UDC 004.62
Original Scientific Paper

A Heuristic Approach to Possibilistic Clustering for Fuzzy Data

Dmitri A. Viattchenin *viattchenin@mail.ru*
National Academy of Sciences of Belarus
United Institute of Informatics Problems

Abstract

The paper deals with the problem of the fuzzy data clustering. In other words, objects attributes can be represented by fuzzy numbers or fuzzy intervals. A direct algorithm of possibilistic clustering is the basis of an approach to the fuzzy data clustering. The paper provides the basic ideas of the method of clustering and a plan of the direct possibilistic clustering algorithm. Definitions of fuzzy intervals and fuzzy numbers are presented and distances for fuzzy numbers are considered. A concept of a vector of fuzzy numbers is introduced and the fuzzy data preprocessing methodology for constructing of a fuzzy tolerance matrix is described. A numerical example is given and results of application of the direct possibilistic clustering algorithm to a set of vectors of triangular fuzzy numbers are considered in the example. Some preliminary conclusions are stated.
Keywords: possibilistic clustering, fuzzy tolerance, allotment among fuzzy clusters, typical point, vector of fuzzy numbers

1. Introduction

The first subsection of this introduction provides some preliminary remarks on fuzzy and possibilistic clustering. A brief review of fuzzy data fuzzy clustering methods is presented in the second subsection.

1.1. Preliminary Remarks

Cluster analysis is a structural approach to solving the problem of object classification without training samples. Clustering is a process aiming at grouping a set of objects into classes according to the characteristics of data so that objects within a cluster have high mutual similarity while objects in different clusters are dissimilar. Clustering methods are called also automatic classification methods and numerical taxonomy methods. In real applications there is very often no sharp boundary between clusters so that fuzzy clustering is often better suited for the data. Membership degrees between zero and one are used in fuzzy clustering instead of crisp assignments of the data to clusters. Areas of application of fuzzy cluster analysis include, for example, data analysis, pattern recognition, and image segmentation. Heuristic methods of fuzzy clustering, hierarchical methods of fuzzy clustering and optimization methods of fuzzy clustering were proposed by different researchers. Fuzzy clustering methods are described in [5], [10], [14] in detail.

The most widespread approach in fuzzy clustering is the optimization approach and the traditional optimization methods of fuzzy clustering are based on the concept of fuzzy partition. The initial set $X = \{x_1, \ldots, x_n\}$ of n objects represented by the matrix of similarity coefficients, the matrix of dissimilarity coefficients or the matrix of object attributes, should be divided into c fuzzy clusters. Namely, the grade $u_{li}, 1 \le l \le c, 1 \le i \le n$, to which an object x_i belongs to the fuzzy cluster A^l should be determined. For each object $x_i, i = 1, \ldots, n$ the grades of membership should satisfy the conditions of a fuzzy partition:

$$\sum_{l=1}^{c} u_{li} = 1, 1 \le i \le n; \quad 0 \le u_{li} \le 1, 1 \le l \le c \tag{1}$$

In other words, the family of fuzzy sets $P(X) = \{A^l \mid l = \overline{1,c}, c \le n\}$ is the fuzzy partition of the initial set of objects $X = \{x_1,...,x_n\}$ if condition (1) is met. The best known optimization approach to fuzzy clustering is the method of fuzzy c-means, developed by Bezdek [2].

If, on the other hand, condition

$$\sum_{l=1}^{c} u_{li} \ge 1, 1 \le i \le n; \quad 0 \le u_{li} \le 1, 1 \le l \le c \tag{2}$$

is met for each object $x_i, 1 \le i \le n$, then the corresponding family of fuzzy sets $C(X) = \{A^l \mid l = \overline{1,c}, c \le n\}$ is the fuzzy coverage of the initial set of objects $X = \{x_1,...,x_n\}$. The concept of fuzzy coverage is used mainly in heuristic fuzzy clustering procedures. The algorithm of Chiang, Yue, and Yin [4] is a very good illustration for the characterization.

A possibilistic approach to clustering was proposed by Krishnapuram and Keller [7]. A concept of possibilistic partition is a basis of possibilistic clustering methods and the possibilistic membership values $\mu_{li}, 1 \le l \le c, 1 \le i \le n$ can be interpreted as the values of typicality degree. For each object $x_i, i = 1,\ldots,n$ the grades of membership should satisfy the conditions of a possibilistic partition:

$$\sum_{l=1}^{c} \mu_{li} > 0, 1 \le i \le n; \quad 0 \le \mu_{li} \le 1, 1 \le l \le c \tag{3}$$

So, the family of fuzzy sets $Y(X) = \{A^l \mid l = \overline{1,c}, c \le n\}$ is the possibilistic partition of the initial set of objects $X = \{x_1,...,x_n\}$ if condition (3) is met. The possibilistic approach to clustering was developed by Łęski [8], Yang and Wu [20] and other researchers. This approach can be considered as a way in the optimization approach in fuzzy clustering because all methods of possibilistic clustering are objective function-based methods.

All methods of possibilistic clustering are objective function-based methods. However, heuristic clustering algorithms display high level of essential clarity and low level of a complexity. Some heuristic clustering algorithms are based on a definition of cluster concept and the aim of these algorithms is cluster detection conform to a given definition. Mandel [9] notes that such algorithms are called algorithms of direct classification or direct clustering algorithms.

Direct heuristic algorithms of fuzzy clustering are simple and very effective in many cases. For example, a direct fuzzy clustering method was outlined by Viattchenin [13], where a basic version of direct fuzzy clustering algorithm was described. Detection of a unique allotment among given number c of fuzzy α-clusters is the aim of classification. The allotment of elements of the set of classified objects among fuzzy clusters can be considered as a special case of possibilistic partition. The fact was demonstrated by Viattchenin [16], [17]. That is why the basic version of the algorithm, which is described by Viattchenin [13], can be considered as a direct algorithm of possibilistic clustering and the algorithm can be called the $D-AFC(c)$-algorithm.

1.2. A Problem of Clustering of Fuzzy Data

Most fuzzy clustering techniques are designed for handling crisp data with their class membership functions. However, the data can be uncertain or fuzzy. For example, the temperature of a room varies as a function of distance from a reference point. Some other examples of imprecise data can be considered. In similar examples, it is imprecision of the observation itself that is of interest rather than the uncertainty due to statistical variation. Fuzzy data is quite a natural type of data, like non-precise data. So, a problem of fuzzy data clustering arises. The problem is very urgent, for example, in medical diagnostics and military applications.

