Available online at www.ijournalse.org

Emerging Science Journal

(ISSN: 2610-9182)

Vol. 5, No. 5, October, 2021

Cluster Data Analysis with a Fuzzy Equivalence Relation to Substantiate a Medical Diagnosis

Abas Hasanovich Lampezhev [1], Elena Yur`evna Linskaya [1], Aslan Adal`bievich Tatarkanov [1*], Islam Alexandrovich Alexandrov [1]

[1] *IDTI RAS Institute for Design - Technological Informatics of RAS, Moscow, Russian Federation*

Abstract

This study aims to develop a methodology for the justification of medical diagnostic decisions based on the clustering of large volumes of statistical information stored in decision support systems. This aim is relevant since the analyzed medical data are often incomplete and inaccurate, negatively affecting the correctness of medical diagnosis and the subsequent choice of the most effective treatment actions. Clustering is an effective mathematical tool for selecting useful information under conditions of initial data uncertainty. The analysis showed that the most appropriate algorithm to solve the problem is based on fuzzy clustering and fuzzy equivalence relation. The methods of the present study are based on the use of this algorithm forming the technique of analyzing large volumes of medical data due to prepare a rationale for making medical diagnostic decisions. The proposed methodology involves the sequential implementation of the following procedures: preliminary data preparation, selecting the purpose of cluster data analysis, determining the form of results presentation, data normalization, selection of criteria for assessing the quality of the solution, application of fuzzy data clustering, evaluation of the sample, results and their use in further work. Fuzzy clustering quality evaluation criteria include partition coefficient, entropy separation criterion, separation efficiency ratio, and cluster power criterion. The novelty of the results of this article is related to the fact that the proposed methodology makes it possible to work with clusters of arbitrary shape and missing centers, which is impossible when using universal algorithms.

Keywords:

Medical Decision Support System;
Fuzzy Logic;
Fuzzy Clustering Algorithms;
k-means Algorithm;
c-means Algorithm.

Article History:

Received:	23	June	2021
Revised:	19	September	2021
Accepted:	27	September	2021
Published:	01	October	2021

1- Introduction

Information processes, primarily including information flows transition, are increasingly affecting the practices of medicine and health care. Making effective medical decisions in the diagnosis, treatment, organization, and management of health issues requires a substantial amount of statistical information that must meet the requirements of reliability, completeness, relevance, and availability. The necessary data relate to patient characteristics (including information about a patient's health status, medical history, distinctive physical features, etc.), medical services (including treatment method, treatment regimen, etc.), and healthcare facility management (including information about doctors, equipment used, medication, and the final cost of treatment).

Databases of statistical medical information incorporate standard medical information, questionnaire data, disease history records, examination results, test results, medical facility reports, assessments, research procedures and results, information about the development of new schemes and models, and more. The medical diagnostic data acquired from

* **CONTACT**: as.tatarkanov@yandex.ru

DOI: http://dx.doi.org/10.28991/esj-2021-01305

these sources may be complex, and their formalization is sometimes very difficult. For example, the information obtained from patients is often characterized by vagueness: "severe pain" or "not severe pain," "weak" or "strong," "recently" or "some time ago." Patients are often unable to accurately recall the exact time of symptoms initial manifesting [1-3].

Thus, medical decision-making often involves the need to analyze a large amount of statistical information, which a priori is not always complete, explicit, or accurate. That is, there is considerable uncertainty in the initial medical data.

Medical Information Systems (MIS) [4-6] are invaluable resources that enable physicians to automatically obtain the comprehensive information necessary for the performance of their professional activities (including establishing diagnoses, describing the problem, and prescribing treatment or courses of rehabilitation) [7-9]. With the help of the core component of the MIS – the Medical Decision Support System (MDSS) – it is possible to collect, structure, store, systematize, analyze, and provide significant amounts of di-verse information on a wide range of processes and problems [10-12].

If the most relevant information identified through MDSS is used timely and reasonably, a wide variety of medical challenges can be addressed qualitatively. An effective means of detecting such information in large datasets accumulated by MDSS is the process of data mining aimed at identifying patterns and trends in the data. Mathematical tools to achieve this goal include a wide range of algorithms for classification, regression, clustering, prediction, and detection of sequences and associations [13-15].

Clustering algorithms are the best method to divide the data into separate groups with certain attributes and make specific conclusions and assumptions about each group. Thanks to the cluster research results, primarily using fuzzy clustering algorithms, where each data object belongs to different clusters with certain values of the fuzzy membership function, it is possible to view large amounts of medical data (including fuzzy data) and reduce it purposefully to effectively resolve the pressing problems of differentiation of significant and unnecessary information, simplifying its further processing [16-18].

Thus, the medical decision-making adequacy is due not only to the systematic accumulation of significant volumes of diverse and various (including semi structured or poorly formalized) medical statistical information for all types of processes and problems, but also to its proper analysis and processing, aimed for reasonable selecting data sets. The last makes it possible to determine the necessary tools for a specific medical problem to describe it, establish a diagnosis, and prescribe treatment [19-21].

In healthcare facilities, the efficiency of using periodically accumulated statistical information in decision support tasks determines the theoretical value of improved methods and algorithms designed to increase the objectivity and reduce the influence of human factors on the decision-making process, especially concerning ambiguity, incompleteness, and uncertainty associated with the initial information. Thus, it is relevant to solve a set of practical problems aimed at MDSS implementation. These include formalizing the problem solution of preparing a rationale for selecting the most appropriate medical diagnostic decision (MDD) from the list of recommended options. This issue represents the primary motivation for the present work to develop a methodology for preparing the abovementioned rationale through the clustering a large volume of statistical information stored in MDSS.

2- Literature Review

An essential role of automated MDSS is to prepare a rationale for selecting the most appropriate MDD for the patient from the list of recommended options [12, 22]. The mandatory first step in performing this task was to analyze the initial set of diverse statistical data and select only that information array which is important or desirable for a particular medical purpose. Notably, the degree of confidence in the selected information and in the results of its further targeted use by the decision-maker largely depend on how logically and mathematically correctly this analysis will be done [23]. It is necessary to have an algorithmic apparatus appropriate to the task and its efficient use in applied medicine to achieve this goal.

