

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

Kopiowanie i powielanie w jakiegokolwiek formie

wymaga pisemnej zgody Wydawcy

© Copyright by Uniwersytet Ekonomiczny we Wrocławiu

Wrocław 2013

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 Pocięcha: 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 <code>pllord</code>	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 Faldziński, 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 <i>unfolding</i> 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 Pawelczyk, 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 Pelka: Rozmyta klasyfikacja spektralna <i>c</i> -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 poziomego 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 Pocięcha: 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 Faldziński, 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 Pawelczyk, 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

Małgorzata Machowska-Szewczyk

Zachodniopomorski Uniwersytet Technologiczny w Szczecinie

KLASYFIKACJA OBIEKTÓW REPREZENTOWANYCH PRZEZ RÓŻNEGO RODZAJU CECHY SYMBOLICZNE

Streszczenie: Większość opracowanych metod klasyfikacji symbolicznej umożliwia grupowanie obiektów opisanych za pomocą cech symbolicznych tego samego typu. W praktycznych zastosowaniach wiele obiektów może być charakteryzowanych przez cechy symboliczne mieszane, czyli o wartościach różnego typu: w postaci zarówno numerycznej, przedziałów liczbowych, listy wartości, jak i list wartości z wagami. Celem pracy jest prezentacja metod klasyfikacji obiektów symbolicznych o cechach mieszanego typu zaproponowanych w pracy [de Carvalho, de Souza 2010] oraz przedstawienie propozycji uogólnienia tych algorytmów do klasyfikacji rozmytej. Główna idea polega na transformacji wartości symbolicznych różnego typu na wartości symboliczne w postaci histogramu.

Słowa kluczowe: analiza danych symbolicznych, dane symboliczne o cechach różnych typów, wartości symboliczne w postaci histogramu, rozmyta klasyfikacja.

1. Wstęp

Większość opracowanych metod klasyfikacji symbolicznej umożliwia grupowanie obiektów opisanych za pomocą cech symbolicznych tego samego typu. W praktycznych zastosowaniach wiele obiektów może być charakteryzowanych przez cechy symboliczne mieszane, czyli o wartościach różnego typu: w postaci zarówno numerycznej, przedziałów liczbowych, listy wartości, jak i listy wartości z wagami.

Celem pracy jest prezentacja zaproponowanych w pracy de Carvalho i de Souza [2010] metod klasyfikacji obiektów symbolicznych o cechach mieszanego typu oraz przedstawienie propozycji uogólnienia tych algorytmów do klasyfikacji rozmytej. Metody te opierają się na metodologii grupowania iteracyjnego z adaptacją odległości euklidesowej. Odległości są zmieniane w każdej iteracji algorytmu i mogą być albo takie same dla wszystkich klas, albo niejednakowe dla poszczególnych grup. W pierwszym kroku dokonuje się transformacji wartości symbolicznych różnego typu na wartości symboliczne w postaci histogramu. Zaproponowana przez autorkę modyfikacja umożliwia przeprowadzenie klasyfikacji zarówno w sensie klasycznym (wówczas jest realizowana metoda klasyfikacji de Carvalho, de Souza), jak i w sensie rozmytym. Klasyfikacja rozmyta jest bardzo użyteczna w sytuacji trudno separo-

wanych klas, obiekty tzw. mieszańce mogą być klasyfikowane do klas z pewnym stopniem przynależności. Klasyfikacja klasyczna wymusza przypisanie obiektu tylko do jednej klasy, przez co nie są rozpoznawane obiekty, których podobieństwo do kilku klas jednocześnie jest dość duże, a jakość otrzymanej klasyfikacji jest wówczas niska. Proponowany algorytm wnosi zatem dodatkową możliwość do analizy danych symbolicznych o cechach mieszanego typu.

