PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu

RESEARCH PAPERS
of Wrocław University of Economics

Nr 327

Taksonomia 22
Klasyfikacja i analiza danych – teoria i zastosowania

Redaktorzy naukowi
Krzysztof Jajuga, Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2014

Redaktor Wydawnictwa: Barbara Majewska
Redaktor techniczny: Barbara Łopusiewicz
Korektor: Barbara Cibis
Łamanie: Beata Mazur
Projekt okładki: Beata Dębska

Publikacja jest dostępna w Internecie na stronach:
www.ibuk.pl, www.ebscohost.com,
w Dolnośląskiej Bibliotece Cyfrowej www.dbc.wroc.pl,
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

Eugeniusz Gatnar, Balance of payments statistics and external competitiveness of Poland 15

Andrzej Sokołowski, Magdalena Czaja, Efektywność metody *k*-średnich w zależności od separowalności grup 23

Barbara Pawełek, Józef Pociecha, Adam Sagan, Wielosektorowa analiza ukrytych przejść w modelowaniu zagrożenia upadłością polskich przedsiębiorstw 30

Elżbieta Gołata, Zróżnicowanie procesu starzenia i struktur demograficznych w Poznaniu i aglomeracji poznańskiej na tle wybranych dużych miast Polski w latach 2002-2011 39

Aleksandra Łuczak, Feliks Wysocki, Ustalanie systemu wag dla cech w zagadnieniach porządkowania liniowego obiektów 49

Marek Walesiak, Wzmacnianie skali pomiaru dla danych porządkowych w statystycznej analizie wielowymiarowej 60

Paweł Lula, Identyfikacja słów i fraz kluczowych w tekstach polskojęzycznych za pomocą algorytmu *RAKE* 69

Mariusz Kubus, Propozycja modyfikacji metody złagodzonego LASSO 77

Andrzej Bąk, Tomasz Bartłomowicz, Wielomianowe modele logitowe wyborów dyskretnych i ich implementacja w pakiecie `DiscreteChoice` programu R 85

Justyna Brzezińska, Wykorzystanie modeli logarytmiczno-liniowych do analizy bezrobocia w Polsce w latach 2004-2012 95

Andrzej Bąk, Marcin Pełka, Aneta Rybicka, Zastosowanie pakietu `dcMNM` programu R w badaniach preferencji konsumentów wódki 104

Barbara Batóg, Jacek Batóg, Analiza stabilności klasyfikacji polskich województw według sektorowej wydajności pracy w latach 2002-2010 113

Małgorzata Markowska, Danuta Strahl, Klasyfikacja europejskiej przestrzeni regionalnej ze względu na filary inteligentnego rozwoju z wykorzystaniem referencyjnego systemu granicznego 121

Kamila Migdał-Najman, Krzysztof Najman, Formalna ocena jakości odwzorowania struktury grupowej na mapie Kohonena 131

Kamila Migdał-Najman, Krzysztof Najman, Graficzna ocena jakości odwzorowania struktury grupowej na mapie Kohonena 139

Beata Basiura, Anna Czapkiewicz, Badanie jakości klasyfikacji szeregów czasowych 148

Michał Trzęsiok, Wybrane metody identyfikacji obserwacji oddalonych 157

Grażyna Dehnel, Tomasz Klimanek, Taksonomiczne aspekty estymacji pośredniej uwzględniającej autokorelację przestrzenną w statystyce gospodarczej ... 167
Michał Bernard Pietrzak, Justyna Wilk, Odległość ekonomiczna w modelowaniu zjawisk przestrzennych z wykorzystaniem modelu grawitacji ... 177
Maciej Beręsewicz, Próba zastosowania różnych miar odległości w uogólnionym estymatorze Petersena ... 186
Marcin Szymkowiak, Tomasz Józefowski, Konstrukcja i praktyczne wykorzystanie estymatorów typu SPREE na przykładzie dwuwymiarowych tabel kontyngencji ... 195
Marcin Pełka, Klasyfikacja pojęciowa danych symbolicznych w podejściu wielomodelowym ... 202
Małgorzata Machowska-Szewczyk, Ocena klas w rozmytej klasyfikacji obiektów symbolicznych ... 210
Justyna Wilk, Problem wyboru liczby klas w taksonomicznej analizie danych symbolicznych ... 220
Andrzej Dudek, Metody analizy skupień w klasyfikacji markerów map Google ... 229
Ewa Roszkowska, Ocena ofert negocjacyjnych w słabo ustrukturyzowanych problemach negocjacyjnych z wykorzystaniem rozmytej procedury SAW 237
Marcin Szymkowiak, Marek Witkowski, Zastosowanie analizy korespondencji do badania kondycji finansowej banków spółdzielczych ... 248
Bartłomiej Jefmański, Budowa rozmytych indeksów satysfakcji klientów z zastosowaniem programu R ... 257
Karolina Bartos, Odkrywanie wzorców zachowań konsumentów za pomocą analizy koszykowej danych transakcyjnych ... 266
Joanna Trzęsiok, Taksonomiczna analiza krajów pod względem dzietności kobiet oraz innych czynników demograficznych ... 275
Beata Bal-Domańska, Próba identyfikacji większych skupisk regionalnych oraz ich konwergencja ... 285
Beata Bieszk-Stolorz, Iwona Markowicz, Wpływ zasiłku na proces poszukiwania pracy ... 294
Marta Dziechciarz-Duda, Klaudia Przybysz, Wykształcenie a potrzeby rynku pracy. Klasyfikacja absolwentów wyższych uczelni ... 303
Tomasz Klimanek, Problem pomiaru procesu dezagraryzacji wsi polskiej w świetle wielowymiarowych metod statystycznych ... 313
Małgorzata Sej-Kolasa, Mirosława Sztemberg-Lewandowska, Wybrane metody analizy danych wzdłużnych ... 321
Artur Zaborski, Zastosowanie miar odległości dla danych porządkowych do agregacji preferencji indywidualnych ... 330
Mariola Chrzanowska, Nina Drejerska, Iwona Pomianek, Zastosowanie analizy korespondencji do badania sytuacji mieszkańców strefy podmiejskiej Warszawy na rynku pracy ... 338

