Розроблено чіткі та нечіткі методи кластеризації, засновані на нечітких бінарних відношеннях. Наведено приклади нечітких бінарних відношень, що характеризують подібність об'єктів за довжиною, кутом та відстанню їх векторних ознак. Проілюстровано роботу представлених методів на прикладі. Використання розроблених методів для розв'язання реальних прикладних задач показало їх ефективність та простоту у застосуванні

Ключові слова: кластерний аналіз, автоматична класифікація об'єктів, інтелектуальний аналіз даних, нечітка кластеризація

Разработаны четкие и нечеткие методы кластеризации объектов, основанные на нечетких бинарных отношениях. Приведены примеры нечетких бинарных отношений, характеризующих сходство объектов по длине, углу и расстоянию их векторных признаков. Проиллюстрирована работа представленных методов на примере. Использование разработанных методов для решения реальных прикладных задач показало их эффективность и простоту в применении

Ключевые слова: кластерный анализ, автоматическая классификация объектов, интеллектуальный анализ данных, нечеткая кластеризация

UDC 519.8
DOI: 10.15587/1729-4061.2017.94961

CLUSTERING METHOD BASED ON FUZZY BINARY RELATION

N. Kondruk
PhD, Associate Professor
Department of Cybernetics and Applied Mathematics
State Higher Education Institution «Uzhhorod National University»
Narodna sq., 3, Uzhgorod, Ukraine, 88000
E-mail: natalia.kondruk@uzhnu.edu.ua

1. Introduction

One of the most interesting and promising approaches to the analysis of multivariate phenomena and processes are methods of cluster analysis or automatic classification of objects. Clustering is one of the key areas of data mining. Its objective is identification of some unknown structure of a group of similar objects in the initial set.

The clear objective of cluster analysis formally refers to the problem of finding some multiple partition (coverage) of the initial set of objects into non-overlapping subsets. In particular, elements of one subset shall differ much less among themselves than those of different subsets. Subsets that possess this property are called clusters. In case of this partition, every object belongs only to one cluster.

Fuzzy clustering allows the same object to belong to several (or even all) clusters at the same time, but to a different extent. Such clustering in many situations is more "natural" than clear, for example, for the objects located on the cluster edge. Besides, the fuzzy approach to solving clustering problems in many cases allows splitting complex clusters and opens up new opportunities for interpretation of clustering results.

Cluster analysis is important and widely used in various fields. It provides an opportunity to find hidden groups of objects, to improve the analysts' perception of group data. Clustering may be applied for economics and sociology data, texts, news feeds, genetic sequences, images, social data, medical and biological indicators, etc.

2. Literature review and problem statement

For a more effective solution of emerging practical problems of cluster analysis, there is always the need for new clustering methods or various modifications of existing ones. Considering the needs of medicine [1], agriculture [2], economy [3] geodemography [4], various modifications of the optimization model of the fuzzy c-means algorithm are presented. The main shortcomings of c-means are the need for setting the optimum number of clusters, computing complexity, "noise" sensitivity, low rate of convergence and "sticking" at local minima. Various modifications eliminate some shortcomings due to complication of the clustering procedure.

In [5] two new optimization models, which allow determining the optimum number of clusters have been developed. The paper [6] deals with developing a clustering method, which is based on the bees optimization algorithm with setting the optimum number of clusters. The shortcomings of clustering optimization approaches are computing complexity, operation only with small data volume, the possibility of clustering only by one object similarity criterion.

In [7] the neutrosophic algorithm based on hierarchical clustering has been proposed. It can be applied to clustering of the data, set in the neutrosophic environment. In [8, 9] the methods of clustering of non-numerical input data based on the genetic approach have been proposed. The particularity of these methods is the possibility of clustering of objects, which are characterized by specifically set features and under uncertainty. Thus, object grouping can be carried only by one similarity criterion.

In [10] the BSP algorithm, which performs clustering not only on the basis of object features, but also depending on the relations between them has been considered. Application of this method for clustering of social networks has shown that its main drawback is high computing resource intensity, which complicates operation with large data sets.

It should be noted that different applied nature of input data, objectives, types of clustering leads to a fundamental impossibility to develop a single universal method of cluster analysis. But even within the same class of problems, the considered clustering methods carry out object grouping only by one similarity criterion (typically using a distance metric). Thus, only sphere- and ellipsoid-shaped clusters are formed. There are a large number of applied problems, including described in [11–13], where this type of object grouping is inadequate and ineffective.

Therefore, development of a flexible mathematical apparatus, which would have a rather wide range of means for object grouping by various criteria of similarity of their features, using distance, length and angular metrics is reasonable. This will allow an effective solution of rather broad classes of problems from various subject areas within the developed approach using various types of clustering. This will also allow clustering not only with ellipsoids but also cones and concentric spheres.

