PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 7 (1207) 2008
TAKSONOMIA 15
Klasyfikacja i analiza danych – teoria i zastosowania

Andrzej Dudek, Marcin Pełka
Uniwersytet Ekonomiczny we Wrocławiu

SYMSCAL: METODA SKALOWANIA WIELOWYMIAROWEGO OBIEKTÓW SYMBOLICZNYCH

1. Wstęp

Ideą skalowania wielowymiarowego obiektów symbolicznych jest przedstawienie relacji zachodzących między obiektami traktowanymi w przypadku obiektów symbolicznych jako hiperprostopadłościany w przestrzeni wielowymiarowej.

Większość metod skalowania wielowymiarowego obiektów symbolicznych (*symbolic multidimensional scaling*) wymaga, aby dane wejściowe stanowiła macierz odległości minimalnych i maksymalnych pomiędzy obiektami symbolicznymi.

W artykule zaprezentowano metodę skalowania wielowymiarowego *SymScal*, którą zaproponowali: P.J.F. Groenen, S. Winsberg, O. Rodríguez, E. Diday [2005]. W artykule porównano ją z innymi metodami skalowania wielowymiarowego obiektów symbolicznych, opisano również problemy, jakie mogą wynikać z zastosowania metody *SymScal*. Opracowanie w części empirycznej prezentuje wyniki skalowania wielowymiarowego uzyskane na przykładzie danych symbolicznych pochodzących z rynku komputerowego.

2. Typy zmiennych w symbolicznej analizie danych

W przypadku obiektów 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. silnik benzynowy, silnik elektryczny;

3) przedziały liczbowe, np. temperatury dzienne w pewnym regionie (–10, ..., 25) czy ilość spalanej benzyny na 100 km (5 litrów, ..., 8 litrów), co w przypadku spalania benzyny oznacza, że samochód spala 5 litrów w cyklu ekonomicznym poza miastem, a 8 litrów w cyklu miejskim;

4) lista kategorii – tu przykładem może być typ nadwozia samochodowego oferowany dla pewnego modelu samochodu (obiektu), określony jako: sedan, hatchback, kombi;

5) lista kategorii z wagami (prawdopodobieństwami), gdzie oprócz listy kategorii występują wagi, z jakimi obiekt posiada wybraną kategorię;

6) zmienne strukturalne [*Analysis of...* 2000, s. 33-37] – w skład tej grupy wchodzą:

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.