Katarzyna Wawrzyniak, Klasyfikacja województw według stopnia realizacji priorytetów Strategii Rozwoju Kraju 2007-2015 z wykorzystaniem wartości centrum wierszowego 346

Summaries

Eugeniusz Gatnar, Statystyka bilansu płatniczego a konkurencyjność gospodarki Polski 22
Andrzej Sokołowski, Magdalena Czaja, Cluster separability and the effectiveness of *k*-means method 29
Barbara Pawełek, Józef Pociecha, Adam Sagan, Multisectoral analysis of latent transitions in bankruptcy prediction models..... 38
Elżbieta Gołata, Differences in the process of aging and demographic structures in Poznań and the agglomeration compared to selected Polish cities in the years 2002-2011 48
Aleksandra Łuczak, Feliks Wysocki, Determination of weights for features in problems of linear ordering of objects 59
Marek Walesiak, Reinforcing measurement scale for ordinal data in multivariate statistical analysis 68
Paweł Lula, Automatic identification of keywords and keyphrases in documents written in Polish..... 76
Mariusz Kubus, The proposition of modification of the relaxed LASSO method..... 84
Andrzej Bąk, Tomasz Bartłomowicz, Microeconomic multinomial logit models and their implementation in the `DiscreteChoice` R package. 94
Justyna Brzezińska, The analysis of unemployment data in Poland in 2004--2012 with application of log-linear models 103
Andrzej Bąk, Marcin Pełka, Aneta Rybicka, Application of the MMLM package of R software for vodka consumers preference analysis..... 112
Barbara Batóg, Jacek Batóg, Analysis of the stability of classification of Polish voivodeships in 2002-2010 according to the sectoral labour productivity 120
Małgorzata Markowska, Danuta Strahl, Classification of the European regional space in terms of smart growth pillars using the reference limit system..... 130
Kamila Migdał Najman, Krzysztof Najman, Formal quality assessment of group structure mapping on the Kohonen's map 138
Kamila Migdał Najman, Krzysztof Najman, Graphical quality assessment of group structure mapping on the Kohonen's map 147
Beata Basiura, Anna Czapkiewicz, Validation of time series clustering 156
Michał Trzęsiok, Selected methods for outlier detection..... 166

Grażyna Dehnel, Tomasz Klimanek, Taxonomic aspects of indirect estimation accounting for spatial correlation in enterprise statistics 176
Michał Bernard Pietrzak, Justyna Wilk, Economic distance in modeling spatial phenomena with the application of gravity model..... 185
Maciej Beręsewicz, An attempt to use different distance measures in the Generalized Petersen estimator 194
Marcin Szymkowiak, Tomasz Józefowski, Construction and practical using of SPREE estimators for two-dimensional contingency tables..... 201
Marcin Pełka, The ensemble conceptual clustering for symbolic data..... 209
Małgorzata Machowska-Szewczyk, Evaluation of clusters obtained by fuzzy classification methods for symbolic objects..... 219
Justyna Wilk, Problem of determining the number of clusters in taxonomic analysis of symbolic data 228
Andrzej Dudek, Clustering techniques for Google maps markers..... 236
Ewa Roszkowska, The evaluation of negotiation offers in ill structure negotiation problems with the application of fuzzy SAW procedure 247
Marcin Szymkowiak, Marek Witkowski, The use of correspondence analysis in analysing the financial situation of cooperative banks..... 256
Bartłomiej Jefmański, The construction of fuzzy customer satisfaction indexes using R program..... 265
Karolina Bartos, Discovering patterns of consumer behaviour by market basket analysis of the transactional data..... 274
Joanna Trzęsiok, Cluster analysis of countries with respect to fertility rate and other demographic factors 284
Beata Bal-Domańska, An attempt to identify major regional clusters and their convergence 293
Beata Bieszk-Stolorz, Iwona Markowicz, The influence of benefit on the job finding process 302
Marta Dziechciarz-Duda, Klaudia Przybysz, Education and labor market needs. Classification of university graduates 312
Tomasz Klimanek, The problem of measuring deagrarianisation process in rural areas in Poland using multivariate statistical methods..... 320
Małgorzata Sej-Kolasa, Mirosława Sztemberg-Lewandowska, Selected methods for an analysis of longitudinal data..... 329
Artur Zaborski, The application of distance measures for ordinal data for aggregation individual preferences 337
Mariola Chrzanowska, Nina Drejerska, Iwona Pomianek, Application of correspondence analysis to examine the situation of the inhabitants of Warsaw suburban area in the labour market 345
Katarzyna Wawrzyniak, Classification of voivodeships according to the level of the realization of priorities of *the National Development Strategy 2007-2015* with using the values of centroid of the rows 355

