

**PRACE NAUKOWE**

Uniwersytetu Ekonomicznego we Wrocławiu

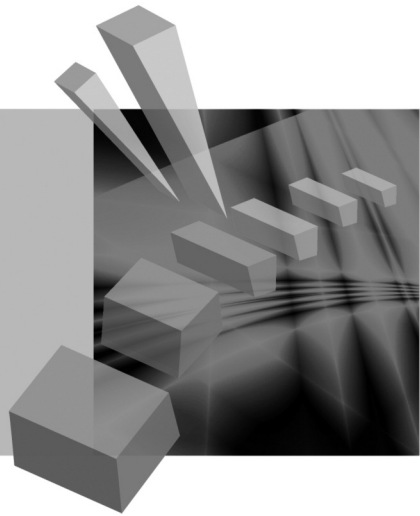
**RESEARCH PAPERS**

of Wrocław University of Economics

**242**

# **Taksonomia 19.**

## **Klasyfikacja i analiza danych – teoria i zastosowania**



Redaktorzy naukowi  
**Krzysztof Jajuga**  
**Marek Walesiak**



Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu  
Wrocław 2012

Recenzenci: Eugeniusz Gatnar, Elżbieta Gołata, Tadeusz Kufel, Józef Pocięcha,  
Miroslaw Szreder, Feliks Wysocki

Redaktor Wydawnictwa: Aleksandra Śliwka

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Małgorzata Czupryńska

Projekt okładki: Beata Dębska

Tytuł sfinansowano ze środków Sekcji Klasyfikacji i Analizy Danych PTS  
i Uniwersytetu Ekonomicznego we Wrocławiu

Publikacja jest dostępna na stronie [www.ibuk.pl](http://www.ibuk.pl)

Streszczenia opublikowanych artykułów są dostępne w międzynarodowej bazie danych  
The Central European Journal of Social Sciences and Humanities <http://cejsh.icm.edu.pl>  
oraz w The Central and Eastern European Online Library [www.ceeol.com](http://www.ceeol.com),  
a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon [http://kangur.uek.krakow.pl/  
bazy\\_ae/bazekon/nowy/index.php](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](http://www.wydawnictwo.ue.wroc.pl)

Kopowanie i powielanie w jakiegokolwiek formie  
wymaga pisemnej zgody Wydawcy

© Copyright by Uniwersytet Ekonomiczny we Wrocławiu  
Wrocław 2012

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

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM  
Nakład: 320 egz.

## Spis treści

<b>Wstęp</b> .....	13
<b>Stanisława Bartosiewicz</b> , Jeszcze raz o skutkach subiektywizmu w analizie wielowymiarowej .....	17
<b>Andrzej Sokolowski</b> , Q uniwersalna miara odległości .....	22
<b>Eugeniusz Gatnar</b> , Jakość danych w systemach statystycznych banków centralnych (na przykładzie NBP) .....	31
<b>Marek Walesiak</b> , Pomiar odległości obiektów opisanych zmiennymi mierzonymi na skali porządkowej – strategię postępowania.....	39
<b>Krzysztof Jajuga, Marek Walesiak</b> , XXV lat konferencji taksonomicznych – fakty i refleksje .....	47
<b>Józef Pocięcha, Barbara Pawelek</b> , Model SEM w analizie zagrożenia bankructwem przedsiębiorstw w świetle koniunktury gospodarczej – problemy teoretyczne i praktyczne .....	50
<b>Paweł Lula</b> , Uczące się systemy pozyskiwania informacji z dokumentów tekstowych .....	58
<b>Ewa Roszkowska</b> , Zastosowanie metody TOPSIS do wspomaganie procesu negocjacji.....	68
<b>Andrzej Młodak</b> , Sąsiedztwo obszarów przestrzennych w ujęciu fizycznym oraz społeczno-ekonomicznym – podejście taksonomiczne .....	76
<b>Andrzej Bąk</b> , Modele kategorii nieuporządkowanych w badaniach preferencji .....	86
<b>Jacek Kowalewski</b> , Zintegrowany model optymalizacji badań statystycznych.....	96
<b>Jan Paradysz, Karolina Paradysz</b> , Obszary bezrobocia w Polsce – problem benchmarkowy.....	106
<b>Tomasz Szubert</b> , W co grać, aby jak najmniej przegrać? Próba klasyfikacji systemów gry w zakładach bukmacherskich.....	116
<b>Izabela Szamrej-Baran</b> , Klasyfikacja krajów UE ze względu na ubóstwo energetyczne .....	126
<b>Sylwia Filas-Przybył, Tomasz Klimanek, Jacek Kowalewski</b> , Analiza dojazdów do pracy za pomocą modelu grawitacji.....	135
<b>Marta Dziechciarz-Duda, Anna Król, Klaudia Przybysz</b> , Minimum egzystencji a czynniki warunkujące skłonność do korzystania z pomocy społecznej. Klasyfikacja gospodarstw domowych .....	144
<b>Hanna Dudek</b> , Subiektywne skale ekwiwalentności – analiza na podstawie danych o satysfakcji z osiągniętych dochodów .....	153

<b>Joanicjusz Nazarko, Ewa Chodakowska, Marta Jaročka</b> , Segmentacja szkół wyższych metodą analizy skupień <i>versus</i> konkurencja technologiczna ustalona metodą DEA – studium komparatywne.....	163
<b>Ewa Chodakowska</b> , Wybrane metody klasyfikacji w konstrukcji ratingu szkół.....	173
<b>Bartosz Soliński</b> , Sektor energetyki odnawialnej w krajach Unii Europejskiej – klasyfikacja w świetle strategii zarządzania zmianą.....	182
<b>Krzysztof Szwarz</b> , Klasyfikacja powiatów województwa wielkopolskiego ze względu na sytuację demograficzną.....	192
<b>Elżbieta Gołata, Grażyna Dehnel</b> , Rejestry administracyjne w analizie przedsiębiorczości.....	202
<b>Katarzyna Chudy, Marek Sobolewski, Kinga Stępień</b> , Wykorzystanie metod taksonomicznych w prognozowaniu wskaźników rentowności banków giełdowych w Polsce.....	212
<b>Katarzyna Dębowska</b> , Modelowanie upadłości przedsiębiorstw przy wykorzystaniu metod dyskryminacji i regresji.....	222
<b>Alina Bojan</b> , Wykorzystanie metod wielowymiarowej analizy danych do identyfikacji zmiennych wpływających na atrakcyjność wybranych inwestycji.....	231
<b>Justyna Brzezińska</b> , Analiza logarytmiczno-liniowa w badaniu przyczyn umieralności w krajach UE.....	240
<b>Aneta Rybicka, Bartłomiej Jefmański, Marcin Pelka</b> , Analiza klas ukrytych w badaniach satysfakcji studentów.....	247
<b>Bartłomiej Jefmański</b> , Pomiar opinii respondentów z wykorzystaniem elementów teorii zbiorów rozmytych i środowiska R.....	256
<b>Julita Stańczuk</b> , Porównanie rezultatów wielostanowej klasyfikacji obiektów ekonomicznych z wykorzystaniem analizy dyskryminacyjnej oraz sieci neuronowych.....	265
<b>Jerzy Krawczuk</b> , Skuteczność metod klasyfikacji w prognozowaniu kierunku zmian indeksu giełdowego S&P500.....	275
<b>Anna Czapkiewicz, Beata Basiura</b> , Symulacyjne badanie wpływu zaburzeń na grupowanie szeregów czasowych na podstawie modelu Copula-GARCH.....	283
<b>Radosław Pietrzyk</b> , Ocena efektywności inwestycji funduszy inwestycyjnych z tytułu doboru papierów wartościowych i umiejętności wykorzystania trendów rynkowych.....	291
<b>Aleksandra Witkowska, Marek Witkowski</b> , Zastosowanie metody Panzara-Rosse’a do pomiaru poziomu konkurencji w sektorze banków spółdzielczych.....	306
<b>Marcin Pelka</b> , Podejście wielomodelowe z wykorzystaniem metody <i>boosting</i> w analizie danych symbolicznych.....	315
<b>Justyna Wilk</b> , Analiza porównawcza oprogramowania komputerowego w klasyfikacji danych symbolicznych.....	323