2. Wstępna homogenizacja danych

Każdy obiekt i ze zbioru $\Omega = \{1, \dots, n\}$, opisywany przez wartości p zmiennych symbolicznych $\{X_1, \dots, X_p\}$, jest utożsamiany z wektorem danych symbolicznych o mieszanych typach cech $\mathbf{x}_i = (x_i^1, x_i^2, \dots, x_i^p)$, $i = 1, \dots, n$. To oznacza, że zmienna symboliczna X_j może przyjąć dla danej jednostki i wartość x_i^j w postaci [Bock, Diday 2000]:

- podzbioru, czyli $X_j(i) = x_i^j \subset A_j$, gdzie $A_j = \{t_1^j, t_2^j, \dots, t_{H_j}^j\}$ jest zbiorem kategorii;
- uporządkowanej listy kategorii, czyli x_i^j jest podlistą uporządkowanej listy kategorii $A_j = [t_1^j, t_2^j, \dots, t_{H_j}^j]$;
- przedziału: $X_j(i) = x_i^j = [a_i^j; b_i^j] \subset [a; b]$, gdzie $[a; b] \in \mathcal{G}$, \mathcal{G} jest zbiorem przedziałów domkniętych w zbiorze liczb rzeczywistych;
- histogramu: $X_j(i) = x_i^j = (S^j(i), \mathbf{q}^j(i))$, gdzie $\mathbf{q}^j(i) = (q_{i1}^j, q_{i2}^j, \dots, q_{iH_j}^j)$ jest wektorem wag definiowanym w $S^j(i)$, takim że waga q_{im}^j odpowiada kategorii m należącej do $S^j(i)$, a $S^j(i)$ jest nośnikiem miary $\mathbf{q}^j(i)$.

Standardowy algorytm klasyfikacji [Diday, Simon 1976] ma na celu znalezienie podziału $P = (C_1, C_2, \dots, C_K)$ zbioru Ω na ustaloną liczbę K klas i odpowiadających im wzorców $\mathbf{G} = (\mathbf{g}_1, \dots, \mathbf{g}_K)$ przez lokalną minimalizację kryterium W , które ocenia dopasowanie między klasami i ich odpowiednimi reprezentantami.

Aby pokonać trudność, jaką jest reprezentacja obiektów za pomocą uporządkowanych lub nieuporządkowanych danych symbolicznych różnego typu, dokonuje się wstępnego przetwarzania, którego celem jest uzyskanie odpowiedniej homogenizacji danych symbolicznych. Polega ono na transformacji danych o mieszanych typach cech na symboliczne dane o wartościach w postaci histogramu.

Jeżeli X_j jest zmienną o wartościach podzbioru, to jej transformacja w symboliczną zmienną \tilde{X}_j o wartościach w postaci histogramu osiągnąca jest w następujący

sposób: $\tilde{X}_j(i) = \tilde{x}_i^j = (A_j(i), \mathbf{q}^j(i))$, gdzie $A_j = \{t_1^j, t_2^j, \dots, t_{H_j}^j\}$ jest dziedziną zmiennej X_j i nośnikiem wektora wag $\mathbf{q}^j(i) = (q_1^j(i), q_2^j(i), \dots, q_{H_j}^j(i))$. Wagi $q_h^j(i)$ ($h=1, \dots, H_j$) kategorii $t_h^j \in A_j$ są zdefiniowane jako [de Carvalho 1995]:

$$q_h^j(i) = \begin{cases} \frac{1}{c(x_i^j)} & \text{jeżeli } t_h^j \in x_i^j \\ 0 & \text{jeżeli } t_h^j \notin x_i^j \end{cases}, \quad (1)$$

gdzie $c(A)$ jest mocą skończonego zbioru kategorii A .

Jeżeli X_j jest zmienną o wartościach w postaci uporządkowanej listy, to przekształca się ją w symboliczną zmienną \tilde{X}_j o wartościach histogramu następująco: $\tilde{X}_j(i) = \tilde{x}_i^j = (A_j, \mathbf{Q}^j(i))$, gdzie $A_j = [t_1^j, t_2^j, \dots, t_{H_j}^j]$ jest nośnikiem wektora skumulowanych wag $\mathbf{Q}^j(i) = (Q_1^j(i), Q_2^j(i), \dots, Q_{H_j}^j(i))$. Skumulowane wagi $Q_h^j(i)$ ($h=1, \dots, H_j$) kategorii t_h^j z listy A_j są zdefiniowane jako [de Carvalho 1995]:

$$Q_h^j(i) = \sum_{r=1}^h q_r^j(i) \quad \text{gdzie } q_r^j(i) = \begin{cases} \frac{1}{l(x_i^j)}, & \text{jeżeli } t_r^j \text{ jest na liście } x_i^j \\ 0 & \text{, w przeciwnym przypadku} \end{cases}, \quad (2)$$

$l(A)$ zaś jest długością uporządkowanej listy kategorii A .

