Communications on Applied Electronics (CAE) – ISSN : 2394-4714
Foundation of Computer Science FCS, New York, USA
Volume 6 – No.2, November 2016 – www.caeaccess.org

Estimation of Bounds of the Set of Potential Number of Fuzzy Clusters in a Sought Clustering Structure

Dmitri A. Viattchenin
United Institute of Informatics Problems of the NAS of Belarus
Minsk, Belarus

Aliaksandr Yaroma
Belarussian State University of Informatics and Radioelectronics
Minsk, Belarus

Aliaksandr Damaratski
United Institute of Informatics Problems of the NAS of Belarus
Minsk, Belarus

ABSTRACT
In this paper, an approach to constructing the set of values of the most possible number of fuzzy clusters in a sought clustering structure is proposed. The proposed approach is based on heuristic possibilistic clustering and fuzzy numbers. For the purpose, fuzzy numbers are described and algorithms of the heuristic approach to possibilistic clustering are considered in brief. A procedure for constructing the set of values of the most possible number of fuzzy clusters is described for the object data set. An application of the proposed technique to the Anderson's iris data set is provided and some concluding remarks are stated.

General Terms
Pattern Recognition, Cluster Validity, Possibility Distribution.

Keywords
Triangular Fuzzy Number, Gaussian Fuzzy Number, Cluster Validity, Heuristic Possibilistic Clustering, Tolerance Threshold.

1. INTRODUCTION
Some notes on fuzzy approach to cluster analysis are presented in the first subsection of the section. A cluster validity problem is considered in the second subsection.

1.1 A Note on Fuzzy Clustering
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. Fuzzy sets theory, which was proposed by Zadeh [1], gives an idea of uncertainty of belonging to a cluster, which is described by a membership function. Fuzzy clustering methods have been applied effectively in image processing, data analysis, symbol recognition and modeling. Heuristic methods of fuzzy clustering, hierarchical methods of fuzzy clustering and optimization methods of fuzzy clustering were proposed by different researchers.

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 c-partition. Objective function-based fuzzy clustering algorithms can in general be divided into two types: object versus relational. The object data clustering methods can be applied if the objects are represented as points in some multidimensional space. The best known optimization approach to fuzzy clustering is the method of fuzzy c-means, developed by Bezdek [2]. The FCM-algorithm is based on an iterative optimization of the fuzzy objective function, which takes the form:

$$Q_{FCM}(P,\mathrm{T}) = \sum_{l=1}^{c}\sum_{i=1}^{n} u_{li}^{\gamma} \left\| x_i - \tau^l \right\|^2, \quad (1)$$

subject to constraints regarding u_{li}

$$\sum_{l=1}^{c} u_{li} = 1,\ 0 \le u_{li} \le 1, \quad (2)$$

where u_{li}, $l = 1,\ldots,c$, $i = 1,\ldots,n$ is the membership degree, x_i, $i \in \{1,\ldots,n\}$ is the data point, $\mathrm{T} = \{\tau^1,\ldots,\tau^c\}$ is the set fuzzy clusters prototypes and $\gamma > 1$ is the weighting exponent. Note that the concept of fuzzy c-partition are defined by the conditions (2). So, the fuzzy c-partition can be arrayed as a $(c \times n)$ matrix $P = [u_{li}]$.

The FCM-algorithm is the basis of the family of fuzzy clustering algorithms. These objective function-based fuzzy clustering algorithms were proposed by different authors and they are described by Höppner, Klawonn, Kruse and Runkler [3] in detail.

However, the condition of fuzzy c-partition is very difficult from essential positions. So, a possibilistic approach to clustering was proposed by Krishnapuram and Keller in [4] and developed by other researchers. This approach can be considered as a way in the optimization approach in fuzzy clustering because major methods of possibilistic clustering are objective function-based methods.

A concept of possibilistic partition is a basis of possibilistic clustering methods and membership values μ_{li}, $l = 1,\ldots,c$, $i = 1,\ldots,n$ 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:

$$\sum_{l=1}^{c} \mu_{li} > 0,\ 0 \le \mu_{li} \le 1. \quad (3)$$

So, the family of fuzzy sets $\mathrm{Y}(X) = \{A^l \mid l = \overline{1,c},\ c \le n\}$ is the possibilistic partition of the initial set of objects $X = \{x_1,\ldots,x_n\}$ if condition (3) is met.

Obviously that the conditions of the possibilistic partition (3) are more flexible than the conditions of the fuzzy c-partition (2). It is should be noted, that a heuristic approach to possibilistic clustering is proposed in [5].

1.2 A Cluster Validity Problem

The most important problem of fuzzy clustering is neither the choice of the numerical procedure nor the distance to use but concerns the number c of fuzzy clusters to look for. Really, lacking in a priori knowledge of the data structure, there is no reason to choose a particular value of c and one must find a way to measure the acceptance with which cluster structure has been identified by a clustering procedure. This is the so-called cluster validity problem.