<b>Tomasz Bartłomowicz, Justyna Wilk</b> , Zastosowanie metod analizy danych symbolicznych w przeszukiwaniu dziedzinowych baz danych.....	333
<b>Kamila Migdał-Najman</b> , Propozycja hybrydowej metody grupowania opartej na sieciach samouczących .....	342
<b>Dorota Rozmus</b> , Porównanie dokładności taksonomii spektralnej oraz zagregowanych algorytmów taksonomicznych opartych na idei metody <i>bagging</i> .....	352
<b>Krzysztof Najman</b> , Grupowanie dynamiczne z wykorzystaniem samouczących się sieci GNG .....	361
<b>Małgorzata Misztal</b> , Wpływ wybranych metod uzupełniania brakujących danych na wyniki klasyfikacji obiektów z wykorzystaniem drzew klasyfikacyjnych w przypadku zbiorów danych o niewielkiej liczebności – ocena symulacyjna .....	370
<b>Mariusz Kubus</b> , Zastosowanie wstępnego uwarunkowania zmiennej objaśnianej do selekcji zmiennych.....	380
<b>Barbara Batóg, Jacek Batóg</b> , Wykorzystanie analizy dyskryminacyjnej do identyfikacji czynników determinujących stopę zwrotu z inwestycji na rynku kapitałowym .....	387
<b>Katarzyna Wójcik, Janusz Tuchowski</b> , Analiza porównawcza miar podobieństwa tekstów opartych na macierzy częstości i tekstów opartych na wiedzy dziedzinowej .....	396
<b>Iwona Staniec</b> , Analiza czynnikowa w identyfikacji obszarów determinujących doskonalenie systemów zarządzania w polskich organizacjach .....	406
<b>Marek Lubicz, Maciej Zięba, Adam Rzechonek, Konrad Pawełczyk, Jerzy Kołodziej, Jerzy Błaszczyk</b> , Analiza porównawcza wybranych technik eksploracji danych do klasyfikacji danych medycznych z brakującymi obserwacjami .....	416
<b>Iwona Foryś</b> , Wykorzystanie analizy log-liniowej do wyboru czynników determinujących atrakcyjność cenową mieszkań w obrocie wtórnym na przykładzie lokalnego rynku mieszkaniowego.....	426
<b>Ewa Genge</b> , Analiza skupień oparta na mieszankach uciętych rozkładów normalnych.....	436
<b>Jerzy Korzeniewski</b> , Ocena efektywności metody uśredniania zmiennych i metody Ichino selekcji zmiennych w analizie skupień .....	444
<b>Andrzej Dudek</b> , SMS – propozycja nowego algorytmu analizy skupień .....	451
<b>Artur Mikulec</b> , Metody oceny wyniku grupowania w analizie skupień.....	460
<b>Małgorzata Machowska-Szewczyk</b> , Algorytm klasyfikacji rozmytej dla obiektów opisanych za pomocą zmiennych symbolicznych oraz rozmytych .....	469
<b>Artur Zaborski</b> , Analiza PROFIT i jej wykorzystanie w badaniu preferencji .....	479
<b>Karolina Bartos</b> , Analiza skupień wybranych państw ze względu na strukturę wydatków konsumpcyjnych obywateli – zastosowanie sieci Kohonena .....	488

<b>Barbara Batóg, Magdalena Mojsiewicz, Katarzyna Wawrzyniak</b> , Klasyfikacja gospodarstw domowych ze względu na bodźce do zawierania umowy o ubezpieczenie z wykorzystaniem modeli zmiennych jakościowych.	496
<b>Izabela Kurzawa</b> , Zastosowanie modelu LA/AIDS do badania elastyczności cenowych popytu konsumpcyjnego w gospodarstwach domowych w relacji miasto–wieś .....	505
<b>Aleksandra Łuczak, Feliks Wysocki</b> , Metody porządkowania liniowego obiektów opisanych za pomocą cech metrycznych i porządkowych .....	513
<b>Agnieszka Sompolska-Rzechuła</b> , Porównanie klasycznej i pozycyjnej taksonomicznej analizy zróżnicowania jakości życia w województwie zachodniopomorskim .....	523
<b>Joanna Banaś, Małgorzata Machowska-Szewczyk</b> , Ocena intensywności wykorzystania skrzynek poczty elektronicznej za pomocą uporządkowanego modelu probitowego .....	532
<b>Iwona Bąk</b> , Segmentacja gospodarstw domowych emerytów i rencistów pod względem wydatków na rekreację i kulturę .....	541
<b>Aneta Becker</b> , Zastosowanie metody ANP do porządkowania województw Polski pod względem dynamiki wykorzystania ICT w latach 2008-2010	552
<b>Katarzyna Dębowska</b> , Klasyfikacja sektorów ze względu na ich kondycję finansową przy użyciu metod wielowymiarowej analizy statystycznej .....	562
<b>Anna Domagała</b> , Propozycja metody doboru zmiennych do modeli DEA (procedura kombinowanego doboru w przód).....	571
<b>Henryk Gierszal, Karina Pawlina, Maria Urbańska</b> , Analiza statystyczna w badaniach zapotrzebowania na usługi teleinformatyczne sieci łączności ruchomej .....	580
<b>Hanna Gruchociak</b> , Konstrukcja estymatora regresyjnego dla danych o strukturze dwupoziomowej.....	590
<b>Tomasz Klimanek, Marcin Szymkowiak</b> , Zastosowanie estymacji pośredniej uwzględniającej korelację przestrzenną w opisie niektórych charakterystyk rynku pracy .....	601
<b>Jarosław Lira</b> , Prognozowanie opłacalności produkcji żywca wieprzowego w Polsce .....	610
<b>Christian Lis</b> , Wykorzystanie metody klasyfikacji w ocenie konkurencyjności portów południowego Bałtyku .....	619
<b>Beata Bieszk-Stolorz, Iwona Markowicz</b> , Wykorzystanie wielomianowego modelu logitowego do oceny szansy podjęcia pracy przez bezrobotnych .	628
<b>Lucyna Przezbórska-Skobiej, Jarosław Lira</b> , Przestrzeń agroturystyczna Polski i ocena jej atrakcyjności.....	637
<b>Paweł Ulman</b> , Model rozkładu wydatków a funkcje popytu.....	646
<b>Maria Urbańska, Tadeusz Mizera, Henryk Gierszal</b> , Zastosowanie metod analizy statystycznej w badaniach mięczaków .....	655

