PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1076 2005
TAKSONOMIA 12
Klasyfikacja i analiza danych – teoria i zastosowania

Marcin Pełka
Akademia Ekonomiczna we Wrocławiu

ZMIENNE HIERARCHICZNE W POMIARZE ODLEGŁOŚCI W SYMBOLICZNEJ ANALIZIE DANYCH

1. Wstęp

Metody klasyfikacji, należące do grupy metod statystycznej analizy wielowymiarowej, mają szerokie zastosowanie w badaniach marketingowych, a szczególnie w segmentacji rynku, pozycjonowania produktu (przedsiębiorstwa) na rynku, identyfikacji rynków testowych oraz określania struktury rynku [Walesiak 1996, s. 117, 120, 151].

Istota metod klasyfikacji sprowadza się do wyróżnienia jednorodnych klas z analizowanego zbioru obiektów, wychwycenia cech charakteryzujących poszczególne klasy oraz wskazania różnic między wyodrębnionymi skupieniami.

Otóż, jednym z podstawowych problemów w każdej metodzie, czy to klasycznej, czy to symbolicznej, jest problematyka pomiaru podobieństwa lub niepodobieństwa. Oprócz problemu odpowiedniej miary odległości, powstaje także problem odpowiedniej miary niepodobieństwa czy podobieństwa dla zmiennych.

2. Typy zmiennych w symbolicznej analizie danych

W obiektach symbolicznych możemy mieć do czynienia z następującymi rodzajami zmiennych [*Analysis of...* 2000, s. 2-3]:

1. Ilorazowe, przedziałowe, porządkowe, nominalne.
2. Kategorie, czyli dane tekstowe, np. biały, zielony.
3. Przedziały liczbowe, które np. określają dochód konsumenta (1000 zł, ..., 2500 zł) czy ilość spalanej benzyny na 100 km (5 l, ..., 11 l), co w przypadku dochodu oznacza, że może on wynieść od 1000 do 2500 zł (np. praca akordowa).

4. Lista kategorii – przykładem może być standard wyposażenia pewnego samochodu (obiektu) określony jako: niski, średni, wysoki, co oznacza, że jest oferowany w trzech standardach wyposażenia.

5. Lista kategorii z wagami (prawdopodobieństwami), na której, oprócz listy kategorii, występują wagi, z jakimi obiekt ma wybraną kategorię, np. jeżeli wybieramy zmienną: typ nadwozia samochodu pewnej grupy osób w postaci listy sedan (0,45), cupe (0,20), pickup (0,25), kabriolet (0,10), inne (0,00), to oznacza to, że 45% samochodów, będących własnością tych osób, to samochody o nadwoziu typu sedan, a 10% to kabriolety. Osoba, która w ankiecie zaznaczyła, że jej samochód to kabriolet, należy właśnie do tej grupy.

6. Zmienne typu NA (ang. *non applicable*), oznaczające, że dla danej zmiennej obiekt nie ma odpowiedniej kategorii czy liczby, przykładem może być tu zmienna *liczba urodzonych dzieci* w przypadku obiektu, którym będzie mężczyzna,

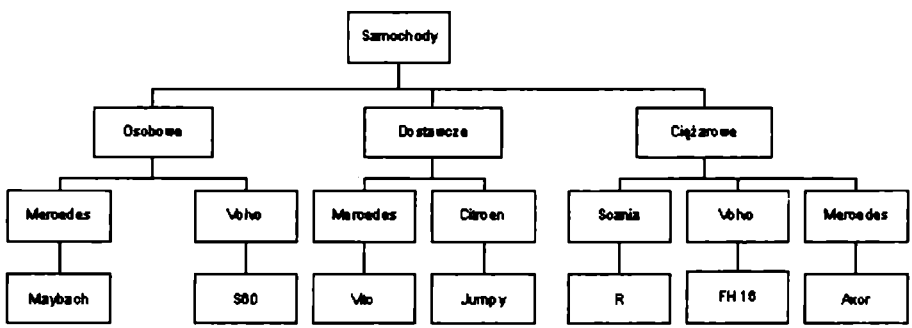
7. Zmienne zależne – Diday [*Analysis of...* 2000, s. 33-37] wyróżnia, oprócz wyżej wymienionych typów zmiennych, także zmienne zależne, przy czym dzieli je na podstawowe podtypy:

