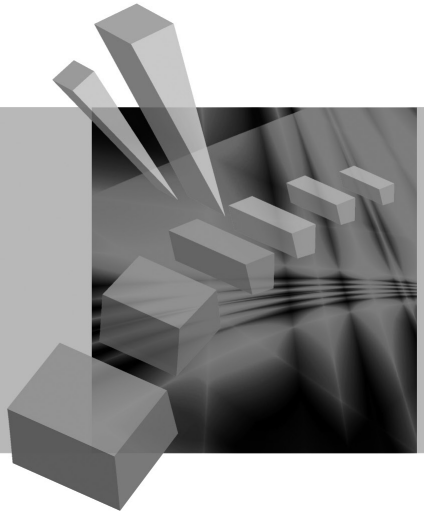
PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu
RESEARCH PAPERS
of Wrocław University of Economics

278

Taksonomia 20

Klasyfikacja i analiza danych – teoria i zastosowania



Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2013

Redaktor Wydawnictwa: Aleksandra Śliwka

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Małgorzata Czupryńska

Projekt okładki: Beata Dębska

Publikacja jest dostępna w Internecie na stronach:
www.ibuk.pl, www.ebscohost.com,
The Central and Eastern European Online Library www.ceeol.com,
a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon
http://kangur.uek.krakow.pl/bazy_ae/bazekon/nowy/index.php

Informacje o naborze artykułów i zasadach recenzowania znajdują się
na stronie internetowej Wydawnictwa
www.wydawnictwo.ue.wroc.pl

Tytuł dofinansowany ze środków Narodowego Banku Polskiego
oraz ze środków Sekcji Klasyfikacji i Analizy danych PTS

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM

Spis treści

Wstęp 9

Józef Pociecha: Wskaźniki finansowe a klasyfikacyjne modele predykcji upadłości firm 15
Eugeniusz Gatnar: Analiza miar adekwatności rezerw walutowych 23
Marek Walesiak: Zagadnienie doboru liczby klas w klasyfikacji spektralnej 33
Joanicjusz Nazarko, Joanna Ejdys, Anna Kononiuk, Anna M. Olszewska: Analiza strukturalna jako metoda klasyfikacji danych w badaniach foresight 44
Andrzej Bąk: Metody porządkowania liniowego w polskiej taksonomii – pakiet `pllord` 54
Aleksandra Łuczak, Feliks Wysocki: Zastosowanie mediany przestrzennej Webera i metody TOPSIS w ujęciu pozycyjnym do konstrukcji syntetycznego miernika poziomu życia 63
Ewa Roszkowska: Zastosowanie rozmytej metody TOPSIS do oceny ofert negocjacyjnych 74
Jacek Batóg: Analiza wrażliwości metody ELECTRE III na obserwacje nietypowe i zmianę wartości progowych 85
Jerzy Korzeniewski: Modyfikacja metody HINoV selekcji zmiennych w analizie skupień 93
Małgorzata Markowska, Danuta Strahl: Wykorzystanie referencyjnego systemu granicznego do klasyfikacji europejskiej przestrzeni regionalnej ze względu na filar inteligentnego rozwoju – kreatywne regiony 101
Elżbieta Sobczak: Inteligentne struktury pracujących a efekty strukturalne zmian zatrudnienia w państwach Unii Europejskiej 111
Elżbieta Gołata, Grażyna Dehnel: Rozbieżności szacunków NSP 2011 i BAEL 120
Iwona Foryś: Wykorzystanie analizy historii zdarzeń do badania powtórnych sprzedaży na lokalnym rynku mieszkaniowym 131
Hanna Dudek, Joanna Landmesser: Wpływ relatywnej deprywacji na subiektywne postrzeganie dochodów 142
Grażyna Łaska: Syntaksonomia numeryczna w klasyfikacji, identyfikacji i analizie przemian zbiorowisk roślinnych 151
Magdalena Osińska, Marcin Fałdziński, Tomasz Zdanowicz: Analiza zależności między procesami fundamentalnymi a rynkiem kapitałowym w Chinach 161

