Available online www.ejaet.com

European Journal of Advances in Engineering and Technology, 2018, 5(12):1099-1105

Research Article ISSN: 2394-658X

Anomaly Detection Techniques in Data Mining

Khirod Chandra Panda

Asurion Insurance, USA

Email id - khirodpanda4bank@gmail.com, Orcid id: 0009-0008-4992-3873

ABSTRACT

Anomaly detection has emerged as a crucial research area for modern researchers, particularly within the realm of data mining. This field is pivotal for future advancements in data mining. Data mining refers to the use of specific methods and algorithms designed to extract and analyze data, uncovering rules and patterns that describe the fundamental properties of data sets. These techniques can be applied to diverse data types, unveiling hidden structures and relationships. In today's data-driven world, massive amounts of data are stored and transferred from one place to another, often exposing the data to potential threats. Although various techniques and applications are deployed to safeguard data, vulnerabilities still exist. To mitigate these risks and identify different types of cyber threats, data mining techniques are increasingly utilized to strengthen data security. Anomaly detection leverages data mining methods to identify unusual or unexpected behaviors within data, enhancing security by reducing the likelihood of intrusion or attack. This paper focuses on the application of anomaly detection in data mining, with a specific emphasis on identifying anomalies in time series data using machine learning techniques.

Keywords: Anomaly Detection; Data Mining; KDD, Time Series Data; Machine Learning Techniques.

INTRODUCTION

The advancement of Information Technology has led to the creation of large databases and the accumulation of extensive data across various fields. This proliferation of data has significantly enhanced approaches to collecting and utilizing this valuable information for further decision-making processes.

Data Mining (DM) is formally recognized as the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [1]. Originally a lesser-known technique used by select credit institutions and retailers, data mining has evolved into a multi-billion-dollar industry. Banks employ data mining to assess the creditworthiness of their clients, while retailers use it to optimize store layouts and insurance companies to detect potentially fraudulent claims [1].

The maturity of data mining is also evident from the fact that most database providers now offer integrated data mining solutions. These tools facilitate the generation of knowledge and contribute to the broader scope of data mining applications. Although still in its nascent stages in areas like crime prevention and bioinformatics, the potential of data mining in these fields is vast. The ongoing enhancements to data mining techniques have improved accuracy, integration with existing databases, and real-time analysis capabilities [1].

Anomaly detection, a key challenge in various research domains and application fields, has seen the establishment of specific techniques for certain applications, while others remain more generalized [2]. Anomaly detection involves identifying patterns in data that deviate from expected behavior, often referred to as anomalies, outliers, discordant observations, exceptions, aberrations, surprises, peculiarities, or contaminants in various application contexts. The terms anomalies and outliers are particularly prevalent in the field of anomaly detection, sometimes used interchangeably [2].

Anomaly detection is widely applicable in numerous areas, including fraud detection in credit card transactions, insurance, healthcare, intrusion detection in cybersecurity, fault detection in safety-critical systems, and military surveillance for detecting enemy activities. The importance of anomaly detection stems from the fact that anomalies often translate into significant and actionable information in a wide range of application domains [2].

In this paper, we propose techniques based on data mining for effective anomaly detection. This new research focuses on analyzing specific data using data mining techniques. The paper concludes with a discussion on future research directions for newcomers to the field.

LITERATURE REVIEW

Data mining is the process of discovering patterns, correlations, and anomalies in large datasets by using statistical methods, machine learning algorithms, and database systems [4]. It involves extracting useful information from vast amounts of data to generate actionable insights and make informed decisions. The objective of data mining is to convert raw data into useful information by identifying and analyzing hidden patterns and relationships that are not readily apparent.