Fuzzy numbers are well used to model the fuzziness of data. Yang and Ko [18] recently proposed a class of fuzzy c-number clustering procedures for fuzzy data clustering. These FCN procedures could be used for handling LR-type, triangular, trapezoidal and Gaussian fuzzy numbers. The approach was developed by Hung and Yang [6] for an exponential-type distance function and the $AFCN$-algorithm was elaborated. However, FCN and $AFCN$ clustering procedures are objective function-based fuzzy clustering algorithms and a fuzzy c-partition is the result of FCN and $AFCN$ clustering procedures application to the data set. Moreover, a set of fuzzy numbers is the input data for FCN and $AFCN$ algorithms of fuzzy clustering. In other words, elements of the initial set $X = \{x_1,...,x_n\}$ of n objects are fuzzy numbers. Notable, that the $D-AFC(c)$-algorithm can also be used for fuzzy numbers clustering. The technique of the $D-AFC(c)$-algorithm application for fuzzy numbers clustering was proposed in [15].

However, in many practical clustering problems objects are described by a set of attributes with fuzzy values. For example, a modification FCM-algorithm for fuzzy attributes was developed by Butkiewicz and Nieradka [3], where for each object $x_i, i = 1,\ldots,n$ and each attribute $x^t, t = 1,\ldots, p$ a membership function $\mu_i(x^t)$ is defined. The discrete membership function $\mu_i(x^t)$ is representing uncertainty of each attribute in the case. Unfortunately, values of objects attributes cannot be represented by the discrete membership function in general.

A continuous membership function for each attribute is more general tool for fuzzy data representation. So, values of objects attributes should be treated as fuzzy numbers. The main goal of the paper is a consideration of possibilities of an application of the $D-AFC(c)$-algorithm clustering of the data where all attributes are represented by fuzzy numbers. For this purpose, the $D-AFC(c)$-algorithm is described. A short consideration of fuzzy numbers definitions is presented and distances for fuzzy numbers are considered. A concept of a vector of fuzzy numbers is introduced and a methodology for the data preprocessing is described in detail. A numerical example is given and results of application of the $D-AFC(c)$-algorithm of fuzzy clustering to a set of vectors of triangular fuzzy numbers are considered in the example. Preliminary conclusions are formulated and perspectives on future investigations are outlined.

2. A Direct Method of Possibilistic Clustering

The basic concepts of the heuristic method of possibilistic clustering based on the allotment concept are considered in the first subsection. A plan of the direct clustering algorithm is considered in the second subsection of the section.

2.1. Basic Concepts

Let us recall the basic concepts of the fuzzy clustering method based on the concept of allotment among fuzzy clusters, which was proposed by Viattchenin [13]. The concept of fuzzy tolerance is the basis for the concept of fuzzy α-cluster. That is why definition of fuzzy tolerance must be considered in the first place.

Let $X = \{x_1,...,x_n\}$ be the initial set of elements and $T : X \times X \to [0,1]$ some binary fuzzy relation on $X = \{x_1,...,x_n\}$ with $\mu_T(x_i, x_j) \in [0,1]$, $\forall x_i, x_j \in X$ being its membership function.

Definition 1. *Fuzzy tolerance is the fuzzy binary intransitive relation which possesses the symmetricity property*

$$\mu_T(x_i, x_j) = \mu_T(x_j, x_i), \forall x_i, x_j \in X, \quad (4)$$

and the reflexivity property

$$\mu_T(x_i, x_i) = 1, \forall x_i \in X. \quad (5)$$

The notions of powerful fuzzy tolerance, feeble fuzzy tolerance and strict feeble fuzzy tolerance were considered by Viattchenin [13], as well. In this context the classical fuzzy tolerance in the sense of definition 1 was called usual fuzzy tolerance. However, the essence of the method here considered does not depend on the kind of fuzzy tolerance. That is why the method herein is described for any fuzzy tolerance T.

Let us consider the general definition of fuzzy cluster, the concept of the fuzzy cluster's typical point and the concept of the fuzzy allotment of objects. The number c of fuzzy clusters can be equal the number of objects, n. This is taken into account in further considerations.

Let $X = \{x_1,...,x_n\}$ be the initial set of objects. Let T be a fuzzy tolerance on X and α be α-level value of T, $\alpha \in (0,1]$. Columns or lines of the fuzzy tolerance matrix T are fuzzy sets $\{A^1,...,A^n\}$ on X.

Definition 2. *The α-level fuzzy set $A^l_{(\alpha)} = \{(x_i, \mu_{A^l}(x_i)) \mid \mu_{A^l}(x_i) \geq \alpha\}$, $x_i \in X$, $l \in \{1,\ldots,n\}$ is fuzzy α-cluster or, simply, fuzzy cluster. So, $A^l_{(\alpha)} \subseteq A^l$, $\alpha \in (0,1]$, $A^l \in \{A^1,\ldots,A^n\}$ and μ_{li} is the membership degree of the element $x_i \in X$ for some fuzzy cluster $A^l_{(\alpha)}$, $\alpha \in (0,1]$, $l \in \{1,\ldots,n\}$. Value of α is the tolerance threshold of fuzzy clusters elements.*

The membership degree of the element $x_i \in X$ for some fuzzy cluster $A^l_{(\alpha)}, \alpha \in (0,1], l \in [1,n]$ can be defined as a

$$\mu_{li} = \begin{cases} \mu_{A^l}(x_i), & x_i \in A^l_\alpha \\ 0, & otherwise \end{cases}, \quad (6)$$

where an α-level $A^l_\alpha = \{x_i \in X \mid \mu_{A^l}(x_i) \geq \alpha\}$, $\alpha \in (0,1]$ of a fuzzy set A^l is the support of the fuzzy cluster $A^l_{(\alpha)}$. So, the α-level A^l_α of a fuzzy set A^l is a crisp set and condition $A^l_\alpha = Supp(A^l_{(\alpha)})$ is met for each fuzzy cluster $A^l_{(\alpha)}, \alpha \in (0,1], l \in [1,n]$.

Membership degree can be interpreted as a degree of typicality of an element to a fuzzy cluster. The value of a membership function of each element of the fuzzy cluster in the sense of definition 2 is the degree of similarity of the object to some typical object of fuzzy cluster. The value zero for a fuzzy set membership function is equivalent to non-belonging of an element to a fuzzy set. That is why values of tolerance threshold α are considered in the interval $(0,1]$.

Definition 3. *If T is a fuzzy tolerance on X, where X is the set of elements, and $\{A^1_{(\alpha)},...,A^n_{(\alpha)}\}$ is the family of fuzzy clusters for some α, then the point $\tau^l_e \in A^l_\alpha$, for which*

$$\tau^l_e = \arg\max_{x_i} \mu_{li}, \forall x_i \in A^l_\alpha \quad (7)$$

is called a typical point of the fuzzy cluster $A^l_{(\alpha)}, \alpha \in (0,1], l \in [1,n]$.

Obviously, a fuzzy cluster can have several typical points. That is why symbol e is the index of the typical point.

