Vadim Nikolayevich Romanov. International Journal of Engineering Research and Applications
www.ijera.com
ISSN: 2248-9622, Vol. 13, Issue 4, April 2023, pp. 187-198

RESEARCH ARTICLE OPEN ACCESS

Use of fuzzy arithmetic in classification and cluster analysis problems

Vadim Nikolayevich Romanov
Doctor of Technical Sciences, Professor
St.-Petersburg, Russia

ABSTRACT
The article examines the use of fuzzy arithmetic in the problems of classification and cluster analysis based on the representation of data in the form of fuzzy gradations proposed by the author. The advantages of the proposed approach are analyzed. This approach allows us to expand the range of solvable problems, to increase the reliability of the distribution of objects by classes and reduce the ambiguity of the distribution of objects by clusters and levels of order. It makes possible to substantiate the choice of a measure of similarity between objects, to smooth the influence of errors associated with data inconsistency; at the same time, the complexity of analysis and calculations is significantly reduced. Examples are considered to illustrate the application of the proposed approach.

Keywords and phrases: fuzzy arithmetic, classification, cluster analysis.

Date of Submission: 03-04-2023 Date of acceptance: 15-04-2023

I. Introduction

A large number of works are devoted to various aspects of classification and cluster analysis problems, due to their importance for obtaining, intelligent analysis, presentation and processing of data and knowledge in various fields, in particular, in biology, medicine, psychology, sociology, economics, etc. [1 – 6]. Currently, various methods of classification and cluster analysis have been developed, for example, hierarchical algorithms, grid methods, partitioning-based technique, density based method, etc. The most advanced are statistical methods of classification and cluster analysis [2]. However, their application requires a large amount of data, which is not always possible in practice. In this case, fuzzy models are applied [1, 3]. The use of fuzzy classification models and cluster analysis in solving diagnostic problems is considered in [1, 3 – 5]. A comparative analysis of clustering methods is given in [6]. Fuzzy clustering methods contain fuzzy C-Shell clustering algorithms [7], Mountain method [8], Gustafson Kessel and Fuzzy-C-means [9]. There is a tendency towards complication of methods, in particular, the use of genetic algorithms [10], artificial intelligence methods [11, 12], neural networks [13, 14]. However, the complication of methods does not lead to an increase in the accuracy and reliability of the results, since the methods of fuzzy classification and clustering are based on the Zadeh generalization principle and the definition of the membership function. Various forms of cluster analysis have the following disadvantages and limitations. 1). The use only quantitative values of criteria and statistical methods of data processing. 2). Strong dependence of the analysis results on the choice of the measure of similarity between objects. 3). The use as a measure of similarity between objects the mean or root-mean-square distance and the approximation of "nearest neighbors" with subsequent averaging over the combined objects, which leads to the loss of useful information and unreasonable conclusions about the number and composition of clusters and levels of order. 4). Analysis errors caused by using a nontransitive resemblance relation instead of a transitive similarity relation.

The purpose of this article is to study fuzzy classification and cluster analysis models based on the approach proposed by the author using fuzzy gradations [15, 16], which eliminates the limitations noted above, simplifies calculations and makes their results more understandable.

The article is organized as follows. First, we briefly consider the rules of fuzzy arithmetic and the evaluation of the reliability of the results. The classification algorithm is generalized for the case of data representation in the form of fuzzy gradations. The advantages of the proposed approach to the classification problem are shown using an example. Further, a generalization of the cluster analysis algorithm based on fuzzy gradations and an example of its application are given. The influence of the choice of the measure of similarity between objects on the results of classification and cluster analysis is investigated. Finally, we discuss the obtained results and give recommendations for their application.

