PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 86 — 2009
Wizualizacja wyników badań marketingowych – podejścia, metody i zastosowania

Marcin Pełka
Uniwersytet Ekonomiczny we Wrocławiu

PODEJŚCIA W SKALOWANIU WIELOWYMIAROWYM OBIEKTÓW SYMBOLICZNYCH

1. Wstęp

Metody skalowania wielowymiarowego obiektów symbolicznych wymagają, aby danymi wejściowymi była albo macierz odległości w postaci przedziałów liczbowych, albo obiekty symboliczne opisywane zmiennymi interwałowymi (por. [Denœux, Masson 2000; Groenen i in. 2005; 2006]). Podejście to nazywa się podejściem „symbolicznym" ze względu na wykorzystywanie pełnej informacji o obiektach. Drugim z podejść jest podejście „klasyczne", znane również jako podejście „symboliczne – numeryczne – symboliczne" (zaproponowane przez Didaya w 1987 r.). W tym podejściu dokonuje się kodowania zmiennych symbolicznych w klasyczne, a następnie przeprowadza się skalowanie wielowymiarowe dla danych klasycznych (metryczne lub niemetryczne). Ostatnim z podejść jest podejście hybrydowe, w którym na podstawie tablicy danych symbolicznych oblicza się macierz odległości w rozumieniu klasycznym i następnie przeprowadza się niemetryczne skalowanie wielowymiarowe dla obiektów w ujęciu klasycznym.

W artykule zaprezentowano możliwe podejścia w skalowaniu wielowymiarowym obiektów symbolicznych. Wskazano również na problemy i ograniczenia, jakie mogą występować w każdym z podejść. W opracowaniu na podstawie symulacyjnych danych symbolicznych wygenerowanych z wykorzystaniem z funkcji `cluster.Gen` pakietu `clusterSim` (zob. [Walesiak, Dudek 2008]) oraz zbiorów danych symbolicznych z programu `SODAS` przedstawiono podejścia w skalowaniu wielowymiarowym obiektów symbolicznych.

2. Zmienne symboliczne

W przypadku obiektów symbolicznych możemy mieć do czynienia z następującymi rodzajami zmiennych [Bock, Diday 2000, s. 2-3; Diday, Billard 2006, s. 16-22] (por. tab. 1):

1) ilorazowe, przedziałowe, porządkowe, nominalne,
2) kategorie, czyli dane tekstowe,
3) zmienne interwałowe – czyli przedziały liczbowe rozłączne lub nierozłączne,
4) zmienne wielowariantowe,
5) zmienne wielowariantowe z wagami,
6) zmienne interwałowe z wagami,
7) zmienne strukturalne [Bock, Diday 2000, s. 33-37; Diday, Billard 2006, s. 30--34]; w literaturze wyróżnia się oprócz wyżej wymienionych typów zmiennych także zmienne strukturalne:

a) zmienne o zależności funkcyjnej lub logicznej pomiędzy zmiennymi, gdzie *a priori* ustalono reguły funkcyjne lub logiczne decydujące o tym, jaką wartość przyjmie dana zmienna,

b) zmienne hierarchiczne, w których *a priori* ustalono warunki, od których zależy, czy zmienna dotyczy danego obiektu, czy też nie,

c) zmienne taksonomiczne, w których *a priori* ustalono systematykę, według której przyjmuje ona swoje realizacje.

Tabela 1. Przykładowe zmienne symboliczne.

Zmienna symboliczna	Realizacje zmiennej symbolicznej	Typ zmiennej symbolicznej
Preferowana cena samochodu	<25000; 36000>, <28000; 37000>, <30000; 50000>, <33000; 58000>, <65000; 80000>, <66000; 90000>	zmienna interwałowa (przedziały liczbowe nierozłączne)
Pojemność silnika	<1000; 1200>, (1200; 1400>, (1400; 1600>, (1600; 1800>, (1800; 2000>, (2000; 2200>	zmienna interwałowa (przedziały liczbowe rozłączne)
Kolor	{zielony, niebieski, biały, żółty, czarny, czerwony}	zmienna wielowariantowa
Preferowana marka samochodu	{60% Honda, 35% Toyota, 5% Audi} {40% Honda, 20% Skoda, 20% Toyota, 20% Audi} {80% Audi, 15% Opel, 5% Toyota}	zmienna wielowariantowa z wagami
Preferowana pojemność bagażnika	{[300; 600) 15%, (600; 850] 25%, (800; 1500] 30%, (1500; 3000] 30%}	zmienna interwałowa z wagami

Źródło: opracowanie własne (dane fikcyjne).