## Summaries

<b>Stanisława Bartosiewicz</b> , The effects of subjectivism in multivariate analysis revisited.....	21
<b>Andrzej Sokółowski</b> , Q universal distance measure .....	30
<b>Eugeniusz Gatnar</b> , Data quality in central banks' statistical systems (NBP example) .....	38
<b>Marek Walesiak</b> , Distance measures for ordinal data – strategies of proceedings.....	46
<b>Krzysztof Jajuga, Marek Walesiak</b> , XXV years of taxonomic conferences – some facts and remarks.....	49
<b>Józef Pocięcha, Barbara Pawelek</b> , General SEM model in researching corporate bankruptcy and business cycles – theoretical and practical problems.....	57
<b>Paweł Lula</b> , Learning-based systems of information extraction from textual resources .....	67
<b>Ewa Roszkowska</b> , The application of the TOPSIS method to support the negotiation process .....	75
<b>Andrzej Młodak</b> , Neighborhood of spatial areas in the physical and socio-economic context – a taxonomic approach.....	85
<b>Andrzej Bąk</b> , Models for unordered categories in preference analysis.....	95
<b>Kowalewski Jacek</b> , An integrated model of optimizing statistical surveys ....	105
<b>Jan Paradysz, Karolina Paradysz</b> , Areas of unemployment in Poland – benchmark problem .....	115
<b>Tomasz Szubert</b> , How to play to lose the least? Classification of systems in sports bets .....	125
<b>Izabela Szamrej-Baran</b> , Classification of EU member states in view of fuel poverty .....	134
<b>Sylvia Filas-Przybył, Tomasz Klimanek, Jacek Kowalewski</b> , An attempt to use the gravity model in the analysis of commuters.....	143
<b>Marta Dziechciarz-Duda, Anna Król, Klaudia Przybysz</b> , Subsistence minimum versus factors influencing tendency to benefit from social care. Classification of households .....	152
<b>Hanna Dudek</b> , Subjective equivalence scales – analysis based on data about satisfaction with incomes.....	162
<b>Joanicjusz Nazarko, Ewa Chodakowska, Marta Jarocka</b> , Segmentation of universities using cluster analysis versus technological competitors determined by the DEA method – a comparative study .....	172
<b>Ewa Chodakowska</b> , Selected methods of classification in schools' rating.....	181
<b>Bartosz Soliński</b> , Renewable energy sector in the European Union – classification in the light of change management strategy .....	191
<b>Krzysztof Szwarz</b> , Classification of Wielkopolska voivodeship due to the demographic situation .....	201

<b>Elżbieta Gołata, Grażyna Dehnel</b> , Administrative registers in business analysis.....	211
<b>Katarzyna Chudy, Marek Sobolewski, Kinga Stępień</b> , Application of taxonomic methods in forecasting the profitability ratios of listed banks in Poland.....	221
<b>Katarzyna Dębowska</b> , Modeling bankruptcy of firms by using discrimination and regression methods.....	230
<b>Alina Bojan</b> , Identification of variables which influence attractiveness of given investments with the usage of multivariate analysis.....	239
<b>Justyna Brzezińska</b> , Log-linear analysis in the study of mortality in EU.....	246
<b>Aneta Rybicka, Bartłomiej Jefmański, Marcin Pelka</b> , Latent class analysis in student satisfaction surveys.....	254
<b>Bartłomiej Jefmański</b> , The respondent's opinions measurement in the R program with an application of fuzzy sets theory.....	264
<b>Julita Stańczuk</b> , A comparison of the results of multistate classification of economic objects using discriminant analysis and artificial neural networks.....	274
<b>Jerzy Krawczuk</b> , Effectiveness of classification methods in S&P500 stock index direction changes forecasting.....	282
<b>Anna Czapkiewicz, Beata Basiura</b> , The simulation study of the utility of the Copula-GARCH models for clustering financial time series.....	290
<b>Radosław Pietrzyk</b> , Timing and selectivity in mutual funds performance measurement.....	305
<b>Aleksandra Witkowska, Marek Witkowski</b> , Use of the Panzar-Rosse method to assess of the competition level in the cooperative banks sector.....	314
<b>Marcin Pelka</b> , Ensemble learning with the application of <i>boosting</i> in symbolic data analysis.....	322
<b>Justyna Wilk</b> , Comparative study of symbolic data classification software.....	332
<b>Tomasz Bartłomowicz, Justyna Wilk</b> , Application of symbolic data analysis methods for domain database searching.....	341
<b>Kamila Migdał-Najman</b> , A proposal of hybrid clustering method based on self-learning networks.....	351
<b>Dorota Rozmus</b> , Comparison of accuracy of spectral clustering and cluster ensembles stability based on bagging idea.....	360
<b>Krzysztof Najman</b> , A dynamic grouping based on self-learning GNG networks.....	369
<b>Małgorzata Misztal</b> , Influence of data imputation methods on the results of object classification using classification trees in the case of small data sets – simulation assessment.....	379
<b>Mariusz Kubus</b> , The application of pre-conditioning of explanatory variable for feature selection.....	386
<b>Barbara Batóg, Jacek Batóg</b> , Application of discriminant analysis to the identification of factors determining the rate of return on the capital market.....	395