The classical approach to cluster validity for fuzzy clustering is based on directly evaluating the fuzzy c-partition. Measures of cluster validity can be used for the purpose. Many authors have proposed several measures of cluster validity associated with fuzzy c-partitions. The cluster validity problem can be illustrated by the method of fuzzy c-means. Various cluster validity indexes for the FCM-algorithm were proposed by different researchers. Among other measures proposed in the literature, the following global validity measures can be found [3]:

- Partition coefficient:

$$V_{PC}(P;c)=\frac{1}{n}\sum_{l=1}^{c}\sum_{i=1}^{n}u_{li}^{2}, \tag{4}$$

- Partition entropy:

$$V_{PE}(P;c)=-\frac{1}{n}\sum_{l=1}^{c}\sum_{i=1}^{n}\left|u_{li}\cdot\ln u_{li}\right|, \tag{5}$$

- Compactness and separation index:

$$V_{CS}(P;c)=\frac{\sum_{l=1}^{c}\sum_{i=1}^{n}u_{li}^{\gamma}\left\|x_i-\tau^{l}\right\|^{2}}{n\times\min\limits_{i\neq l}\left\|x_i-\tau^{l}\right\|^{2}}. \tag{6}$$

The number of clusters that minimizes $V_{PE}(P;c)$ and $V_{CS}(P;c)$ or maximizes $V_{PC}(P;c)$ is taken as the optimal number c of fuzzy clusters in the sought in the constructing fuzzy c-partition P.

So, for the determination of the numbers of clusters, $X=\{x_1,\ldots,x_n\}$ a following procedure [3] can be applied to the data set. An assumption, that some FCM-like clustering procedure and some corresponding validity measure $V_*(P;c)$ are taken into account, will be useful for the following consideration.

1. Set $c:=c_{\min}$ and $c_{opt}:=c$;
2. Run the clustering algorithm with the input $(X;c)$;
3. Calculate the value of a validity measure $V_*(P;c)$;
4. The following condition is checked:

   if the value of $V_*(P;c)$ is better than the value of $V_*(P;c_{opt})$ then $c_{opt}:=c$, $c:=c+1$ and go to step 2 else go to step 5;
5. The following condition is checked:

   if a condition $c_{opt}<c_{\max}$ is met, then go to step 2, else go to step 6;
6. The optimum cluster number is equal c_{opt} and stop.

Traditionally, a lower bound for the number of clusters $c_{\min}$ is equal to 2 and an upper bound for the number of clusters $c_{\max}$ is equal to $(n-1)$, where $n=card(X)$. For the large data sets, the determination of the actual number of fuzzy clusters is computationally very expensive because the number of running the clustering procedure is equal to $(n-2)$. So, the lower bound for the number of clusters $c_{\min}$ and the upper bound for the number of clusters $c_{\max}$ must be estimated and the set $\{c_{\min},\ldots,c_{\max}\}$ of most possible cluters in the sought clustering structure should be constructed.

The formulated problem can be solved by using the heuristic D-AFC-TAGA-algorithm of possibilistic clustering [6] and fuzzy numbers. For this purpose, a short consideration of heuristic algorithms of possibilistic clustering is presented, basic types of fuzzy numbers are considered, the general plan of the procedure for constructing the set of values of the most possible number of fuzzy clusters is proposed, an illustrative example is given and preliminary conclusions are formulated.

2. A BACKGROUND FOR THE PROPOSED APPROACH

Heuristic algorithms of possibilistic clustering are considered in brief in the first subsection of the section. The second subsection includes a consideration of basic types of fuzzy numbers. Methods of the data preprocessing are described in the third subsection of the section.

2.1 Heuristic Algorithms of Possibilistic Clustering: A Survey

A heuristic approach to possibilistic clustering is proposed in [5]. The essence of the heuristic approach to possibilistic clustering is that the sought clustering structure of the set of observations is formed based directly on the formal definition of fuzzy cluster and possibilistic memberships are determined also directly from the values of the pairwise similarity of observations. A concept of the allotment among fuzzy clusters is basic concept of the approach and the allotment among fuzzy clusters is a special case of the possibilistic partition (3).

Direct heuristic algorithms of possibilistic clustering can be divided into two types: relational versus prototype-based. A fuzzy tolerance relation matrix is a matrix of the initial data for the direct heuristic relational algorithms of possibilistic clustering and a matrix of attributes is a matrix for the prototype-based algorithms. In particular, the group of direct relational heuristic algorithms of possibilistic clustering includes:

- the D-AFC(c)-algorithm which is based on the construction of an allotment among an a priori given number c of partially separate fuzzy clusters [5];

- the D-PAFC-algorithm which is based on the construction of an principal allotment among an unknown minimal number of at least c fully separate fuzzy clusters [5];
- the D-AFC-PS(c)-algorithm which is based on the construction of an allotment among an a priori given number c of partially separate fuzzy clusters in the presence of labeled object [5];
- the D-AFC(u)-algorithm which is based on the construction of an allotment among an a priori unknown number c of partially separate fuzzy clusters with respect to the given maximal number of elements in every class [7].

On the other hand, the family of prototype-based heuristic algorithms of possibilistic clustering includes:

- the D-AFC-TC-algorithm which is based on the construction of an allotment among an a priori unknown number c of fully separate fuzzy clusters [5];
- the D-PAFC-TC-algorithm which is based on the construction of a principal allotment among an a priori unknown minimal number of at least c fully separate fuzzy clusters [5];
- the D-AFC-TC(α)-algorithm which is based on the construction of an allotment among an a priori unknown number c of fully separate fuzzy clusters with respect to the minimal value α of the tolerance threshold [5];
- the H-AFC-TC-algorithm which is based on the construction of an hierarchy of allotments among an a priori unknown number c of fully separate fuzzy clusters [5].

