Methods of Anomalous Data Detection in Datasets

Akbar Rashidov
Department of Artificial intelligence
and Information systems
Samarkand state university
Samarkand, Uzbekistan
researcher.are@gmail.com

Akmal Akhatov
Department of Artificial intelligence
and Information systems
Samarkand state university
Samarkand, Uzbekistan
akmalar@ramler.ru

Akbar Soliev
Department of Artificial intelligence
and Information systems
Samarkand state university
Samarkand, Uzbekistan
akbar.soliev25@gmail.com

***Abstract*— It is known that the accuracy of data analysis and artificial intelligence models that trained and tuned on the basis of data is closely related to the quality of the data set. The quality of the data set depends on several factors, one of the most important of which is the absence or elimination of anomalous data in the data set. Anomalous data has such a property that artificial intelligence models work normally with a data set with this anomalous data. That is, artificial intelligence models do not notice at all that they are working with incorrect data. As a result, the artificial intelligence model returns an incorrect results, which may lead to incorrect conclusions about the object. Therefore, today, the detection of anomalous data in the datasets is one of the studies that has retained its relevance. This research paper discusses anomalous data, their negative consequences, and the types of anomalies in the data set. It also studies methods for detecting anomalous data in datasets and analyzes their use cases.**

Keywords*— *anomalous data, types of anomalous data, anomalous data detections

I. Introduction

Today, data analysis is gaining ground in all fields [1]. The increase in data per second, the interest in identifying hidden patterns in this data is motivating the development of the field of data analysis [2]. It can also be said that the development of artificial intelligence and big data technologies has brought data analysis to a new level [3-5]. The combination of these three fields, data analysis, artificial intelligence and big data technologies, provides solutions to complex problems in all fields [6-10]. In other words, the combination of these three fields increases the possibilities of making the right decisions in all fields. However, the increase in the volume of data also increases the likelihood of encountering unexpected information in this data [11]. It is known that the occurrence of unexpected, that is, anomalous data can negatively change the result of conventional data analysis or cause the artificial intelligence model to make the wrong decision [12]. Therefore, the detection of anomalies in the data set remains one of the current issues in the field of data analysis and artificial intelligence today.

Anomalous data is data that does not comply with the well-defined rules of normal values in the data set [13]. Typically, anomalous data in the data set can arise due to the following reasons [14]:

- due to errors in the process of collecting the data set;
- due to the use of different units of measurement or errors in the measurement process;
- as a result of the addition of various noises; - as a result of incorrect filling in of missing data.

Anomalous data can not only be negative, but also provide an opportunity to identify hidden patterns in the data set and identify new concepts. Therefore, the process of detecting and processing anomalous data is complex. Moreover the main difficulties in detecting anomalous data are followings:

- the difficulty of knowing in advance all possible normal values in the data set;
- the difficulty of determining a clear boundary between normal and anomalous values;
- the fact that sometimes anomalous values can be adjusted to normal values;
- the fact that values in many areas of data can take on variable values, which have the property of increasing or decreasing;
- the need to change anomaly detection models or threshold values as the data set changes;
- the lack of labeled data that allows detecting anomalous data using artificial intelligence models.

Due to the above difficulties and the high negative impact of anomalous data on data quality and artificial intelligence algorithms, research on detecting anomalous data in data sets remains relevant. Therefore, this research work is dedicated to the topic of anomalous data detection, and methods for detecting anomalous data will be analyzed throughout the research work.

II. Anomalous Data Detection Methods

A. Types of anomalous data

Before describing the methods for identifying anomalous data, it is necessary to know the types of anomalous data. Usually, anomalous data is divided into 3 types [15]:

- Point anomaly; - Contextual anomaly; - Collective anomaly.

Point anomaly indicates that an anomaly is observed for the values of exactly one field in the data set. For example, the occurrence of values of 15 or - 5 from the normal values in the range [1:10] represents a point anomaly.

Contextual anomaly indicates that the value in one field does not correspond to the second field in two or more related fields. For example, in a data set with time and temperature fields, the value of -50 C in June indicates an anomaly. If this situation were observed only for the temperature field, this value would not be considered anomalous. Because the temperature field of -50 C is present in the data set.

Collective anomaly refers to an anomaly in a subset of the data set, rather than a single piece of data within the data set.

B. Types of anomalous data detection methods

Detecting anomalies in data sets is a long-standing research area, and there are several methods for anomaly detection. These methods can be summarized as the hierarchy of anomaly detection methods shown in Figure 1.