3. Metody skalowania wielowymiarowego obiektów symbolicznych

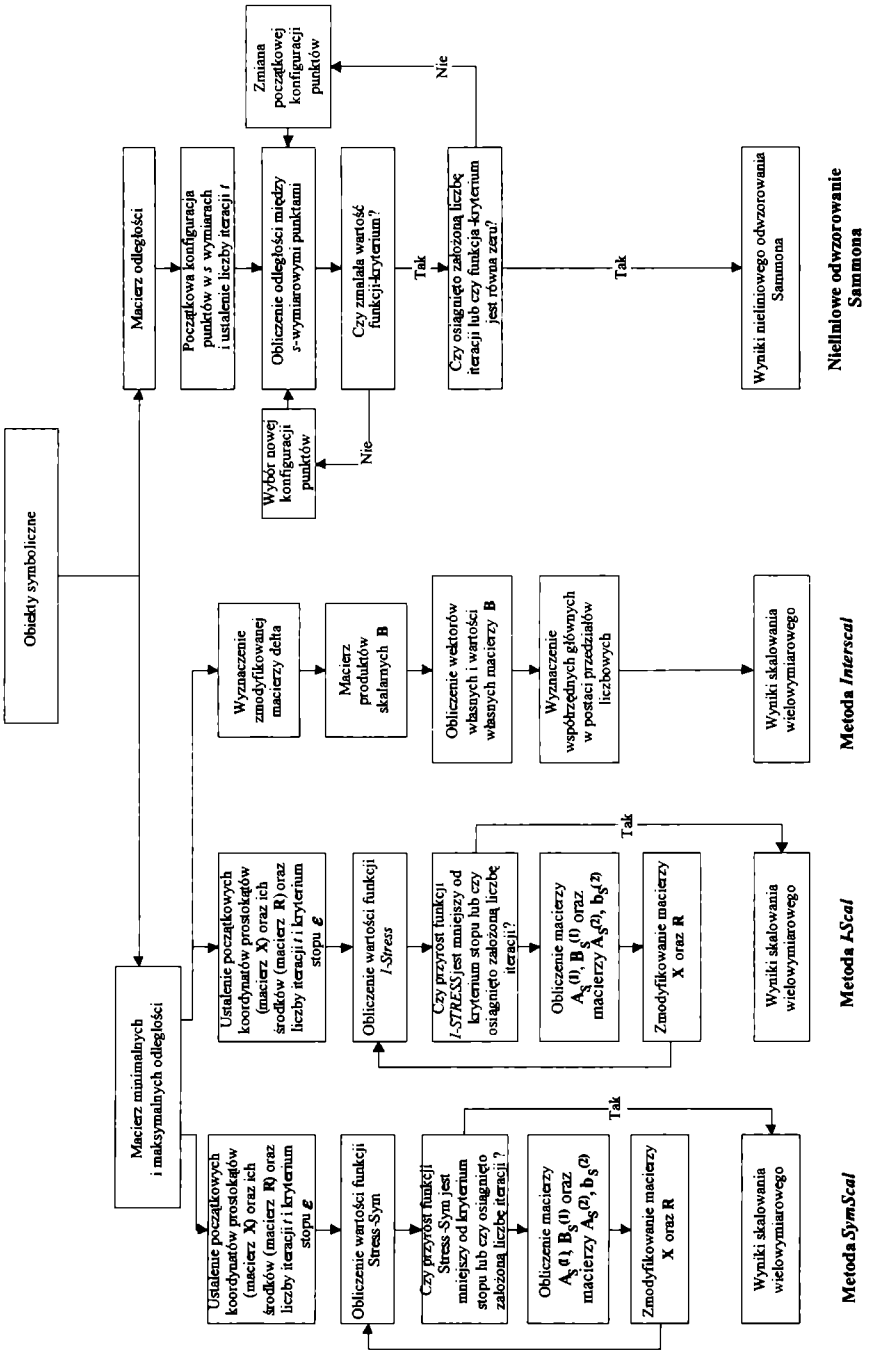
W przypadku nieliniowego odwzorowania Sammona danymi wejściowymi są albo obiekty opisywane przez dowolne zmienne symboliczne, albo macierz odległości. Wynikiem interpretacji są obiekty symboliczne traktowane w przestrzeni wielowymiarowej jako punkty. Szczegółowe omówienie tej procedury zawarto w pracy [Pełka 2007a].

Metody *Interscal*, *SymScal*, *I-Scal* oraz *3-Way SymScal* wymagają, aby danymi wejściowymi były albo obiekty symboliczne opisywane wyłączne przez przedziały liczbowe (co oznacza znaczne ograniczenie możliwości opisu zjawiska), albo macierz odległości minimalnych i maksymalnych (zwana w metodzie *Interscal* macierzą delta). Metoda *Interscal* została omówiona w pracy [Pełka 2007a]. Metoda *3-Way SymScal* jest modyfikacją metody *SymScal* zaproponowaną w 2006 r. przez zespół pod kierownictwem P.J.F. Groenena. Modyfikacja ta pozwala na wprowadzane wag do skalowania wielowymiarowego oraz proponuje nieznormalizowany miernik oceny dopasowania odwzorowania *I-Stress*, który następnie, po normalizacji, został zaproponowany przez zespół pod kierownictwem P.J.F. Groenena w 2007 r.

Na rys. 1 dokonano zestawienia metod skalowania wielowymiarowego obiektów symbolicznych.

