International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 03 Mar. 2024, pp: 97-101 www.ijaem.net ISSN: 2395-5252

A Review on Clustering Based Real Time Anomaly Detection in Big Data

Prashant V. Chauhan[1], Vijay K. Vyas[2], Darshan P. Upadhyay[3]
[1,2,3] Assistant Professor, Department of Information Technology, VVP Engineering College, Rajkot, India.

Date of Submission: 06-03-2024 Date of Acceptance: 16-03-2024

ABSTRACT: Recently, there has been a sharp increase in the amount of real-time network data due to the growing use of linked Internet-of-things devices. In particular, in the field of network anomaly detection, which is seen to be essential for network security. However, initial research has shown that the methods currently in use to identify network abnormalities are insufficiently successful, especially when it comes to real-time anomaly detection. The fundamental cause of the inefficiency of present methods is the vast amounts of data that are gathered via linked devices. Network dangers become imminent at the same time, making it essential to spot abnormalities in real-time network data. Currently, the majority of anomaly detection methods concentrate mostly on batch processing machine learning techniques. In the meanwhile, real-time analytics-focused detection techniques have a tendency to be less accurate in their detections while using more memory and taking longer to execute. Utilizing established techniques like K-means, hierarchical density-based spatial clustering of applications with noise (HDBSCAN), isolation forest, and spectral clustering, a comprehensive comparison analysis is conducted to complete the study.

KEYWORDS: Anomaly Detection, K-Means, Spectral Clustering, IsolationForest, HDBSCAN

I. INTRODUCTION

The purpose of real-time anomaly detection is to quickly identify anomaly in a system's behaviour. The anomaly / anomalies emerges as malware infection, malicious network intrusion, increased system resource usage as a result of design errors, etc.

The process of finding anomaly involves gathering a constant supply of data, analyzing it, and, if necessary, taking remedial action. Because of the volume, pace, and potential complexity of the data being analyzed, this is a very difficult task. Fail to complete this activity in a timely manner might have disastrous consequences that would severely affect business continuity.[1]Numerous anomaly detection areas of applications, including network monitoring, power distribution, Internet of Things (IoT) sensor devices, healthcare, fraud detection, transportation traffic anomalies, and many more, have integrated machine learning methods.

With the increasing number of real-time analytics applications, unsupervised machine learning algorithms utilized in this context encounter significant limitations and difficulties. Unlike batch processing, which allows data to be processed and analyzed in finished datasets, these applications create data continually, necessitating processing and analysis in an operating setting. One of the most important aspects of real-time analysis is computational performance. However, the methods now in use are not able to analyse large amounts of data due to their high memory consumption and low detection efficiency.[2] This paper has addressed some traditional anomaly detection techniques for real-time big data analytics.

II. MACHINE LEARNING BASED BIG DATA TECHNOLOGIES

The majority of existing anomaly detection techniques have been applied to the batch processing methodology, which requires manual setup and training in order to identify potential hazards. These models, however, are employed for non-real time detection and have poor scalability and dependability.[3]

Additionally, the aforementioned work developed a hierarchical temporal memory model that is capable of real-time data flow prediction based on the status from prior learning. However, the study did not utilize an attack dataset that is publicly available; instead, it only used simulated datasets for assessment; additionally, processing performance and efficiency can still be increased.[3]

Another study has developed a framework for anomaly detection in real-time network traffic using machine learning algorithm, which deals with massive volume of real-time data in a scalable and durable manner.[4] The objective is to do real-time processing and assess the real-time network flow by integrating existing machine learning with big data processing framework. They were able to obtain encouraging real-time network anomaly detection results by using such technology. For understanding dynamic behaviour, visualization tools must be included, and accuracy must still be increased.

User group profiling has been a significant danger in harmful mobile assaults in recent years. In this regard, a study has suggested a scalable method for real-time anomaly detection to anticipate targeted malware in real-time.[2] In addition, the suggested solution relies on alerting mobile device users to the target malware in order to reduce the quantity of concurrent dynamic analyses by combining the behavior-triggering probability technique with user groups. The outcome displays personalized, proactive notifications to each user, but it takes a long time to process.