a) zmienne o zależności funkcyjnej, gdzie pomiędzy zmienną A i B zachodzi zależność funkcyjna, np. jeżeli $a = 5$, $b = 2$ *to pole prostokąta* $P = a * b = 2 * 5 = 10$ lub jeżeli *suma ubezpieczenia wynosi 10 000 zł,* to *składka podstawowa wynosi 30 zł,*

b) zmienne o zależności logicznej – pomiędzy zmiennymi A i B występuje zależność logiczna, np. jeżeli *waga danej osoby* ≤ 50 *kg, to wzrost* ≤ 180 *cm,* zmienna ta nie ma zależności *stricte* funkcyjnej, lecz ma zależność logiczną,

c) zmienne zależnie stochastycznie, w których (w przeciwieństwie do tradycyjnych procesów stochastycznych), nie mówi się o elemencie losowym, lecz o zależności pomiędzy zmiennymi, wynikającej z korelacji pomiędzy tymi zmiennymi,

d) zmienną hierarchiczną, która ma następującą postać. $Y = \{A, B, \ldots\}$ można przedstawić za pomocą drzewa H.



Rys. 1. Zmienna hierarchiczna – samochody
Źródło: opracowanie własne.

Każda kategoria nazywana jest **węzłem** (ang. *node*) drzewa (hierarchii) H, dana kategoria C jest **potomkiem** (ang. *descendent*) dla kategorii A, jeżeli $C \prec A$, np. kategoria Maybach jest potomkiem kategorii Mercedes. Kategorię A nazywamy **rodzicem, przodkiem** (ang. *ancestor*) kategorii C[1]. Kategorię możemy nazwać także **bezpośrednim potomkiem, bezpośrednim następcą** (ang. *successor, direct descendent*) kategorii, jeżeli nie znajduje się żadna pośrednia kategoria pomiędzy nimi.

Każde hierarchia ma unikalny **korzeń** (ang. *root*), który nie ma żadnych rodziców, a kategorię nazywamy **liściem, węzłem końcowym** (ang. *leaf, terminal node*), jeżeli nie ma ona żadnych potomków. Węzłami na rys. 1 są m. in. kategorie: Axor, FH 16 czy R.

Hierarchia dla takiej zmiennej jest ustalana *a priori* przez prowadzącego badanie. Dodatkowo nie są podawane żadne kryteria ważności poszczególnych elementów takiej zmiennej. Znaczenie wyboru badacza będzie przedstawione w dalszej części artykułu na podstawie pomiaru odległości pomiędzy obiektami, w których wystąpią tego typu zmienne.

3. Symboliczne miary odległości

Po zaprezentowaniu najważniejszych typów danych (a przed przedstawieniem problemów, które niesie wybór takiej, a nie innej struktury dla zmiennej hierarchicznej) przedstawione zostaną miary odległości, stosowane w symbolicznej analizie danych.

Dla wszystkich wzorów użytych w dalszej części artykułu zastosowano następujące oznaczenia:

- $O_1 = (A_1, \ldots, A_n)$ to obiekt symboliczny numer 1, składający się z p zmiennych, a $O_2 = (B_1, \ldots, B_n)$ obiekt symboliczny numer 2, składający się z p zmiennych,
- j to zmienna z przedziału $(1,\ldots,p)$,
- w_j to waga związana z j-tą zmienną przy czym suma wag równa jest 1,
- q to liczba całkowita większa lub równa 1.

1. **Miara odległości Gowdy-Didaya** [*Analysis of...* 2000, s. 166–170]

Jest to jedna z pierwszych i najprostszych miar odległości, które można zastosować w analizie symbolicznej. Miara Gowdy-Didaya nie może być zastosowana w przypadku zmiennych hierarchicznych, dlatego nie będzie zastosowana w przykładzie empirycznym. Odległość Gowdy-Didaya jest odległością nieznormalizowaną.

2. **Miara odległości Ichino-Yacuchiego** [*Analysis of...* 2000, s. 170-173]

Miara ta ma możliwości pomiaru odległości dla zmiennych w postaci zmiennej hierarchicznej, a ponadto istnieje jej wariant znormalizowany. Jest ona oparta na pojęciach operatorów połączenia (ang. *cartesian join*) i przekroju (ang. *cartesian meet*), będącymi rozszerzeniami operatorów $\cup$ i $\cap$ na wszystkie typy zmiennych w obiektach symbolicznych. Miara Ichino-Yacuchiego w przypadku obiektów wyraża się wzorem:

[1] Trzeba zaznaczyć, że niektóre kategorie należy traktować raz jako rodziców, a raz jako potomków.

$$d_q(O_1,O_2)=\left(\sqrt[q]{\sum_{i=1}^{p}\phi(A_j,B_j)^q}\right). \tag{1}$$

Element $\phi(A_j,B_j)$ opisany jest następującym wzorem.