W przypadku zmiennej X_j o wartościach w postaci przedziałów jest ona transformowana w symboliczną zmienną \tilde{X}_j o wartościach w postaci histogramu następująco: $\tilde{X}_j(i) = \tilde{x}_i^j = (\tilde{A}_j, \mathbf{Q}^j(i))$, gdzie $\tilde{A}_j = \{I_1^j, I_2^j, \dots, I_{H_j}^j\}$ jest listą elementarnych przedziałów, stanowiących nośnik wektora skumulowanych wag $\mathbf{Q}^j(i) = (Q_1^j(i), Q_2^j(i), \dots, Q_{H_j}^j(i))$. Skumulowana waga $Q_h^j(i)$ ($h=1, \dots, H_j$) elementarnego przedziału I_h^j jest zdefiniowana jako [de Carvalho 1995]:

$$Q_h^j(i) = \sum_{r=1}^h q_r^j(i) \quad \text{gdzie } q_r^j(i) = \frac{l(I_r^j \cap x_i^j)}{l(x_i^j)}, \quad (3)$$

$l(I)$ zaś jest długością domkniętego przedziału I .

Można pokazać, że: $0 \leq q_h^j(i) \leq 1$ ($h=1, \dots, H_j$) i $\sum_{h=1}^{H_j} q_h^j(i) = 1$. Ponadto także $q_1^j(i) = Q_1^j(i)$ i $q_h^j(i) = Q_h^j(i) - Q_{h-1}^j(i)$ ($h=2, \dots, H_j$).

Granice elementarnych przedziałów I_h^j ($h=1, \dots, H_j$) są uzyskiwane z uporządkowanych granic $n+1$ przedziałów $\{x_1^j, x_2^j, \dots, x_n^j, [a; b]\}$ i liczba elementarnych przedziałów H_j wynosi co najwyżej $2n$. Przedziały elementarne mają następujące własności [de Carvalho 1995]:

- $\sum_{h=1}^{H_j} I_h^j = [a; b]$,
- $I_h^j \cap I_{h'}^j = \emptyset$ jeżeli $h \neq h'$,
- $\forall h \exists i \in \Omega$ takie, że $I_h^j \cap x_i^j \neq \emptyset$,
- $\forall i \exists S_i^j \subset \{1, \dots, H_j\} : \bigcup_{h \in S_i^j} I_h^j = x_i^j$.

Po etapie wstępnego przetwarzania każdy obiekt i ($i=1, \dots, n$) jest reprezentowany przez wektor danych symbolicznych o wartościach w postaci histogramu $\tilde{x}_i = (\tilde{x}_i^1, \dots, \tilde{x}_i^p)$, przy czym $\tilde{x}_i^j = (D_j, \mathbf{u}^j(i))$, gdzie D_j (dziedzina zmiennej \tilde{X}_j) w zależności od typu pierwotnej zmiennej jest zbiorem kategorii, uporządkowaną listą kategorii lub listą elementarnych przedziałów, $\mathbf{u}^j(i) = (u_1^j(i), \dots, u_{H_j}^j(i))$ jest wektorem wag lub skumulowanych wag. Wzorzec klasy C_k ($k=1, \dots, K$) jest także reprezentowany przez wektor danych symbolicznych o wartościach w postaci histogramu $\mathbf{g}_k = (g_k^1, \dots, g_k^p)$, $g_k^j = (D_j, \mathbf{v}^j(k))$ ($j=1, \dots, p$), gdzie $\mathbf{v}^j(k) = (v_1^j(k), \dots, v_{H_j}^j(k))$ jest wektorem wag lub skumulowanych wag, D_j jest zbiorem kategorii, listą kategorii lub listą elementarnych przedziałów. Warto zauważyć, że dla każdej zmiennej ($j=1, \dots, p$) nośnik jest taki sam dla wszystkich jednostek i wzorców.

