Natalia Kondruk
Candidate of Technical Sciences, Associate Professor of the Department of Cybernetics and Applied Mathematics
Uzhhorod National University

SEGMENTATION OF DATA SETS BY DIFFERENT TYPES OF CLUSTERS

Clustering is a powerful tool in the field of Data mining, when there is no a priori information about the relationships between data. Currently, many cluster analysis algorithms are successfully used in various application areas, where there is a need to divide similar in certain features objects into subsets [1-3]. A crisp split into clusters is possible only with very different features of clustering objects. Therefore, fuzzy methods are increasingly used to solve real problems, in which the division of objects is carried out to determine the degree of belonging of objects to clusters.

All existing methods can be classified according to the similarity measures they use [3, 4]. On the other hand, it determines the different geometric shape of the formed clusters and allows obtaining qualitatively different applied interpretations of the obtained homogeneous segments of data sets. Therefore, it is the specifics of applied problems that make it impossible to automatically transfer methods to another application area without the risk of deliberately obtaining a bad solution. Therefore, it is advisable to develop an information system that would have a fairly wide range of tools for grouping objects by different similarity measures. This makes it possible to effectively solve a lot of applied problems in different subject areas. The main works in which the technology is presented, which allows to solve this problem are presented in [4-7].

The focus of the system is a single-level clustering method based on fuzzy binary relations described in [5]. The flexibility of this algorithm allows you to form different geometric shapes of clusters of datasets by simply changing the appearance of the degree of similarity of objects. In this case, the similarity of the objects O_i and O_j by some criterion is characterized by a fuzzy binary relation R on the set of vector features with the membership function μ_R. The closer the value μ_R is to 1, the more similar the objects will be to this criterion. Thus, in [4-7], three types of similarity measures of objects are proposed: length-based, angular and distance.

To form elliptically similar clusters, it is expedient to use the "distance" similarity measure, which is described by a fuzzy binary relation R^V [4]. The fuzzy binary relation R^K [7] characterizes the angle of deviation between the feature vectors. Its use makes it possible to carry out clustering with conical clusters. The length-based similarity measure R^D allows splitting the feature vectors of objects into clusters by concentric spheres [5].

Conical clustering can be effectively used to solve multi-criteria linear programming problems with a large-scale criterion space [7], which arise, in particular, in mathematical modeling of balanced nutrition problems. One of the steps in solving such problems is to cluster their criteria space. In this case, the relationships between the criteria are determined by their angular similarity R^K. Clustering by elliptical

clusters is most common in many application problems, as the similarity of objects is based on a "distance" similarity measure. Also in [6] two synthetic sets of two-dimensional data of Gaussian type are generated and efficiency of application of a clustering method based on fuzzy binary relations at various indices of an estimation of quality of partition is investigated. Clustering by concentric clusters (clusters in the form of concentric spheres) [5] made it possible to group objects by length-based similarity of their feature vectors and to obtain a qualitatively new applied meaningful interpretation of the formed homogeneous groups in practice. In addition, this approach allows for both crisp and fuzzy data clustering.

In perspective researches the combined index of an estimation of clustering quality which is adapted to use of various similarities measures of a fuzzy binary relations method will be created; development of a software system that will ensure the segmentation of data sets into different geometric shapes clusters without prior determination of the clustering threshold.

References:

1. T. Sajana, C. S. Rani, K. V Narayana, A survey on clustering techniques for big data mining. Indian journal of Science and Technology, 9(3), 2016, pp. 1-12. doi: 10.17485 / ijst / 2016 / v9i3 / 75971

2. A. Amelio, A. Tagarelli, Data mining: clustering. Encyclopedia of Bioinformatics and Computational Biology, 2018, pp. 437-48. doi: 10.1016 / B978-0-12-809633-8.20489-5

3. K. Chitra, D. Maheswari, A comparative study of various clustering algorithms in data mining. International Journal of Computer Science and Mobile Computing, 6(8), 2017, pp. 109-115.

4. N. Kondruk, Clustering method based on fuzzy binary relation, Eastern-European Journal of Enterprise Technologies, 2017, pp. 10–16. doi:10.15587/1729–4061.2017.94961

5. N. Kondruk, Use of length-based similarity measure in clustering problems, Radio Electronics, Computer Science, Control, 2018, pp. 98–105. doi:10.15588/1607-3274-2018-3-11.

6. N. E. Kondruk, A comparative study of cluster validity indices, Radio Electronics. Computer Science. Control, 4, 2019, pp. 59 – 67. doi: 10.15588/1607-3274-2019-4-6.

7. N. E. Kondruk, M. M. Malyar, Structuring of the ctriterional space by an angle similarity measure, Scientific Bulletin of Uzhhorod University. Series of Mathematics and Informatics, 2020, pp. 85 – 91. doi: 10.24144/2616-7700.2020.1(36).85-91