Definition 4. *Let $R^\alpha_z(X) = \{A^l_{(\alpha)} \mid l = \overline{1,c}, 2 \leq c \leq n\}$ be a family of fuzzy clusters for some value of tolerance threshold $\alpha \in (0,1]$, which are generated by some fuzzy tolerance T on the initial set of elements $X = \{x_1,...,x_n\}$. If condition*

$$\sum_{l=1}^{c} \mu_{li} > 0, \quad \forall x_i \in X \quad (8)$$

is met for all fuzzy clusters $A^l_{(\alpha)}, l = \overline{1,c}, c \leq n$, then the family is the allotment of elements of the set $X = \{x_1,...,x_n\}$ among fuzzy clusters $\{A^l_{(\alpha)} \mid l = \overline{1,c}, 2 \leq c \leq n\}$ for some value of the tolerance threshold α.

It should be noted that several allotments $R^\alpha_z(X)$ can exist for some tolerance threshold α. That is why symbol z is the index of an allotment.

The condition (8) requires that every object $x_i, i=\overline{1,n}$ must be assigned to at least one fuzzy cluster $A^l_{(\alpha)}, l=\overline{1,c}, c \le n$ with the membership degree higher than zero. The condition $2 \le c \le n$ requires that the number of fuzzy clusters in $R^\alpha_z(X)$ must be more than two. Otherwise, the unique fuzzy cluster will contain all objects, possibly with different positive membership degrees.

Obviously, the definition of the allotment among fuzzy clusters (8) is similar to the definition of the possibilistic partition (3). So, the allotment among fuzzy clusters can be considered as the possibilistic partition and fuzzy clusters in the sense of definition 2 are elements of the possibilistic partition. However, the concept of allotment will be used in further considerations. The concept of allotment is the central point of the method. But the next concept introduced should be paid attention to, as well.

Definition 5. *Allotment $R^\alpha_I(X)=\{A^l_{(\alpha)} \mid l=\overline{1,n}\}$ of the set of objects among n fuzzy clusters for some tolerance threshold α is the initial allotment of the set $X=\{x_1,...,x_n\}$.*

In other words, if initial data are represented by a matrix of some fuzzy T then lines or columns of the matrix are fuzzy sets $A^l \subseteq X, l=\overline{1,n}$ and level fuzzy sets $A^l_{(\alpha)}, l=\overline{1,n}, \alpha \in (0,1]$ are fuzzy clusters. These fuzzy clusters constitute an initial allotment for some tolerance threshold α and they can be considered as clustering components.

The problem of clustering can be defined in general as the problem of discovering the unique allotment $R^*(X)$, resulting from the classification process, which corresponds to either most natural allocation of objects among fuzzy clusters or to researcher's opinion about classification. In the first case, the number of fuzzy clusters c is not fixed. In the second case, the researcher's opinion determines the kind of the allotment sought. So, the classification problem formulation depends on the parameters of classification and these parameters are determined for a problem of classification in a concrete case. The parameters of classification are considered in [16].

If some allotment $R^\alpha_z(X)=\{A^l_{(\alpha)} \mid l=\overline{1,c}, c \le n\}$ corresponds to the formulation of a concrete problem, then this allotment is an adequate allotment. In particular, if condition

$$\sum_{l=1}^{c} card(A^l_\alpha) \ge card(X), \forall A^l_{(\alpha)} \in R^\alpha_z(X), \alpha \in (0,1], card(R^\alpha_z(X)) = c, \tag{9}$$

and condition

$$card(A^l_\alpha \cap A^m_\alpha) \le w, \forall A^l_{(\alpha)}, A^m_{(\alpha)}, l \ne m, \alpha \in (0,1], \tag{10}$$

are met for all fuzzy clusters $A^l_{(\alpha)}, l=\overline{1,c}$ of some allotment $R^\alpha_z(X)=\{A^l_{(\alpha)} \mid l=\overline{1,c}, c \le n\}$ then the allotment is the allotment among particularly separate fuzzy clusters and $0 \le w \le n$ is the maximum number of elements in the intersection area of different fuzzy clusters. Obviously, if $w=0$ in conditions (9) and (10) then the intersection area of any pair of different fuzzy cluster is an empty set and fuzzy clusters are fully separate fuzzy clusters.

The adequate allotment $R^\alpha_z(X)$ for some value of tolerance threshold α is a family of fuzzy clusters which are elements of the initial allotment $R^\alpha_I(X)$ for the value of α and the family of fuzzy clusters should satisfy the conditions (9) and (10). So, the construction of adequate allotments $R^\alpha_z(X)=\{A^l_{(\alpha)} \mid l=\overline{1,c}, c \le n\}$ for every α is a trivial problem of combinatorics.

Detection of fixed c number of fuzzy clusters can be considered as the aim of classification. So, the adequate allotment $R^\alpha_z(X)$ is any allotment among c fuzzy clusters in the case. Several adequate allotments can exist. Thus, the problem consists in the selection of the unique adequate allotment $R^*(X)$ from the set $B(c)=\{R^\alpha_z(X)\}$ of adequate allotments, which is the class of possible solutions of the concrete classification problem and

$B(c) = \{R_z^\alpha(X)\}$ depends on the parameters the classification problem. The selection of the unique adequate allotment $R^*(X)$ from the set $B(c) = \{R_z^\alpha(X)\}$ of adequate allotments must be made on the basis of evaluation of allotments. The criterion

$$F_1(R_z^\alpha(X), \alpha) = \sum_{l=1}^{c} \frac{1}{n_l} \sum_{i=1}^{n_l} \mu_{li} - \alpha \cdot c, \quad (11)$$

where c is the number of fuzzy clusters in the allotment $R_z^\alpha(X)$ and $n_l = card(A_\alpha^l), A_{(\alpha)}^l \in R_z^\alpha(X)$ is the number of elements in the support of the fuzzy cluster $A_{(\alpha)}^l$, can be used for evaluation of allotments. The criterion

$$F_2(R_z^\alpha(X), \alpha) = \sum_{l=1}^{c} \sum_{i=1}^{n_l} (\mu_{li} - \alpha), \quad (12)$$

can also be used for evaluation of allotments. Both criteria were proposed in [11].

Maximum of criterion (11) or criterion (12) corresponds to the best allotment of objects among c fuzzy clusters. So, the classification problem can be characterized formally as determination of the solution $R^*(X)$ satisfying

$$R^*(X) = \arg \max_{R_z^\alpha(X) \in B(c)} F(R_z^\alpha(X), \alpha), \quad (13)$$

where $B(c) = \{R_z^\alpha(X)\}$ is the set of adequate allotments and criteria (11) and (12) are denoted by $F(R_z^\alpha(X), \alpha)$. The condition (13) must be met for the some unique allotment $R_z^\alpha(X) \in B(c)$. Otherwise, the number c of fuzzy clusters in the allotment sought $R^*(X)$ is suboptimal [12].