Wszystkie te metody zakładają, że obiekty w przestrzeni wielowymiarowej są hiperprostopadłościanami, natomiast w przestrzeni o mniejszej liczbie wymiarów s (zazwyczaj $s = 2$ lub $s = 3$) są to prostopadłościany (dla $s = 3$) lub prostokąty (dla $s = 2$).

Jeżeli danymi wejściowymi ma być macierz odległości minimalnych i maksymalnych pomiędzy obiektami, to może być ona uzyskana na podstawie:



Rys. 1. Metody skalowania wielowymiarowego obiektów symbolicznych

Źródło: opracowanie własne na podstawie [Denoeux, Masson 2000; Groenen i in. 2005; Groenen, Winsberg 2006; Sammon 1969].

1) zmiennych opisujących obiekty. Jeżeli w zbiorze zmiennych opisujących obiekty dominują zmienne w postaci przedziałów liczbowych, to na podstawie propozycji zawartej w metodzie *Interscal* (por. [*Scientific Report*... 2001, s. 57; Pełka 2007a, s. 180-181]) można obliczyć macierz odległości minimalnych i maksymalnych. Podejście to wydaje się problematyczne w zastosowaniu, jeżeli przedziały liczbowe nie są jedynym typem zmiennych symbolicznych opisujących obiekty. W takim przypadku zastosowanie tego podejścia wymaga usunięcia pozostałych zmiennych i utratę informacji o obiektach;

2) wiedzy i opinii ekspertów lub respondentów, którzy oceniają podobieństwo (lub niepodobieństwo) obiektów. Najczęściej ocena taka jest podawana dla poszczególnych par obiektów symbolicznych.

Jeżeli mamy do czynienia z wiedzą i opinią ekspertów lub respondentów, to najczęściej korzystamy z wielu źródeł takich opinii. Oznacza to, że otrzymujemy wiele różnych osądów (ocen) co do wzajemnego podobieństwa obiektów. W tej sytuacji możemy (por. [*Scientific Report*... 2001, s. 56]):

a) ustalić na podstawie własnej wiedzy ostateczną macierz odległości minimalnych i maksymalnych między obiektami,

b) przyjąć 5% długości przedziału otrzymanych ocen jako odległość minimalną, a 95% długości tego przedziału jako odległość maksymalną,

c) przyjąć kwartyl pierwszy ocen ekspertów jako odległość minimalną, a kwartyl trzeci jako odległość maksymalną.

Ideą metody *SymScal*, a także metod *3-Way SymScal*, *I-Scal* jest majoryzacja funkcji dopasowania odwzorowania. W przypadku metody *SymScal* miarą dopasowania jest nieznormalizowana funkcja *Stress-Sym*, dla metody *3-Way SymScal* jest to nieznormalizowany miernik *I-Stress*, a dla metody *I-Scal* jest to znormalizowany w przedziale [0; 1] miernik *I-Stress*.

Niezależnie od rodzaju funkcji oceniającej dopasowanie kluczowymi krokami są: ustalenie liczby iteracji t (w literaturze najczęściej proponuje się używanie krotności 100 iteracji) oraz ustalenie macierzy początkowych współrzędnych prostokątów – macierz $\mathbf{X}$, macierzy początkowych długości boków tych prostokątów – macierz $\mathbf{R}$ oraz kryterium stopu ε (najczęściej proponowaną wartością jest wartość 10^{-6}).

Elementy macierzy $\mathbf{X}$ oraz $\mathbf{R}$ można ustalić na podstawie własnej wiedzy, wiedzy i opinii ekspertów, w sposób losowy (np. z rozkładu normalnego). Autorzy metody *SymScal* i *I-Scal* proponują także dokonać skalowania metodą *Interscal* i wykorzystać jej wyniki do budowy macierzy $\mathbf{X}$ i $\mathbf{R}$. Kolejnym istotnym krokiem jest obliczenie miary *Stress-Sym*:

$$Stress\text{-}Sym\left(\mathbf{X},\mathbf{R}\right)=\sum_{i<j}^{n}\omega_{ij}\left[\overline{\delta}_{ij}-\overline{d}_{ij}(\mathbf{X},\mathbf{R})\right]^2+\sum_{i<j}^{n}\omega_{ij}\left[\underline{\delta}_{ij}-\underline{d}_{ij}(\mathbf{X},\mathbf{R})\right]^2, \quad (1)$$

gdzie: $\mathbf{X},\mathbf{R}$ – macierze początkowych współrzędnych prostokątów,

