Application of the C-Means Fuzzy Clustering Method for the Patient's State Recognition Problems in the Medical Monitoring Systems

Nina Bakumenko[0000-0003-3496-7167], Viktoriia Strilets[0000-0002-2475-1496],

Mykhaylo Ugryumov[0000-0003-0902-2735]

V. N. Karazin Kharkiv National University, Ukraine
n.bakumenko@karazin.ua, miss_victoria@ukr.net,
ugryumov.mykhaylo52@gmail.com

Abstract. Systematization and grouping of information about objects allows improving the quality of decisions. This paper discusses using of pattern recognition methods in medical monitoring systems. The analysis of clustering methods for processing biomedical data was carried out. The problem of stratification of patients using the c-means method, a feature of which is the possibility of designing clusters that intersect, is considered. Particular attention is paid to the choice of the structure and parameters of fuzzifiers to achieve the best clustering accuracy.

Keywords: clustering, c-means fuzzy clustering, fuzzifier, medical data processing.

1 Introduction

In medicine clustering is one of the tools of experimental data and clinical observations analyzing. This mathematical instrument is widely used for diagnostic purposes, for the classification problems solving and the search for new patterns, and for formulating new scientific hypotheses [1]. A Significant advantage of cluster analysis is that it allows doing a breakdown of objects by not only one parameter, but by a whole set of factors. In addition, cluster analysis, unlike most mathematical and statistical methods, does not impose any limitations on the kind of objects under consideration, and allows considering a lot of initial data of arbitrary nature. This is essential for the objects classification problems in medicine.

2 Problem statement

Currently there are many different approaches and specific heuristic algorithms for solving cluster analysis problems (taxonomies, or classifications without a teacher) when it is necessary to find natural groups of similar objects (clusters) according to a

given sample of their vector descriptive descriptions [2]. Like any other method, cluster analysis has certain disadvantages and limitations: in particular, the composition and number of clusters depends on the selection criteria for the partition. When the output data array is reduced to a more compact form, certain distortions may occur and individual features of certain objects may be lost due to the replacement of their characteristics with the generalized values of the cluster parameters. When the objects are classified, the possibility of the absence of any cluster values in this set is often ignored. Solutions found by different algorithms can vary significantly, which requires careful selection of the clustering method. The purpose of this work is to improve the quality of diagnosing diseases by applying methods of fuzzy clustering of data [3].

3 Publications review

The problem of automatic objects classification consists of dividing the whole set of analyzed objects into a relatively small number of homogeneous, in a certain sense, classes. Different methods can be used for this purpose. One of them is based on constructing a hierarchy of classes, which is defined as the sequence of embedded partitions [4].

The algorithms for organizing data of this type proceed from the fact that some kind of objects is characterized by a certain degree of connectivity. It is assumed to have attached groups (clusters of different order). Algorithms, in their turn, are divided into agglomerative (unifying) and divisive (separating). In separating clustering all the initial data sets are considered as one cluster, which splits into two, those in their turn for two more etc., until each of them will consist of a single object.

In agglomerate clustering the hierarchical tree is also formed, but by combining objects into larger clusters of smaller ones. First each object of the initial set is treated as a separate cluster, then two objects are searched, the distance between which is minimal, and are combined into one and so on. This procedure continues until all the objects are assembled into a single cluster. The disadvantages of this approach include the lack of clear recommendations for choosing the number of clusters, the relatively large amount of computations and the impossibility of individual accounting of certain elements when clustering is combined.

Another approach to solving the automatic classification problem is given by probabilistic clustering models [5], such as EM-algorithm and Bayesian models. Methods of the EM-algorithm family assume that there is some cluster mathematical model in the data space and seek to maximize the similarity of this model and the available data. Often, this apparatus uses mathematical statistics.

The EM algorithm [6] is based on the assumption that the studied data set can be modeled using a linear combination of multidimensional distributions. Its goal is to evaluate the distribution parameters that maximize the likelihood function used as a safety feature of the model. In other words, it is assumed that data in each cluster is subject to certain distribution laws. Taking into account this assumption, it is possible to determine the optimal parameters of the distribution law – the mathematical expec-

tation and the variance in which the probability function is maximal. Thus, it is assumed that any object belongs to all clusters, but with different probabilities. Then the task will be to "fit" the set of data distributions, and then to determine the probabilities of belonging to each cluster. Obviously, the object must be attributed to the cluster for which this feature is higher.