It should be noted that these direct prototype-based heuristic possibilistic clustering algorithms are based on a transitive closure of an initial fuzzy tolerance relation. On the other hand, a family of direct prototype-based heuristic possibilistic clustering algorithms based on a transitive approximation of a fuzzy tolerance is proposed in [6]. The family of clustering procedure is based on using the TAGA-algorithm [8]. So, the family of prototype-based algorithms includes:

- the D-AFC-TAGA-algorithm which is based on the construction of an allotment among an a priori unknown number c of fully separate fuzzy clusters;
- the D-PAFC-TAGA-algorithm which is based on the construction of a principal allotment among an a priori unknown minimal number of at least c fully separate fuzzy clusters;
- the D-AFC-TAGA(α)-algorithm which is based on the construction of an allotment among an a priori unknown number c of fully separate fuzzy clusters with respect to the minimal value α of the tolerance threshold.

All direct prototype-based heuristic possibilistic clustering algorithms based on a transitive closure of an initial fuzzy tolerance relation are particular versions of corresponding prototype-based heuristic possibilistic clustering algorithms which based on the calculation of a transitive approximation of a fuzzy tolerance.

2.2 Fuzzy Numbers

Fuzzy intervals and fuzzy numbers can be considered as a special kind of fuzzy sets. Fuzzy numbers are useful tool for constructing a possibility distribution in the formulated problem which will be considered below.

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 described, for example, in [5] and [9].

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

$$\mu_V(x)=\begin{cases} L\left(\dfrac{\underline{m}-x}{a}\right), & x\le \underline{m} \\ 1, & \underline{m}\le x\le \overline{m}, \\ R\left(\dfrac{x-\overline{m}}{b}\right), & x\ge \overline{m} \end{cases} \tag{7}$$

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

$$T(x)=\begin{cases} 1-x, & 0\le x\le 1 \\ 0, & otherwise \end{cases}, \tag{8}$$

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:

$$\mu_V(x)=\begin{cases} 1-\dfrac{\underline{m}-x}{a}, & x\le \underline{m} \\ 1, & \overline{m}\le x\le \underline{m}. \\ 1-\dfrac{x-\overline{m}}{b}, & x\ge \overline{m} \end{cases} \tag{9}$$

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

$$\mu_V(x)=\begin{cases} L\left(\dfrac{m-x}{a}\right), & x\le m \\ R\left(\dfrac{x-m}{b}\right), & x\ge m \end{cases}, \tag{10}$$

where m is called the mean value of V and a and b are called the left and right spreads. 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

$$T(x)=\begin{cases}1-x, 0\le x\le 1\\ 0, \quad otherwise\end{cases}, \tag{11}$$

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

$$\mu_V(x)=\begin{cases}1-\dfrac{m-x}{a}, x\le m\\ 1-\dfrac{x-m}{b}, x\ge m\end{cases}. \tag{12}$$

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

$$\mu_V(x)=\exp\left(-\frac{(x-m)^2}{\sigma^2}\right), -\infty<x<\infty. \tag{13}$$

2.3 A Note on the Data Preprocessing

Clustering algorithms can in general be divided into two types: object and relational. The object data clustering methods can be applied if the objects are represented as points in some multidimensional space $I^{m_1}(X)$. In other words, the data which is composed of n objects and m_1 attributes is denoted as $\hat{X}_{n\times m_1}=[\hat{x}_i^{t_1}]$, $i=1,\ldots,n$, $t_1=1,\ldots,m_1$ and the data are called sometimes the two-way data [10]. Let $X=\{x_1,\ldots,x_n\}$ is the set of objects. So, the two-way data matrix can be represented as follows:

$$\hat{X}_{n\times m_1}=\begin{pmatrix}\hat{x}_1^1 & \hat{x}_1^2 & \ldots & \hat{x}_1^{m_1}\\ \hat{x}_2^2 & \hat{x}_2^2 & \ldots & \hat{x}_2^{m_1}\\ \ldots & \ldots & \ldots & \ldots\\ \hat{x}_n^1 & \hat{x}_n^2 & \ldots & \hat{x}_n^{m_1}\end{pmatrix}. \tag{14}$$

So, the two-way data matrix can be represented as $\hat{X}=(\hat{x}^1,\ldots,\hat{x}^{m_1})$ using n-dimensional column vectors $\hat{x}^{t_1}$, $t_1=1,\ldots,m_1$, composed of the elements of the t_1-th column of $\hat{X}$.

In the relational approach to fuzzy clustering, the problem of the data classification is solved by expressing a relation which quantifies either similarity, or dissimilarity, between pairs of objects. So, the data matrix taken a form

$$\hat{\rho}_{n\times n}=\begin{pmatrix}\hat{\rho}_{11} & \hat{\rho}_{12} & \ldots & \hat{\rho}_{1n}\\ \hat{\rho}_{21} & \hat{\rho}_{22} & \ldots & \hat{\rho}_{2n}\\ \ldots & \ldots & \ldots & \ldots\\ \hat{\rho}_{n1} & \hat{\rho}_{n2} & \ldots & \hat{\rho}_{nn}\end{pmatrix}, \tag{15}$$

where a general notation $\hat{\rho}_{ij}$ used for designation of pair wise dissimilarities $d(x_i,x_j)$ or the similarity coefficients $r(x_i,x_j)$. In general, the values $\hat{\rho}_{ij}$ are not normalized.

Relational clustering procedures can be used with the two-way data (14), by choosing a suitable metric to measure similarity. Moreover, heuristic possibilistic relational clustering procedures can be used for the three-way data processing. The fact was shown in [11], where the corresponding dissimilarity measures were proposed.