Andrzej Bąk, Tomasz Bartłomowicz: Mikroekonometryczne modele wielomianowe i ich zastosowanie w analizie preferencji z wykorzystaniem programu R 169
Andrzej Dudek, Bartosz Kwaśniewski: Przetwarzanie równoległe algorytmów analizy skupień w technologii CUDA 180
Michał Trzęsiok: Wycena rynkowej wartości nieruchomości z wykorzystaniem wybranych metod wielowymiarowej analizy statystycznej 188
Joanna Trzęsiok: Wybrane symulacyjne techniki porównywania nieparametrycznych metod regresji..... 197
Artur Mikulec: Kryterium Mojeny i Wisharta w analizie skupień – przypadek skupień o różnych macierzach kowariancji 206
Artur Zaborski: Analiza *unfolding* z wykorzystaniem modelu grawitacji 216
Justyna Wilk: Identyfikacja obszarów problemowych i wzrostowych w województwie dolnośląskim w zakresie kapitału ludzkiego 225
Karolina Bartos: Analiza ryzyka odejścia studenta z uczelni po uzyskaniu dyplomu licencjata – zastosowanie sieci MLP 236
Ewa Genge: Segmentacja uczestników Industriady z wykorzystaniem analizy klas ukrytych 246
Izabela Kurzawa: Wielomianowy model logitowy jako narzędzie identyfikacji czynników wpływających na sytuację mieszkaniową polskich gospodarstw domowych 254
Marek Lubicz, Maciej Zięba, Konrad Pawełczyk, Adam Rzechonek, Jerzy Kołodziej: Modele eksploracji danych niezbilansowanych – procedury klasyfikacji dla zadania analizy ryzyka operacyjnego..... 262
Aleksandra Łuczak: Zastosowanie rozmytej hierarchicznej analizy w tworzeniu strategii rozwoju jednostek administracyjnych 271
Marcin Pełka: Rozmyta klasyfikacja spektralna *c*-średnich dla danych symbolicznych interwałowych..... 282
Małgorzata Machowska-Szewczyk: Klasyfikacja obiektów reprezentowanych przez różnego rodzaju cechy symboliczne 290
Ewa Chodakowska: Indeks Malmquista w klasyfikacji podmiotów gospodarczych według zmian ich względnej produktywności działania 300
Beata Bieszk-Stolorz, Iwona Markowicz: Wykorzystanie modeli proporcjonalnego i nieproporcjonalnego hazardu Coxa do badania szansy podjęcia pracy w zależności od rodzaju bezrobocia 311
Marcin Salamaga: Weryfikacja teorii poziomu rozwoju gospodarczego J.H. Dunninga w ujęciu sektorowym w wybranych krajach Unii Europejskiej 321
Justyna Wilk, Michał Bernard Pietrzak, Stanisław Matusik: Sytuacja społeczno-gospodarcza jako determinanta migracji wewnętrznych w Polsce. 330
Hanna Gruchociak: Delimitacja lokalnych rynków pracy w Polsce na podstawie danych z badania przepływów ludności związanych z zatrudnieniem 343

Radosław Pietrzyk: Efektywność inwestycji polskich funduszy inwestycyjnych z tytułu doboru papierów wartościowych i umiejętności wykorzystania trendów rynkowych 351
Sabina Denkowska: Procedury testowań wielokrotnych 362

Summaries

Józef Pociecha: Financial ratios and classification models of bankruptcy prediction 22
Eugeniusz Gatnar: Analysis of FX reserve adequacy measures 32
Marek Walesiak: Automatic determination of the number of clusters using spectral clustering 43
Joanicjusz Nazarko, Joanna Ejdys, Anna Kononiuk, Anna M. Olszewska: Structural analysis as a method of data classification in foresight research 53
Andrzej Bąk: Linear ordering methods in Polish taxonomy – `pllord` package 62
Aleksandra Łuczak, Feliks Wysocki: The application of spatial median of Weber and the method TOPSIS in positional formulation for the construction of synthetic measure of standard of living 73
Ewa Roszkowska: Application of the fuzzy TOPSIS method to the estimation of negotiation offers..... 84
Jacek Batóg: Sensitivity analysis of ELECTRE III method for outliers and change of thresholds 92
Jerzy Korzeniewski: Modification of the HINoV method of selecting variables in cluster analysis 100
Małgorzata Markowska, Danuta Strahl: Implementation of reference limit system for the European regional space classification regarding smart growth pillar – creative regions 110
Elżbieta Sobczak: Smart workforce structures versus structural effects of employment changes in the European Union countries 119
Elżbieta Gołata, Grażyna Dehnel: Divergence in National Census 2011 and LFS estimates..... 130
Iwona Foryś: Event history analysis in the resale study on the local housing market 141
Hanna Dudek, Joanna Landmesser: Impact of the relative deprivation on subjective income satisfaction 150
Grażyna Łaska: Numerical syntaxonomy in classification, identification and analysis of changes of secondary communities 160
Magdalena Osińska, Marcin Fałdziński, Tomasz Zdanowicz: Analysis of relations between fundamental processes and capital market in China..... 166
Andrzej Bąk, Tomasz Bartłomowicz: Microeconomic polynomial models and their application in the analysis of preferences using R program..... 179

