

Marcin Pelka

MIARY PODOBIENSTWA OBIEKTÓW SYMBOLICZNYCH. IDEA GOWDY I DIDAYA

1. Wstęp

Pomiar odległości odgrywa ważną rolę w wielu metodach, których celem jest np. klasyfikacja. W niniejszym artykule przedstawione zostaną pojęcia obiektu symbolicznego oraz idea pomiaru odległości zaprezentowana przez K.C. Gowdę i E. Didaya. Miara ta została opisana w pracy *Analysis of Symbolic Data. Explanatory Methods for Extracting Statistical Information from Complex Data* [Bock, Diday 2000] obok miary Ichino-Yaguchiego oraz miary de Carvalho, którą zaprezentował A. Dudek w zeszycie *Ekonometria* nr 14. Przedstawimy także problemy mogące wystąpić przy pomiarze odległości symbolicznej tą metodą.

2. Obiekt symboliczny i zmienne symboliczne

Według E. Gatnara [1998, s. 62], obiekt symboliczny opisywany jest przez koniunkcję wartości poszczególnych jego cech, gdzie w opisie obiektów i ich klas stosuje się logikę V L1 Michalskiego. Przykładowo obiekt

[posiadany samochód = Skoda] $\wedge$ [pojemność = 1400cm³] $\wedge$ [zakupione opony = Dębica]

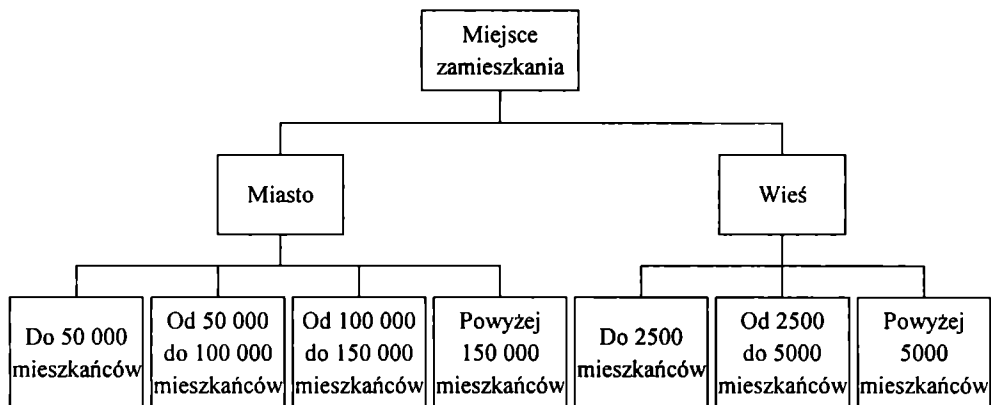
reprezentuje samochód marki Skoda o pojemności silnika 1400 cm³, z oponami marki Dębica. Taki opis w logice V L1 jest nazywany kompleksem [Gatnar 1998, s. 67-69]. Zastosowanie pojęć naturalnych pozwala na interpretowanie obiektów czy ich klas łatwiejsze niż w metodach tradycyjnych.

W obiekcie symbolicznym zmienne mogą przyjmować nieco inną postać niż w ujęciu tradycyjnym. Dokładnie przedstawia to tab. 1.

Tabela 1. Zmienne występujące w obiektach symbolicznych

Nazwa zmiennych	Przykład
Liczbowe (nazywane numerycznymi)	wzrost = 175
Tekstowe (jakościowe)	kolor = zielony
Przedziały wartości	zarobki = (1560,...,2500)
Lista wartości	wybrane programy = {TVP1, TVN, TVP2}
Lista wartości z wagami (prawdopodobieństwami)	miejsce zamieszkania = {miasto 65%, wieś 35%}

Źródło: opracowanie własne.



Rys. 1. Zmienna hierarchiczna – miejsce zamieszkania

Źródło: opracowanie własne.