KDD (Knowledge discovery process), is a process of discovering useful knowledge and insights from large and complex datasets [5]

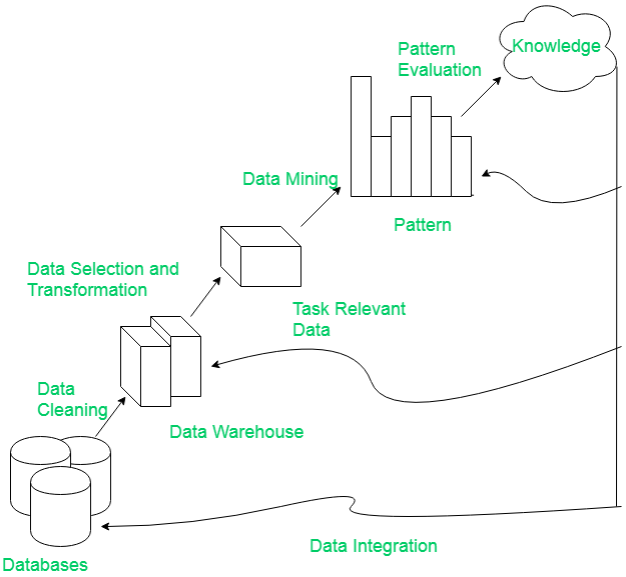

Figure 1: KDD Process

The overall process of finding and interpreting patterns from data involves the repeated application of the following steps [6]:

[1]. Developing an understanding of
   A. the application domain
   B. the relevant prior knowledge
   C. the goals of the end-user

[2]. Creating a target data set: selecting a data set, or focusing on a subset of variables, or data samples, on which discovery is to be performed.

[3]. Data cleaning and preprocessing.
   A. Removal of noise or outliers.
   B. Collecting necessary information to model or account for noise.
   C. Strategies for handling missing data fields.
   D. Accounting for time sequence information and known changes.

[4]. Data reduction and projection.
   A. Finding useful features to represent the data depending on the goal of the task.
   B. Using dimensionality reduction or transformation methods to reduce the effective number of variables under consideration or to find invariant representations for the data.

[5]. Choosing the data mining task.
   A. Deciding whether the goal of the KDD process is classification, regression, clustering, etc.

[6]. Choosing the data mining algorithm(s).
   A. Selecting method(s) to be used for searching for patterns in the data.
   B. Deciding which models and parameters may be appropriate.
   C. Matching a particular data mining method with the overall criteria of the KDD process.

[7]. Data mining.
   A. Searching for patterns of interest in a particular representational form or a set of such representations as classification rules or trees, regression, clustering, and so forth.

[8]. Interpreting mined patterns.

[9]. Consolidating discovered knowledge.

The terms knowledge discovery and data mining are distinct.
KDD refers to the overall process of discovering useful knowledge from data. It involves the evaluation and possibly interpretation of the patterns to make the decision of what qualifies as knowledge. It also includes the choice of encoding schemes, preprocessing, sampling, and projections of the data [7][8] prior to the data mining step. Data mining refers to the application of algorithms for extracting patterns from data without the additional steps of the KDD process.

ANOMALY DETECTION

Anomalies are patterns in dataset which do not conform to a well-defined notion of normal behavior [9]. Anomalies cannot always be characterized as attack, but it can be a surprising or unexpected behavior which is previously not known. It may or may not be harmful. Figure 2 illustrates anomalies in a simple two-dimensional data set. The data has two normal regions N1 and N2 since most observations lie in these two regions. Points that are necessarily far away from these regions, for example, the points o1 and o2, and points in region O3 are anomalies [10].