According to the author, the use of fuzzy binary relations for a mathematical description of various object similarity criteria will ensure the implementation of different types of clustering within a single approach. In addition, this will allow not only identifying a relationship between objects by different criteria, but also determining its degree, which gives the opportunity to develop a fuzzy clustering approach with all its benefits.

3. Mathematical statement of the fuzzy problem of cluster analysis

The lack of a single conventional statement of fuzzy modification of the cluster analysis problem requires a clear factual description of the research problem.

Let us consider the general problem of fuzzy cluster analysis in the following statement.

Let there be given some objects $O_1,..., O_m$, characterized by n quantitative features. Let $\left(c_1^i, c_2^i, \ldots, c_n^i\right)$ be the feature vector, characterizing the object with the number i. Thus, each object O_i, $i=\overline{1,m}$ definitely corresponds to the feature vector $\overline{c_i}\left(c_1^i, c_2^i, \ldots, c_n^i\right)$, $i=\overline{1,m}$.

It is necessary to partition the given objects O_i, $i=\overline{1,m}$ into homogeneous "similarity" groups (clusters) by all n features according to a particular one or more similarity criteria of objects and determine the degree of their "membership" in the obtained clusters. To do this from the mathematical point of view, the problem of fuzzy clustering of feature vectors $\overline{c_i}\left(c_1^i, c_2^i, \ldots, c_n^i\right)$, $i=\overline{1,m}$ shall be solved.

4. Goal and objectives

The research goal is to improve the efficiency of solving automatic classification (clustering) problems by developing clear and fuzzy clustering methods, which allow object grouping by the distance, angular and length similarity of vector features.

To achieve the goal, the following problems shall be solved:

– to give examples of membership functions, described by fuzzy binary relations that characterize different object similarity criteria;

– to develop clear and fuzzy methods for single-level and multi-level clustering based on fuzzy binary relations;

– to illustrate the methods on real applied problems.

5. Examples of constructing object similarity measures based on fuzzy binary relations

The cluster analysis method is based on the concepts of similarity of objects in their features. Determination of the most "similar" objects helps to partition a set into clusters (groups). In the majority of clustering methods, the similarity criterion of objects is "similarity" of their feature vectors, expressed by the distance between them, which is the basis for constructing various types of metrics.

The research [14] has shown that the apparatus of fuzzy binary relations is useful in setting measures of similarity between objects. Let some fuzzy binary relation R be given on the set of vector features

$$C=\left\{\overline{c_i} \mid i=\overline{1,m}\right\}$$

with the membership function $\mu\left(\overline{c_i}, \overline{c_j}\right)$, where $\mu: C^2 \rightarrow [0,1]$, characterizing object similarity by some criterion. The closer the value of $\mu\left(\overline{c_i}, \overline{c_j}\right)$ to 1, the more similar the objects O_i and O_j according to this criterion. In addition, Core(R)≠Ø, since each object is similar to itself.

Setting fuzzy sets similarity functions is not unequivocal and can be expressed by different kinds of functions. In particular, to determine the "distance" similarity measure, it is proposed to introduce the fuzzy binary relation R^V with the membership function

$$\mu_{R^V}: C^2 \rightarrow \left[\frac{1}{e}, 1\right]:$$

$$\mu_{R^V}\left(\overline{c_i}, \overline{c_j}\right)=e^{-\left(\frac{\rho\left(\overline{c_i}, \overline{c_j}\right)}{\Delta}\right)}, \tag{1}$$

where

$$\Delta=\max_{i,j=\overline{1,m}} \rho\left(\overline{c_i}, \overline{c_j}\right),$$

where ρ is some distance metric (Euclidean, Manhattan, Chebyshev, etc). Thus, set function μ_{R^V} will belong to the class of Harrington type desirability functions [15]. The closer the points corresponding to the feature vectors of two objects O_i and O_j, the closer the value $\mu_{R^V}\left(\overline{c_i}, \overline{c_j}\right)$ to 1. For the most "distant" objects, this value will be equal to $\frac{1}{e}$.

There are classes of applied problems, the solution of which requires clustering based on the "angular", "length" [11–13] and other measures of similarity between the feature vectors of the corresponding objects.

To determine the "length" similarity measure of feature vectors of objects, it is proposed to use the binary relation R^D with the Harrington type membership function

$$\mu_{R^D}: C^2 \rightarrow \left[\frac{1}{e}, 1\right]:$$

$$\mu_{R^D}\left(\overline{c_i}, \overline{c_j}\right)=e^{-\frac{\left|\left\|\overline{c_i}\right\|-\left\|\overline{c_j}\right\|\right|}{\Delta}}, \tag{2}$$

where

$$\Delta=\max_i \left\|\overline{c_i}\right\|, \quad i=\overline{1,m}, \quad j=\overline{1,m}.$$

The defined fuzzy binary relation R^D characterizes the difference between the lengths of vectors $\overline{c_i}$ and $\overline{c_j}$. Moreover, the smaller the difference between the lengths, the closer μ_{R^D} to 1.