ω_{ij} – wagi związane ze wzajemnym położeniem i-tego i j-tego obiektu symbolicznego (najczęściej ustalone wagi są sobie równe),

$\overline{\delta}_{ij}, (\underline{\delta}_{ij})$ – odległość maksymalna (odpowiednio odległość minimalna) między i-tym a j-tym obiektem symbolicznym wynikająca z macierzy odległości minimalnych i maksymalnych,

$\overline{d}_{ij}(\mathbf{X},\mathbf{R}), (\underline{d}_{ij}(\mathbf{X},\mathbf{R}))$ – odległość maksymalna (odpowiednio odległość minimalna) między i-tym a j-tym obiektem symbolicznym obliczona na podstawie macierzy $\mathbf{X}$ i $\mathbf{R}$ (zob. wzór (2) i (3)),

$$\overline{d}_{ij}(\mathbf{X},\mathbf{R}) = \sqrt{\sum_{s=1}^{p}\left[\left|x_{is} - x_{js}\right| + \left(r_{is} + r_{js}\right)\right]^2}, \tag{2}$$

$$\underline{d}_{ij}(\mathbf{X},\mathbf{R}) = \sqrt{\sum_{s=1}^{p}\max\left[0, \left|x_{is} - x_{js}\right| + \left(r_{is} + r_{js}\right)\right]^2}, \tag{3}$$

$s = 1, \ldots, p$ – wymiar przestrzeni, w której dokonywana jest reprezentacja obiektów (zazwyczaj $s = 1 \vee 2$).

Metoda majoryzacji funkcji *Stress-Sym* jest przedstawiona w pracy [Groenen, Winsberg 2006, s. 18-26]. Algorytm metody *SymScal* został oprogramowany przez autorów publikacji w skrypcie dla programu *R*[1]. W obecnej wersji pliku macierze $\mathbf{X}$ i $\mathbf{R}$ są wyznaczane za pomocą procedury `sample` programu *R*, liczbę iteracji ustalono na 200, a kryterium stopu jako 10^{-6}.

4. Zastosowanie algorytmu *SymScal* w ocenie rynku monitorów LCD

W badaniu poproszono 116 osób zajmujących się sprzedażą i serwisowaniem sprzętu klasy PC o ocenę podobieństwa między sześcioma modelami 17" monitorów ciekłokrystalicznych (LCD). Eksperci oceniali podobieństwo ofert monitorów poszczególnych firm w skali od 0% (oferty firm X oraz Y są różne) do 100% (oferty firm X i Y są identyczne). W badaniu tym nieznana jest początkowa liczba wymiarów, ponieważ każdy z ekspertów oceniał monitory pod względem innego, subiektywnie dobranego zestawu cech.