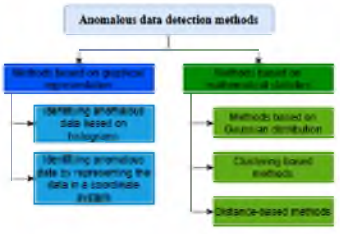

Fig. 1. Types of anomalous data detection methods

As can be seen from Figure 1, anomalous data detection methods are divided into two groups. The first group of methods is based on graphical representation. They allow detecting anomalous data by visually representing the data in the form of diagrams, tables, histograms, graphs, etc. Among these approaches, the most common and effective are histograms and coordinate system approaches.

The second group of methods are methods based on mathematical statistics, which are based on the distribution of data, its deviation from normality, the differences in distances between data, and similar statistics. There are several methods based on mathematical statistics, the most widely used of which are the following:

- Gaussian methods are methods based on the Gaussian distribution, an example of which is the Box-plots method [16]. In the Box-plots method, boundary values based on quartiles are determined, and data outside this boundary are considered anomalous data. Usually, Q1-1.5xIQR is taken as the lower boundary, and Q3+1.5xIQR is taken as the upper boundary. Here, Q1 represents the first quantile value in the data set, and Q3 represents the third quantile value, IQR=Q3-Q1. This Box-plots method is used to identify Point anomalies;

- Clustering-based methods. These methods are based on clustering data and identify data that does not belong to any cluster as anomalous data. A vivid example of these methods is the DBSCAN algorithm [17-19]. The DBSCAN clustering algorithm uses 2 main concepts in the process of accepting data as anomalous. These are Eps and MinPts. Eps is the maximum distance between neighboring points to belong to a cluster. MinPts is the minimum number of points required to be a cluster. In short, in this method, if a piece of data is located far from the eps distance of a cluster and the number of data around it that are smaller than the eps distance is less than MinPts, this data is considered anomalous data. These Clustering-based methods is used to identify all types anomalies.

Distance-based methods. There are several methods that fall into this class, a clear example of which is the ABOD method [20]. This method is called the ABOD Angle-Based Outlier Detection method. This method, which is used to detect anomalies, is based on calculating the angular variations between points. In this, the angular dispersion of each selected object is monitored.

III. CONCLUSIN

In general, the study revealed that the detection of anomalous data in a dataset has played an important role in modern data analysis and artificial intelligence. The study also revealed that there are several methods for detecting anomalous data today. However, one of the main shortcomings of these methods is the problem of defining clear boundaries. In this case, as the dataset changes, the boundaries also change, complicating the study. In general, the study of methods for detecting anomalous data in a dataset and the development of rules for using these methods is one of the current research areas in the field of science.

REFERENCES

[1] Taherdoost, H. Different Types of Data Analysis; Data Analysis Methods and Techniques in Research Projects / H. Taherdoost // International Journal of Academic Research in Management (IJARM). – 2020. – Vol. 9(1). – P. 1-9.

[2] Demidova, L.A. An Approach to Identify the Hidden Patterns in the Datasets for Patients with the Multiple Chronic Diseases / L.A. Demidova, N.V. Doroshina // Procedia Computer Science. – 2021. – Vol. 186. – P. 620-627. DOI: 10.1016/j.procs.2021.04.184.

[3] Rashidov, A. Real-Time Big Data Processing Based on a Distributed Computing Mechanism in a Single Server / A. Rashidov, A.R. Akhatov, F.M. Nazarov // Stochastic Processes and Their Applications in Artificial Intelligence / ed. by C. Ananth, N. Anbazhagan, M. Goh. – IGI Global, 2023. – P. 121-138. DOI: 10.4018/978-1-6684-7679-6.ch009.

[4] Akhatov, A. Mechanisms of Information Reliability in Big Data and Blockchain Technologies / A. Akhatov, F. Nazarov, A. Rashidov // 2021 International Conference on Information Science and Communications Technologies (ICISCT). – 2021. DOI: 10.1109/ICISCT52966.2021.9670052.

[5] Makhmadiyarovich, N.F. Development of Algorithms for Predictive Evaluation of Investment Projects Based on Machine Learning / N.F. Makhmadiyarovich, Y. Sherzodjon // Artificial Intelligence, Blockchain, Computing and Security. – Vol. 2. – 2023. – P. 681-685.