II. Arithmetic operations on fuzzy gradations

To describe the object area we use the fuzzy gradations in the range VL…VH. The range includes gradations VL – very low value, (VL-L) – between very low and low, L – low, (L-M) – between low and middle, M – middle (medium), (M-H) – between middle and high, H – high, (H-VH) – between high and very high, VH – very high. Boundary gradations out of range are also known, namely VVL (lowest value) and VVH (highest value). Depending on condition of the task gradation VVL can be interpreted as zero, lower bound, exact lower bound etc. and gradation VVH – as unit, infinity, upper bound, exact upper bound, etc. On the set of fuzzy gradations, a relation of a non-strict order (preference relation) is established, which has the properties of reflexivity, antisymmetry and transitivity. Fuzzy gradations form an ordinal scale in which an admissible transformation is arbitrary monotone function that does not change the order of gradations. In particular, all gradations can be simultaneously multiplied or divided, as well as increased or decreased by the same number. The summation and multiplication operations on fuzzy gradations are performed in the same way as in ordinary arithmetic. For instance, for summation we have VL + VL = (VL-L), (VL-L) + VL = L, etc., (H-VH) + VL = VH, VH + VL = VVH. Similarly, summation is performed for other fuzzy gradations. For multiplication operation we have VL * VL = VVL, VL * (VL-L) = VVL, etc., VL * M = VL, etc., VL * VH = VL. Similarly, multiplication is performed for other fuzzy gradations. Note that operation of multiplication belong to Archimedean operations and the summation operation within the range under consideration belongs to nilpotent operations. Fuzzy gradations form an Abelian addition group and an Abelian semigroup for multiplication. In calculations, it makes no sense to introduce small gradation shares, and rounding should be used towards the nearest gradation, since this does not affect the accuracy of the final results. When the number of factors (summands) is more than two, the result is determined similarly. The process quickly converges as the number of components (factors or summands) increases; so for three to four components, the extreme limits of the range are reached. We can also determine the results for inverse operations (subtraction and division). When performing calculations, multiplication or division by an integer or rational number are defined using the summation operation. Subtraction and division operations are defined through the operations of addition and multiplication, respectively. The exponentiation and root extraction operations are defined through the multiplication operation. Calculations can be performed directly in fuzzy gradations or using modal values corresponding to fuzzy gradations, with the subsequent representation of the numerical results in the form of fuzzy gradations. Calculations on fuzzy gradations are greatly simplified if all gradations are expressed in terms of the smallest gradation (VL), namely, (VL-L) = 2VL, L = 3VL, etc., VH = 9VL, VVH = 10VL. This representation makes it possible to extend calculations formally outside the range 0 ... 1. When weighting factors are taken into account, normalization to 1 is not required. In calculations, each gradation is considered as a whole. If necessary, we can introduce a gradation structure using the membership function, but this is not required for the tasks under consideration.

We use this technique in subsequent calculations. The initial quantitative and qualitative information about objects and criteria, obtained using measurements and expert methods, is transformed into fuzzy gradations as follows. Each named variable is assigned a standardized (normalized) variable, varying in the interval [0, 1]. Then a fuzzy gradation is assigned to the standardized variable. In this case, the value 0 corresponds to the gradation VVL (the lowest value), and the value 1 corresponds to the gradation VVH (the highest value). A value of 0.1 corresponds to the modal value of the VL gradation (very low value); similarly a value of 0.3 – gradation of L (low value); a value of 0.5 – gradation M (middle value); the value of 0.7 – gradation H (high value), the value of 0.9 – gradation VH (very high value). The transition from physical to standardized variable is determined by the ratio $x = (z - z_{\min})/(z_{\max} - z_{\min}) \pm 0.1$, where the plus sign corresponds to the value of $z_{\min}$, and the minus sign to the value of $z_{\max}$. Here x is a standardized variable from the interval (0, 1); z is a "physical" variable, determined by measurement or expert method, which takes values in the interval $[z_{\min}, z_{\max}]$. Named numbers or dimensionless estimates represent values of z.

III. Classification problem

The problem is formulated as follows. We designate a set of objects as $X = \{x_1, ..., x_m\}$, a set of fuzzy criteria used to describe objects and their states as $\{K_1, ..., K_n\}$, a set of standards represented by fuzzy criteria as $Y = \{y_{01}, ..., y_{0l}\}$ and a set of classes as $Z = \{z_1, ..., z_p\}$ characterized by the range of values of criteria in a fuzzy form $\{K_{z1}\}, ..., \{K_{zp}\}$. It is required to determine the belonging of objects to classes. It should be borne in mind that several standards can correspond to one class, and the classes themselves are distributed over a certain range of criteria values. All criteria are given by fuzzy gradations in the range [VL, VH]. Without loss of generality, it can be assumed that the criteria are measured in a direct scale, and the solution closest to the standard is considered as the best. The case of specifying the critical area can be considered similarly. The degree of agreement α_j of the object x_u with the standard y_{0v} by criterion *j* is determined using the correspondence matrix (see table 1). In particular, $\alpha_j = \text{VH}$ if K_{uj} is equal to K_{0vj}; $\alpha_j = \text{H-VH}$ if K_{uj} and K_{0vj} differ by one gradation in one direction or another, and so on. When determining the critical area, it is more convenient to use the distance measure between the object and the standard $d_j = \bar{\alpha}_j$. We assume $d_j = \text{VL}$ if $\alpha_j = \text{VH}$, i.e. K_{uj} is equal to K_{0vj}, etc.

Table 1
Correspondence matrix

fuzzy gradations x, y	VL	VL-L	L	L-M	M	M-H	H	H-VH	VH
	$\alpha(x, y)$								
VL	VH	H-VH	H	M-H	M	L-M	L	VL-L	VL
VL-L		VH	H-VH	H	M-H	M	L-M	L	VL-L
L			VH	H-VH	H	M-H	M	L-M	L
L-M				VH	H-VH	H	M-H	M	L-M
M					VH	H-VH	H	M-H	M
M-H						VH	H-VH	H	M-H
H							VH	H-VH	H
H-VH								VH	H-VH
VH									VH

Note. Since $\alpha(x, y) = \alpha(y, x)$, only the values above the main diagonal are given in the table.