<b>Katarzyna Wójcik, Janusz Tuchowski</b> , Comparative analysis of text documents similarity measures based on frequency matrix and based on domain knowledge.....	405
<b>Iwona Staniec</b> , Factor analysis in the identification of areas that determine the improvement of management systems in Polish organizations.....	415
<b>Marek Lubicz, Maciej Zięba, Adam Rzechonek, Konrad Pawełczyk, Jerzy Kołodziej, Jerzy Błaszczyk</b> , Comparative analysis of selected data mining approaches to the classification of medical data with missing values (covariates).....	425
<b>Iwona Foryś</b> , The log-linear analysis using to select the factors determining the attractiveness of the price of flats on the secondary market on the example of local housing market.....	435
<b>Ewa Genge</b> , Trimming approach to the mixtures of normal distributions.....	443
<b>Jerzy Korzeniewski</b> , Efficiency assessment of Ichino method and mean value method of selecting variables in cluster analysis.....	450
<b>Andrzej Dudek</b> , SMS – proposal of new clustering algorithm.....	459
<b>Artur Mikulec</b> , Evaluation methods for the grouping result in cluster analysis.....	468
<b>Małgorzata Machowska-Szewczyk</b> , Fuzzy clustering algorithm for objects described by symbolic or fuzzy variables.....	478
<b>Artur Zaborski</b> , PROFIT analysis and its using in the research of preferences.....	487
<b>Karolina Bartos</b> , Cluster analysis of selected countries due to the structure of their citizens' consumer expenditures – the use of Kohonen networks.....	495
<b>Barbara Batóg, Magdalena Mojsiewicz, Katarzyna Wawrzyniak</b> , Classification of households according to the impulses of concluding the insurance contract by means of qualitative variable models.....	504
<b>Izabela Kurzawa</b> , The application of LA/AIDS model to examine price elasticities of demand of households in the urban-rural relationship.....	512
<b>Aleksandra Luczak, Feliks Wysocki</b> , Linear ordering methods of objects described by a set of metric and ordinal characteristics.....	522
<b>Agnieszka Sompolska-Rzechuła</b> , The comparison of the classical and positional taxonomic analysis of the quality of life differentiation in Zachodniopomorskie voivodeship.....	531
<b>Joanna Banaś, Małgorzata Machowska-Szewczyk</b> , Evaluation of intensity of mailboxes using with the ordered probit model.....	540
<b>Iwona Bąk</b> , Segmentation of pensioners and annuitants households in terms of expenditures on recreation and culture.....	551
<b>Aneta Becker</b> , Application of ANP method to organize Polish voivodships in terms of dynamics of the use of ICT in 2008-2010.....	561
<b>Katarzyna Dębowska</b> , The classification of sectors' financial situation using the methods of multivariate statistical analysis.....	570

---

<b>Anna Domagała</b> , Proposal of a new method for variable selection in DEA models (combined forward stepwise selection method).....	579
<b>Henryk Gierszal, Karina Pawlina, Maria Urbańska</b> , Statistical analysis in demand research of ICT services in mobile networks.....	589
<b>Hanna Gruchociak</b> , Construction of regression estimator for two-level data	600
<b>Tomasz Klimanek, Marcin Szymkowiak</b> , Application of spatial models in indirect estimation of some labor market characteristics .....	609
<b>Jarosław Lira</b> , Forecasting of hog livestock production profitability in Poland .....	618
<b>Christian Lis</b> , The utilization of taxonomic methods in the appraisal of competitiveness of south Baltic ports .....	627
<b>Beata Bieszk-Stolorz, Iwona Markowicz</b> , The application of the multinomial logit model in evaluating employment odds for the unemployed job seekers .....	636
<b>Lucyna Przezbórska-Skobiej, Jarosław Lira</b> , Agritourism space of Poland and its valuation.....	645
<b>Paweł Ulman</b> , Model of expenses distribution and demand functions.....	654
<b>Maria Urbańska, Tadeusz Mizera, Henryk Gierszal</b> , Methods of statistical analysis in research of molluscs .....	663

**Małgorzata Machowska-Szewczyk**

Zachodniopomorski Uniwersytet Technologiczny w Szczecinie

---

## **ALGORYTM KLASYFIKACJI ROZMYTEJ DLA OBIEKTÓW OPISANYCH ZA POMOCĄ ZMIENNYCH SYMBOLICZNYCH ORAZ ROZMYTYCH**

---

**Streszczenie:** Większość opracowanych metod klasyfikacji umożliwia grupowanie obiektów opisanych za pomocą zmiennych ustalonego typu. W praktycznych zastosowaniach wiele obiektów może być charakteryzowanych przez różne typy cech. Celem pracy jest prezentacja rozmytego algorytmu klasyfikacji obiektów, które mogą być opisane jednocześnie za pomocą zmiennych numerycznych, symbolicznych lub rozmytych. Algorytm ten został zaproponowany przez Yanga, Hwanga i Chena, którzy zdefiniowali miarę niepodobieństwa między mieszanymi obiektami oraz zmodyfikowali rozmytą metodę *c*-środków. W artykule przedstawiono także numeryczny przykład zastosowania tej metody do obiektów o cechach mieszanych na podstawie danych rzeczowych.

**Słowa kluczowe:** rozmyta klasyfikacja, dane rozmyte, dane symboliczne, miara niepodobieństwa.

### **1. Wstęp**

Zmienne symboliczne służą do opisu obiektów o złożonej strukturze, w której mogą występować powiązania logiczne, hierarchiczne. Od 1980 r. rozwijane są dość dynamicznie metody klasyfikacji dla danych symbolicznych (np. [Diday 1988]). Na podstawie zdefiniowanych nowych miar niepodobieństwa między obiektami [Gowda, Ravi 1995] El-Sonbaty i Ismail zaproponowali rozmytą metodę *k*-średnich dla danych symbolicznych FCM (*Fuzzy C-Means*), która przypisuje poszczególnym obiektom symbolicznym stopień przynależności do klas [El-Sonbaty, Ismail 1998]. Liczby rozmyte z kolei są stosowane do opisu nieprecyzyjnych informacji, w których źródłem niepewności nie jest przypadkowość, lecz subiektywizm oceny. Tego typu informacje można znaleźć w naturalnym języku, w naukach społecznych, reprezentacji wiedzy itp. Dla obiektów rozmytych zostały również opracowane metody klasyfikacji rozmytej (np. [Hathaway i in. 1996]).

Większość proponowanych metod umożliwia grupowanie obiektów opisanych za pomocą zmiennych ustalonego typu. W praktycznych zastosowaniach wiele obiektów

tów może być charakteryzowanych przez zmienne różnego typu, zarówno symboliczne, jak i rozmyte. Celem pracy jest prezentacja rozmytej metody  $MVFCM^1$  klasyfikacji obiektów, które mogą być opisane jednocześnie za pomocą zmiennych numerycznych, symbolicznych lub rozmytych, zaproponowanej przez Yanga, Hwanga i Chena [2004] jako modyfikacja metody FCM. W artykule przedstawione zostanie zastosowanie tej metody do obiektów o cechach mieszanych na podstawie danych rzeczywistych.

## 2. Miara niepodobieństwa dla obiektów o cechach mieszanego typu

Dowolny obiekt  $O_i \in \{O_1, \dots, O_N\}$  może być utożsamiany z wektorem zaobserwowanych wartości zmiennych  $\mathbf{x}_i = [x_{i1}, \dots, x_{iM}]$ . Dla dowolnych dwóch wektorów  $\mathbf{x}_i = [x_{i1}, \dots, x_{iM}]$  i  $\mathbf{x}_j = [x_{j1}, \dots, x_{jM}]$  niepodobieństwo między obiektami  $O_i, O_j$  może być zdefiniowane następująco:

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sum_{m=1}^M \alpha_m d(x_{im}, x_{jm}),$$

gdzie  $\alpha_1, \dots, \alpha_M$  – wagi odpowiadające poszczególnym zmiennym,  $d(x_{im}, x_{jm})$  – odległość między obiektami  $O_i, O_j$  ze względu na cechę  $X_m$  zdefiniowana zależnie od jej typu.