Andrzej Dudek, Bartosz Kwaśniewski: Parallel processing of clustering algorithms in CUDA technology 187
Michał Trzęsiok: Real estate market value estimation based on multivariate statistical analysis 196
Joanna Trzęsiok: On some simulative procedures for comparing nonparametric methods of regression 205
Artur Mikulec: Mojena and Wishart criterion in cluster analysis – the case of clusters with different covariance matrices 215
Artur Zaborski: Unfolding analysis by using gravity model 224
Justyna Wilk: Determination of problem and growth areas in Dolnośląskie Voivodship as regards human capital 235
Karolina Bartos: Risk analysis of bachelor students' university abandonment – the use of MLP networks 245
Ewa Genge: Clustering of industrial holiday participants with the use of latent class analysis 253
Izabela Kurzawa: Multinomial logit model as a tool to identify the factors affecting the housing situation of Polish households 261
Marek Lubicz, Maciej Zięba, Konrad Pawełczyk, Adam Rzechonek, Jerzy Kołodziej: Modelling class imbalance problems: comparing classification approaches for surgical risk analysis 270
Aleksandra Łuczak: The application of fuzzy hierarchical analysis to the evaluation of validity of strategic factors in administrative districts 281
Marcin Pełka: A spectral fuzzy c-means clustering algorithm for interval-valued symbolic data 289
Małgorzata Machowska-Szewczyk: Clustering algorithms for mixed-feature symbolic objects 299
Ewa Chodakowska: Malmquist index in enterprises classification on the basis of relative productivity changes 310
Beata Bieszk-Stolorz, Iwona Markowicz: Using proportional and non proportional Cox hazard models to research the chances for taking up a job according to the type of unemployment 320
Marcin Salamaga: Verification J.H. Dunning's theory of economic development by economic sectors in some EU countries 329
Justyna Wilk, Michał Bernard Pietrzak, Stanisław Matusik: Socio-economic situation as a determinant of internal migration in Poland 342
Hanna Gruchociak: Delimitation of local labor markets in Poland on the basis of the employment-related population flows research 350
Radosław Pietrzyk: Selectivity and timing in Polish mutual funds performance measurement 361
Sabina Denkowska: Multiple testing procedures 369

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 278 • 2013

Taksonomia 20. Klasyfikacja i analiza danych – teoria i zastosowania ISSN 1899-3192

Marcin Pełka
Uniwersytet Ekonomiczny we Wrocławiu

ROZMYTA KLASYFIKACJA SPEKTRALNA *C*-ŚREDNICH DLA DANYCH SYMBOLICZNYCH INTERWAŁOWYCH

Streszczenie: Celem artykułu jest zaproponowanie nowej metody klasyfikacji rozmytej na potrzeby analizowania danych symbolicznych interwałowych. W artykule przedstawiono podstawowe pojęcia z zakresu analizy danych symbolicznych, klasyfikacji spektralnej oraz rozmytej klasyfikacji *c*-średnich. W części empirycznej przedstawiono wyniki badań symulacyjnych dla sztucznych zbiorów danych wygenerowanych w programie R.

Słowa kluczowe: klasyfikacja spektralna, rozmyta klasyfikacja *c*-średnich, dane symboliczne interwałowe.

1. Wstęp

Metodę rozmytej klasyfikacji *c*-średnich dla danych w rozumieniu klasycznym zaproponował Dunn [1973], następnie jej modyfikację wprowadził Bezdek [1981]. W pracach El-Sonbaty'ego i Ismaila [1998], Yanga i in. [2004] przedstawiono rozmyte metody klasyfikacji danych symbolicznych różnych typów. W pracach de Carvalho [2007] oraz de Carvalho i Tenório [2010] zaproponowano kolejne adaptacje i modyfikacje różnych metod klasyfikacji rozmytej na potrzeby analizy danych symbolicznych interwałowych.

Klasyfikacja spektralna, którą zaproponowali w swej pracy Ng, Jordan i Weiss, jest tak naprawdę nie tyle nową metodą klasyfikacji, ile nowym podejściem do przygotowywania danych na potrzeby klasyfikacji, która wykorzystuje ideę dekompozycji spektralnej.