2.2. General Plan of the D-AFC(c)-algorithm

Detection of fixed c number of fuzzy clusters is the aim of classification. There is the $D\text{-}AFC(c)$-algorithm [16], [17]:

1. Calculate α-level values of the fuzzy tolerance T and construct the sequence $0 < \alpha_0 < \alpha_1 < \ldots < \alpha_\ell < \ldots < \alpha_Z \le 1$ of α-levels;
2. Construct the initial allotment $R_I^\alpha(X) = \{A_{(\alpha)}^l \mid l = \overline{1,n}\}, \alpha = \alpha_\ell$ for every value α_ℓ from the sequence $0 < \alpha_0 < \alpha_1 < \ldots < \alpha_\ell < \ldots < \alpha_Z \le 1$;
3. Let $w := 0$;
4. Construct allotments $R_z^\alpha(X) = \{A_{(\alpha)}^l \mid l = \overline{1,c}\}, \alpha = \alpha_\ell$, which satisfy conditions (9) and (10) for every value α_ℓ from the sequence $0 < \alpha_0 < \alpha_1 < \ldots < \alpha_\ell < \ldots < \alpha_Z \le 1$;
5. Construct the class of possible solutions of the classification problem $B(c) = \{R_z^\alpha(X)\}$, $\alpha \in \{\alpha_1, \ldots, \alpha_Z\}$ for the given number of fuzzy clusters c and different values of the tolerance threshold α as follows:

   if for some allotment $R_z^\alpha(X), \alpha \in \{\alpha_1, \ldots, \alpha_Z\}$ the condition $card(R_z^\alpha(X)) = c$ is met

   then $R_z^\alpha(X) \in B(c)$

   else let $w := w + 1$ and go to step 4;
6. Calculate the value of some criterion $F(R_z^\alpha(X), \alpha)$ for every allotment $R_z^\alpha(X) \in B(c)$;
7. The result $R^*(X)$ of classification is formed as follows:

if for some unique allotment $R_z^\alpha(X)$ from the set $B(c)$ the condition (13) is met
then the allotment is the result of classification
else the number c of classes is suboptimal.

The allotment $R_z^\alpha(X) = \{A_{(\alpha)}^l \mid l = \overline{1,c}\}$ among the given number c of fuzzy clusters and the corresponding value of tolerance threshold α are the results of classification. The value of α is very important for the interpretation of results from the epistemological position.

3. A Method of the Fuzzy Data Clustering

The first subsection of the section includes a short consideration of definitions of fuzzy intervals and fuzzy numbers. Distances for fuzzy numbers and distances for fuzzy intervals are considered in the second subsection. A method of the fuzzy data preprocessing is considered in the third subsection of the section.

3.1. Fuzzy Intervals and Fuzzy Numbers

Fuzzy data is easy to be found in natural language, psychometrics, biology, econometrics, social science and some other domains.

Usually, LR-type fuzzy intervals and LR-type fuzzy numbers are used to represent fuzzy data. So, the concept of a LR-type fuzzy interval and the concept of a LR-type fuzzy number must be defined in the first place. These concepts were considered by Bandemer and Näther [1].

Let L or R be decreasing, shape functions from $\Re^+$ to $[0,1]$ with $L(0)=1$ and $\forall x > 0, L(x) < 1, \forall x < 1, L(x) > 0$; $L(1) = 0$ or $L(x) > 0, \forall x$ and $L(+\infty) = 0$. Then a fuzzy set V is called a LR-type fuzzy interval $V = (\underline{m}, \overline{m}, a, b)_{LR}$ with $a > 0$, $b > 0$ if a membership function $\mu_V(x)$ of V is defined as

$$\mu_V(x) = \begin{cases} L\left(\dfrac{\underline{m} - x}{a}\right), & x \le \underline{m}, \\ 1, & \underline{m} \le x \le \overline{m}, \\ R\left(\dfrac{x - \overline{m}}{b}\right), & x \ge \overline{m}, \end{cases} \tag{14}$$

where $\underline{m}$ is called the lower mean value of V and $\overline{m}$ is called the upper mean value of V. Parameters a and b are called the left and right spreads, respectively.

For a LR-type fuzzy interval $V = (\underline{m}, \overline{m}, a, b)_{LR}$, if L and R are of the form

$$T(x) = \begin{cases} 1 - x, & 0 \le x \le 1, \\ 0, & otherwise \end{cases} \tag{15}$$

then V is called a trapezoidal fuzzy interval. The trapezoidal fuzzy interval will be denoted by $V = (\underline{m}, \overline{m}, a, b)_{TI}$ and its membership function is defined as follows:

$$\mu_V(x) = \begin{cases} 1 - \dfrac{\underline{m} - x}{a}, & for\ x \le \underline{m}, \\ 1, & for\ \overline{m} \le x \le \underline{m}, \\ 1 - \dfrac{x - \overline{m}}{b}, & for\ x \ge \overline{m}, \end{cases} \tag{16}$$

Let $V = (\underline{m}, \overline{m}, a, b)_{LR}$ be a LR-type fuzzy interval. If a condition $\underline{m} = \overline{m} = m$ is met, then a LR-type fuzzy interval V is called a LR-type fuzzy number and its membership function is defined as

$$\mu_V(x) = \begin{cases} L\left(\frac{m-x}{a}\right), \ for \ x \le m, \\ R\left(\frac{x-m}{b}\right), \ for \ x \ge m, \end{cases}, \tag{17}$$

where m is called the mean value of V and a and b are called the left and right spreads. Symbolically, a fuzzy number of LR-type is denoted by $V = (m, a, b)_{LR}$.

In LR-type fuzzy numbers, the triangular and Gaussian fuzzy numbers are most commonly used. In particular, for a LR-type fuzzy number $V = (m, a, b)_{LR}$ if L and R are of the form

$$T(x) = \begin{cases} 1 - x, \ 0 \le x \le 1, \\ 0, \quad otherwise, \end{cases} \tag{18}$$

then V is called a triangular fuzzy number, denoted by $V = (m, a, b)_T$ and its membership function is defined as

$$\mu_V(x) = \begin{cases} 1 - \frac{m-x}{a}, \ for \ x \le m, (a > 0) \\ 1 - \frac{x-m}{b}, \ for \ x \ge m, (b > 0) \end{cases}. \tag{19}$$

Let us consider a definition of the Gaussian fuzzy numbers. If $L(x) = R(x) = \exp\left(-((x-m)/\sigma)^2\right)$ for a LR-type fuzzy number $V = (m, a, b)_{LR}$, then V is called a Gaussian fuzzy number, denoted by $V = (m, \sigma)_G$. A membership function of a Gaussian fuzzy number $V = (m, \sigma)_G$ is defined as

$$\mu_V(x) = \exp\left(-\frac{(x-m)^2}{\sigma^2}\right), \ for \ -\infty < x < \infty. \tag{20}$$

Triangular fuzzy numbers can be considered as a special kind of trapezoidal fuzzy intervals. Moreover, trapezoidal fuzzy intervals are called trapezoidal fuzzy numbers and LR-type fuzzy intervals are called tolerant fuzzy numbers sometimes.