The EM-algorithm is simple and easy to implement, not sensitive to isolated objects and quick converges with successful initialization. However, it requires indication of the clusters number k for initialization, which implies the presence of priori knowledge about the data. In addition, if the initialization fails, the convergence of the algorithm may be slow or a poor result can be obtained. Obviously, such algorithms are not applicable to spaces with high dimensionality since in this case it is extremely difficult to assume a mathematical model of data distribution in this space.

Among the fast-acting algorithms using the concept of the masses center the most common algorithms are k-means and the FOREL algorithm. The FOREL algorithm (FORmal ELement) proposed by Zagoruiko and Yolkina [7] has numerous variations, described in detail in [8, 9]. The basis of all these variations is the following basic procedure. Let some point $x_0 \in X$ and parameter R are given. Specify all the sample points $x_i \in X^l$ that fall into the sphere $\rho(x_i, x_0) \leq R$, and the point x_0 is transferred to the center of the selected points gravity. This procedure is repeated until the composition of the selected points, and hence the position of the center, will not cease to change. It is proved that this procedure converges for a finite number of steps. In this case the sphere moves to the place of local points thickening. In the general case the sphere center x_0 is not the subject of the sample, therefore it is called a formal element. The algorithm is very sensitive to the choice of the starting position of the point x_0 for each new cluster. To remove this disadvantage in [10] it is proposed to generate several (about 10…20) clustering centers. Since the starting position of the centers is chosen randomly, these clustering will vary difference. Finally, the clustering that supplies the optimal value to the given quality function is selected.

The k-means algorithm [11] builds k located at possibly large distances from each other. The main problem type solved by the k-means algorithm is the presence of assumptions (hypotheses) as to the clusters number, while they must be different as far as possible. The choice of the number k can be based on the results of previous studies, theoretical considerations or intuitions.

The general idea of the algorithm: the given fixed number k of observation clusters is compared to the clusters so that the averages in the clusters (for all variables) differ as much as possible from each other. The algorithm disadvantage is that the algorithm is too sensitive to emissions, which can distort the average; slow work at large databases; it is needed to set the clusters number.

The considered clustering algorithms provide the separation of objects into disjoint sets, while medical data have the property of overlaying one another. This disadvantage is deprived of fuzzy clustering algorithms, for example, the c-means algorithm, the application of which for medical and biological data is considered in this article.

4 Application of the k-means fuzzy clustering method for the patient's state recognition problems in the medical monitoring systems

As an example of a complex system the medical-biological system is considered, this includes the following elements: a physician, patients and a subsystem diagnosing patients.

The system model of the diagnostic process of the medical and biological system is presented in Fig. 1. Designations in the picture are: S is an inputs adder, Controller is a control body, which is an attending medical doctor, who develops patient care scenarios, u are control variables, f is external influences (perturbation), Object of control - the object of control (patients), Z is variable of patients states, W is criteria for the quality of patients states. The beginning and end of the stages of the patient's life-cycle will be determined as the set of end-states of patients. The number of states accepted for consideration is set by an expert in the subject area based on the results of the clustering analysis. It is accepted hypothesis of local equilibrium as working, according to which the patient's state is uniquely determined by the fundamental system of its variables. We will assume that a critical condition on a set of end states is a resistant state in which the patient loses control in the process of treatment due to the progressive development of defects in functional parts. The objective problem is that there is no structural decision rule for the transition to a resistant state for the patient under consideration.

