Fuzzy Clustering Algorithms with Unknown Number of Clusters

Yuriy Zaychenko[1], Aydin Həsənov[2]
[1, 2]Institute for Applied System Analysis KPI, Kyiv, Ukraine
[1]zaychenkoyuri@ukr.net, [2]ayding44@rambler.ru

***Abstract*– In the paper the problem of number of clusters in fuzzy clustering algorithms is considered. A novel method for its determination is suggested. The method is based on modification of differential grouping algorithm for determination of centers of fuzzy clusters. The experimental investigations of the suggested algorithm were carried out in the problem UNO counties clustering problem on sustainable development indices.**

Keywords– fuzzy clustering, fuzzy neural network, index GINI

I. INTRODUCTION

In recent years the cluster analysis is widely used in the intellectual analysis of data (Data Mining), as one of the principal methods. All of them can be subdivided on hierarchical and not hierarchical [1].

In not hierarchical algorithms their work and conditions of stop need to be regulated in advance often with large number of parameters that is sometimes difficult, especially at the initial stage of investigation. But in such algorithms big flexibility in a variation of a clustering is reached and usually the number of clusters is defined On the other hand, when objects are characterized by a large number of features (parameters), a task of grouping features is important. Initial information contains in a square matrix of features interconnections, in particular, in a correlation matrix. Basis of the successful solution of a grouping task is the informal hypothesis of a small number of the hidden factors which define structure of an interconnection between features.

In hierarchical algorithms one actually refuses to define a number of clusters, building a full tree of the enclosed clusters (so-called dendrogram). The number of clusters is defined from the assumptions, in principle, which aren't relating to work of algorithms, for exampleon dynamics of change of a threshold of splitting (merge) of clusters.

Difficulties of such algorithms are well studied: choice of measures of proximity of clusters, problem of inversions of indexation in the dendrogramms, inflexibility of hierarchical classifications which is sometimes undesirable. Nevertheless, representation of a clustering in the form of a dendrogramm allows to gain the most complete display of structure of clusters.

Hierarchical algorithms are connected with dendrogramms construction and divided on:

1. agglomerative, characterized by consecutive merge of initial elements and the corresponding reduction of number of clusters (creation of clusters from below to top);

2. divisional (divided) in which the number of clusters increases, starting with one cluster therefore the sequence of the splitting groups is constructed (creation of clusters from top to down).

Clustering algorithms are also may be divided on non-fuzzy and fuzzy ones.

In non- fuzzy algorithms the objects belong only to one class, while in fuzzy objects may belong to several classes simultaneously with different degree of membership based on membership functions. which is more general case and more proper in many real cases.

To the date there are a lot of fuzzy clustering algorithms were developed, most known of them are Fuzzy C-means[2] and Gustavsson- Kesel algorithms [3, 4]. Last years were suggested new class of fuzzy algorithms so-called possibilistic algorithms [5].

The main drawbacks of all these algorithms is that they may work only when the number of clusters are known a–priori. But in many cases the number of clusters isn't known for decision-maker, it depends on objects grouping in multi dimensional space and should be determined in the process of algorithm run.

There are several approaches for determining number of clusters. One of the mostly used is so-called criterion Hi-Beni [5].

$$\alpha = d_{av} \Big/ D_{av},$$

where d_{av} is the average intra-cluster distance, D_{av} - average inter-cluster distance. This indicator should be minimized. It's main drawback is that it may have several minima and reaches the global minimum in trivial case when the number of clusters K become equal to number of objects n.

Therefore is the main goal of this paper is to developed new approaches and methods for detrmining the number of clusters which more adequately (proper) fit to real grouping of objects in n-dimensional space.

A. Algorithm of clusters number determination

Assume there are given n objects in space $x_i = \|x_{ij}\| i = i, n,\ j = 1, m.$, where x_{ij} is j-th feature of i-th

DOI: 10.25045/NCSoftEng.2017.30

object. They should be split into *K* clusters where K isn't known a priori by minimizing criterion

$$E = \sum_k \sum_j w_{kj}^{\beta} d^2(x_k, c_j) \quad (1)$$

Where the distance between the center *c* and vectors x is defined using scaling matrix A by a formula:

$$d(x,c) = \|x-c\| = \sqrt{(x-c)^T A(x-c)} \quad (2)$$

As scaling usually the positive-definite matrix is used, that is a matrix, at which all own numbers are real and positive.

