The problem of distribution of a set of objects, the state of which is determined by a set of controlled parameters, into a set of subsets of objects maximally homogeneous in their properties is considered. Relevance of the problem and important advantage of the clustering procedure: when its implementing it is possible to reduce the initial difficult problem of high dimensionality objects analysing to the soluti

DOI: 
http://creativecommons.org/licenses/by/4.0

Abstract. The problem of distribution of a set of objects, the state of which is determined by a set of controlled parameters, into a set of subsets of objects maximally homogeneous in their properties is considered. Relevance of the problem and important advantage of the clustering procedure: when its implementing it is possible to reduce the initial difficult problem of high dimensionality objects analysing to the solution of a number of simpler problems of lower dimensionality. This circumstance acquires additional attractiveness and importance if the initial data of the problem contain uncertainty, for example, are vaguely defined. Research object is the procedure of partitioning a set of objects into clusters under conditions of uncertainty. In this regard, the purpose of the study is to develop a method for solving the problem of clustering in conditions where the initial data on objects controlled parameters the values contain uncertainty. The method of solving the problem is based on clustering procedure mathematical model development, containing analytical expressions for the criterion of its effectiveness, written in the form of a twice fractionally quadratic function. The impossibility of mathematical programming problem direct solution initiated the development of a heuristic algorithm for its solution. As a result, an iterative method was obtained and applied to solve the clustering problem under conditions of fuzzy initial data. The developed computational procedure is based on a reasonable system of rules for performing operations on fuzzy numbers. The situations when the belonging functions of problem fuzzy parameters are defined on infinite or compact media are considered. The developed system of rules allows to correctly perform operations in the metric of fuzzy defined states between clustering objects. The proposed method is easily...
Анотація. Розглядається завдання розподілу множини об’єктів, стан яких визначається набором контролюваних параметрів, на сукупність підмножин об’єктів максимально однорідних за своїми властивостями. Актуальність проблеми обумовлена важливою перевагою процедури кластеризації: за її реалізації з’являється можливість виходу важку задачу аналізу об’єктів високої розмірності звести до роз’їжджа низки простіших задач меншої розмірності. В першому випадку набуває додаткової привабливості та важливості, якщо вхідні дані задачі містять невизначеність, наприклад, визначені нечітко. Об’єкт дослідження – процедура розбиття множини об’єктів на кластери у вумовах невизначеності. У зв’язку з цим мета дослідження – розроблення методу розв’язання задачі кластеризації в умовах, коли вхідні дані про значення контролюваних параметрів об’єктів містять невизначеність. Метод розв’язання задачі буде здійснюватися у рамках математичної моделі процедури кластеризації, що містить аналітичні вирази для критерію ефективності, записаного у формі двічі дробово-квадратичної функції. Неможливість безпосереднього розв’язання даних задач знайдеться на нескінченних або компактних носіях. Розроблена система правил дає змогу коректно виконувати операції в метриці нечітко визначених станів між об’єктами кластеризації. Запропонований метод легко узагальнюється на випадок, коли невизначеність вихідних даних є ієрархічною.

Key words: clustering problem, uncertainty of initial data, rules of operations execution, step-by-step computational procedure of solution.

Ключові слова: задача кластера, нечіткість вихідних даних, правила виконання операцій, покрокова обчислювальна процедура розв’язання.

Introduction. Clustering is a multivariate statistical procedure that divides a set of objects into a set of subsets containing maximally homogeneous groups of similar controlled features (parameters) of objects.

Clustering technology is widely used in economics, medicine, psychology, chemistry, sociology, etc. Its practical application allows:
- better understand and explain the mechanism of typical features and parameters of objects on the nature of their functioning;
- improve the quality of objects analysis taking into account their differences in groups;
- to perform data compression by selecting the most typical representatives of objects and to divide them into subsets according to the type (nature) of their state, to select atypical objects that cannot be attached to any of the groups.

The clustering procedure can be hierarchical, when subsets of objects included in each cluster are again subjected to clustering.

Analysis of clustering problem formulations, publication and solution. Formal formulation of the clustering problem has the following form [1.2]. Set of objects is given whose position in the phase space of the controlled parameters and the state of these objects is determined by a set of coordinates. The task is to distribute a set of objects into a number of compact subsets in accordance with some selected criterion.