W analizie danych symbolicznych mamy do czynienia albo z obiektami symbolicznymi pierwszego rzędu (*first-order objects, simple objects*) – są to elementarne jednostki badania rozumiane w podobny sposób jak obiekty w ujęciu klasycznym (pojedynczy respondent, produkt itp.), lecz obiekty te są opisywane przez zmienne symboliczne, albo też z obiektami symbolicznymi drugiego rzędu (*complex objects, aggregate objects, super-individuals*) – to mniej lub bardziej homogeniczne grupy (agregaty, złożenia) obiektów w ujęciu klasycznym lub obiektów symbolicznych pierwszego rzędu.

3. Skalowanie wielowymiarowe obiektów symbolicznych

W skalowaniu wielowymiarowym obiektów symbolicznych, podobnie jak w przypadku innych metod analizy danych symbolicznych, można zastosować dwa główne podejścia:

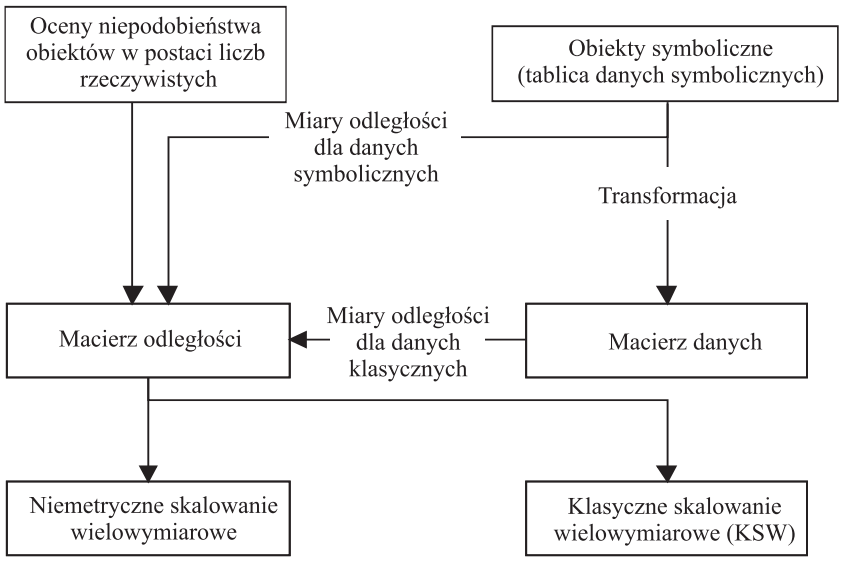
1. Podejście „klasyczne", w którym dokonywana jest transformacja zmiennych symbolicznych w klasyczne, a następnie przeprowadzane jest skalowanie wielowymiarowe z zastosowaniem metod skalowania wielowymiarowego obiektów w ujęciu klasycznym (metody metryczne lub niemetryczne) (por. rys. 1). Problemem w tym przypadku jest dobór metod transformacji zmiennych symbolicznych w klasyczne, a dodatkowo podejście to prowadzi do utraty części informacji o obiektach (zob. tab. 2).

Tabela 2. Transformacja zmiennych symbolicznych w klasyczne

Zmienne symboliczne/ klasyczne	Realizacje zmiennej symbolicznej/klasycznej	Typ zmiennej symbolicznej/klasycznej	Transformacja
Preferowana cena w tys. zł	<25; 36>, <28; 37>, <30; 50>, <33; 58>, <65; 80>, <66; 90>	zmienna interwałowa	kodowanie rozmyte
Wybrane pojemności silnika	<1000; 1200>, (1200; 1400>, (1400; 1600>, (1600; 1800>, (1800; 2000>, (2000; 2200>	zmienna interwałowa	kodowanie przedziałów, np.: (1000; 1200> = 1; (1200; 1400> = 2 (1400; 1600> = 3; (1600; 1800> = 4 (1800; 2000> = 5; (2000; 2200> = 6
Preferowany kolor	{zielony, niebieski, biały, żółty, czarny, czerwony}	zmienna wielowariantowa	wprowadzenie zmiennych binarnych w liczbie równej liczbie kategorii
Preferowana marka	{60% Honda, 40% Toyota} {40% Honda, 20% Skoda, 20% Toyota, 20% Audi} {80% Audi, 20% Opel}	zmienna wielowariantowa z wagami	wprowadzenie zmiennych ilorazowych w liczbie równej liczbie kategorii, ich realizacjami będą prawdopodobieństwa (wagi)
Liczba drzwi	2, 3, 4, 5	zmienna klasyczna (ilorazowa)	bez zmian

