# Information Technology of Evaluation and Improvement the Quality of Cluster Analysis

Marina Sidorova

*Abstract* - **In this paper the information technology of evaluation and improvement the quality of cluster analysis has been proposed. It allows to support the decision making in choosing the best partition in the face of uncertainty of the cluster analysis.**

*Keywords* – **cluster analysis, quality, data mining, information technology.**

## I. INTRODUCTION

Cluster analysis is an important branch of data mining, which is used to identify groups, hierarchical structures and patterns in the data set. The objective of cluster analysis is to part a data set into groups (clusters) so that the samples within the same cluster are more similar to each other than samples from different clusters.

At the moment, there are many different approaches and methods to solve this problem. However, the result of clustering is highly dependent on the choice of features, measures of proximity, ways of formalizing the notion of similarity between objects and clusters. Clustering schemes, obtained by different methods or parameter values may be very different or irrelevant to the objectively existing groups. Therefore, one of the most important issues of cluster analysis is to evaluate the results and find the partition that best fits the data structure.

Another promising direction of research in this area is the development of collective methods of cluster analysis (construction an ensemble of algorithms) that allows to receive the most consistent and stable solutions [1].

## II. BASIC MATERIAL

In this paper we propose information technology for evaluating and improving the quality of clustering, which combines the approaches described above and consists of the following steps:

1. Perform the data clustering by different methods or with different parameter values.
2. Assess the validity of the obtained results using the quality functionals [2].
3. Apply methods of decision theory, using the estimates, obtained in step 2. It allows to take into account the different quality criteria at the same time, which provides more accurate assessment of the results [3].
4. If among the clustering schemes there is one that better than others fits the data structure, it can be considered as the solution of the problem. However, most often, in assessing the quality of results, several methods, showing, in general, different partitions can be identified. In this case, we propose to combine their results by constructing a collective decision.

The proposed technology has become a part of the program «Medisa», developed by the authors, and has found practical implementation in the data analysis of medical and hydrochemical monitorings.

The core of the program consists of the procedures of cluster analysis based on hierarchical methods (single link, complete link, average distance, Ward's method), fast hierarchical methods, K-means (Ball-Hall and Mack-Keen variants), graph clustering method. Three types of metrics are proposed: Euclidean, Manhattan, Chebyshev.

The results of clustering algorithms are evaluated by quality functionals such as the sum of variances in classes for all parameters, the sum of squared distances to the centers of classes, the sum distances in classes, the average ratio medium distances in and out classes. Decision-making methods such as multiple analysis, Bord's and pluralitar procedures were proposed to determine the method that gives the best partitioning.

## III. CONCLUSION

Thus, the information technology of decision making support in choosing the best partition in the face of uncertainty of the cluster analysis, that allows to take into account the different quality criteria at the same time and can combine the best results by constructing a collective decision, has been proposed.

## REFERENCES

[1] А.С. Бирюков, В.В. Рязанов, А.С. Шмаков «Решение задач кластерного анализа коллективами алгоритмов», *Журнал вычислительной математики и математической физики*. – 2008. – Т. 48, N 1. – С. 176– 192.

[2] Мандель И.Д. «Кластерный анализ» – М., 1988. – 176с.

[3] О.П. Приставка, М.Г. Сидорова «Підтримка прийняття рішень в задачах кластерного аналізу» *Актуальні проблеми автоматизації та інформаційних технологій* : зб. наук. праць. – 2011. – Т.15. – С.117–125.

Marina Sidorova - Dniepropetrovsk National University named after Oles Honchar, Gagarin Avenue 72, Dniepropetrovsk, 49050, UKRAINE,
E-mail: Sidorova.M.G@gmail.com
Research supervisor – prof. Baybuz O.G.