Additionally, for an ultra-high speed big data environment, a real-time anomaly detection system built on Apache Hadoop has been presented. This system uses a machine learning technique to identify assaults on unknown networks.

El. at [6] proposed a framework for searching finding unusual behaviour from http log. It also compared algorithms for anomaly detection using high-dimensional data in real time, namely random projection, principal component analysis, and diffusion map; the speed and memory consumption of all three algorithms were evaluated; however, the scalability of the framework remains unknown, as simulated logs were used instead of publicly available datasets.

In the same way, a new real-time anomaly detection technique built on variable cloud resource scheduling has been presented.[7] The purpose is to track the performance of the virtual machine's stream data, including CPU load, memory utilization, and I/O. A distributed system similar to Apache Storm was employed in the study to handle performance stream data and make fast decisions. The efficacy of the proposed approach has been demonstrated by providing real-time sophisticated analytical capabilities over stream data.

Furthermore, a different research assessed several real-time anomaly detection algorithms according to their execution time, CPU use, and anomaly count. Monitoring the streaming log data produced by the national educational network is the goal.[8] Nevertheless, it did not emphasize efficiency or scalability, and it only employed tiny data sets for analysis.

A novel approach has been presented here that enables anomalies to be analysed multidimensionality and to be detected in real time. The suggested approach focuses on identifying abnormal behaviour in real time and learning the system's typical behaviour from historical data. The method completely changes how quality control is implemented in complex structures with a large number of nonlinearly associated components.[9]

In addition, a real-time threat detection system built on stream processing and machine learning techniques has been developed. This system aids in the identification and classification of known threats. However, this method has a considerable reaction time lag.

III. CLUSTERING TECHNIQUES

There are various clustering algorithms used to identify anomaly from real-time data. An unsupervised algorithm does not require a separate training and testing phase; instead, it concentrates on data that missing labelling information. For anomaly detection, several unsupervised machine learning techniques have been applied. Clustering contributes in the categorization of patterns into groups and further aids in data compression, understanding, and classification.

K-Means algorithm is most commonly used in clustering due to a number of factors, including its simplicity, flexibility, time and storage complexity, invariance to data ordering, and assured coverage. Furthermore, it is also used in some other algorithms like Spectral Clustering and Hierarchical Density Based Spatial Clustering of Application with Noise.[10]

K-Means

Depending on object properties, the items are grouped together into K disjoint clusters using the K-means clustering method.[11] The items with the same properties are gathered in the same cluster. Using k-means clustering algorithm a new flow-base anomaly detection method is design by el. at [11]. In order to enable the efficient detection of anomaly based on distance in new monitoring data, the basic cluster characteristics are utilized as patterns. Although producing a promising outcome, the system did not address the memory consumption and execution time of the suggested strategy.

A new approach is developed a multi-level hybrid intrusion detection model that makes use of K-means modifications and support vector machines.[12] This model significantly improved the classifier's performance and shortened the training period. Nevertheless, memory overflow was another issue this approach had to deal with.

Spectral Clustering

Compared to other conventional clustering methods like K-means and hierarchical clustering, spectral clustering has demonstrated superior performance.[13] Additionally, spectral clustering has been utilised to create a unique spectral ranking method for anomaly detection. This method provides an anomalous ranking based on two primary patterns or the majority class. In addition, the ranking reference is generated by a positive or negative smaller class order. The suggested method was further assessed in relation to Receiver Operating Characteristic curves. However, the evaluation measures did not address memory use or execution time.