In the first place, the two-way data can be normalized as follows:

$$x_i^{t_1}=\frac{\hat{x}_i^{t_1}}{\max\limits_i \hat{x}_i^{t_1}}. \tag{16}$$

In the second place, the two-way data can be normalized using a formula

$$x_i^{t_1}=\frac{\hat{x}_i^{t_1}-\min\limits_i \hat{x}_i^{t_1}}{\max\limits_i \hat{x}_i^{t_1}-\min\limits_i \hat{x}_i^{t_1}}. \tag{17}$$

So, each object can be considered as a fuzzy set x_i, $i=1,\ldots,n$ and $x_i^{t_1}=\mu_{x_i}(x^{t_1})\in[0,1]$, $i=1,\ldots,n$, $t_1=1,\ldots,m_1$ are their membership functions. The matrix of coefficients of pair wise dissimilarity between objects $I=[\mu_I(x_i,x_j)]$, $i,j=1,\ldots,n$ can be obtained after application of some distance function to the matrix of normalized data $X_{n\times m_1}=[\mu_{x_i}(x^{t_1})]$, $i=1,\ldots,n$, $t_1=1,\ldots,m_1$. The most widely used distances for fuzzy sets x_i, x_j, $i,j=1,\ldots,n$ in $X=\{x_1,\ldots,x_n\}$ are:

- the normalized Hamming distance:

$$l(x_i,x_j)=\frac{1}{m_1}\sum_{t_1=1}^{m_1}|\mu_{x_i}(x^{t_1})-\mu_{x_j}(x^{t_1})|, \tag{18}$$

- the normalized Euclidean distance:

$$e(x_i,x_j)=\sqrt{\frac{1}{m_1}\sum_{t_1=1}^{m_1}\left(\mu_{x_i}(x^{t_1})-\mu_{x_j}(x^{t_1})\right)^2}, \tag{19}$$

- the squared normalized Euclidean distance:

$$\varepsilon(x_i,x_j)=\frac{1}{m_1}\sum_{t_1=1}^{m_1}\left(\mu_{x_i}(x^{t_1})-\mu_{x_j}(x^{t_1})\right)^2. \tag{20}$$

These distances were considered by Kaufmann [12] in detail.

The matrix of fuzzy tolerance $T=[\mu_T(x_i,x_j)]$, $i,j=1,\ldots,n$ can be obtained after application of the complement operation

$$\mu_T(x_i,x_j)=1-\mu_I(x_i,x_j),\ i,j=1,\ldots,n \quad (21)$$

to the matrix of fuzzy intolerance $I=[\mu_I(x_i,x_j)]$, $i,j=1,\ldots,n$.

The complement operation (21) was also considered in detail by Zadeh in [1] and Kaufmann in [12].

3. A PROPOSED TECHNIQUE

A method for constructing triangular and Gaussian fuzzy numbers over the initial data set by using the D-AFC-TAGA-algorithm of possibilistic clustering is considered in the first subsection of the section. The second subsection includes a consideration of a technique for constructing the set of values of most possible number of fuzzy clusters in the sought clustering structure.

3.1 Constructing Fuzzy Numbers through Heuristic Possibilistic Clustering

The allotment $R_c^*(X)$ among either an a priori given or an unknown number c of fuzzy clusters and the value of tolerance threshold $\alpha\in(0,1]$ are principal results of classification obtained from all direct heuristic algorithms of possibilistic clustering. So, the value c and the value $\alpha\in(0,1]$ can be used for estimating the lower bound for the number of clusters $c_{\min}$ and the upper bound for the number of clusters $c_{\max}$. A method for constructing a triangular fuzzy number over the set of elements $X=\{x_1,\ldots,x_n\}$ of the initial data set should be considered in the first place.

Let $X=\{x_1,\ldots,x_n\}$ be the initial set of elements and c is the number of fuzzy clusters in the obtained allotment $R_c^*(X)$. So, a triangular fuzzy number $V=(m,a,b)_T$ can be constructed immediately and its membership function is defined by (12), where $m=c$, $1\le c\le n$, $a=c-1$, $b=n-c$ and $x=i$, $i\in\{1,\ldots,n\}$. This situation is presented by Fig.1.

Let us consider a technique for constructing the Gaussian fuzzy number over the initial data set. The technique is outlined in [13]. For the goal, the triangular fuzzy number over the initial data set should be constructed and some parameters should be calculated.

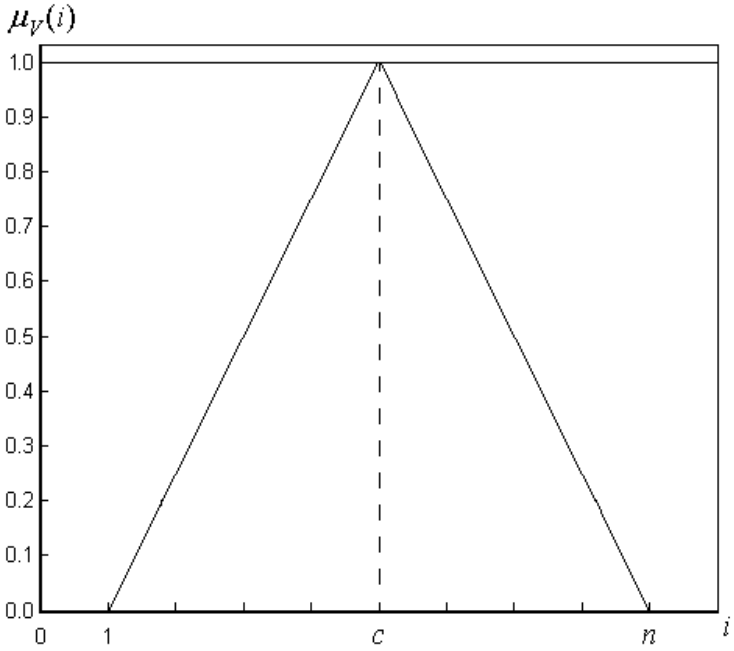

Fig 1: Constructing the Triangular Fuzzy Number

Let $\alpha\in(0,1]$ is the value of tolerance threshold obtained from the D-AFC-TAGA-algorithm application to the data set. So, the parameter $\hat{\underline{c}}$ can be defined from the conditions