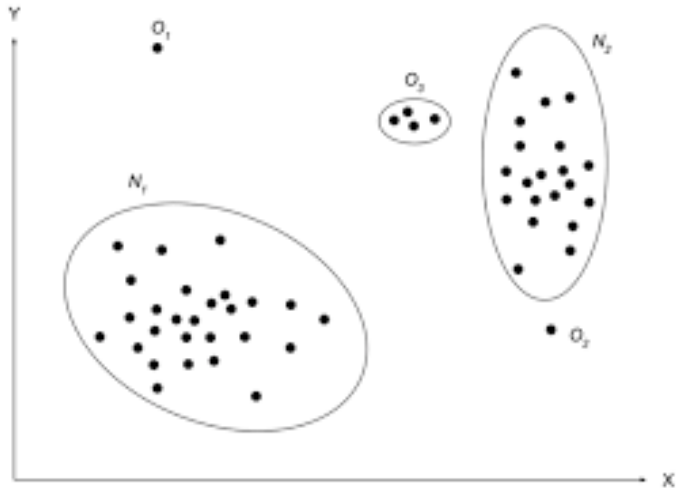

Figure 2: Anomaly in two-dimensional data set

Anomalies might be driven in the dataset for a variety of reasons, like some malicious activity, for example, credit card fraud, cyber-intrusion, terrorist activity or breakdown of a system, but all of the reasons have the common characteristic that they are interesting to the analyst. The interestingness or real-life importance of anomalies is a key feature of anomaly detection [2]. Anomaly detection is defined as the process of finding the patterns in a dataset whose behavior is not normal or expected [11]. It is the identification of data points, items, observations, or events that do not conform to the expected or known pattern of a given group. These anomalies occur very rarely but may signify a large and significant threat like cyber intrusions or fraud. Anomaly detection is extremely used in behavioral analysis and further methods of analysis such that help in learning about the detection, identification, and prediction of the happening of these anomalies [12]. Anomaly detection is primarily a process of data mining and is used to determine the types of anomalies happening in a given data set and to define details about their happenings. It is applicable in domains such as fraud detection, intrusion detection, fault detection, system health monitoring and event detection systems in sensor networks. In the situation of fraud and intrusion detection, the anomalies or interesting patterns are not necessarily the rare items but those unexpected torrents of activities. These types of anomalies do not conform to the definition of anomalies or outliers as rare incidences, so many anomaly detection methods do not work in these instances unless they have been suitably combined or trained. So, in these cases, a cluster analysis algorithm may be more suitable for detecting the micro cluster patterns created by these data points [2]. Figure 3 illustrated the key components associated with an anomaly detection technique.

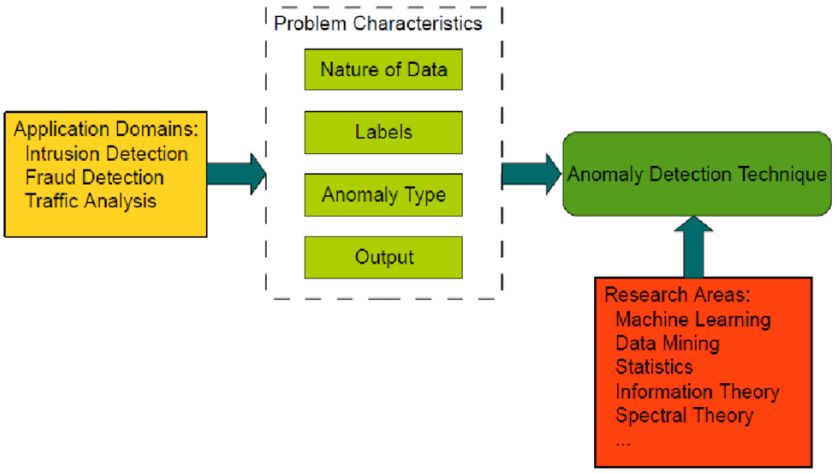

Figure 3: Key Components of Anomaly detection

[1]. Types of Anomalies

A. Point Anomalies: A point anomaly is where a single datapoint stands out from the expected pattern, range, or norm. In other words, the datapoint is unexpected.

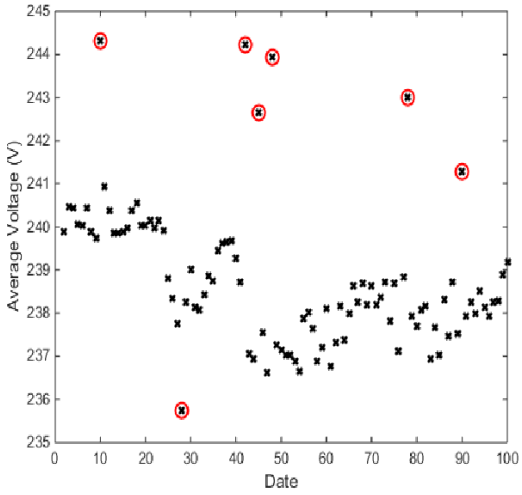

Figure 4: Point Anomalies

B. Contextual Anomalies: -Instead of looking at specific datapoints or groups of data, [13] an algorithm looking for contextual anomalies will be interested in unexpected results that come from what appears to be normal activity.