### 2.1. Niepodobieństwo obiektów ze względu na cechy typu symbolicznego

Gowda i Diday podzielili cechy symboliczne na ilościowe oraz jakościowe i zdefiniowali sposób mierzenia niepodobieństwa między obiektami ze względu na ustaloną zmienną symboliczną, uwzględniając jej rodzaj [Gowda, Diday 1991]. Yang, Hwang i Chen [2004] zaproponowali modyfikację tej miary, która daje rezultaty zgodne z intuicją. Niepodobieństwo między dwiema wartościami cechy  $x_{im}$  i  $x_{jm}$  jest zdefiniowane jako suma niepodobieństw odpowiadających: pozycji  $d_p(x_{im}, x_{jm})$ , rozpiętości  $d_s(x_{im}, x_{jm})$  oraz zawartości  $d_c(x_{im}, x_{jm})$ .

a) Typ ilościowy. Niech  $\underline{x}_{im}$  – dolna granica  $x_{im}$ ,  $\bar{x}_{im}$  – górna granica  $x_{im}$ ,  $\underline{x}_{jm}$  – dolna granica  $x_{jm}$ ,  $\bar{x}_{jm}$  – górna granica  $x_{jm}$ ,  $inters$  – długość przecięcia  $x_{im}$  i  $x_{jm}$ ,  $l_s = |\max(\bar{x}_{im}, \bar{x}_{jm}) - \min(\underline{x}_{im}, \underline{x}_{jm})|$ ,  $U_m$  – różnica między najwyższą

<sup>1</sup> *Mixed-type Variables Fuzzy C-Means.*

a najniższą wartością  $m$ -tej cechy we wszystkich obiektach,  $l_i = |\bar{x}_{im} - \underline{x}_{im}|$ ,  
 $l_j = |\bar{x}_{jm} - \underline{x}_{jm}|$ .

Te trzy miary niepodobieństwa są zdefiniowane następująco [Yang i in. 2004]:

$$d_p(x_{im}, x_{jm}) = \frac{|(\bar{x}_{im} + \underline{x}_{im}) / 2 - (\bar{x}_{jm} + \underline{x}_{jm}) / 2|}{U_m},$$

$$d_s(x_{ik}, x_{jk}) = \frac{|l_i - l_j|}{U_m + l_i + l_j - inters}, \quad d_c(x_{im}, x_{jm}) = \frac{|l_i + l_j - 2 \cdot inters|}{U_m + l_i + l_j - inters}$$

$$i \quad d^2(x_{im}, x_{jm}) = d_p^2(x_{im}, x_{jm}) + d_s^2(x_{im}, x_{jm}) + d_c^2(x_{im}, x_{jm}).$$

b) Typ jakościowy

Niech  $l_i$  – liczba elementów w  $x_{im}$ ,  $l_j$  – liczba elementów w  $x_{jm}$ ,  $l_s$  – liczba elementów sumy  $x_{im}$  i  $x_{jm}$ ,  $inters$  – liczba elementów części wspólnej  $x_{ik}$  i  $x_{jk}$ .

Składowe miary niepodobieństwa są zdefiniowane następująco [Yang i in. 2004]:

$$d_s(x_{im}, x_{jm}) = \frac{|l_i - l_j|}{l_s}, \quad d_c(x_{im}, x_{jm}) = \frac{|l_i + l_j - 2 \cdot inters|}{l_s}.$$

$$\text{Zatem } d^2(x_{im}, x_{jm}) = d_s^2(x_{im}, x_{jm}) + d_c^2(x_{im}, x_{jm}).$$

## 2.2. Niepodobieństwo obiektów ze względu na cechy typu rozmytego

W rzeczywistych zastosowaniach liczb rozmytych najczęściej do reprezentacji informacji nieprecyzyjnej i modelowania niedokładności wykorzystuje się trapezowe liczby rozmyte (TFN). Hathaway, Bezdek i Pedrycz [1996] zaproponowali rozmyte grupowanie FCM (*Fuzzy C-Mean*) dla symetrycznych liczb trapezowych (TFN), stosując podejście parametryczne. Aby dokonać klasyfikacji FCM dla dowolnych liczb rozmytych w reprezentacji typu LR (włączając symetryczne TFN), Yang, Hwang i Chen określili dowolną trapezową liczbę rozmytą  $A$  za pomocą czterech parametrów  $A = (a_1, a_2, a_3, a_4)_T$ , gdzie  $a_1$  oznacza środek,  $a_2$  jest średnicą wewnętrzną,  $a_3$  lewy promień zewnętrzny,  $a_4$  prawy promień zewnętrzny. Stosując tę reprezentację parametryczną, można zapisać zarówno liczby rzeczywiste, przedziały, trójkątne jak i trapezowe liczby rozmyte.

Dla dwóch liczb rozmytych w często stosowanej reprezentacji typu LR Yang i Ko [1996] podali sposób pomiaru odległości między nimi.

Niech  $L(R)$  będzie malejącą funkcją z  $\mathbb{R}^+$  do  $\langle 0,1 \rangle$ , taką że  $L(0) = 1$ ,  $L(x) < 1$  dla wszystkich  $x > 0$ ,  $L(x) > 0$  dla wszystkich  $x < 1$ ,  $L(1) = 0$  (lub  $L(x) > 0$  dla wszystkich  $x$  i  $L(+\infty) = 0$ ). Rozmyta liczba  $X$  z jej funkcją przynależności  $\mu_X$ , określoną następująco [Zimmermann 1991]:

$$\mu_X(x) = \begin{cases} L\left(\frac{m_1 - x}{\alpha}\right) & \text{dla } x \leq m_1 \\ 1 & \text{dla } m_1 \leq x \leq m_2 \\ R\left(\frac{x - m_2}{\beta}\right) & \text{dla } x \geq m_2 \end{cases}$$

jest nazywana trapezową liczbą rozmytą typu  $LR$ . Symbolicznie  $X$  można zapisać jako  $X = (m_1, m_2, \alpha, \beta)_{LR}$ , gdzie  $\alpha > 0$ ,  $\beta > 0$  są nazywane lewym oraz prawym rozrzutem odpowiednio. Dla ustalonych  $A = (m_{1a}, m_{2a}, \alpha_a, \beta_a)_{LR}$  i  $B = (m_{1b}, m_{2b}, \alpha_b, \beta_b)_{LR}$  Yang i Ko zdefiniowali odległość  $d_{LR}(A, B)$  następującym wzorem [Yang, Ko 1996]:

$$d_{LR}^2(A, B) = (m_{1a} - m_{1b})^2 + (m_{2a} - m_{2b})^2 + ((m_{1a} - l\alpha_a) - (m_{1b} - l\alpha_b))^2 + \\ + ((m_{2a} + r\beta_a) - (m_{2b} + r\beta_b))^2$$

gdzie  $l = \int_0^1 L^{-1}(w)dw$  i  $r = \int_0^1 R^{-1}(w)dw$ .