The most effective mathematical tool for analyzing a large volume of statistical in-formation is clustering, especially for the uncertainty associated with initial medical data. There are many explicit and fuzzy clustering algorithms, each with its distinct advantages, disadvantages, and specific implementation details. There are hierarchical and genetic versions based on fuzzy clustering and others [24-26].

Explicit clustering algorithms subdivide the initial set of objects X into several disjointed subsets. In this case, any object from X belongs to only one cluster. Fuzzy clustering algorithms allow the same object to belong to several (or even all) clusters simultaneously, though with varying degrees. Fuzzy clustering is more natural than explicit clustering in real situations because objects that correspond precisely to one or another category or class are rarely found. A particular object may have some of the attributes, while another part may be absent. Thus, the membership of such an object to any class turns out to be fuzzy. Formulas for setting the membership functions of fuzzy variables in the general case [27] take the form (1). Modal values of membership functions coincide with the centers of clusters, as shown in Figure 1.

$$\mu A_{11}(x_1)=\begin{cases}1 \text{ for } x_1 < a_{11},\\ \dfrac{x_1-a_{12}}{a_{11}-a_{12}} \text{ for } a_{11}\le x_1 < a_{12},\end{cases}$$

$$\mu A_{12}(x_1)=\begin{cases}\dfrac{x_1-a_{11}}{a_{12}-a_{11}} \text{ for } a_{11}\le x_1 < a_{12},\\ \dfrac{x_1-a_{13}}{a_{12}-a_{13}} \text{ for } a_{12}\le x_1 < a_{13},\end{cases}$$

$$\mu A_{13}(x_1)=\begin{cases}\dfrac{x_1-a_{12}}{a_{13}-a_{12}} \text{ for } a_{12}\le x_1 < a_{13},\\ 1 \text{ for } x_1 \ge a_{13},\end{cases}$$

$$\mu A_{21}(x_2)=\begin{cases}1 \text{ for } x_2 < a_{21},\\ \dfrac{x_2-a_{22}}{a_{21}-a_{22}} \text{ for } a_{21}\le x_2 < a_{22},\end{cases}$$

$$\mu A_{22}(x_2)=\begin{cases}\dfrac{x_2-a_{21}}{a_{22}-a_{21}} \text{ for } a_{21}\le x_2 < a_{22},\\ \dfrac{x_2-a_{23}}{a_{22}-a_{23}} \text{ for } a_{22}\le x_2 < a_{23},\end{cases} \qquad (1)$$

$$\mu A_{23}(x_2)=\begin{cases}\dfrac{x_2-a_{22}}{a_{23}-a_{22}} \text{ for } a_{22}\le x_2 < a_{23},\\ 1 \text{ for } x_2 \ge a_{23},\end{cases}$$

$$\mu B_1(y)=\begin{cases}1 \text{ for } y < b_1,\\ \dfrac{y-b_2}{b_1-b_2} \text{ for } b_1\le y < b_2,\end{cases}$$

$$\mu B_2(y)=\begin{cases}\dfrac{y-b_1}{b_2-b_1} \text{ for } b_1\le y < b_2,\\ \dfrac{y-b_3}{b_2-b_3} \text{ for } b_2\le y < b_3,\end{cases}$$

$$\mu B_3(y)=\begin{cases}\dfrac{y-b_2}{b_3-b_2} \text{ for } b_2\le y < b_3,\\ 1 \text{ for } y \ge b_3.\end{cases}$$

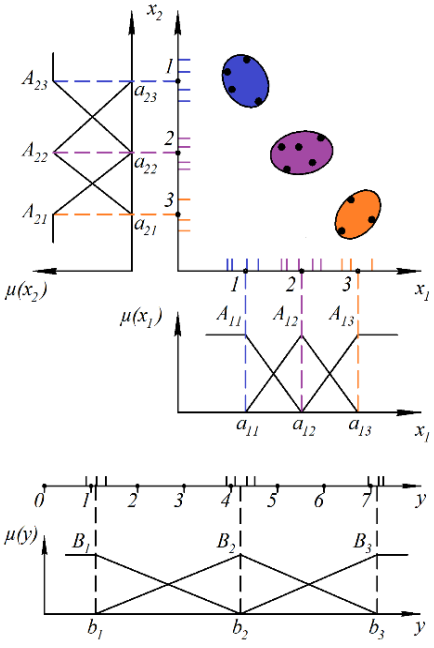

Figure 1. Example of projection of results on individual axes, cluster coordinates, and membership functions located in the centers of clusters.

Then the set of candidate rules Ri, constructed based on all possible combinations of input and output fuzzy sets A_{1i}, A_{2j}, B_k, is formed. These fuzzy rules, following which the clustering is performed, have the following form:

$$\begin{aligned}
&R_1: && \text{IF } (x_1=A_{11}) \text{ AND } (x_2=A_{21}) \text{ TO } (y=B_1),\\
&R_2: && \text{IF } (x_1=A_{11}) \text{ AND } (x_2=A_{21}) \text{ TO } (y=B_2),\\
&R_3: && \text{IF } (x_1=A_{11}) \text{ AND } (x_2=A_{21}) \text{ TO } (y=B_3),\\
&R_4: && \text{IF } (x_1=A_{11}) \text{ AND } (x_2=A_{22}) \text{ TO } (y=B_1),\\
&R_5: && \text{IF } (x_1=A_{11}) \text{ AND } (x_2=A_{22}) \text{ TO } (y=B_2),\\
&R_6: && \text{IF } (x_1=A_{11}) \text{ AND } (x_2=A_{22}) \text{ TO } (y=B_3),\\
&\vdots && \vdots\\
&R_{27}: && \text{IF } (x_1=A_{13}) \text{ AND } (x_2=A_{23}) \text{ TO } (y=B_3).
\end{aligned} \tag{2}$$

Thus, the set contains 27 fuzzy rules for each of which confidence coefficients are calculated corresponding to specific elements of the sample, and then the maximum values of the confidence coefficient are determined.