To determine the degree of agreement between objects and standards and between standards and classes, we use three types of functions: function by the greatest difference, the Hamming function and Euclidean function. When using the measure by the greatest difference, the degree of agreement of the object x_u with the standard y_{0v} by all criteria is given by the ratio

$$\alpha_{uv} \equiv \alpha(x_u, y_{0v}) = \min_j \alpha_j . \quad (1)$$

For Hamming function, we have

$$\alpha_{uv} = \frac{1}{n} \sum_j \alpha_j , \quad (2a)$$

where *n* is the number of criteria. For the Euclidean function, we have

$$\alpha_{uv} = \sqrt{\frac{1}{n} \sum_j \alpha_j^2} . \quad (2b)$$

For the degree of consistency between the standard y_{0v} and the class z_w, we have a similar relationships. We define the standard y_{0v*}, with which the object x_u is most consistent, by the expression

$$\alpha_{uv*} = \max_v \alpha_{uv} \tag{3}$$

We define class z_{w*}, with which the standard y_{0v*} is most consistent, by the expression

$$\alpha_{v*w*} = \max_w \alpha_{v*w} \tag{4}$$

Class z_{w*} is also the class with which object x_u is most consistent. The degree of agreement of the object x_u with the class z_{w*} is determined by the expression

$$\alpha_{uw*} = \min(\alpha_{uv*}, \alpha_{v*w*}) \tag{5}$$

Thus, solving the fuzzy classification problem establishes a hierarchical relationship between the external level (classes), the system level (standards), and the subsystem level (objects and facts). We now estimate the reliability of the solution to the fuzzy classification problem. We determine the reliability of correspondence of the object x_u to the standard y_{0v} by the ratio

$$\alpha_{uv} > \nu_{uv} \ , \tag{6}$$

where ν_{uv} is the fuzziness index, determined by the expression

$$\nu_{uv} \in (\min_j \min(\alpha_j, \bar{\alpha}_j), \max_j \min(\alpha_j, \bar{\alpha}_j)) \tag{7}$$

We define the reliability of correspondence of the standard y_{0v} to the class z_w by the similar relation

$$\alpha_{vw} > \nu_{vw} \ , \tag{8}$$

where ν_{vw} is determined by the expression

$$\nu_{vw} \in (\min_j \min(\alpha_{j0vw}, \bar{\alpha}_{j0vw}), \max_j \min(\alpha_{j0vw}, \bar{\alpha}_{j0vw})) \tag{9}$$

Thus, the indices of fuzziness are distributed over a certain area bounded by fuzzy gradations. We consider that the correspondence of object x_u to class z_w is reliable if both inequalities (6) and (8) are simultaneously satisfied.

III. Example of study for classification problem

Consider the diagnostic problem as an example of a fuzzy classification. The set of objects consists of six elements $X = \{x_1, ..., x_6\}$, each of which is described by criteria in the form of fuzzy gradations $\{K_1, ..., K_8\}$. The type of object is not specified; it can be a technical system, a person, a firm, a social system, or a mixed-type system. The set of representative standards consists of nine elements $Y = \{y_{01}, ..., y_{09}\}$, evaluated by the same criteria as objects. The set of classes consists of three classes $Z = \{z_1, z_2, z_3\}$, interpreted depending on the subject area. We assume that z_1 corresponds to the normal state of objects (suitable for use, healthy, etc.), z_2 corresponds to a risk group (requires prevention, observation, rest, etc.), z_3 corresponds to an abnormal group (emergency state, breakdown, illness, crisis, etc.). Each class is characterized by a distributed range of values of fuzzy criteria. The choice of criteria depends on the subject area. We use the following groups of criteria: functional, technical and economic, ergonomic, social. The criteria take into account both objective and subjective factors, for example, well-being, prejudice, prestige, personal or group benefit, etc. It is required to determine the belonging of objects to classes. The initial data are given in table 2. It should be borne in mind that the belonging of the standards to the classes is indicated presumably and should be checked in the process of solving the problem. The initial data are selected in such a way as to explore the possibilities of our approach in solving the problem of fuzzy classification. The data are contradictory, and objects, standards and classes overlap, so it is difficult to determine the solution, and in the classical formulation, the problem is unsolvable; its solution using membership functions is also very difficult.