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 327 • 2014
Taksonomia 22 ISSN 1899-3192
Klasyfikacja i analiza danych – teoria i zastosowania

Małgorzata Machowska-Szewczyk
Zachodniopomorski Uniwersytet Technologiczny w Szczecinie

OCENA KLAS W ROZMYTEJ KLASYFIKACJI OBIEKTÓW SYMBOLICZNYCH

Streszczenie: W artykule przedstawiono propozycję oceny wyników klasyfikacji rozmytej, zawartej w pracy Machowskiej-Szewczyk [2013]. Zdefiniowano w tym celu wskaźniki ogólnej niejednorodności danych symbolicznych, heterogeniczności wewnątrz klas oraz między klasami, znaczenie każdej cechy przy tworzeniu danej klasy w klasyfikacji rozmytej. Praca jest kontynuacją prowadzonych wcześniej badań nad modyfikacją procedury de Carvalho i de Souza [2010], pozwalającą wykorzystać dany algorytm do utworzenia rozmytej klasyfikacji obiektów symbolicznych.

Słowa kluczowe: analiza danych symbolicznych, klasyfikacja rozmyta, wartości symboliczne w postaci histogramu, heterogeniczność wewnątrzklasowa.

1. Wstęp

Narzędzia interpretacji klas pozwalają ocenić ogólną niejednorodność danych, heterogeniczność oraz homogeniczność klas, udział każdej zmiennej w tworzeniu danej klasy itp. Dla zwykłych ilościowych danych, podzielonych za pomocą klasycznego algorytmu grupowania, Celeux i in. [1989] wprowadzili rodzinę wskaźników przeznaczonych do interpretacji klas opartych na miarach dyspersji. Później de Carvalho i de Souza [2010] dostosowali te wskaźniki do interpretacji podziałów i odpowiadających im klas dla danych symbolicznych o wartościach w postaci histogramów, otrzymanych po etapie wstępnego przetwarzania i podzielonych za pomocą algorytmu klasyfikacji iteracyjnej, którego kryterium dopasowania jest oparte na adaptacyjnych odległościach.

W artykule przedstawiono propozycję oceny wyników rozmytej klasyfikacji obiektów, opisanych za pomocą cech symbolicznych różnego typu, zawartej w pracy Machowskiej-Szewczyk [2013]. Zdefiniowano w tym celu wskaźniki ogólnej niejednorodności danych symbolicznych, heterogeniczności wewnątrz klas oraz między klasami, znaczenie każdej cechy przy tworzeniu danej klasy w klasyfikacji rozmytej. Praca jest kontynuacją prowadzonych wcześniej badań nad modyfikacją

procedury de Carvalha i de Souzy [2010], w której wykorzystano funkcję przynależności obiektu do danej klasy, co pozwoliło zastosować dany algorytm do klasyfikacji rozmytej.

2. Ocena klas rozmytej klasyfikacji obiektów symbolicznych o różnych typach cech

Niech $\Omega = \{1,...,n\}$ oznacza zbiór wszystkich obiektów opisywanych przez zmienne $X_1,...,X_p$, które mogą przyjmować wartości symboliczne różnego typu. Dzięki przeprowadzonej transformacji [de Carvalho, de Souza 2010] każdy obiekt i (i = 1,…, n) jest reprezentowany przez wektor danych symbolicznych o wartościach w postaci histogramu $\tilde{\mathbf{x}}_i = \left(\tilde{x}_i^1,...,\tilde{x}_i^p\right)$, przy czym $\tilde{x}_i^j = \left(D_j, \mathbf{u}^j(i)\right)$, gdzie D_j (dziedzina zmiennej $\tilde{X}_j$ o wartościach w postaci histogramu) w zależności od typu pierwotnej zmiennej jest zbiorem kategorii, uporządkowaną listą kategorii lub listą elementarnych przedziałów, $\mathbf{u}^j(i) = \left(u_1^j(i),...,u_{H_j}^j(i)\right)$ jest wektorem wag lub skumulowanych wag, natomiast H_j liczbą elementów zbioru D_j.

