_1 \oplus B_1, A_2 \oplus B_2, \dots, A_p \oplus B_p), \quad (17)$$

$$O_1 \otimes O_2 = (A_1 \otimes B_1, A_2 \otimes B_2, \dots, A_p \otimes B_p), \quad (18)$$

gdzie $O_1 = (A_1, A_2, \dots, A_p)$, $O_2 = (B_1, B_2, \dots, B_p)$ oznaczają obiekty symboliczne opisane p zmiennymi.

Wzór (19) przedstawia postać nieznormalizowaną miary de Carvalho, wzory (20) i (21) przedstawiają postacie znormalizowane:

$$d_1''(O_1, O_2) = [\pi(O_1 \oplus O_2) - \pi(O_1 \otimes O_2) + \gamma \cdot (2\pi(O_1 \otimes O_2) - \pi(O_1) - \pi(O_2))], \quad (19)$$

$$d_2''(O_1, O_2) = \frac{\pi(O_1 \oplus O_2) - \pi(O_1 \otimes O_2) + \gamma \cdot (2\pi(O_1 \otimes O_2) - \pi(O_1) - \pi(O_2))}{\pi(D^E)}, \quad (20)$$

$$d_3''(O_1, O_2) = \frac{\pi(O_1 \oplus O_2) - \pi(O_1 \otimes O_2) + \gamma \cdot (2\pi(O_1 \otimes O_2) - \pi(O_1) - \pi(O_2))}{\pi(O_1 \oplus O_2)}, \quad (21)$$

gdzie O_1, O_2 oznaczają obiekty symboliczne, D^E – obiekt maksymalny w analizowanej dziedzinie.

5. Właściwości metryczne miar de Carvalho

Pojęcie miary niepodobieństwa obiektów symbolicznych nie jest równoważne z pojęciem metryki, gdyż nie wszystkie właściwości są dla nich spełnione. W tab. 3 przedstawiono właściwości metryczne dla odległości de Carvalho zdefiniowanych w (12), (19), (20), (21).

Tabela 3. Właściwości metryczne miar de Carvalho

Odległość	Numer wzoru	Nieujemność	Zwrotność	Symetryczność	Nierówność trójkąta
d_1	(6), (11), (12)	+	+	+	+
d_2	(7), (11), (12)	+	+	+	–
d_3	(8), (11), (12)	+	+	+	+
d_4	(9), (11), (12)	+	+	+	–
d_5	(10), (11), (12)	+	+	+	–
d'_1	(19)	+	+	+	–
d'_2	(20)	+	+	+	–
d'_3	(21)	+	+	+	+

Źródło: opracowanie własne na podstawie: [1; 2; 10; 12].

6. Zastosowanie miar de Carvalho

Najważniejsze zastosowania miar de Carvalho można podzielić na dwie grupy:

- wykorzystanie miar niepodobieństwa de Carvalho w metodach analizy danych symbolicznych,
- wykorzystanie miar niepodobieństwa de Carvalho w „tradycyjnych” metodach klasyfikacyjnych opartych na macierzach odległości.

W pierwszej grupie zastosowań należy wymienić takie zastosowania miar niepodobieństwa de Carvalho, jak:

- analiza dyskryminacyjna obiektów symbolicznych [1, s. 234-293],
- drzewa klasyfikacyjne dla obiektów symbolicznych [4],
- metody dynamicznej klasyfikacji dla obiektów symbolicznych [11].

Druga grupa zastosowań jest związana z wykorzystaniem miar de Carvalho do stworzenia danych wejściowych dla hierarchicznych metod klasyfikacyjnych opartych na macierzy odległości, a w szczególności dla metod: pojedynczego połączenia, kompletnego połączenia, średniej klasowej, średniej odległości, Warda (por. np. [7, s. 326; 8, s. 79]).

Przykład 4

W tab. 4 przedstawiono dane symboliczne o francuskich winach. Dane te pochodzą z repozytorium danych symbolicznych dla pakietu SODAS i są dostępne pod adresem www.ceremade.dauphine.fr/~touati/sodas-pagegarde.htm.

Kolumna 1 określa, czy dane wino było klasyfikowane w rankingach branżowych, kolumna 2 przedstawia wynik tej klasyfikacji, kolumny 3-14 przedstawiają cechy smakowe win według ekspertów, kolumna 15 – podobieństwo do znanych rodzajów win a kolumna 16 – obszar wytwarzania.

Obliczoną macierz odległości d_3 de Carvalho przedstawiono w tab. 5.

