QUANTITATIVE METHODS IN ECONOMICS
Vol. XIV, No. 2, 2013, pp. 51 – 60

FUZZY CLASSIFICATION OF SYMBOLIC OBJECTS

Małgorzata Machowska–Szewczyk
Department of Artificial Intelligence Methods and Applied Mathematics
West Pomeranian University of Technology in Szczecin
e-mail: mmachowska@wi.zut.edu.pl

Abstract: The aim of this work is to present fuzzy clustering algorithm for objects, which can be described by mixed feature-type symbolic data and fuzzy data. The main idea is the transformation of mixed feature-type symbolic data and fuzzy data into histogram-valued symbolic data. Fuzzy classification is very useful in case, when classes are difficult separated, mixed objects can be classified into class with the fixed degree of membership.

Keywords: fuzzy classification, symbolic data, histogram-valued symbolic data

INTRODUCTION

Clustering algorithms of symbolic objects (e.g. only numeric values or interval-valued data) most often assume that the variables used for description are of the same type. However, there is a lot of real objects requiring the symbolic features for description which can be both numeric-valued, interval-valued, a set of categories-valued and an ordered list-valued with weights.
The aim of this study is presenting the proposal of a generalization of the classification methods of symbolic objects, characterized by mixed type features, proposed in the work de Carvalho and de Souza [2010], relating to the case of fuzzy classification and the possibility of including fuzzy features to describe objects. These methods are based on iterative clustering methodology with adaptation of the Euclidean distance. Distances are changed in each iteration of the algorithm, and can either be the same for all classes, or different for particular groups. In the first step the transformation of symbolic values of various types is made to histogram-valued symbolic data. The modification proposed by the author allows to carry out the classification in both the classical sense (then the

classification method of de Carvalho and de Souza is used) as well as in terms of fuzzy classification. The fuzzy classification is very useful in a situation of classes separated with difficulty, the so called mixed objects can be classified into classes with a certain degree of membership. The classical classification forces the assigning of an object only to one class, therefore the objects whose similarity to several classes at the same time is quite high are not recognized, and the quality of the classification obtained is then low. The proposed algorithm, therefore, contributes to an additional opportunity for mixed-value symbolic data analysis.

TRANSFORMATION INTO HISTOGRAM – VALUED SYMBOLIC DATA

Each object i from the set $\Omega = \{1,...,n\}$, described by the p-values of symbolic variables $\{X_1,...,X_p\}$, is identified with the vector of mixed-value symbolic data $\mathbf{x}_i = (x_i^1, x_i^2,...,x_i^p)$, $i = 1,...,n$. This means that the symbolic variable X_j can assume for a given unit i the value x_i^j in the form of [Bock, Diday 2000]:

- set–valued, if given an item i, $X_j(i) = x_i^j \subset A_j$, where $A_j = \{t_1^j, t_2^j,...,t_{H_j}^j\}$ is set of categories;
- ordered list–valued, if given an item i, x_i^j is set-list of ordered list of categories $A_j = [t_1^j, t_2^j,...,t_{H_j}^j]$;
- interval–valued, if given an item i, $X_j(i) = x_i^j = [a_i^j; b_i^j] \subset [a;b]$, where $[a;b] \in \wp$ and $\wp$ is set closed intervals defined from R;
- histogram–valued, if given an item i, $X_j(i) = x_i^j = \left(S^j(i), \mathbf{q}^j(i)\right)$, where $\mathbf{q}^j(i) = (q_{i1}^j, q_{i2}^j,...,q_{iH_j}^j)$ is the vector of weights defined in $S^j(i)$, such that a weight q_{im}^j corresponds to the category m from $S^j(i)$ and $S^j(i)$ is support of measure $\mathbf{q}^j(i)$.

The aim of standard clustering algorithm [Diday i Simon 1976] is to find a partition $P = \left(C_1, C_2,...,C_K\right)$ of the set Ω into a fixed number K of classes and their corresponding patterns $\mathbf{G} = \left(\mathbf{g_1},...,\mathbf{g}_K\right)$ by the local minimization of criterion function W. This criterion assesses the fitting between classes and their respective representatives.

To overcome the difficulty, which is the representation of objects using ordered or non-ordered symbolic data of various types, the pre-processing is made, whose purpose is to obtain a suitable homogenization of symbolic data. It consists

in the transformation of mixed-value symbolic data to histogram-valued symbolic data.