Zaproponowana w pracy [Machowska-Szewczyk 2013] metoda tworzenia rozmytej klasyfikacji zbioru obiektów symbolicznych polega na poszukiwaniu takiego wektora $[\mu_1,...,\mu_K]$ stopni przynależności do klas, wartości wektora wzorców klas $\mathbf{g}_1^{(t)},...,\mathbf{g}_K^{(t)}$ oraz wektora wag dla każdej klasy $\boldsymbol{\lambda}_k = \left(\lambda_k^1,...,\lambda_k^p\right)\left(k = 1,...,K\right)$, aby funkcja kryterialna osiągnęła minimum:

$$\sum_{k=1}^{K}\sum_{i=1}^{n}\left[\mu_k(i)\right]^r d\left(\tilde{\mathbf{x}}_i, \mathbf{g}_k / \boldsymbol{\lambda}_k\right) \to \min,$$

przyjmując, że $r > 1$ oznacza stopień rozmycia, $\mu_k(i)$ zaś stopień przynależności obiektu i do klasy C_k oraz $\sum_{k=1}^{K}\mu_k(i) = 1$.

W kolejnych krokach algorytmu iteracyjnego wyznaczane są:

1. wartości wektora wzorców klas $\mathbf{g}_1,...,\mathbf{g}_K$, przy czym $\mathbf{g}_k = (g_k^1, g_k^2,..., g_k^p)$, $k \in \{1,...,K\}$ oraz $g_k^j = \left(D_j, \mathbf{v}_j(k)\right)$, $j \in \{1,...,p\}$, za pomocą równości:

$$v_h^j(k) = \frac{\sum_{i=1}^{n}\left[\mu_k(i)\right]^r u_h^j(i)}{\sum_{i=1}^{n}\left[\mu_k(i)\right]^r},$$

2. wartości wektora wag dla poszczególnych zmiennych oraz klas ze wzoru:

$$\lambda_k^j = \frac{\left\{\chi \prod_{l=1}^{p}\left(\sum_{i=1}^{n}\left[\mu_k(i)\right]^r \sum_{h=1}^{H_l}\left(u_h^l(i)-v_h^l(k)\right)^2\right)\right\}^{\frac{1}{p}}}{\sum_{i=1}^{n}\left[\mu_k(i)\right]^r \sum_{h=1}^{H_j}\left(u_h^j(i)-v_h^j(k)\right)^2},$$

3. wartości nowych stopni przynależności $\mu^{(t+1)} = \left\{\mu_1^{(t+1)},...,\mu_K^{(t+1)}\right\}$:

$$\mu_k(i) = \frac{\left[\sum_{j=1}^{p}\lambda_k^j \sum_{h=1}^{H_j}\left(u_h^j(i)-v_h^j(i)\right)^2\right]^{-1/(r-1)}}{\sum_{q=1}^{K}\left[\sum_{j=1}^{p}\lambda_q^j \sum_{h=1}^{H_j}\left(u_h^j(i)-v_h^j(i)\right)^2\right]^{-1/(r-1)}}.$$

Poszczególne kroki tego algorytmu, począwszy od zadanego wstępnie podziału rozmytego, są powtarzane do momentu, aż suma wartości bezwzględnych różnic w stopniach przynależności danym kroku i kroku poprzednim nie różni się o więcej niż ustalona z góry liczba bliska zeru. Metoda ta, uwzględniająca częściową przynależność obiektów do klas, jest uogólnieniem metody de Carvalha i de Souzy.

Niech $\mu_1,...,\mu_K$ będzie rozmytym podziałem zbioru Ω na K klas, który otrzymano za pomocą adaptacyjnego algorytmu rozmytej klasyfikacji iteracyjnej, prezentowanego w pracy [Machowska-Szewczyk 2013]. Niech $\mathbf{g}_k = (g_k^1,...,g_k^p)$, $g_k^j = (D_j, \mathbf{v}^j(k))$, $(j = 1,...,p)$ będzie symbolicznym opisem reprezentującym klasę C_k o wartościach w postaci histogramu, gdzie $\mathbf{v}^j(k) = (v_1^j(k),...,v_{H_j}^j(k))$. Ponadto wektor histogramów $\mathbf{g} = (g^1,...,g^p)$, $g^j = (D_j, \mathbf{v}^j)$, $(j = 1,...,p)$, gdzie $\mathbf{v}^j = (v_1^j,...,v_{H_j}^j)$, jest ogólnym reprezentantem obiektów należących do Ω.

W dalszej części będą zdefiniowane trzy sumy kwadratów dla tego podziału: ogólna *T*, wewnątrz klas *W*, między klasami *B*. Miary te są podstawą zdefiniowania narzędzi oceny klas.

Ogólna heterogeniczność obiektów należących do Ω jest mierzona przez *ogólną sumę kwadratów* według zastosowanej funkcji odległości:

$$T = \sum_{i=1}^{n}\sum_{k=1}^{K}\left[\mu_k(i)\right]^r d\left(\tilde{\mathbf{x}}_i, \mathbf{g} / \boldsymbol{\lambda}_k\right), \tag{1}$$

gdzie $\boldsymbol{\lambda}_k = \left(\lambda_k^1,...,\lambda_k^p\right)\left(k = 1,...,K\right)$ są wektorami wag zmieniającymi się w każdej iteracji i mogą być niejednakowe dla poszczególnych klas, $r > 1$ oznacza stopień rozmycia, $\mu_k(i)$ stopień przynależności obiektu *i* do klasy C_k oraz zachodzi równość: $\sum_{k=1}^{K}\mu_k(i) = 1$.