Zgodnie z ogólnym schematem algorytm klasyfikacji iteracyjnej [Diday, Simon 1976] poszukuje podziału $P^* = (C_1^*, C_2^*, \dots, C_K^*)$ zbioru Ω na ustaloną liczbę K klas, odpowiadającego K wzorcom $\mathbf{G}^* = (\mathbf{g}_1^*, \dots, \mathbf{g}_K^*)$, reprezentującym klasy w P^* , oraz K wektorów wag $\mathbf{D}^* = (\boldsymbol{\lambda}_1^*, \dots, \boldsymbol{\lambda}_K^*)$ parametryzujących kwadraty adaptacyjnych odległości euklidesowych, dla których minimalna będzie wartość funkcji kryterialnej:

$$W(\mathbf{G}, \mathbf{D}, P) = \sum_{k=1}^K \sum_{i \in C_k} d(\tilde{x}_i, \mathbf{g}_k / \boldsymbol{\lambda}_k). \quad (4)$$

We wzorze (4) rozważa się:

a) kwadraty adaptacyjnych odległości euklidesowych parametryzowane przez jednaki wektor wag $\lambda_k = \lambda (k=1, \dots, K)$, gdzie $\lambda = (\lambda^1, \dots, \lambda^p)$ zmienia się w każdej iteracji, ale jest taki sam dla wszystkich klas:

$$d(\tilde{\mathbf{x}}_i, \mathbf{g}_k / \lambda) = \sum_{j=1}^p \lambda^j \varphi^2(\mathbf{u}^j(i), \mathbf{v}^j(k)) = \sum_{j=1}^p \lambda^j \sum_{h=1}^{H_j} (u_h^j(i) - v_h^j(k))^2, \quad (5)$$

b) kwadraty adaptacyjnych odległości euklidesowych parametryzowane przez wektory wag $\lambda_k = (\lambda_k^1, \dots, \lambda_k^p) (k=1, \dots, K)$, które zmieniają się w każdej iteracji i są niejednakowe dla poszczególnych klas:

$$d(\tilde{\mathbf{x}}_i, \mathbf{g}_k / \lambda_k) = \sum_{j=1}^p \lambda_k^j \varphi^2(\mathbf{u}^j(i), \mathbf{v}^j(k)) = \sum_{j=1}^p \lambda_k^j \sum_{h=1}^{H_j} (u_h^j(i) - v_h^j(k))^2. \quad (6)$$

W pierwszym przypadku wektor wag jest szacowany globalnie dla wszystkich klas od razu, podczas gdy w drugim wagi są estymowane lokalnie dla każdej klasy.

3. Algorytm klasyfikacji dla danych symbolicznych o różnych typach zmiennych

Algorytm klasyfikacji, rozpoczynając od rozwiązania początkowego $\mathbf{v}_0 = (\mathbf{G}^0, \mathbf{D}^0, \mathbf{P}^0)$, stosuje na przemian trzy kroki aż do uzyskania zbieżności, tzn. gdy kryterium W osiąga stałą wartość, reprezentującą lokalne minimum.

2.1 Krok 1: Najlepszy wzorzec

Twierdzenie 1 [de Carvalho, de Souza 2010]. Jeżeli $P = (C_1, \dots, C_K)$ oraz $\mathbf{D} = (\lambda_1, \dots, \lambda_K)$ są ustalone, to niezależnie od funkcji odległości (równania (5) i (6)) wektor wzorców $\mathbf{G} = (\mathbf{g}_1, \dots, \mathbf{g}_K)$, gdzie $\mathbf{g}_k = (g_k^1, \dots, g_k^p) (k=1, \dots, K)$ z $g_k^j = (D_j, \mathbf{v}^j(k)) (j=1, \dots, p)$, który minimalizuje kryterium klasyfikacji W , jest taki, że elementy $v_h^j(k) (h=1, \dots, H_j)$ wektora wag $\mathbf{v}^j(k) = (v_1^j(k), \dots, v_{H_j}^j(k))$ są obliczane zgodnie ze wzorem:

$$v_h^j(k) = \frac{1}{n_k} \sum_{i \in C_k} u_h^j(i) \quad (7)$$

gdzie n_k jest liczebnością klasy C_k .

