CLUSTERING OF HOUSEHOLD GOODS BASED ON DEMAND TIME SERIES INDICATORS

Bezchastna M. V., Kostiuk O. V., Heliarovska O. A.
National Technical University «Kharkiv polytechnic institute», Kharkiv

With the development of technology and the availability of large datasets, one of the important tasks of business is to conduct detailed analysis of consumer behavior and preferences in order to understand the market better and optimize marketing strategies. The need to group household goods based on indicators of time series of demand requires the use of cluster analysis methods of time series, which allows to identify regularities and trends of demand for various consumer goods over time for further adaptation of product offers, prices and promotions to specific groups of consumers.

One of the main problems of clustering objects represented by time series is the choice of appropriate distance metrics and clustering algorithms [1]. There are three approaches to calculate distance measures between object indicators: by input data, by descriptive features, and by the results of model fitting. The choice of metric depends on the nature of the data and the chosen clustering method. Methods for clustering time series are modified methods of statistical data clustering, which can be divided into the following classes: hierarchical and flat, clear and fuzzy. Qualitative assessments such as sum of squared errors, silhouette coefficient, Calinski-Harabasz index, Davies-Bouldin index play an important role in determining the effectiveness of the obtained clustering [2].

The purpose of this work is to divide a set of household goods, each of which is represented by a time series of demand for a certain period of time, into groups. The paper proposes an algorithm that consistently uses the methods of analysis, filtering, and subsequent clustering of time series of household goods sales number based on two approaches: using the k-means method, as well as the c-means method, for different values of clusters' number. Approaches to justifying the optimal number of clusters are proposed, the used methods are compared, and their advantages and disadvantages are indicated. The result of the work is a software product containing the implementation of the proposed algorithm, which can be used to solve similar problems of dividing objects represented by time series into the optimal number of groups based on similar characteristics.

References (translated):

1. Guojun Gan, Chaoqun Ma, Jianhong Wu. Theory, Algorithms, and Applications // Data Clustering, 2007.
2. Ian H. Witten, Eibe Frank, Mark A. Hall. Data Mining // Practical Machine Learning Tools and Techniques, 2012.