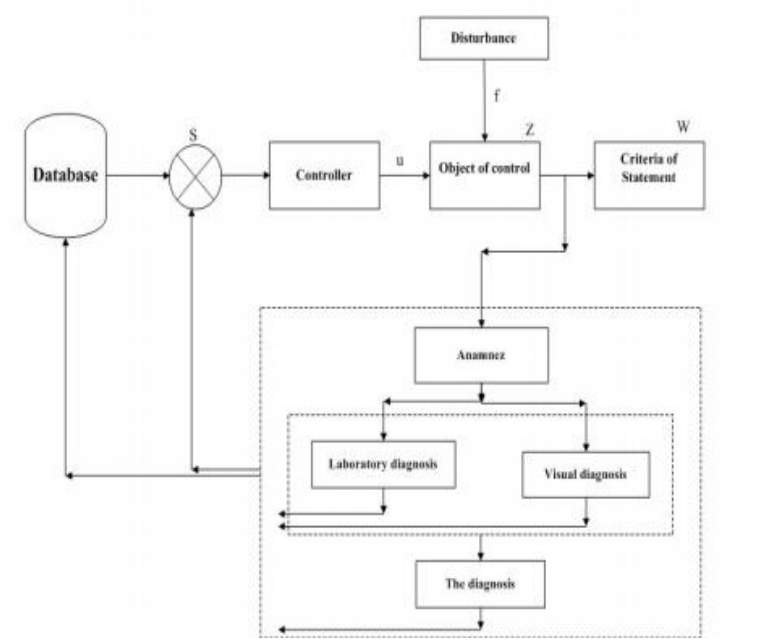

Fig. 1. System model of the diagnostic process of the medical and biological system

A characteristic feature of class recognition in medical data is the presence of intersecting classes, which leads to the need to use fuzzy clustering methods since it allows evaluating the membership grade of a class object.

For clustering it is suggested to use an adaptive algorithm of fuzzy data clustering for a batch or sequential processing of information [12-13].

The initial information is a sample of observations generated from the N n-dimensional vector of factors $X = \{x(1), x(2), \dots, x(N)\}, x(k) \in X, k = 1,2, \dots, N$. The result of the method is to split the initial data array into m classes with a certain level $w_j(k)$ of belonging to the *k*-th vector of factors of the *j*-th cluster. The objective function to be minimized is:

$$E(\mathrm{w}(k), \mathrm{c}) = \sum_{k=1}^{N} \sum_{j=1}^{m} w_j^{\beta}(k) d^2(x(k), c_j) \to min \quad (1)$$

with limitations:

$$\sum_{j=1}^{m} w_j(k) = 1, k = 1, \dots, n, 0 < \sum_{k=1}^{N} w_j(k) < N, j = 1, \dots, m. \quad (2)$$

Here $w_j(k) \in [0,1]$ is the membership level of the vector to the j-th cluster; c_j is the centroid of the j-th cluster; $d^2(x(k), c_j)$ is the distance between $x(k)$ and c_j in the accepted metric, β is an integral parameter called "fuzzifier" (in the case of use as $d^2(x(k), c_j)$ an Euclidean distance, is taken as 2).

The work of the algorithm begins with the definition of the initial random matrix of fuzzy partition W_0. According to its values, the initial set of prototype centers c_j^0 is calculated according to the formula

$$c_j = \frac{\sum_{k=1}^{N} w_j^{\beta}(k) x(k)}{\sum_{k=1}^{N} w_j^{\beta}(k)} \quad (3)$$

Based on the calculated prototype centers c_j^0, the matrix W_1 is calculated in accordance with the formula:

$$w_j = \frac{(d^2(x(k), c_j))^{\frac{1}{1-\beta}}}{\sum_{l=1}^{m} (d^2(x(k), c_l))^{\frac{1}{1-\beta}}} \quad (4)$$

Another method of calculating the coefficients of the matrix (denoted by its μ-matrix) was also used in the work:

$$\mu_j = \frac{1}{1 + \left[\frac{d^2(x(k), c_j(k))}{r^2}\right]^{\frac{1}{1-\beta}}}, r = \min_{i,j} \|c_i - c_j\| \quad (5)$$

After that a batch mode $c_j^1, W_2, \dots, W_t, c_j^t, W_{t+1}, \dots so\ on$ calculates, until the difference between the current and the next values of the matrix W is not be less than the

given threshold of accuracy. Thus, all available sample data is processed multiple times.

As a result of the algorithm, we obtain a matrix of fuzzy clustering, in which patients will be divided into clusters (diagnoses). The shape of the clusters can vary from the hypersphere to the hyperellipsoid, depending on the form of the input data, that is from the choice of the distance between $x(k)$ and c_j:

$$d^2\big(x(k), c_j\big) = \sqrt{\big(x(k) - c_j\big)^T A_j (x(k) - c_j)} \tag{6}$$

where A_j is a matrix that can be defined as the inverse fuzzy covariance matrix of each cluster.