2.2 Krok 2: Najlepsze wagi

Twierdzenie 2 [de Carvalho, de Souza 2010]. Jeżeli $P = (C_1, \dots, C_K)$ oraz $\mathbf{G} = (\mathbf{g}_1, \dots, \mathbf{g}_K)$ są ustalone, to K wektorów wag $\mathbf{D} = (\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_K)$, które minimalizują kryterium W , są obliczane w zależności od zastosowanej funkcji odległości.

a) Jeżeli funkcja odległości jest dana przez równość (5), to wektory wag $\boldsymbol{\lambda}_k = \boldsymbol{\lambda}(k = 1, \dots, K)$, gdzie $\boldsymbol{\lambda} = (\lambda^1, \dots, \lambda^p)$, które minimalizują kryterium klasyfikacji W , przy czym $\lambda^j > 0$ i $\prod_{j=1}^p \lambda^j = \eta$, gdzie $\eta \in \mathbf{R}$ jest stałe, mają swoje wagi λ^j obliczane zgodnie ze wzorem:

$$\lambda^j = \frac{\left\{ \eta \prod_{l=1}^p \left(\sum_{k=1}^K \left[\sum_{i \in C_k} \left(\sum_{h=1}^{H_l} (u_h^l(i) - v_h^l(k))^2 \right) \right] \right) \right\}^{\frac{1}{p}}}{\sum_{k=1}^K \left[\sum_{i \in C_k} \left(\sum_{h=1}^{H_j} (u_h^j(i) - v_h^j(k))^2 \right) \right]}. \quad (8)$$

b) Jeżeli funkcja odległości jest dana przez równość (6), to wagi $\boldsymbol{\lambda}_k = (\lambda_k^1, \dots, \lambda_k^p)$ ($k = 1, \dots, K$), które minimalizują kryterium klasyfikacji W , przy czym $\lambda_k^j > 0$ i $\prod_{j=1}^p \lambda_k^j = \chi$, gdzie $\chi \in \mathbf{R}$ jest stałe, mają swoje wagi λ_k^j obliczane następująco:

$$\lambda_k^j = \frac{\left\{ \chi \prod_{l=1}^p \left(\sum_{i \in C_k} \left(\sum_{h=1}^{H_l} (u_h^l(i) - v_h^l(k))^2 \right) \right) \right\}^{\frac{1}{p}}}{\sum_{i \in C_k} \left(\sum_{h=1}^{H_j} (u_h^j(i) - v_h^j(k))^2 \right)}. \quad (9)$$

2.3 Krok 3: Najlepszy podział

Twierdzenie 3 [de Carvalho, de Souza 2010]. Jeżeli $\mathbf{G} = (\mathbf{g}_1, \dots, \mathbf{g}_K)$ i $\mathbf{D} = (\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_K)$ są ustalone, to podział $P = (C_1, \dots, C_K)$, który minimalizuje kryterium W , jest aktualizowany zgodnie z następującą regułą alokacji:

$$C_k = \left\{ i \in \Omega : d(\tilde{\mathbf{x}}_i, \mathbf{g}_k / \boldsymbol{\lambda}_k) < d(\tilde{\mathbf{x}}_i, \mathbf{g}_m / \boldsymbol{\lambda}_m) \text{ lub } d_k(\tilde{\mathbf{x}}_i, \mathbf{g}_k / \boldsymbol{\lambda}_k) = d_k(\tilde{\mathbf{x}}_i, \mathbf{g}_m / \boldsymbol{\lambda}_m) \right. \\ \left. \text{o ile } k < m \forall m \neq k (m = 1, \dots, K) \right\} \quad (10)$$

2.4. Schemat algorytmu

1. Dla $i = 1, \dots, n$ i $j = 1, \dots, p$ obliczyć $\tilde{x}_i^j = (D_j, \mathbf{u}^j(i))$, stosując równość (1), (2) lub (3) w zależności od typu zmiennej symbolicznej.

