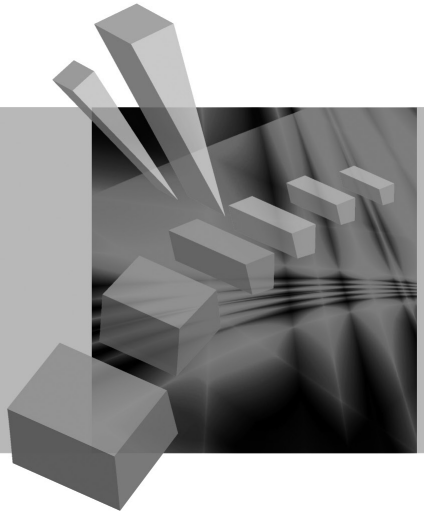
PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu
RESEARCH PAPERS
of Wrocław University of Economics

278

Taksonomia 20

Klasyfikacja i analiza danych – teoria i zastosowania



Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2013

Redaktor Wydawnictwa: Aleksandra Śliwka

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Małgorzata Czupryńska

Projekt okładki: Beata Dębska

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

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

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

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

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM

Spis treści

Summaries

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

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

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 $\Omega = \{1, ..., n\}$, opisywany przez wartości p zmiennych symbolicznych $\{X_1, ..., X_p\}$, jest utożsamiany z wektorem danych symbolicznych o mieszanych typach cech $\mathbf{x}_i = \left(x_i^1, x_i^2, ..., x_i^p\right), i = 1, ..., n$. 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 $A_j = \left\{t_1^j, t_2^j, ..., t_{H_j}^j\right\}$ jest zbiorem kategorii;
- uporządkowanej listy kategorii, czyli x_i^j jest podlistą uporządkowanej listy kategorii $A_j = \left[t_1^j, t_2^j, ..., t_{H_j}^j\right]$;
- przedziału: $X_j(i) = x_i^j = \left[a_i^j; b_i^j\right] \subset [a; b]$, gdzie $[a; b] \in \wp$, $\wp$ jest zbiorem przedziałów domkniętych w zbiorze liczb rzeczywistych;
- histogramu: $X_j(i) = x_i^j = \left(S^j(i), \mathbf{q}^j(i)\right)$, gdzie $\mathbf{q}^j(i) = \left(q_{i1}^j, q_{i2}^j, ..., q_{iH_{ij}}^j\right)$ 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 $P = \left(C_1, C_2, ..., C_K\right)$ zbioru Ω na ustaloną liczbę K klas i odpowiadających im wzorców $\mathbf{G} = \left(\mathbf{g}_1, ..., \mathbf{g}_K\right)$ 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 $\widetilde{X}_j$ o wartościach w postaci histogramu osiągana jest w następujący

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

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

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 $\widetilde{X}_j$ o wartościach histogramu następująco: $\tilde{X}_j(i) = \tilde{x}_i^j = \left(A_j, \mathbf{Q}^j(i)\right)$, gdzie $A_j = \left[t_1^j, t_2^j, ..., t_{H_j}^j\right]$ jest nośnikiem wektora skumulowanych wag $\mathbf{Q}^j(i) = \left(Q_1^j(i), Q_2^j(i), ..., Q_{H_j}^j(i)\right)$. Skumulowane wagi $Q_h^j(i)\left(h = 1, ..., H_j\right)$ kategorii t_h^j z listy A_j są zdefiniowane jako [de Carvalho 1995]:

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

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

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

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

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

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

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

a) $\sum_{h=1}^{H_j} I_h^j = [a;b]$,

b) $I_h^j \cap I_{h'}^j = \varnothing$ jeżeli $h \neq h'$,

c) $\forall h \exists i \in \Omega$ takie, że $I_h^j \cap x_i^j \neq \varnothing$,

d) $\forall i \exists S_i^j \subset \{1, ..., H_j\} : \bigcup_{h \in S_i^j} I_h^j = x_i^j$.

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

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

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

We wzorze (4) rozważa się:

a) kwadraty adaptacyjnych odległości euklidesowych parametryzowane przez jednakowy wektor wag $\boldsymbol{\lambda}_k = \boldsymbol{\lambda}(k = 1, ..., K)$, gdzie $\boldsymbol{\lambda} = (\lambda^1, ..., \lambda^p)$ zmienia się w każdej iteracji, ale jest taki sam dla wszystkich klas:

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

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

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

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 $v_0 = (\mathbf{G}^0, \mathbf{D}^0, \mathbf{P}^0)$, 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 $P = (C_1, ..., C_K)$ oraz $\mathbf{D} = (\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_K)$ są ustalone, to niezależnie od funkcji odległości (równania (5) i (6)) wektor wzorców $\mathbf{G} = (\mathbf{g}_1, ..., \mathbf{g}_K)$, gdzie $\mathbf{g}_k = (g_k^1, ..., g_k^p)(k = 1, ..., K)$ z $g_k^j = (D_j, \mathbf{v}^j(k))$ $(j = 1, ..., p)$, który minimalizuje kryterium klasyfikacji *W*, jest taki, że elementy $v_h^j(k)(h = 1, ..., H_j)$ wektora wag $\mathbf{v}^j(k) = (v_1^j(k), ..., v_{H_j}^j(k))$ 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 $P=\left(C_1, ..., C_K\right)$ oraz $\mathbf{G}=\left(\mathbf{g}_1, ..., \mathbf{g}_K\right)$ są ustalone, to *K* wektorów wag $\mathbf{D}=\left(\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_K\right)$, które minimalizują kryterium *W*, są obliczane w zależności od zastosowanej funkcji odległości.

a) Jeżeli funkcja odległości jest dana przez równość (5), to wektory wag $\boldsymbol{\lambda}_k=\boldsymbol{\lambda}\left(k=1, ..., K\right)$, gdzie $\boldsymbol{\lambda}=\left(\lambda^1, ..., \lambda^p\right)$, 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:

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

b) Jeżeli funkcja odległości jest dana przez równość (6), to wagi $\boldsymbol{\lambda}_k=\left(\lambda_k^1, ..., \lambda_k^p\right)\left(k=1, ..., K\right)$, 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:

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

2.3 *Krok 3: Najlepszy podział*

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

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

2.4. *Schemat algorytmu*

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

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

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

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

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

5. Przydzielić 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: $\tilde{W}\left(\mathbf{G}, \mathbf{D}, \mu\right) = \sum_{k=1}^{K}\sum_{i=1}^{n}\left[\mu_k(i)\right]^r d\left(\tilde{\mathbf{x}}_i, \mathbf{g}_k / \boldsymbol{\lambda}_k\right) \to \min$, przyjmując, że $r > 1$ oznacza stopień rozmycia, $\mu_k(i)$ zaś stopień przynależności obiektu i do klasy C_k oraz $\sum_{k=1}^{K}\mu_k(i) = 1$.

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:

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

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

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

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

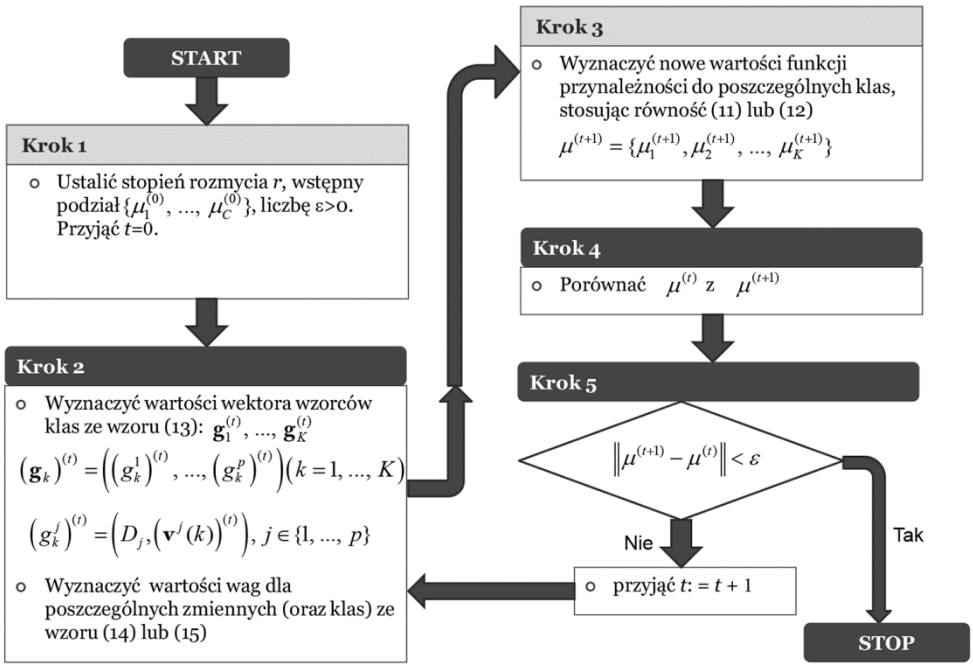
Podobnie można wyznaczyć najlepsze wagi, dla których funkcja kryterium osiąga minimum lokalne, przy czym $\lambda^j>0$ i $\prod_{j=1}^{p}\lambda^j=\eta$, gdzie $\eta\in\mathbf{R}$ jest stałe:

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

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:

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

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 mieszanego 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 „mieszańce", 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 Clasification*, 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.