If, as a matrix A_j, we take a unit matrix, then the result is an Euclidean distance $d^2\big(x(k), c_j\big) = \sqrt{\big(x(k) - c_j\big)^T (x(k) - c_j)}$ and the shape of the clusters will be rounded (hypersphere).

To give clusters the form of hyperellipsoids as a matrix A_j, one can use a symmetric positive definite matrix, that is a matrix in which all eigenvalues are real and positive and $A_j = F_j^{-1}$, where

$$F_j = \frac{\sum_{j=1}^{m} w_j^{\beta}(k)(x(k)-c_j)(x(k)-c_j)^T}{\sum_{j=1}^{m} w_j^{\beta}(k)}. \tag{7}$$

As a result of the clustering algorithm, we obtain the division of our data into homogeneous clusters, which may take the form of arbitrarily-oriented hyperellipsoid in the space and are able to intersect in the space of signs. Also, as a result of the algorithm's operation, the degree of belonging of each object to each of the clusters $w_j(k)$ will be known.

The training sample consisted of 180 objects, which are data from laboratory studies of patients aged from 46 to 78 years old, where 50 belonged to the class "healthy" - benign tumor, 45 people belonged to the “nonmetastable class ", 52 people to the class" metastasize "and 33 people to the class "hormone-resistant "[14]. To describe the objects in use 24 controlled variable states (attributes) were used. Their values are valid and take different meanings. The attributes are presented in Table 1. A set of characteristic is required for the identification of 4 groups of patients.

The initial information was a dataset of 24-attributed set describing the stage of the Prostate cancer stage. The task of recognizing the disease group was to automatically assign set of 24 attributes, describing the status of a patient, to one of the four classes mentioned above.

The data was transformed to a normalized form to improve the algorithm's performance. As data has been normalized, the Mahalanobis distance instead of the Euclidean distance can be used (suppose that there are no correlations between variables).

Three datasets were selected for conducting experiments. The first dataset contains the Prostate cancer data (the main sample), the second one is Fischer's irises and the third is the breast cancer data. The first experiment was performed on a Prostate can-

cer data containing 25 parameters, 4 groups of patients and 180 samples. In addition, it was decided to choose different methods for calculating fuzzy clustering matrix (W-matrix and μ-matrix). To verify the accuracy of clustering with a fuzzy c-means method the obtained clustering is compared in percentage to the original dataset by the number of samples of the particular disease stage.

Table 1. Attributes of objects and their description

Atribute	Description	Details
Age	Age	Years
VASScale	VAS scale	0-10 points
UrinationCount	Count of urination	times
NumberOfUrgency	Number of imperative urinations	times
NightUrination	Night urination	times
Strangury	Stranguria	yes/no
OZM	Acute retention of urine	yes/no
HZM	Chronic urinary retention	yes/no
ResidualUrine	The amount of residual urine	ml.
LvSided	Uropathy double sided	yes/no
ProstateVolume	Volume of prostate	cm3/mm
PSA	prostate-specific antigen (PSA)	ng/ml
Hemoglobin	Hemoglobin	gram / ltr
ESR	Erythrocyte Sedimentation Rate	mm per hour
Leukocytes	leucocytes	10^12/ltr
Lymphocytes	Lymphocytes	%
SpecificGravity	Specific weight	Density
Eritrotsyty	erythrocytes	Items in sight
LeukocytesUrine	Leukocytes in urine	Items in sight
Lymphadenopathy	Lymphadenopathy	yes/no
Bones	Metastasis in bones	yes/no
Vertebrates	Metastasis in spine	yes/no
G	Stage	1-2-3
Glisson	Glisson scale	1-10

In the first case with the W matrix, 43 patients of «Without Metastases» group of 45 were included in the first cluster (95% accuracy), the second - 31 patients from «With Metastases» group from 52 with 52 (60% accuracy), in the third - 49 patients in the «Healthy» group from 50 (98% accuracy) and in the fourth - 13 patients in the group "4th" from 33 (39% accuracy). This clustering was performed with an accuracy of 72.3%. In the second case with the μ matrix, 34 patients from the Without Metastases group from 45 (75% accuracy), the second - 14 patients in the group "4th" from 33 (42% accuracy), in the third - 25 patients with the "With Metastases" group from 52 were in the first cluster (49% accuracy) and in the fourth - 50 patients of the

"Healthy" group from 50 (100% accuracy). This clustering was carried out with an accuracy of 66.5%.