IsolationForest

An unsupervised, tree-based ensemble technique called the isolation forest (iForest) uses the innovative idea of isolation to discover anomalies. Isolation is the process of separating one instance from all the others. Because this technique just needs the tree topologies of the trained ensemble to produce anomaly scores, it eliminates the computationally costly calculations of density or distance measurements.[22] Another approach uses a sliding window to stream data using the isolation forest technique. In order to identify the idea drift, a score was created for each instance of sliding windows. The effectiveness of the evaluation findings is demonstrated, although no comparison with current methods was made, and computational complexity was not assessed for performance measures.[16] Similar to this, in a different research, the isolated forest model's interpretability and anomaly detection performance were improved by using it for dimensionality reduction in the pre-processing stage.[15]

HDBSCAN

HDBSCAN - Hierarchical Density-Based Spatial Clustering of Applications with Noise can handle clusters with varying densities and forms, and it automatically calculates the number of clusters.[17] Moreover, it was simple to identify noise and anomaly. A malware behaviour detection method based on HDBSCAN that classifies malware samples was created by Abdullah, J., et al. [17]. This method uses single-linkage clustering algorithms to identify anonymous malware types and handles the problems of asymmetric and anonymous malware from the data. The assessment parameters employed were recall and precision. However, using that approach, none of the computation performance criteria were assessed.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

The real-time analytics for clustering techniques are shown in this section. In this experiment, four distinct types of algorithms—K-means, Isolation Forest, Spectral Clustering, and HDBSCAN are used in a cloud platform to assess the accuracy, memory use, and execution time of real-time analytics for anomaly detection.Three datasets in all were utilized for the experiment: (i) The DARPA Intrusion Detection Dataset, which comprises tagged attacks and actual network data. (ii) The dataset for the Mid-Atlantic Collegiate Cyber Defence Competition, which includes information gathered from the network during the competition as well as assaults carried out by certain teams. and (iii) the DEF CON dataset, which contains several tasks including capture the flag, scavenger hunts, and lock picking. The current real-time anomaly detection techniques have been assessed based on the performance parameters listed below: (i) Accuracy: this measures how successfully the cluster points were sorted into the appropriate groups. (ii) Memory consumption: Shows how much memory, measured in megabytes, is needed to complete a certain job. (iii) Execution time: Indicates how many seconds it took to finish the execution. The method used and the amount of the data both have an impact.

Accuracy

We evaluated each algorithm using the modified rand index approach in order to determine its accuracy. A comparison of the accuracy of three data sets and four algorithms is presented in Figure 1. The Isolation Forest method outperforms all other algorithms in terms of accuracy, as can be seen in the above figure. Furthermore, K-Means and Isolation Forest provide higher accuracy with smaller datasets, whereas spectral clustering outperforms other datasets when applied to the MACCDC dataset. Conversely, spectral clustering works best when used to medium-sized datasets. In conclusion, the DEF CON dataset yields superior results for the HDBSCAN method compared to the DARPA and MACCDC datasets.

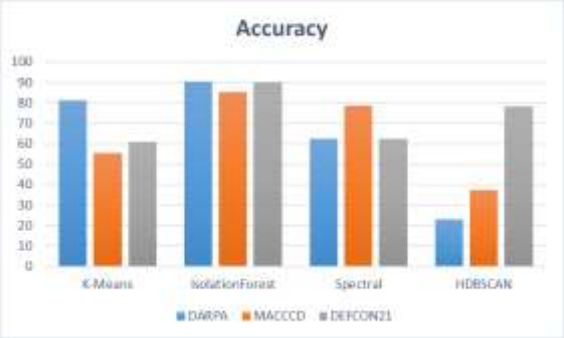

Figure 1: Accuracy Analysis of Existing Algorithms.

Memory Consumption

A comparison of each algorithm's memory use is shown in Figure 2. It is important to consider memory utilization while real-time data streaming is occurring. An application crash is a possibility if the algorithm uses a lot of memory and receives data continuously. Isolation Forest uses the most RAM when it comes to the DARPA dataset. In the MACCDC dataset, the HDBSCAN uses the least amount of memory.

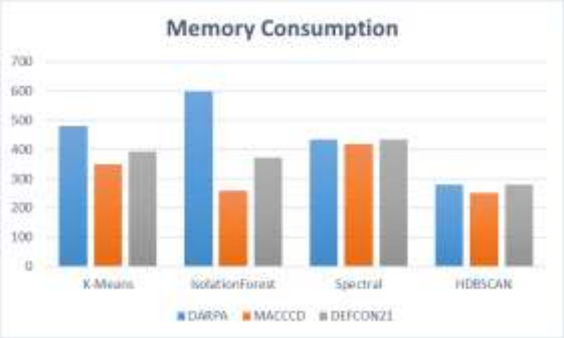

Figure 2: Memory Consumption Analysis of Existing Algorithms.