For the solution of this problem Gustavsson- Kessel algorithm is used. But prior its application number clusters should be determined. The suggested algorithm is based on method of clusters centers grouping- differential grouping [1].

The algorithm of differential grouping is a modification of the previous algorithm, in which vectors x_j are considered as the potential centers. Peak function $D(x_i)$ in this case takes the form [1]:

$$D(x_i) = \sum_{j=1}^{N} \exp\left\{ - \frac{\|x_i - x_j\|^{2b}}{\left(r_a / 2\right)^2} \right\}, \quad (3)$$

where value of coefficient r_a defines the sphere of the neighborhood. On value $D(x_i)$ considerably influence only vectors x_j, which are inside this sphere.

At the big density of points near x_i function value $D(x_i)$ is large. After calculation of values of peak function for each point x_i, the vector x is found, for which density measure $D(x)$ will appear to be the greatest. This point becomes the first center c_1.

Choice of the following center c_2 is performed after an exception of the previous center and all points which lie in its vicinity.

As well as in the previous case peak function is redefined so

$$D_{new}(x_i) = D(x_i) - D(c_1) \exp\left\{ - \frac{\|x_i - c_1\|^{2b}}{\left(r_b / 2\right)^2} \right\}, \quad (4)$$

At new definition of function *D* coefficients r_b designate new values of a constant which sets the sphere of the neighborhood of the following center. Usually a condition $r_b \geq r_a$ is used.

After modification of value of peak function a search of a new point x, for which $D_{new}(x_i) \to \max$ is performed It becomes the new center.

Process of finding of the next center is resumed after the exception of all already selected points.The algorithm runs till the following stop criterion is fulfilled:

$$\max\left[D(X_i) - D(X_{i)} \right] \leq \varepsilon, \quad (5)$$

where $\varepsilon = (0{,}1 - 0{,}2) \max D(X_i)$ or another value given by decision –maker.

Such approach enable to determine number of clusters which properly maps the real groups of objects.

The experimental investigations of this approach were carried at the problem clustering the countries of the United Nations sustainable development indicators. For this, the data of the World Data Center in Ukraine (WDC) were used. As sustainable development indicators the following indices were taken:

- Index GINI- GINI;
- Ihd- index of health status;
- Iql - standard of living index;
- Isd- index of sustainable development.

As algorithm of initial centers placement the suggested modified algorithm of differential grouping was applied. The real number of clusters were d etermined and clustering was performed. For comparison Hi- Beni index was used Clustering was carried out for a different number of clusters K = 3,4,5.

The results of the experiments are presented and discussed in report.

CONCLUSION

1. Algorithm for determining of clusters in the problem of fuzzy clustering was suggested which adequately maps real groups of objects in multidimensional space. The suggested algorithm may be used in cases where number of clusters isn't known a priori which is more real case in practice.

2. The experimental investigations of the suggested algorithm and comparison with algorithm using HI-Beni criterion were carried out.

REFERENCES

[1]. B. Durant, G.Smith. Cluster analysis. – M. Ctatistica. –1987. – 289 p. (rus)

[2]. D.C. Park, I. Dagher. Gradient based fuzzy C–means (GBFCM) algorithm. In: Proceedings of IEEE International Conference on Neural Networks, 1984.–pp. 1626–1631.

DOI: 10.25045/NCSoftEng.2017.30

[3]. R. R. Yager, D. P. Filev. Approximate clustering via the mountain method IEEE Trans. on Syst., Man and Cybern. – 1994. – 24. – pp. 1279–1284.

[4]. S.Osovsky. Neural networks for information processing, transl. from pol. – M.: Publ. house Finance and Statistics.– 2002.–344 p. (rus)

[5]. M. Zgurovsky, Yu.P. Zaychenko. The Fundamentals of Computational Intelligence: System Approach. Springer.–2016. – 375 p.

DOI: 10.25045/NCSoftEng.2017.30