Można przyjąć, że odległość *d* między obiektem symbolicznym a wzorcem zbioru lub wzorcami klas może być wyrażona za pomocą kwadratu odległości euklidesowej, wtedy:

$$d\left(\tilde{\mathbf{x}}_i, \mathbf{g} / \boldsymbol{\lambda}_k\right) = \sum_{j=1}^{p} \lambda_k^j \sum_{h=1}^{H_j} \left(u_h^j(i) - v_h^j\right)^2, \tag{2}$$

$$d\left(\tilde{\mathbf{x}}_i, \mathbf{g}_k / \boldsymbol{\lambda}_k\right) = \sum_{j=1}^{p} \lambda_k^j \sum_{h=1}^{H_j} \left(u_h^j(i) - v_h^j(k)\right)^2. \tag{3}$$

Twierdzenie 2.1[1]

Wektor ogólnego reprezentanta zbioru obiektów $\mathbf{g} = \left(g^1, ..., g^p\right)$, $g^j = \left(D_j, \mathbf{v}^j\right)$, $\left(j = 1, ..., p\right)$, który minimalizuje ogólną dyspersję *T* ma składowe $v_h^j \left(h = 1, ..., H_j\right)$ wektora wag $\mathbf{v}^j = \left(v_1^j, ..., v_{H_j}^j\right)$ obliczane według wzoru:

$$v_h^j = \frac{\sum_{k=1}^{K} \lambda_k^j \sum_{i=1}^{n} \left[\mu_k(i)\right]^r u_h^j(i)}{\sum_{k=1}^{K} \lambda_k^j \sum_{i=1}^{n} \left[\mu_k(i)\right]^r}. \tag{4}$$

Ogólną sumę kwadratów *T* można przedstawić w jednej z następujących postaci:

a) $T = \sum_{k=1}^{K} T_k$, gdzie $T_k = \sum_{i=1}^{n} \left[\mu_k(i)\right]^r d\left(\tilde{\mathbf{x}}_i, \mathbf{g} / \boldsymbol{\lambda}_k\right)$,

b) $T = \sum_{j=1}^{p} T_j$, gdzie $T_j = \sum_{i=1}^{n} \sum_{k=1}^{K} \lambda_k^j \left[\mu_k(i)\right]^r \sum_{h=1}^{H_j} (u_h^j(i) - v_h^j)^2$,

c) $T = \sum_{k=1}^{K} \left(\sum_{j=1}^{p} T_{kj}\right)$, gdzie $T_{kj} = \lambda_k^j \sum_{i=1}^{n} \left[\mu_k(i)\right]^r \sum_{h=1}^{H_j} (u_h^j(i) - v_h^j)^2$.

Podobnie możemy mierzyć heterogeniczność wewnątrz klas za pomocą *sumy kwadratów wewnątrz klas*:

$$W = \sum_{i=1}^{n} \sum_{k=1}^{K} \left[\mu_k(i)\right]^r d\left(\tilde{\mathbf{x}}_i, \mathbf{g}_k / \boldsymbol{\lambda}_k\right). \tag{5}$$

Suma kwadratów wewnątrzklasowa *W* rozkłada się jako:

a) $W = \sum_{k=1}^{K} W_k$, gdzie $W_k = \sum_{i=1}^{n} \left[\mu_k(i)\right]^r d\left(\tilde{\mathbf{x}}_i, \mathbf{g}_k / \boldsymbol{\lambda}_k\right)$;

b) $W = \sum_{j=1}^{p} W_j$, gdzie $W_j = \sum_{k=1}^{K} \lambda_k^j \sum_{i=1}^{n} \left[\mu_k(i)\right]^r \sum_{h=1}^{H_j} (u_h^j(i) - v_h^j(k))^2$;

c) $W = \sum_{k=1}^{K} \left(\sum_{j=1}^{p} W_{kj}\right)$, gdzie $W_{kj} = \lambda_k^j \sum_{i=1}^{n} \left[\mu_k(i)\right]^r \sum_{h=1}^{H_j} (u_h^j(i) - v_h^j(k))^2$.

Suma kwadratów między klasami jest dana za pomocą równości:

$$B = \sum_{k=1}^{K} \sum_{i=1}^{n} \left[\mu_k(i)\right]^r d\left(\mathbf{g}_k, \mathbf{g} / \boldsymbol{\lambda}_k\right). \tag{6}$$

[1] Dowody twierdzeń pominięto ze względu na ograniczoną liczbę stron artykułu.

Mierzy ona dyspersję, jaka jest między reprezentantami klas a ogólnym reprezentantem zbioru Ω. Sumę kwadratów między grupami można rozłożyć jako:

a) $B = \sum_{k=1}^{K} B_k$, gdzie $B_k = \sum_{i=1}^{n} [\mu_k(i)]^r \sum_{j=1}^{p} \lambda_k^j \sum_{h=1}^{H_j} (v_h^j(k) - v_h^j)^2$;

b) $B = \sum_{j=1}^{p} B_j$, gdzie $B_j = \sum_{k=1}^{K} \lambda_k^j \sum_{i=1}^{n} [\mu_k(i)]^r \left[\sum_{h=1}^{H_j} (v_h^j(k) - v_h^j) \right]^2$;

c) $B = \sum_{k=1}^{K} \left(\sum_{j=1}^{p} B_{kj} \right)$, gdzie $B_{kj} = \sum_{i=1}^{n} [\mu_k(i)]^r \lambda_k^j \left[\sum_{h=1}^{H_j} (v_h^j(k) - v_h^j)^2 \right]$.