3.2. Distances for Fuzzy Numbers

Let us consider distances for fuzzy intervals and fuzzy numbers. These distances were proposed by Yang and Ko [18]. A method of the fuzzy data preprocessing is based on these distances.

The set of all LR-type fuzzy intervals will be denoted by $\mathcal{F}_{(LR)FI}(\Re)$ and a set of p fuzzy intervals in $\mathcal{F}_{(LR)FI}(\Re)$ will be denoted by $X_{(LR)FI} = \{V_1, \ldots, V_p\}$. A distance $d^2_{(LR)FI}(V_i, V_j)$ for any $V_i = (\underline{m}_i, \overline{m}_i, a_i, b_i)_{LR}$ and $V_j = (\underline{m}_j, \overline{m}_j, a_j, b_j)_{LR}$ in $\mathcal{F}_{(LR)FI}(\Re)$ is defined as follows:

$$\begin{aligned} d^2_{(LR)FI}(V_i, V_j) &= (\underline{m}_i - \underline{m}_j)^2 + (\overline{m}_i - \overline{m}_j)^2 + \\ &+ \left((\underline{m}_i - la_i) - (\underline{m}_j - la_j)\right)^2 + \left((\overline{m}_i + rb_i) - (\overline{m}_j + rb_j)\right)^2, \end{aligned} \tag{21}$$

where $l = \int_0^1 L^{-1}(\omega)d\omega$ and $r = \int_0^1 R^{-1}(\omega)d\omega$.

Let $\mathcal{F}_{(T)FI}(\Re)$ be a space of all trapezoidal fuzzy intervals and $X_{(T)FI} = \{V_1, \ldots, V_p\}$ be a set of p trapezoidal fuzzy intervals in $\mathcal{F}_{(T)FI}(\Re)$. According to the distance $d^2_{(LR)FI}(V_i, V_j)$ considered before, a distance $d^2_{(T)FI}(V_i, V_j)$ for any two trapezoidal fuzzy intervals $V_i = (\underline{m}_i, \overline{m}_i, a_i, b_i)_{TI}$ and $V_j = (\underline{m}_j, \overline{m}_j, a_j, b_j)_{TI}$ can be defined as follows:

$$d^2_{(T)FI}(V_i,V_j) = (\underline{m}_i - \underline{m}_j)^2 + (\overline{m}_i - \overline{m}_j)^2 + \\ + \left((\underline{m}_i - \frac{1}{2}a_i) - (\underline{m}_j - \frac{1}{2}a_j) \right)^2 + \left((\overline{m}_i + \frac{1}{2}b_i) - (\overline{m}_j + \frac{1}{2}b_j) \right)^2 . \quad (22)$$

Let $\mathcal{F}_{(LR)FN}(\Re)$ denote the set of all LR-type fuzzy numbers and $X_{(LR)FN} = \{V_1,\ldots,V_p\}$ be a set of p fuzzy numbers in $\mathcal{F}_{(LR)FN}(\Re)$. A distance $d^2_{(LR)FN}(V_i,V_j)$ for any $V_i = (m_i,a_i,b_i)_{LR}$ and $V_j = (m_j,a_j,b_j)_{LR}$, $V_i,V_j \in \mathcal{F}_{(LR)FN}(\Re)$ can be defined as follows:

$$d^2_{(LR)FN}(V_i,V_j) = (m_i - m_j)^2 + \\ + \left((m_i - la_i) - (m_j - la_j)\right)^2 + \left((m_i + rb_i) - (m_j + rb_j)\right)^2 , \quad (23)$$

where $l = \int_0^1 L^{-1}(\omega)d\omega$ and $r = \int_0^1 R^{-1}(\omega)d\omega$.

Based on the distance $d^2_{(T)FI}(V_i,V_j)$ defined on $\mathcal{F}_{(T)FI}(\Re)$ before, a distance $d^2_{(T)FN}(V_i,V_j)$ for any two triangular fuzzy numbers $V_i = (m_i,a_i,b_i)_T$ and $V_j = (m_j,a_j,b_j)_T$ in the space $\mathcal{F}_{(T)FN}(\Re)$ of all triangular fuzzy numbers can be defined as follows:

$$d^2_{(T)FN}(V_i,V_j) = (m_i - m_j)^2 + \left((m_i - m_j) - \frac{1}{2}(a_i - a_j) \right)^2 + \\ + \left((m_i - m_j) + \frac{1}{2}(b_i - b_j) \right)^2 . \quad (24)$$

Notable, that $X_{(T)FN} = \{V_1,\ldots,V_p\}$ is a set of p triangular fuzzy numbers in $\mathcal{F}_{(T)FN}(\Re)$.

Now let $\mathcal{F}_{(G)FN}(\Re)$ be the set of all Gaussian fuzzy numbers. Let $X_{(G)FN} = \{V_1,\ldots,V_p\}$ be a set of p Gaussian fuzzy numbers in $\mathcal{F}_{(G)FN}(\Re)$. Then a distance $d^2_{(G)FN}(V_i,V_j)$ for any two Gaussian fuzzy numbers $V_i = (m_i,\sigma_i)_G$ and $V_j = (m_j,\sigma_j)_G$ in $\mathcal{F}_{(G)FN}(\Re)$ is defined as follows:

$$d^2_{(G)FN}(V_i,V_j) = 3(m_i - m_j)^2 + \frac{\pi}{2}(\sigma_i - \sigma_j)^2 . \quad (25)$$

Notable, that distances (21) – (25) are metrics in corresponding spaces. The fact was demonstrated in [18].

3.3. The Fuzzy Data Preprocessing

In general, the data can be presented as a matrix of attributes $X_{n \times p} = [x_i^t]$, $i = 1,\ldots,n$, $t = 1,\ldots,p$, where the value x_i^t is the value of the t-th attribute for i-th object. However, we often have to deal with objects that cannot be described by precise values of attributes. So, a set $X_{(LR)FI} = \{V^1,\ldots,V^p\}$ of LR-type fuzzy intervals, a set $X_{(T)FI} = \{V^1,\ldots,V^p\}$ of trapezoidal fuzzy intervals, a set $X_{(LR)FN} = \{V^1,\ldots,V^p\}$ of LR-type fuzzy numbers, a set $X_{(T)FN} = \{V^1,\ldots,V^p\}$ of triangular fuzzy numbers and a set $X_{(G)FN} = \{V^1,\ldots,V^p\}$ of Gaussian fuzzy numbers can be considered as sets of values of objects attributes.