Jeżeli  $L$  i  $R$  są liniowe, to  $l = r = \frac{1}{2}$ . Zatem dla dowolnych dwóch trapezowych liczb rozmytych  $A = (a_1, a_2, a_3, a_4)_T$  i  $B = (b_1, b_2, b_3, b_4)_T$  można wyznaczyć odległość bazującą na odległości Yanga  $d_f(A, B)$ :

$$d_f^2(A, B) = \left(\frac{2a_1 - a_2}{2} - \frac{2b_1 - b_2}{2}\right)^2 + \left(\frac{2a_1 + a_2}{2} - \frac{2b_1 + b_2}{2}\right)^2 + \\ + \left(\left(\frac{2a_1 - a_2}{2} - \frac{1}{2}a_3\right) - \left(\frac{2b_1 - b_2}{2} - \frac{1}{2}b_3\right)\right)^2 + \\ + \left(\left(\frac{2a_1 + a_2}{2} + \frac{1}{2}a_4\right) - \left(\frac{2b_1 + b_2}{2} + \frac{1}{2}b_4\right)\right)^2.$$

### 3. Rozmyty algorytm grupowania

Niech  $A = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$  będzie zbiorem  $N$  wektorów cech w przestrzeni  $\mathbb{R}^M$  oraz  $C$  będzie dodatnią liczbą całkowitą większą niż jeden. Podział zbioru  $A$  na  $C$  grup może być przedstawiony za pomocą wzajemnie rozłącznych zbiorów  $A_1, A_2, \dots, A_C$ , takich że  $A_1 \cup A_2 \cup \dots \cup A_C = A$  lub równoważnie za pomocą funkcji przynależności  $\mu_1, \mu_2, \dots, \mu_C$ , takich że  $\mu_k(\mathbf{x}_i) = 1$ , jeżeli  $\mathbf{x}_i \in A_k$ ,  $\mu_k(\mathbf{x}_i) = 0$ , jeżeli  $\mathbf{x}_i \notin A_k$  dla dowolnego  $k \in \{1, 2, \dots, C\}$ . Jest to tzw. twardy podział  $\{\mu_1, \mu_2, \dots, \mu_C\}$  zbioru  $A$  na  $C$  klas  $A_1, A_2, \dots, A_C$ . Jeżeli wartości  $\mu_k(\mathbf{x}_i)$  mogą pochodzić z przedziału  $\langle 0, 1 \rangle$  oraz  $\sum_{k=1}^C \mu_k(\mathbf{x}_i) = 1$ , to  $\{\mu_1, \mu_2, \dots, \mu_C\}$  jest rozmytym  $C$  podziałem zbioru  $A$ . Algorytm  $C$  środków (FCM) poszukuje takiego podziału rozmytego  $\{\mu_1, \mu_2, \dots, \mu_C\}$ , dla którego osiągnię minimum następująca funkcja:

$$J(\mu, \mathbf{w}) = \sum_{k=1}^C \sum_{i=1}^N \mu_k^r(\mathbf{x}_i) \|\mathbf{x}_i - \mathbf{w}_k\|^2, \quad (1)$$

gdzie:  $r$  jest ustaloną liczbą większą niż jeden, przedstawiającą stopień rozmycia,  $\{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_C\}$  jest zbiorem środków klas,  $\|\cdot\|$  normą euklidesową w przestrzeni  $\mathbb{R}^M$ .

W przypadku obiektów symbolicznych nie da się bezpośrednio przeprowadzić procedury grupowania FCM. Aby pokonać problemy związane z wyznaczaniem średniej ważonej, El-Sonbaty i Ismail [1998] zaproponowali nowy sposób reprezentacji środków klas. Środek  $k$ -tej klasy  $\mathbf{w}_k$  powstaje jako wektor wartości cech  $\mathbf{w}_k = [w_{1/k}, w_{2/k}, \dots, w_{M/k}]$ , z których każda wartość może być symboliczna, liczbowa lub rozmyta. W przypadku cechy symbolicznej wartość może składać się z kilku zdarzeń. Niech  $w_{mp/k}$  będzie  $p$ -tym zdarzeniem  $m$ -tej cechy w  $k$ -tej klasie i niech  $e_{mp/k}$  będzie stopniem przynależności  $p$ -tego zdarzenia do  $m$ -tej cechy w  $k$ -tej klasie. Zatem wartość  $m$ -tej cechy w środku  $k$ -tej klasy  $w_{m/k}$  może być przedstawiona jako:

$$w_{m/k} = [(w_{m1/k}, e_{m1/k}), \dots, (w_{mP/k}, e_{mP/k})]. \quad (2)$$

Wówczas:  $0 \leq e_{mp/k} \leq 1$  i  $\sum_p e_{mp/k} = 1$ ,  $\bigcap_p w_{mp/k} = \emptyset$  i  $\bigcup_p w_{mp/k} = \bigcup_i x_{im}$ ,

gdzie:  $e_{mp/k} = 0$ , jeżeli nie zaszło zdarzenie  $w_{mp/k}$  dla  $m$ -tej cechy środka  $k$ -tej klasy, i  $e_{mp/k} = 1$ , jeżeli żadne inne zdarzenie poza  $w_{mp/k}$  nie zaszło dla  $m$ -tej cechy środka

$k$ -tej klasy  $w_{m/k}$ . Funkcja przynależności  $e_{m'p/k}$  jest bardzo ważną funkcją wskaźnikową, która umożliwi zastosowanie algorytmu FCM do symbolicznych danych.

W pracy rozważane jest zastosowanie algorytmu FCM nie tylko do danych symbolicznych, ale również do cech rozmytych, przy czym przyjęto założenie, że każda rozmyta cecha przyjmuje wartości w postaci trapezowej liczby rozmytej. Niech  $w_{m/k} = (w_{m1/k}, w_{m2/k}, w_{m3/k}, w_{m4/k})_T$  będzie wartością  $m$ -tej rozmytej cechy środka  $k$ -tej klasy.