The binary relation R^K with the membership function

$$\mu_{R^K}: C^2 \to [0,1]$$

is determined by the formula [14]:

$$\mu_{R^k}\left(\overline{c_i},\overline{c_j}\right)=\frac{1+\dfrac{\overline{c_i}\cdot\overline{c_j}}{\left\|\overline{c_i}\right\|\cdot\left\|\overline{c_j}\right\|}}{2},\quad i=\overline{1,m},\quad j=\overline{1,m}. \tag{3}$$

It describes the angle between the vectors $\overline{c_i}$ and $\overline{c_j}$. Obviously, the smaller the angle between $\overline{c_i}$ and $\overline{c_i}$, the closer the value μ_{R^K} to 1, and vice versa, the larger the angle, the closer μ_{R^K} to zero.

6. Clear single-level clustering method

Let it be required to perform single-level object clustering by a certain similarity criterion, described by some fuzzy binary relation R with the membership function $\mu_R\left(\overline{c_i},\overline{c_j}\right)$. Also, the specified numerical value $\mu_R^* \in [0,1]$ is the clustering threshold. It characterizes the required similarity degree of objects within one cluster. If $\mu_R^*=0$, the object similarity degree is considered the weakest, which in turn will lead to the formation of a cluster, which will include all objects. If $\mu_R^*=1$, then, on the contrary, each object will form a separate cluster as the object similarity degree will be very high. Thus, the closer μ_R^* to unity, the more clusters formed as a result of clustering.

For a more accurate determination of numerical values of thresholds for the Harrington type membership functions, it is proposed to be guided by the psychophysical Harrington desirability scale [15]. It establishes a correspondence between physical and psychophysical parameters and is widely and effectively used in solving practical problems. It can be considered that the threshold values provide:

– from the interval [0.8, 1] – high object similarity degree;

– from the interval [0.63, 0.8] – average;

– from the interval [0.37, 0.63] – low similarity degree.

It should be noted that this scale is only tentative and needs to be clarified for each specific problem.

In the development of the method, the following heuristics were also taken into account: a new cluster shall be formed based on two "most similar" ungrouped objects.

The following describes a clear single-level object clustering method.

Let

$$\Omega_1=\{1, 2, ..., m\}.$$

Next, the l-th iteration is described step by step.

Step 1. We choose the dominant centroid vector $\overline{c_l^*}$ from the set $\left\{\overline{c_i} \mid i\in\Omega_l\right\}$, around which the l-th cluster will be built. This vector can be arbitrary or determined by a specific rule. In particular, given the heuristics, it is recommended to assign as dominant one of the vectors $\left\{\overline{c_i} \mid i\in\Omega_l\right\}$, which provides the maximum value of the function μ_R in relation to all other vectors from $\left\{\overline{c_i} \mid i\in\Omega_l\right\}$. All $\overline{c_i}\in\Omega_l$, for which

$$\mu_R\left(\overline{c_i},\overline{c_l^*}\right)\geq\mu_R^*$$

is valid are included in the conventional cluster U^l.

Step 2. The procedure of alignment of the cluster U^l is performed. For this, the centroid vector is recalculated, for example, by the formula:

$$\overline{c_l^*}=\frac{\sum_{\overline{c_i}\in U^l}\overline{c_i}}{\left\|U^l\right\|}.$$

Items with U^l: $U^l=\varnothing$ are cleared. Further, a new refined cluster is formed

$$U^l=\left\{\overline{c_i}\in\Omega_l \,\middle|\, \mu_R\left(\overline{c_i},\overline{c_l^*}\right)\geq\mu_R^*\right\}.$$

The procedure of alignment of the cluster U^l is held until the $\overline{c_l^*}$ coordinates change in recalculation. Otherwise, we proceed to the next step.

Step 3. The procedure of refinement of the cluster U^l is performed. For this, the vector $\overline{c_p}$, which is "the most similar" to the vector $\overline{c_l^*}$, is determined from the set Ω_l/U^l, i. e.

$$\mu_R\left(\overline{c_p},\overline{c_l^*}\right)=\max_{i\in\Omega_l/U^l}\mu_R\left(\overline{c_i},\overline{c_l^*}\right).$$

And the possibility of joining $\overline{c_p}$ to U^l is checked. For the set $U^l\cup\left\{\overline{c_p}\right\}$, the centroid vector is calculated, for example

$$\overline{c_l^{**}}=\frac{\sum_{\overline{c_i}\in U^l\cup\left\{\overline{c_p}\right\}}\overline{c_i}}{\left\|U^l\right\|+1}.$$

If the set

$$\left\{\overline{c_i}\in U^l\cup\left\{\overline{c_p}\right\} \,\middle|\, \mu_R\left(\overline{c_i},\overline{c_l^{**}}\right)\geq\mu_R^*\right\}$$

is equal to the set $U^l\cup\left\{\overline{c_p}\right\}$, the vector $\overline{c_p}$ is included in the set U^l, and the centroid vector $\overline{c_l^*}$ is recalculated, i. e. $\overline{c_l^*}=\overline{c_l^{**}}$. If

$$\left\{\overline{c_i}\in U^l\cup\left\{\overline{c_p}\right\} \,\middle|\, \mu_R\left(\overline{c_i},\overline{c_l^{**}}\right)\geq\mu_R^*\right\}\neq U^l\cup\left\{\overline{c_p}\right\},$$

the procedure of refinement of the cluster U^l is completed.