[6] Zaripova, R. Unlocking the Potential of Artificial Intelligence for Big Data Analytics / R. Zaripova, V. Kosulin, M. Shkinderov, I. Rakhmatullin // E3S Web of Conferences. – 2023. – Vol. 460. – P. 04011. DOI: 10.1051/e3sconf/202346004011.

[7] Akhatov, A. Optimization of the Database Structure Based on Machine Learning Algorithms in Case of Increased Data Flow / A. Akhatov, A. Renavikar, A. Rashidov // Proceedings of the International Conference on Artificial Intelligence, Blockchain, Computing And Security (ICABCS 2023). – 2023.

[8] Zaynidinov, H. Intelligent Algorithms of Digital Processing of Biomedical Images in Wavelet Methods / H. Zaynidinov, L. Khuramov, D. Khodjaeva // Proceedings of the International Conference on Artificial Intelligence, Blockchain, Computing and Security (ICABCS 2023). – 2024. – Vol. 2. – P. 648-653.

[9] Nazarov, F.M. Optimization of Prediction Results Based on Ensemble Methods of Machine Learning / F.M. Nazarov, S. Yarmatov // 2023 International Russian Smart Industry Conference (SmartIndustryCon). – 2023. – P. 181-185. DOI: 10.1109/SmartIndustryCon57312.2023.10110726.

[10] Benzidia, S. The Impact of Big Data Analytics and Artificial Intelligence on Green Supply Chain Process Integration and Hospital Environmental Performance / S. Benzidia, N. Makaoui, O. Bentahar // Technological Forecasting and Social Change. – 2021. – Vol. 165. – P. 120557.

[11] Abghari, S. Data Mining Approaches for Outlier Detection Analysis / S. Abghari. – Blekinge Institute of Technology Doctoral Dissertation Series, 2020. – Vol. 2020:09.

[12] Rashidov, A.E. Selecting Methods of Significant Data from Gathered Datasets for Research / A.E. Rashidov, J.S. Sayfullaev // International Journal of Advanced Research in Education, Technology and Management. – 2024. – Vol. 3(2). – P. 289-296. DOI: 10.5281/zenodo.10781255.

[13] Chinn, C.A. The Role of Anomalous Data in Knowledge Acquisition: A Theoretical Framework and Implications for Science Instruction / C.A. Chinn, W.F. Brewer // Review of Educational Research. – 1993. – Vol. 63(1). – P. 1-49. DOI: 10.2307/1170558.

[14] Meleshko, A.V. [Название статьи] / A.V. Meleshko [et al.] // IOP Conference Series: Materials Science and Engineering. – 2020. – Vol. 709. – P. 033034.

[15] Singh, K. Outlier Detection: Applications And Techniques / K. Singh, S. Upadhyaya // International Journal of Computer Science Issues (IJCSI). – 2012. – Vol. 9, Issue 1(3). – P. 307-323.

[16] Stevens, D.L. Recommended Methods for Outlier Detection and Calculations of Tolerance Intervals and Percentiles – Application to RMP data for Mercury-, PCBs-, and PAH-contaminated Sediments / D.L. Stevens. – 2011.

[17] Rashidov, A. The Same Size Distribution of Data Based on Unsupervised Clustering Algorithms / A. Rashidov, A. Akhatov, F. Nazarov // Advances in Artificial Systems for Logistics Engineering III. ICAILE 2023. Lecture Notes on Data Engineering and Communications Technologies. – 2023 – Vol. 180. DOI: 10.1007/978-3-031-36115-9_40.

[18] Bolikulov, F. Effective Methods of Categorical Data Encoding for Artificial Intelligence Algorithms / F. Bolikulov, R. Nasimov, A. Rashidov, F. Akhmedov, Y.-I. Cho // Mathematics. – 2024. – Vol. 12(16). – P. 2553. DOI: 10.3390/math12162553.

[19] Rashidov, A. The Distribution Algorithm of Data Flows Based on the BIRCH Clustering in the Internal Distribution Mechanism / A. Rashidov, A. Akhatov, D. Mardonov // 2024 International Russian Smart Industry Conference (SmartIndustryCon). – 2024. – P. 923-927. DOI: 10.1109/SmartIndustryCon61328.2024.10516193.

[20] Hodge, V. A Survey of Outlier Detection Methodologies / V. Hodge, J. Austin // Artificial Intelligence Review. – 2004. – Vol. 22. – P. 85-126. DOI: 10.1023/B:AIRE.0000045502.10941.a9.