Methods of statistical information clustering in MDSS have already been widely used, and some examples of relevant research works are presented below. Thong et al. [28], developed a hybrid model that combines fuzzy clustering of images and intuitionistic fuzzy recommendation systems for medical diagnosis. The authors focused on improving the quality of medical diagnosis, and as a result, the accuracy of the hybrid model they developed was better than that of other relevant algorithms. The high accuracy of the hybrid model has been experimentally verified on the UCI machine learning reference dataset. The disadvantages of the proposed hybrid model are its limited application area related to image processing. Masulli and Schenone [29], developed a similar system for segmentation based on fuzzy clustering to support diagnosis in medical imaging. Due to noise, there is uncertainty in the medical imaging. In particular, the boundaries between tissues are not precisely defined, and the belonging to boundary regions is fuzzy. Thus, computer methods of uncontrolled fuzzy clustering prove to be particularly suitable for processing the decision-making process regarding the segmentation of multimodal medical images. The authors applied a widely used c-means algorithm as the basis for neural network-based clustering. The resulting solution is designed to work with images, and this defines the area of its use. Poczeta et al. [30] considers the task of processing multivariate medical data related to Parkinson's disease, for which the authors use fuzzy cognitive maps and k-means clustering. They used the k-means method to group the data and then constructed a separate fuzzy cognitive map for each cluster to improve the accuracy of predictions.

The range of fuzzy clustering algorithms is broad enough: fuzzy k-means algorithm, fuzzy c-means (FCM) algorithm, fuzzy decision trees, fuzzy Petri nets, fuzzy associative memory, fuzzy self-organizing maps, and others [31-33]. The k-means algorithm, the basis of a more advanced method of fuzzy c-means clustering [34, 35], is fundamental. These algorithms became the basis for many other ones in this class, and they have enough multiprogram implementations, for example, the FCM algorithm built into MATLAB.

The k-means method works well when clusters are significantly separated compact clouds. It is effective for processing large amounts of data, but it is not applicable for detecting clusters of nonconvex shape or very different sizes. The fuzzy c-means clustering method can be seen as an improved k-means one: in it for each element in the considered set, the degree of its belonging to each of the clusters is calculated. The fuzzy c-means clustering method has limited application due to a significant disadvantage – the impossibility of correct partitioning into clusters when they have different variance on different dimensions (axes) of elements (for example, if the cluster is elliptical). FCM algorithm is an unsupervised fuzzy clustering method, which does not require human intervention in algorithm implementation. For the FCM algorithm, "c" is identical to "k" for k-means relating to the number of clusters. "F" is a fuzzy value referring to the incident degree. The disadvantage of the algorithm is that some initial parameters must be set. The invalid initial choice of parameters may affect the correctness of the clustering results. When the data sample set and the number of functions are large, the real-time performance of the algorithm is low.

Based on the above information-analytical review, the following hypotheses were formulated to achieve the aim of the study:

2-1- Hypothesis 1 (H1)

Simplicity, a high implementation speed, and the effectiveness of initial partitioning into clusters are the advantages of fuzzy clustering algorithms in solving many practical problems. However, their use in solving problems with the need to analyze large amounts of semi structured medical information in many cases provides unreasonable decisions. This is since insufficiently versatile tools of these algorithms fail to account for the fact that, usually, the form of clusters can be any, and cluster centres may be absent or unidentified. Thus, the procedures of partitioning objects into clusters are based only on identifying the interrelation between objects and cluster centres but not on the dependence of data objects on each other.

2-2- Hypothesis 2 (H2)

For the analysis of semi structured medical information, the use of an algorithm developed through the fuzzy clustering method, based on the fuzzy relation of equivalence, and generated by the properties of the data under study, seems promising [36]. This algorithm, in which the attribute relationship of the data under study is considered as fuzzy object relationships, makes it possible to identify clusters of arbitrary shapes productively. Selecting the best solution to the fuzzy clustering problem is performed without using additional information about the clusters.

2-3- Hypothesis 3 (H3)

When using the fuzzy clustering method based on the fuzzy relation of equivalence, its adjustment and adaptation for each specific type of medical diagnostic task is required. Furthermore, it may require the addition of other algorithms. Therefore, it is of interest to create a generalized methodology for preparing a rationale for making appropriate MDD based on the clustering of a large volume of statistical information stored in MDSS.

3- Research Methodology

The workability and efficiency of the fuzzy clustering algorithm based on the fuzzy equivalence relation make it possible to use for the hardware implementation of MDSS in many areas of the medical field [37]. The following procedure is aimed at ensuring efficiency when this method is used to analyze the statistical data required for making decisions in applied medicine. A flowchart explaining the methodology is shown in Figure 2. The proposed approach is based on the clustering of initial statistical data using a fuzzy equivalence relation and includes a mandatory sequence of steps:

3-1- Preliminary Data Preparation

The preparatory process involves the selection of the object set for analysis and attributes selection. It is essential that they clearly and fully reflect the considered set. During this stage, the medical technologies to be applied and the procedures involved will be formalized.

3-2- Establishment of Goals of Data Cluster Analysis

Possible goals include:

- Determining the number of clusters and identifying their composition for determining cluster composition of the data under study;
- Identifying the elements of the object set that are not part of the clusters (the deviations found show the pathology in the ongoing process);
- Data preparation based on cluster analysis results to solve the problem of classifying and processing results.

3-3- Defining the Representation form of Results

The results of fuzzy clustering data analysis, depending on the type of data, can be represented as:

- Simple enumeration (a universal method of representation where each cluster is identified by its elements);
- Tables (the most appropriate way to represent the results of fuzzy clustering: the rows of the table correspond to data objects, columns indicate the clusters, and the values in table cells correspond with values of the membership function).

3-4- Data Normalization

Data normalization is the conversion of ordinal and categorical data into numerical values. When normalizing numerical data in the range of 0 to 1, all weighting coefficients must be equal when comparing data. Consequently, when the attribute weights are different, a single variable needs to be used to process the data. Data normalization is usually carried out based on peer reviews.