We obtain an analytical relation for the quality criterion of the clustering procedure. Let’s introduce:
\[ N = \{1, 2, ..., j, ..., n\} – set of objects to be grouped; \]
\[ S = \{1, 2, ..., s, ..., S\} – multiple clusters; \]
\[ L = \{1, 2, ..., l, ..., L\} – set of controlled parameters of clustering objects; \]
\[ x_{ij} – value of the l – th coordinate of the j – th object corresponding to the e – th value of the controlled parameter of this object, l \in L, j \in N; \]
\[ r_{i, j} = \left( \sum_{k=1}^{t} (x_{ik} - x_{jk})^2 \right)^{1/2} – \text{is the Euclidean distance between points in the parameter space corresponding to objects} \]
\[ j_i \text{ and } j_2, j_1 \in N, j_2 \in N; \]
\[ V_{si} – value of the l – th coordinate of the center of the grouping of the s - th cluster, s \in S; \]
\[ r_w = \left( \sum_{k=1}^{t} (V_{si} - x_{ik})^2 \right)^{1/2} – \text{distance between the } j – th object and the s - th grouping center; \]
\[ u_{sj} = \begin{cases} 1, & \text{if } j - \text{th object belongs to the } s - \text{ th cluster,} \\ 0, & \text{otherwise.} \end{cases} \]

Let some distribution of objects by clusters be obtained. Then
$R_s = \sum_{h=1}^{s} \sum_{j_2=1}^{s} r_{h,j_1} u_{h,j_1} u_{h,j_2}$ – the sum of the distances between the objects in $s$-th cluster.

$U_s = \sum_{h=1}^{s} \sum_{j_2=1}^{s} u_{h,j_1} u_{h,j_2}$ – the number of pairs of objects in $s$-th cluster.

$R_c = \sum_{h=1}^{s} \sum_{j_2=1}^{s} \sum_{j_3=1}^{s} r_{h,j_1} u_{h,j_1} u_{h,j_2} u_{h,j_3} u_{h,j_3}$ – sum of the distances between objects in different clusters.

$U_c = \sum_{h=1}^{s} \sum_{j_2=1}^{s} \sum_{j_3=1}^{s} u_{h,j_1} u_{h,j_2} u_{h,j_3} u_{h,j_3}$ – number of pairs of objects.

It is clear that the result of clustering is naturally considered to be of higher quality, the further apart the objects that fall into different clusters are, and the closer they are to each other if they belong to the same cluster. In accordance with these, we introduce

$R_s \over U_s$ – the average distance between pairs of objects caught in $s$-th cluster.

$R_c \over U_c$ – the average distance between pairs of objects belonging to different clusters.

Now the clustering quality criterion can be calculated by the formula

$$\tau = \frac{R_c \over U_c}{\max_s R_s \over U_s}$$

(1)

The meaning of criterion (1) is clear: its value is higher the greater the average distance between objects from different clusters and, at the same time, the smaller the average distance between objects for the most "loose" (least compact) cluster.

Thus, the clustering problem can be formulated as follows: find a set $\{U_s\}$ that maximizes

$$\tau = \frac{R_c \over U_c}{\max_s R_s \over U_s}$$

(2)

and satisfying natural constraints

$$\sum_{j=1}^{n} u_{j,s} = 1, \quad j = 1, 2, \ldots, n,$$

$$\sum_{h=1}^{s} u_{j,s} \geq 2, \quad s = 1, 2, \ldots, s.$$

(3)

The resulting task (2), (3) is a fractional-quadratic programming problem. Let's consider possible methods for solving such tasks. The fractional-polynomial structure of the criterion function (3) leads to a significant complication of the analytical description of the derivatives of this function. Therefore, the use of optimization methods of the first and second order is not feasible in this case. The possibility of using universal zero-order methods here is limited by the high dimension of practical clustering problems.