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

gdzie[2]: A_j,B_j – zmienne dowolnego typu,

$\oplus$ – operator połączenia,

$\otimes$ – operator przekroju,

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

γ – parametr z przedziału <0, ½ >.

Nadmienić należy, że operator przekroju odpowiada $A\otimes B=A\cap B$ dla zmiennej dowolnego typu, a działania podejmowane przy operatorze połączenia są zależne od typu zmiennej, z którą mamy do czynienia.

Dla zmiennej numerycznych (liczbowej) operator połączenia jest zdefiniowany zgodnie ze wzorem:

$$A_j\oplus B_j=\begin{cases}(A_j,B_j)\Leftrightarrow A_j<B_j\\(B_j,A_j)\Leftrightarrow A_j\geq B_j\end{cases}, \tag{3}$$

gdzie: A_j,B_j – zmienne numeryczne,

(A_j,B_j) – przedział liczbowy.

Dla zmiennej w postaci przedziału liczbowego operator połączenia jest zdefiniowany zgodnie ze wzorem:

$$(a_1,a_2)\oplus(b_1,b_2)=\left(\min(a_1,b_1),\max(a_2,b_2)\right), \tag{4}$$

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

Dla zmiennych o strukturze gałęziowej operator połączenia jest zdefiniowany zgodnie ze wzorem:

$$A_j\oplus B_j=\begin{cases}A_j\cup B_j\Leftrightarrow N(A_j)=N(B_j)\\L(N(A_j\cup B_j))\Leftrightarrow N(A_j)\neq N(B_j)\end{cases}, \tag{5}$$

gdzie: A_j,B_j – zmienne o strukturze gałęziowej,

$N(X)_{\text{dla } x=A,B \text{ lub } A\cup B}$ – najniżej położony element drzewa, który jest rodzicem dla wszystkich zmiennych X,

$L(Y)_{\text{dla } Y=N(A\cup B)}$ – zbiór wszystkich liści drzewa położonych poniżej Y.

[2] Szerzej na temat odległości Ichino-Yacuchiego zob. [Dudek 2004].

Dla pozostałych typów zmiennych operator połączenia jest zdefiniowany zgodnie ze wzorem:

$$A \oplus B = A \cup B. \tag{6}$$

Widać, że w przypadku zmiennej hierarchicznej ważną rolę odgrywa początkowy wybór postaci drzewa przyjmowany przez badacza.

3. **Miary De Carvalho** [*Analysis of*... 2000, s. 173-177]

Stanowią one rozszerzenie miary Ichino i Yacuchiego oparte na pojęciach funkcji porównującej (*CF – comparision function*) i funkcji agregującej (*AF – agregation function*). Miary odległości De Carvalho uwzględniają także logiczne zależności między zmiennymi. Z tego względu zostaną one jedynie przedstawione w formie tabeli zbiorczej (tab. 1).