3-5- Criteria Selection for Assessing the Quality of Decisions

The aim is to assess the quality of fuzzy clustering results so that effective medically related decisions can be made. Therefore, partition coefficients, entropy partition criteria, partition efficiency coefficients, and cluster power criteria should be used.

3-6- Application of Data Fuzzy Clustering

A cluster analysis method based on fuzzy equivalence relation will be applied to medical statistics.

3-7- Analysis of the Results and Recommendations for Their Utilization in Further Work

A brute-force search of values from a given range of the number of clusters and calculating of criteria taken for analysis are carried out. Then the best partitioning is selected by analyzing the set of criteria extremums. The next operation is measuring either deviations or results preparation for classification, depending on the goals.

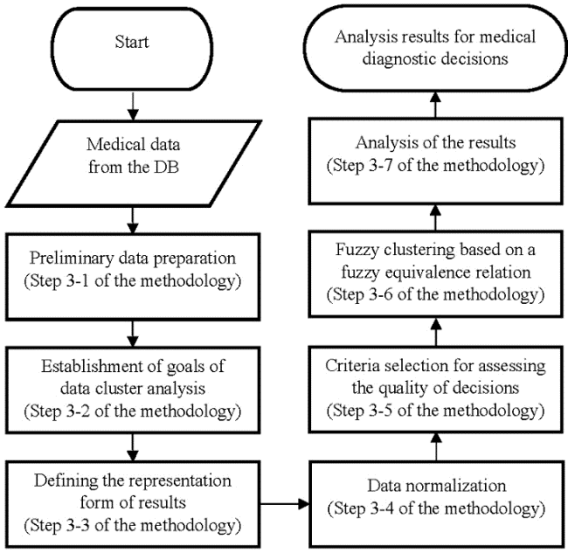

Figure 2. A flowchart explaining the methodology.

4- Results

Fuzzy clustering of medical data based on fuzzy equivalence relations under the proposed algorithm is carried out consistently according to the steps outlined below.

4-1- Step 1

Determination of the normal similarity measure by distance for each attribute $p_j \in P$ of the set of all attributes $\boldsymbol{P}$ by the formula*:*

$$\mu_{p_i}(p_j) = 1 - \frac{d(p_i, p_j)}{\max\limits_{p_k \in P} d(p_i, p_k)}, \tag{3}$$

where $d(p_i, p_j)$ – the distance between attributes p_i and p_j.

Obviously $\mu_{p_i}(p_j) \in [0,1]$. If $\mu_{p_i}(p_j) = 0$, then p_i differs from p_j. In the case if p_i is absolutely similar to p_j, $\mu_{p_i}(p_j) = 1$.

Thus, in the process of calculating the normal similarity measure of attribute p_i by distance for each attribute $p_j \in P$, fuzzy subsets of attributes similar to it are formed.

4-2- Step 2

Determination of relative similarity measure $\xi_{P_i}(p_j, p_k)$ of pair of attributes $p_j, p_k \in P$ regarding the third attribute $p_i \in P$ of the set of all attributes P by the formula:

$$\xi_{P_i}(p_j, p_k) = 1 - \left| \mu_{p_i}(p_j) - \mu_{p_i}(p_k) \right|, \tag{4}$$

where $\mu_{p_i}(p_j)$ and $\mu_{p_i}(p_k)$ – normal similarity measure relative to p_j and p_k.

4-3- Step 3

Determination of similarity measure $\xi(p_j, p_k)$ of pair of attributes $p_j, p_k \in P$ on the set $\boldsymbol{P}$ *by the formula*:

$$\xi(p_j, p_k) = T\left[\xi_{P_i^{|P|}}(p_j, p_k) \right], i = 1, \ldots, |P|, \tag{5}$$

where $\xi_{P_i^{|P|}}(p_j, p_k)$ – relative similarity measure; T – t-norm binary operation.

When using the Zadeh t-norm cluster, similarity measure $\xi(p_j, p_k)$ of pair of attributes $p_j, p_k \in P$ on the set $\boldsymbol{P}$ is determined by the formula:

$$\xi(p_j, p_k) = \min\left[\xi_{P_i^{|P|}}(p_j, p_k)\right], i = 1,...,|P|. \quad (6)$$

4-4- Step 4

Determination of the fuzzy equivalence relation $R_\xi^{|P|}$ based on to the calculation results of the transitive closure of the fuzzy relation in the cycle by the formula:

$$R_\xi^t = R_\xi^{t-1} \circ R \quad (7)$$

where $R_\xi = \xi(p_j, p_k);\ t = 2,...,|P|;\ R_\xi^{|P|}$.

4-5- Step 5: Gradation of Fuzzy Equivalence Relation G_ξ by Ranking Elements of Its Matrix $R_\xi^{|P|}$

Gradation of fuzzy equivalence relation G_ξ creates many equivalence relations, and they all make it possible to partition the initial family into classes of equivalence. The size of detailed partitioning of the initial set P directly depends on the level of the relation. A more detailed partitioning of the set P corresponds to a higher level of relation.

4-6- Step 6: Selection of the Level of Fuzzy Equivalence Relation L_i for Partitioning the Initial Set Into Clusters

Partitioning into clusters depends on the selected level of fuzzy relation L_i; in this case, the number and composition of clusters change. According to the presented algorithm of fuzzy clustering using fuzzy equivalence relation, the best partitioning into clusters should be considered the result that meets the quality criteria of fuzzy clustering. To assess the quality of fuzzy clustering based on fuzzy equivalence relation, the following criteria and some of their modifications are most effective.

4-6-1- Partition Coefficient K_{pc}

Calculated by the formula:

$$K_{pc} = \sum_{i=1}^{|P|}\sum_{j=1}^{|CL|}\frac{r_{ij}^2}{|P|},\ K_{pc} \in \left[1/|CL|, 1\right], \quad (8)$$

where P is the initial set of attributes; CL – set of clusters; r_{ij} – element of fuzzy equivalence relation matrix $R_\xi^{|P|}$. The maximum value of the coefficient K_{pc}=1 indicates the maximum uncertainty; therefore, the obtained partitioning is considered to be the worst.

