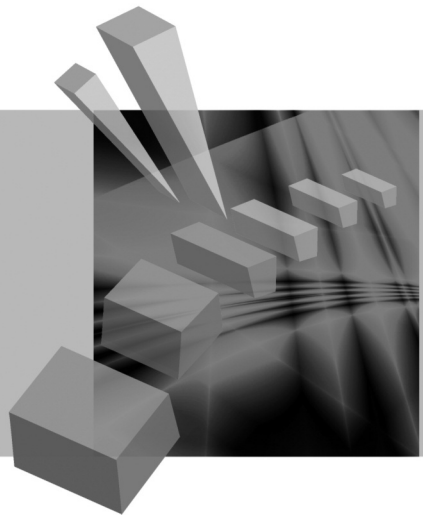
PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu
RESEARCH PAPERS
of Wrocław University of Economics

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 Pociecha,
Mirosław 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

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,
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

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

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 242
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS 2012

Taksonomia 19. Klasyfikacja i analiza danych – teoria i zastosowania ISSN 1899-3192

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 rzeczywistych.

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 poponowanych metod umożliwia grupowanie obiektów opisanych za pomocą zmiennych ustalonego typu. W praktycznych zastosowaniach wiele obiek-

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 $O_i \in \{O_1, ..., O_N\}$ może być utożsamiany z wektorem zaobserwowanych wartości zmiennych $\mathbf{x}_i = [x_{i1},...,x_{iM}]$. Dla dowolnych dwóch wektorów $\mathbf{x}_i = [x_{i1},...,x_{iM}]$ i $\mathbf{x}_j = [x_{j1},...,x_{jM}]$ 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 $\alpha_1, ..., \alpha_M$ – 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}, $\overline{x}_{im}$ – górna granica x_{im}, $\underline{x}_{jm}$ – dolna granica x_{jm}, $\overline{x}_{jm}$ – górna granica x_{jm}, *inters* – długość przecięcia x_{im} i x_{jm}, $l_s = \left|\max(\overline{x}_{im}, \overline{x}_{jm}) - \min(\underline{x}_{im}, \underline{x}_{jm})\right|$, U_m – różnica między najwyższą

[1] *Mixed-type Variables Fuzzy C-Means.*

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

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

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

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

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

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_{jk}, *inters* – liczba elementów części wspólnej x_{ik} i x_{jm}.

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

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

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 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 μ_X, określoną następująco [Zimmermann 1991]:

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

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 $A = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$ będzie zbiorem *N* wektorów cech w przestrzeni 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 $A_1, A_2, ..., A_C$, takich że $A_1 \cup A_2 \cup ... \cup A_C = A$ lub równoważnie za pomocą funkcji przynależności $\mu_1, \mu_2, ..., \mu_C$, 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 $k \in \{1, 2, ..., C\}$. Jest to tzw. twardy podział $\{\mu_1, \mu_2, ..., \mu_C\}$ zbioru *A* na *C* klas $A_1, A_2, ..., A_C$. 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 $\{\mu_1, \mu_2, ..., \mu_C\}$ jest rozmytym *C* podziałem zbioru *A*. Algorytm *C* środków (*FCM*) poszukuje takiego podziału rozmytego $\{\mu_1, \mu_2, ..., \mu_C\}$, dla którego osiągnie minimum następująca funkcja:

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

gdzie: *r* jest ustaloną liczbą większą niż jeden, przedstawiającą stopień rozmycia, $\{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_C\}$ jest zbiorem środków klas, $\|\cdot\|$ normą euklidesową w przestrzeni 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 $\mathbf{w}_k = [w_{1/k}, w_{2/k}, ..., w_{M/k}]$, 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:

$$w_{m/k} = \left[(w_{m1/k}, e_{m1/k}), ..., (w_{mP/k}, e_{mP/k}) \right]. \tag{2}$$

Wówczas: $0 \le e_{mp/k} \le 1$ i $\sum_{p} e_{mp/k} = 1$, $\bigcap_{p} w_{mp/k} = \varnothing$ 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 $e_{mp/k} = 1$, jeśli żadne inne zdarzenie poza $w_{mp/k}$ nie zaszło dla *m*-tej cechy środka

k-tej klasy $w_{m/k}$. Funkcja przynależności $e_{mp/k}$ jest bardzo ważną funkcją wskaźnikową, która umożliwia 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 $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$ będzie zbiorem obiektów reprezentowanych przez cechy różnych typów. Funkcja kryterium FCM jest zdefiniowana jako:

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

gdzie

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

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

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

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 **e** i przyrównując ją do zera, można otrzymać wzór:

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

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*):

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

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

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

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

Interaktywny algorytm grupowania *C*-środków (*MVFCM*) dla danych mieszanego typu:

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

Krok 2: Dla cechy symbolicznej m' obliczyć środek *k*-tej klasy $w_{km'}^{(t)} = \left[(w_{m'1/k}^{(t)}, e_{m'1/k}^{(t)}), ..., (w_{m'p/k}^{(t)}, e_{m'p/k}^{(t)})\right]$, stosując równości (2), (6). Dla cechy rozmytej m obliczyć środek *k*-tej klasy $w_{km}^{(t)} = \left(w_{m1}^{(t)}, w_{m2}^{(t)}, w_{m3}^{(t)}, w_{m4}^{(t)}\right)$, 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 $\left\|\mu^{(t+1)} - \mu^{(t)}\right\| \geq \varepsilon$, 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, Zł, 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, Zł – złoty, Ż – żółty.

Źródło: opracowanie własne na podstawie www.skoda-auto.pl; www.fiat.pl; www.renault.pl; www.citroen.pl; 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
μ_{1j}	0,9941	0,9354	0,0084	0,9257	0,5691	0,8977	0,9922	0,0301	0,9417	0,9303
μ_{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 = {Fabia, Oktawia, Panda, Bravo, C3 Picasso, C1, Thalia, Megane}, C_2 = {Superb, C5}. Fiat Bravo jest obiektem mieszańcem, 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 pozbawione 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.