Tabela 4. Obiekty symboliczne dotyczące piętnastu gatunków francuskich win

Lp.	Classification	Classement	AC	BY	CG	CQ	DG	EZ	FT	HY
1	1: no(1)	NA	[56; 74]	[83; 85]	[84; 90]	[80; 91]	[70; 80]	[76; 85]	[35; 70]	[48; 72]
2	2: yes(1)	1: 1er Grand Cru(1)	[75; 92]	[89; 94]	[86; 92]	[85; 93]	[75; 90]	[76; 88]	[40; 78]	[53; 91]
3	2: yes(1)	2: Grand Cru(1)	[83; 90]	[83; 89]	[87; 93]	[85; 92]	[75; 95]	[82; 93]	[58; 90]	[51; 88]
4	2: yes(1)	2: Grand Cru(1)	[47; 82]	[48; 87]	[82; 86]	[73; 76]	[65; 85]	[76; 89]	[40; 50]	[69; 80]
5	2: yes(1)	1: 1er Grand Cru(1)	[68; 86]	[84; 91]	[88; 95]	[85; 92]	[75; 85]	[78; 92]	[68; 82]	[55; 88]
6	2: yes(1)	1: 1er Grand Cru(1)	[64; 82]	[81; 92]	[87; 90]	[85; 91]	[70; 95]	[85; 90]	[75; 78]	[77; 89]
7	1: no(1)	NA	[63; 83]	[89; 91]	[86; 90]	[75; 88]	[67; 87]	[79; 87]	[30; 45]	[63; 78]
8	2: yes(1)	2: Grand Cru(1)	[70; 84]	[75; 93]	[87; 93]	[80; 93]	[80; 95]	[88; 89]	[65; 90]	[51; 85]
9	2: yes(1)	2: Grand Cru(1)	[66; 91]	[69; 88]	[82; 88]	[71; 86]	[80; 92]	[82; 87]	[70; 85]	[31; 87]
10	1: no(1)	NA	[42; 90]	[81; 91]	[86; 90]	[81; 91]	[60; 90]	[85; 86]	[30; 89]	[62; 83]
11	1: no(1)	NA	[61; 93]	[85; 87]	[85; 90]	[83; 89]	[77; 97]	[76; 82]	[80; 89]	[62; 65]
12	1: no(1)	NA	[65; 83]	[80; 86]	[88; 92]	[92; 92]	[65; 85]	[85; 90]	[70; 88]	[62; 90]
13	2: yes(1)	3: Cru(1)	[62; 81]	[83; 90]	[88; 91]	[84; 91]	[80; 92]	[83; 87]	[60; 74]	[38; 78]
14	2: yes(1)	2: Grand Cru(1)	[69; 86]	[79; 94]	[85; 92]	[79; 90]	[80; 95]	[73; 85]	[45; 87]	[33; 53]
15	2: yes(1)	1: 1er Grand Cru(1)	[84; 88]	[80; 86]	[90; 92]	[88; 92]	[80; 95]	[87; 92]	[75; 83]	[72; 87]

IK	JY	KO	LP	Cepage	Area
[84; 98]	[70; 85]	[60; 70]	[65; 84]	1: cabernet_franc(0.5) 2: merlot(0.5)	1 : saint emilion
[94; 98]	[78; 93]	[78; 88]	[84; 92]	1: cabernet_franc(0.4) 2: merlot(0.6)	1 : saint emilion
[89; 98]	[78; 88]	[75; 92]	[86; 91]	2: merlot(0.4) 3: cabernet_sauvignon(0.6)	2 : saint estephe
[82; 90]	[59; 83]	[40; 81]	[80; 92]	1: cabernet_franc(0.05) 2: merlot(0.25) 3: cabernet_sauvignon(0.65) 4: petit_verdot(0.05)	3 : saint julien
[93; 98]	[82; 93]	[65; 90]	[90; 93]	1: cabernet_franc(0.25) 2: merlot(0.25) 3: cabernet_sauvignon(0.5)	4 : graves
[94; 98]	[73; 85]	[71; 83]	[78; 88]	1: cabernet_franc(0.04) 2: merlot(0.2) 3: cabernet_sauvignon(0.75)	5 : pauillac
[89; 96]	[84; 89]	[51; 84]	[52; 89]	1: cabernet_franc(0.5) 2: merlot(0.5)	6 : pomerol
[95; 96]	[71; 88]	[73; 84]	[78; 92]	1: cabernet_franc(0.2) 2: merlot(0.35) 3: cabernet_sauvignon(0.4) 4: petit_verdot(0.05)	7 : haut medoc
[91; 96]	[73; 76]	[58; 83]	[84; 94]	1: cabernet_franc(0.13) 2: merlot(0.19) 3: cabernet_sauvignon(0.65)	3 : saint julien
[89; 97]	[79; 93]	[56; 90]	[85; 92]	2: merlot(0.71) 3: cabernet_sauvignon(0.29)	6 : pomerol
[90; 98]	[79; 89]	[79; 90]	[87; 91]	1: cabernet_franc(0.1) 2: merlot(0.15) 3: cabernet_sauvignon(0.75)	5 : pauillac
[85; 99]	[72; 89]	[77; 85]	[89; 92]	1: cabernet_franc(0.025) 2: merlot(0.2) 3: cabernet_sauvignon(0.75) 4: petit_verdot(0.025)	8 : margaux
[88; 98]	[71; 86]	[82; 89]	[76; 92]	1: cabernet_franc(0.1) 2: merlot(0.25) 3: cabernet_sauvignon(0.65)	4 : graves
[85; 95]	[66; 88]	[59; 88]	[64; 94]	1: cabernet_franc(0.1) 2: merlot(0.25) 3: cabernet_sauvignon(0.65)	2 : saint estephe
[91; 98]	[84; 93]	[71; 73]	[82; 95]	1: cabernet_franc(0.08) 2: merlot(0.07)	5 : pauillac

