

Marcin Pełka

Akademia Ekonomiczna we Wrocławiu

METODY WIELOWYMIAROWEGO SKALOWANIA OBIEKTÓW SYMBOLICZNYCH

1. Wstęp

Zadaniem skalowania wielowymiarowego obiektów symbolicznych, podobnie jak w klasycznym skalowaniu wielowymiarowym, jest przedstawienie relacji zachodzących między obiektami traktowanymi w przypadku obiektów symbolicznych jako hiperprostokąściany w przestrzeni wielowymiarowej. Ze względu na możliwości geometrycznej interpretacji otrzymanych wyników jest to najczęściej przestrzeń dwuwymiarowa.

Celem artykułu jest zaprezentowanie metody skalowania wielowymiarowego *INTERSCAL*, którą zaproponowali: Denoeux i Masson w swoim artykule (por. [Denoeux, Masson 2000]). W artykule porównano ją z drugą z metod skalowania wielowymiarowego obiektów symbolicznych – metodą *SYMSCAL*, zaproponowaną przez Groenena, Winsberga, Rodrígueza i Didaya (por. [Groenen i in. 2005]). Obie metody skalowania wielowymiarowego dla obiektów symbolicznych (*symbolic multidimensional scaling*) wymagają, aby danymi wejściowymi była macierz minimalnych i maksymalnych odległości pomiędzy obiektami symbolicznymi.

W artykule zaprezentowane zostaną również problemy, jakie mogą wynikać z zastosowania metody *INTERSCAL*. Artykuł 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 [Bock, Diday 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, ..., 2 500 zł) czy ilość spalanej benzyny na 100 km (5 litrów, ..., 11 litrów), co w przypadku dochodu oznacza, że może on wynieść od 1000 zł 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 oferowany jest on w trzech rodzajach standardów wyposażenia,
- 5) lista kategorii z wagami (prawdopodobieństwami), w której oprócz listy kategorii występują wagi, z jakim obiekt ma wybraną kategorię, np. jeżeli wybierze się zmienną „typ nadwozia samochodu” pewnej grupy osób w postaci listy: sedan (0,45), *coupé* (0,20), *pick-up* (0,25), kabriolet (0,10), inne (0,00), to oznacza to, że 45% samochodów posiadanych przez te osoby to samochody o nadwoziu typu sedan, 10% zaś to kabriolety. osoba, która w ankiecie zaznaczyła, że jej samochód to kabriolet, należy właśnie do tej grupy,
- 6) zmienne strukturalne [Bock, Diday 2000, s. 33-37], 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.