$$\mu_V(\hat{\underline{c}})=(1-\alpha),\ \mu_V(1)=0, \quad (22)$$

and the parameter $\hat{\overline{c}}$ can be defined from the conditions

$$\mu_V(\hat{\overline{c}})=(1-\alpha),\ \mu_V(n)=0, \quad (23)$$

where $\mu_V(i)$, $i\in\{1,\ldots,n\}$ is the membership function of the triangular fuzzy number $V=(m,a,b)_T$. A method of calculating the parameters $\hat{\underline{c}}$ and $\hat{\overline{c}}$ is illustrated by Fig.2.

So, the parameters of the Gaussian fuzzy number $V=(m,\sigma)_G$ which is generated from the D-AFC-TAGA-algorithm results can be calculated as follows.

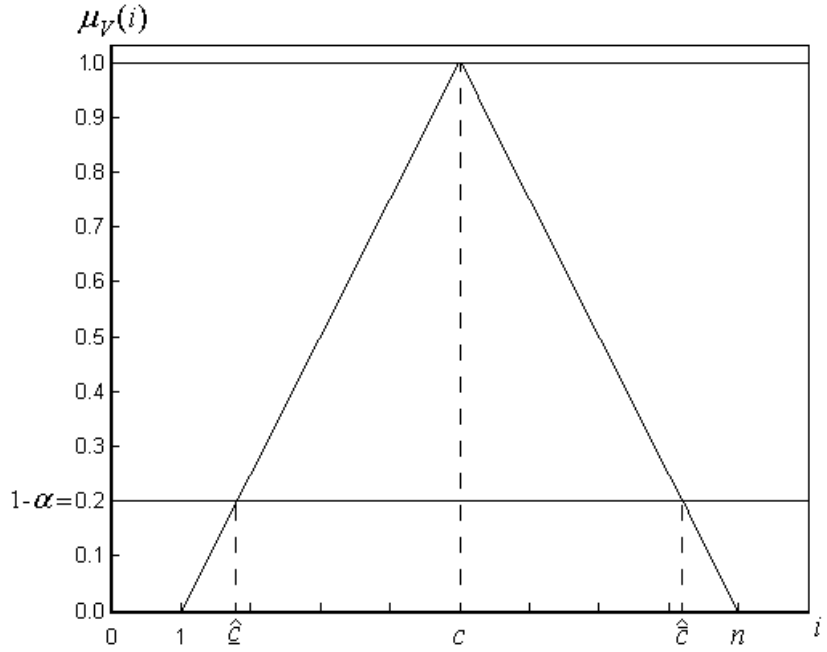

Fig 2: Calculating the parameters for constructing the Gaussian fuzzy number

In general, asymmetric Gaussian membership function $\mu_V(i)$ of fuzzy number $V=(m,\sigma)_G$ can be defined as

$$\mu_V(i)=\begin{cases}\exp\left[-\frac{1}{2}\left(\frac{i-c}{\sigma_L}\right)^2\right], i<c\\ \exp\left[-\frac{1}{2}\left(\frac{i-c}{\sigma_R}\right)^2\right], i\geq c\end{cases}, \quad (24)$$

where σ_L and σ_R represent the left and right spreads. So, values of σ_L and σ_R can be defined as

$$\sigma_L=\frac{\hat{\underline{c}}-c}{\sqrt{-2\ln(1-\alpha)}}, \quad (25)$$

and

$$\sigma_R=\frac{\hat{\bar{c}}-c}{\sqrt{-2\ln(1-\alpha)}}. \quad (26)$$

That is why symmetric Gaussian membership function $\mu_V(i)$ of fuzzy number $V=(m,\sigma)_G$ can be defined as

$$\mu_V(i)=\left\{\exp\left[-\frac{1}{2}\left(\frac{i-c}{\sigma}\right)^2\right], -\infty<i<\infty \right. \quad (27)$$

where $\sigma=\max\{\sigma_L,\sigma_R\}$. The symmetric Gaussian membership function $\mu_V(i)$ of fuzzy number $V=(m,\sigma)_G$ is presented in Fig. 3.

3.2 Constructing the Set of Values of Most Possible Number of Fuzzy Clusters

Triangular or Gaussian fuzzy numbers obtained from the D-AFC-TAGA-algorithm of possibilistic clustering can be useful for constructing the set of values of most possible number of fuzzy clusters in the sought clustering structure. There is the four-step procedure for constructing the set of values.