Źródło: opracowanie własne.

Pewną odmianą strategii klasycznej jest strategia „hybrydowa". Polega ona na zastosowaniu do tablicy danych symbolicznych miary odległości adekwatnej dla tego typu danych [Bock, Diday 2000, s. 166-183; Diday, Billard 2006, s. 231-248]. Następnie na podstawie otrzymanej macierzy odległości przeprowadzane jest skalowanie wielowymiarowe (metody metryczne lub niemetryczne) (por. rys. 1). W tym



Rys. 1. Podejście „klasyczne" w skalowaniu wielowymiarowym

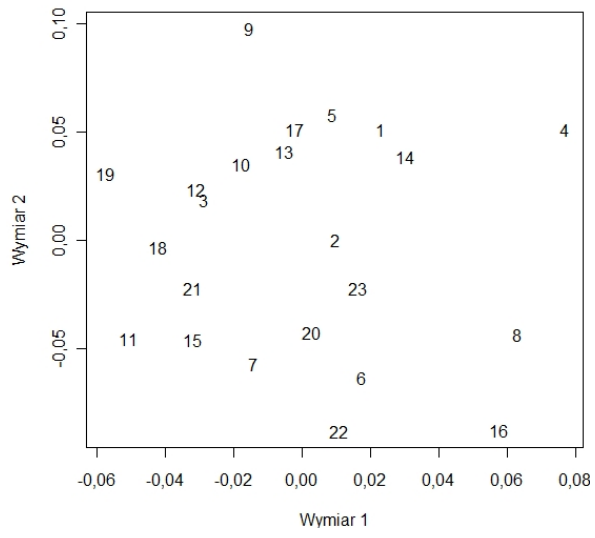
Źródło: opracowanie własne na podstawie [Groenen i in. 2005; 2006].

przypadku nie mamy do czynienia z utratą informacji o obiektach, a wynikiem skalowania wielowymiarowego są punkty reprezentujące obiekty symboliczne. Problematyczne wydaje się przedstawianie obiektów symbolicznych jako punktów, podczas gdy w przestrzeni wielowymiarowej ze względu na zmienne je opisujące (zob. tab. 1) obiekty te nie są punktami.

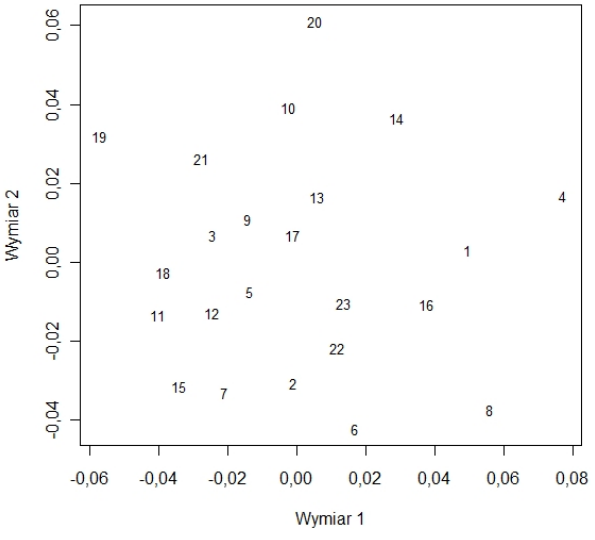
Podejście tego typu dostępne jest w programie SODAS po nazwą *bi-dimensional mapping* i jest adaptacją odwzorowania Sammona dla danych symbolicznych (por. [Pełka 2007a; i 2007b]).