Twierdzenie 2.2

Jeżeli odległości między obiektem symbolicznym a wzorcem zbioru lub klasy są dane za pomocą wzorów (2) lub (3), to spełnione są następujące równości:

$$\begin{aligned} T &= W + B, \\ T_k &= W_k + B_k \ (k = 1,..., K), \\ T_j &= W_j + B_j \ (j = 1,..., p), \\ T_{kj} &= W_{kj} + B_{kj} \ (k = 1,..., K, \ j = 1,..., p). \end{aligned} \tag{7}$$

Ogólny wskaźnik heterogeniczności podziału *R* jest zdefiniowany jako:

$$R = \frac{B}{T} = \frac{B}{W + B} = 1 - \frac{W}{T}. \tag{8}$$

Wyraża on, jaka część ogólnej sumy kwadratów została wyjaśniona przez podział $\mu_1,..., \mu_K$. Większa wartość *R* prowadzi do bardziej jednorodnych klas i lepszej reprezentacji elementów z klasy C_k przez ich wzorzec $\mathbf{g}_k \ (k = 1,..., K)$.

Siłę dyskryminacyjną ustalonej zmiennej symbolicznej w klasyfikacji rozmytej można ocenić za pomocą *wskaźnika heterogeniczności zmiennej*:

$$COR(j) = \frac{B_j}{T_j} = \frac{B_j}{W_j + B_j}. \tag{9}$$

Porównując wartość $COR(j)$ z wartością ogólnego wskaźnika heterogeniczności *R*, który mierzy średnią siłę dyskryminacyjną wszystkich zmiennych, można ocenić, czy siła dyskryminacyjna zmiennej X_j jest powyżej, czy poniżej poziomu średniego.

Względny udział zmiennej X_j w międzygrupowej sumie kwadratów *B* jest dany przez równość:

$$CTR(j) = \frac{B_j}{B}. \tag{10}$$

Zauważmy, że $\sum_{j=1}^{p} CTR(j) = 1$. Wysoka wartość $CTR(j)$ wskazuje, że zmienna X_j ma bardzo duże znaczenie w tworzeniu wzorca klas. Interesująca jest sytuacja, gdy $COR(j)$ ma niską wartość, zaś $CTR(j)$ wysoką – oznacza to, że zmienna X_j ma słabą siłę dyskryminacyjną, chociaż ma duży udział w międzygrupowej sumie kwadratów [Celeux i in. 1989].

Udział klasy C_k w ogólnej sumie kwadratów jest określony przez:

$$T(k) = \frac{T_k}{T}. \tag{11}$$

Udział klasy C_k w międzygrupowej sumie kwadratów jest mierzony stosunkiem:

$$B(k) = \frac{B_k}{B}. \tag{12}$$

Wysoka wartość $B(k)$ wskazuje, że klasa C_k jest dość odległa od globalnego centrum.

Udział klasy C_k w wewnątrzgrupowej sumie kwadratów jest dany jako:

$$W(k) = \frac{W_k}{W}. \tag{13}$$

Stosunkowo duża wartość $W(k)$ wskazuje, że klasa C_k jest dość zróżnicowana w porównaniu z innymi klasami.

Udział siły dyskryminacyjnej zmiennej X_j w odniesieniu do klasy C_k jest określony przez równość:

$$COR(j,k) = \frac{B_{kj}}{T_j}. \tag{14}$$

Zauważmy, że $\sum_{k=1}^{K} COR(j,k) = COR(j)$. Wysoka wartość $COR(j,k)$ oznacza, że zdolność dyskryminacyjna zmiennej X_j jest niewielka w klasie C_k.

Wpływ zmiennej X_j na heterogeniczność klasy C_k jest mierzony za pomocą wskaźnika:

$$CTR(j,k) = \frac{B_{kj}}{B_k}. \tag{15}$$

Można również rozważyć względny udział zmiennej X_j i klasy C_k w międzygrupowej sumie kwadratów jako:

$$CE(j,k) = \frac{B_{kj}}{B}. \tag{16}$$

Jeżeli $CE(j,k)$ jest bliskie 1, to zmienna X_j ma duży wpływ na profil klasy C_k.

3. Ocena eksperymentalna

Jako przykład wykorzystano zbiór 37 miast opisanych za pomocą 12 zmiennych symbolicznych o wartościach w postaci przedziałów, które zostały utworzone na podstawie minimalnej oraz maksymalnej temperatury w stopniach Celsjusza w poszczególnych miesiącach ustalonego roku [Guru i in. 2004]. Klasyfikacja miast otrzymana za pomocą algorytmu de Carvalha i de Souzy z odległościami parametryzowanymi przez wagi jednakowe w każdej klasie dała następujący podział na cztery klasy:

Klasa 1: Bahrajn, Bombaj, Kair, Kalkuta, Colombo, Dubaj, Hongkong, Kuala Lumpur, Madras, Manila, New Delhi, Singapur.