The crucial element here is context: Are the results out of context?

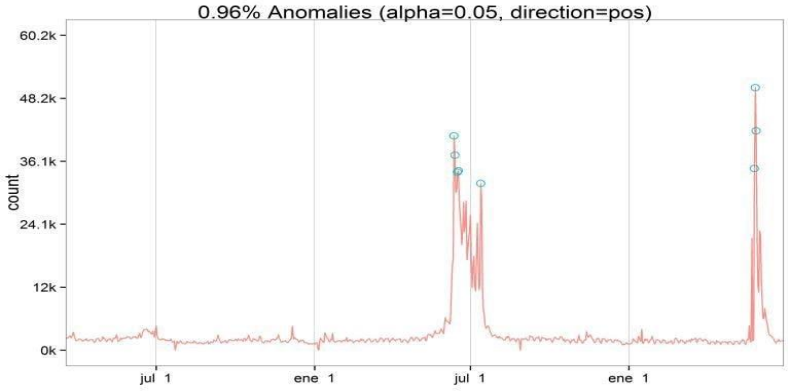

Figure 5: Contextual anomaly detection using the Twitter AnomalyDetection package in R

C. CollectiveAnomalies: - A collective anomaly occurs where single datapoints looked at in isolation appear normal. When you look at a group of these datapoints, however, unexpected patterns, behaviors, or results become clear.
Those unexpected occurrences could be, for example, events occurring in an order or in a combination that is unexpected.

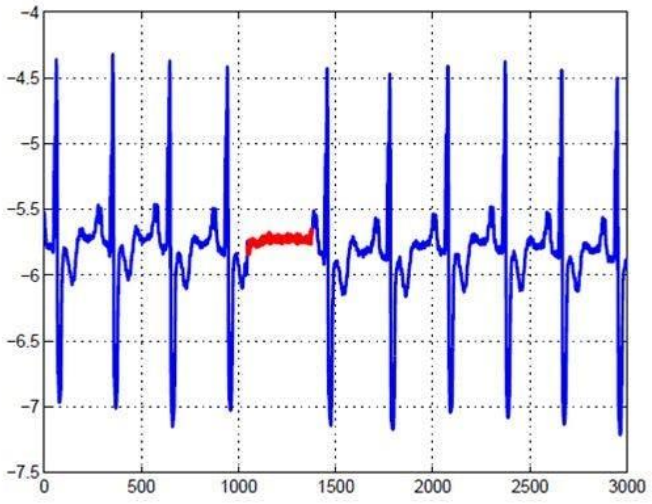

Figure 6: Collective Anomaly

We can use potential credit card fraud as our example again. This would occur where multiple purchases appear to fit within normal spending activity when looked at individually. When you look at these purchases as a group, however, unusual patterns and behavior can appear.

[2]. Data Labels

The labels related with a data point indicate whether that point is normal or anomalous [14]. It should be illustrious that obtaining labeled data that is exact as well as characteristic of all types of behaviors is often too expensive. Labeling is frequently complete by a human expert and hence considerable effort is required to obtain the labeled training data set [15]. Normally, getting a labeled set of anomalous data instances / point that covers all possible type of anomalous behavior is more difficult than receiving labels for normal behavior [16]. Based on the scope which the labels are available, anomaly detection techniques can work on one of the following three modes [17][18]:

A. Supervised Anomaly Detection: In supervised mode techniques trained with assume the availability of a training data set which has labeled instances for normal as well as anomaly classes.
B. Semi-supervised Anomaly Detection: In semi-supervised mode techniques operate with assume that only for the normal class the training data has labeled instances. Then they do not require labels for the anomaly class, they are more broadly valid than supervised techniques.
C. Unsupervised Anomaly Detection: In unsupervised mode techniques operate those do not require training data and hence are most widely applicable. The techniques in this approach type the contained assumption that normal instances are far more frequent than anomalies in the test data. If this assumption [19][20] is not true then such techniques suffer from high false alarm rate.