If X_j is a set-valued variable, its transformation into a symbolic histogram-valued variable $\tilde{X}_j$ is achieved as follows: $\tilde{X}_j(i) = \tilde{x}_i^j = (A_j, \mathbf{q}^j(i))$, where $A_j = \{t_1^j, t_2^j, ..., t_{H_j}^j\}$ is a domain of variable X_j and the support of the weight vector $\mathbf{q}^j(i) = (q_1^j(i), q_2^j(i), ..., q_{H_j}^j(i))$. The weight $q_h^j(i)\,(h = 1,...,H_j)$ of the category $t_h^j \in A_j$ is defined as [de Carvalho 1995]:

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

where $c(x_i^j)$ is the cardinality of a finite set of category $c(x_i^j)$.

If X_j is an ordered list-valued variable, then it is transformed into a histogram-valued symbolic variable $\tilde{X}_j$ as follows: $\tilde{X}_j(i) = \tilde{x}_i^j = \left(A_j, \mathbf{Q}^j(i)\right)$, where $A_j = [t_1^j, t_2^j, ..., t_{H_j}^j]$ is support of the vector of cumulative weights $\mathbf{Q}^j(i) = (Q_1^j(i), Q_2^j(i), ..., Q_{H_j}^j(i))$. The cumulative weights $Q_h^j(i)\,(h = 1,...,H_j)$ of category t_h^j from the list A_j are defined as [de Carvalho 1995]:

$$Q_h^j(i) = \sum_{r=1}^{h} q_r^j(i) \text{ where } q_r^j(i) = \begin{cases} \dfrac{1}{l(x_i^j)} & \text{if } t_r^j \text{ is from the list } x_i^j \\ 0 & \text{otherwise} \end{cases}, \tag{2}$$

$l(x_i^j)$ is the length of an ordered list of category x_i^j.

In the case of interval-valued variable X_j it is transformed to histogram-valued symbolic variable $\tilde{X}_j$ as follows: $\tilde{X}_j(i) = \tilde{x}_i^j = (\tilde{A}_j, \mathbf{Q}^j(i))$, where $\tilde{A}_j = \{I_1^j, I_2^j, ..., I_{H_j}^j\}$ is the list of elementary intervals, constituting support of the cumulative weight vector $\mathbf{Q}^j(i) = (Q_1^j(i), Q_2^j(i), ..., Q_{H_j}^j(i))$. The cumulative weight $Q_h^j(i)\,(h = 1,...,H_j)$ of the elementary interval I_h^j is defined as [de Carvalho 1995]:

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

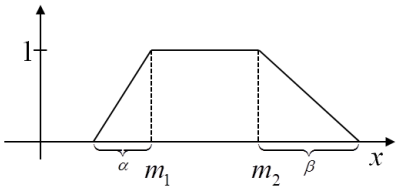
$l(I)$ is the length of the closed interval I.

It can be shown that: $0 \le q_h^j(i) \le 1 \, (h = 1,...,H_j)$ and $\sum_{h=1}^{H_j} q_h^j(i) = 1$. In addition, that: $q_1^j(i) = Q_1^j(i)$ i $q_h^j(i) = Q_h^j(i) - Q_{h-1}^j(i) \, (h = 2,...,H_j)$.

Limits of elementary intervals $I_h^j \, (h = 1,...,H_j)$ are derived from the ordered limits of n+1 intervals $\left\{x_1^j, x_2^j, ..., x_n^j, [a;b]\right\}$ and the number of elementary intervals is at most $2n$. The elementary intervals have the following properties [de Carvalho 1995]:

1. $\sum_{h=1}^{H_j} I_h^j = [a;b]$,
2. $I_h^j \cap I_{h'}^j = \varnothing$ if $h \ne h'$,
3. $\forall h \exists i \in \Omega$ that $I_h^j \cap x_i^j \ne \varnothing$,
4. $\forall i \exists S_i^j \subset \{1,...,H_j\} : \bigcup_{h \in S_i^j} I_h^j = x_i^j$.