Klasa 2: Ateny, Madryt, Rzym, Seul, Tokio, Lizbona, Nowy Jork, San Francisco, Teheran.

Klasa 3: Amsterdam, Frankfurt, Londyn, Monachium, Sztokholm, Wiedeń, Kopenhaga, Genewa, Moskwa, Paryż, Toronto, Zürich.

Klasa 4: Mauritius, Nairobi, Meksyk, Sydney.

Do tego zbioru zastosowano również procedurę klasyfikacji rozmytej z odległościami parametryzowanymi przez wagi jednakowe w każdej klasie, opisaną w [Machowska-Szewczyk 2013]. Fragment macierzy stopni przynależności do poszczególnych klas przedstawiono w tabeli 1.

Tabela 1. Stopnie przynależności do klas w klasyfikacji rozmytej

Miasta	Klasa 1	Klasa 2	Klasa 3	Klasa 4
Amsterdam	0,0196	0,0919	0,8484	0,0401
Ateny	0,0919	0,5461	0,0840	0,2780
Bahrajn	0,5900	0,1148	0,0617	0,2335
Bombaj	0,9154	0,0219	0,0124	0,0503
Kair	0,3399	0,1479	0,0547	0,4575
…	…	…	…	…
Sydney	0,1577	0,2259	0,1685	0,4479
Teheran	0,2010	0,3615	0,1749	0,2626
…	…	…	…	…
Zürich	0,0478	0,3401	0,4909	0,1212

Źródło: obliczenia własne w programie Excel.

Część ogólnej sumy kwadratów wyjaśniana przez podział na cztery klasy w klasyfikacji de Carvalha i de Souzy wyniosła $R_1 = 0{,}796$, natomiast dla klasyfikacji rozmytej $R_2 = 0{,}856$ (patrz równość (8)). Porównując wartości *COR* (patrz tab. 2)

dla poszczególnych zmiennych z wartościami R_2 dla podziału na cztery klasy otrzymanego przez zastosowanie iteracyjnej metody klasyfikacji rozmytej, możemy wywnioskować, że siła dyskryminacyjna zmiennych: kwiecień, maj, czerwiec, wrzesień, październik jest powyżej średniej, podczas gdy wszystkie inne zmienne mają siłę dyskryminacyjną niższą od średniej. Co więcej, zmienne: kwiecień, maj i październik mają duży wpływ na rozdzielenie klas (*CTR* >10%).

Tabela 2. Wartości wskaźników heterogeniczności dla zmiennych w klasyfikacji rozmytej

	1	2	3	4	5	6	7	8	9	10	11	12
COR	0,782	0,812	0,849	0,888	0,891	0,861	0,799	0,814	0,857	0,898	0,782	0,812
CTR	0,051	0,061	0,080	0,113	0,116	0,088	0,056	0,062	0,085	0,125	0,051	0,061

Źródło: opracowanie własne w programie Excel.

Na podstawie analizy wartości wskaźników heterogeniczności klas przedstawionych w tabeli 3 można wywnioskować, że obiekty reprezentujące klasę 4 są najbardziej zbliżone do reprezentanta całego zbioru miast ($B(4) = 0{,}065$, $W(4) = 0{,}222$).

Tabela 3. Wartości wskaźników heterogeniczności klas w klasyfikacji rozmytej

Klasa	1	2	3	4	Wartości sum kwadratów dla całego zbioru miast
$T(k)$	0,147	0,210	0,555	0,088	747,911
$B(k)$	0,129	0,205	0,601	0,065	843,741
$W(k)$	0,255	0,242	0,281	0,222	191,054

Źródło: opracowanie własne w programie Excel.

Tabela 4. Wartości wskaźników niejednorodności klas dotyczące zmiennych

	Klasa 1			Klasa 2			Klasa 3			Klasa 4		
	COR	*CTR*	*CE*	*COR*	*CTR*	*CE*	*COR*	*CTR*	*CE*	*COR*	*CTR*	*CE*
1	0,076	0,097	0,005	0,177	0,226	0,012	0,496	0,634	0,032	0,033	0,043	0,002
2	0,091	0,112	0,007	0,180	0,222	0,014	0,505	0,621	0,038	0,037	0,045	0,003
3	0,094	0,111	0,009	0,205	0,241	0,019	0,508	0,598	0,048	0,043	0,051	0,004
4	0,109	0,122	0,014	0,212	0,239	0,027	0,512	0,577	0,065	0,056	0,063	0,007
5	0,120	0,135	0,016	0,195	0,219	0,025	0,512	0,575	0,066	0,064	0,072	0,008
6	0,129	0,150	0,013	0,167	0,194	0,017	0,493	0,573	0,050	0,072	0,084	0,007
7	0,136	0,170	0,010	0,116	0,145	0,008	0,467	0,584	0,033	0,080	0,100	0,006
8	0,143	0,175	0,011	0,116	0,142	0,009	0,475	0,584	0,036	0,081	0,099	0,006
9	0,113	0,131	0,011	0,136	0,159	0,014	0,548	0,640	0,055	0,060	0,070	0,006
10	0,115	0,128	0,016	0,177	0,197	0,025	0,555	0,617	0,077	0,052	0,058	0,007
11	0,097	0,111	0,011	0,192	0,220	0,021	0,534	0,612	0,060	0,051	0,058	0,006
12	0,090	0,110	0,007	0,182	0,221	0,014	0,511	0,623	0,040	0,038	0,046	0,003

Źródło: opracowanie własne w programie Excel.