Podejście klasyczne oparte na transformacji zmiennych oraz hybrydowe bazujące na macierzy odległości zastosowano do danych dotyczących 23 gatunków win (obiekty symboliczne drugiego rzędu), które są opisywane 28 zmiennymi symbolicznymi różnych typów (dane te zawiera plik `vine.sds` dostępny w programie `SODAS` w wersji `2.0`). Wyniki skalowania wielowymiarowego dla obu podejść przedstawiono na rys. 2.

Dodatkową zaletą metody „klasycznej" i „hybrydowej" jest możliwość skorzystania z funkcji kary w skalowaniu wielowymiarowym w celu uniknięcia rozwiązań zdegenerowanych. Z rysunku 2 wynika, że lepszym rozwiązaniem, w rozumieniu funkcji dopasowania STRESS, jest podejście hybrydowe. Jednakże w podejściu zarówno hybrydowym, jak i klasycznym trudno zidentyfikować jednoznacznie grupy (klasy) obiektów podobnych.



a)



b)

Rys. 2. Wyniki skalowania wielowymiarowego – podejście klasyczne i hybrydowe

a) wyniki skalowania wielowymiarowego – podejście z transformacją zmiennych („klasyczne"), STRESS = 14,11%

b) wyniki skalowania wielowymiarowego – podejście bez transformacji zmiennych („hybrydowe"), STRESS = 5%.

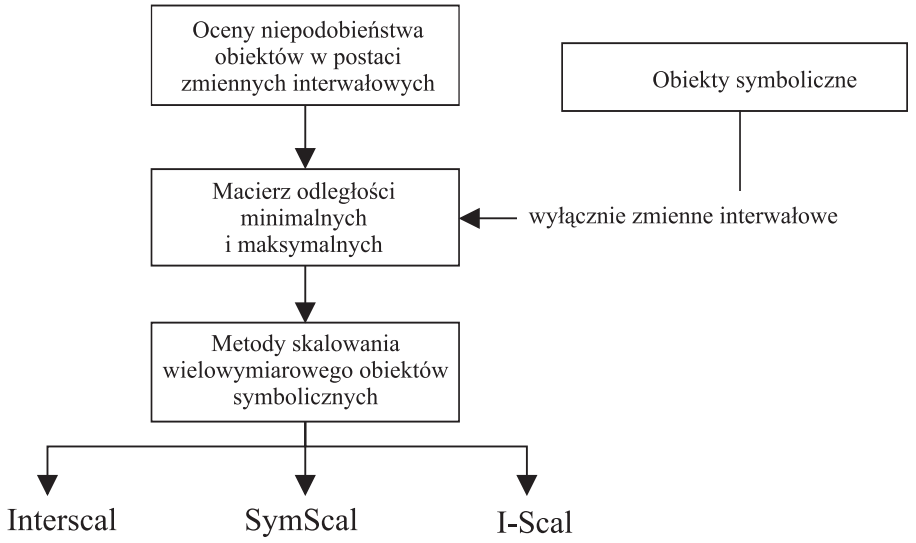
Źródło: obliczenia własne z wykorzystaniem programów SODAS i R.

2. Podejście symboliczne, w którym wykorzystywana jest macierz odległości minimalnych i maksymalnych (por. [Pełka, Dudek 2008, s. 455]). Macierz ta może być otrzymana albo na podstawie tablicy danych symbolicznych (w której obiekty opisywane są wyłącznie zmiennymi interwałowymi), albo oceny niepodobieństw między obiektami symbolicznymi (por. rys. 3). Szczegółowo podejście to zaprezentowane jest w pracy [Pełka, Dudek 2008]. W podejściu symbolicznym wykorzystywane są następujące metody skalowania wielowymiarowego (zob. [Pełka, Dudek 2008, s. 457; Denœux, Masson 2000; Groenen i in. 2005; 2006]) (por. rys. 4):

a) Interscal, która jest adaptacją klasycznego skalowania wielowymiarowego dla danych symbolicznych,

b) SymScal, która jest adaptacją niemetrycznego skalowania wielowymiarowego dla danych symbolicznych,