Table 2
Initial data for the classification example

Objects	Criteria							
	K_1	K_2	K_3	K_4	K_5	K_6	K_7	K_8
x_1	VH	H	M	M	H	M-H	M	L
x_2	M	M	M	H	H	M	M	M
x_3	M	L	M	H	M-H	L	H	M
x_4	M	H	L-M	L	M	M	VH	VL
x_5	H	L-M	M	H	H	M	M	H
x_6	H	M	M	L	H	M	H	H
Standards	Criteria							
	K_1	K_2	K_3	K_4	K_5	K_6	K_7	K_8
$y_{01}(z_1)$	VH	VH	H	H	M	M	VH	H
$y_{02}(z_1)$	H	M	VH	M	VH	H	H	M
$y_{03}(z_1)$	M	H	M	VH	H	VH	H	M
$y_{04}(z_2)$	H	H	M	M	L	L	H	M
$y_{05}(z_2)$	M	L	H	L	H	M	M	L
$y_{06}(z_2)$	L	M	L	H	M	H	M	L
$y_{07}(z_3)$	M	M	L	L	VL	VL	M	L
$y_{08}(z_3)$	L	VL	M	VL	M	L	L	VL
$y_{09}(z_3)$	VL	L	VL	M	L	M	L	VL
Classes	Criteria							
	K_{z1}	K_{z2}	K_{z3}	K_{z4}	K_{z5}	K_{z6}	K_{z7}	K_{z8}
z_1	VH, H or M							
z_2	H, M or L							
z_3	M, L or VL							

First, we use the measure of the greatest difference. Below are the results of calculations in an abbreviated form; detailed calculations are shown only for typical cases to make the conclusions clear. When determining the belonging of objects to standards and standards to classes should take into account the most distant gradations, which allows us to determine the degree of agreement without tedious calculations. For the object x_1, we have from the relation (1): $\alpha(x_1, y_{01})$ = $\alpha(x_1, y_{02})$ = $\alpha(x_1, y_{03})$ = $\alpha(x_1, y_{04})$ = $\alpha(x_1, y_{05})$ = M; $\alpha(x_1, y_{06})$ = $\alpha(x_1, y_{07})$ = $\alpha(x_1, y_{08})$ = L; $\alpha(x_1, y_{09}) = \mathrm{VL}$. It is seen that the object x_1 has the maximum degree of agreement with the standards y_{01}, y_{02}, y_{03}, y_{04}, y_{05}. Now we define the agreement of these standards with classes. From relation (1) we similarly obtain $\alpha(y_{02}, z_1) = \mathrm{VH} \vee \mathrm{H} \vee \mathrm{M}$, $\alpha(y_{02}, z_2) = \mathrm{H} \vee \mathrm{M} \vee \mathrm{L}$, $\alpha(y_{02}, z_3) = \mathrm{M} \vee \mathrm{L} \vee \mathrm{VL}$. Here and below, the sign $\vee$ means "or". The same estimates are obtained for the standards y_{01} and y_{03}. From relation (5) it follows that according to the standards y_{01}, y_{02}, y_{03} the degree of agreement of the object x_1 with the class z_1 is middle. We have $\alpha(x_1, z_1) = \min(\mathrm{M}, \mathrm{VH} \vee \mathrm{H} \vee \mathrm{M}) = \mathrm{M}$. For the standard y_{05} we have $\alpha(y_{05}, z_1) = \mathrm{H} \vee \mathrm{M} \vee \mathrm{L}$, $\alpha(y_{05}, z_2) = \mathrm{VH} \vee \mathrm{H} \vee \mathrm{M}$, $\alpha(y_{05}, z_3) = \mathrm{H} \vee \mathrm{M} \vee \mathrm{L}$. The same estimates are obtained for the standards y_{04} and y_{06}. It follows that according to the standards y_{04} and y_{05} $\alpha(x_1, z_2) = \min(\mathrm{M}, \mathrm{VH} \vee \mathrm{H} \vee \mathrm{M}) = \mathrm{M}$ but according to the standard y_{06} we have $\alpha(x_1, z_2) = \min(\mathrm{L}, \mathrm{VH} \vee \mathrm{H} \vee \mathrm{M}) = \mathrm{L}$. Therefore, object x_1 agrees with both class z_1 and class z_2 to the same degree. However, the agreement according to the standard y_{02} is more reliable, since it is confirmed by two other standards y_{01} and y_{03}. The agreement according to standard y_{05} is confirmed by only one standard y_{04}. Therefore, we conclude that the object x_1 belongs to the class z_1. Calculations for other objects are performed in the same way. The results are given in tables 3, 4. We summarize the results of fuzzy classification using the measure of the greatest difference: the object x_1 belongs to the class z_1, objects x_2, x_3, x_4, x_5 and x_6 belong to the class z_2. The class z_3 is empty. We evaluate the reliability of classification. The calculation of the fuzzy index according to (7), (9) shows that it is distributed in the interval $\nu \in (\mathrm{VL}, \mathrm{M})$ both when comparing objects with standards, and when comparing standards with classes; therefore, the results of classification are reliable. Estimates show that the center of the distribution of the fuzzy index is near the gradation VL. In particular, when comparing the group of standards y_{01}, y_{02}, y_{03} with the class z_1, the center of distribution is to the left of the gradation L (less than L); when comparing this group with the class z_2 it is equal to L, and when comparing this group with the class z_3 it is to the right of L (more than L). Consequently, the most reliable is the agreement of this group of standards with the class z_1. When comparing the group of standards y_{04}, y_{05}, y_{06} with the class z_1, the center of distribution of the fuzzy index is equal to L, with the class z_2 – it is less than L, with the class z_3 – it is more than L. When comparing the group of standards y_{07}, y_{08}, y_{09} with the class z_1, the center of distribution of the fuzzy index is more than L, with the class z_2 – it is equal to L, and with the class z_3 – it is less than L. When comparing objects with standards, we obtain the following results. For objects x_1, x_2, the center of distribution of the fuzzy index when compared with y_{06} is equal to (VL-L), and when compared with other standards it is equal to L. For the object x_3, when compared with y_{04}, the center of distribution of the fuzzy index is less than L, and when compared with other standards it is equal to L. For the object x_4, the center of distribution of the fuzzy index is (VL-L) at comparison with y_{08}, it is equal to (L-M) when compared with y_{02}, and when compared with other standards it is equal to L. For objects x_5, x_6, when compared with all standards, the center of distribution of the fuzzy index is L. Comparison of the estimates obtained with the results given in tables 3 and 4 confirms the correctness of the assignment of standards to classes and objects to standards, and also makes it possible to increase the reliability of the classification results and identify errors in the initial data. If we use the softer measures of Hemming and Euclid to assess the degree of agreement, then the calculations according to the rules of fuzzy arithmetic (see above) show that in this case the boundaries of intersection of objects with standards and standards with classes become more vague. The results are given in tables 5, 6. From relation (5) we determine the belonging of objects to classes. The object x_1 should be attributed to the class z_2 (although according to Hamming's measure it can also be attributed to a class z_1); objects x_2, x_3, x_4, x_5, x_6 belong to the class z_2 (although according to Hamming's measure, the object x_5 can also be attributed to the class z_1). Classes z_1 and z_3 are empty. As we can see, the classification results for different choice of fuzzy measures are close. The measure of greatest difference allows us to determine quickly the "critical area", i.e. classes in which the object obviously does not belong. The Hemming and Euclid measures make it possible to specify the classification results. In the case of ambiguous classification results, when making a final decision on the belonging of objects to classes, external goals (priorities) must be taken into account. In particular, in our example, when an object is assigned to the class z_2 and, moreover, to the class z_3, direct costs (operating costs) increase, and when an object is assigned to the class z_1, indirect costs (consequences of a sudden failure) increase.