3. Ogólna procedura skalowania wielowymiarowego obiektów symbolicznych

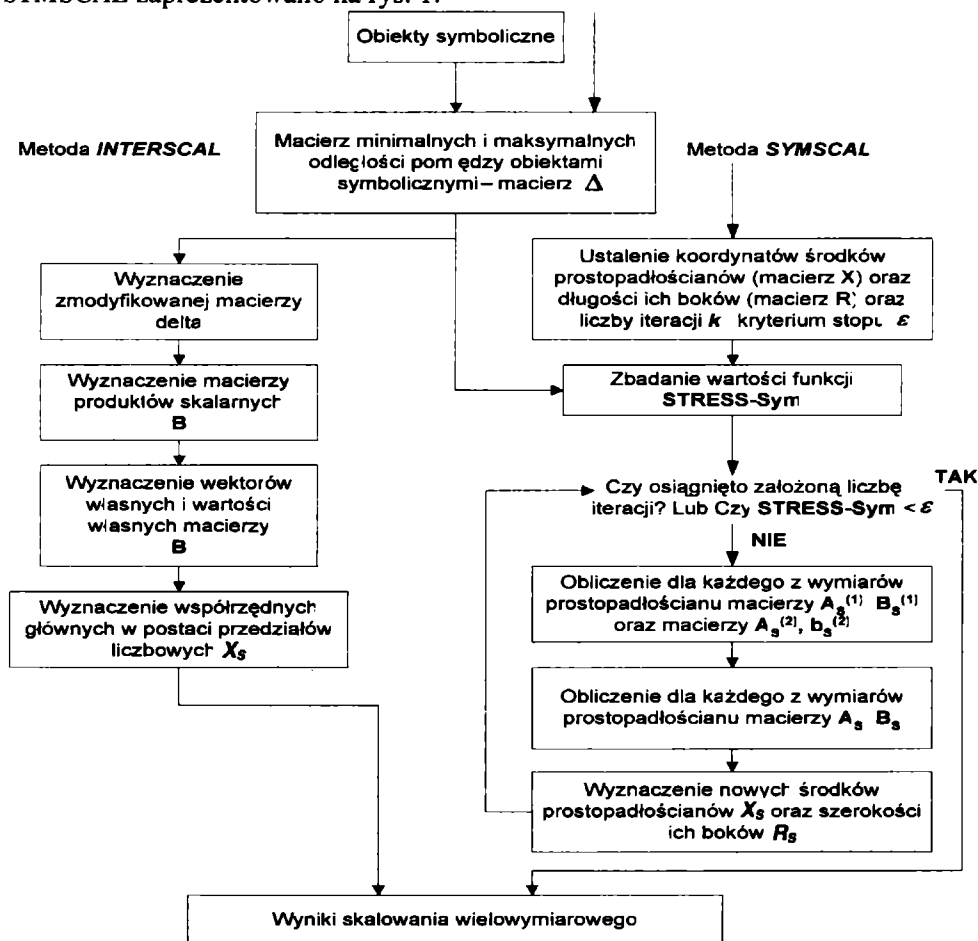
Dane wejściowe do skalowania wielowymiarowego obiektów symbolicznych są zależne od metody, którą mamy zamiar zastosować. Zależność między metodą a dopuszczalnymi przez nią danymi wejściowymi prezentuje tab. 1.

Tabela 1. Dane wejściowe do skalowania wielowymiarowego obiektów symbolicznych zależnie od metody

<i>INTERSCAL</i>	<i>SYMSCAL</i>
Obiekty opisywane zmiennymi symbolicznymi w postaci przedziałów liczbowych	obiekty opisywane zmiennymi symbolicznymi dowolnego typu
lub	
Macierz minimalnych i maksymalnych odległości pomiędzy obiektami symbolicznymi	macierz minimalnych i maksymalnych odległości pomiędzy obiektami symbolicznymi

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

Ogólną procedurę skalowania wielowymiarowego dla metody *INTERSCAL* i *SYMSCAL* zaprezentowano na rys. 1.



Rys. 1. Ogólna procedura skalowania wielowymiarowego
Źródło: opracowanie własne.

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

- 1) zmiennych opisujących obiekty (problemem może być tu uzyskanie macierzy, gdy obiekty są opisywane przez zmienne różnych typów),
- 2) wiedzy i opinii ekspertów (podstawowym problemem jest tu wybór ekspertów z danej dziedziny oraz decyzja, ilu ekspertów należy wybrać).

Gdy mamy do czynienia z wiedzą i opinią ekspertów, to często nie korzystamy z jednego źródła takich opinii, lecz z wielu źródeł. W tym przypadku możemy

otrzymać wiele różnych osądów co do wzajemnego podobieństwa obiektów. W tej sytuacji możemy (por. [Lechevallier 2000, s. 56]):

a) ustalić na podstawie własnej wiedzy ostateczną macierz minimalnych i maksymalnych odległości między obiektami; jednakże w takim przypadku wielu badaczy może otrzymać różne wyniki na podstawie tych samych źródeł,