We set the task of developing a special method for solving the fractional-quadratic optimization problem. Let's consider the simplest example of such a task. We introduce the criterion function

$$F(x) = \frac{\sum_{j=1}^{n} a_j x_j^2}{\sum_{j=1}^{n} x_j^2}.$$

(4)

and restriction

$$\sum_{j=1}^{n} x_j = C.$$

(5)
We set the task of finding a set \( x_j, j = 1, 2, ..., n \), maximizing (4) and satisfying (5). Let's introduce a variable

\[
y_0^2 = \frac{1}{\sum_{j=1}^{n} x_j^2}
\]

From here

\[
y_0^2 = \sum_{j=1}^{n} x_j^2 = \sum_{j=1}^{n} (y_0 x_j)^2 = 1. \tag{6}
\]

Now let's introduce new variables

\[
y_j = y_0 x_j, \quad j = 1, 2, ..., n. \tag{7}
\]

Then

\[
x_j = \frac{y_j}{y_0}, \quad j = 1, 2, ..., n. \tag{8}
\]

Transform (4) and (5) considering (8). At the same time

\[
F(x) = \sum_{j=1}^{n} a_j x_j^2 = y_0^2 \sum_{j=1}^{n} a_j \left( \frac{y_j}{y_0} \right)^2 = \sum_{j=1}^{n} a_j y_j^2, \tag{9}
\]

from where

\[
\sum_{j=1}^{n} y_j = y_0 C. \tag{10}
\]

Now the original problem (5), (6) is transformed to the form: find a set \( \{ y \}, j = 1, 2, ..., n \), maximizing (10) and satisfying (10). Let's solve this problem by the method of indefinite Lagrange multipliers.

Let’s introduce

\[
\varphi(y) = \sum_{j=1}^{n} a_j y_j - \lambda \left( \sum_{j=1}^{n} y_j - y_0 C \right).
\]

Further

\[
\frac{\partial \varphi(y)}{\partial y_j} = 2a_j y_j - \lambda = 0, \quad y_j = \frac{\lambda}{2a_j}, \quad j = 1, 2, ..., n \tag{11}
\]

We will \( \lambda \) find an indefinite multiplier using (10). At the same time

\[
\sum_{j=1}^{n} \frac{\lambda}{2a_j} = \frac{\lambda}{2} \sum_{j=1}^{n} \frac{1}{a_j} = y_0 C.
\]

From here

\[
\frac{\lambda}{2} = \frac{y_0 C}{\sum_{j=1}^{n} \frac{1}{a_j}}. \tag{12}
\]

At the same time, substituting (11) into (12), we get

\[
y_j = \frac{y_0 C}{\sum_{j=1}^{n} \frac{1}{a_j}} \cdot \frac{1}{a_j}, \quad j = 1, 2, ..., n. \tag{13}
\]

Then, substituting (13) into (8), we have

\[
x_j = \frac{y_j}{y_0} = \frac{y_j C}{y_0 \sum_{j=1}^{n} \frac{1}{a_j}} \cdot \frac{1}{a_j} = \frac{C}{y_j C} \cdot \frac{1}{a_j} \sum_{j=1}^{n} \frac{1}{a_i}, \quad j = 1, 2, ..., n.
\]

The solution has been received. Note, however, that the technique used in this problem, which transformed the fractional–quadratic problem (5), (6) into an ordinary mathematical programming problem with a quadratic objective function, cannot be used to solve the problem (2)-(4). The fact is that the structure of the objective function (2) is doubly fractional–quadratic, that is, the components of the fractional–quadratic function themselves are fractional–quadratic. In this regard, to solve the problem (2)-(3), we use one of the well–known approximate clustering methods [3,4], for
example, the following.

Let be given a set of $n$ objects to be clustered, whose position in the $L$-dimensional space of controlled parameters is determined by sets of coordinates $\{(x_{11}, x_{12}, \ldots, x_{1L}), (x_{21}, x_{22}, \ldots, x_{2L}), \ldots, (x_{n1}, x_{n2}, \ldots, x_{nL})\}$, as well as coordinates of $m$-cluster grouping centers $\{(x_{11}, x_{12}, \ldots, x_{1L}), (x_{21}, x_{22}, \ldots, x_{2L}), \ldots, (x_{m1}, x_{m2}, \ldots, x_{mL})\}$.