1. The matrix of the initial data $\hat{X}_{n\times m_1}=[\hat{x}_i^{t_1}]$, $i=1,\ldots,n$, $t_1=1,\ldots,m_1$ after the normalizations processed by the D-AFC-TAGA-algorithm by choosing a suitable distance $d(x_i,x_j)$; the number c of fully separated fuzzy clusters in the obtained allotment $R_c^*(X)$ and the corresponding value of tolerance threshold α are main results of classification;
2. Construct the triangular $V=(m,a,b)_T$ or Gaussian fuzzy number $V=(m,\sigma)_G$ over the initial data set $X=\{x_1,\ldots,x_n\}$;
3. Construct the fuzzy set $\hat{V}=\{\hat{c}_g,\mu_{\hat{V}}(\hat{c}_g)\}$ from the triangular or Gaussian fuzzy number V as follows: a subset of integer values $\hat{C}=\{\hat{c}_*,\ldots,\hat{c}^*\}$ where $\hat{c}_*=2$ and $\hat{c}^*=n-1$ should be extracted from the continuum $(1,n)$ and the value of the membership degree $\mu_{\hat{V}}(\hat{c}_g)$, $\hat{c}_g\in\hat{C}$ of the fuzzy set $\hat{V}$ is equal to the membership function value $\mu_V(i)$ of corresponding fuzzy number V in the case $i=\hat{c}_g$;

4. Construct the α -level fuzzy set for $\hat{V}$ as follows: $\hat{V}_{(\alpha)} = \left\{(\hat{c}_g \in \hat{V}_\alpha, \mu_{\hat{V}_{(\alpha)}}(\hat{c}_g) = \mu_{\hat{V}}(\hat{c}_g))\right\}$, where $\hat{V}_\alpha = \{\hat{c}_g \in \hat{C} \mid \mu_{\hat{V}}(\hat{c}_g) \geq \alpha\}$ is the α -level of the fuzzy set $\hat{V}$.

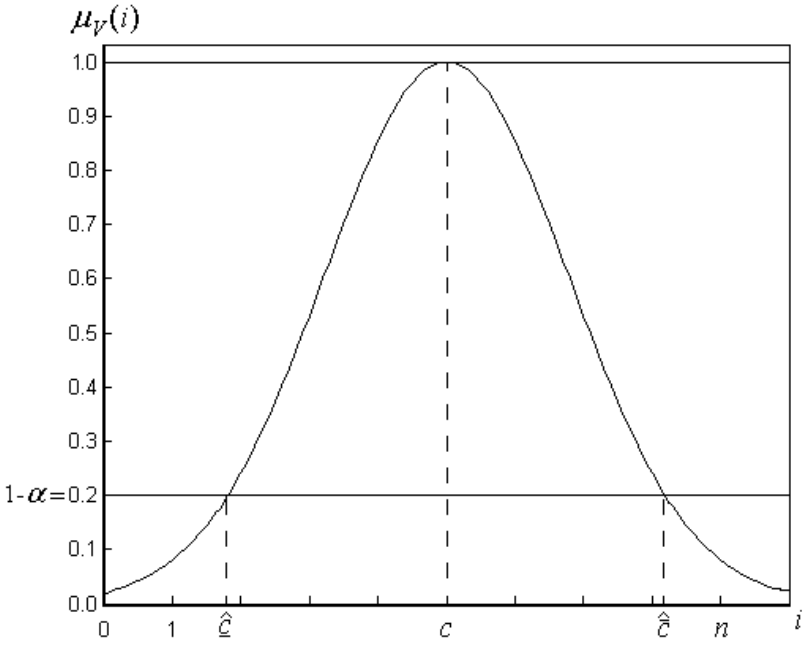

Fig 3: Constructing the Gaussian fuzzy number

Thus, the proposed technique for constructing the set of values of most possible number of fuzzy clusters in the sought clustering structure can be considered as a simplified version of the corresponding technique for a case of the interval-valued data [5].

The matter of the proposed technique is illustrated by Fig. 4.

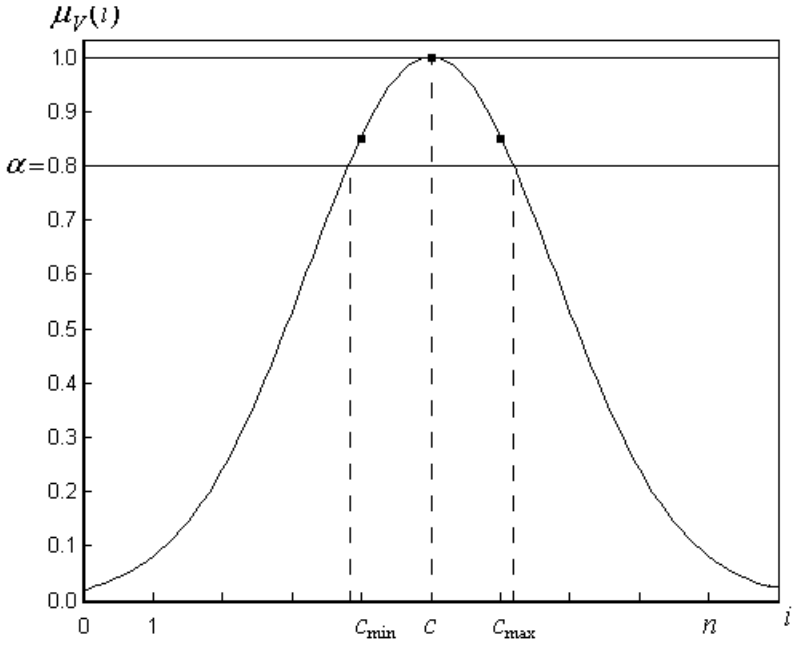

Fig 4: Constructing the Set of Values of Most Possible Number of Fuzzy Clusters from the Gaussian Fuzzy Number

The set $\hat{V}_\alpha = Supp(\hat{V}_{(\alpha)})$ is the set of values of most possible number of fuzzy clusters in some sought clustering structure. So, bounds $c_{\min}$ and $c_{\max}$ for the number of clusters c can be estimated. The membership function $\mu_{\hat{V}_{(\alpha)}}(\hat{c}_g)$ can be interpreted as a possibility distribution π [5] and possibility degrees $\pi(\hat{c}_g)$ express the extent to which the number $\hat{c}_g \in \hat{V}_\alpha$ of fuzzy clusters is plausible. Values of the possibility degrees $\pi(\hat{c}_g)$ are denoted in Fig. 4 by ■.