Figure 1. Parameterization of *TFN*



Source: own elaboration

The trapezoidal fuzzy numbers *TFN* (see Fig. 1) are in real applications often, represented as *L-R* fuzzy numbers. Let *L* (*R*) be decreasing, shape function from R^+ to $[0,1]$, with $L(0) = 1$, $L(x) < 1$ for all $x > 0$, $L(x) > 0$ for all $x < 1$, $L(x) = 0$ ($L(x) > 0$ for all x and $L(+\infty) = 0$). A fuzzy number A with its membership function μ_A [Zimmermann 1991]:

$$\mu_A(x) = \begin{cases} L\left(\dfrac{m_1 - x}{\alpha}\right) & \text{for } x < m_1 \\ 1 & \text{for } m_1 \le x \le m_2 \\ R\left(\dfrac{x - m_2}{\beta}\right) & \text{for } x > m_2 \end{cases} \tag{4}$$

is called an *L-R* type *TFN*. Symbolically, A can be denoted by $A = (m_1, m_2, \alpha, \beta)_{LR}$, where $\alpha > 0, \beta > 0$ are called left and right spreads, respectively. Using this

parametric representation can be presented four kinds of *TFN*s with real numbers, interval, triangular and trapezoidal fuzzy numbers.

If X_j is variable of trapezoidal fuzzy value $x_i^j = (m_1^j(i), m_2^j(i), \alpha^j(i), \beta^j(i))_{LR}$, its transformation into symbolic histogram-valued variable $\tilde{X}_j$ is accomplished in the following way (author's proposal): $\tilde{X}_j(i) = \tilde{x}_i^j = (A_j, \mathbf{Q}^j(i))$, where $\tilde{A}_j = \{I_1^j, I_2^j, ..., I_{H_j}^j\}$ is the list of interval fuzzy numbers constructed on elementary intervals, constituting support of the cumulative weight vector $\mathbf{Q}^j(i) = (Q_1^j(i), Q_2^j(i), ..., Q_{H_j}^j(i))$. Cumulative weight $Q_h^j(i)\,(h = 1,...,H_j)$ interval fuzzy number I_h^j is defined as:

$$Q_h^j(i) = \sum_{r=1}^{h} q_r^j(i), \text{ where } q_r^j(i) = \frac{l(I_r^j \cap x_i^j)}{l(x_i^j)}, \tag{5}$$

$l(I)$ is a area under a membership function of fuzzy number I.

It can be show, that: $0 \le q_h^j(i) \le 1\,(h = 1,...,H_j)$ and $\sum_{h=1}^{H_j} q_h^j(i) = 1$. Moreover, again $q_1^j(i) = Q_1^j(i)$ i $q_h^j(i) = Q_h^j(i) - Q_{h-1}^j(i)\,(h = 2,...,H_j)$.

The boundaries of fuzzy numbers $I_h^j\,(h = 1,...,H_j)$ are obtained from the ordered boundaries of supports and cores of all considered fuzzy numbers $\{x_1^j, x_2^j, ..., x_n^j\}$.

After the pre-processing step every object i (i=1,…,n) is represented by a histogram-valued symbolic data vector $\tilde{\mathbf{x}}_i = \left(\tilde{x}_i^1, ..., \tilde{x}_i^p\right)$, and $\tilde{x}_i^j = \left(D_j, \mathbf{u}^j(i)\right)$, where D_j (a domain of variable $\tilde{X}_j$) depending on the type of the primary variable is the set of categories, an ordered list of categories or a list of elementary intervals, a list of fuzzy numbers with supports resulting from the elementary intervals, $\mathbf{u}^j(i) = (u_1^j(i), ..., u_{H_j}^j(i))$ is a vector of weights or of cumulative weights. The pattern of class $C_k\left(k = 1,...,K\right)$ is also represented by a histogram-valued symbolic data vector $\mathbf{g}_k = (g_k^1, ..., g_k^p)$, $g_k^j = (D_j, \mathbf{v}^j(k))\ (j = 1,...,p)$ with a vector of weights or of cumulative weights $\mathbf{v}^j(k) = (v_1^j(k), ..., v_{H_j}^j(k))$, where D_j is the set of categories, list of categories or a list of elementary intervals. It is noteworthy that for each variable $\tilde{X}_j\ (j = 1,...,p)$ the support is the same for all units and patterns.