For each of the clustering objects $j$, $j = 1, 2, \ldots, n$, the distance to each of the grouping centers is calculated

$$R_{js} = \left[ \sum_{i=1}^{L} (x_{ij} - x_{ij'})^2 \right]^{1/2}, \quad j = 1, 2, \ldots, n, \quad s = 1, 2, \ldots, m.$$ 

Now, by going through $s$, the nearest of the grouping centers is determined for each object, and this object is attached to the corresponding cluster. If the cluster grouping centers are not known, various algorithms for their sequential determination are used. At the same time, the order of objects joining becomes particularly important. Well-known methods successfully solve such a clustering problem with a good level of approximation. However, the task becomes significantly more complicated if objects coordinates are not precisely defined, but are given, for example, in terms of fuzzy mathematics [5-6].

**Clustering Method in the Conditions of Fuzzy Source Data.** Let the objects coordinates be fuzzy numbers with a probability function $(L-R)$ – type [7,8]. In this case, the belonging function is uniquely determined by the set $<m, \alpha, \beta>$, where $m$ is the modal value of the fuzzy number, $\alpha$ and $\beta$, respectively, are the left and right fuzzy coefficients. The analytical relations describing the corresponding belonging functions depend on the selected type of functions. If, in particular, these are Gaussian functions given on an infinite interval, then the analytical description of the belonging function has the form:

$$\mu(x) = \begin{cases} \exp \left\{ -\frac{(m-x)^2}{2\alpha^2} \right\}, & x \leq m, \\ \exp \left\{ -\frac{(x-m)^2}{2\beta^2} \right\}, & x > m. \end{cases} \quad (14)$$

Let’s introduce

$$\mu_{ij}(x_{ij}) = \begin{cases} \exp \left\{ -\frac{(m_{ij} - x_{ij})^2}{2\alpha_{ij}^2} \right\}, & x_{ij} \leq m_{ij}, \\ \exp \left\{ -\frac{(x_{ij} - m_{ij})^2}{2\beta_{ij}^2} \right\}, & x_{ij} > m_{ij}. \end{cases}$$

the function of belonging of the $e$-th controlled parameter of the $j$-object, $e=1, 2, \ldots, L, \quad j = 1, 2, \ldots, n$. Now the coordinates of the cluster grouping centers $(x_{ij})$ are sequentially substituted into the relations $\mu_{ij}(x_{ij})$ for each of the clustered objects and the one for which the value $\min \mu_{ij}(x_{ij})$, $e=1, 2, \ldots, L$, is the largest is selected.

Certainly, this method is unacceptable if the belonging functions of object controlled parameters are set on compact media.

Such typical function has a triangular form, and its description is defined by the relation

$$\mu(x) = \begin{cases} 0, & x < m - \alpha, \\ \frac{x - (m - \alpha)}{\alpha}, & m - \alpha \leq x < m, \\ \frac{(m + \beta) - x}{\beta}, & m < x < m + \beta, \\ 0, & x > m + \beta. \end{cases} \quad (15)$$

In this case, the fuzzy description of the clustering object parameters belonging functions in the form (15) necessitates the appropriate mathematical support development. During the implementation of the clustering procedure, arithmetic operations of addition, subtraction, multiplication, as well as square root extraction are performed. Here are the rules for performing these operations, taking into account the description of the belonging functions of fuzzy parameters in the form of $(L-R)$-type functions.

Let the belonging functions of two fuzzy $(L-R)$-type numbers be given by the values:

$$x_1 = <m_1, \alpha_1, \beta_1>, \quad x_2 = <m_2, \alpha_2, \beta_2>.$$ \quad (16)

Let’s introduce a binary operation

$$z = x_1 \otimes x_2 = <m, \alpha, \beta>.$$
In accordance with the arithmetic rules of fuzzy numbers of \((L - R)\)-type [9], the formulas for calculating the components of the belonging function of a fuzzy calculation result have the following form:

Addition
\[
m = m_1 + m_2, \quad \alpha = \alpha_1 + \alpha_2, \quad \beta = \beta_1 + \beta_2; \]
(17)

Subtraction
\[
m = m_1 - m_2, \quad \alpha = \alpha_1 + \beta_2, \quad \beta = \beta_1 + \alpha_2; \]
(18)

Multiplication
\[
m = m_1m_2, \quad \alpha = m_1\alpha_1 + m_2\alpha_2 - \alpha_1\alpha_2, \quad \beta = m_1\beta_1 + m_2\beta_2 + \beta_1\alpha_2. \quad \quad \beta = \beta_1 + \alpha_2; \]
(19)

