Fuzzy Clustering Using C-Means Method

Georgi Krastev [1], Tsvetozar Georgiev [1]

[1]*Department of Computing, University of Ruse, Ruse, Bulgaria*

***Abstract* – The cluster analysis of fuzzy clustering according to the fuzzy c-means algorithm has been described in this paper: the problem about the fuzzy clustering has been discussed and the general formal concept of the problem of the fuzzy clustering analysis has been presented. The formulation of the problem has been specified and the algorithm for solving it has been described.**

***Keywords* – cluster analysis, fuzzy clustering, c-means.**

1. Introduction

It has been accepted to designate the cluster of algorithms, approaches and procedures, developed for solving the problems about forming homogeneous classes in a random problem area [1], [2], [3], [4], [5], [6] by the term cluster analysis.
In general, the problem of fuzzy clustering consists of finding the fuzzy grouping or the fuzzy coverage of the multitude of elements of the researched set, which form the structure of the fuzzy clusters, appearing in the data in question. This problem is reduced to finding the degree of the belonging of the elements of the universe toward the sought fuzzy clusters, which in combination define the fuzzy grouping or the fuzzy coverage of the output multitude of the researched elements [16], [17], [18], [19],[20].

The problem of fuzzy clustering can be formulated in the following way: for the specified matrix data ***D***, the quantity of the fuzzy clusters c ($c \in N$ и $c>1$) and the parameter m, the task is to define the matrix U by the values of the functions for belonging of the objects of the clustering $a_i \in A$ with fuzzy clusters A_k ($k \in \{2, ..., c\}$), which minimize the target function (3) and comply with the restrictions (1) and (2), as well as the additional restrictions (4) and (5):

$$\sum_{k=1}^{c} \mu_{A_k}(a_i) = 1 \qquad (\forall a_i \in A), \tag{1}$$

$$v_j^k = \frac{\sum_{i=1}^{n} (\mu_{A_k}(a_i))^m \cdot x_j^i}{\sum_{i=1}^{n} (\mu_{A_k}(a_i))^m} \quad (\forall k \in \{2,\cdots,c\}, \forall p_j \in P) \tag{2}$$

$$f(A_k, v_j^k) = \sum_{i=1}^{n}\sum_{k=1}^{c} (\mu_{A_k}(a_i))^m \sum_{j=1}^{q} (x_j^i - v_j^k)^2 \tag{3}$$

$$\sum_{i=1}^{n} \mu_{A_k}(a_i) > 0 \qquad (\forall k \in \{2,..., c\}), \tag{4}$$

$$\mu_{A_k}(a_i) \geq 0 \ (\forall k \in \{2,..., c\}, \forall a_i \in A), \tag{5}$$

where c is the total quantity of the fuzzy clusters A_k ($k \in \{2, ..., c\}$), which is considered for previously given ($c \in N$ и $c>1$), m is a parameter, called exponential weight and equal to a real number ($m>1$). Each of the cluster centres represents a vector $v_k=(v_1^k, v_2^k, \ldots, v_q^k)$ in q-dimensional norm space, isomorphic R^q, i.e. $v_k^j \in R^q$, if all signs are measured in the ratio scale [11].

Condition (4) excludes the occurrence of empty fuzzy clusters in the fuzzy clustering which is being sought. The last condition (5) has an entirely formal character, because it follows immediately from the definition of the membership function of the fuzzy multitudes. In this case the minimization of the objective function (3) minimizes the deviation of all the clustering objects from the centres of the fuzzy clusters, proportionally to the values of the membership functions of these objects with the corresponding fuzzy clusters [9], [10].
Due to the fact that the objective function (2) is not protruding, and the restrictions (1), (2), (4) and (5) in their totality form a non-protruding set of feasible alternatives, therefore in the general case the problem of fuzzy clustering refers to the multiextreme problems of nonlinear programming.

2. Algorithm for solving the problem of fuzzy clustering according to the method of fuzzy c-means

The main concepts of the algorithm for solving the formulated problem of fuzzy clustering were proposed by J. C. Dunn in 1974. Initially, this algorithm received the designation of the fuzzy algorithm ISODATA (fuzzy ISODATA or F-ISODATA) [15], [16]. J. C. Bezdek proved theoretically the convergence of this algorithm in 1980. Later on, in 1981, J. C. Bezdek summarized the algorithm ISODATA about a case with a random fuzzy variety and proposed the fuzzy c-means (FCM, Fuzzy C-Means) [13], [14] for its designation. This algorithm is most popular namely with this designation.

