On the Computational Accuracy of the Heuristic Method of Possibilistic Clustering

Aliaksandr Damaratski[1)], **Dzmitry Novikau**[2)]

GIS Research and Engineering Center of the National Academy of Sciences of Belarus, Surganov St. 6, 220012 Minsk, Belarus 1) e-mail: xandr.hell@gmail.com 2) e-mail: novikov_di@tut.by

Abstract: *The problem of computational efforts of a heuristic method of possibilistic clustering based on the concept of allotment among fuzzy clusters is considered in the paper. The basic concepts of the method of clustering based on the allotment concept are considered. The dependence of the classification results upon the accuracy of the computation is illustrated on the example of Anderson's Iris data. Preliminary conclusions are formulated.*

***Keywords*:** clustering, fuzzy cluster, fuzzy tolerance, allotment among fuzzy cluster, typical point

1. INTRODUCTION

Cluster analysis is a structural approach to solving the problem of object classification without training samples. Clustering methods are called also automatic classification methods and numeral taxonomy methods.

Fuzzy clustering has been applied successfully in various fields, including image segmentation, decision analysis and social sciences. Despite the successful applications, there are a number of issues that must be dealt with in practical applications of fuzzy clustering algorithms. Heuristic methods of fuzzy clustering, hierarchical methods of fuzzy clustering and optimization methods of fuzzy clustering were proposed by different authors. Different fuzzy clustering methods are considered by Miyamoto [1]. Objective function-based clustering algorithms are the most widespread methods in fuzzy clustering. For example, the most popular FCM-algorithm of fuzzy clustering is based on the minimization of the objective function [2]

$$Q(P) = \sum_{l=1}^{c} \sum_{i=1}^{n} u_{li}^{\gamma} \left\| x_i - \overline{\tau}^l \right\|^2 \qquad (1)$$

under the fuzzy c-partition constraint

$$\sum_{l=1}^{c} u_{li} = 1,\, l = 1,\ldots,c\ , \qquad (2)$$

where $X = \{x_1,\ldots,x_n\}$ is the initial set of points, c is the number of fuzzy clusters $A^l \in P$, $l = 1,\ldots,c$, $0 \le u_{li} \le 1$ is the membership degree of object x_i to fuzzy cluster A^l, $\overline{\tau}^l$ is the prototype for fuzzy cluster A^l, and the parameter $\gamma > 1$ is the fuzziness index. The fuzziness index is called the fuzzifier also. The condition (2) is very difficult from profound point of view. That is why a possibilistic approach to clustering was proposed by Krishnapuram and Keller [3]. The approach can be considered as a special case of fuzzy clustering because membership degrees defined a possibility distribution function for some cluster over the domain of discourse consisting of all objects of the data set. The condition (2) is replaced in the possibilistic methods by the constraint of possibilistic partition

$$\sum_{l=1}^{c} \mu_{li} > 1,\, l = 1,\ldots,c\ . \qquad (3)$$

where $0 \le \mu_{li} \le 1$ is the possibilistic memberships and the parameter $\psi > 1$ is the analog of the fuzziness index.

All possibilistic clustering algorithms are objective function-based procedures. However, a heuristic method of possibilistic clustering was proposed in [4] and the method is called the D-AFC(c)-algorithm. The possibilistic clustering method based on the concept of allotment among fuzzy clusters. The D-AFC(c)-algorithm is very precise and effectual method for images segmentation. The fact was demonstrated in [5].

The D-AFC(c)-algorithm display low level of a complexity. However, the D-AFC(c)-algorithm is sensitive to the accuracy of the initial data. The main goal of the paper is a preliminary consideration of the problem of the computational accuracy of the D-AFC(c)-algorithm. The contents of this paper is as follows: in the second section basic concepts of the possibilistic clustering method based on the concept of allotment among fuzzy clusters are outlined and the general plan of the D-AFC(c)-algorithm is described, in the third section the data preprocessing method are given, in the fourth section a simple illustrative example is considered, in the fifth section the results of application of the D-AFC(c)-algorithm with different values of accuracy to the Anderson's Iris data example are given, in the sixth section some final remarks are stated.