The result of performing the operation of extracting the square root from a fuzzy number \(< m, \alpha, \beta >\) has the form:
\[
m = \sqrt{m}, \quad \alpha = \sqrt{m} - \alpha, \quad \beta = \sqrt{m} + \beta. \quad \quad \beta = \beta_1 + \alpha_2; \]
(20)

When solving many practical problems, the uncertainty about the values of the controlled parameters of these tasks has a hierarchical character. In this case, the components \(< m, \alpha, \beta >\) of the belonging function of fuzzy parameters are themselves fuzzy quantities, that is
\[
m = < m_a, \alpha_a, \beta_a >, \quad \alpha = < m_a, \alpha_a, \beta_a >, \quad \beta = < m_a, \alpha_a, \beta_a >. \]

Let’s write down the rules for performing arithmetic (20) operations on fuzzy numbers defined by (20):

Addition
\[
m = < m_a, \alpha_a, \beta_a >, \quad \alpha = < m_a, \alpha_a, \beta_a >, \quad \beta = < m_a, \alpha_a, \beta_a >. \quad \quad \beta = \beta_1 + \alpha_2; \]
(21)

Subtraction
\[
m_1 - m_2 = < m_1, \alpha_1, \beta_1 > - < m_2, \alpha_2, \beta_2 >; \quad \alpha = < m_1, \alpha_1, \alpha_2 >, \quad \beta = < m_1, \beta_1, \beta_2 >. \]
(24)

Let’s get the rules for performing the multiplication operation. In accordance with (19), the parameters of the fuzzy number \(x_1x_2\) belonging function are calculated by calculating the fuzzy number belonging functions \(m_1m_2, m_1\alpha_2, m_2\alpha_1, \alpha_1\alpha_2, m_1\beta_2, m_2\beta_1, \beta_1\beta_2\).

The corresponding operations are performed according to the formulas:
\[
m_1m_2 = < m_1m_2, \alpha_1\alpha_2 + m_1\alpha_m + m_2\alpha_m + \alpha_1\alpha_2, \quad m_1\beta_2 + m_2\beta_1; \quad \alpha = < m_1m_2, \alpha_1\alpha_2 + m_1\alpha_m + m_2\alpha_m + \alpha_1\alpha_2, \quad \beta = < m_1m_2, \alpha_1\alpha_2 + m_1\alpha_m + m_2\alpha_m + \alpha_1\alpha_2; \]
(27)

\[
m_1\alpha_1 = < m_1, \alpha_1m_2 + m_1\alpha_m + m_2\alpha_m - \alpha_1\alpha_2, \quad m_1\alpha_m + m_2\beta_1 + m_2\beta_1; \quad \alpha = < m_1, \alpha_1m_2 + m_1\alpha_m + m_2\alpha_m - \alpha_1\alpha_2, \quad \beta = < m_1, \alpha_1m_2 + m_1\alpha_m + m_2\alpha_m - \alpha_1\alpha_2. \]
(28)

\[
m_1\alpha_2 = < m_1m_2, \alpha_1m_2 + m_1\alpha_m + m_2\alpha_m - \alpha_1\alpha_2, \quad m_1m_2 + m_1\beta_2 + m_2\beta_1; \quad \alpha = < m_1m_2, \alpha_1m_2 + m_1\alpha_m + m_2\alpha_m - \alpha_1\alpha_2, \quad \beta = < m_1m_2, \alpha_1m_2 + m_1\alpha_m + m_2\alpha_m - \alpha_1\alpha_2. \]
(29)

\[
m_2\alpha_1 = < m_1m_2, \alpha_1m_2 + m_1\alpha_m + m_2\alpha_m - \alpha_1\alpha_2, \quad m_1m_2 + m_1\beta_2 + m_2\beta_1; \quad \alpha = < m_1m_2, \alpha_1m_2 + m_1\alpha_m + m_2\alpha_m - \alpha_1\alpha_2, \quad \beta = < m_1m_2, \alpha_1m_2 + m_1\alpha_m + m_2\alpha_m - \alpha_1\alpha_2. \]
(30)