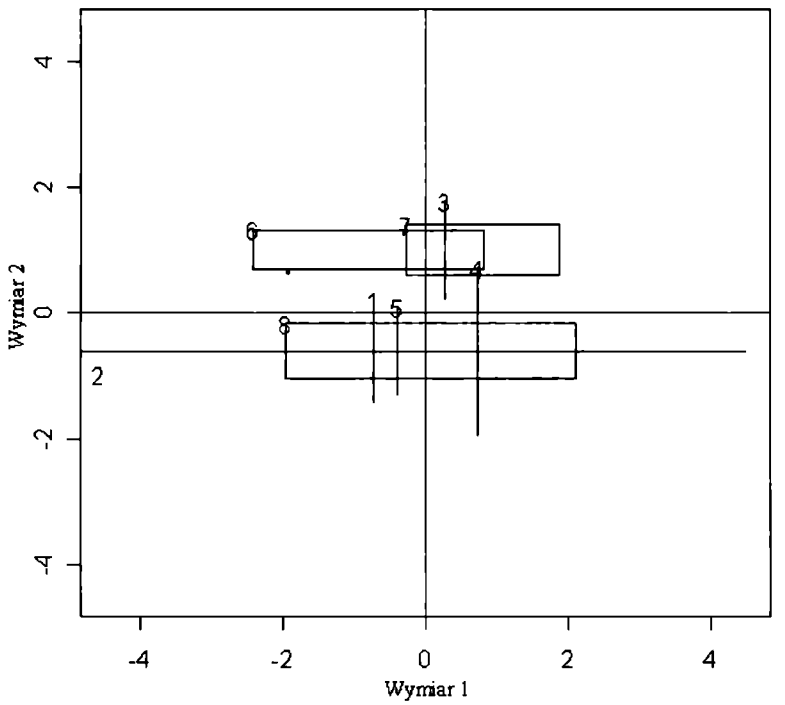
W tym przypadku 0% oznacza ofertę ubogą, o niewielkim stopniu zróżnicowania, a 100% – ofertę bogatą, o znacznym stopniu zróżnicowania. Podobieństwa przekształcono na niepodobieństwa, odejmując 100% od otrzymanych ocen. W badaniu przyjęto, że kwartyl pierwszy otrzymanych niepodobieństw stanowi odległości minimalne, a kwartyl trzeci stanowi odległości maksymalne. Do badania

[1] Skrypt dostępny jest na stronie Katedry Ekonometrii i Informatyki – http://wgrit.ae.jgora.pl/keii/.

wybrano osiem firm o największym udziale w rynku w 2006 r.: Samsung, LG, Maxdata, Philips, BenQ, NEC, Neovo, Hyundai (zob. [Kuśmierz 2006, s. 10]). Wyniki skalowania wielowymiarowego przedstawiono na rys. 2.

Miara *Stress-Sym* wyniosła w przybliżeniu 241. Ponieważ nie jest to miara znormalizowana, to nie można podać jednoznacznej interpretacji tego wyniku. Autorzy artykułów o metodach skalowania obiektów symbolicznych *SymScal*, *I-Scal* oraz *Interscal* nie podali żadnych wskazówek co do sposobu interpretacji osi skalowania. W przypadku nieliniowego odwzorowania Sammona zaznaczono, że nie można dokonać w tym przypadku interpretacji osi skalowania. Autorzy niniejszego artykułu posiłkowali się metodami interpretacji osi skalowania znanymi z metod skalowania wielowymiarowego, opracowanymi dla statystycznej analizy wielowymiarowej. Metody interpretacji osi można podzielić na:

1) metody obiektywne, polegające na badaniu (najczęściej poprzez współczynniki korelacji) współzależności (relacji) między poziomem zmiennych a wartością wymiarów. Metody te nie znajdują zastosowania wówczas, gdy danymi wejściowymi jest macierz odległości minimalnych i maksymalnych;



1 – Samsung; 2 – LG; 3 – Maxdata; 4 – Philips;
5 – Benq; 6 – NEC; 7 – Neovo; 8 – Hyundai

Rys. 2. Wyniki skalowania wielowymiarowego metodą *SymScal*
Źródło: opracowanie własne z wykorzystaniem pakietu *R*.

2) metody subiektywne, polegające na podaniu kryteriów stosowanych przy ocenie niepodobieństwa obiektów przez ekspertów czy respondentów. Kryteria te mogą być także wynikiem oceny dokonanej przez badacza. W tym przypadku wymiar 1 zinterpretowano jako rozwiązania technologiczne, a wymiar 2 jako estetykę produktu.

W przypadku gdy nieznana jest początkowa liczba wymiarów, nie można podać jednoznacznej interpretacji położenia samych obiektów. Identyfikować można natomiast grupy obiektów podobnych (klasy).