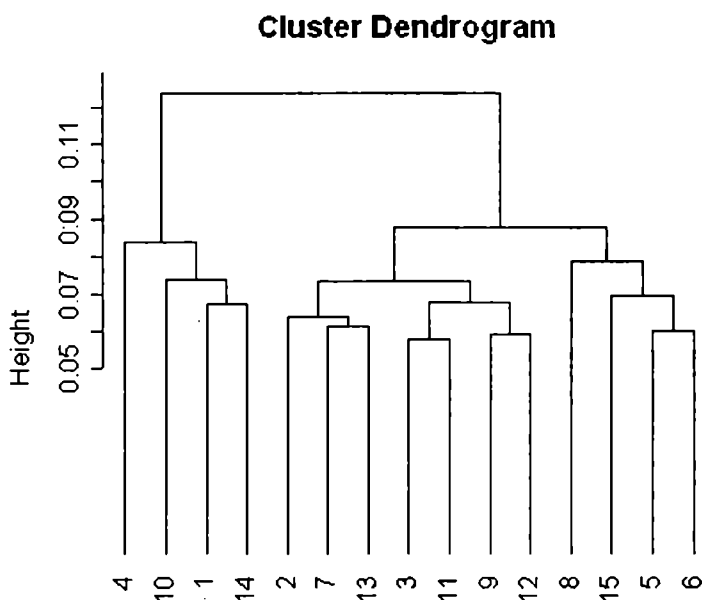
Źródło: www.ceremade.dauphine.fr/~touati/sodas-pagegarde.htm.

Tabela 5. Macierz odległości de Carvalho dla danych symbolicznych z tabeli 4

0,000	0,085	0,084	0,078	0,083	0,074	0,079	0,077	0,083	0,077	0,081	0,085	0,079	0,067	0,087
0,085	0,000	0,062	0,100	0,065	0,063	0,065	0,073	0,064	0,076	0,064	0,070	0,062	0,073	0,071
0,084	0,062	0,000	0,096	0,060	0,062	0,065	0,082	0,060	0,069	0,058	0,060	0,060	0,070	0,068
0,078	0,100	0,096	0,000	0,093	0,082	0,098	0,084	0,096	0,083	0,099	0,100	0,096	0,078	0,102
0,083	0,065	0,060	0,093	0,000	0,060	0,068	0,074	0,068	0,075	0,068	0,070	0,070	0,069	0,064
0,074	0,063	0,062	0,082	0,060	0,000	0,067	0,063	0,067	0,070	0,064	0,067	0,062	0,063	0,071
0,079	0,065	0,065	0,098	0,068	0,067	0,000	0,080	0,065	0,076	0,076	0,066	0,061	0,073	0,067
0,077	0,073	0,082	0,084	0,074	0,063	0,080	0,000	0,083	0,082	0,078	0,082	0,075	0,070	0,085
0,083	0,064	0,060	0,096	0,068	0,067	0,065	0,083	0,000	0,068	0,068	0,059	0,064	0,070	0,070
0,077	0,076	0,069	0,083	0,075	0,070	0,076	0,082	0,068	0,000	0,069	0,073	0,072	0,067	0,078
0,081	0,064	0,058	0,099	0,068	0,064	0,076	0,078	0,068	0,069	0,000	0,066	0,063	0,071	0,080
0,085	0,070	0,060	0,100	0,070	0,067	0,066	0,082	0,059	0,073	0,066	0,000	0,066	0,077	0,073
0,079	0,062	0,060	0,096	0,070	0,062	0,061	0,075	0,064	0,072	0,063	0,066	0,000	0,064	0,079
0,067	0,073	0,070	0,078	0,069	0,063	0,073	0,070	0,070	0,067	0,071	0,077	0,064	0,000	0,079
0,087	0,071	0,068	0,102	0,064	0,071	0,067	0,085	0,070	0,078	0,080	0,073	0,079	0,079	0,000

Źródło: opracowanie własne.