Celem artykułu jest zaprezentowanie nowej metody klasyfikacji rozmytej dla danych symbolicznych interwałowych – rozmytej klasyfikacji spektralnej *c*-średnich, test to propozycja stanowiąca autorskie połączenie dwóch istniejących rozwiązań w zakresie klasyfikacji danych – tj. dekompozycji spektralnej i rozmytej klasyfikacji *c*-średnich.

W części empirycznej przedstawiono wyniki symulacji z wykorzystaniem sztucznych zbiorów danych symbolicznych interwałowych wygenerowanych z wy-

korzystaniem pakietów `clusterSim` oraz `clusterGeneration` programu R oraz rzeczywistych zbiorów danych.

2. Dane symboliczne

Obiekty symboliczne mogą być opisywane przez następujące rodzaje zmiennych symbolicznych [Bock, Diday (red.) 2000, s. 2-3]:

a) ilorazowe,

b) przedziałowe,

c) porządkowe,

d) nominalne,

e) interwałowe, których realizacją są przedziały liczbowe rozłączne lub nierozłączne,

f) wielowariantowe, gdzie realizacją zmiennej jest więcej niż jeden wariant (liczba lub kategoria),

g) wielowariantowe z wagami, gdzie realizacją zmiennej oprócz wielu wariantów są dodatkowo wagi (lub prawdopodobieństwa) dla każdego z wariantów zmiennej dla danego obiektu.

Niezależnie od typu zmiennej w analizie danych symbolicznych możemy mieć do czynienia ze zmiennymi strukturalnymi [Bock, Diday (red.) 2000, s. 2-3; 33-37]. Do tego typu zmiennych zalicza się **zmienne hierarchiczne** – w których *a priori* ustalone są reguły decydujące o tym, czy dana zmienna opisuje dany obiekt czy nie; **zmienne taksonomiczne** – w których ustalone są *a priori* realizacje danej zmiennej; **zmienne logiczne** – tj. takie, dla których ustalono *a priori* reguły logiczne lub funkcyjne, które decydują o wartościach zmiennej.

W analizie danych symbolicznych wyróżnia się dwa typy obiektów symbolicznych:

- **obiekty symboliczne pierwszego rzędu** – obiekty rozumiane w sensie „klasycznym" (obiekty elementarne), np. konsument, przedsiębiorstwo, produkt, pacjent czy gospodarstwo domowe,
- **obiekty symboliczne drugiego rzędu** – obiekty utworzone w wyniku agregacji zbioru obiektów symbolicznych pierwszego rzędu, np. grupa konsumentów preferująca określony produkt, region geograficzny (jako wynik agregacji podregionów).

3. Rozmyta klasyfikacja spektralna *c*-średnich

W literaturze przedmiotu zaproponowano trzy rozmyte metody klasyfikacji, które mają zastosowanie wyłącznie dla danych symbolicznych interwałowych – są to rozmyta klasyfikacja *c*-średnich, rozmyta adaptacyjna klasyfikacja *c*-średnich de Carvalho [2007] (por. [Pełka 2010]) oraz rozmyta klasyfikacja *k*-średnich de Carvalho i Tenório [2010].

Rozmyta klasyfikacja *c*-średnich oraz rozmyta adaptacyjna klasyfikacja *c*-średnich dla danych symbolicznych interwałowych bazują w swej konstrukcji na adaptacji odległości euklidesowej (zob. [de Carvalho 2007, s. 425-426; Pełka 2010]). Metody te są modyfikacją klasycznej rozmytej klasyfikacji *c*-średnich na potrzeby danych interwałowych.

Rozmyta klasyfikacja *k*-średnich dla danych symbolicznych bazuje w swej konstrukcji na adaptacji odległości Mahalanobisa (por. [de Carvalho, Tenório 2010, s. 2980]). Na potrzeby analizy danych symbolicznych interwałowych w artykule de Carvalho i Tenório [2010] zaproponowano różne warianty obliczania macierzy kowariancji **M**. Podstawowe podobieństwa i różnice między tymi metodami zaprezentowano w tab. 1.