2. Losowo wybrać podział $P^{(0)} = (C_1^{(0)}, \dots, C_K^{(0)})$ lub K różnych obiektów $\mathbf{g}_1^{(0)}, \dots, \mathbf{g}_K^{(0)}$ należących do Ω i przypisać poszczególne obiekty do najbliższego

wzorca $\mathbf{g}_{k^*}^{(0)}$, gdzie $k^* = \arg \min_{k=1, \dots, K} \left\{ \sum_{j=1}^p \sum_{h=1}^{H_j} \left((u_h^j(i))^{(0)} - (v_h^j(k))^{(0)} \right)^2 \right\}$. Przyj-
 jąć $t = 1$.

3. Dla $k = 1, \dots, K$ obliczyć $\mathbf{g}_k^{(t)} = \left((g_k^1)^{(t)}, \dots, (g_k^p)^{(t)} \right)$, $(g_k^j)^{(t)} = \left(D_j, (\mathbf{v}^j(k))^{(t)} \right)$,
 $(j = 1, \dots, p)$, gdzie $(\mathbf{v}^j(k))^{(t)} = \left((v_1^j(k))^{(t)}, \dots, (v_{H_j}^j(k))^{(t)} \right)$, wykorzystując równość
 (7).

4. Obliczyć $\boldsymbol{\lambda}_k^{(t)} = \boldsymbol{\lambda}_k^{(t)} = \left((\lambda^1)^{(t)}, \dots, (\lambda^p)^{(t)} \right)$, gdzie $\boldsymbol{\lambda} = (\lambda^1, \dots, \lambda^p)$ zgodnie z równo-
 ścią (8) lub $\boldsymbol{\lambda}_k^{(t)} = \boldsymbol{\lambda}_k^{(t)} = \left((\lambda^1)^{(t)}, \dots, (\lambda^p)^{(t)} \right)$ zgodnie z (9) w zależności od stosowanej
 metody.

5. Przydzielć poszczególne obiekty do klas zgodnie z regułą (10). Przyjąć $t = t + 1$.

6. Jeżeli nie nastąpiła zmiana w przydzieleniu obiektów do klas to STOP, w przeciwnym przypadku idź do 3.

4. Algorytm klasyfikacji rozmytej dla danych symbolicznych o różnych typach cech

Zaproponowane przez autorkę uogólnienie procedury de Carvalho i de Souza [2010] na przypadek klasyfikacji rozmytej pozwoli w sytuacji trudno separowanych klas wykorzystać częściową przynależność do klas obiektów, których podobieństwo do kilku klas jednocześnie jest duże. Uwzględniając stopnie przynależności do poszczególnych klas, można zdefiniować funkcję, stanowiącą kryterium klasyfikacji następująco: $\tilde{W}(\mathbf{G}, \mathbf{D}, \boldsymbol{\mu}) = \sum_{k=1}^K \sum_{i=1}^n [\mu_k(i)]^r d(\tilde{\mathbf{x}}_i, \mathbf{g}_k / \boldsymbol{\lambda}_k) \rightarrow \min$, przyjmując, że $r > 1$ oznacza stopień rozmycia, $\mu_k(i)$ zaś stopień przynależności obiektu i do klasy C_k oraz $\sum_{k=1}^K \mu_k(i) = 1$.

Zakładając, że wagi są jednakowe w każdej klasie lub niejednakowe, można, korzystając z metody mnożników Lagrange'a i rozwiązując odpowiednie układy równań, wyznaczyć wartości stopni przynależności poszczególnych obiektów do klas odpowiednio:

$$\mu_k(i) = \frac{\left[\sum_{j=1}^p \lambda^j \sum_{h=1}^{H_j} (u_h^j(i) - v_h^j(k))^2 \right]^{-1/(r-1)}}{\sum_{q=1}^K \left[\sum_{j=1}^p \lambda^j \sum_{h=1}^{H_j} (u_h^j(i) - v_h^j(q))^2 \right]^{-1/(r-1)}}, \quad (11)$$

$$\mu_k(i) = \frac{\left[\sum_{j=1}^p \lambda_k^j \sum_{h=1}^{H_j} (u_h^j(i) - v_h^j(k))^2 \right]^{-1/(r-1)}}{\sum_{q=1}^K \left[\sum_{j=1}^p \lambda_q^j \sum_{h=1}^{H_j} (u_h^j(i) - v_h^j(q))^2 \right]^{-1/(r-1)}}. \quad (12)$$