The algorithm FCM for solving the problem of fuzzy clustering in the type of (1)–(5) has an iterative character of consistently improving the output fuzzy clustering $R(A)=\{A_k|A_k \ \ A\}$, which is given by the user or is automatically formed by a heuristic rule. The values of the membership functions of the fuzzy clusters and their typical representatives are recalculated recurrently at each iteration [8], [9], [12].

The algorithm FCM will stop operating in case there is implementation of the a priori specified finite number of iterations, or when the minimum absolute difference between the values of the membership functions of two successive iterations becomes less than some a priori specified value.
Formally, the FCM algorithm is defined in the form of iterative execution of the following sequential steps:

1. It is necessary that the following values must be given in advance: the quantity of the fuzzy clusters that are sought c ($c \in N$ и $c>1$), the maximum quantity of iterations of the algorithm s ($s \in N$), the convergence parameter of the algorithm ε($\varepsilon \in R+$), as well as the exponential weight of the target function and the cluster centres m (as a rule $m=2$). The output fuzzy clustering $R(A)=\{A_k|A_k \ \ A\}$ of c non-empty fuzzy clusters, which are described by the sum of the membership functions $\mu_k(a_i)$ $(\forall k \in \{2,...,c\}, \forall a_i \in A)$ is given as the *current fuzzy* clustering of the first iteration of the algorithm for a matrix of data ***D***.
2. The centres of the fuzzy clusters v_j^k $(\forall k \in \{2,...,c\}, \forall p_j \in P)$ are calculated for the output fuzzy clustering $R(A)=\{A_k|A_k \ \ A\}$ according to formula (2) and the value of the target function $f(A_k, v_j^k)$ is calculated according to formula (3).
3. A new fuzzy breaking $R(A)=\{A_k|A_k \ \ A\}$ is formed of the output multitude of objects of the A clustering of c non-empty fuzzy clusters, characterized by the sum of the membership functions $\mu'_k(a_i)$ $(\forall k \in \{2,...,c\}, \forall a_i \in A)$,which are determined according to the formula:

$$\mu' A_k(a_i) = \left(\sum_{l=1}^{c} \left(\frac{(\sum_{j=1}^{q}(x_j^i - v_j^k)^2)^{\frac{1}{2}}}{(\sum_{j=1}^{q}(x_j^i - v_j^l)^2)^{\frac{1}{2}}} \right)^{\frac{2}{m-1}} \right)^{-1} \quad (\forall k \in \{2,...,c\}, \forall a_i \in A). \tag{6}$$

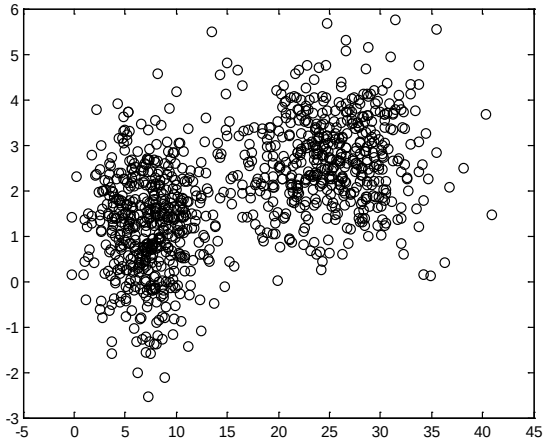
4. Moreover, if for any $k \in \{2,...,c\}$ and for any $a_i \in A$ the value $\sum_{j=1}^{q}(x_j^i - v_j^k)^2 = 0$, then for the corresponding fuzzy cluster A_k we exert $\mu'_k(a_i)=1$, and for the rest A $_l$ $(\forall l \in \{2,...,c\}, l \neq k)$ we exert $\mu'_l(a_i)=0$. If $k \in \{2,...,c\}$ for any $a_i \in A$ turn out to be several, i.e. for them the value $\sum_{j=1}^{q}(x_j^i - v_j^k)^2 = 0$, then heuristically for the smallest k we exert $\mu'_k(a_i)=1$, and for the rest $l \in \{2,...c\}, l \neq k$ п we exert $\mu'_l(a_i)=0$.
5. The centres of the fuzzy clusters v_j^k $(\forall k \in \{2,...,c\}, \forall p_j \in P)$ are calculated for the new fuzzy breaking $R(A)=\{A_k|A_k \ \ A\}$ according to formula (2) and the value of the target function $f'(A_k, v_j^k)$ is calculated according to formula (3).
6. If the quantity of the implemented iterations exceeds the previously given number s or the difference module $|f(A_k, v_j^k) - f'(A_k, v_j^k)| \leq \varepsilon$, i.e. it does not exceed the value of the convergence parameter of the algorithm ε, then for the result which is sought for the fuzzy clustering, the fuzzy grouping $R'(A)=\{A_k|A_k \ \ A\}$ is accepted and the implementation of the algorithm finishes. Otherwise, it is considered to be a current fuzzy clustering $R(A)=R'(A)$ and one must proceed with the next step 3 of the algorithm, increasing the quantity of the implemented iterations with 1.