It is also worth noting that when there are not enough clusters, the obtained value of the partition coefficient is inadequate for its range of values. In this case, it is reasonable to use a modified partition coefficient K_{mpc} calculated by the Equation 9. The essence of this modification is to move only its range of values.

$$K_{mpc} = \sum_{i=1}^{|P|}\sum_{j=1}^{|CL|}\frac{r_{ij}^2}{|P|} - \frac{1}{|CL|},\ K_{mpc} \in \left[0, \frac{|CL|-1}{|CL|}\right]. \quad (9)$$

In this case, the dependence of the modified partition coefficient K_{mpc} on the number of clusters resulted from the end of the partition coefficient range of values.

4-6-2- Entropy Partition Criterion K_{ep}

Calculated by the formula:

$$K_{ep} = \frac{\sum_{i=1}^{|P|}\sum_{j=1}^{|CL|}\left|r_{ij}\ln(r_{ij})\right|}{|P|},\ K_{ep} \in \left[0, \ln(|CL|)\right], \quad (10)$$

A lower value of K_{ep} corresponds to a greater degree of an element belonging to one cluster. K_{ep}= ln $|CL|$ is considered the worst partitioning, and K_{ep} =0 is the best partitioning. The number of clusters strongly affects values of K_{ep} =0. Thus, a small number of clusters corresponds to low values of K_{ep} =0, and a large number of clusters corresponds to large values of K_{ep} =0. More correct results of partitioning assessment can be obtained using modified entropic criterion K_{mep} calculated by the formula:

$$K_{mep} = \frac{\sum_{i=1}^{|P|}\sum_{j=1}^{|CL|}\left|r_{ij}\ln(r_{ij})\right|}{|P|\ln(|CL|)},\ K_{mep} \in [0,1]. \quad (11)$$

where K_{mep} is not linked to the number of clusters, so if the number of clusters is different, it can be used to compare the results of different clustering methods.

4-6-3- Partition Efficiency Coefficient

Partition efficiency coefficient K_{pe} is determined by the difference between the coefficient of intra-cluster differences K_{pei} and coefficient of cross-cluster differences K_{pec} by the formula:

$$K_{pe} = \sum_{i=1}^{|P|}\sum_{j=1}^{|CL|} r_{ij}^2 d^2(c_j, \bar{p}) - \sum_{i=1}^{|P|}\sum_{j=1}^{|CL|} r_{ij}^2 d^2(p_i, c_j) \quad (12)$$

where: P is the initial set of attributes; p_i – the i-th component of the set P; p – the average value of p_i components; CL – the set of clusters; c_j – the center of the j-th cluster $cl_j \in CL$; r_{ij}– the element of the fuzzy equivalence relation matrix $R_{\xi}^{|P|}$; $d(p_i, c_j)$ representing the distance between the two objects p_i and c_j.

A higher value of the coefficient K_{pec} corresponds to a more qualitative partitioning – that is, at the optimal value of the number of clusters, the value of K_{pec} tends to the maximum. Modified partition coefficient and entropy criterion have no link to the number of clusters. Therefore, using them, it is possible to assess the quality of clustering both on a large and a low number of clusters and obtain the results of assessments in the range [0, 1].

4-6-4- Cluster Power Criterion

Cluster power criterion is based on the concept of *powerful cluster* understood as a cluster considered being of practical use at some importance of the equivalence class in the fuzzy equivalence relation gradation. This quality assessment algorithm, using this criterion, is based on the concepts of equivalence relation level of powerful clusters and intermediate coefficient.

The above criteria make it possible to meet the ambiguous clustering problems effectively. For example (Figure 3), two clusters, clearly separated in two-dimensional attribute space $X \times Y$, overlap when projected on the x-axis, with the result that one-dimensional analysis leads to the conclusion about the existence of one cluster. This makes it possible to set in space one cluster A_1, whose center a_1 does not correspond to any of the centers of two-dimensional clusters. A similar case of complete or partial overlapping of clusters may arise for the y-axis, thus depriving the possibility to correctly determine the number of clusters and coordinates of their centers without using the criteria of clustering quality assessment.

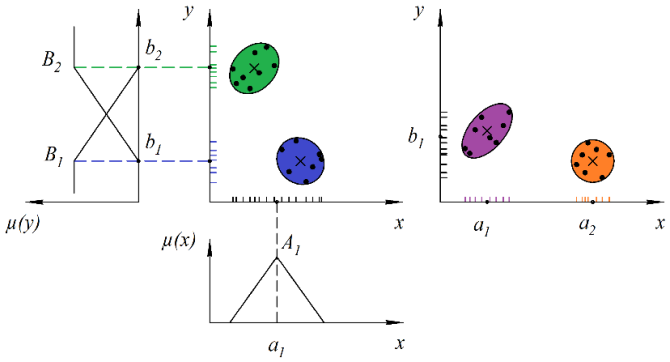

Figure 3. Example of overlapping clusters when projected on different axes.

5- Discussion

5-1- Main Findings of the Present Study

The discussion focuses on the results of the data analysis to prepare a rationale for MDD selection. The main result of using the proposed methodological approach of fuzzy clustering of medical statistical data based on fuzzy equivalence relations is the partitioning large volume clusters of statistical information stored in MDSS which corresponds to a particular clustering goal in conditions of the uncertain initial medical data. Such a goal could be, for example, preparing a rationale for MDD selection. To solve this problem, the methodological approach of fuzzy clustering of medical statistical data based on a fuzzy equivalence relation can be formalized more specifically as follows.

The medical statistical information used in MDSS for clustering is a health card (HC) for each patient, displaying his or her health status (e.g., the patient's body temperature and the results of blood and urine tests, etc.). Each HC is characterized by classification attributes (measure, value, patient characteristics, importance, and norms). The system must store the HC for each patient in a normalized form [38, 39]. For fuzzy clustering, an initial set P is formed based on HC values for each patient.