Thus, some fuzzy number or some fuzzy interval V_i^t, $i \in \{1,\ldots,n\}$, $t \in \{1,\ldots,p\}$ is the value of the t-th fuzzy attribute for i-th object. In the context of this approach, a concept of a vector of fuzzy numbers can be defined as follows:

Definition 6. *A crisp set $\overline{V} = \{V^t \mid t = 1,\ldots,p\}$ of p fuzzy numbers of the same type is the vector of fuzzy numbers.*

A kind of a vector of fuzzy numbers depends on the kind of its elements. In other words, fuzzy numbers of the same type can be elements of the vector $\overline{V}$. In particular, if a fuzzy number V^t is a fuzzy number of LR-type for all $t = 1,\ldots,p$, then the vector $\overline{V} = \{V^t \mid t = 1,\ldots,p\}$ is the vector of fuzzy numbers of LR-type. Notable, that the definition 6 is general definition, because fuzzy intervals of the same type can be elements of a vector $\overline{V}$, as well.

From other hand, a concept of a fuzzy vector was considered by Bandemer and Näther in [1]. A fuzzy set A on p-dimensional vector space $\Re^p$ is called a fuzzy vector if A is convex and there exists one point $y \in \Re^p$ with membership function $\mu_A(y) = 1$. Obviously, the definition of the vector of fuzzy numbers is different from the definition of the fuzzy vector.

Let $X = \{x_1,...,x_n\}$ be the initial set of elements and a vector of fuzzy numbers $\overline{V_i}$, $i = 1,\ldots,n$ corresponds to each object $x_i \in X$. In general, a distance between different objects $x_i = (V_i^1,\ldots,V_i^p)$ and $x_j = (V_j^1,\ldots,V_j^p)$ can be defined as an average of the sum of distances between attributes:

$$D(x_i, x_j) = \frac{1}{p}\sum_{t=1}^{p} d^2(V_i^t, V_j^t), \tag{26}$$

where $d^2(V_i^t, V_j^t)$ is a distance for two fuzzy numbers which represents values of the same t-th fuzzy attribute for different objects $x_i, x_j \in X$.

The distance (26) is a linear combination of p distances between attributes. So, the distance depends on the kind of fuzzy numbers which represents objects attributes. For example, if objects attributes represented by triangular fuzzy numbers $V_i^t = (m_i^t, a_i^t, b_i^t)_T$, $i = 1,\ldots,n$, $t = 1,\ldots,p$, then a distance (26) can be rewritten as follows:

$$D(x_i, x_j) = \frac{1}{p}\sum_{t=1}^{p} d^2_{(T)FN}(V_i^t, V_j^t), \tag{27}$$

where $d^2_{(T)FN}(V_i^t, V_j^t)$ is the distance for triangular fuzzy numbers (24):

$$\begin{aligned} d^2_{(T)FN}(V_i^t, V_j^t) &= (m_i^t - m_j^t)^2 + \left((m_i^t - m_j^t) - \frac{1}{2}(a_i^t - a_j^t)\right)^2 + \\ &+ \left((m_i^t - m_j^t) + \frac{1}{2}(b_i^t - b_j^t)\right)^2 \end{aligned} \tag{28}$$

for all $i, j = 1,\ldots,n$, $t = 1,\ldots,p$.

After application of a distance (26) to the data set $X = \{x_1,...,x_n\}$ a matrix of coefficients of pair wise dissimilarity between objects $d_{n\times n} = [d_{ij}],\ i, j = 1,\ldots,n$ can be obtained. Fuzzy tolerance matrix can be obtained as follows. In the first place, a matrix of fuzzy intolerance $I = [\mu_I(x_i, x_j)],\ i, j = 1,\ldots,n$ must be obtained for construction of the fuzzy tolerance matrix. For the purpose, the matrix of dissimilarity coefficients should be normalized as follows:

$$\mu_I(x_i, x_j) = \frac{d_{ij}}{\max\limits_{i,j} d_{ij}}, \tag{29}$$

where $d_{ij}, \forall i, j = 1,\ldots,n$ are dissimilarity coefficients. The matrix of fuzzy tolerance $T = [\mu_T(x_i, x_j)],\ i, j = 1,\ldots,n$ can be obtained after application of complement operation

$$\mu_T(x_i, x_j) = 1 - \mu_I(x_i, x_j),\ \forall i, j = 1,\ldots,n \tag{30}$$

to the matrix of fuzzy intolerance $I = [\mu_I(x_i, x_j)],\ i, j = 1,\ldots,n$. The matrix of fuzzy tolerance $T = [\mu_T(x_i, x_j)],\ i, j = 1,\ldots,n$ is the matrix of initial data for the $D\!-\!AFC(c)$-algorithm.

4. An Illustrative Example

The first subsection includes the data set description. Results of processing of these data by the $D-AFC(c)$-algorithm are presented and discussed in the second subsection of the section.

4.1. The Data

Let us consider an application of the $D-AFC(c)$-algorithm to the fuzzy data classification problem. For the purpose, the Yang and Ko's data set of thirty triangular fuzzy numbers [18] were modified. The artificial data set is presented in Table 1.

Numbers of objects, i	Attributes								
	$V^1=(m^1,a^1,b^1)_T$			$V^2=(m^2,a^2,b^2)_T$			$V^3=(m^3,a^3,b^3)_T$		
	m^1	a^1	b^1	m^2	a^2	b^2	m^3	a^3	b^3
1	3.34	1.46	1.30	19.78	1.47	0.42	32.77	0.63	0.47
2	9.56	0.27	1.00	20.67	1.34	1.10	34.88	1.08	0.66
3	10.56	1.95	1.93	21.45	0.92	1.60	35.45	1.48	1.26
4	10.89	0.56	1.17	22.34	0.04	1.58	35.88	1.79	0.16
5	13.89	0.89	0.88	23.47	0.81	0.51	38.88	0.66	0.64
6	14.78	0.12	1.21	24.67	0.14	1.09	40.25	0.52	1.71
7	14.90	1.19	0.41	25.78	0.39	1.51	40.47	1.95	0.15
8	15.67	1.82	0.90	26.45	1.61	0.92	43.56	0.92	0.63
9	16.87	1.90	1.85	28.34	1.95	0.12	43.98	1.74	1.69
10	17.45	1.79	1.95	32.29	1.66	1.64	45.77	1.71	0.79

Table 1. The data set for numerical experiments.