Niech  $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$  będzie zbiorem obiektów reprezentowanych przez cechy różnych typów. Funkcja kryterium FCM jest zdefiniowana jako:

$$J(\mu, \mathbf{e}, \mathbf{w}) = \sum_{k=1}^C \sum_{i=1}^N [\mu_k(\mathbf{x}_i)]^r \cdot d^2(\mathbf{x}_i, \mathbf{w}_k), \quad (3)$$

gdzie

$$d^2(\mathbf{x}_i, \mathbf{w}_k) = \sum_{m' \text{ symbol}} \left( \sum_p d^2(x_{im'}, w_{m'p/k}) \cdot e_{m'p/k} \right) + \sum_{m \text{ rozmyty}} d_f^2(x_{im}, w_{m/k}). \quad (4)$$

Wykorzystując metodę mnożników Lagrange'a poszukiwania rozwiązań optymalnych, można znaleźć wartość funkcji przynależności:

$$\mu_k(\mathbf{x}_i) = \left( \sum_{q=1}^C \frac{(d^2(\mathbf{x}_i, \mathbf{w}_k))^{1/(r-1)}}{(d^2(\mathbf{x}_i, \mathbf{w}_q))^{1/(r-1)}} \right)^{-1}, \quad k \in \{1, \dots, C\}, i \in \{1, \dots, N\}, \quad (5)$$

gdzie:  $d^2(\mathbf{x}_i, \mathbf{w}_k)$  jest zdefiniowane przez równość (4), dla której  $d^2(x_{im'}, w_{m'p/k})$  i  $d_f^2(x_{im}, w_{m/k})$  są niepodobieństwami między symbolicznymi oraz rozmytymi danymi odpowiednio zaproponowanymi w punkcie 2.

(a) Niech  $m'$  będzie dowolnym numerem cechy symbolicznej. Stosując miarę odległości zaproponowaną w pracy El-Sonbaty i Ismaila [1998] oraz wyznaczając pochodną funkcji  $J$  względem  $\mathbf{e}$  i przyrównując ją do zera, można otrzymać wzór:

$$e_{m'p/k} = \frac{\sum_{i=1}^N \mu_k^r(\mathbf{x}_i) \cdot \theta}{\sum_{i=1}^N \mu_k^r(\mathbf{x}_i)}, \quad (6)$$

gdzie  $\mu_k(\mathbf{x}_i)$  jest funkcją przynależności obiektu  $\mathbf{x}_i$  do klasy  $A_k$ ,  $\theta \in \{0, 1\}$  i  $\theta = 1$ , jeżeli dla  $m$ -tej cechy  $i$ -tego obiektu  $\mathbf{x}_i$  zaszło  $p$ -te zdarzenie, w przeciwnym przypadku  $\theta = 0$ .



(b) Dla tych  $m$ , które oznaczają numery cech rozmytych, wyznacza się pochodne cząstkowe funkcji  $J$  względem  $w_{km1}, w_{km2}, w_{km3}, w_{km4}$  i przyrównuje je do zera. Na tej podstawie można wyznaczyć parametry poszczególnych liczb rozmytych, które są wartościami  $m$ -tej cechy rozmytej środka  $k$ -tej klasy w iteracji ( $t$ ):

$$w_{m1/k}^{(t)} = \frac{\sum_{i=1}^N [\mu_k^{(t-1)}(\mathbf{x}_i)]^r (8x_{im1} - x_{im3} + x_{im4} + w_{m3/k}^{(t-1)} - w_{m4/k}^{(t-1)})}{8 \sum_{i=1}^N [\mu_k^{(t-1)}(\mathbf{x}_i)]^r}, \quad (7)$$

$$w_{m2/k}^{(t)} = \frac{\sum_{i=1}^N [\mu_k^{(t-1)}(\mathbf{x}_i)]^r (4x_{im2} + x_{im3} + x_{im4} - w_{m3/k}^{(t-1)} - w_{m4/k}^{(t-1)})}{4 \sum_{i=1}^N [\mu_k^{(t-1)}(\mathbf{x}_i)]^r}, \quad (8)$$

$$w_{m3/k}^{(t)} = \frac{\sum_{i=1}^N [\mu_k^{(t-1)}(\mathbf{x}_i)]^r (-2x_{im1} + x_{im2} + x_{im3} + 2w_{m1/k}^{(t)} - w_{m2/k}^{(t)})}{\sum_{i=1}^N [\mu_k^{(t-1)}(\mathbf{x}_i)]^r}, \quad (9)$$

$$w_{m4/k}^{(t)} = \frac{\sum_{i=1}^N [\mu_k^{(t-1)}(\mathbf{x}_i)]^r (2x_{im1} + x_{im2} + x_{im4} - 2w_{m1/k}^{(t)} - w_{m2/k}^{(t)})}{\sum_{i=1}^N [\mu_k^{(t-1)}(\mathbf{x}_i)]^r}. \quad (10)$$

Interaktywny algorytm grupowania  $C$ -środków ( $MVFCM$ ) dla danych mieszane-go typu:

*Krok 1:* Niech będzie dane  $\varepsilon > 0$ . Ustalić stopień rozmycia  $r$  i liczbę klas  $C$  oraz określić wstępny rozmyty podział na  $C$  klas:  $\mu^{(0)} = \{\mu_1^{(0)}, \dots, \mu_C^{(0)}\}$ . Przyjąć  $t = 0$ .

*Krok 2:* Dla cechy symbolicznej  $m'$  obliczyć środek  $k$ -tej klasy  $w_{km'}^{(t)} = [(w_{m'1/k}^{(t)}, e_{m'1/k}^{(t)}), \dots, (w_{m'p/k}^{(t)}, e_{m'p/k}^{(t)})]$ , stosując równości (2), (6). Dla cechy rozmytej  $m$  obliczyć środek  $k$ -tej klasy  $w_{km}^{(t)} = (w_{m1}^{(t)}, w_{m2}^{(t)}, w_{m3}^{(t)}, w_{m4}^{(t)})$ , stosując równości (7)-(10).

*Krok 3:* Zaktualizować  $\mu^{(t+1)}$ , stosując równości (4) i (5).

*Krok 4:* Porównać  $\mu^{(t+1)}$  z  $\mu^{(t)}$  za pomocą normy macierzowej. Jeżeli  $\|\mu^{(t+1)} - \mu^{(t)}\| \geq \varepsilon$ , to przejść do kroku 2, przyjmując  $t = t + 1$ , w przeciwnym przypadku koniec obliczeń.

## 4. Przykład empiryczny

Zbiór obiektów składa się z 10 marek samochodów z czterech firm: Skoda, Fiat, Citroen oraz Renault. Każda marka jest charakteryzowana przez sześć cech: firmę, pojemność silnika, cenę, dostępny kolor, komfort, zużycie paliwa. Cechy: firma, pojemność silnika, kolor przyjmują wartości symboliczne, cena jest wartością rzeczywistą, komfort i bezpieczeństwo zaś są danymi rozmytymi. Zbiór danych przedstawiono w tab. 1.