Step 4. We form the cluster

$$K^l=U^l$$

and

$$\Omega_{l+1}=\Omega_l/K^l.$$

If $\Omega_{l+1}\neq\varnothing$, we proceed to step 1. Otherwise, the clustering procedure is completed.

The result of the method will be clear clusters $K^1, K^2, ..., K^z$, where $z\leq m$ with the appropriate representatives $\overline{c_l^*}$.

The convergence of the method is guaranteed by the condition $\mathrm{Core}(R)\neq\varnothing$.

It should be noted that the formulas of centroid vectors of cluster representatives can be modified depending on the considered types of similarity measures (angular, length).

The method, unlike the majority of cluster analysis methods, requires no prior information on the desired number of clusters, but needs information about the threshold value μ_R^*. Thus, gradually changing the numerical value of μ_R^*, it is possible to get complete information about the formation dynamics of clusters, changes in their number and degree of relationships between objects.

7. Multi-level sequential clear clustering method

Let it be required to perform multi-level clear clustering of objects $O_1,\ldots, O_m$ by some s object similarity criteria, specified by fuzzy binary relations R_k, $k=\overline{1,s}$ with the membership functions $\mu_{R_k}\left(\overline{c_i},\overline{c_j}\right)$ and the clustering thresholds $\mu_{R_k}^* \in [0,1]$, $k=\overline{1,s}$. be determined for each R_k, $k=\overline{1,s}$.

This method requires object similarity criteria pre-ordering by preference in the form of a certain ranking of the respective binary relations: $R_1 \succ R_2 \succ \ldots \succ R_s$.

The first iteration.

Clustering of the initial set of objects is performed by the clear single-level clustering method for the binary relation R_1 with the membership function $\mu_{R_1}\left(\overline{c_i},\overline{c_j}\right)$ and the threshold $\mu_{R_1}^*$. Let the clusters $K_1^1, K_1^2,\ldots,K_1^{z_1}$. be obtained.

Then the l-th iteration is described step by step.

To each cluster of the previous iteration

$$K_{l-1}^p,\ p=\overline{1,z_{l-1}}$$

the clear single-level clustering method is applied: filtering objects of each cluster by the criterion, expressed by the binary relation R_l with the membership function $\mu_{R_l}\left(\overline{c_i},\overline{c_j}\right)$, which is not lower than the threshold $\mu_{R_l}^*$. Thus, the z_{l-1} clustering problem with the original set Ω_l, consisting only of the numbers of objects included in the separate cluster K_{l-1}^p is solved. As a result, we get z_{l-1} sets of new smaller clusters. By re-enumerating all the obtained clusters we get $K_l^1, K_l^2,\ldots,K_l^{z_l}$.

Exactly *s* described iterations shall be performed in this method. So, by re-assigning the resultant clusters, we get clear sets of partition of the initial set of objects – K^1, K^2, ..., K^z, where $z \le m$ with the appropriate representatives $\overline{c_l^*}$.

8. Fuzzification of clusters

Let it be required to perform single-level fuzzy object clustering by a certain object similarity criterion, specified by some fuzzy binary relation R with the membership function $\mu_R\left(\overline{c_i},\overline{c_j}\right)$ and the threshold $\mu_R^* \in [0,1]$. Then, according to the clear single-level method in p. 6, we perform clustering into clear clusters K^1, K^2, ..., K^z, $z \le m$ with the appropriate representatives $\overline{c_1^*},\overline{c_2^*},\ldots,\overline{c_z^*}$. The membership functions

$$\tilde{\mu}_j : C \to [0,1]$$

of the fuzzified clusters $\tilde{K}^j$, $j=\overline{1,z}$ are proposed to be determined by the formulas:

$$\tilde{\mu}_j\left(\overline{c_i}\right)=\mu_R\left(\overline{c_i},\overline{c_j^*}\right), \quad (5)$$

or

$$\tilde{\mu}_j\left(\overline{c_i}\right)=\exp\left(-\frac{\left[1-\mu_R\left(\overline{c_i},\overline{c_j^*}\right)\right]^2}{\beta}\right), \quad (6)$$

where β is an extension factor.