The fuzzy clustering algorithm performs partitioning of the HC set by classifying attributes into clusters representing subsets of the initial set P. Based on the results of fuzzy clustering (set *CL* of HC clusters, membership matrix r_{ij}), it is possible to conduct MDD selection. MDSS must store MDD templates for each diagnosis. These templates are compiled based on the results of analysis of medical statistics by a panel. For each template, a set of HCs with certain values of weight coefficients ω_{ij} should be stored. If a patient's diagnosis is defined, the MDD corresponds to the template stored in the system. Otherwise, it is possible to get an MDD from a general list without linking it to a diagnosis. In this case, the MDD list will be much larger, but the accuracy of the proposed decision will be lower. In this regard, it is necessary to compare the value of the patient's HC weight coefficients r_{ij}, obtained by fuzzy clustering, with the weight coefficients of the templates. Then, based on this comparison, a ranked list of possible MDDs should be generated. A ranked list of possible MDDs is generated based on the similarity measure assessment between the patient's HC and HC of templates. The score $\phi(S_k)$, where S is the set of templates that specifies the MDD similarity measure of the current patient's health status (PHS), is determined by the following formula:

$$\phi(S_k) = \sum_{j=1}^{N} r_j \sum_{i=1}^{M} \omega_i \mu(C_i, C_j), \tag{13}$$

where N is the number of concepts belonging to the PHS model; M is the number of concepts belonging to template S_k; r_j is the importance of the concept in the patient's situation; ω_i is the importance of the concept in the template; and $\mu(C_j, C_i)$ is the similarity of the i-th and j-th concepts. A higher value of $\phi(S_k)$corresponds to the template that is closer to the patient's situation and that has greater importance in the set of actions for the patient. The number $\phi(S_k)$, belonging to template S_k, is called a criterion score, and the generated scale is a criterion scale. Thus, the desired template will be a set of smaller templates that meet the condition $\max_{S_k \in S} \phi(S_k)$.

When compared with other templates in the medical area, the maximum criterion score value templates are included in the list of selected templates in ranked order from highest to lowest value. The doctor selects the most appropriate option from the recommended list according to the patient's situation. Then, the patient's treatment method is formed according to the selected MDD.

The main result of using the proposed methodological approach to the analysis of initial medical data is the appropriate clustering of a large volume of statistical information stored in MDSS, including in the context of the uncertainty of the initial medical data.

5-2- Comparison with Other Studies

The formalized methodology of motivational base preparation proposed in this paper should provide a uniform choice of the most appropriate options of MDD based on the clustering of a large volume of statistical information stored in MDSS. The opportunity to work with clusters of arbitrary shape and missing centers provides an advantage over known universal algorithms.

6- Conclusion, Recommendation, and Future Direction

Computer technology is becoming an integral part of all areas of medicine and health care [40]. Decision support systems, which accumulate significant volumes of statistical information of a medical nature, make it possible to obtain in automated mode only the information that is required to provide a motivational basis for selecting the most appropriate MDD option for a particular patient from a recommended list.

The most effective mathematical tool for selecting (from the entire array of accumulated data) information suitable for a specialist, especially in the uncertainty of the initial medical data, is clustering. Among the wide range of known algorithms of explicit and fuzzy clustering, the most suitable to solve practical problems related to the need for analysis of poorly formalized and semi-structured information is the algorithm developed through the fuzzy clustering method and based on the fuzzy relation of equivalence generated by the properties of the data under study. It has proven its effectiveness in solving many practical problems.

The mathematical apparatus implemented within this fuzzy clustering algorithm, based on fuzzy equivalence relation, forms the basis of the proposed methodological approach to the initial statistical data analysis necessary to make medical decisions. This approach consistently implements the following procedures: preliminary data preparation, goal selection of data cluster analysis, definition of the resulting representation form, data normalization, selection of decision quality evaluation criteria, application of fuzzy clustering of data, assessment of the sample results, and their use in further work.

The formalized methodology of motivational base preparation proposed in this paper should provide a uniform choice of the most appropriate options of MDD based on the clustering of a large volume of statistical information stored in MDSS. The application of the proposed methodology and algorithms has a limitation due to the following disadvantage – the inability to correctly partition into clusters when they have significantly different variance in different dimensions. Thus, eliminating this disadvantage determines the prospects for further research to improve the approaches outlined in this paper.

6-1- Strengths and Limitations

It should be noted that, when implementing and using the proposed methodology to process medical data, the following should be taken into account:

- Statistical medical data for analysis should be preliminarily checked for outliers and incorrect elements by experts in the field of data engineering;
- The variance across different dimensions (axes) in medical data clusters should not differ significantly (approximately no more than 25%).

7- Declarations

7-1- Author Contributions

Conceptualization, I.A.; methodology, I.A.; writing – original draft preparation, A.L.; writing – review and editing, E.L.; supervision, A.T.; project administration, A.T. All authors have read and agreed to the published version of the manuscript.

7-2- Data Availability Statement

The data presented in this study are available in article.

7-3- Funding

Selected findings of this work were obtained under the Grant Agreement in the form of subsidies from the federal budget of the Russian Federation for state support for the establishment and development of world-class scientific centers performing R&D on scientific and technological development priorities (internal number 00600/2020/56890) dated November 13, 2020, No. 075-15-2020-929.

7-4- Acknowledgements

The authors are grateful to Professor L. Chervyakov for a careful discussion of this paper.

7-5- Ethical Approval

The article follows the guidelines of the Committee on Publication Ethics (COPE) and involves no studies on human or animal subjects. Consent to participate is not applicable, since the research doesn't involve studies on humans.

7-6- Conflicts of Interest

The authors declare that there is no conflict of interests regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancies have been completely observed by the authors.

8- References

[1] Bricon-Souf, N., C. Verdier, A. Flory, and M.C. Jaulent. "Theme C: Medical Information Systems and Databases – Results and Future Work." IRBM 34, no. 1 (February 2013): 9–10. doi:10.1016/j.irbm.2012.12.010.