The fuzzifier was chosen manually to select the best result, therefore, on the basis of experiments with the fuzzifier, the best result was found in the matrix W in the values of 1.1, in the case of the matrix μ fuzzifier β= - 2. The choice of the fuzzifier depends on the sample data and the matrix. Clusterization results for diagnosing prostate cancer are shown in Figure 2.

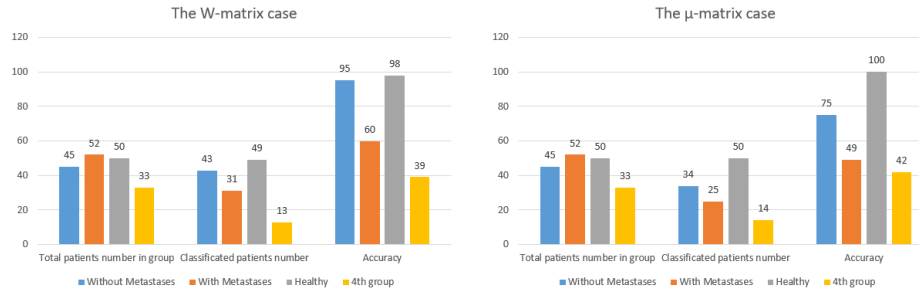

Fig. 2. Clusterization results for diagnosing prostate cancer

From the experiments conducted, we can conclude that clustering with a W matrix gives better results. Therefore, it was decided to finally accept the W matrix for further experiments.

The second experiment was conducted on a Irises dataset, which contains 4 parameters, 3 groups of the iris type and 150 samples. In the second experiment different fuzzifier values were used. In the first case with fuzzifier 1.1, 36 flowers from 50 of the group "virginica" were in included the first cluster (72% accuracy), in the second - 50 flowers from 50 of setosa group (100% accuracy) and in the third - 47 flowers from 50 of the group "versicolor" (94% accuracy). This clustering was carried out with an accuracy of hit 87%. In the second case with the fuzzifier 2, 50 flowers of the setosa group from 50 were included in the first cluster (100% accuracy), the second - 43 flowers from 50 of the group "versicolor" (86% accuracy) and in the third - 45 flowers from 50 of the group "virginica" (90% accuracy). This clustering was performed with an accuracy of 92%. The experiment showed that using the fuzzifier 2 for the sample with irises is more accurate. Clusterization results for the Irises dataset are shown in Figure 3.

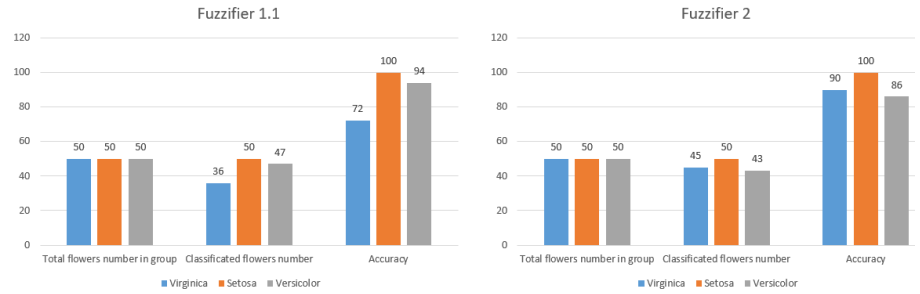

Fig. 3. Clusterization results for the Iris dataset

Then let us consider the latest experiment with breast cancer data. The second experiment was conducted on a sample of cancer patients, containing 30 parameters, 2 groups of tumor type and 569 samples. In the first case with the fuzzifier 2, 350 patients from 357 of the "benign" group (98% accuracy) were included in the first cluster, the other 142 patients from 212 of the "malignant" group (66% accuracy) were included in the second cluster. This clustering was performed at an accuracy of 82%. In the second case with the fuzzifier 3, 324 patients from 357 of the "benign" group (90% accuracy) and the other 171 patients from 212 of the "malignant" group (80% accuracy) were in the second cluster. This clustering was performed with an accuracy of 85%. The experiment has shown that using the fuzzifier 3 for a cancer patients dataset is more accurate. Clusterization results for the breast cancer data are shown in Figure 4.