2. BASIC NOTIONS

Basic concepts of the D-AFC(c)-algorithm 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. Fuzzy tolerance is the fuzzy binary intransitive relation which possesses the symmetricity property

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

and the reflexivity property

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

Let T be a fuzzy tolerance on X and α be the α-level value of T, $\alpha \in (0,1]$. So, columns or lines $\{A^1,...,A^n\}$ of a matrix of T be fuzzy sets on X. The

α -level fuzzy set $A^l_{(\alpha)} = \left\{(x_i, \mu_{A^l}(x_i)) \mid \mu_{A^l}(x_i) \geq \alpha\right\}$ is fuzzy cluster, and μ_{li} is the membership degree of the element $x_i \in X$ for some fuzzy cluster $A^l_{(\alpha)}$, $\alpha \in (0,1]$, $l \in [1, n]$. 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

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

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 $A^l_{(\alpha)}$, $A^l_\alpha = Supp\left(A^l_{(\alpha)}\right)$. So, the value of α is the tolerance threshold of fuzzy clusters elements. The number c of fuzzy clusters can be equal the number of objects, n.

From other hand, the point $\tau^l_e \in A^l_\alpha$, for which

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

is called a typical point of the fuzzy cluster $A^l_{(\alpha)}$, $\alpha \in (0,1]$. The symbol e is the index of the typical point because a fuzzy cluster can have a few typical points. A set $K(A^l_{(\alpha)}) = \{\tau^l_1, \ldots, \tau^l_{|l|}\}$ of typical points of the fuzzy cluster $A^l_{(\alpha)}$ is a kernel of the fuzzy cluster and $card\left(K(A^l_{(\alpha)})\right) = |l|$ is a cardinality of the kernel. If the fuzzy cluster has a unique typical point, then the fuzzy cluster is fuzzy cluster with a center. Otherwise, the fuzzy cluster is the fuzzy cluster with a kernel. Note that the concept of the typical point is not equal to the concept of the prototype for fuzzy cluster in the objective function-based methods of clustering because any typical point is some real object always.

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 α, which are generated by some fuzzy tolerance T on the set X. If condition

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

is met for all $A^l_{(\alpha)}$, $l = \overline{1,c}$, $c \leq n$, then the family is the allotment of elements of the set $X = \{x_1, ..., x_n\}$ among fuzzy clusters $\{A^l_{(\alpha)}, l = \overline{1,c}, 2 \leq c \leq n\}$ for some value of the tolerance threshold α. It should be noted that the condition (8) is equal to the condition of the possibilistic partition, but fuzzy clusters in the sense of the expression (6) are elements of the possibilistic partition. Moreover, several allotments $R^\alpha_z(X)$ can exist for some tolerance threshold α. So, symbol z is the index of an allotment.

Detection of fixed c number of fuzzy clusters can be considered as the aim of classification. So, the classification problem can be characterized formally as determination of the unique solution $R^*(X) \in B(c)$ where $B(c) = \{R^\alpha_z(X)\}$ is the set of allotments corresponding to the formulation of a concrete classification problem. The plan of the D-AFC(c)-algorithm of possibilistic clustering is described in [4].

The allotment $R^*(X) = \{A^l_{(\alpha)} \mid l = \overline{1,c}\}$ among the given number of fuzzy clusters and the corresponding value of tolerance threshold α are the results of classification.

3. THE DATA PREPROCESSING

The D-AFC(c)-algorithm can be applied directly to the matrix of tolerance coefficients. However, the initial data are contained in the matrix of attributes very often. Let us consider a method for the data preprocessing.

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. However, the data can be presented as a matrix of attributes $\hat{X}_{n \times m} = [\hat{x}^t_i]$, $i = 1, \ldots, n$, $t = 1, \ldots, m$, where the value $\hat{x}^t_i$ is the value of the t -th attribute for i -th object. The data can be normalized as follows:

$$x^t_i = \frac{\hat{x}^t_i}{\max\limits_i \hat{x}^t_i}, \tag{9}$$

So, each object can be considered as a fuzzy set x_i, $i = 1, \ldots, n$ and $x^t_i = \mu_{x_i}(x^t) \in [0,1]$, $i = 1, \ldots, n$, $t = 1, \ldots, m$ 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 to the matrix of normalized data $X_{n \times m} = [\mu_{x_i}(x^t)]$, $i = 1, \ldots, n$, $t = 1, \ldots, m$. The most widely used distances for fuzzy sets x_i and x_j, $i, j = 1, \ldots, n$ in $X = \{x_1, ..., x_n\}$ are [6] the normalized Hamming distance

$$l(x_i, x_j) = \frac{1}{m} \sum_{t=1}^{m} |\mu_{x_i}(x^t) - \mu_{x_j}(x^t)|, \ i, j = \overline{1,n}, \tag{10}$$

the normalized Euclidean distance

$$e(x_i, x_j) = \sqrt{\frac{1}{m} \sum_{t=1}^{m} \left(\mu_{x_i}(x^t) - \mu_{x_j}(x^t)\right)^2}, \ i, j = \overline{1,n}, \tag{11}$$

and the squared normalized Euclidean distance

$$\varepsilon(x_i, x_j) = \frac{1}{m} \sum_{t=1}^{m} \left(\mu_{x_i}(x^t) - \mu_{x_j}(x^t)\right)^2, \ i, j = \overline{1,n}. \tag{12}$$