Dalej, postępując analogicznie, można wyznaczyć wektor wzorców klas, który minimalizuje kryterium klasyfikacji:

$$v_h^j(k) = \frac{\sum_{i=1}^n [\mu_k(i)]^r u_h^j(i)}{\sum_{i=1}^n [\mu_k(i)]^r}. \quad (13)$$

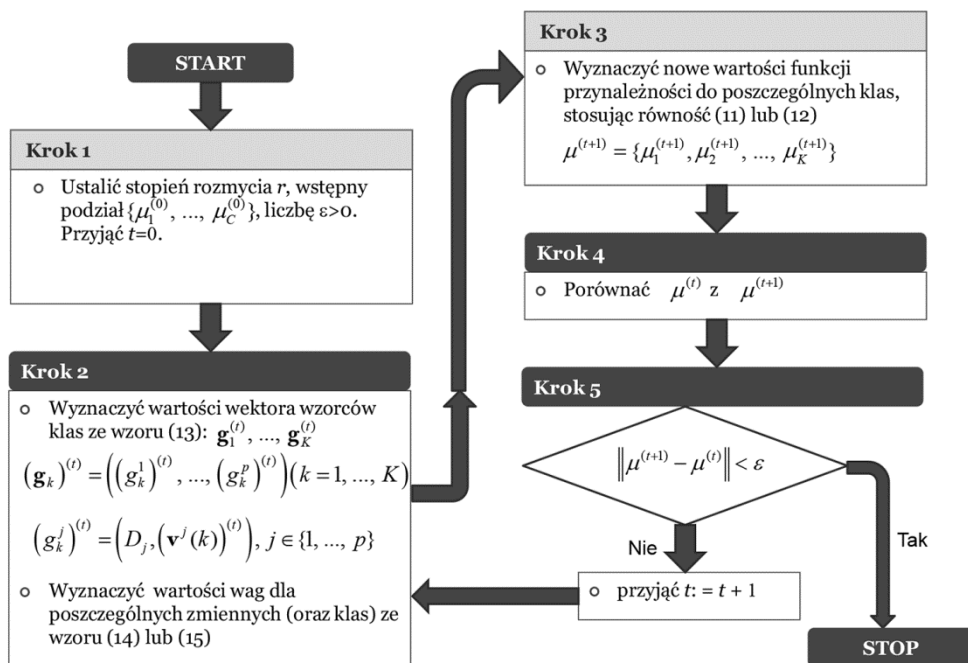
Podobnie można wyznaczyć najlepsze wagi, dla których funkcja kryterium osiąga minimum lokalne, przy czym $\lambda^j > 0$ i $\prod_{j=1}^p \lambda^j = \eta$, gdzie $\eta \in \mathbf{R}$ jest stałe:

$$\lambda^j = \frac{\left\{ \eta \prod_{l=1}^p \left(\sum_{k=1}^K \left[\sum_{i=1}^n [\mu_k(i)]^r \sum_{h=1}^{H_l} (u_h^l(i) - v_h^l(k))^2 \right] \right) \right\}^{\frac{1}{p}}}{\sum_{k=1}^K \left[\sum_{i=1}^n [\mu_k(i)]^r \sum_{h=1}^{H_j} (u_h^j(i) - v_h^j(k))^2 \right]}. \quad (14)$$

Jeżeli w funkcji kryterialnej W uwzględniony zostanie kwadrat odległości euklidesowej, parametryzowany wagami, które mogą być dla poszczególnych klas niejednakowe oraz zmieniają się w każdej iteracji, to przy założeniu, że $\lambda_k^j > 0$ i $\prod_{j=1}^p \lambda_k^j = \chi$, gdzie $\chi \in \mathbf{R}$ jest stałe, do wyznaczenia wag minimalizujących kryterium W można wykorzystać metodę mnożników Lagrange'a oraz pewne elementy algebry i otrzymać wzór:

$$\lambda_k^j = \frac{\left\{ \chi \prod_{l=1}^p \left(\sum_{i=1}^n [\mu_k(i)]^r \sum_{h=1}^{H_l} (u_h^l(i) - v_h^l(k))^2 \right) \right\}^{\frac{1}{p}}}{\sum_{i=1}^n [\mu_k(i)]^r \sum_{h=1}^{H_j} (u_h^j(i) - v_h^j(k))^2}. \quad (15)$$

Poszczególne kroki algorytmu klasyfikacji rozmytej dla danych symbolicznych o różnych typach cech przedstawiono na rys. 1.