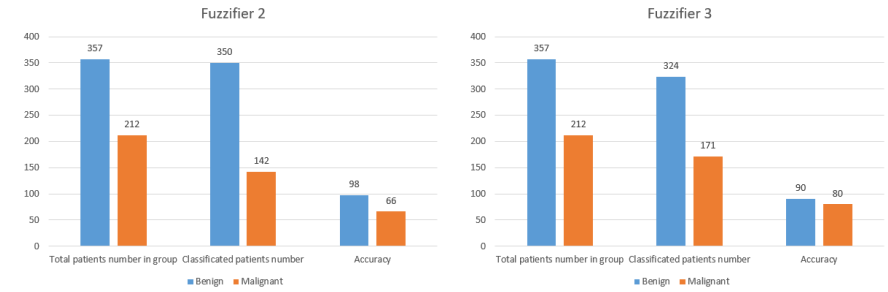

Fig. 4. Clusterization results for the breast cancer data

5 Conclusion

The problem of stratification of patients in medical monitoring systems using of the of c-means fuzzy clustering method has been considered. The two models of membership grade have been investigated. It is shown that the W-matrix model is more effective in terms of image recognition. From the experiments carried out, it has been concluded that using different values of the distance fuzzifier allows improving the quality of clustering, as well as obtaining a more accurate picture of the conducted cluster-

ing. As it can be seen from the three experiments, the quality of clustering depends not only on the fuzzifier, but also on the set of data and the number of clusters. This is the reason for a more detailed study of the method, its modifications, and the choice of the fuzzifier.

6 References

1. Kochetov, A. G., Lyang, O.V., Masenko, V. P.: Metody statisticheskoy obrabotki meditsin-skih dannyih. RKNPK, Moskva (2012)
2. Kovalenko, O. S.: Obzor problem i perspektiv analiza dannyih. Informatika, vychislitel'naja tehnika I inzhenernoe obrazovanie (2), 15-31 (2010)
3. Bezdek, J.: Fuzzy models and algorithms for pattern recognition and image processing. Springer, New York (1999)
4. Zhambyu, M.: Ierarhicheskiy klaster-analiz i sootvetstviya. Finansyi i statistika, Moskva (1988)
5. Nicholas, O. A., Fox, E. A.: Recent Developments in Document Clustering (2007)https://www.researchgate.net/publication/228350017_Recent_Developments_in_Document_Clustering, last accessed 2019/03/31
6. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM Algorithm. Journal of the Royal Statistical Society. Series B (Methodological)(39), 1–38 (1977)
7. Zagoruyko, N. G., Yolkina, V. N. Lbov, G.: Algoritmyi obnaruzheniya empiricheskih zako-nomernostey. Nauka, Novosibirsk (1985).
8. Zagoruyko, N. G.: Prikladnyie metodyi analiza dannyih i znaniy. IM SO RAN, Novosibirsk (1999)
9. Malyutov, M. B.: Statisticheskie metodyi dlya EVM. Nauka, Moskva (1986)
10. Lance, G.N., Willams W.T.: A general theory of classificatory sorting strategies. 1. Hierarchical systems. Computer Journal (9), pp. 373-380 (1967) https://doi.org/10.1093/comjnl/9.4.373, last accessed 2019/03/31.
11. Ayvazyan, S. A., Bezhaeva, Z. I., Staroverov, O. V.: Klassifikatsiya mnogomernyih nablyudeniy. Statistika, Moskva (1974)
12. Wass, J. A.: How Statistical Software Can Be Assessed. Scientific Computing & Automation (October), 14–24 (1996)
13. Darmoni, S. J., Massari, P., Droy, J. M.: SETH: an expert system for the management on acute drug poisoning in adults. Computer Methods and Programs in Biomedicine (Jun), 171–176 (1994)
14. Antonyan, I. M., Ugryumov, M. L., Zelenskiy A. I.: Zabolevaniya predstatelnoy zhelezyi. Diagnosticheskaya model i metod klassifikatsii sostoyaniy. Urologiya (1), pp. 31–38 (2015)