4. AN ILLUSTRATIVE EXAMPLE

Let us consider an application of proposed technique to the classification problem for the well-known Anderson's Iris data set [14]. The four attribute values represent the sepal length, sepal width, petal length and petal width measured for 150 irises. It has three classes Setosa, Versicolor and Virginica, with 50 samples per class.

The problem is to classify the plants into three subspecies on the basis of this information. It is known that two classes Versicolor and Virginica have some amount of overlap while the class Setosa is linearly separable from the other two.

The Anderson's Iris data form the matrix of attributes $\hat{X}_{150\times4} = [\hat{x}_i^{t_1}]$, $i = 1,\ldots,150$, $t_1 = 1,\ldots,4$, where the sepal length is denoted by $\hat{x}^1$, sepal width – by $\hat{x}^2$, petal length – by $\hat{x}^3$ and petal width – by $\hat{x}^4$. The data was preprocessed according to the formula (16).

So, each object can be considered as a fuzzy set x_i, $i = 1,\ldots,150$ and $x_i^{t_1} = \mu_{x_i}(x^{t_1}) \in [0,1]$, $i = 1,\ldots,150$, $t_1 = 1,\ldots,4$, are their membership functions. The distance (19) was applied to the normalized data as the parameter for the D-AFC-TAGA-algorithm in experiments.

In a case of automatic constructing an acceptable transitive approximation $\tilde{T}_k$, $k = 1,\ldots6$ of the fuzzy tolerance T, the result obtained from the D-AFC-TAGA-algorithm is equal to the result obtained by using the mean operator. The fact is explained by Table 1 where values of the Kuzmin's distance [15]

$$d(T,\tilde{T}_k) = \sum_{(x_i,x_j)} \left| \mu_T(x_i,x_j) - \mu_{\tilde{T}_k}(x_i,x_j) \right|, \quad \tilde{T}_k \in \{\tilde{T}_1,\ldots,\tilde{T}_6\} \tag{28}$$

are given.

Table 1. Values of the distance between fuzzy tolerance and its transitive approximations

A type of the aggregation operator	Values of the distance between fuzzy relations
maximum	3348.760
minimum	4156.602
mean	1298.728
median	1546.668
upmedian	1577.377
downmedian	1674.286

So, the condition $\min_k d(T,\tilde{T}_k)$, $k = 1,\ldots6$ is met for the transitive approximation obtained by using the mean operator. The allotment among $c = 2$ fuzzy clusters was obtained for the value $\alpha = 0.80384$. The set of values of most possible number of fuzzy clusters in the sought clustering structure with corresponding possibility degrees obtained using the triangular fuzzy number is presented in Fig. 5.

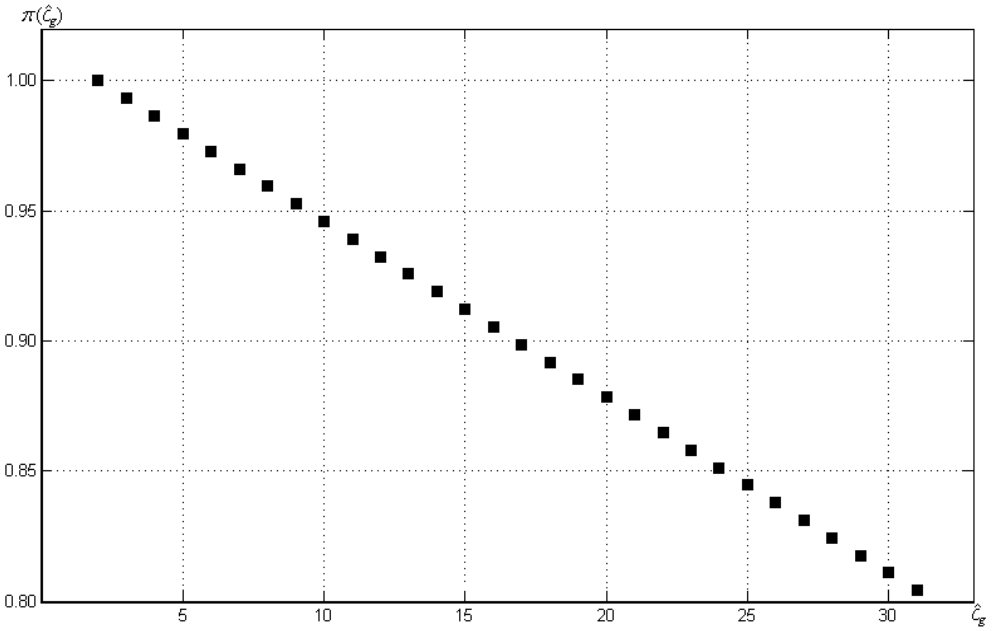

Fig 5: Possibility degrees obtained by using the triangular fuzzy number

Thus, the set of values of most possible number of fuzzy clusters in the sought clustering structure is $\{c_{\min} = 2, \ldots, c_{\max} = 31\}$.

On the other hand, the set of values of most possible number of fuzzy clusters in the sought clustering structure with corresponding possibility degrees obtained by using the Gaussian fuzzy number is presented in Fig. 6.