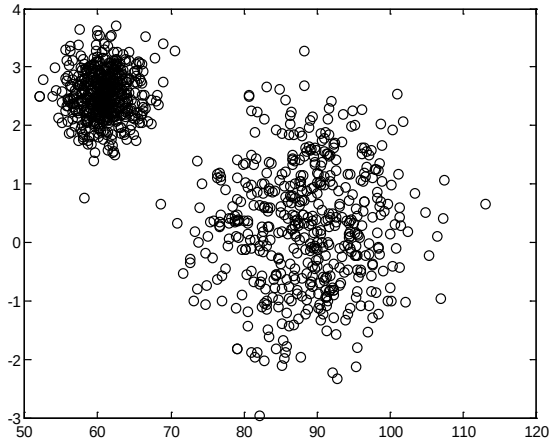
The FCM algorithm in its character belongs to the approximate algorithms for searching the extreme for the target function (3) by the presence of the restrictions (1), (2), (4) and (5). Therefore, as a result of the implementation of that particular algorithm, the optimal fuzzy clustering $R^*(A)$, is determined locally, and which is described by the sum of the membership functions $\mu_k(a_i)$ $(\forall k \in \{2,...,c\}, \forall a_i \in A)$, as well as by the centres or by the typical representatives of each of the fuzzy clusters v_j^k $(\forall k \in \{2,...,c\}, \forall p_j \in P)$.

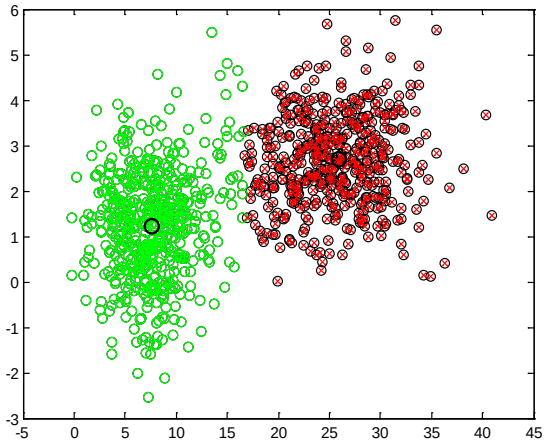
The experiment for solving applied problems for fuzzy clustering shows that the most efficient way for receiving adequate results comprises the multiple implementation of the FCM algorithm for different output fuzzy clustering, and even if the quantity of the fuzzy clusters is not known for the different values c ($c \in N$ and $c>1$). The received results for identical values c are compared with the values of the target function of the received fuzzy clustering in order to determine the final solution about the fuzzy clustering, which is sought.

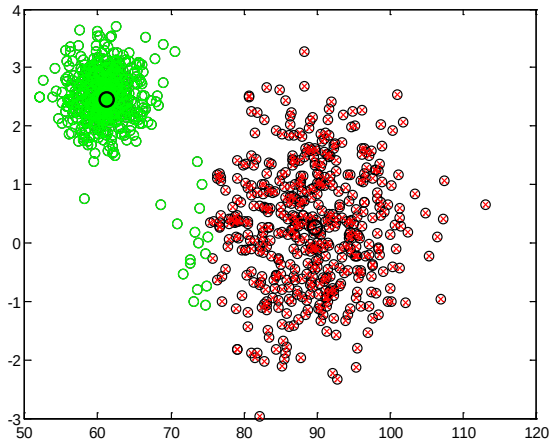
3. Received results

After the conducted research, there are loads of available data, which were received during the four experiments and which are used as a test set of objects for fuzzy clustering. These data comprise a matrix of ***D*** data with measurements (1000x2). In that case the matrix of ***D*** data corresponds to 1000objects about each of which measurements according to two criteria are fulfilled and which appears to be suitable for visualizing the output data and results from the fuzzy clustering in the two dimensional space. The result of the solution of the problem of fuzzy clustering for 2 fuzzy clusters by applying the described algorithm has been visualized on Figure 1.