So, the vector of triangular fuzzy numbers $\overline{V_i}=(V_i^1,V_i^2,V_i^3)$ corresponds to each element of the set of objects $X=\{x_1,...,x_{10}\}$. A matrix of the fuzzy tolerance $T=[\mu_T(x_i,x_j)]$, $i,j=1,\dots,10$ was obtained after an application of the formulae (27) – (30) to the data set. The $D-AFC(c)$-algorithm can be applied directly to the matrix of tolerance coefficients.

4.2. Experimental Results

Intuitively, the number of fuzzy clusters $c=2$ is suitable for the data set $X=\{x_1,...,x_{10}\}$ from Table 1. From other hand, the triangular fuzzy number $V_1^1=(m_1^1,a_1^1,b_1^1)_T$ is far away from the other triangular fuzzy numbers $V_i^1=(m_i^1,a_i^1,b_i^1)_T$, $i=2,\dots,10$. That is why the first object can be regarded as an element of a separate class. That is why the number of fuzzy clusters $c=3$ in the allotment sought should be taken into account in experiments. So, the data was processed by the $D-AFC(c)$-algorithm for $c=2$ and $c=3$.

Let us consider every experiment in detail. In the first place, the $D-AFC(c)$-algorithm was applied to the data with the number of classes $c=2$. The allotment $R^*(X)$ among two fully separate fuzzy clusters, which corresponds to the result, is obtained for the tolerance threshold $\alpha=0.706096$. The value of the membership function of the fuzzy cluster, which corresponds to the first class is maximal for the first object and is equal one. So, the first object is the typical point of the fuzzy cluster which corresponds to the first class. The membership value of the tenth object is equal one and the value is maximal for the fuzzy cluster which corresponds to the second class. That is why the tenth object is the typical point of the fuzzy cluster which corresponds to the second class. Membership functions of two classes of the allotment are presented in Figure 1 and values which equal zero are not shown in the figure. Membership values of the first class are represented by ○ and membership values of the second class are represented by ■.

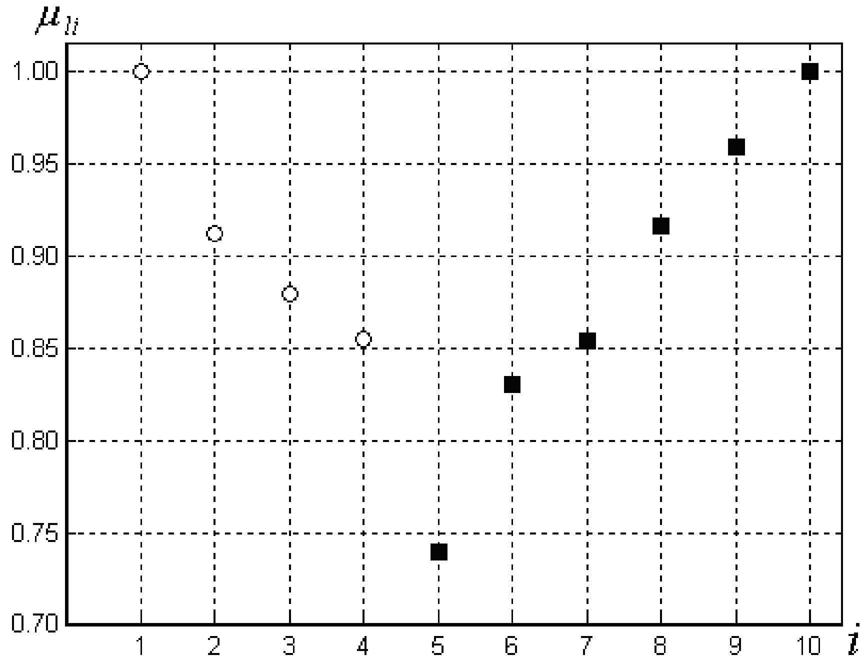

Figure 1. Membership functions of two classes obtained from the D-AFC(c)-algorithm.

By executing the $D-AFC(c)$-algorithm for three classes, we obtain that the first class is formed by a unique element, the second class is composed of four elements and the third class is formed by five elements. Membership functions of three classes of the allotment are presented in Figure 2.

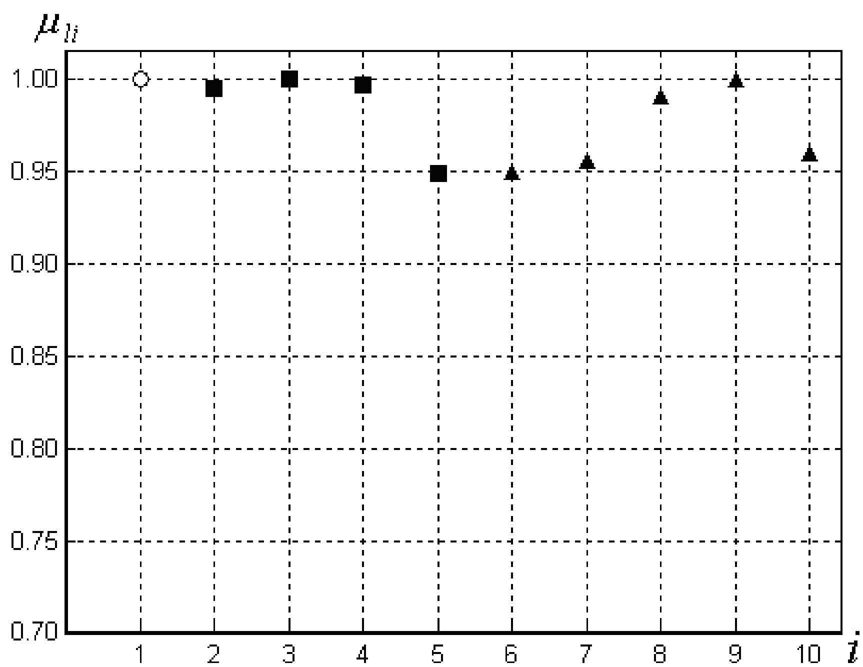

Figure 2. Membership functions of three classes obtained from the D-AFC(c)-algorithm.

The allotment $R^*(X)$ among three fully separate fuzzy clusters, which is corresponds to the result, was obtained for the tolerance threshold $\alpha = 0.915996$. Membership values of the first class are represented in Figure 2 by ○, membership values of the second class are represented in Figure 2 by ■, and membership values of the third class are represented in Figure 2 by ▲.

The value of the membership function of the fuzzy cluster, which corresponds to the first class is maximal for the first object and is equal one. So, the first object is the typical point of the fuzzy cluster which corresponds to the first class. The membership value of the third object is equal one and the value is maximal for the fuzzy cluster which corresponds to the second class. Thus, the third object is the typical point of the fuzzy cluster which corresponds to the second class. The membership function of the third fuzzy cluster is maximal for the ninth object and is equal one. That is why the ninth object is the typical point of the fuzzy cluster which corresponds to the third class.