**Tabela 1.** Zbiór danych o samochodach

Lp.	Marka	Firma	Pojemność silnika	Cena	Kolor*	Komfort	Zużycie paliwa
1	Fabia	Skoda	1,4	43,35	B, Br, C, Cz, Cz1, F, M, M1, N, N1, P, S, S1, S2, Z, Ż	[6.6;0.2;1.5;0.7]	[5.9;1;0.7;1.6]
2	Oktavia	Skoda	1,4	54,5	B, Br, C, Cz, M, M1, N, N1, P, S, S1, S2,	[7.35;0;0.65;0.65]	[6.4;1;0.8;1.6]
3	Superb	Skoda	1,8	88,3	B, Br, C, Cz, M, M1, N, N1, P, S, S1, S2,	[8.8;1;0;0]	[8.1;1.2;0.9;1.9]
4	Panda	Fiat	1,2	26,99	B, C, , Cz, F, M1, N, N1, Ps, S2, Z1, Ż	[6.5;1;0.1;0.4]	[4.9;0;0.9;1.5]
5	Bravo	Fiat	1,6	66,99	B, B1, C, Cz, N, N1, S1, S2	[7.5;0;0.4;0.7]	[4.9;0;0.8;1.4]
6	C3 Picasso	Citroen	1,4	56,6	B1,C, Cz, N, Ps,S2, Z	[7;1;0.2;0.2]	[6.6;1;1.1;1.6]
7	C1	Citroen	1	43,1	B, C, Cz, N, P, Ps,S1, S2	[6.8;0;1;1]	[4.5;0;0.6;1]
8	C5	Citroen	1,6	101,7	B, B1, Br, Cz1, Gr, M, S, S1, S2	[9;1;0;0]	[7.1;1;1.1;1.2]
9	Thalia	Renault	1,2	29,9	B, C, Cz1,N1, Ps, S, S1, S2	[6.8;0;0.7;0.8]	[5.9;0;1.1;1.7]
10	Megane	Renault	1,6	54,45	B, Cz, N, S, S1, S2,	[7.7;1;0;0.1]	[6.8;1;0.8;1.8]

\* W cesze kolor przyjęto następującą notację: B – biały, B1 – biały perłowy, Br – bordowy, C – czerwony, Cz – czarny, Cz1 – czarny perła, F – fiolet, Gr – grafitowy, M – morski, M1 – morski jasny, N – niebieski, N1 – niebieski jasny, P – pistacjowy, Ps – piaskowy, S – srebrny, S1 – szary jasny, S2 – szary, Z – zielony, Z1 – złoty, Ż – żółty.

Źródło: opracowanie własne na podstawie [www.skoda-auto.pl](http://www.skoda-auto.pl); [www.fiat.pl](http://www.fiat.pl); [www.renault.pl](http://www.renault.pl); [www.citroen.pl](http://www.citroen.pl); [opinie.auto.com.pl](http://opinie.auto.com.pl).

**Tabela 2.** Wartości funkcji przynależności do dwóch klas otrzymane za pomocą metody MVFCM

Lp.	1	2	3	4	5	6	7	8	9	10
$\mu_{1j}$	0,9941	0,9354	0,0084	0,9257	0,5691	0,8977	0,9922	0,0301	0,9417	0,9303
$\mu_{2j}$	0,0059	0,0646	0,9916	0,0743	0,4309	0,1023	0,0078	0,9699	0,0583	0,0697

Źródło: opracowanie własne.

W tabeli 2 przedstawiono wyniki funkcji przynależności dla 10 obiektów, otrzymane w rezultacie zastosowania algorytmu MVFCM dla zmiennych mieszanych, przy założeniu, że  $r = 2$ ,  $C = 2$  i  $\varepsilon = 0,0001$ . Analizując wyniki w tab. 4, można wyodrębnić dwie ostre klasy:  $C_1 = \{\text{Fabia, Oktawia, Panda, Bravo, C3 Picasso, C1, Thalia, Megane}\}$ ,  $C_2 = \{\text{Superb, C5}\}$ . Fiat Bravo jest obiektem mieszkańcem, którego przynależność do obu klas jest dość duża.

## 5. Podsumowanie

Większość algorytmów umożliwia grupowanie obiektów, które są reprezentowane tylko przez zmienne tego samego typu. Zaproponowany algorytm MVFCM pozwala zastosować różne typy cech, takie jak: numeryczne, symboliczne oraz rozmyte. Wyniki eksperymentalne przeprowadzone przez autorów wykazały, że algorytm MVFCM jest efektywny w przypadku zmiennych mieszanego typu. W realnych sytuacjach często mamy do czynienia z cechami mieszanymi: numerycznymi, symbolicznymi lub rozmytymi. W takim przypadku zaproponowany algorytm MVFCM może stanowić użyteczne i efektywne narzędzie analizy danych.

Wadą proponowanego algorytmu, podobnie jak FCM, jest wrażliwość na wstępny podział. Jednak jeżeli klasy danych są dobrze odseparowane od siebie, to algorytmy te nie są wrażliwe na rozpoczęcie podziału. Gdy natomiast stopień nakładania się poszczególnych klas rośnie, to rośnie także wrażliwość tych algorytmów na podział początkowy. Kolejną wadą jest brak możliwości zastosowania zmiennych symbolicznych z wagami. Jako ograniczenie tej metody wymienić można także zawężenie wartości rozmytych jedynie do postaci liczb trapezowych. Jednak symboliczne metody klasyfikacji oraz rozmytość w zagadnieniach klasyfikacyjnych stale są przedmiotem badań, należy zatem spodziewać się kolejnych propozycji, które pozabawione będą tych wad.

## Literatura

- Diday E., *The Symbolic Approach in Clustering*, [w:] *Classification and Related Methods of Data Analysis*, H.H. Bock (red.), North-Holland, Amsterdam 1988.
- El-Sonbaty Y., Ismail M.A., *Fuzzy clustering for symbolic data*, „IEEE Trans. Fuzzy Systems 6 (2)” 1998.
- Gowda K.C., Diday E., *Symbolic clustering using a new dissimilarity measure*, „Pattern Recognition” 1991, no 24 (6).

- Gowda K.C., Ravi T.V., *Divisive clustering of symbolic objects using the concepts of both similarity and dissimilarity*, „Pattern Recognition” 1995, no 28 .
- Hathaway R.J., Bezdek J.C., Pedrycz W., *A parametric model for fusing heterogeneous fuzzy data*, „IEEE Trans. Fuzzy Systems” 1996, no 4 (3) .
- Yang M.-S., Hwang P.-Y., Chen D.-H., *Fuzzy clustering algorithms for mixed feature variables*, „Fuzzy Sets and Systems” 2004, no 141.
- Yang M.-S., Ko C.H., *On a class of fuzzy c-numbers clustering procedures for fuzzy data*, „Fuzzy Sets and Systems” 1996, no 84.
- Zimmermann H.J., *Fuzzy Set Theory and Its Applications*, Kluwer, Dordrecht 1991.

## **FUZZY CLUSTERING ALGORITHM FOR OBJECTS DESCRIBED BY SYMBOLIC OR FUZZY VARIABLES**

**Summary:** The majority of discussed classification methods allow to cluster objects described by variables of the same type. In real applications many objects can be characterized by mixed feature types. The aim of this work is to present fuzzy clustering algorithm for objects, which can be described at the same time by numerical, symbolic and fuzzy data. This algorithm was presented by Yang, Hwang and Chen, who defined dissimilarity measure between objects represented by mixed features and they modified fuzzy c-means algorithm. This article also includes a numerical example based on real data, which illustrates the application of this method for objects with mixed features.

**Keywords:** fuzzy classification, fuzzy data, symbolic data, dissimilarity measure.