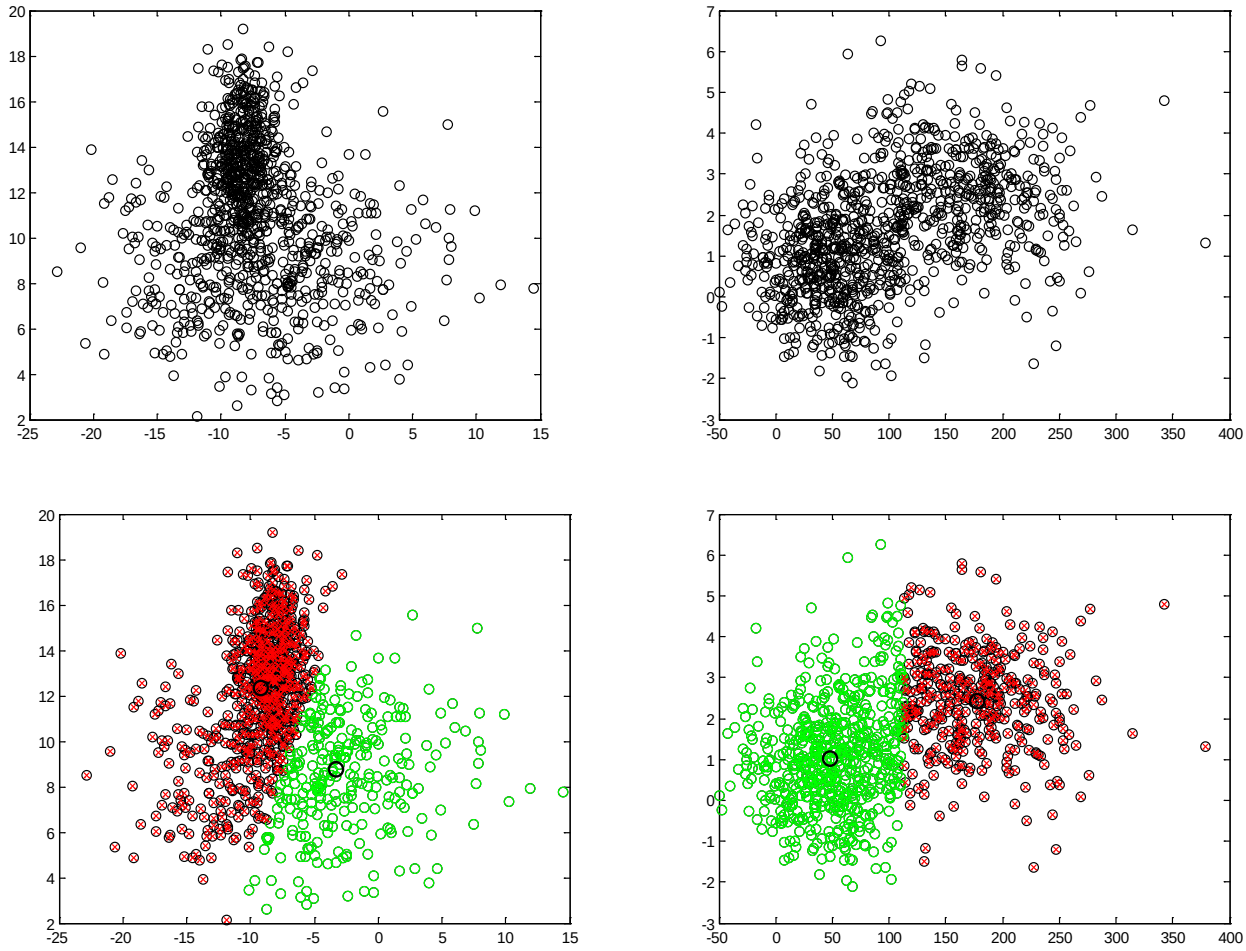






Coordinates of the centres of fuzzy clusters in question
7.57004035736911 1.22902153750007
25.8924428464167 2.69117623296698