[2] Wagholikar, Kavishwar Balwant, Kathy L MacLaughlin, Thomas M Kastner, Petra M Casey, Michael Henry, Robert A Greenes, Hongfang Liu, and Rajeev Chaudhry. "Formative Evaluation of the Accuracy of a Clinical Decision Support System for Cervical Cancer Screening." Journal of the American Medical Informatics Association 20, no. 4 (July 2013): 749–757. doi:10.1136/amiajnl-2013-001613.

[3] Piibe, Quinn, Erica Kane, Marlene Melzer-Lange, and Kathleen Beckmann. "Patient at Risk: Emergency Medical Service Providers' Opinions on Improving an Electronic Emergency Information Form Database for the Medical Care of Children with Special Health Care Needs." Disability and Health Journal 13, no. 2 (April 2020): 100852. doi:10.1016/j.dhjo.2019.100852.

[4] Andrikov, D.A., and A.S. Kuchin. "Development of a Prototype of a Medical Information System for a Clinical Diagnostic Center." Procedia Computer Science 186 (2021): 287–292. doi:10.1016/j.procs.2021.04.147.

[5] Chang, Wenjun, Qian Zhang, Chao Fu, Weiyong Liu, Guangquan Zhang, and Jie Lu. "A Cross-Domain Recommender System through Information Transfer for Medical Diagnosis." Decision Support Systems 143 (April 2021): 113489. doi:10.1016/j.dss.2020.113489.

[6] Anifah, Lilik, and Haryanto. "Decision Support System Two Dimensional Cattle Weight Estimation Using Fuzzy Rule Based System." 2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT) (April 9, 2021). doi:10.1109/eiconcit50028.2021.9431911.

[7] L, Arokia Jesu Prabhu, Sudhakar Sengan, Kamalam G K, Vellingiri J, Jagadeesh Gopal, Priya Velayutham, and Subramaniyaswamy V. "Medical Information Retrieval Systems for e-Health Care Records Using Fuzzy Based Machine Learning Model." Microprocessors and Microsystems (October 2020): 103344. doi:10.1016/j.micpro.2020.103344.

[8] Zhao, Yan, Li Liu, Yanbo Qi, Fengge Lou, Jingdan Zhang, and Wenhui Ma. "Evaluation and Design of Public Health Information Management System for Primary Health Care Units Based on Medical and Health Information." Journal of Infection and Public Health 13, no. 4 (April 2020): 491–496. doi:10.1016/j.jiph.2019.11.004.

[9] Tashkandi, Araek, Ingmar Wiese, and Lena Wiese. "Efficient In-Database Patient Similarity Analysis for Personalized Medical Decision Support Systems." Big Data Research 13 (September 2018): 52–64. doi:10.1016/j.bdr.2018.05.001.

[10] Cibella, Fabio, Simona Panunzi, Valerio Cusimano, and Andrea De Gaetano. "Decision Support for Medical Disasters: Evaluation of the IMPRESS System in the Live Palermo Demo." International Journal of Disaster Risk Reduction 50 (November 2020): 101695. doi:10.1016/j.ijdrr.2020.101695.

[11] Shaikh, Faiq, Jamshid Dehmeshki, Sotirios Bisdas, Diana Roettger-Dupont, Olga Kubassova, Mehwish Aziz, and Omer Awan. "Artificial Intelligence-Based Clinical Decision Support Systems Using Advanced Medical Imaging and Radiomics." Current Problems in Diagnostic Radiology 50, no. 2 (March 2021): 262–267. doi:10.1067/j.cpradiol.2020.05.006.

[12] Katzmann, Alexander, Oliver Taubmann, Stephen Ahmad, Alexander Mühlberg, Michael Sühling, and Horst-Michael Groß. "Explaining Clinical Decision Support Systems in Medical Imaging Using Cycle-Consistent Activation Maximization." Neurocomputing 458 (October 2021): 141–156. doi:10.1016/j.neucom.2021.05.081.

[13] Li, Haoran, Fazhi He, and Yilin Chen. "Learning Dynamic Simultaneous Clustering and Classification via Automatic Differential Evolution and Firework Algorithm." Applied Soft Computing 96 (November 2020): 106593. doi:10.1016/j.asoc.2020.106593.

[14] Yang, Chao-Lung, and Nguyen Thi Phuong Quyen. "Data Analysis Framework of Sequential Clustering and Classification Using Non-Dominated Sorting Genetic Algorithm." Applied Soft Computing 69 (August 2018): 704–718. doi:10.1016/j.asoc.2017.12.019.

[15] Wang, Yulong, Yuan Yan Tang, Cuiming Zou, Luoqing Li, and Hong Chen. "Modal Regression Based Greedy Algorithm for Robust Sparse Signal Recovery, Clustering and Classification." Neurocomputing 372 (January 2020): 73–83. doi:10.1016/j.neucom.2019.09.056.

[16] Xu, Kaijie, Witold Pedrycz, Zhiwu Li, and Weike Nie. "Optimizing the Prototypes with a Novel Data Weighting Algorithm for Enhancing the Classification Performance of Fuzzy Clustering." Fuzzy Sets and Systems 413 (June 2021): 29–41. doi:10.1016/j.fss.2020.05.009.

[17] Prajapati, Purvi, and Amit Thakkar. "Performance Improvement of Extreme Multi-Label Classification Using K-Way Tree Construction with Parallel Clustering Algorithm." Journal of King Saud University - Computer and Information Sciences (March 2021). doi:10.1016/j.jksuci.2021.02.014.

[18] Mouton, Jacques P., Melvin Ferreira, and Albertus S.J. Helberg. "A Comparison of Clustering Algorithms for Automatic Modulation Classification." Expert Systems with Applications 151 (August 2020): 113317. doi:10.1016/j.eswa.2020.113317.

[19] Luo, Kangqi, Jinyi Lu, Kenny Q. Zhu, Weiguo Gao, Jia Wei, and Meizhuo Zhang. "Layout-Aware Information Extraction from Semi-Structured Medical Images." Computers in Biology and Medicine 107 (April 2019): 235–247. doi:10.1016/j.compbiomed.2019.02.016.