When using the Harrington type membership functions (1), (2), the formula (5) is not conventional for fuzzification of data, such as

$$\tilde{\mu}_j : C \to \left[\frac{1}{e};1\right],$$

i. e. the degree of membership of the most distant objects in the j-th cluster will be not less than the number $\frac{1}{e}$. But in this case, it is possible to make an effective analysis of clustering results and numerical values of $\tilde{\mu}_j$ according to the Harrington desirability scale. The use of the Gaussian type membership function (6) leads to its normalization, i. e. $\tilde{\mu}_j : C \to (0;1]$, and the coefficient β in the exponent argument is proposed to be determined by the three-sigma rule. In particular, the value of β is calculated for the Harrington type membership functions (1), (2) – β=0.0882.

Let it be required to perform multi-level fuzzy clustering by some $s \ge 2$ object similarity criteria, specified by fuzzy binary relations R_k, $k=\overline{1,s}$ with the membership functions $\mu_{R_k}\left(\overline{c_i},\overline{c_j}\right)$ and the clustering thresholds $\mu_{R_k}^* \in [0,1]$, $k=\overline{1,s}$ be determined for each R_k, $k=\overline{1,s}$. By fuzzy clustering we mean partition of the initial set of objects O_i, $i=\overline{1,m}$ into fuzzy clusters $\tilde{K}^j$, $j=\overline{1,z}$, where each object O_i is included in each fuzzy cluster $\tilde{K}^j$ with the corresponding membership function $\tilde{\mu}_j : C \to [0,1]$.

Let the sequential multi-level clustering method in p. 7 provide clear clusters K^1, K^2,..., K^z, $z \le m$. To determine the numerical value of the membership degree of the cluster K_j, $j=\overline{1,z}$ of the object O_i, $i=\overline{1,m}$, it is proposed to use the Gaussian membership function:

$$\tilde{\mu}_j\left(\overline{c_i}\right)=\exp\left(-\frac{\sqrt{\sum_{l=1}^{s}\alpha_l\left[\mu_{R_l}\left(\overline{c_i},\overline{c_j^*}\right)-1\right]^2}}{\beta}\right), \quad (7)$$

where

$$\alpha_l=\frac{\mu_{R_l}^*}{\sum_{k=1}^{s}\mu_{R_k}^*},\quad l=\overline{1,s},\quad i=\overline{1,m},\quad j=\overline{1,z},$$

where β is an extension factor. The values α_l, $l=\overline{1,s}$ specify the necessary normalized object similarity degree by the l-th criterion within each cluster.

9. Computing experiment

The single-level clustering method for the angular similarity of objects R^k was used in solving multicriteria linear programming problems with a high-dimensional criteria space. They occur in mathematical modeling of balanced diet problems [11, 13]. One of the stages of solving such problems is clustering of their criteria space. In this case, the relations between criteria are determined by their angular similarity.

Application of the proposed approach has proved to be effective and convenient in solving the problems of efficiency criteria grouping into sets of strongly related and contradictory criteria [12].

A software system that enables to solve various cluster analysis problems, both single-level, and multi-level clear and fuzzy clustering, was also developed based on the presented methods.

Verification of the presented methods was conducted on many clustering problems in a two-dimensional space where the object similarity groups can be represented graphically, making it possible to estimate the results.

In addition, from the epistemological standpoint, the quality of machine clustering is determined by its compliance with the man-done classification. So, in case of significant similarity of man-done and machine clustering, general conclusions about the procedure correctness can be drawn. This leads to the need for testing clustering methods on real data with the known structure. Also [16] summarizes the data quality requirements for verification of experimental studies of clustering and presents the applied problem that satisfies them (Table 1).

Table 1

Some specifications of the USAF aircrafts

Object No.	Aircraft type	Flight performance				
		Wing span, m	Length, m	Height, m	Estimated takeoff weight, kg	Maximum range, km
1	F–104G	6.68	16.69	4.11	9428	1200
2	F–5A	7.70	14.68	4.06	6080	898
3	F–5E	8.13	14.68	4.06	7030	1083
4	F–16A	9.45	14.52	5.01	10205	925
5	F–15A	13.05	19.43	5.63	18824	1050
6	F–4B	11.70	17.76	4.96	20865	1600
7	F–4E	11.77	19.19	5.02	18818	1266
8	F–106A	11.67	21.56	6.18	16100	920
9	RA–5C	16.15	23.11	5.92	30300	1600
10	SR–71A	16.95	32.74	5.64	63505	1930
11	B–58A	17.32	29.49	9.53	68000	2600

According to [16], the specifications presented in Table 1 allowed experts to conventionally group these aircrafts into 3 classes (clusters):

– class 1 – light aircrafts: objects 1–4;
– class 2 – average: objects 5–9;
– class 3 – heavy aircrafts: objects 10, 11.