W tym przypadku za marki podobne eksperci ocenili Maxdata – 3, NEC – 6, Neovo – 7 (klasa 1) oraz Samsung – 1, BenQ – 5, Hyundai – 8 oraz Philips – 4 (klasa 2).

5. Podsumowanie

Metody skalowania wielowymiarowego obiektów symbolicznych dążą do przedstawienia zależności zachodzących między obiektami w przestrzeni wielowymiarowej w przestrzeni o mniejszej liczbie wymiarów (zazwyczaj dwu- lub trzywymiarowej), przy jak najmniejszej utracie informacji o zależnościach między tymi obiektami.

Obiekty symboliczne w metodach skalowania wielowymiarowego są traktowane nie jako punkty, lecz jako hiperprostopadłościany przestrzeni wielowymiarowej. Dlatego ich graficzną reprezentacją w dwóch wymiarach są prostokąty, a w trzech wymiarach prostopadłościany.

Istotną wadą metody *SymScal* (podobnie jak metod *I-Scal*, *Interscal*) jest ograniczenie opisu obiektów symbolicznych tylko do zmiennych w postaci przedziałów liczbowych. Jedynym sposobem na analizowanie wzajemnych relacji między obiektami opisywanymi przez zmienne różnych typów jest wykorzystanie wiedzy i opinii ekspertów.

Do oceny otrzymanych wyników w przypadku metody *SymScal* autorzy sugerują wykorzystanie nieznormalizowanej miary *Stress-Sym*. Lepszym rozwiązaniem z pewnością będzie wykorzystanie znormalizowanej miary *I-Stress*.

Bardzo istotnym krokiem w algorytmach *SymScal*, *I-Scal* oraz *3-Way SymScal* jest ustalenie macierzy **X** i **R**. Problematyka optymalnego ich ustalenia wymaga dalszych badań.

Literatura

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

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

Groenen P.J.F., Winsberg S. (2006), *3WaySym-Scal: Three-way Symbolic Multidimensional Scaling*, Econometric Report EI 2006-49, Erasmus University, Rotterdam.

Groenen P.J.F., Winsberg S., Rodríguez O., Diday E. (2005), *SymScal: Symbolic Multidimensional Scaling of Interval Dissimilarities*, Econometric Report EI 2005-15, Erasmus University, Rotterdam.

Groenen P.J.F., Winsberg S., Rodríguez O., Diday E. (2007), *I-Scal: Multidimensional Scaling of Interval Dissimilarities*, „Computational Statistics and Data Analysis", vol. 51, issue 1, s. 360-378.

Kuśmierz M. (2006), *Monitory ciekłokrystaliczne zdobywają rynek*, „Gazeta Prawna", nr 164 (1782), s. 10.

Pełka M. (2007a), *Metody skalowania wielowymiarowego obiektów symbolicznych*, [w:] Taksonomia 14, red. K. Jajuga, M. Walesiak, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1169, AE, Wrocław, s. 178-185.

Pełka M. (2007b), *Zastosowanie nieliniowego odwzorowania Sammona do wstępnej oceny konkurencyjności*, Prace Naukowe AE w Krakowie (w druku).

Sammon J. W. (1969), *A Nonlinear Mapping for Data Structure Analysis*, „IEEE Transactions on Computers", vol. C-18, nr 5, s. 401-409.

Scientific Report for Unsupervised Classification, Validation and Cluster Representation. Analysis System of Symbolic Official Data Technical Report (2000), red. Y. Lechevallier, Raport IST-2000-25161 projektu SODAS.

SYMSCAL: SYMBOLIC MULTIDIMENSIONAL SCALING METHOD

Summary

The aim of this paper is to present one of symbolic multidimensional scaling methods, the *SymScal* method along with problems which could be encountered while applying it. The article compares this method with other multidimensional scaling methods for symbolic objects. The paper presents in the empirical part the symbolic multidimensional scaling results based on LCD monitors market with the application of *R* programme script.