Tabela 1. Miary niepodobieństwa w analizie symbolicznej

Lp.	Miara niepodobieństwa zmiennych	Miara niepodobieństwa obiektów
1	$D^{(j)}(A,B) = D_\pi(A,B) + D_S(A,B) + D_C(A,B)$	$d(O_1,O_2) = \sum_{j=1}^{p} D^{(j)}(A_j,B_j)$
2	$\phi(A,B) = \lvert A \oplus B\rvert - \lvert A \otimes B\rvert + \gamma\left(2.\lvert A \oplus B\rvert - \lvert A\rvert - \lvert B\rvert\right)$	$d_q(O_1,O_2) = \left(\sqrt[q]{\sum_{i=1}^{p} \varphi(A_i,B_i)^q}\right)$
3	$\psi(A,B) = \frac{\varphi(A,B)}{\lvert Y\rvert}$	$d_q(O_1,O_2) = \left(\sqrt[q]{\sum_{i=1}^{p} \psi(A_i,B_i)^q}\right)$
4	$\psi(A,B) = \frac{\varphi(A,B)}{\lvert Y\rvert}$	$d_q^i(O_1,O_2) = \left(\sqrt[q]{\sum_{i=1}^{p}\left[w_j\psi(A_i,B_i)\right]^q}\right)$
5	$d_i(A,B)\ i = 1,2,\ldots,5$	$d_q(O_1,O_2) = \left(\sqrt[q]{\sum_{i=1}^{p}\left[w_j d_i(A_i,B_i)\right]^q}\right)$
6	$\psi'(A,B) = \frac{\phi(A,B)}{\mu(A \oplus B)}$	$d'_q(O_1,O_2) = \left(\sqrt[q]{\sum_{i=1}^{p}\frac{1}{p}\left[\psi'(A_i,B_i)\right]^q}\right)$
7		$d'_1(O_1,O_2) = \left[\pi(O_1 \oplus O_2) - \pi(O_1 \otimes O_2) + \gamma.(2\pi(O_1 \otimes O_2) - \pi(O_1) - \pi(O_2))\right]$
8		$d'_2(O_1,O_2) = \left[\pi(O_1 \oplus O_2) - \pi(O_1 \otimes O_2) + \gamma.(2\pi(O_1 \otimes O_2) - \pi(O_1) - \pi(O_2))\right] / \pi(O_1^E)$
9		$d'_3(O_1,O_2) = \left[\pi(O_1 \oplus O_2) - \pi(O_1 \otimes O_2) + \gamma.(2\pi(O_1 \otimes O_2) - \pi(O_1) - \pi(O_2))\right] / \pi(O_1 \oplus O_2)$
10	$d_i(A,B)\ i = 1,2,\ldots,5$	$d_q^i(O_1,O_2) = \left(\sqrt[q]{\frac{\sum_{i=1}^{p}\left[w_j d_i(A_i,B_i)\right]^q}{\sum_{j=1}^{p}\delta(j)}}\right)$

Źródło: [Dudek 2004].

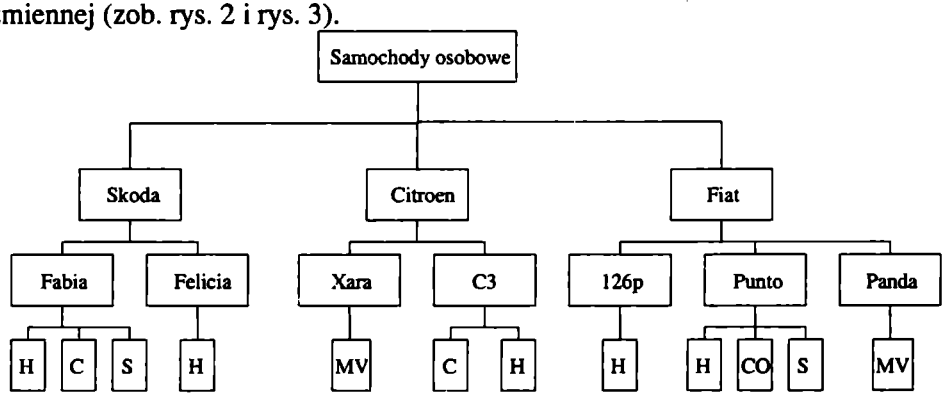
W tabeli 1 miara Gowdy-Didaya to pozycja 1, pozycje od 2 do 4 to miary Ichino-Yacuchiego, a pozycje od 5 do10 to miary De Carvalho.

4. Przykład empiryczny

W przykładzie empirycznym wykorzystano wyniki badania ankietowego przeprowadzonego od lipca do sierpnia 2004 r. na 167 respondentach. Ankietę przeprowadzono metodą „kuli śniegowej" – kolejne osoby przekazywały ankietę innym osobom, co zwiększyło liczbę respondentów. Ostatecznie do analizy przyjęto 150 prawidłowo wypełnionych ankiet. Do prezentacji znaczenia przyjętej struktury zmiennej hierarchicznej w symbolicznej analizie danych, a zwłaszcza w pomiarze odległości, przyjęto następujące zmienne:

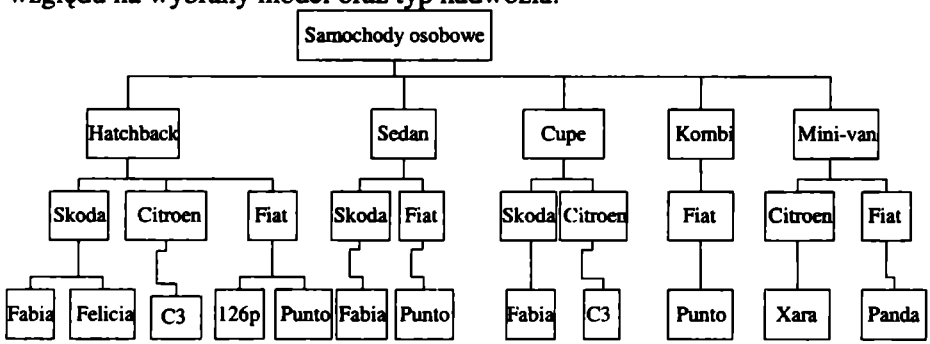
1. Kwotę, którą respondent byłby skłonny wydać na zakup samochodu – przedział liczbowy.
2. Zamiar zakupu samochodu w ciągu dwóch najbliższych lat – zmienna w postaci listy kategorii.
3. Czynniki wybrane przez respondenta przy ocenie samochodu – lista kategorii.
4. Cel wykorzystania samochodu – lista kategorii.
5. Źródła, które respondent wykorzystuje do pozyskiwania interesujących go informacji o samochodach – lista kategorii,
6. Typ nadwozia samochodowego wybranego przez respondenta – zmienna hierarchiczna „samochody".

Konieczne jest wstępne ustalenie zmiennej hierarchicznej. W celu pokazania jej znaczenia w symbolicznej analizie danych przyjęto dwa różne warianty takiej zmiennej (zob. rys. 2 i rys. 3).



Rys. 2. Zmienna hierarchiczna samochody – przypadek pierwszy
H – hatchback, C – cupe, S – sedan, MV – minivan, CO – kombi
Źródło: opracowanie własne na podstawie zebranych ankiet.

W tym przypadku za pierwszy przyjęto podział zmiennej hierarchicznej ze względu na wybraną markę samochodu, a dopiero w następnej kolejności – ze względu na wybrany model oraz typ nadwozia.



Rys. 3. Zmienna hierarchiczna samochody – przypadek drugi
Źródło: opracowanie własne na podstawie zebranych ankiet.

W tym przypadku podziału zmiennej hierarchicznej na samym początku dokonano ze względu na wybrany typ nadwozia, a w następnej kolejności – ze względu na markę oraz model samochodu. Do analizy wybrano siedem obiektów symbolicznych reprezentujących siedem marek samochodów. Dane wykorzystano do przedstawienia różnic pomiędzy pomiarami odległości dla obiektów i sześciu zmiennych (tab. 2).

Tabela 2. Macierz informacji

Zmienne obiekty	Kwota (w tys. zł)	Zakup auta	Wybrane czynniki	Cel wykorzystania	Źródła informacji	Typ nadwozia
1	{21,...,60}	nie	B, E, M	prywatny	portale, internet, gazety	H, S, C
2	{21,....,60}	nowy, nie	E, C	prywatny służbowy	TV, Internet	H
3	{41,...,85}	nie	M, T, K, O	prywatny	Internet, TV	H, C
4	{20,...,40}	nowy, używany	C, M, K	prywatny	gazety, TV, portale	H
5	{21,...,85}	nie, nowy	K, M, T, B	prywatny służbowy	Internet, gazety	H, S, K
6	{41,...,85}	nie	B, T, M, P	prywatny	TV, Internet, gazety	MV
7	{41,...,110}	nie	T, M, B	prywatny służbowy	portale, gazety, internet	MV

B – bezpieczeństwo, E – estetyka, T – technologia, C – cena, M – marka, K – komfort, O – osiągi, P – promocje, S – sedan, C – cupe, H – hatchback, K – kombi, MV – minivan

Źródło: opracowanie własne.

Dla obiektów z tab. 2 obliczono odległości symboliczne miarą Ichino-Yacuchiego dla przypadków, gdy zmienna hierarchiczna samochody przyjmuje postać jak na rys. 3 oraz dla przypadku, gdy postać tej zmiennej przedstawia rys. 2. Macierze odległości prezentu-

ją tab. 3 i 4. W tabelach wytłuszczonym drukiem zaznaczono odległości obiektu drugiego od czwartego w celu wskazania dalszych różnic.

Tabela 3 przedstawia macierz odległości w przypadku, gdy zmienna „samochody" przybiera postać przedstawioną na rys. 2.

Tabela 3. Macierz odległości

	1	2	3	4	5	6	7
1	0,00	4,00	9,67	5,39	9,67	9,49	10,63
2	4,00	0,00	9,70	**5,29**	9,82	9,87	11,11
3	9,67	9,70	0,00	11,68	7,11	4,58	8,25
4	5,39	5,29	11,68	0,00	12,79	10,86	12,75
5	9,67	9,82	7,11	13,17	0,00	6,67	9,33
6	9,49	9,87	4,58	11,18	6,67	0,00	7,52
7	10,63	11,11	8,25	12,75	9,33	7,52	0,00

Źródło: opracowanie własne.