To illustrate the proposed method in the examined cluster analysis problem, single-level clustering from p. 6 was chosen. Further, the results were fuzzified according to p. 8. The fuzzy binary relation R^V "distance" with the membership function described by the formula (1) was used. The Euclidean distance was taken as ρ.

The input data were normalized data in Table 1. For the interpretation of fuzzy clustering results, matrices of fuzzy object distribution among clusters were used. For a visual representation of the results, the line chart of fuzzy partitions was constructed. The membership values are given on the y-axis of the chart, and the numbers of objects – on the x-axis. The cluster object membership is determined by the intersection point of the lines that corresponds to the object number and object degree of membership in a cluster. The cluster number is specified near the point.

In the study of the clustering process dynamics, given data set, different thresholds (accurate to 0.01) were chosen. Sample results of clear clustering are given in Table 2.

Table 2

The results of clear clustering of the considered set

Numerical values of the clustering threshold	Clustering results
$\mu^*_{R^V} \in [0.57;0.72]$	Cluster 1: objects 1–9 Cluster 2: objects 10, 11
$\mu^*_{R^V} \in [0.73;0.83]$	Cluster 1: objects 1–8; Cluster 2: object 9; Cluster 3: objects 10, 11
$\mu^*_{R^V} = 0.84$	Cluster 1: objects 1–4; Cluster 2: objects 5–8; Cluster 3: object 9; Cluster 4: objects 10, 11
$\mu^*_{R^V} \in [0.85;0.88]$	Cluster 1: objects 1–4; Cluster 2: objects 5–8; Cluster 3: object 9; Cluster 4: object 10; Cluster 5: object 11

According to the Harrington desirability scale, the results of clustering in the transition from the average object similarity degree to high (in the transition of numerical values of the thresholds $\mu^*_{R^V}$ from the interval [0.63, 0.8] to [0.8, 1]) are the most significant. In the context of this problem, a high object similarity degree is achieved with $\mu^*_{R^V} \geq 0,84,$ so clustering with $\mu^*_{R^V} = 0,84.$ can be considered the solution of the problem. Let us analyze the matrix of the corresponding fuzzy partition (Table 3) and its line chart (Fig. 1) using the formula (5).

Table 3

The matrix of fuzzy partition of the examined set of objects with $\mu^*_{R^V} = 0,84$ with a degree of membership determined by the formula (5)

Object No.	Cluster 1	Cluster 2	Cluster 3	Cluster 4
1	0.92	0.77	0.42	0.61
2	0.95	0.76	0.40	0.60
3	0.97	0.78	0.41	0.62
4	0.91	0.82	0.43	0.65
5	0.76	0.94	0.51	0.78
6	0.77	0.88	0.53	0.78
7	0.80	0.96	0.51	0.77
8	0.77	0.89	0.50	0.74
9	0.62	0.78	0.63	1.00
10	0.44	0.54	0.84	0.66
11	0.38	0.47	0.84	0.57

In Fig. 1 dashed lines are conventional boundaries of different degrees of object similarity within one cluster, clarified for the considered problem. The chart in Fig. 1 shows that the object 9 has an average degree of similarity with the objects of the second cluster. And the objects of the second cluster have an average degree of similarity with the objects of the first cluster. The similarity of the objects in the fourth cluster is high, but the weakest among others.

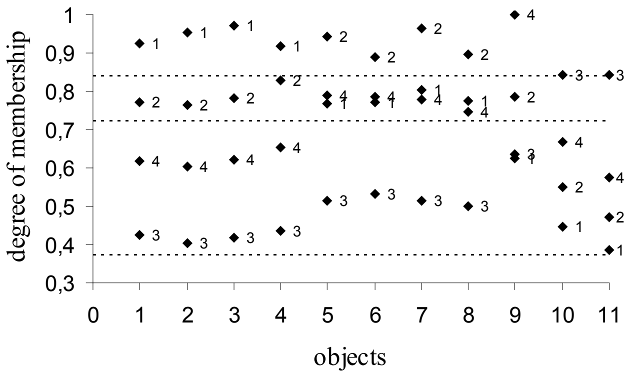

Fig. 1. The line chart of fuzzy partition of the examined set with a degree of membership determined by the formula (5)

Further, the matrix of the corresponding fuzzy partition (Table 4) and its line chart (Fig. 2) are presented using the normalization formula (6).