Oprócz tego zmienne mogą mieć postać zmiennej zależnej, np.:

- zmienne o zależności funkcyjnej**, gdy $a = 5$, $b = 10$, wówczas zmienna *pole prostokąta* przyjmie wartość 50,
- zmienne o zależności logicznej**, jeżeli *wiek* > 70 , to wówczas zmienna *emerytura (renta)* = *tak*,
- zmienne reprezentujące strukturę hierarchiczną** (*taxonomic variable*, por. [Wójcik 1965, s. 47, 55-57; Dudek 2004a]); przykładem zmiennej tego typu może być miejsce zamieszkania (rys. 1).

3. Miara odległości symbolicznej według K.C. Gowdy i E. Didaya

Miara Gowdy–Didaya, zwana także miarą wzajemnego sąsiedztwa (*mutual neighborhood value*), została zaproponowana dla obiektów symbolicznych w artykule *Symbolic Clustering Using a New Dissimilarity Measure* [Gowda, Diday 1991].

Gowda i Diday [Bock, Diday 2000, s. 166-167] zaproponowali odległość, która wyraża się wzorem (1).

$$D(A, B) = \sum_{j=1}^p D(A_j, B_j), \quad (1)$$

gdzie: A, B – obiekty symboliczne,

A_j, B_j – odpowiednio zmienne w obiektach A i B , dla których liczona jest miara niepodobieństwa zmiennych $D(A_j, B_j)$,

$j = 1, 2, \dots, p$ oznacza numer kolejnej zmiennej.

Aby móc dokonać obliczeń za pomocą miary odległości Gowdy–Didaya, konieczne jest zdefiniowanie miary niepodobieństwa dla zmiennych, zależących od typu zmiennej, z którą mamy do czynienia.

1. **Zmienne liczbowe** (zmienne mierzone na skali przedziałowej lub ilorazowej) – dla tego typu zmiennych miara niepodobieństwa dla zmiennych składa się z następujących elementów:

$$D(A_j, B_j) = D_{II}(A_j, B_j) + D_S(A_j, B_j) + D_C(A_j, B_j). \quad (2)$$

H.H. Bock i E. Diday [2000, s. 167] opisują elementy wzoru 2 następująco:

- $D_{II}(A_j, B_j)$ odzwierciedla względną, relatywną, pozycję dwóch zmiennych, w tym przypadku są to zmienne w postaci przedziału liczbowego, w dziedzinie.
- $D_S(A_j, B_j)$ oznacza relatywną wielkość zmiennych bez uwzględnienia elementów wspólnych dla obu zmiennych.
- $D_C(A_j, B_j)$ oznacza relatywną wielkość różnicy symetrycznej zmiennych.

Zmienna liczbowa A_j ma postać:

$$A_j = [\underline{a_j}, \overline{a_j}], \text{ gdzie } \overline{a_j} \text{ to górna granica przedziału zmiennej, a } \underline{a_j} \text{ to dolna}$$

granica przedziału zmiennej.

Zmienna B_j ma postać:

$$B_j = [\underline{b_j}, \overline{b_j}], \text{ gdzie granice przedziału zmiennej opisane są tak samo, jak}$$

w przypadku zmiennej A_j .

Obserwacje, które nie są przedziałami, lecz punktami, zapisuje się jako zmienne, których początek i koniec przedziału są takie same.

Wzory (3)–(5) opisują sposób obliczania elementów D_{II} , D_S oraz D_C .

$$D_{II}(A_j, B_j) = \frac{|\underline{a_j} - \underline{b_j}|}{|Y_j|}, \quad |Y_j| \text{ oznacza długość dziedziny dla } j\text{-tej zmiennej}, \quad (3)$$

$$D_S(A_j, B_j) = \frac{|l_a - l_b|}{l_S}, \quad (4)$$

gdzie odpowiednio w ramach D_S [Bock, Diday 2000, s. 167]:

- $l_a = |\overline{a_j} - \underline{a_j}|$,
- $l_b = |\overline{b_j} - \underline{b_j}|$,
- $l_S = |\max(\overline{a_j}, \overline{b_j}) - \min(\underline{a_j}, \underline{b_j})|$,

$$D_C(A_j, B_j) = \frac{l_a + l_b - 2l_{a \cap b}}{l_S}, \quad (5)$$

gdzie $l_{a \cap b}$ oznacza długość przedziału liczbowego będącego częścią wspólną dla zmiennej A_j oraz zmiennej B_j .