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

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

Jeżeli danymi wejściowymi w przypadku metody *INTERSCAL* nie jest macierz minimalnych i maksymalnych odległości pomiędzy obiektami, lecz obiekty opisywane przez zmienne w postaci przedziałów liczbowych, to minimalne i maksymalne odległości pomiędzy nimi wyznacza się następująco (zob. [Lechevallier 2000, s. 57]):

$$\bar{d}_{ij} = \frac{1}{2} \sqrt{\sum_{k=1}^n \left[(\bar{x}_{ik} - \underline{x}_{ik}) + (\bar{x}_{jk} - \underline{x}_{jk}) + 2 \left| \frac{\bar{x}_{ik} - \underline{x}_{ik}}{2} - \frac{\bar{x}_{jk} - \underline{x}_{jk}}{2} \right| \right]^2}, \quad (1)$$

gdzie: $k = (1, \dots, n)$ – liczba zmiennych symbolicznych opisujących obiekty,

$i, j = (1, \dots, m)$ – liczba obiektów symbolicznych,

$\bar{x}_{ij}$ – górny kraniec przedziału zmiennej,

$\underline{x}_{ij}$ – dolny kraniec przedziału zmiennej,

$\bar{d}_{ij}$ – maksymalna odległość między i -tym a j -tym obiektem symbolicznym.

$$\underline{d}_{ij} = \frac{1}{4} \sqrt{\sum_{k=1}^n \left[(\bar{x}_{ik} - \underline{x}_{ik}) + (\bar{x}_{jk} - \underline{x}_{jk}) + 2 \left| \frac{\bar{x}_{ik} - \underline{x}_{ik}}{2} - \frac{\bar{x}_{jk} - \underline{x}_{jk}}{2} \right| \right] \sim \left[(\bar{x}_{ik} - \underline{x}_{ik}) + (\bar{x}_{jk} - \underline{x}_{jk}) + 2 \left| \frac{\bar{x}_{ik} - \underline{x}_{ik}}{2} - \frac{\bar{x}_{jk} - \underline{x}_{jk}}{2} \right| \right]^2}, \quad (2)$$

gdzie: $\underline{d}_{ij}$ – minimalna maksymalna odległość między i -tym a j -tym obiektem symbolicznym, pozostałe oznaczenia jak we wzorze (1).

Macierz minimalnych i maksymalnych odległości nazywana jest też macierzą delta (Δ). Na jej podstawie obliczana jest zmodyfikowana macierz delta ($\bar{\Delta}$) o $2m$ elementach (por. [Lechevallier 2000, s. 54]):

$$\tilde{\Delta} = \begin{bmatrix} 0 & \overline{\delta_{11}} & \underline{\delta_{12}} & \frac{\overline{\delta_{12}} + \underline{\delta_{12}}}{2} & \underline{\delta_{13}} & \frac{\overline{\delta_{13}} + \underline{\delta_{13}}}{2} & \underline{\delta_{1m}} & \frac{\overline{\delta_{1m}} + \underline{\delta_{1m}}}{2} \\ \overline{\delta_{11}} & 0 & \frac{\overline{\delta_{12}} + \underline{\delta_{12}}}{2} & \overline{\delta_{12}} & \frac{\overline{\delta_{13}} + \underline{\delta_{13}}}{2} & \overline{\delta_{13}} & \dots & \frac{\overline{\delta_{1m}} + \underline{\delta_{1m}}}{2} & \overline{\delta_{1m}} \\ \underline{\delta_{21}} & \frac{\overline{\delta_{21}} + \underline{\delta_{21}}}{2} & 0 & \overline{\delta_{22}} & \underline{\delta_{23}} & \frac{\overline{\delta_{23}} + \underline{\delta_{23}}}{2} & \underline{\delta_{2m}} & \frac{\overline{\delta_{2m}} + \underline{\delta_{2m}}}{2} \\ \frac{\overline{\delta_{21}} + \underline{\delta_{21}}}{2} & \overline{\delta_{21}} & \overline{\delta_{22}} & 0 & \frac{\overline{\delta_{23}} + \underline{\delta_{23}}}{2} & \overline{\delta_{23}} & \dots & \frac{\overline{\delta_{2m}} + \underline{\delta_{2m}}}{2} & \overline{\delta_{2m}} \\ \underline{\delta_{31}} & \frac{\overline{\delta_{31}} + \underline{\delta_{31}}}{2} & \underline{\delta_{32}} & \frac{\overline{\delta_{32}} + \underline{\delta_{32}}}{2} & 0 & \overline{\delta_{33}} & \underline{\delta_{3m}} & \frac{\overline{\delta_{3m}} + \underline{\delta_{3m}}}{2} \\ \frac{\overline{\delta_{31}} + \underline{\delta_{31}}}{2} & \overline{\delta_{31}} & \frac{\overline{\delta_{32}} + \underline{\delta_{32}}}{2} & \overline{\delta_{32}} & \overline{\delta_{33}} & 0 & \dots & \frac{\overline{\delta_{3m}} + \underline{\delta_{3m}}}{2} & \overline{\delta_{3m}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \underline{\delta_{m1}} & \frac{\overline{\delta_{m1}} + \underline{\delta_{m1}}}{2} & \underline{\delta_{m2}} & \frac{\overline{\delta_{m2}} + \underline{\delta_{m2}}}{2} & \underline{\delta_{m3}} & \frac{\overline{\delta_{m3}} + \underline{\delta_{m3}}}{2} & 0 & \underline{\delta_{mm}} \\ \frac{\overline{\delta_{m1}} + \underline{\delta_{m1}}}{2} & \overline{\delta_{m1}} & \frac{\overline{\delta_{m2}} + \underline{\delta_{m2}}}{2} & \overline{\delta_{m2}} & \frac{\overline{\delta_{m3}} + \underline{\delta_{m3}}}{2} & \overline{\delta_{m3}} & \overline{\delta_{mm}} & 0 \end{bmatrix}, \quad (3)$$