Table 4

The matrix of fuzzy partition of the examined set of objects with $\mu^*_{R^V} = 0,84$ with a degree of membership determined by the formula (6)

Object No.	Cluster 1	Cluster 2	Cluster 3	Cluster 4
1	0.94	0.55	0.02	0.19
2	0.98	0.53	0.02	0.17
3	0.99	0.59	0.02	0.19
4	0.93	0.71	0.02	0.25
5	0.54	0.96	0.07	0.6
6	0.55	0.86	0.08	0.59
7	0.64	0.98	0.06	0.57
8	0.56	0.89	0.05	0.47
9	0.2	0.59	0.22	1
10	0.03	0.09	0.75	0.28
11	0.01	0.04	0.75	0.12

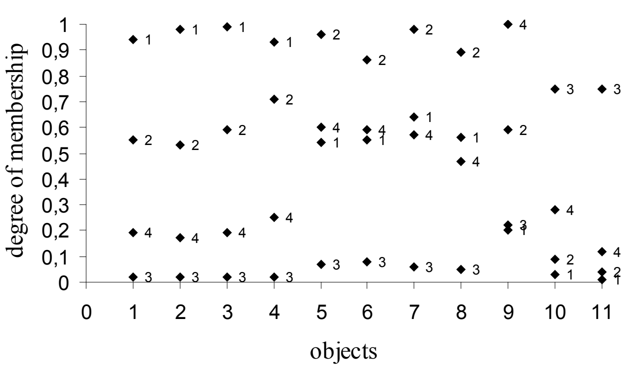

Fig. 2. The line chart of fuzzy partition of the examined set with a normalized degree of membership determined by the formula (6)

For the analysis of the results, it will be appropriate to present the main results of fuzzy clustering of the examined set described in [16]. Fuzzy c-means clustering into three clusters coincided fully with the expert one. Partition into 4 clusters is the same as with $\mu^*_{R^V} = 0,84$ (Table 2). The conclusion is made that the object 9 takes an intermediate position between the cluster of objects 1–8 and the cluster of objects 10, 11.

Using the Pedrycz optimization method [17] with the objects marked with numbers 11, 3, 6 and the number of clusters 3, the following results are obtained:

– cluster 1 – objects 1–4 and 8;
– cluster 2 – objects 5–7 and 9;
– cluster 3 – objects 10 and 11.

The drawback of the method is that the results of clustering depend on the initially marked objects.

According to the Windham optimization method [18], only clusters with objects 1–4 and 9–11 were clearly identified when partition into three and four clusters, all other objects appeared evenly distributed among all clusters.

All these optimization methods require pre-setting the desired number of clusters and can be used with a small number of objects.

The result of the hierarchical version of the Tamura-Higuchi-Tanaka algorithm [19] that the most corresponds to the expert one:

– cluster 1 – objects 1–4;
– cluster 2 – objects 5–8;
– cluster 3 – object 9;
– cluster 4 – object 10;
– cluster 5 – object 11.

10. Discussion of the experimental results

The diverse nature of the data and the purposes of clustering, the existence of various geometric forms of clusters leads to impossibility of existence of a universal method of clustering. Therefore, comparison of cluster analysis methods is not quite correct. Each has advantages and disadvantages and can be effective in solving a certain class of problems.

In particular, the experimental studies demonstrated the convenience and efficiency of single-level and multi-level sequential clustering methods for solving various classes of clustering problems. It is possible to determine their main advantages:

– the developed methods allow object clustering not only by the distance degree of similarity, but also length and angular;

– application of the Harrington type membership functions allows an effective analysis of the resulting object partition;

– a variety of clustering thresholds provides an additional opportunity to observe the formation dynamics, structural changes of clusters and reveal hidden relations between objects;

– the multi-level sequential clustering method enables object clustering by various similarity criteria;

– the developed methods can be used for preliminary data analysis and for holding the clustering procedure.

All this makes it possible to apply single-level and multi-level sequential clustering methods for solving a wider range of applied problems.

The research is a continuation and generalization of [12, 14]. Also, refinement of the developed mathematical apparatus is planned:

– for implementing parallel multi-level clustering and its application;

– for solving an even wider range of problems involving the length similarity measure of objects.

11. Conclusions

1. Examples of membership functions described by fuzzy binary relations, which characterize the distance, angular and length similarity measures of vector features of objects are given. Application of the Harrington type functions gave an additional opportunity of effective interpretation of the obtained results of clustering.

2. Single-level and multi-level clustering methods are developed. They made it possible to perform clear and fuzzy clustering of objects according to one or more similarity criteria. Thus, certain values, such as clustering thresholds, characterizing the degree of similarity of objects in a cluster are set. By changing the clustering thresholds, one can analyze the formation dynamics of clusters, investigate their structure and relations between objects. The presented methods allow clustering in the absence of additional a priori information, so they can be used in preliminary data analysis.

3. The software systems that implement the presented methods are developed. Testing of systems showed their effectiveness in dealing with a fairly broad class of applied problems of cluster analysis. The single-level fuzzy method is illustrated on a real problem.