Tabela 1. Podobieństwa i różnice w metodach klasyfikacji rozmytej dla danych symbolicznych interwałowych

Kryterium porównania	Rozmyta klasyfikacja *c*-średnich	Rozmyta adaptacyjna klasyfikacja *c*-średnich	Rozmyta klasyfikacja *k*-średnich
Funkcja--kryterium	Metoda minimalizuje funkcję-kryterium, w której wykorzystywany jest stopień przynależności obiektu do klasy (μ_{ik})		
Zmienne symboliczne	Wyłącznie zmienne symboliczne interwałowe		
Miara odległości	Funkcja-kryterium wykorzystuje adaptację odległości euklidesowej		Funkcja-kryterium wykorzystuje adaptację odległości Mahalanobisa
Liczba klas	Liczba klas jest parametrem, który ustala badacz		
Wybór liczby klas	Można zastosować różnorodne miary jakości klasyfikacji bazujące na przynależności obiektu do klasy, a także skorygowany indeks Randa dla klasyfikacji rozmytych		

Źródło: opracowanie własne na podstawie prac [de Carvalho 2007; de Carvalho, Tenório 2010; Pełka 2010].

Proponowana w niniejszym opracowaniu rozmyta klasyfikacja spektralna *c*-średnich dla danych symbolicznych interwałowych składa się z dwóch zasadniczych elementów:

1. Klasyfikacji spektralnej, która tak naprawdę jest nie tyle nową metodą klasyfikacji, ile nowym podejściem do przygotowania danych na potrzeby klasyfikacji (por. [Ng i in. 2001; Walesiak, Dudek 2009]). W wyniku zastosowania tego podejścia otrzymuje się nową macierz danych (macierz **Y**), która jest podstawą do zastosowania wybranej metody klasyfikacji.

2. Rozmytej klasyfikacji c-średnich, w której macierzą danych jest macierz **Y** otrzymana dzięki zastosowaniu klasyfikacji spektralnej.

Klasyfikacja spektralna dla danych symbolicznych interwałowych składa się z następujących kroków (zob. [Walesiak, Dudek 2009, s. 12-14]):

1. Konstrukcja tablicy danych symbolicznych $\mathbf{V} = \left[v_{ij} \right]$ o wymiarach $n \times m$ ($i = 1, \ldots, n$ – numer obiektu, $j = 1, \ldots, m$ – numer zmiennej).

2. Zastosowanie estymatora jądrowego do obliczenia macierzy podobieństw $\mathbf{A} = \left[A_{ik} \right]$ (*affinity matrix*) między obiektami. Najczęściej do wyznaczenia macierzy $\mathbf{A}$ wykorzystywany jest estymator gaussowski (zob. [Karatzoglou 2006, s. 26]):

$$A_{ik} = \exp\left(-\sigma \cdot d_{ik}\right), \quad i,k = 1, \ldots, n, \tag{1}$$

gdzie: d_{ik} – odległość między *i*-tym i *k*-tym obiektem symbolicznym,
σ – parametr skali (szerokość pasma – *kernel width*).

3. Obliczenie diagonalnej macierzy **D**, na głównej przekątnej tej macierzy znajdują się sumy każdego wiersza z macierzy **A**, a poza nią są zera.

4. Konstrukcja znormalizowanej macierzy Laplace'a:

$$\mathbf{L} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}. \tag{2}$$

Własności tej macierzy zaprezentowano m.in. w pracy [von Luxburg 2006].

5. Obliczenie wartości własnych i odpowiadających im wektorów własnych (o długości równej jeden) dla macierzy **L**. Uporządkowanie wektorów własnych według malejących wartości własnych. Pierwsze *u*, gdzie *u* – liczba klas, wektorów własnych tworzy macierz $\mathbf{E} = \left[e_{ij} \right]$ o wymiarach $n \times u$.

6. Przeprowadzenie normalizacji macierzy **E** zgodnie ze wzorem:

$$y_{ij} = e_{ij} \Big/ \sqrt{\sum_{j=1}^{u} e_{ij}^{\,2}}, \tag{3}$$

gdzie: $i = 1, \ldots, n$ – numer obiektu, $j = 1, \ldots, u$ – numer zmiennej, *u* – liczba klas.

Dzięki tej normalizacji długość każdego wektora wierszowego macierzy $\mathbf{Y} = \left[y_{ij} \right]$ jest równa jeden.

7. Macierz **Y** stanowi punkt wyjścia zastosowania klasycznych metod analizy skupień (proponuje się tutaj wykorzystanie rozmytej klasyfikacji *c*-średnich).

Zasadnicze znaczenie dla klasyfikacji spektralnej ma parametr σ. Zagadnienie wyboru odpowiedniej wartości parametru zawarto w pracy Walesiaka i Dudka [2009] oraz Karatzoglou [2006]. Drugim ważnym zagadnieniem w przypadku danych symbolicznych jest wybór odpowiedniej miary odległości (zob. wzór (1)). Miary odległości dla danych symbolicznych omówione są m.in. w pracach Gatnara i Walesiaka [2011], Bocka i Didaya [2000].