W tabeli 4 umieszczono wartości wskaźników heterogeniczności klas, dotyczące pojedynczych zmiennych. Na tej podstawie można zauważyć, że zmienne: sierpień, marzec, wrzesień, lipiec odgrywają bardzo ważną rolę w heterogeniczności klas 1, 2, 3 i 4 odpowiednio ($CTR(8,1) = 17{,}52\%$, $CTR(3,2) = 24{,}1\%$, $CTR(9,3) = 64\%$, $CTR(7,4) = 10\%$). Ponadto zmienne: sierpień, kwiecień, październik, sierpień mają najbardziej jednorodne zachowania w klasach odpowiednio: 1, 2, 3, 4 ($COR(8,1) = 14{,}27\%$, $COR(4,2) = 21{,}19\%$, $COR(10,3) = 55{,}45\%$, $COR(8,4) = 8{,}07\%$). Wreszcie temperatury w październiku, kwietniu, maju miały największy udział w profilowaniu klas odpowiednio 1, 2, 3 i 4 ($CE(10,1) = 1{,}6\%$, $CE(4,2) = 2{,}7\%$, $CE(10,3) = 7{,}74\%$, $CE(5,4) = 0{,}83\%$).

Porównując podział uzyskany metodą klasyfikacji de Carvalha i de Souzy z podziałem rozmytym, można zauważyć dość dużą zgodność. Jednak dzięki zastosowaniu częściowej przynależności do klas można wykryć, że obiekty: Teheran i Sydney są w dużym stopniu podobne do kilku klas. Podobnie Kair należy do klasy czwartej z najwyższym stopniem przynależności, jednak jest również w znacznym stopniu podobny do obiektów z klasy pierwszej (tabela 1). W wyniku klasycznego podziału wymuszona jest przynależność obiektów do jednej klasy i w przypadku występowania „mieszańców" następuje strata informacji.

4. Podsumowanie

Algorytm rozmytej klasyfikacji obiektów reprezentowanych przez cechy symboliczne różnego typu pozwala wykryć obiekty o dużym podobieństwie do kilku klas jednocześnie.

W pracy zaprezentowano propozycje oceny jakości klasyfikacji uzyskanych metodą klasyfikacji rozmytej z uwzględnieniem wag oraz pokazano na przykładzie, że w przypadku zbioru obiektów trudno separowalnych klasyfikacja rozmyta może dać lepszą jakość, mierzoną za pomocą narzędzi wykorzystujących sumy kwadratów zmienności międzygrupowej, wewnątrzgrupowej oraz ogólnej. Opracowane wskaźniki heterogeniczności klas, zmiennych oraz podziału umożliwiają ocenę wyników klasyfikacji. Wyniki eksperymentalne potwierdzają dobrą jakość klasyfikacji w przypadku trudno separowalnych klas.

Kierunkiem dalszych badań będzie poszukiwanie wskaźnika służącego do porównania wyników rozmytej klasyfikacji uwzględniającej wagi z wynikami otrzymanymi za pomocą innych metod klasyfikacji rozmytej [Machowska-Szewczyk 2013].

Literatura

Celeux G., Diday E., Govaert G., Lechevallier Y., Ralambondrainy H. (1989), *Classification Automatique des Données*, Bordas, Paris.

De Carvalho F.A.T., de Souza R. (2010), *Unsupervised pattern recognition models for mixed feature-type symbolic data*, Pattern Recognition Letters 31, s. 430-443.

Guru D.S., Kiranagi B.B., Nagabhushan P. (2004), *Multivalued type dissimilarity measure and concept of mutual dissimilarity value for clustering symbolic patterns*, Pattern Recognition 38, s. 1203-1213.

Machowska-Szewczyk M. (2013), *Klasyfikacja obiektów reprezentowanych przez różnego rodzaju cechy symboliczne*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 278, Taksonomia 20, Wydawnictwo UE, Wrocław, s. 290-299.

EVALUATION OF CLUSTERS OBTAINED BY FUZZY CLASSIFICATION METHODS FOR SYMBOLIC OBJECTS

Summary: The aim of this work is to present the evaluation proposition of classes of fuzzy classification algorithm. For this purpose overall heterogeneity indexes of symbolic data, intra-cluster heterogeneity and between clusters heterogeneity as well as the importance of every variable in the formation of a given cluster in fuzzy classification were discussed. The work is a continuation of previous studies on the modification of Carvalho and Souza' algorithm [2010] that allows using the algorithm to create a fuzzy classification of symbolic objects.

Keywords: symbolic data analysis, fuzzy classification, histogram-valued symbolic data, intra-cluster heterogeneity.