\[
m_2\alpha_2 = < m_1m_2, \alpha_1m_2 + m_1\alpha_m + m_2\alpha_m - \alpha_1\alpha_2, \quad m_1m_2 + m_1\beta_2 + m_2\beta_1; \quad \alpha = < m_1m_2, \alpha_1m_2 + m_1\alpha_m + m_2\alpha_m - \alpha_1\alpha_2, \quad \beta = < m_1m_2, \alpha_1m_2 + m_1\alpha_m + m_2\alpha_m - \alpha_1\alpha_2. \]
(31)

\[
m_1\beta_1 = < m_1m_2, \alpha_1m_2 + m_1\alpha_m + m_2\alpha_m - \alpha_1\alpha_2, \quad m_1m_2 + m_1\beta_2 + m_2\beta_1; \quad \alpha = < m_1m_2, \alpha_1m_2 + m_1\alpha_m + m_2\alpha_m - \alpha_1\alpha_2, \quad \beta = < m_1m_2, \alpha_1m_2 + m_1\alpha_m + m_2\alpha_m - \alpha_1\alpha_2. \]
(32)

\[
m_1\beta_2 = < m_1m_2, \alpha_1m_2 + m_1\alpha_m + m_2\alpha_m - \alpha_1\alpha_2, \quad m_1m_2 + m_1\beta_2 + m_2\beta_1; \quad \alpha = < m_1m_2, \alpha_1m_2 + m_1\alpha_m + m_2\alpha_m - \alpha_1\alpha_2, \quad \beta = < m_1m_2, \alpha_1m_2 + m_1\alpha_m + m_2\alpha_m - \alpha_1\alpha_2. \]
(33)

Now, in accordance with (19), it is required to determine the parameters of the belonging functions of the following fuzzy numbers:
\[
m_1\alpha_2 + m_2\alpha_1 - \alpha_1\alpha_2, \quad m_1\beta_2 + m_2\beta_1 + \beta_1\beta_2. \]

Acting similarly to the previous one, we get:
\[
m_1\alpha_2 + m_2\alpha_1 - \alpha_1\alpha_2 = < m_1m_2, \alpha_1m_2 + m_1\alpha_m + m_2\alpha_m - \alpha_1\alpha_2, \quad m_1m_2 + m_1\beta_2 + m_2\beta_1; \quad \alpha = < m_1m_2, \alpha_1m_2 + m_1\alpha_m + m_2\alpha_m - \alpha_1\alpha_2, \quad \beta = < m_1m_2, \alpha_1m_2 + m_1\alpha_m + m_2\alpha_m - \alpha_1\alpha_2. \]
(34)
Thus, the obtained relations (17)-(34) allow us to calculate the parameters values of the belonging functions of fuzzy and non-fuzzy results of arithmetic operations on fuzzy numbers of \((L−R)\)-type. At the same time, the clustering problem solution in conditions of uncertainty can be performed using a simple step-by-step algorithm. When implementing this algorithm, the distances from this object to each of the grouping centers are sequentially calculated for each clustering object. After that, this object is attached to the cluster specified by the "nearest" grouping center. The meaning of the term "nearest" should be clarified. When comparing the distances from the clustered object to the two grouping centers, the second of the obtained fuzzy numbers is subtracted from the first. If the resulting difference \(R\) is positive, then the second distance is smaller than the first and, therefore, the second grouping center is "closer" to the object than the first. The sign of a fuzzy number \(R\) is determined by the formula

\[
\xi = \frac{\int \mu(R)\alpha R}{\int \mu(R)\alpha R}.
\]

At the same time, if \(\xi < 0.5\), then the number \(R\) is negative, if \(\xi = 0.5\), then \(R\) is a fuzzy zero, if \(\xi > 0.5\), then \(R\) is positive.

**Conclusions.** The problem of clustering a set of objects whose position in the phase space is given by their coordinates is considered. Analytical expression for clustering efficiency criterion in the form of a doubly fractional–quadratic function is obtained. Step-by-step method for solving the problem is proposed. The resulting method is extended to the case when the coordinates of objects subject to clustering are given indistinctly by their belonging of \((L−R)\)-type functions on infinite or compact carriers.

**References**


Отримана в редакції 03.01.2024. Прийнята до друку 14.02.2024. Received 02 January 2024. Approved 14 February 2024. AvailableinInternet 12 April 2024