According to the general scheme, the iterative classification algorithm [Diday i Simon 1976] is searching for a set Ω partition $P^* = (C_1^*, C_1^*, ..., C_K^*)$ into a fixed number K of classes, corresponding to K patterns $\mathbf{G}^* = (\mathbf{g}_1^*, ..., \mathbf{g}_K^*)$

representing the classes in P^*, and K of weight vectors parametrizing the squares of adaptive Euclidean distances, for which the criterion function value is minimum:

$$W(\mathbf{G},\mathbf{D},P)=\sum_{k=1}^{K}\sum_{i\in C_k} d(\tilde{\mathbf{x}}_i,\mathbf{g}_k/\boldsymbol{\lambda}_k). \tag{6}$$

In the formula (4) the following is considered:

- squares of adaptive Euclidean distances parameterized by the same vector of weights $\boldsymbol{\lambda}_k=\boldsymbol{\lambda}(k=1,...,K)$, where $\boldsymbol{\lambda}=(\lambda^1,...,\lambda^p)$ changes at each iteration but is the same for all classes:

$$d(\tilde{\mathbf{x}}_i,\mathbf{g}_k/\boldsymbol{\lambda})=\sum_{j=1}^{p}\lambda^j\phi^2\left(\mathbf{u}^j(i),\mathbf{v}^j(k)\right)=\sum_{j=1}^{p}\lambda^j\sum_{h=1}^{H_j}\left(u_h^j(i)-v_h^j(k)\right)^2, \tag{7}$$

- squares of adaptive Euclidean distances parameterized by the weight vectors $\boldsymbol{\lambda}_k=(\lambda_k^1,...,\lambda_k^p),(k=1,...,K)$, that change with each iteration and are different for particular classes:

$$d(\tilde{\mathbf{x}}_i,\mathbf{g}_k/\boldsymbol{\lambda}_k)=\sum_{j=1}^{p}\lambda_k^j\phi^2\left(\mathbf{u}^j(i),\mathbf{v}^j(k)\right)=\sum_{j=1}^{p}\lambda_k^j\sum_{h=1}^{H_j}\left(u_h^j(i)-v_h^j(k)\right)^2. \tag{8}$$

In the first case the weight vector is estimated globally for all classes at once, while in the second case the weights are estimated locally for each class.

FUZZY CLUSTERING ALGORITHM FOR SYMBOLIC OBJECTS

The generalization of de Carvalho and de Souza procedure [2010] proposed by the author in this paper for the case of the fuzzy classification will permit, in a situation of classes separated with difficulty, to use the partial membership of classes of objects whose similarity to several classes at the same time is high. Given the degree of membership to particular classes, one can define a function representing the classification criterion as follows:

$$\tilde{W}(\mathbf{G},\mathbf{D},\mu)=\sum_{k=1}^{K}\sum_{i=1}^{n}[\mu_k(i)]^r\, d(\tilde{\mathbf{x}}_i,\mathbf{g}_k/\boldsymbol{\lambda}_k)\rightarrow\min, \tag{9}$$

assuming that $r>1$ is the degree of fuzziness, whereas $\mu_k(i)$ is the degree of the object i membership of class C_k and $\sum_{k=1}^{K}\mu_k(i)=1$.

Assuming that the weights are the same in each class or different, one can use the method of Lagrange multipliers and solve the corresponding systems of equations, to determine the degree of individual objects membership of the classes as follows, respectively:

Using the method of Lagrange multipliers can solve the corresponding systems of equations and determine the degree of individual objects membership of the classes as formula (10) when the weights are the same in each class or formula (11), when the weights are different:

$$\mu_k(i) = \frac{\left[\sum_{j=1}^{p} \lambda^j \sum_{h=1}^{H_j} \left(u_h^j(i) - v_h^j(k)\right)^2\right]^{-1/(r-1)}}{\sum_{q=1}^{K}\left[\sum_{j=1}^{p} \lambda^j \sum_{h=1}^{H_j} \left(u_h^j(i) - v_h^j(k)\right)^2\right]^{-1/(r-1)}}, \tag{10}$$

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

Next, proceeding in the analogous manner, one can designate vector of class patterns that minimizes the classification criterion:

$$v_h^j(k) = \frac{\sum_{i=1}^{n} [\mu_k(i)]^r u_h^j(i)}{\sum_{i=1}^{n} [\mu_k(i)]^r}. \tag{12}$$