Coordinates of the centres of fuzzy clusters in question
61.1867694239986 2.46175886064993
89.6915488860185 0.272222272501953



Coordinates of the centres of fuzzy clusters in question

-3.29820380278446	*8.74729032363265*
-9.19242030617352	*12.3454434647511*

Coordinates of the centres of fuzzy clusters in question

47.6689791526396	*1.0309781667094*
177.773326979059	*2.41751282560672*

Figure 1. The result of the solution of the problem for fuzzy clustering for 4 groups of experiments, each of them with 2 fuzzy clusters

4. Conclusion

The results of the fuzzy clustering have approximate character and can only serve for preliminary structuring of the information, which exists in the multitude of output data. By solving problems of fuzzy clustering, it is necessary to remember the peculiarities and the restrictions of the process for measuring the signs in the set of the objects of the clustering. Due to the fact that the fuzzy clusters are formed on the basis of the Euclidean metric, the corresponding space of signs must meet the axioms of metrical space. In the meantime, for searching regularities in a problem area, which have non-metrical character, it is necessary to use the special means and tools, developed for intellectual data analysis (Data Mining).

References

[1]. Rayzina, J.(2008). *Classification and cluster*. Mir.

[2]. Kofman, A. (1982). *Introduction to the theory of fuzzy sets*. Radio and Communications.

[3]. Kuzmin, V.B.(1982). *Construction group solutions in spaces of clear and fuzzy binary relations*. Nauka.

[4]. Leonenkov, A.V. (2003).*Fuzzy modeling in MATLAB and fuzzyTECH*. BHV-Petersburg.

[5]. Leonenkov, A.V.(2010). Algorithm of fuzzy cluster analysis in problems of structuring complex systems.*The collection of algorithms and programs of common tasks. Vol. 11*.74-79.

[6]. Mandel, N.D. (2008).*Cluster analysis*. Finance and Statistics.

[7]. Didz et al. (2005). *The method of analysis of spatial information.* Finance and Statistics.

[8]. Pospelov, D.A. (1986)*Fuzzy sets in management models and artificial intelligence*.Nauka.

[9]. Yager, R.R. (1986). *Fuzzy sets and possibility theory*. Radio and Communications.

[10]. Borisov, A.N. et al. (1989).*Treatment of fuzzy information in the decision-making systems.* Radio and Communications.
[11]. Orlovsky, S.A. (1981).*Decision making with fuzzy initial information*. Nauka.
[12]. Tzrano, T., K. Asai, M. Sugeno. (1993). *Applied fuzzy systems.* Mir.
[13]. Bezdek, J. C. (1980). A convergence theorem for the fuzzy ISODATA clustering algorithm. *IEEE Transactions on Pattern Analysis and Mashine Intelligence, vol. 2, no. 1*, 1-8.
[14]. Bezdek, J.C. (1983). Some recent applications of fuzzy c-means in pattern recognition and image processing. *IEEE Workshop on Languages for Automation*, 247-252.
[15]. Dunn, J. C. (1974). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal on Cybernetics, vol. 3, no. 3*, 32-37
[16]. Ross, T.J. (2005).*Fuzzy logic with engineering applications.* McGraw-Hill.
[17]. Sugeno, M. (1977). Fuzzy measures and fuzzy integrals: a survey, *Fuzzy Automata and Decision Processes*, North-Holland, New York,89-102.
[18]. Takagi, T., M. Sugeno. (1985). Fuzzy identification of systems and its applications to modeling and control. *IEEE Transactions on Systems, Man, and Cybernetics, vol. 15, no.1*, 116-132.
[19]. Windham, M. P. (2002). Cluster validity for the fuzzy c-means clustering algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 4, no. 4*, 357-363.
[20]. Windham, M. P. (2003). Cluster validity for the fuzzy c-means clustering algorithm. *Fuzzy Sets and Systems, vol. 10, no. 3*, 271-279.

Corresponding author: Georgi Krastev
Institution: Department of Computing,
University of Ruse, Bulgaria
E-mail: GKrastev@ecs.uni-ruse.bg