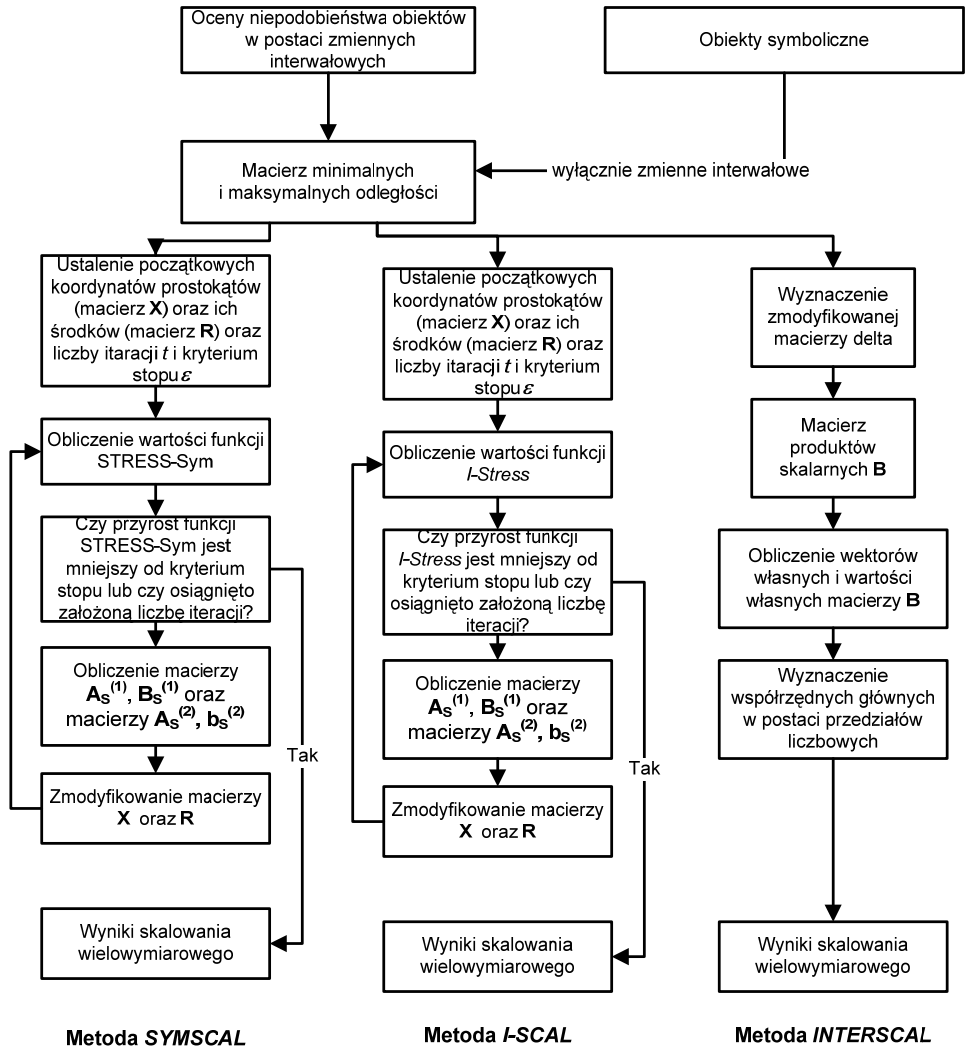
c) I-Scal, która również jest adaptacją niemetrycznego skalowania wielowymiarowego dla danych symbolicznych. Jednakże w odróżnieniu od metod Interscal i SymScal w metodzie tej zaproponowano miarę oceny otrzymanego rozwiązania – I-STRESS, który jest interpretowany podobnie jak zwykły STRESS.



Rys. 3. Podejście symboliczne w skalowaniu wielowymiarowym

Źródło: opracowanie własne na podstawie [Groenen 2005; 2006; Denœux, Masson 2000].

W tym podejściu można również wykorzystać tablicę danych symbolicznych, w której obiekty są opisywane różnymi zmiennymi. Następnie tablica taka mogłaby stanowić podstawę do oceny niepodobieństw między obiektami symbolicznymi przez ekspertów.