Tabela 4. Macierz odległości

	1	2	3	4	5	6	7
1	0,00	4,80	9,46	5,20	9,46	9,43	10,58
2	4,80	0,00	9,59	**4,36**	9,67	9,77	11,02
3	9,46	9,59	0,00	11,27	6,96	4,58	8,25
4	5,20	4,36	11,27	0,00	13,10	11,20	12,77
5	9,46	9,67	6,96	13,10	0,00	7,18	9,27
6	9,43	9,77	4,58	11,20	7,18	0,00	6,67
7	10,58	11,02	8,25	12,77	9,27	6,67	0,00

Źródło: opracowanie własne.

Tabela 4 prezentuje macierz odległości oraz wyróżnioną odległość obiektu drugiego od czwartego, gdy zmienna „samochody" przyjmuje postać zaprezentowaną na rys. 3. W tym przypadku odległości obiektów 2 od 4 i 1 od 4 nie są już do siebie tak podobne jak w przypadku poprzednim.

5. Wnioski

Przyjęcie z góry ustalonej struktury zmiennej hierarchicznej (zmienna „samochody") powoduje powstanie znacznych różnic w odległości pomiędzy obiektami. W przykładzie empirycznym wskazano, że przyjęcie dwóch różnych zmiennych hierarchicznych powoduje znaczne różnice w uzyskanych wynikach. Różnice te zwiększyłyby się, gdyby zmienne były dodatkowo ważone, a wagi przypisane zmiennej hierarchicznej byłyby duże. Zastosowana miara jest miarą nieznormalizowaną, spełniającą kryteria metryki. Jeżeli zastosujemy miarę normalizowaną, to problem ten nie się zmieni.

Powstawanie różnic, wynikających z przybierania różnej postaci zmiennej hierarchicznej, może powodować znaczne różnice w pomiarze odległości, tworzeniu

klas. W efekcie te same obiekty mogą być przyłączane raz do jednej, raz do innej klasy, co z kolei może utrudniać klasyfikację czy walidację.

Aby uniknąć problemów związanych z arbitralnie przyjmowaną strukturą hierarchiczną w prowadzeniu symbolicznej analizy danych,sugeruje się, aby:

1. Przyjąć strukturę zmiennej hierarchicznej, którą uzyskano w innych badaniach – np. drzewa klasyfikacyjne.
2. Przeprowadzić badania dla różnych struktur tej samej zmiennej i uśrednić wyniki.
3. Przyjmować strukturę klas zgodnie z celem badania lub zgodnie z warunkami, które stawia badaczowi zamawiający badanie.

Literatura

Analysis of Symbolic Data. Explanatory Methods for Extracting Statistical Information from Complex Data, red. H.H. Bock, E. Diday, Springer Verlag 2000.

Dudek A., *Miary podobieństwa obiektów symbolicznych. Odległość Ichino-Yacuchiego*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu, Ekonometria 15, 2004 (w druku).

Gatnar E., *Symboliczne metody klasyfikacji danych*, Wydawnictwo Naukowe PWN, Warszawa 1998.

Ichino M., Yacuchi H., *Generalized Minkowski Metrics for Mixed Feature-Type Data Analysis*, IEEE Transactions on Systems, Man, and Cybernetics, vol. 24, nr 4, 1994 s. 698-707.

Malerba D., Espozito F., Giovalle V., Tamma V., *Comparing Dissimilarity Measures for Symbolic Data Analysis*, materiały konferencyjne „New Techniques and Technologies for Statistcs" and „Exchange of Technology and Know-How" (ETK-NTTS'01), 2001, s. 473-481.

Pełka M., *Miary podobieństwa obiektów symbolicznych. Idea Gowdy i Didaya*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu, Ekonometria 15, 2004 (w druku).

Walesiak M., *Metody analizy danych marketingowych*, PWN, Warszawa 1996.

TAXONOMIC VARIABLES IN SYMBOLIC DISSIMILARITY MEASURE IN SYMBOLIC DATA ANALYSIS

Summary

The aim of this article is to present all types of dissimilarity measures in symbolic data analysis (such as Gowdy-Diday, Ichino-Yacuchi and De Carvalho measures) with problems which can be found in each of them while using specified variables types, especially taxonomic variables. Article presents polish suggestions for terms used in dissimilarity measures. For empirical examples analysis of car marked was used.