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 to the matrix of fuzzy intolerance $I = [\mu_I(x_i, x_j)]$, $i, j = 1, \ldots, n$ obtained from previous operations:

$$\mu_T(x_i,x_j)=1-\mu_I(x_i,x_j)\,,\ i,j=\overline{1,n}\,. \qquad (13)$$

So, the matrix of tolerance coefficients $T=[\mu_T(x_i,x_j)]$, $i,j=1,\ldots,n$ can be constructed by normalizing the initial data and choosing a suitable distance for fuzzy sets. However, tolerance coefficients $\mu_T(x_i,x_j)$, $i,j=1,\ldots,n$ depend on the accuracy of measurement of the data set and values of the tolerance coefficients depend on the distance selection in the process of the data preprocessing.

The membership function (6) of each element of the fuzzy cluster is obtained from the matrix of fuzzy tolerance. In general, the results of application of the D-AFC(c)-algorithm to the data depend on the tolerance coefficients because the sequence $0<\alpha_0<\alpha_1<\ldots<\alpha_\ell<\ldots<\alpha_Z\le 1$ of α-levels is constructed from the values of $\mu_T(x_i,x_j)$, $i,j=1,\ldots,n$. The fact can be explained by a simple example.

4. AN ILLUSTRATIVE EXAMPLE

Let us consider the two-dimensional data set which is presented in Fig. 1. The simple example will be useful in further considerations.

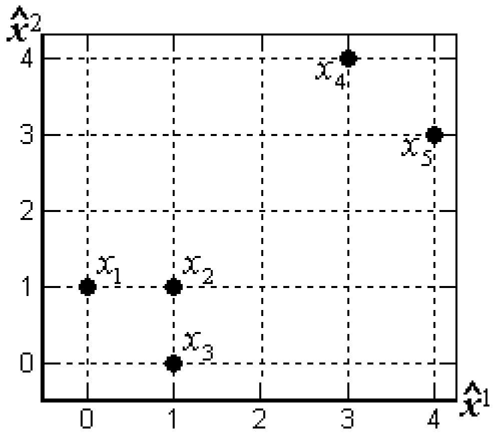

Fig. 1 – The data set

The matrix of a fuzzy tolerance T was obtained from the matrix of the initial data using the normalized Euclidean distance (11) in the process of the data preprocessing for the accuracy value $\varepsilon=0.0001$. The matrix is given in Table 1.

Table 1. The fuzzy tolerance relation on the object set for the accuracy value ε=0.0001

$T1$	x_1	x_2	x_3	x_4	x_5
x_1	1.0000				
x_2	0.8232	1.0000			
x_3	0.7500	0.8232	1.0000		
x_4	0.2500	0.3626	0.2094	1.0000	
x_5	0.2094	0.3626	0.2500	0.7500	1.0000

The initial set of objects $X=\{x_1,x_2,x_3,x_4,x_5\}$ can be divided into two classes, $A^1=\{x_1,x_2,x_3\}$ and $A^2=\{x_4,x_5\}$. By executing the D-AFC(c)-algorithm for two classes and for the accuracy value $\varepsilon=0.1$, we obtain that the allotment $R^*(X)$, which is corresponds to the result, was obtained for the tolerance threshold $\alpha=0.8$. The first object is the typical point τ^1 of the fuzzy cluster which corresponds to the first class and the fourth object is the typical point τ^2 of the second fuzzy cluster. Memberships of both classes are presented in Table 2.

Table 2. The allotment among two classes obtained for the accuracy value ε=0.1

Numbers of objects	Membership degrees	
	Class 1	Class 2
x_1	1.0	0.0
x_2	0.8	0.0
x_3	0.8	0.0
x_4	0.0	1.0
x_5	0.0	0.8

From other hand, the matrix of the fuzzy tolerance T was processed by D-AFC(c)-algorithm for the accuracy value $\varepsilon=0.0001$ and the result is presented in Table 3.