gdzie: $\underline{\delta}_{ij}$ – minimalna odległość między i -tym a j -tym obiektem symbolicznym,

$\overline{\delta}_{ij}$ – maksymalna odległość między i -tym a j -tym obiektem symbolicznym.

Zmodyfikowana macierz delta ($\tilde{\Delta}$) służy do obliczenia macierzy produktów skalarnych (macierz **B**), której elementy są obliczane zgodnie ze wzorem (por. [Lechevallier 2000, s. 55]):

$$b_{ij} = -\frac{1}{2} \left(\tilde{\delta}_{ij}^2 - \frac{1}{2m} \sum_{r=1}^{2m} \tilde{\delta}_{rj}^2 - \frac{1}{2m} \sum_{s=1}^{2m} \tilde{\delta}_{sj}^2 + \frac{1}{(2m)^2} \sum_{r=1}^{2m} \sum_{s=1}^{2m} \tilde{\delta}_{rs}^2 \right), \quad (4)$$

gdzie: $i, j = (1, \dots, 2m)$ – kolejne elementy macierzy **B**,

$r, s = (1, \dots, 2m)$ – elementy zmodyfikowanej macierzy delta,

$\tilde{\delta}$ – element zmodyfikowanej macierzy delta ($\tilde{\Delta}$).

Ostatecznie macierz **B** jest macierzą o $2m$ elementach. Następnie konieczne jest obliczenie wartości własnych $\lambda_1, \lambda_2, \dots, \lambda_{2m}$ i odpowiadających im wektorów własnych $v_1, v_2, \dots, v_{2m}$ macierzy **B**.

Wykorzystując wartości własne i wektory własne macierzy **B**, obliczamy współrzędne $2m$ punktów w przestrzeni R^n (por. [Lechevallier 2000, s. 55]):

$$x_{ir} = \sqrt{|\lambda_r|} * v_{ir}, \quad (5)$$

gdzie: $r = (1, \dots, 2m)$ – numer kolejny punktu w przestrzeni,

$i = (1, \dots, n)$ – wymiar przestrzeni.