Table 3
The degree of agreement of objects with standards according to (1)

Standards	Objects					
	x_1	x_2	x_3	x_4	x_5	x_6
y_{01}	M	M	L	L	L-M	M
y_{02}	M	M	M	L-M	M	M
y_{03}	M	M	L	L	M	L
y_{04}	M	M	M	M	M	M
y_{05}	M	M	M	M	M	M
y_{06}	L	H	M	M	M	M
y_{07}	L	L	L-M	M	L	L
y_{08}	L	L	L	L	L	L
y_{09}	VL	M	M	L	L	L

Table 4
The degree of agreement of standards with classes according to (1)

Standards	Classes and range of values of criteria		
	z_1 $\mathrm{VH} \vee \mathrm{H} \vee \mathrm{M}$	z_2 $\mathrm{H} \vee \mathrm{M} \vee \mathrm{L}$	z_3 $\mathrm{M} \vee \mathrm{L} \vee \mathrm{VL}$
y_{01}, y_{02}, y_{03}	$\mathrm{VH} \vee \mathrm{H} \vee \mathrm{M}$	$\mathrm{H} \vee \mathrm{M} \vee \mathrm{L}$	$\mathrm{M} \vee \mathrm{L} \vee \mathrm{VL}$
y_{04}, y_{05}, y_{06}	$\mathrm{H} \vee \mathrm{M} \vee \mathrm{L}$	$\mathrm{VH} \vee \mathrm{H} \vee \mathrm{M}$	$\mathrm{H} \vee \mathrm{M} \vee \mathrm{L}$

y_{07}, y_{08}, y_{09}	M ∨ L ∨ VL	H ∨ M ∨ L	VH ∨ H ∨ M

Note. The order of the gradation in the table cells corresponds to the order of the gradation in the classes. ∨ – "or" sign, which means that the values are distributed over the given area.

Table 5
The degree of agreement of objects with the standards according to (2a), (2b)

Standards	Objects					
	x_1	x_2	x_3	x_4	x_5	x_6
y_{01}	H	H	H	H	H	H
y_{02}	H	H	H	M-H	H	H (H-VH)
y_{03}	H	H-VH	H (H-VH)	H	H	H
y_{04}	H (H ∨ H-VH)	H	H-VH	H	H	H ∨ H-VH (H-VH)
y_{05}	H (H ∨ H-VH)	H-VH	H (H ∨ H-VH)	H	H (H ∨ H-VH)	H ∨ H-VH (H-VH)
y_{06}	H	H-VH	H	H	H	H (M-H ∨ H)
y_{07}	M-H (M-H ∨ H)	H	H	H	M-H	M-H ∨ H (H)
y_{08}	M-H	M-H	H	H	M-H	M-H
y_{09}	M-H	M-H	M-H	M-H	M-H	M ∨ M-H

Note. The values for the Euclidean function are shown in parentheses. If the values for the Hamming and Euclidean functions are the same, then only one value is specified.