Rozmyta klasyfikacja *c*-średnich jest metodą iteracyjno-optymalizacyjną, której idea jest bardzo mocno zbliżona do klasycznej metody *k*-średnich. Głównym celem tej metody jest znalezienie takich środków ciężkości klas, które zminimalizują funkcję-kryterium w postaci:

$$J_m = \sum_{i=1}^{c}\sum_{j=1}^{n} \mu_{ij}^m d_{ij}^2, \tag{4}$$

gdzie: μ_{ij} – stopień przynależności j-tego obiektu do i-tej klasy rozmytej,
d_{ij} – odległość euklidesowa między środkiem ciężkości i-tej klasy rozmytej a j-tym obiektem,
m – parametr rozmycia, przy czym $m > 1$.

Algorytm rozmytej klasyfikacji c-średnich składa się z następujących kroków:

1. Ustalenie początkowych przynależności obiektów do poszczególnych klas rozmytych – otrzymujemy macierz $\mathbf{U} = [\mu_{ik}]$. Określenie maksymalnej liczby iteracji T oraz kryterium stopu ε (np. $\varepsilon = 10^{-6}$), $t = 1$.

2. Ustalenie środków ciężkości klas zgodnie ze wzorem:

$$c_i = \frac{\sum_{k=1}^{n} \mu_{ik}^m y_k}{\sum_{k=1}^{n} \mu_{ik}^m}, \tag{5}$$

gdzie: oznaczenia jak we wzorach (4) oraz (3).

3. Obliczenie nowej macierzy $\mathbf{U}_N$ zgodnie ze wzorem:

$$\mu_{ij} = \frac{1}{\sum_{l=1}^{c}\left(\frac{d_{ij}}{d_{lj}}\right)^{\frac{2}{m-1}}}, \tag{6}$$

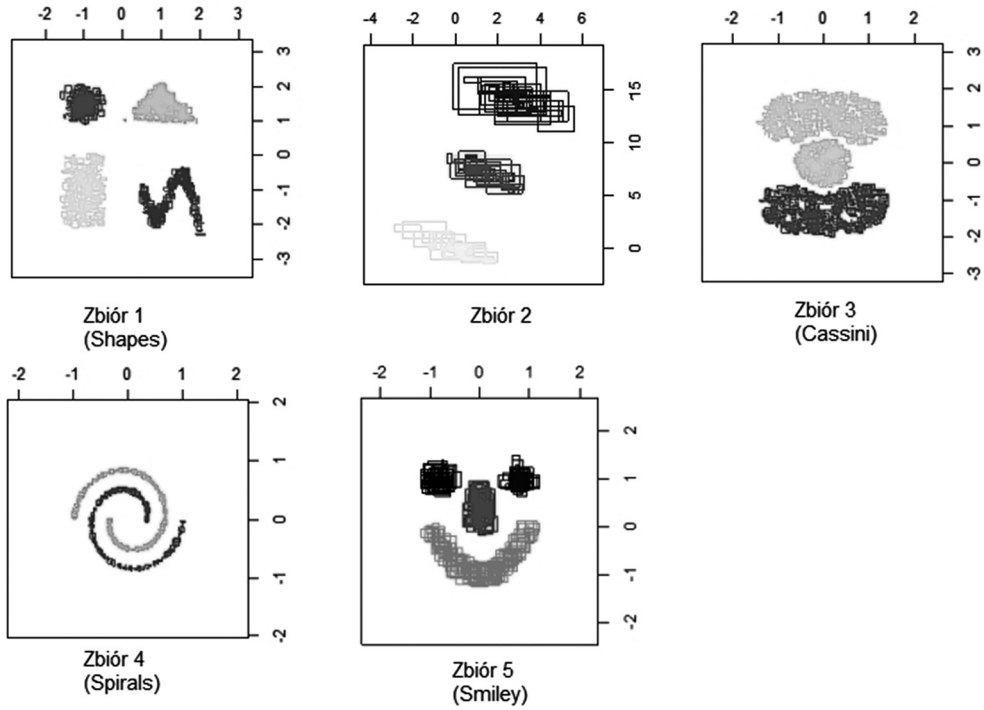
gdzie: d_{ij} – odległość między l-tym środkiem ciężkości klas a j-tym obiektem, pozostałe oznaczenia jak we wzorze (4).