As can be seen from Figure 1 and Figure 2, the membership function obtained from the $D{-}AFC(c)$-algorithm for $c = 3$ is sharper than the membership function obtained for $c = 2$. Moreover, the first object is the unique element of the first class. That is why the first element can be considered as an outlier.

Three fuzzy clusters of the allotment $R^*(X)$ are compact and well-separated fuzzy clusters. As can be seen from Table 1 and Figure 2, the allotment $R^*(X)$ among three fuzzy clusters can be very well interpreted from essential positions. So, the allotment among three fuzzy clusters is the appropriate result of classification.

5. Concluding Remarks

Preliminary conclusions are discussed in the first subsection. The second subsection deals with the perspectives on future investigations.

5.1. Discussion

The results of application of the possibilistic clustering method based on the allotment concept can be very well interpreted. Moreover, the clustering method based on the allotment concept depends on the set of adequate allotments only. That is why the clustering results are stable.

The methodology of heuristic possibilistic clustering of the fuzzy data is outlined in the paper. The approach is based on the concept of the vector of fuzzy numbers. A matrix of dissimilarity coefficients between vectors of fuzzy numbers can be constructed using the proposed distance and a matrix of fuzzy tolerance can be obtained.

The results of numerical experiments seem to be satisfactory. The results of application of the proposed methodology for the fuzzy data preprocessing and its processing by the $D{-}AFC(c)$-algorithm to the set of vectors of triangular fuzzy numbers show that the methodology and the $D{-}AFC(c)$-algorithm are a precise and effective technique for the fuzzy data possibilistic clustering.

5.2. Perspectives

In the first place, the $D{-}AFC(c)$-algorithm is the basic version of the clustering procedure. Other parameters of a clustering procedure were introduced by Viattchenin [16]. Moreover, a heuristic for the detection of an unknown number of fuzzy clusters in the sought allotment was also proposed in [16]. So, the corresponding versions of the algorithm can be developed.

In the second place, fuzzy numbers of the same type can be elements of the vector of fuzzy numbers. So, vectors of fuzzy numbers in the sense of definition 6 can be called homogeneous vectors of fuzzy numbers. A concept of a heterogeneous vector of fuzzy numbers can be introduced. Thus, LR-type fuzzy intervals, trapezoidal fuzzy intervals, LR-

type fuzzy numbers, triangular fuzzy numbers and Gaussian fuzzy numbers can be elements of some heterogeneous vector of fuzzy numbers. So, the described approach of the fuzzy data clustering can be generalized for a case of heterogeneous vectors of fuzzy numbers.

These perspectives for investigations are of great interest both from the theoretical point of view and from the practical one as well.

Acknowledgements

I would like to thank the anonymous referees and Prof. Neven Vrcek, Editor, for their constructive reviews and comments. I also thank the Director of the Systems Research Institute of the Polish Academy of Sciences for the possibility of conducting the investigations and the Mianowski Fund for financial support. Partial support by the Presidium of the National Academy of Sciences of Belarus under Grant 21 is also acknowledged. I am grateful to Prof. Janusz Kacprzyk and Dr. Jan W. Owsinski for their interest in my investigations and fruitful discussions during the paper preparation. I also thank Mr. Aliaksandr Damaratski for elaborating the experimental software.

References

[1] Bandemer, H; Näther, W. *Fuzzy data analysis*. Kluwer Academic Publishers, Dordrecht, 1992.

[2] Bezdek, J.C. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York, 1981.

[3] Butkiewicz, B; Nieradka, G. Fuzzy clustering for fuzzy features. In De Baets, B; De Tré, G; Fodor, J; Kacprzyk, J; Zadrozny, S, editors. *Current Issues in Data and Knowledge Engineering*, EXIT, Warszawa, pages 102-107, 2004.

[4] Chiang, J.-H; Yue, S; Yin, Z.-X. A new fuzzy cover approach to clustering. *IEEE Transactions on Fuzzy Systems*, 12(2):199-208, 2004.

[5] Höppner, F; Klawonn, F; Kruse, R; Runkler, T. *Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition*. John Wiley & Sons, Chichester, 1999.

[6] Hung, W.-L; Yang, M.-S. Fuzzy clustering on LR-type fuzzy numbers with an application in Taiwanese tea evaluation. *Fuzzy Sets and Systems*, 150(3):561-577, 2005.

[7] Krishnapuram, R; Keller, J.M. A possibilistic approach to clustering. *IEEE Transactions on Fuzzy Systems*, 1(2):98-110, 1993.

[8] Łęski, J.M. Robust possibilistic clustering. *Archives of Control Sciences*, 10(3/4):141-155, 2000.

[9] Mandel, I.D. *Clustering Analysis*. Finansy i Statistica, Moscow, 1988. (in Russian)

[10] Miyamoto, S. *Introduction to Cluster Analysis: Theory and Applications of Fuzzy Clustering*. Morikita-Shuppan, Tokyo, 1990. (in Japanese)

[11] Viattchenin, D.A. Criteria of quality of allotment in fuzzy clustering. In *Proceedings of the Third International Conference on Neural Networks and Artificial Intelligence*, pages 91-94, Minsk, Belarus, 2003.

[12] Viattchenin, D.A. On the number of fuzzy clusters in the allotment. In *Proceedings of the Seventh International Conference on Computer Data Analysis and Modeling*, vol. 1, pages 198-201, Minsk, Belarus, 2004.

[13] Viattchenin, D.A. A new heuristic algorithm of fuzzy clustering. *Control & Cybernetics*, 33(2):323-340, 2004.

[14] Viattchenin, D.A. *Fuzzy Methods of Automatic Classification*. Technoprint Publishing House, Minsk, 2004. (in Russian)

[15] Viattchenin, D.A. A proximity-based fuzzy clustering of fuzzy numbers. In *Proceedings of the Eighth International Conference on Computer Data Analysis and Modeling*, vol. 1, pages 182-185, Minsk, Belarus, 2007.

[16] Viattchenin, D.A. An outline for a heuristic approach to possibilistic clustering. Working Paper WP-2-2007, Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland, 2007.

[17] Viattchenin, D.A. A direct algorithm of possibilistic clustering with partial supervision. *Journal of Automation, Mobile Robotics and Intelligent Systems*, 1(3):29-38, 2007.

[18] Yang, M.-S; Ko, C.-H. On a class of fuzzy C-numbers clustering procedures for fuzzy data. *Fuzzy Sets and Systems*, 84(1):49-60, 1996.

[19] Yang, M.-S; Liu, H.-H. Fuzzy clustering procedures for conical fuzzy vector data. *Fuzzy Sets and Systems*, 106(2):189-200, 1999.

[20] Yang, M.-S; Wu, K.-L. Unsupervised possibilistic clustering. *Pattern Recognition*, 39(1):5-21, 2006.