Przykład 1. Miara niepodobieństwa oraz odległości dla zmiennej w postaci przedziału liczbowego

	Liczba mieszkańców domu (zmienna X_1)
A	(2, ..., 7)
B	(3, ..., 9)

$$Y_j = |(2; 9)| = |9 - 2| = 7,$$

$$l_a = |7 - 2| = 5,$$

$$l_b = |9 - 3| = 6;$$

$$l_S = |\max(9; 7) - \min(2; 3)| = |9 - 2| = 7,$$

$$l_{a \cap b} = |(3, \dots, 7)| = |7 - 3| = 4,$$

$$D_H(A_1, B_1) = \frac{|2 - 3|}{7} = 0,14,$$

$$D_S(A_1, B_1) = \frac{|5 - 6|}{7} = 0,14,$$

$$D_C(A_1, B_1) = \frac{5 + 6 - 2 \cdot 4}{7} = 0,43.$$

Ostatecznie miara niepodobieństwa zmiennych $D(A_1, B_1) = 0,14 + 0,14 + 0,43 = 0,71$, a odległość obiektu A od B to: $D(A, B) = \sum_{j=1}^P D(A_j, B_j) =$

$$= \sum_{j=1}^1 D(A_1, B_1) = 0,71.$$

2. **Zmienne tekstowe** (jakościowe; zmienne mierzone na skali porządkowej lub nominalnej). Dla tego typu zmiennych miara niepodobieństwa dla zmiennych składa się z następujących elementów [Bock, Diday 2000, s. 170]:

$$D(A_j, B_j) = D_S(A_j, B_j) + D_C(A_j, B_j). \quad (6)$$

Wzory (7) i (8) prezentują sposób obliczania elementów D_S oraz D_C .

$$D_S(A_j, B_j) = \frac{|l_a - l_b|}{l_S}, \quad (7)$$

gdzie odpowiednie elementy to:

- l_a to liczba kategorii w $A_j = |a_j|$; symbol $|a|$ nie oznacza wartości bezwzględnej lecz liczbę odpowiednich wariantów (kategorii) zaobserwowanych dla obiektu A w zmiennej X_j ,
- l_b to liczba kategorii w $B_j = |b_j|$,
- $l_{a \cap b}$ oznacza liczbę kategorii wspólnych dla zmiennych,
- l_S opisywane jest jako $l_S = l_a + l_b - l_{a \cap b}$;

$$D_C(A_j, B_j) = \frac{l_a + l_b - 2l_{a \cap b}}{l_S}. \quad (8)$$

Przykład 2. Miara niepodobieństwa oraz odległości dla zmiennej w postaci listy kategorii (wartości)

	Akceptowane ryzyko inwestycyjne (zmienna X_1)
A	{niskie, średnie}
B	{niskie, średnie, wysokie}

$$l_a = |\{\text{niskie, średnie}\}| = 2,$$

$$l_b = |\{\text{niskie, średnie, wysokie}\}| = 3,$$

$$l_{a \cap b} = |\{\text{niskie, średnie}\} \cap \{\text{niskie, średnie, wysokie}\}| = |\{\text{niskie, średnie}\}| = 2,$$

$$l_S = 2 + 3 - 2 = 3.$$

Po podstawieniu tych danych do wzorów (7) i (8) otrzymujemy:

$$D_S(A_1, B_1) = \frac{|2 - 3|}{3} = 0,33,$$

$$D_C(A_1, B_1) = \frac{2 + 3 - 2 \cdot 2}{3} = \frac{1}{3} = 0,33.$$

W efekcie końcowym miara niepodobieństwa dla zmiennej w postaci listy kategorii wynosi:

$$D(A_1, B_1) = D_S(A_1, B_1) + D_C(A_1, B_1) = 0,33 + 0,33 = 0,66,$$

a odległość A od B wynosi:

$$D(A, B) = \sum_{j=1}^p D(A_j, B_j) = \sum_{j=1}^1 D(A_1, B_1) = 0,66.$$

Ważne jest, ażeby zaznaczyć, że odległość dla tego typu zmiennych w dużej mierze zależy od liczby kategorii w obiektach A i B dla tej samej zmiennej. Jeżeli liczba kategorii w ramach tej samej zmiennej będzie taka sama w obiektach A i B , to: $|A_j| = |B_j| \rightarrow D_S(A_j, B_j) = 0$.

4. Wnioski

Pomiar odległości miarą odległości Gowdy i Didaya był jednym z pierwszych, jakie wprowadzono do analizy symbolicznej. Jest to miara względnie łatwa do obliczenia. Jedną z najważniejszych wad, które można znaleźć w tej idei, jest to, że nie pozwala ona na zastosowanie zmiennych w postaci list kategorii (wartości) z wagami; inną wadą jest brak wskazań co do możliwości zastosowania zmiennych w postaci zmiennej hierarchicznie zależnej (por. [Bock, Diday 2000, s. 166]).

Jeżeli chodzi o własności metryki, to nie jest dowiedzione, czy odległość mierzona według idei Gowdy i Didaya spełnia warunek trójkąta. Wiadomo, że odległość mierzona dla zmiennych ze „słabych” skal pomiarowych spełnia wszystkie warunki metryki, oprócz warunku trójkąta, odległość zaś mierzona dla zmiennych ilorazowych lub przedziałowych wymaga wyznaczenia $|Y_j|$ [Bock, Diday 2000, s. 169].

Literatura

- Bock H.-H., Diday E. (2000), *Analysis of Symbolic Data. Explanatory Methods for Extracting Statistical Information from Complex Data*, Springer Verlag.
- Dudek A. (2004a), *Miary podobieństwa obiektów symbolicznych. Odległość Ichino-Yaguchiego*, [w:] *Ekonometria 14, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1021*, AE, Wrocław.
- Dudek A. (2004b), *Tworzenie obiektów symbolicznych z baz danych*, [w:] *Ekonometria 14, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1021*, AE, Wrocław.
- Gatnar E. (1998), *Symboliczne metody klasyfikacji danych*, Wydawnictwo Naukowe PWN, Warszawa.
- Gowda K.C., Diday E. (1991), *Symbolic Clustering Using a New Dissimilarity Measure*, [w:] R. Ledley (red.) „Pattern Recognition” nr 24, Elsevier, Amsterdam, s. 567-578.
- Hair E., Anderson E., Tatham L., Black C. (1998), *Multivariate Data Analysis*, New Jersey, Prentice Hall.
- Wójcik T. (1965), *Zarys teorii klasyfikacji*, Wydawnictwo Naukowe PWN, Warszawa.

SYMBOLIC OBJECTS DISSIMILARITY MEASURES. GOWDA AND DIDAY IDEA

Summary

This article presents symbolic dissimilarity measure proposed by Gowda and Diday. The aim of the article is to present all the parts of this idea and to characterize them. Article presents also suggestions of polish translations for words used by Gowda and Diday. At the end of the article some weaknesses of this idea were presented the possible usage of this idea was presented too.

Mgr Marcin Pelka jest pracownikiem Katedry Ekonometrii i Informatyki (Wałbrzych) Akademii Ekonomicznej we Wrocławiu.