Dane te posłużyły jako dane wejściowe dla metody klasyfikacji hierarchicznej Warda, a otrzymany dendrogram przedstawiony jest na rys. 2.



Rys. 2. Klasyfikacja danych z tab. 4 metodą Warda

Źródło: opracowanie własne (wydruk z programu R).

7. Wnioski końcowe

Mierzenie niepodobieństwa obiektów symbolicznych jest jednym z najważniejszych zagadnień symbolicznej analizy danych. Ze względu na specyfikę reprezentacji danych w obiektach symbolicznych nie jest możliwe zastosowanie do tego celu odległości używanych w przypadku danych numerycznych. Miary de Carvalho pozwalają na mierzenie niepodobieństwa między obiektami symbolicznymi zawierającymi zmienne dowolnych typów. Mogą być one wykorzystywane w metodach znanych tylko z symbolicznej analizy danych, w niektórych metodach klasyfikacyjnych wykorzystujących jako dane wejściowe macierze odległości oraz w metodach hierarchicznych.

Zagadnieniem otwartym pozostaje używanie miar de Carvalho w metodach klasyfikacyjnych nie opartych na macierzach odległości, ale na macierzach danych, takich jak metoda k -średnich. Metody te oprócz liczenia odległości wymagają znajomości środków ciężkości klas, które dla klas składających się z obiektów symbolicznych nie są określone.

Literatura

- [1] Bock H.H., Diday E. (red.), *Analysis of Symbolic Data. Explanatory Methods for Extracting Statistical Information from Complex Data*, Springer-Verlag, Heidelberg 2000.
- [2] de Carvalho F.A.T., Souza R., *Statistical Proximity Functions of Boolean Symbolic Objects Based on Histograms*, Advances in Data Science and Classification, Springer-Verlag, Heidelberg 1998, s. 391-396.
- [3] Diday E., *An introduction to symbolic data analysis and the SODAS software*, J.S.D.A., International E-Journal 2002.
- [4] Diday E., Mballo C., *Kolmogorov-Smirnov for Decision Trees on Interval and Histogram Variables*, Classification, Clustering and Data Mining, Springer-Verlag, Berlin 2004, s. 341-350.
- [5] Dudek A., *Miary podobieństwa obiektów symbolicznych. Odległość Ichino-Yaguchiego*, Ekonomia XIV, Prace Naukowe Akademii Ekonomicznej nr 1021, AE, Wrocław 2004, 100-106.
- [6] Gatnar E., *Symboliczne metody klasyfikacji danych*, Wydawnictwo Naukowe PWN, Warszawa 1998.
- [7] Gatnar E., Walesiak M. (red.), *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*, AE, Wrocław 2004.
- [8] Gordon A.D., *Classification* (Second Edition), Chapman and Hall / CRC, London 1999.
- [9] Ichino M., Yaguchi H., *Generalized Minkowski metrics for mixed feature-type data analysis*, „IEEE Transactions on Systems, Man, and Cybernetics” 1994, Vol. 24, No. 4, 698-707.
- [10] Malerba D., Espozito F., Giovalle V., Tamma V., *Comparing dissimilarity measures for symbolic data analysis*, materiały konferencyjne „New Techniques and Technologies for Statistics” i „Exchange of Technology and Know-how” (ETK-NTTS'01) 2001, s. 473-481.
- [11] Verde R., *Clustering methods in symbolic data analysis*, [w:] Banks D., House L., McMorris F.R., Araabie P., Gaul W. (red.), *Classification, Clustering and Data Mining Application*, Springer-Verlag, Berlin 2004, s. 299-318.

- [12] Walesiak M., *Uogólniona miara odległości w statystycznej analizie wielowymiarowej*, AE, Wrocław 2002.
- [13] SODAS, *Documentation*, dokumentacja pakietu SODAS w wersji 1.20, dostępna pod adresem www.ceremade.dauphine.fr/~touati/aidedoc/.

SYMBOLIC OBJECTS SIMILARITY MEASURES. DE CARVALHO MEASURES

Summary

This paper presents main measures used for measuring similarity and dissimilarity between symbolic (e.g. not necessarily numeric) objects. This second part is a continuation of the article from *Ekonometria XII'* and describes de Carvalho dissimilarity measures along with definitions of comparison and aggregation functions, description potential of symbolic object with an example of using de Carvalho measures matrix as an input to Ward clustering algorithm.

Dr Andrzej Dudek jest pracownikiem Katedry Ekonometrii i Informatyki w Akademii Ekonomicznej we Wrocławiu.