Table 6
The degree of agreement of standards with classes according to (2a), (2b)

Standards	Classes and range of values of criteria		
	z_1 VH ∨ H ∨ M	z_2 H ∨ M ∨ L	z_3 M ∨ L ∨ VL
y_{01}, y_{02}, y_{03}	H ∨ H-VH ∨ H	H-VH ∨ H ∨ M	H ∨ M ∨ L
y_{04}, y_{05}, y_{06}	M ∨ H ∨ H-VH	H ∨ H-VH ∨ H	H-VH ∨ H ∨ M
y_{07}, y_{08}, y_{09}	L ∨ M ∨ H	M ∨ H ∨ H-VH	H ∨ H-VH ∨ H

Note. The results for the Euclidean function coincide within a quarter of the basic gradation with the results for the Hamming function, so the difference is not significant.

IV. Cluster analysis

Cluster analysis is a kind of classification problem when there is no information about a set of standards. It consists in combining objects into groups (clusters) depending on the degree of similarity, determined by a number of criteria (features, properties). The task is formulated as follows. A set of objects $X = \{x_1,\ldots, x_m\}$ are given, evaluated by a set of criteria $K_1,\ldots, K_n$, represented by fuzzy gradations in the range [VL, VH]. It is required to distribute objects into clusters and order levels. The algorithm for solving the problem includes the following steps:

1. The initial quantitative and qualitative information about objects and criteria, obtained by measurements and expert methods, is transformed into fuzzy gradations in the same way as in the classification problem (see above).

2. The measure of proximity α_{ij}^{l} between objects x_i and x_j by the criterion K_l is determined in the same way as in the classification problem, using the correspondence matrix (table 1). The distance d_{ij}^{l} that determines the degree of difference between the objects x_i, x_j by the criterion K_l is the reciprocal (opposite) of the proximity measure α_{ij}^{l}, namely, $d_{ij}^{l} = \bar{\alpha}_{ij}^{l}$.

3. The degree of proximity of objects is determined by all criteria. To do this, it is necessary to aggregate the results for each criterion using a suitable decision-making model, which should be chosen to take into account the small values of the degrees of proximity. The use of averages (Hamming and Euclidean functions), which, generally speaking, do not satisfy this requirement, leads to the loss of useful information and the ambiguity of combining objects into clusters. To preserve all the useful information and to minimize the ambiguity of combining objects into clusters, we use the measure of proximity by the greatest difference

$$\alpha_{ij} = \min_{l} \alpha_{ij}^{l}. \tag{10}$$

From (10) it follows that the degree of proximity α_{ij} of objects x_i, x_j for all criteria is determined by the smallest value of the measure of proximity of these objects according to individual criteria. Such an assessment is the most reliable and allows us to reduce the influence of the inadequacy of the decision-making model on the analysis results. Sometimes it is convenient to use distances, then (10) is written in the form

$$d_{ij} = \max_{l} d_{ij}^{l}. \tag{10a}$$

4. The result of the analysis is the unification of objects into clusters, distributed over the levels of order, in accordance with the values of the fuzzy measure of proximity.

5. If for a group of homogeneous objects the results of the analysis turn out to be the same, then to separate the objects, we can use the Hamming function, as it is most convenient for calculations. Its use is correct for objects of the same level. In this case, we have

$$\alpha_{ij} = 1/n \sum_{l} \alpha_{ij}^{l}. \tag{11}$$

V. Example of study for cluster analysis

Consider the problem of determining the quality of objects as an example of cluster analysis. In this case, depending on the subject area, the objects can be solution methods, strategies, software or hardware, knowledge bases, treatment methods, drugs, etc. To evaluate objects, we use the same groups of criteria as in the classification problem. The initial data for the example are given in table 7.

Table 7
Initial data for the example of cluster analysis

Objects	Criteria				
	K_1	K_2	K_3	K_4	K_5
x_1	M	L-M	M	L	L-M
x_2	M	L	H-VH	VL	L-M
x_3	M-H	L-M	VL	L-M	L-M
x_4	L	M-H	VH	L-M	VL
x_5	H	VL	H	L	L-M
x_6	VL-L	M-H	H	L-M	L
x_7	VL	H	VL	VL-L	H
x_8	VL	VH	VL	L-M	VL

x_9	VL	H-VH	VL	VL-L	L
x_{10}	VH	VL	H	H	VL-L
x_{11}	H-VH	VL	M-H	VH	L-M
x_{12}	VH	VL	H	L-M	L-M
x_{13}	M	VL	H	L-M	VH
x_{14}	VL-L	M-H	H	M-H	L

As can be seen from table 7, the initial data are heterogeneous, and their estimates by the criteria are not consistent, which corresponds to the real situation. First, we determine the value of the measure of proximity of objects for each criterion, using table 1 and table 7. Then we determine the degree of proximity between objects for our example by all criteria from relation (10). The results are given in table 8.