[3]. Output of Anomaly Detection

A significant feature for any anomaly detection technique is the way in which the anomalies are described. Typically, the outputs made by anomaly detection techniques are one of the following two types [21]:

A. Scores: Scoring techniques assign an anomaly score to each point in the test data depending on the degree to which that point is considered an anomaly. Thus, the output of such techniques is a ranked list of anomalies [22]. An analyst may select to either analyze the topmost anomalies or use a cutoff threshold to select the anomalies.
B. Labels: In this category, techniques assign a label – 'normal' or 'anomalous' [23] [24] to each test instance.

CONCLUSION

This paper has comprehensively addressed the multifaceted aspects of anomaly detection within the realm of data mining, offering a detailed exploration of the types of anomalies, methodologies for detection, and the practical implications of these technologies in various fields. We have systematically categorized anomalies into point anomalies, contextual anomalies, and collective anomalies, each requiring unique strategies for identification and analysis.

Through the lens of data mining, we delved into advanced detection techniques such as statistical methods, machine learning algorithms, and neural networks. Each method brings distinct advantages and is suitable for specific types of data and anomaly detection tasks. Machine learning techniques, particularly supervised and unsupervised learning, have shown great promise in enhancing the accuracy and efficiency of anomaly detection systems. Moreover, the use of deep learning has emerged as a powerful tool for handling high-dimensional data and complex anomaly patterns that traditional methods might not effectively address.

The practical applications of anomaly detection in data mining are vast and impactful. Industries ranging from finance to healthcare benefit significantly from the insights provided by these techniques. In cybersecurity, for example, anomaly detection is crucial for identifying potential threats and vulnerabilities, thereby safeguarding sensitive information and systems. In healthcare, these techniques are used to detect aberrations in patient data, which can be indicative of critical, unforeseen medical conditions.

However, despite the advancements and successes, challenges such as the handling of high-dimensional space, the dynamic nature of data, and the differentiation between noise and anomalies continue to impose constraints on the effectiveness of anomaly detection systems. Future research should thus focus on refining these techniques, possibly through the integration of AI advancements and more sophisticated algorithms, to enhance their adaptability and accuracy in real-world scenarios.

In summary, as data continues to grow both in size and complexity, the role of anomaly detection in data mining becomes increasingly essential. This paper not only highlights current methodologies and applications but also sets the stage for future innovations that could redefine how we approach anomalies in vast datasets. By continuing to evolve and adapt these techniques, we can look forward to more robust, efficient, and precise anomaly detection systems that can meet the demands of tomorrow's data-driven world.

REFERENCES

[1]. J. Huysmans, B. Baesens, D. Martens, K. Denys and J. Vanthienen, New Trends in Data Mining, TijdschriftvoorEconomieen Management, Vol. L, 4, 2005: 1-14.

[2]. Varun Chandola, Arindam Banerjee and Vipin Kumar, Anomaly Detection: A Survey, ACM Computing Surveys, Vol. 41, No. 3, Article 15, 2009: 1-58.

[3]. AnimeshPatcha, Jung-Min Park, An overview of anomaly detection techniques: Existing solutions and latest technological trends, ScienceDirect 2007.

[4]. G. Babu and 2T. Bhuvaneswari, "A Data Mining Technique to Find Optimal Customers for Beneficial Customer Relationship Management", Journal of Computer Science 8 (1): 89-98, 2012

[5]. Fayyad, Piatetsky-Shapiro, Smyth, "From Data Mining to Knowledge Discovery: An Overview", in Fayyad, Piatetsky-Shapiro, Smyth, Uthurusamy, Advances in Knowledge Discovery and Data Mining, AAAI Press / The MIT Press, Menlo Park, CA, 1996, pp.1-34

[6]. Kalyani M Raval, Data Mining Techniques, International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE) Volume 2, Issue 10, 2012: 439-442.