Rys. 4. Metody skalowania wielowymiarowego w podejściu symbolicznym

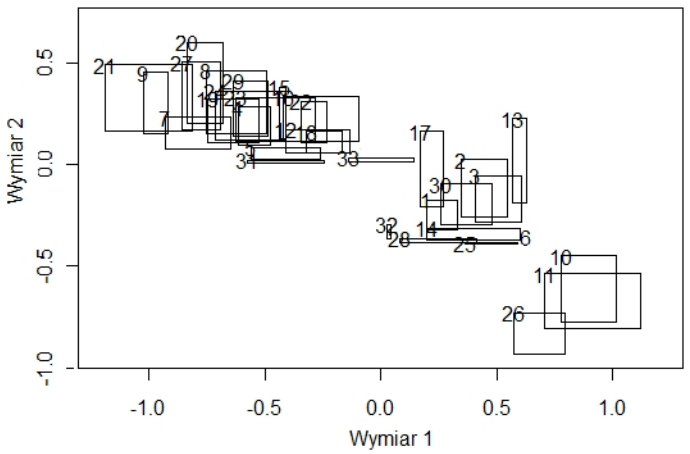
Źródło: [Pełka, Dudek 2008, s. 457].

Wynikiem skalowania wielowymiarowego obiektów symbolicznych w tym podejściu są prostokąty reprezentujące obiekty symboliczne. Pewnym problemem może być tu ostateczna interpretacja wyników oraz brak funkcji kary dla rozwiązań zdegenerowanych.

Podejście symboliczne zastosowano do danych dotyczących 33 marek samochodów osobowych (obiekty symboliczne drugiego rzędu), które są opisywane 8 interwałowymi zmiennymi symbolicznymi (dane te zawiera plik `car.sds` dostępny

w programie SODAS w wersji 2.0). Wyniki skalowania wielowymiarowego dla obu podejść przedstawiono na rys. 5.

W tym przypadku można wyróżnić trzy klasy. Wartość funkcji I-STRESS wyniosła 3,13%, co pozwala uznać to rozwiązanie za bardzo dobre.



Rys. 5. Wyniki skalowania wielowymiarowego – podejście symboliczne

Źródło: obliczenia własne z zastosowaniem programu R.

Tabela 3. Wyniki badań symulacyjnych

Lp. modelu	Metoda	Liczba zmiennych zakłócających			
		0	1	2	5
1	Interscal	0,1783	0,2114	0,2871	0,3926
	I-Scal[a]	0,0802	0,1002	0,1826	0,2395
	I-Scal[b]	0,0712	0,1012	0,1434	0,2212
2	Interscal	0,1721	0,2113	0,2711	0,4091
	I-Scal[a]	0,0531	0,0982	0,1394	0,2092
	I-Scal[b]	0,0530	0,1002	0,1232	0,1983
3	Interscal	0,1812	0,2345	0,2873	0,3178
	I-Scal[a]	0,0923	0,1099	0,1666	0,2433
	I-Scal[b]	0,0931	0,0991	0,1582	0,2277
4	Interscal	0,1639	0,2478	0,2764	0,3674
	I-Scal[a]	0,0982	0,1333	0,1877	0,2369
	I-Scal[b]	0,0801	0,0988	0,1594	0,2109

[a] – metoda I-Scal, w której początkowe współrzędne prostokątów są dobierane losowo.
[b] – metoda I-Scal, w której początkowe współrzędne prostokątów pochodzą z metody Interscal.

Źródło: obliczenia własne z wykorzystaniem programu R.

Na potrzeby niniejszego artykułu przeprowadzono skalowanie wielowymiarowe obiektów symbolicznych z wykorzystaniem metod Interscal oraz I-Scal na podstawie symulacyjnych danych symbolicznych wygenerowanych z wykorzystaniem z funkcji `cluster.Gen` pakietu `clusterSim` (zob. [Walesiak, Dudek 2008]), zawierających różną liczbę klas i zmiennych zakłócających. Zastosowano również rozwiązanie, w którym wyniki uzyskane za pomocą metody Interscal wykorzystywane są jako rozwiązanie początkowe w metodzie I-Scal. Wyniki tych symulacji zawarto w tab. 3.

Wyniki zestawione w tab. 3 nie wskazują jednoznacznie, która z metod skalowania wielowymiarowego obiektów symbolicznych daje lepsze wyniki w sytuacji, gdy w zbiorze zmiennych znajdują się zmienne zakłócające. Podobnie niejednoznaczne wyniki dały symulacje dla zbiorów ze zmiennymi zakłócającymi oraz obserwacjami odstającymi.