[20] Tekli, Gilbert. "A Survey on Semi-Structured Web Data Manipulations by Non-Expert Users." Computer Science Review 40 (May 2021): 100367. doi:10.1016/j.cosrev.2021.100367.

[21] Hasan, Abul, Mark Levene, and David Weston. "Learning Structured Medical Information from Social Media." Journal of Biomedical Informatics 110 (October 2020): 103568. doi:10.1016/j.jbi.2020.103568.

[22] Tashkandi, Araek, Ingmar Wiese, and Lena Wiese. "Efficient In-Database Patient Similarity Analysis for Personalized Medical Decision Support Systems." Big Data Research 13 (September 2018): 52–64. doi:10.1016/j.bdr.2018.05.001.

[23] Shaikh, Faiq, Jamshid Dehmeshki, Sotirios Bisdas, Diana Roettger-Dupont, Olga Kubassova, Mehwish Aziz, and Omer Awan. "Artificial Intelligence-Based Clinical Decision Support Systems Using Advanced Medical Imaging and Radiomics." Current Problems in Diagnostic Radiology 50, no. 2 (March 2021): 262–267. doi:10.1067/j.cpradiol.2020.05.006.

[24] Galvani, Marta, Agostino Torti, Alessandra Menafoglio, and Simone Vantini. "FunCC: A New Bi-Clustering Algorithm for Functional Data with Misalignment." Computational Statistics & Data Analysis 160 (August 2021): 107219. doi:10.1016/j.csda.2021.107219.

[25] Nooraeni, Rani, Muhamad Iqbal Arsa, and Nucke Widowati Kusumo Projo. "Fuzzy Centroid and Genetic Algorithms: Solutions for Numeric and Categorical Mixed Data Clustering." Procedia Computer Science 179 (2021): 677–684. doi:10.1016/j.procs.2021.01.055.

[26] Dong, Yihong, Yueting Zhuang, Ken Chen, and Xiaoying Tai. "A Hierarchical Clustering Algorithm Based on Fuzzy Graph Connectedness." Fuzzy Sets and Systems 157, no. 13 (July 2006): 1760–1774. doi:10.1016/j.fss.2006.01.001.

[27] Piegat, Andrzej. "Fuzzy Control." Studies in Fuzziness and Soft Computing (2001): 495–607. doi:10.1007/978-3-7908-1824-6_7.

[28] Thong, Nguyen Tho, and Le Hoang Son. "HIFCF: An Effective Hybrid Model between Picture Fuzzy Clustering and Intuitionistic Fuzzy Recommender Systems for Medical Diagnosis." Expert Systems with Applications 42, no. 7 (May 2015): 3682–3701. doi:10.1016/j.eswa.2014.12.042.

[29] Masulli, Francesco, and Andrea Schenone. "A Fuzzy Clustering Based Segmentation System as Support to Diagnosis in Medical Imaging." Artificial Intelligence in Medicine 16, no. 2 (June 1999): 129–147. doi:10.1016/s0933-3657(98)00069-4.

[30] Poczeta, Katarzyna, Łukasz Kubuś, and Alexander Yastrebov. "Multidimensional Medical Data Modeling Based on Fuzzy Cognitive Maps and k-Means Clustering." Procedia Computer Science 176 (2020): 118–127. doi:10.1016/j.procs.2020.08.013.

[31] Wang, Xizhao, Bin Chen, Guoliang Qian, and Feng Ye. "On the Optimization of Fuzzy Decision Trees." Fuzzy Sets and Systems 112, no. 1 (May 2000): 117–125. doi:10.1016/s0165-0114(97)00386-2.

[32] Crockett, Keeley, Zuhair Bandar, and David Mclean. "On the Optimization of T-Norm Parameters within Fuzzy Decision Trees." 2007 IEEE International Fuzzy Systems Conference (June 2007). doi:10.1109/fuzzy.2007.4295348.

[33] Haas, Peter J. "Colored Stochastic Petri Nets." Stochastic Petri Nets (2002): 385–445. doi:10.1007/0-387-21552-2_9.

[34] Budayan, Cenk, Irem Dikmen, and M. Talat Birgonul. "Comparing the Performance of Traditional Cluster Analysis, Self-Organizing Maps and Fuzzy C-Means Method for Strategic Grouping." Expert Systems with Applications 36, no. 9 (November 2009): 11772–11781. doi:10.1016/j.eswa.2009.04.022.

[35] Askari, Salar. "Fuzzy C-Means Clustering Algorithm for Data with Unequal Cluster Sizes and Contaminated with Noise and Outliers: Review and Development." Expert Systems with Applications 165 (March 2021): 113856. doi:10.1016/j.eswa.2020.113856.

[36] Ćirić, Miroslav, Aleksandar Stamenković, Jelena Ignjatović, and Tatjana Petković. "Fuzzy Relation Equations and Reduction of Fuzzy Automata." Journal of Computer and System Sciences 76, no. 7 (November 2010): 609–633. doi:10.1016/j.jcss.2009.10.015.

[37] Dumka, Ankur. "Smart Information Technology for Universal Healthcare." Healthcare Data Analytics and Management (2019): 211–226. doi:10.1016/b978-0-12-815368-0.00008-7.

[38] Iyawa, Gloria Ejehiohen, Collins Oduor Ondiek, and Jude Odiakaosa Osakwe. "mHealth." Smart Medical Data Sensing and IoT Systems Design in Healthcare (2020): 1–21. doi:10.4018/978-1-7998-0261-7.ch001.

[39] N., Ambika. "Methodical IoT-Based Information System in Healthcare." Smart Medical Data Sensing and IoT Systems Design in Healthcare (2020): 155–177. doi:10.4018/978-1-7998-0261-7.ch007.

[40] Badaev, F.I., and T.V. Filippovskaya. "Health Digitalization Alternative: Is There One or Not?" Proceedings of the International Scientific and Practical Conference on Digital Economy (ISCDE 2019) (7-8 November 2019): 150-153. doi:10.2991/iscde-19.2019.28.