Kirichenko L., Kobziev V., Fedorenko Y.

DATA MINING METHODS FOR DETECTION OF COLLECTIVE ANOMALIES IN TIME SERIES

Data Mining [1] is a set of methods for identifying previously unknown, non-trivial, practically useful and accessible interpretations of knowledge needed for decision-making in various areas of human activity. One of the tasks of intellectual analysis is to detect anomalies - to identify rare and unusual elements, events or observations that cause suspicion, because they differ significantly from most data.

A time series is a time-ordered sequence of values of some process (for example, the value of a sensor). The need to detect unusual observations (emissions, or anomalies) in time series often arises in situations such as monitoring the condition of equipment, number of participants in certain events, accounting for patient health indicators, etc. [2]. There are three main types of anomalies in time series: point, contextual, and collective anomalies. In this paper, we will review collective anomalies, as they are the most common for time series. Collective anomalies occur when a sequence of related instances of data (e.g., a time series section) is anomalous with respect to an entire data set. A single instance of data in such a sequence may be random deviation, but the co-occurrence of such instances is a collective anomaly [3].

The basic idea of collective anomalies is that such grouped points cannot be anomalies alone. There are many methods for identifying such anomalies, including clustering methods, which allow you to select set of anomalous values as a separate cluster [4].

For a large number of observations, it would be rational to use the method of k-means. This method is one of the simplest and most popular clustering methods for separate a set of elements of a vector space into a predetermined number of clusters k for a certain number of iterations. The general concept of this method is: at each iteration recalculate the center of mass for each cluster obtained in the previous step, then the vectors are divided into clusters again according to which of the new centers was closer in selected metric. The algorithm ends when there is no change intra-cluster distance on any iterations. This occurs for a finite number of iterations, since the number of possible partitions of the finite set is finite, and at each step the total quadratic deviation decreases, so the loop is impossible.

The paper considers the approach to the detection of collective anomalies in time series, based on the use of clustering methods, in particular the method of k-means, as well as the effectiveness of their application.

References

1. Han, Jiawei. Data mining: concepts and techniques / Jiawei Han, Micheline Kamber, Jian Pei. – 3rd ed. - Morgan Kaufmann Publishers is an imprint of Elsevier. 2012. – 740p.

2. Mohammad Braei, Dr.-Ing. Sebastian Wagner. Anomaly detection in univariate time-series: a survey on the state -of-the-art, 2020: https://arxiv.org/pdf/2004.00433.pdf

3. Y. Jiang, C. Zeng, J. Xu and T. Li. Real time contextual collective anomaly detection over multiple data streams, 2014: https://api.semanticscholar.org/CorpusID:18868065

4. F. Anguilli and F. Fassetti. Detecting distance-based outliers in streams of data. CIKM '07: Proceedings of the sixteenth ACM conference on information and knowledge management. 2007. P. 811–820.