References

1. Alsmadi, M. K. A hybrid firefly algorithm with fuzzy-c mean algorithm for MRI brain segmentation [Text] / M. K. Alsmadi // American Journal of Applied Sciences. – 2014. – Vol. 11, Issue 9. – P. 1676–1691. doi: 10.3844/ajassp.2014.1676.1691
2. Giusti, E. A Fuzzy Decision Support System for irrigation and water conservation in agriculture [Text] / E. Giusti, S. Marsili-Libelli // Environmental Modelling & Software. – 2015. – Vol. 63 – P. 73–86. doi: 10.1016/j.envsoft.2014.09.020
3. Egrioglu, E. Fuzzy time series forecasting with a novel hybrid approach combining fuzzy c-means and neural networks [Text] / E. Egrioglu, C. H. Aladag, U. Yolcu // Expert Systems with Applications. – 2013. – Vol. 40, Issue 3. – P. 854–857. doi: 10.1016/j.eswa.2012.05.040
4. Son, L. H. A novel intuitionistic fuzzy clustering method for geo-demographic analysis [Text] / L. H. Son, B. C. Cuong, P. L. Lanzi, N. T. Thong // Expert Systems with Applications. – 2012. – Vol. 39, Issue 10. – P. 9848–9859. doi: 10.1016/j.eswa.2012.02.167
5. Saha, S. Exploring differential evolution and particle swarm optimization to develop some symmetry-based automatic clustering techniques: application to gene clustering [Text] / S. Saha, R. Das // Neural Computing and Applications. – 2017. doi: 10.1007/s00521-016-2710-0
6. Ouadfel, S. Handling Fuzzy Image Clustering with a Modified ABC Algorithm [Text] / S. Ouadfel, S. Meshoul // International Journal of Intelligent Systems and Applications. – 2012. – Vol. 4, Issue 12. – P. 65–74. doi: 10.5815/ijisa.2012.12.09
7. Sahin, R. Neutrosophic Hierarchical Clustering Algoritms [Text] / R. Sahin // Neutrosophic Sets and Systems. – 2014. – Vol. 2. – P. 18–24.
8. Abad, M. J. F. H. Heuristic Model of Cellular Learning Automata for Fuzzy Rule Extraction [Text] / M. J. F. H. Abad, P. Derakhshan-Barjoei // Research Journal of Applied Sciences, Engineering and Technology. – 2012. – Vol. 4, Issue 12. – P. 1701–1707.
9. Demin, D. A. Mathematical modeling in the problem of selecting optimal control of obtaining alloys for machine parts in uncertainty conditions [Text] / D. A. Demin // Problems of mechanical engineering. – 2013. – Vol. 16, Issue 6. – P. 15–23.
10. Shmalyuk, I. Yu. Software implementation of the BSP algorithm for clusterization of social networks [Text] / I. Yu. Shmalyuk, I. M. Bushyn // Technology audit and production reserves. – 2015. – Vol. 2, Issue 2 (22). – P. 21–26. doi: 10.15587/2312-8372.2015.40779
11. Kondruk, N. E. Application of multicriteria models for the problems of a balanced diet [Text] / N. E. Kondruk, M. M. Malyar // Bulletin of Cherkasy State Technological University. Series: Engineering Sciences. – 2010. – Issue 1. – P. 3–7.
12. Kondruk, N. E. Some applications of clustering criterion space for selection tasks [Text] / N. E. Kondruk, N. N. Malyar // Computer Mathematics. – 2009. – Issue 2. – P. 142–149.
13. Kondruk, N. E. Decision Support Systems for automated diets [Text] / N. E. Kondruk // Management of Development of Complex Systems. – 2015. – Issue 23 (1). – P. 110–114.
14. Kondruk, N. E. Some methods of automatic grouping of objects [Text] / N. E. Kondruk // Eastern-European Journal of Enterprise Technologies. – 2014. – Vol. 2, Issue 4 (68). – P. 20–24. – Available at: http://journals.uran.ua/eejet/article/view/22930/20939
15. Harrington, E. C. The Desirability Function [Text] / E. C. Harrington // Industrial Quality Control. – 1965. – P. 494–498.
16. Vyatchenin, D. A. Nechetkie metody avtomaticheskoj classifikacii [Text]: monografiya / D. A. Vyatchenin. – Minsk: UP Tehnoprint, 2004. – 219 p.
17. Pedrycz, W. Algorithms of fuzzy clustering with partial supervision [Text] / W. Pedrycz // Pattern Recognition Letters. – 1985. – Vol. 3, Issue 1. – P. 13–20. doi: 10.1016/0167-8655(85)90037-6
18. Windham, M. P. Cluster validity for fuzzy clustering algorithms [Text] / M. P. Windham // Fuzzy Sets and Systems. – 1981. – Vol. 5, Issue 2. – P. 177–185. doi: 10.1016/0165-0114(81)90015-4
19. Tamura, S. Pattern Classification Based on Fuzzy Relations [Text] / S. Tamura, S. Higuchi, K. Tanaka // IEEE Transactions on Systems, Man, and Cybernetics. – 1971. – Vol. SMC-1, Issue 1. – P. 61–66. doi: 10.1109/tsmc.1971.5408605