[7]. [Philippe Esling and Carlos Agon, Time-Series Data Mining, ACM Computing Surveys, Volume 45, No. 1, Article 12 (2012),: 1- 34.

[8]. Varun Chandola, Deepthi Cheboli, and Vipin Kumar, Detecting Anomalies in a Time Series Database, ACM, Technical Report (2009).

[9]. Victoria J. Hodge & Jim Austin, A Survey of Outlier Detection Methodologies, Artificial Intelligence Review 22 (2004): 85–126.

[10]. AnvardhNanduri and Lance Sherry, Anomaly Detection In Aircraft Data Using Recurrent Neural Networks (RNN), IEEE Integrated Communications Navigation and Surveillance (ICNS) Conference, 5C2-8(2016):19-21.

[11]. Vrushali D. Mane and S.N. Pawar, Anomaly based IDS using Backpropagation Neural Network, International Journal of Computer Applications (0975 – 8887) Volume 136 – No.10 (2016):29-34.

[12]. Pavel Kachurka and Vladimir Golovko, Neural Network Approach to Real-Time Network Intrusion Detection and Recognition, The 6th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications 15-17 (2011): 393-397.

[13]. Usman Ahmed and Asif Masood, Host Based Intrusion Detection Using RBF Neural Networks, IEEE 2009 International Conference on Emerging Technologies (2009): 48-51.

[14]. TetianaGladkykh, TarasHnot and Volodymyr Solskyy, Fuzzy Logic Inference for Unsupervised Anomaly Detection, IEEE First International Conference on Data Stream Mining & Processing 23-27 (2016): 42-47.
[15]. HesamIzakian and Witold Pedrycz, Anomaly Detection in Time Series Data using a Fuzzy C-Means Clustering, IEEE (2013): 1513-1518.
[16]. LinquanXie, Ying Wang, Liping Chen, and Guangxue Yue, An Anomaly Detection Method Based on Fuzzy Cmeans Clustering Algorithm, Proceedings of the Second International Symposium on Networking and Network Security (ISNNS '10), Academy Publisher, 2-4 (2010): 089-092.
[17]. Saeed Aghabozorgi and Teh Ying Wah, Effective Clustering of Time-Series Data Using FCM, International Journal of Machine Learning and Computing, Vol. 4, No. 2, (2014): 170-176.
[18]. MunaMhammad T. Jawhar and Monica Mehrotra, Design Network Intrusion Detection System using hybrid Fuzzy-Neural Network, International Journal of Computer Science and Security, Volume 4, Issue 3(2010): 285- 294.
[19]. Sampada Chavan, Khusbu Shah, Neha Dave and Sanghamitra Mukherjee, Ajith Abraham and Sugata Sanyal, Adaptive Neuro-Fuzzy Intrusion Detection Systems, IEEE Computer Society Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'04) (2004).
[20]. Gang Wang, Jinxing Hao, Jian Ma, and Lihua Huang, A new approach to intrusion detection using Artificial Neural Networks and fuzzy clustering, Elsevier Expert Systems with Applications 37 (2010): 6225–6232.
[21]. Dahlia Asyiqin Ahmad Zainaddin and ZurinaMohdHanapi, Hybrid of Fuzzy Clustering Neural Network over Nsl Dataset for Intrusion Detection System, Journal of Computer Science, Volume 9, No. 3 (2013): 391-403.
[22]. Prof. D.P. Gaikwad, Sonali Jagtap, Kunal Thakare and Vaishali Budhawant, Anomaly Based Intrusion Detection System Using Artificial Neural Network and Fuzzy Clustering, International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 9 (2012): 1-6.
[23]. Swain Sunita, Badajena J Chandrakanta and Rout Chinmayee, A Hybrid Approach of Intrusion Detection using ANN and FCM, European Journal of Advances in Engineering and Technology, 3(2), (2016): 6-14.
[24]. Bhavana Jain and Vaishali Kolhe, Hybrid Approach for Classification using Multilevel Fuzzy Min-Max Neural Network, International Journal of Innovative Research in Computer and Communication Engineering, Volume 4, Issue 5 (2016): 8636-8640.