Table 8
The degree of proximity of objects by all criteria, determined from (10)

Objects	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}
x_1	VH	M-H	M	M	M-H	M-H	M	L-M	M	M	L	M	L-M	M-H
x_2		VH	VL-L	M-H	H	M-H	VL-L	VL-L	VL-L	L	VL	M	L-M	L-M
x_3			VH	VL	L	L	L-M	L-M	L-M	L	L-M	L	L	L
x_4				VH	L-M	H	VL	VL	VL	L	L-M	L	VL	H
x_5					VH	L-M	L	VL	VL-L	C	L	H	L-M	L-M
x_6						VH	L	L	L	VL-L	L	VL-L	L	H
x_7							VH	L	C	VL	VL-L	VL	L	L
x_8								VH	H	VL	VL	VL	VL	L
x_9									VH	VL	VL-L	VL	VL-L	L
x_{10}										VH	H	M-H	VL-L	VL-L
x_{11}											VH	L-M	L-M	L
x_{12}												VH	L-M	VL-L
x_{13}													VH	L
x_{14}														VH

From the data in table 8, we determine the sequence of combining objects into clusters and their distribution by order levels. The results are given in table 9. The results of the analysis at the levels VH and H-VH can be specified using the Hamming function (11). This procedure is correct, since it is applied to objects of the same level $\{x_5, x_{12}\}$, $\{x_4, x_6, x_{14}\}$, $\{x_8, x_9\}$, $\{x_{10}, x_{11}\}$ with the same smallest degree of proximity H in each cluster. The degree of proximity of these objects by criteria, determined using the data from tables 1 and 7, is given in table 10. From the data of table 10, we determine the degree of proximity of objects by all criteria according to relation (11), using the rules of fuzzy arithmetic. We show how the calculations are performed. For i = 6, j = 14 we have α_{ij} = 1/5(VH + VH + VH + H + VH) = VH.

For $i = 5$, $j = 12$ we have α_{ij} = 1/5(H + VH + VH + H-VH + VH) = 1/5(7VL + 9VL + 9VL + 8VL + 9VL) = 42/5VL = 8,4VL = H-VH (with rounding). The rest of the calculations are performed similarly. For example, for i = 4, j = 14 we obtain α_{ij} = 1/5((H-VH) + VH + H + H + H) = 1/5(8VL + 9VL + 7VL + 7VL + 7VL) = 38/5VL = 7,6VL= H-VH (with rounding) and so on.

The results for the levels of VH and H-VH are given in table 11. Thus, using the Hamming measure, it is possible to determine clusters at higher order levels (VH and H-VH), although these results are less reliable than for the measure of the greatest difference.

Table 9
Results of cluster analysis

Degree of objects proximity	Number of clusters	Composition (structure) of clusters
VH	0	All objects are separated from each other
H-VH	0	All objects are separated from each other
H	4	$\{x_5, x_{12}\}$, $\{x_4, x_6, x_{14}\}$, $\{x_8, x_9\}$, $\{x_{10}, x_{11}\}$
M-H	5	$\{x_1, x_2\}$, $\{x_5, x_{12}\}$, $\{x_4, x_6, x_{14}\}$, $\{x_8, x_9\}$, $\{x_{10}, x_{11}\}$
M	4	$\{x_1, x_2, x_5, x_{12}\}$, $\{x_4, x_6, x_{14}\}$, $\{x_8, x_9\}$, $\{x_{10}, x_{11}\}$
L-M	4	$\{x_1, x_2, x_5, x_{12}, x_{13}\}$, $\{x_4, x_6, x_{14}\}$, $\{x_3, x_8, x_9\}$, $\{x_{10}, x_{11}\}$
L	4	$\{x_1, x_2, x_5, x_{12}, x_{13}\}$, $\{x_4, x_6, x_{14}\}$, $\{x_3, x_8, x_9, x_7\}$, $\{x_{10}, x_{11}\}$
VL-L	3	$\{x_1, x_2, x_5, x_{12}, x_{13}\}$, $\{x_4, x_6, x_{14}, x_{10}, x_{11}\}$, $\{x_3, x_8, x_9, x_7\}$
VL	1	All objects are combined into one cluster

Note. A cluster is a subset consisting of at least two objects. Clusters are in braces.