Similarly, the best weights can be determined for which the criterion function reaches a local minimum, and $\lambda^j > 0$ and $\prod_{j=1}^{p} \lambda^j = \eta$, where $\eta \in \mathsf{R}$ is constant:

$$\lambda^j = \frac{\left\{\eta \prod_{l=1}^{p}\left(\sum_{k=1}^{K}\left[\sum_{i=1}^{n} [\mu_k(i)]^r \sum_{h=1}^{H_l} \left(u_h^l(i) - v_h^l(k)\right)^2\right]\right)\right\}^{\frac{1}{p}}}{\sum_{k=1}^{K}\left[\sum_{i=1}^{n} [\mu_k(i)]^r \sum_{h=1}^{H_j} \left(u_h^j(i) - v_h^j(k)\right)^2\right]}. \tag{13}$$

If in the criterion function *W* the squared Euclidean distance is considered, parameterized by weights, which may be different for particular classes, and change with each iteration, then assuming that $\lambda_k^j > 0$ and $\prod_{j=1}^{p} \lambda_k^j = \chi$, where $\chi \in \mathsf{R}$ is constant, to determine the weights that minimize the criterion *W* one can use the method of Lagrange multipliers and some elements of algebra and obtain the formula:

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

The particular steps in the algorithm for fuzzy classification of symbolic data with different types of features are as follows:

1. For $i = 1,...,n$ and $j = 1,...,p$ calculate $\tilde{x}_i^j = (D_j, \mathbf{u}^j(i))$, using equality (1), (2), (3), (5) depending on the type of symbolic variable.
2. Assume $t = 0$.
3. Fix the degree of fuzziness $r > 1$, the initial fuzzy partition $\mu^{(t)} = \left\{\mu_1^{(t)},...,\mu_K^{(t)}\right\}$ and the number $\varepsilon > 0$.

4. Calculate the vector of class patterns $\mathbf{g}_1^{(t)},...,\mathbf{g}_K^{(t)}$, where $\mathbf{g}_k^{(t)} = \left(\left(g_k^1\right)^{(t)},\left(g_k^2\right)^{(t)},...,\left(g_k^p\right)^{(t)}\right)$, $k \in \{1,...,K\}$, $\left(g_k^j\right)^{(t)} = \left(D_j,(\mathbf{v}_j(k))^{(t)}\right)$, $j \in \{1,...,p\}$, is given by (10).
5. Determine the weight vector values for particular variable and classes using the formulas (13) or (14).
6. Calculate the new value of membership function $\mu^{(t+1)} = \left\{\mu_1^{(t+1)},...,\mu_K^{(t+1)}\right\}$ using formulas (10) or (11).
7. If $\left\|\mu^{(t+1)} - \mu^{(t)}\right\| > \varepsilon$ then assume $t := t+1$ and go back to step 4. Otherwise STOP.

EMPIRICAL EXAMPLE

A set of objects consists of 10 brands of cars from four companies: Skoda, Fiat, Citroen and Renault. Each brand is characterized by six features: company, engine capacity, price, available colour, comfort, fuel consumption. The features: company, engine capacity, colour take symbolic values, the price is an real valued, comfort and safety are fuzzy data. The data set is shown in Table 1.

Table 1. Data set of cars

Lp.	Brands	Company	Engine capacity	Price	Colour[1]	Comfort	Fuel consumption
1	Fabia	Skoda	1,4	43,35	B, Br, C, Cz, Cz1, F, M, M1, N, N1, P, S, S1, S2, Z, Ż	[6.5;6.7; 1.5; 0.7]	[5.4; 6.4; 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;7.35; 0.65; 0.65]	[5.9;6.9; 0.8; 1.6]
3	Superb	Skoda	1,8	88,3	B, Br, C, Cz, M, M1, N, N1, P, S, S1, S2,	[8.3;9.3; 0;0]	[7.5;8.7; 0.9;1.9]

[1] In the colour feature the following notation is adopted: B-white, B1-pearl white, Br-burgundy, C-red, Cz-black, Cz1-black pearl, F-violet, Gr-graphite, M- sea green, M1-light sea green, N-blue, N1-light blue, P- pistachio-green, Ps- sand-coloured, S-silver, S1-light gray, S2-gray, Z-green, Zł-golden, Ż-yellow.

