

Р. В. КОРОТКЕВИЧ

Научный руководитель А. И. ЯКИМОВ, канд. техн. наук, доц.

Консультант Е. Н. МЕЛЬНИКОВА

ГУ ВПО «Белорусско-Российский университет»

The issue of analyzing large amounts of multidimensional data is encountered in many fields of science and technology. In recent years there has been increased attention to this problem, as there is a trend towards using considerable volumes and variety of data, as well as replacing routine human labor by technological advances and innovations. Clustering is one of the possible ways to solve this problem.

Cluster analysis or clustering is a task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters).

Clustering is the main task of exploratory data mining and a common technique for statistical data analysis which is used in many fields, including image analysis, information retrieval, bioinformatics, machine learning, pattern recognition, data compression and computer graphics.

The notion of a ‘cluster’ cannot be precisely defined, which is one of the reasons why there are so many clustering algorithms. There is a common part between all variants of this notion: a group of data objects. However, different researchers employ different cluster models, and for each of these cluster models different algorithms can be given. The notion of a cluster, as found by different algorithms, varies significantly in its properties.

A ‘clustering’ is essentially a set of such clusters, usually containing all objects in the data set. Additionally, it may specify the relationship of the clusters to each other. There are two types of clustering:

- hard clustering: each object belongs or does not belong to a cluster;
- soft clustering (or fuzzy clustering): each object belongs to each cluster to a certain degree.

K-means clustering method is a method of vector quantization, originally used in signal processing, which is a prevailing method applied for cluster analysis in data mining. K-means clustering algorithm aims to partition given number of observations into a certain number of clusters in which each observation belongs to the cluster with the nearest mean, which serves as a prototype of the cluster.