4. Wnioski końcowe

Skalowanie wielowymiarowe obiektów symbolicznych, podobnie jak skalowanie wielowymiarowe dla obiektów w ujęciu klasycznym, może służyć m.in. do:
- określania pozycji rynkowej produktu na tle produktów konkurencyjnych,
- poszukiwania luk na rynku,
- segmentacji rynku,
- oceny nowo wprowadzanych produktów na rynek w stosunku do produktów konkurencyjnych,
- określania struktury rynku,
- oceny haseł reklamowych.

W metodach skalowania wielowymiarowego obiektów symbolicznych wykorzystywać można zarówno podejście klasyczne (z transformacją danych lub bez), jak i podejście symboliczne. W podejściu klasycznym obiekty symboliczne są przedstawiane na płaszczyźnie jako punkty. Rozwiązanie to nie jest idealne ze względu na to, że obiekty te nie są punktami w przestrzeni wielowymiarowej. Jednakże jest to jedyne rozwiązanie w sytuacji, gdy obiekty są opisywane przez zmienne symboliczne różnych typów, a nie godzimy się na utratę części informacji przez odrzucenie innych zmiennych niż interwałowe przy założeniu, że nie możemy pozyskać ocen obiektów symbolicznych w postaci macierzy odległości interwałowych.

W podejściu symbolicznym obiekty symboliczne traktowane są jako prostokąty. Podejście to jest adekwatne w sytuacji, gdy mamy do czynienia wyłącznie ze zmiennymi interwałowymi w tablicy danych symbolicznych lub możemy otrzymać ocenę podobieństwa obiektów w postaci macierzy odległości interwałowych.

Wśród przedstawionych przykładów i zastosowanych w nich metod za najlepsze rozwiązanie w podejściu klasycznym należy uznać metody bazujące na miarach odległości adekwatnych dla danych symbolicznych. W podejściu symbolicznym na

podstawie wcześniejszych doświadczeń i symulacji za najbardziej adekwatną należy uznać metodę I-Scal.

Obszarem dla przyszłych badań powinna stać się funkcja kary dla rozwiązań zdegenerowanych w skalowaniu wielowymiarowym obiektów symbolicznych.

Literatura

Bock H.-H., Diday E. (red.), *Analysis of Symbolic Data. Explanatory Methods for Extracting Statistical Information from Complex Data*, Springer Verlag, Berlin-Heidelberg 2000.

Denœux T., Masson M., *Multidimensional Scaling of Interval-Valued Dissimilarity Data*. „Pattern Recognition Letters" 2000 vol. 21, issue 1, s. 83-92.

Diday E., Billard L., *Symbolic Data Analysis. Conceptual Statistics and Data Mining*, Wiley, Chichester 2006.

Groenen P.J.F., Winsberg S., Rodríguez O., Diday E., *Multidimensional Scaling of Interval Dissimilarities*. „Econometric Report" 2005 nr 15, Erasmus University, Rotterdam 2005.

Groenen P.J.F., Winsberg S., Rodriguez O., Diday E., *I-Scal: Multidimensional Scaling of Interval Dissimilarities*, „Computational Statistics and Data Analysis" 2006 vol. 51, s. 360-378.

Pełka M., *Metody skalowania wielowymiarowego obiektów symbolicznych*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1169, AE, Wrocław 2007a, s. 178-185.

Pełka M., *Zastosowanie nieliniowego odwzorowania Sammona dla obiektów symbolicznych do wstępnej oceny konkurencyjności*, [w:] *Badanie konkurencji i konkurencyjności przedsiębiorstw i produktów na rynku*, S. Mynarski (red.), UE, Kraków 2007b, s. 135-142.

Pełka M., Dudek A., *SymScal: metoda skalowania wielowymiarowego obiektów symbolicznych*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 7 (1207), UE, Wrocław 2008, s. 454--461.

Walesiak M., Dudek A., *ClusterSim package*, http://www.R-project.org, 2008.

APPROACHES IN SYMBOLIC MULTIDIMENSIONAL SCALING

Summary

The aim of the paper is to present and compare approaches in symbolic multidimensional scaling for symbolic data. The article presents basic terms of symbolic data analysis, the classical, hybrid and symbolic approach for symbolic multidimensional scaling.

The article also presents results obtained with these approaches on artificial symbolic data and data from SODAS software.