Na podstawie $2m$ punktów przestrzeni R^n konstruowane są współrzędne główne (*principal coordinates*) w postaci przedziałów liczbowych (por. [Lechevallier 2000, s. 55]):

$$\underline{x}_{ij} = \min_{k \in \{2i-1, 2i\}} (x_{kj}), \quad (6)$$

gdzie: $i = (1, \dots, m)$ – numer kolejny obiektu symbolicznego,

$j = (1, \dots, n)$ – wymiar przestrzeni,

$\underline{x}_{ij}$ – dolny kraniec przedziału współrzędnej głównej i -tego obiektu dla j -tego wymiaru przestrzeni.

$$\overline{x}_{ij} = \max_{k \in \{2i-1, 2i\}} (x_{kj}), \quad (7)$$

gdzie: $\overline{x}_{ij}$ – górny kraniec przedziału współrzędnej głównej i -tego obiektu dla j -tego wymiaru przestrzeni.

4. Przykład empiryczny

W przykładzie empirycznym wykorzystano wyniki badania ankietowego dotyczącego oceny sześciu wybranych marek monitorów ciekłokrystalicznych (Philips, Sony, LG, BenQ, Actina, NEC) pod względem ich użyteczności i funkcjonalności. Badanie przeprowadzono wśród 83 klientów sklepów komputerowych w lipcu i sierpniu 2006 r. W tym samym okresie dwanaście osób zajmujących się serwisowaniem i sprzedażą sprzętu komputerowego poproszono o ocenę niepodobieństwa pomiędzy monitorami ciekłokrystalicznymi firm: Philips, Sony, LG, BenQ, Actina, NEC pod kątem ich użyteczności i funkcjonalności. Obie próby zostały dobrane według wygody. Celem tych badań było sprawdzenie, jak monitory wybranych firm są postrzegane przez respondentów i ekspertów.

Do analizy przyjęto 76 prawidłowo wypełnionych ankiet, w których respondenci oceniali monitory LCD pod względem następujących cech:

- X_1 – przekątna ekranu monitora (przedział liczbowy);
- X_2 – rozmiar plamki (przedział liczbowy);
- X_3 – czas odpowiedzi – tzw. bezwładność (przedział liczbowy);
- X_4 – maksymalna rozdzielczość obrazu (przedział liczbowy);
- X_5 – waga monitora (przedział liczbowy);
- X_6 – kąt obrotu monitora (przedział liczbowy);
- X_7 – cena (przedział liczbowy).

W przypadku ocen ekspertów ze względu na niewielkie zróżnicowanie postrzegania niepodobieństw pomiędzy wybranymi markami monitorów do analizy przyjęto macierz maksymalnych i minimalnych odległości otrzymaną na podstawie ich wiedzy oraz opinii. Na podstawie ankiet obliczono, zgodnie ze wzorami (1) i (2), macierz minimalnych i maksymalnych odległości.

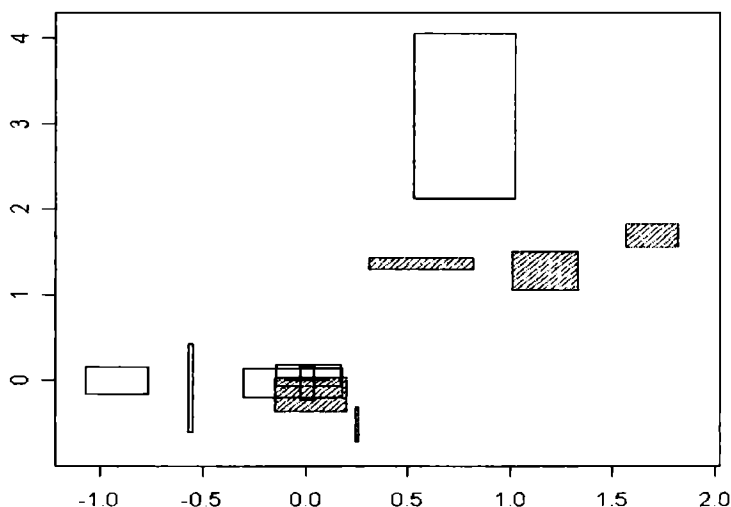
Następnie dla obu macierzy minimalnych i maksymalnych odległości przeprowadzono procedurę skalowania wielowymiarowego metodą *INTERSCAL* (zob. rys. 1) z założeniem przedstawienia jej rezultatów w dwóch wymiarach (R^2). W efekcie otrzymano współrzędne główne, które zawarto w tab. 2.