Lp.	Brands	Company	Engine capacity	Price	Colour[1]	Comfort	Fuel consumption
4	Panda	Fiat	1,2	26,99	B, C, , Cz, F, M1, N, N1, Ps, S2, Zł, Ż	[6.0;7.0; 0.1;0.4]	[4.9;4.9; 0.9; 1.5]
5	Bravo	Fiat	1,6	66,99	B, B1, C, Cz, N, N1, S1, S2	[7.5;7.5; 0.4; 0.7]	[4.9;4.9; 0.8; 1.4]
6	C3 Picasso	Citroen	1,4	56,6	B1,C, Cz, N, Ps,S2, Z	[6,5;7,5; 0.2;0.2]	[6.1;7.1; 1.1; 1.6]
7	C1	Citroen	1	43,1	B, C, Cz, N, P, Ps,S1, S2	[6.8;6.8; 1;1]	[4.5;4.5; 0.6;1]
8	C5	Citroen	1,6	101,7	B, B1, Br, Cz1, Gr, M, S, S1, S2	[8.5;9.5; 0;0]	[6.6;7.6; 1.1; 1.2]
9	Thalia	Renault	1,2	29,9	B, C, Cz1,N1, Ps, S, S1, S2	[6.8;6.8; 0.7; 0.8]	[5.9;5.9; 1.1; 1.7]
10	Megane	Renault	1,6	54,45	B, Cz, N, S, S1, S2,	[7.2;8.2; 0; 0.1]	[6.3;7.3; 0.8; 1.8]

Source: the author's own elaboration on the basis of www.skoda-auto.pl; www.fiat.pl; www.renault.pl; www.citroen.pl; opinie.auto.com.pl

Table 2. Membership for cars calculated by fuzzy clustering algorithm

Lp.	1	2	3	4	5	6	7	8	9	10
$\mu_1(i)$	0,747	0,694	0,186	0,769	0,640	0,721	0,706	0,177	0,761	0,523
$\mu_2(i)$	0,253	0,306	0,814	0,231	0,360	0,279	0,294	0,823	0,239	0,477

Source: the author's own elaboration

Table 2 shows the values of the membership function of two classes for 10 objects, obtained as a result of the application of fuzzy classification algorithm for various types of symbolic and fuzzy variables using different weights for particular classes, assuming that $r = 2$, $K = 2$ i $\varepsilon = 0{,}0001$.

Analyzing the results in Table 2, two determined classes can be isolated: C1={Fabia, Oktawia, Panda, Bravo, C3 Picasso, C1, Thalia, Megane}, C2={Superb, C5}. Renault Megane is a mixed object whose belonging to both classes is considerably high. It is also possible to notice that Fiat Bravo and Skoda Octavia have a fairly high degree of belonging to the second class, which means that there is a relatively high degree of similarity of these vehicles to the objects belonging to class 2.

CONCLUDING REMARKS

The presented iterative algorithm of the classical and fuzzy classification permits to cluster objects featured by mixed-value symbolic data. The algorithm

using distances with different weights for particular classes is able to identify the classes of different shapes and sizes, which is a definite advantage. The disadvantage is that they are dependent on the initial partition. The experimental evaluations for the interval-valued data have showed the superiority of classification algorithm applying the same weights, in terms of class recognition quality (assessed using the corrected Rand index) in the configuration of data with a priori almost equal dispersion of classes, and the superiority of the algorithm using different weights for particular classes where dispersion of classes is preset in advance as a different one. The proposed fuzzy classification methods for symbolic data with different types of features are a generalization of the methods presented in the work of de Carvalho and de Souza [2010], and therefore they have the same advantages and disadvantages. They allow, however, to assign to the individual objects the degrees of membership of different classes in the range from 0 to 1. This is of particular importance when the classes are separated with difficulty and the classical clustering forces the assignment of a given object to one class only. Therefore, in this case, the fuzzy classification may give better results, identifying the mixed objects whose similarity to several classes at the same time is high.

REFERENCES

Bock H. H., Diday E. (2000) Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data, Springer-Verlag, Berlin, Heidelberg.

De Carvalho F.A.T. (1995) Histograms in symbolic data analysis. Annals of Operations Research 55, 229–322.

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

Diday E., Simon J.C. (1976) Clustering analysis. In: Fu, K.S. (Ed.), Digital Pattern Clasification. Springer, Berlin, 47–94.

Zimmermann H.J. (1991) Fuzzy Set Theory and Its Applications, Kluwer, Dordrecht.