Table 10
Degree of proximity α_{ij}^{l} of objects $\{x_5, x_{12}\}$, $\{x_4, x_6, x_{14}\}$, $\{x_8, x_9\}$, $\{x_{10}, x_{11}\}$

i	j	$l = 1$	$l = 2$	$l = 3$	$l = 4$	$l = 5$
6	14	VH	VH	VH	VH	VH
5	12	H	VH	VH	H-VH	VH
8	9	VH	H-VH	VH	H	H
10	11	H-VH	VH	H-VH	H	H
4	6	H-VH	VH	H	VH	H
4	14	H-VH	VH	H	H	H

Table 11
Results of cluster analysis at the levels VH and H-VH according to (11)

Degree of objects proximity	Number of clusters	Composition (structure) of clusters
VH	1	$\{x_6, x_{14}\}$
H-VH	4	$\{x_4, x_6, x_{14}\}$, $\{x_5, x_{12}\}$, $\{x_8, x_9\}$, $\{x_{10}, x_{11}\}$

VI. Discussion of the results

The results obtained show a certain analogy between cluster analysis and connectivity analysis (topological analysis), the application of which to complexes of object differences can be useful, although it requires additional efforts. Calculations carried out using a numerical measure of proximity determined by the Euclidean function and the nearest-neighbor approximation show that the results are similar but not identical. In particular, there is an increase in the number of order levels up to 13 (instead of 9 in our analysis), and the difference between these additional levels is less than 0.01 (the relative difference ranges from 0.7% to 1.5%), i.e. does not exceed the error of the initial data. The appearance of such additional levels is difficult to justify and interpret.

VII. Conclusion

Thus, the study confirmed by calculations, shows that the proposed approach to the classification problem allows us to expand the area of solvable problems, increase the reliability of the solution and reduce the complexity of calculations.

The noted advantages are especially perceptible when applying the approach to classification problems in large systems. The application of the proposed approach in cluster analysis allows to reduce errors and ambiguity in combining objects into clusters, makes it more reasonable to determine the levels of order and number of clusters at each level in comparison with the analysis using an exact numerical measure of similarity or membership function. In addition, the complexity of the analysis is significantly reduced.

References

[1]. Applied fuzzy systems. Ed. by T. Terano. Moscow, Publ. house "Mir", 1993 (in Russian translation).
[2]. Applied statistics. Ed. by S.A. Ayvazyan. Moscow, Publishing house "Finance and Statistics", 1989 (in Russian).
[3]. Fuzzy sets and theory of possibilities. Ed. by R.R. Jager. Moscow, Publ. house "Radio and Communication", 1986 (in Russian translation).
[4]. Chauhan R., Kaur H., Alam M. Data Clustering Method for Cluster Detection in Spatial Cancer Databases," International Journal of Computer Applications, Issue 10, No. 6, 2010, pp. 9 – 14.
[5]. Yelipe U., Porica S., Golla M. An Efficient Approach for the Imputation and Classification of Medical Data Values Using Class-Based Clustering of Medical Records. Computers and Electrical Engineering, V. 66, 2018, pp. 487 – 504.
[6]. Hammuda C., Carrey F. Comparative Study of Data Clustering Techniques. University of Waterloo, 2000.
[7]. Bezdek James K. Pattern Recognition with Fuzzy Objective Function Algorithms. Springer Science & Business Media, 2013.
[8]. Jager R., Dimitar P. Approximate Clustering Using the Rock Method. IEEE Transactions on Systems, Man and Cybernetics, V. 24, No. 8, 1994, pp. 1279 – 1284.
[9]. Zhang K., Yang L.T., Chen Z., Li P. A Privacy-Preserving Possible High Order C-Means Algorithm for Clustering Big Data with Cloud Computing. IEEE Transactions on Big Data, 2017.
[10]. Yan X., Zhu Y., Zou W., Wang L. A New Approach to Data Clustering Using the Hybrid Artificial Bee Colony Algorithm. Neurocomputing, V. 97, 2012, pp. 241 – 250.
[11]. Pakrashi A., Chaudhury B. B. Kalman Filtering Induced by Heuristic Optimization Based on Partitioned Data Clustering. Information Science, 2016.
[12]. Hatamlow A. Black hole. A New Heuristic Optimization Approach for Data Clustering. Information Science, V. 222, 2013, pp. 175 – 184.
[13]. Tsujinishi D., Abe S. Fuzzy least squares support vector machines for multiclass problems. Neural Networks, V.16, 2003, pp.785 – 792.
[14]. Binu, D., Kariyappa B. A new rider optimization algorithm-based neural network for fault diagnosis in analog circuits. IEEE Transactions on Instrumentation and Measurement, V. 68, 2018, No. 1, pp.2 – 26.
[15]. Romanov V.N. Fuzzy arithmetic and its application in problems of optimal choice. Int. Journal of Engineering Research and Applications, V.7, Issue 11, 2017, pp. 7 – 15.
[16]. Romanov V.N. The use of fizzy arithmetic in decision-making problems. SN Applied sciences, V.4, 1:367, 2019.