4. Jeżeli $\|\mathbf{U}_N - \mathbf{U}\| > \varepsilon$, gdzie $\|\mathbf{U}_N - \mathbf{U}\|$ to odległość euklidesowa, wówczas $\mathbf{U} = \mathbf{U}_N$ i należy wrócić do kroku 2 algorytmu, zwiększając liczbę iteracji o jeden $(t = t + 1)$. Całość postępowania kończy się, gdy zostanie osiągnięta założona liczba iteracji T lub gdy $\|\mathbf{U}_N - \mathbf{U}\| < \varepsilon$.

4. Badania symulacyjne

Na potrzeby badań symulacyjnych przygotowano w programie R pięć zbiorów danych o znanej strukturze klas. Zbiory danych wygenerowano z wykorzystaniem pakietu `clusterSim` (funkcja `cluster.Gen`) oraz pakietu `mlbench` (funkcje `mlbench.shapes`, `mlbench.cassini`, `mlbench.spirals` oraz `mlbench.smiley`). W celu otrzymania danych symbolicznych interwałowych

z wykorzystaniem pakietu `mlbench` otrzymane dane traktowane są jako środki zmiennej symbolicznej interwałowej. Rozstęp zmiennych jest dobierany w taki sposób, aby zachować oryginalny kształt danych. Najczęściej jest on dobierany losowo z przedziału $[0;1]$. Wygenerowane zbiory danych zaprezentowano na rys. 1.



Rys. 1. Zbiory danych symulacyjnych

Źródło: opracowanie własne z wykorzystaniem programu R.

Dla każdego zbioru danych wykonano 20 symulacji i obliczono średnią wartość skorygowanego indeksu Randa dla klasyfikacji rozmytych (M_R) oraz wartość odchylenia standardowego dla tego indeksu (S_R). Indeks ten zaprezentowano w pracy [Hüllermier i Rifqi 2009].

W badaniach symulacyjnych zastosowano cztery różne warianty miar odległości na potrzeby klasyfikacji spektralnej – nieznormalizowaną odległość Ichino-Yaguchiego (U_2), odległość Hausdorffa (H), odległość de Carvalho bazującą na potencjale opisowym obiektu symbolicznego (SO_3) oraz odległość de Carvalho bazującą na mierze Ichino-Yaguchiego (SO_2) (zob. [Bock, Diday 2000, s. 139-185]). Wyniki symulacji zawarto w tab. 2.

Tabela 2. Wyniki badań symulacyjnych

Miara odległości	Zbiór 1 (Shapes)	Zbiór 2	Zbiór 3 (Cassini)	Zbiór 4 (Spirals)	Zbiór 5 (Smiley)
Ichino--Yaguchiego (U_2)	$M_R = 1$ $S_R = 7{,}04E\text{-}11$	$M_R = 1$ $S_R = 3{,}63E\text{-}07$	$M_R = 1$ $S_R = 7{,}41E\text{-}09$	$M_R = 0{,}99999$ $S_R = 1{,}90E\text{-}07$	$M_R = 1$ $S_R = 1{,}23E\text{-}09$
Hausdorffa (H)	$M_R = 0{,}99999$ $S_R = 8{,}50E\text{-}08$	$M_R = 1$ $S_R = 2{,}84E\text{-}08$	$M_R = 1$ $S_R = 3{,}79E\text{-}08$	$M_R = 0{,}99999$ $S_R = 5{,}85E\text{-}07$	$M_R = 0{,}999999$ $S_R = 9{,}68E\text{-}08$
de Carvalho (SO_2)	$M_R = 1$ $S_R = 1{,}09E\text{-}08$	$M_R = 0{,}99999$ $S_R = 2{,}60E\text{-}07$	$M_R = 1$ $S_R = 1{,}18E\text{-}08$	$M_R = 0{,}99999$ $S_R = 3{,}61E\text{-}07$	$M_R = 0{,}999996$ $S_R = 3{,}29E\text{-}06$
de Carvalho (SO_3)	$M_R = 1$ $S_R = 9{,}27E\text{-}10$	$M_R = 1$ $S_R = 4{,}09E\text{-}10$	$M_R = 0{,}99999$ $S_R = 6{,}95E\text{-}08$	$M_R = 0{,}99999$ $S_R = 5{,}99E\text{-}07$	$M_R = 1$ $S_R = 4{,}97E\text{-}08$

Źródło: obliczenia własne w programie R.