Table 3. The allotment among two classes obtained for the accuracy value ε=0.1

Numbers of objects	Membership degrees	
	Class 1	Class 2
x_1	0.8232	0.0000
x_2	1.0000	0.0000
x_3	0.8232	0.0000
x_4	0.0000	1.0000
x_5	0.0000	0.7500

So, by executing the D-AFC(c)-algorithm for two classes and for the accuracy value $\varepsilon=0.0001$, we obtain the allotment $R^*(X)$ among two fully separated fuzzy clusters. The allotment was obtained for the tolerance threshold $\alpha=0.75$. The second object is the typical point τ^1 of the fuzzy cluster which corresponds to the first class and the fourth object is the typical point τ^2 of the second fuzzy cluster.

5. EXPERIMENTAL RESULTS

The Anderson's Iris data set consists of the sepal length, sepal width, petal length and petal width measured for 150 irises [7]. The problem is to classify the plants into three subspecies on the basis of this information. The Anderson’s Iris data form the matrix of attributes $X_{4\times150}=[\hat{x}_i^t]$, $i=1,\ldots,150$, $t=1,\ldots,4$ 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 Iris database is the most known database to be found in the pattern recognition literature. The method of the data preprocessing which was described in the third section can be used for constructing the matrix of fuzzy tolerance and the matrix of fuzzy tolerance can be processed by the D-AFC(c)-algorithm. The squared normalized Euclidean distance (12) was selected for the data preprocessing. The results the data processing by D-AFC(c)-algorithm is presented in the Table 4.

Table 4. Results of application of the D-AFC(c)-algorithm to the Iris data for different values of the accuracy threshold

Characteristics	The value of accuracy threshold	
	ε=0.001	ε=0.0001
A number of misclassified objects	6	6
The value of the tolerance threshold	$\alpha = 0.965$	$\alpha = 0.9642$
Typical points of fuzzy clusters	$x_{95} = \tau_1^1$, $x_1 = \tau_2^1$, $x_{36} = \tau_3^1$, $x_{64} = \tau_4^1$, $x_{106} = \tau_5^1$, $x_{134} = \tau_6^1$, $x_{145} = \tau_7^1$, $x_{98} = \tau_1^2$, $x_{33} = \tau_2^2$, $x_{73} = \tau^3$	$x_{95} = \tau^1$, $x_{98} = \tau^2$, $x_{73} = \tau^3$

Obviously, the number of typical points of fuzzy clusters in the allotment $R^*(X)$ and the value of the tolerance threshold α depend on the accuracy threshold ε, which was used in the data preprocessing. However, the results, when these membership assignments are converted into hard outputs, are similar. So, the D-AFC(c)-algorithm leads to $6/150$ error rate in both experiments.

6. FINAL REMARKS

The D-AFC(c)-algorithm is a precise and effective numerical procedure for solving classification problems. The results of application of the clustering method based on the allotment concept can be very well interpreted and the clustering results are stable because the D-AFC(c)-algorithm depending on the set $B(c) = \{R_z^\alpha(X)\}$ of possible solutions of the classification problem. However, the results of application of the D-AFC(c)-algorithm to the data depend on the computational accuracy.

The test of the D-AFC(c)-algorithm on the Anderson's Iris data shows a sensitivity of the algorithm to the accuracy threshold ε selection. Obviously, the membership function obtained from the D-AFC(c)-algorithm for small values of is sharper for increasing the accuracy threshold ε. So, the accuracy threshold ε can be considered as an analog of fuzzifier γ in (1). The accuracy threshold ε can be considered as a sufficient parameter of classification in some cases. So, the accuracy threshold ε must be taken into account in the process of the data preprocessing.

7. REFERENCES

[1] S. Miyamoto. *Fuzzy Sets in Information Retrieval and Cluster Analysis.* Kluwer Academic Publishers. Dordrecht, 1990.

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

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

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

[5] D.A. Viattchenin, D. Novikau, A. Damaratski. An application of algorithms based on the concept of allotment among fuzzy clusters to image segmentation. *Proceedings of the 9th International Conference "Pattern Recognition and Information Processing (PRIP'2007)"*, Minsk, Belarus, 22-24 May 2007, Vol. 2, pp. 226-231.

[6] A. Kaufmann. *Introduction to the Fuzzy Subsets Theory.* Academic Press. New York, 1975.

[7] E. Anderson. The irises of the Gaspe Peninsula. Bulletin of the American Iris Society 59 (1) (1934). p 2-5.