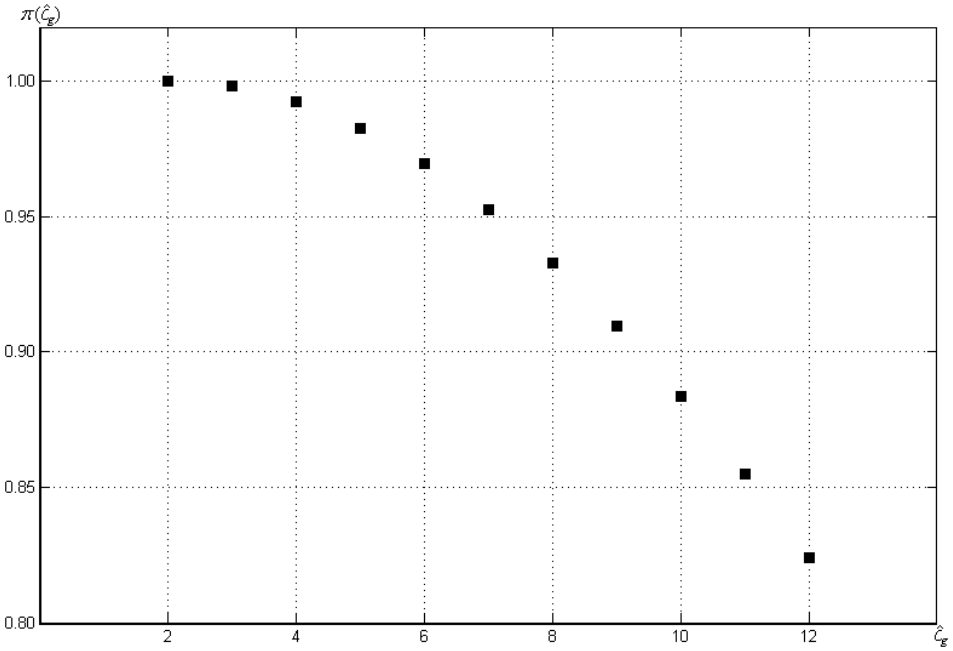

Fig 6: Possibility degrees obtained by using the Gaussian fuzzy number

So, the set of values of most possible number of fuzzy clusters in the sought clustering structure is $\{c_{\min} = 2, \ldots, c_{\max} = 12\}$.

5. CONCLUDING REMARKS

The technique for estimation of a lower bound for the number of clusters $c_{\min}$ and an upper bound for the number of clusters $c_{\max}$ for the set $\{c_{\min}, \ldots, c_{\max}\}$ of most possible number of fuzzy clusters in the sought clustering structure is proposed in the paper. The heuristic D-AFC-TAGA-algorithm and fuzzy numbers are a basis of the proposed technique.

So, a clustering procedure can be applied to the data set for estimated set $\{c_{\min}, \ldots, c_{\max}\}$ of clusters in the sought clustering structure. The proposed technique can be simply generalized for a case of the relational data set (15) by using the heuristic D-PAFC-algorithm of possibilistic clustering [5].

Numerical experiments are show, that the proposed technique is a useful tool for the exploratory data analysis.

6. REFERENCES

[1] Zadeh, L.A. 1965. Fuzzy Sets. Information and Control. 8, 3, 338-353.

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

[3] Höppner, F., Klawonn, F., Kruse, R. and Runkler, T. 1999. Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition. Chichester: Wiley.

[4] Krishnapuram, R. and Keller, J.M. 1993. A Possibilistic Approach to Clustering. IEEE Transactions on Fuzzy Systems. 1, 2, 98-110.

[5] Viattchenin, D.A. 2013. A Heuristic Approach to Possibilistic Clustering: Algorithms and Applications. Heidelberg: Springer.

[6] Viattchenin, D.A. and Damaratski A. 2013. Direct Heuristic Algorithms of Possibilistic Clustering Based on Transitive Approximation of Fuzzy Tolerance. Informatica Economicá. 17, 3, 5-15.

[7] Viattchenin, D.A., Yaroma, A. and Damaratski, A. 2014. A Novel Direct Relational Heuristic Algorithm of Possibilistic Clustering. International Journal of Computer Applications. 107, 18, 15-21.

[8] Dawyndt, P., De Meyer, H. and De Baets, B. 2006. UPGMA Clustering Revisited: A Weight-Driven Approach to Transitive Approximation. International Journal of Approximate Reasoning. 42, 3, 174-191.

[9] Yi, X., Miao, Y., Zhou, J. and Wang, Y. 2016. Some Novel Inequalities for Fuzzy Variables of the Variance and Its Rational Upper Bounds. Journal of Inequalities and Applications. 2016, 41.

[10] Sato-Ilic, M. and Jain, L.C. 2006. Innovations in Fuzzy Clustering: Theory and Applications. Heidelberg: Springer.

[11] Viattchenin, D.A. 2009. An Outline for a Heuristic Approach to Possibilistic Clustering of the Three-Way Data. Journal of Uncertain Systems. 3, 1, 64-80.

[12] Kaufmann, A. 1975. Introduction to the Theory of Fuzzy Subsets. New York: Academic Press.

[13] Viattchenin, D.A., Tati, R., and Damaratski, A.V. 2013. Designing Gaussian Membership Functions for Fuzzy Classifier Generated by Heuristic Possibilistic Clustering. Journal of Information and Organizational Sciences. 37, 2, 127-139.

[14] Anderson, E. 1935. The Irises of the Gaspe Peninsula. Bulletin of the American Iris Society. 59, 1, 2-5.

[15] Kuzmin, V.B. 1982. Constructing of Group Decisions in Spaces of Crisp and Fuzzy Binary Relations. Moscow: Nauka. (in Russian)