5. Podsumowanie

Dane symboliczne interwałowe mają tendencję do tworzenia klas nierozłącznych (rozmytych) o różnorodnych kształtach. Zaproponowana w artykule rozmyta klasyfikacja spektralna *c*-średnich pozwala analizować dane tego typu. Dodatkowo zaproponowana metoda może znaleźć zastosowanie dla danych symbolicznych dowolnego typu. Wówczas należy zastosować jedynie odpowiednią miarę odległości dla tych danych.

Przeprowadzone zostały badania symulacyjne z wykorzystaniem czterech wybranych miar odległości dla danych symbolicznych. Zbliżone wyniki otrzymano przy zastosowaniu nieznormalizowanej odległości Ichino-Yaguchiego, miary de Carvalho bazującej na potencjale opisowym obiektu symbolicznego oraz miary de Carvahlo bazującej na mierze Ichino-Yaguchiego.

W badaniach symulacyjnych nie sprawdzano, czy rozmyta klasyfikacja spektralna *c*-średnich pozwala na odkrycie właściwej struktury klas. Niemniej jednak dotychczasowe rezultaty sugerują, że w przypadku, gdy liczba klas zadana przez badacza jest większa od rzeczywistej liczby klas, wówczas przynależność obiektu do tych „dodatkowych" klas bardzo szybko się zbliża się do zera.

Dla wszystkich analizowanych w badaniu symulacyjnym zbiorów danych otrzymano mocno stabilne rezultaty.

Celem dalszych badań będzie porównanie rozmytej klasyfikacji spektralnej *c*-średnich z innymi metodami klasyfikacji rozmytej dla danych symbolicznych interwałowych z zastosowaniem różnorodnych zbiorów danych.

Literatura

Bezdek J.C. (1981), *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York.

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

De Carvalho F.A.T. (2007), *Fuzzy c-means clustering methods for symbolic interval data*, "Pattern Recognition Letters" 28(4), s. 423-437.

De Carvalho F.A.T., Tenório C.P. (2010), *Fuzzy k-means clustering algorithms for interval-valued data based on adaptive quadric distances*, "Fuzzy Sets and Systems", 161 (23), s. 2978-2999.

El-Sonbaty Y., Ismail M.A. (1998), *Fuzzy clustering for symbolic data*, "IEEE Transactions on Fuzzy Systems", vol. 6, issue 2, s. 195-204.

Gatnar E., Walesiak M. (red.) (2011), *Analiza danych jakościowych i symbolicznych z wykorzystaniem programu R*, C.H. Beck, Warszawa.

Hüllermeir E., Rifqi M. (2009), *A fuzzy variant of the Rand Index for comparing clustering structures*, Proceedings of the IFSA/EUSFLAT Conference 2009, s. 1294-1298.

Karatzoglou A. (2006), Kernel Methods. Software, Algorithms and Applications, rozprawa doktorska, Uniwersytet Techniczny we Wiedniu.

Ng A., Jordan M., Weiss Y. (2001), *On Spectral Clustering: Analysis and an Algorithm*, [w:] T. Dietrich, S. Becker, Z. Ghahramani (eds.), *Advances in Neural Information Processing Systems*, MIT Press, s. 849-856.

Pełka M. (2010), *Rozmyta klasyfikacja k-średnich dla danych symbolicznych interwałowych*, PN UE we Wrocławiu nr 107, s. 190-196.

Walesiak M., Dudek A. (2009), *Odległość GDM dla danych porządkowych a klasyfikacja spektralna*, PN UE we Wrocławiu nr 84, s. 9-19.

von Luxburg U. (2006), *A Tutorial on Spectral Clustering*, Max Planck Institute for Biological Cybernetics, Technical Report TR-149.

Yang M.-S., Hwang P.-Y., Chen D.-H. (2004), *Fuzzy clustering algorithms for mixed feature types*, "Fuzzy Sets and Systems" 141, s. 301-317.

A SPECTRAL FUZZY *C*-MEANS CLUSTERING ALGORITHM FOR INTERVAL-VALUED SYMBOLIC DATA

Summary: The main aim of the paper is to present a proposal of new fuzzy clustering method for symbolic interval-valued data. The paper presents basic terms of symbolic data, spectral clustering and fuzzy *c*-means clustering. In the empirical part results of simulation study with application of artificial data sets obtained from R software are presented.

Keywords: spectral clustering, fuzzy c-means clustering algorithm, symbolic interval-valued data.