Tabela 2. Współrzędne główne otrzymane w procedurze *INTERSCAL*

Współrzędne uzyskane na podstawie wyników ankiety												
	1 – Philips		2 – Sony		3 – LG		4 – BenQ		5 – Actina		6 – NEC	
	min.	max	min.	max	min.	max	min.	max	min.	max	min.	max
X	-0,03	0,04	-0,14	0,17	-0,30	0,18	-0,57	-0,55	-1,08	-0,77	0,53	1,02
Y	-0,23	0,17	-0,06	0,18	-0,20	0,14	-0,60	0,43	-0,16	0,15	2,12	4,06
Współrzędne uzyskane na podstawie opinii ekspertów												
	1 – Philips		2 – Sony		3 – LG		4 – BenQ		5 – Actina		6 – NEC	
	min.	max	min.	max	min.	max	min.	max	min.	max	min.	max
X	-0,10	0,11	-0,15	0,20	0,24	0,26	0,31	0,82	1,01	1,33	1,56	1,82
Y	0,002	0,02	-0,36	0,03	-0,71	-0,31	1,30	1,43	1,05	1,50	1,56	1,82

Źródło: obliczenia własne (otrzymane za pomocą pakietu *Euler*).

Na rysunku 2 przedstawiono graficzną reprezentację wzajemnego położenia obiektów uzyskanego na podstawie ankiet (prostokąty białe) oraz na podstawie ocen ekspertów (prostokąty wypełnione).



Rys. 2. Graficzna prezentacja otrzymanych wyników

Źródło: obliczenia własne (pakiet statystyczny *R*).

Na podstawie graficznej reprezentacji wyników można stwierdzić, że dla klientów większość monitorów LCD jest podobna do siebie, z wyjątkiem monitorów marki NEC. Ekspersi podobnie oceniają zaś marki Philips, Sony, LG oraz marki BenQ, Actina, NEC. Różnica w postrzeganiu marek może wynikać z liczby zmien-nych, którymi klienci oceniali monitory.

5. Podsumowanie

Istotną wadą metody *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.

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

Do oceny otrzymanych wyników w przypadku metody *INTERSCAL* autorzy sugerują wykorzystanie miary inercji. Lepszym rozwiązaniem w tym względzie wydaje się wykorzystanie miary *STRESS-Sym* (standaryzowana suma kwadratów reszt dla obiektów symbolicznych). W przypadku współrzędnych uzyskanych na podstawie ocen ekspertów wartość miary *STRESS-Sym* wyniosła 0,02394. Dla współrzędnych uzyskanych na podstawie ankiet miara ta wyniosła 0,03582.

Literatura

- Bock H.-H., Diday E. (red.) (2000), *Analysis of Symbolic Data. Explanatory Methods for Extracting Statistical Information From Complex Data*, 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., Rodríguez O., Diday E. (2005), *SymScal: Symbolic Multidimensional Scaling of Interval Dissimilarities*, Econometric Report EI 2005-15, Erasmus University, Rotterdam.
- Lechevallier Y. (red.) (2000), *Scientific Report for Unsupervised Classification, Validation and Cluster Representation. Analysis System of Symbolic Official Data Technical Report*, raport IST-2000-25161 projektu SODAS.

METHODS OF MULTIDIMENSIONAL SCALING FOR SYMBOLIC OBJECTS

Summary

The aim of this article is to present one of symbolic clustering methods, the *INTERSCAL* method with problems, which will be encountered while applying this method. Article compares also this method with another multidimensional scaling method for symbolic objects (*SYMSCAL*). The article will compare in the empirical part the symbolic multidimensional scaling results based on LCD monitors market.