Rys. 1. Algorytm rozmytej klasyfikacji dla danych symbolicznych o cechach symbolicznych różnego typu

Źródło: opracowanie własne.

5. Podsumowanie

Przedstawiony algorytm iteracyjny klasyfikacji klasycznej oraz rozmytej pozwala grupować obiekty o cechach symbolicznych mieszane go typu. Algorytm stosujący odległości z różnymi wagami dla poszczególnych klas jest w stanie rozpoznać klasy o różnych kształtach i wielkościach, co stanowi niewątpliwą zaletę. Wadą zaś jest to, że wyniki klasyfikacji są uzależnione od podziału wstępnego. Oceny eksperymentalne dla danych przedziałowych wykazały wyższość algorytmu klasyfikacji stosującego jednakowe wagi pod względem jakości rozpoznawania klas (ocenionej za pomocą skorygowanego indeksu Rand) w konfiguracji danych z prawie jednakową dyspersją klas *a priori*, zaś wyższość algorytmu wykorzystującego wagi niejednakowe dla poszczególnych klas w przypadku, gdy dyspersja klas z góry zadana jest niejednakowa. Zaproponowane metody rozmytej klasyfikacji dla danych symbolicznych o różnych typach cech są uogólnieniem przedstawionych metod de Carvalho i de Souza, zatem mają te same zalety oraz wady. Umożliwiają jednak przypisanie poszczególnym obiektom stopni przynależności do poszczególnych klas w zakresie od 0 do 1. Ma to szczególne znaczenie, gdy klasy są trudno separowane i sztywna klasy-

fikacja wymusza przypisanie obiektu tylko do jednej klasy. Zatem w takim przypadku klasyfikacja rozmyta może dać lepsze rezultaty, rozpoznając obiekty „mieszane”, których podobieństwo do kilku klas jednocześnie jest duże.

Kierunkiem dalszych działań będzie przeprowadzenie badań eksperymentalnych, które pozwolą ocenić skuteczność tych metod na tle innych, w przypadku gdy klasy są trudno separowane oraz prezentacja osiągniętych korzyści praktycznych, wynikających z uogólnienia algorytmów na przypadek klasyfikacji rozmytej.

Literatura

- Bock H.H., Diday E., *Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data*, Springer-Verlag, Berlin, Heidelberg, 2000.
- De Carvalho F.A.T., *Histograms in symbolic data analysis*, Annals of Operations Research 55, 1995, 229-322.
- De Carvalho F.A.T., de Souza R., *Unsupervised pattern recognition models for mixed feature-type symbolic data*, "Pattern Recognition Letters" 31, 2010, 430-443.
- Diday E., Simon J.C., *Clustering Analysis*, [in:] K.S. Fu (ed.), *Digital Pattern Classification*, Springer, Berlin, 1976, 47-94.

CLUSTERING ALGORITHMS FOR MIXED-FEATURE SYMBOLIC OBJECTS

Summary: The majority of discussed classification methods allow clustering of symbolic objects described by variables of the same type. In real applications many objects can be characterized by symbolic mixed feature types: both numeric-valued, interval-valued, set of categories-valued and ordered list-value with weights. The aim of this work is to present clustering algorithms discussed in paper [de Carvalho, de Souza 2010] for objects, which can be described simultaneously by mixed type symbolic data and to propose generalization of these algorithms for fuzzy classification. The main idea is the transformation of mixed feature-type symbolic data into histogram-valued symbolic data.

Keywords: symbolic data analysis, mixed feature-type symbolic data, histogram-valued symbolic data, fuzzy classification.