Execution Time

A comparison of each algorithm's execution times is displayed in Figure 3.In the MACCDC dataset, it is clear from Figure 3 that Isolation Forest requires the least amount of time. In the DARPA dataset, HDBSCAN uses the most time. On the other hand, K-Means uses about the same amount of time on the MACCDC and DARPA datasets.

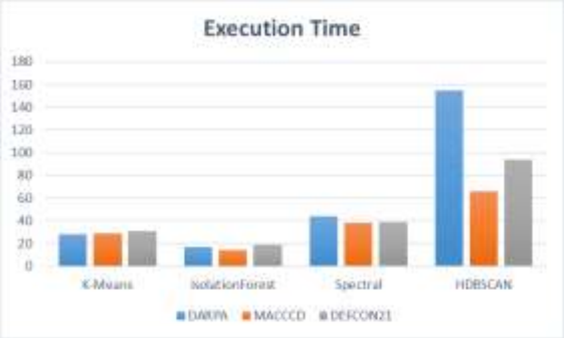

Figure 3: Execution Time Analysis of Existing Algorithms.

Figures 1, 2, and 3 display the graphs we created to represent our findings. Figure 1 shows that the Isolation Forest method yields the highest accuracy. Nevertheless, Isolation Forest uses more memory than other programmes and is not a good choice for real-time processing (Figure 2). Additionally, K-Means consumes a lot of memory and has a maximum accuracy (Figure 1), which is insufficient for anomaly identification (Figure 2). Lastly, as shown in figure 1, Spectral Clustering and HDBSCAN are not appropriate for real-time anomaly identification since their accuracy is less across all datasets.

V. CONCLUSION

In this survey paper we had represented different clustering algorithms and machine learning algorithm used for anomaly detection in big data. Various algorithms are compared based on various parameters for different data sets. Since, these algorithms lack accuracy, memory consumption or execution time. Future research will concentrate on a variety of anomaly detection applications, including industrial IoT sensors, transportation anomalies, and healthcare, and evaluate them using heterogeneous data sets.

REFERENCES

[1]. Solaimani, M., et al. Spark-based anomaly detection over multi-source VMware performance data in real-time. in Computational Intelligence in Cyber Security (CICS), 2014 IEEE Symposium on. 2014. IEEE.

[2]. Jia, B., et al., A novel real-time ddos attack detection mechanism based on MDRA algorithm in big data. Mathematical Problems in Engineering, 2016.

[3]. Wang, C., et al., A Distributed Anomaly Detection System for In-Vehicle Network Using HTM. IEEE ACCESS, 2018. 6: p. 9091-9098.

[4]. Zhao, S., et al. Real-time network anomaly detection system using machine learning. in Design of Reliable Communication Networks (DRCN), 2015 11th International Conference on the. 2015. IEEE.

[5]. McNeil, P., et al., SCREDENT: Scalable Real-time Anomalies Detection and Notification of Targeted Malware in Mobile Devices. Procedia Computer Science, 2016. 83: p. 1219-1225.

[6]. Juvonen, A., T. Sipola, and T. Hämäläinen, Online anomaly detection using dimensionality reduction techniques for HTTP log analysis. Computer Networks, 2015. 91: p. 46-56.

[7]. Solaimani, M., L. Khan, and B. Thuraisingham. Real-time anomaly detection over VMware performance data using storm. in Information Reuse and Integration (IRI), 2014 IEEE 15th International Conference on. 2014. IEEE.

[8]. Hasani, Z. Robust anomaly detection algorithms for real-time big data: Comparison of algorithms. in Embedded Computing (MECO), 2017 6th Mediterranean Conference on. 2017. IEEE.

[9]. Stojanovic, L., et al. Big-data-driven anomaly detection in industry (4.0): An approach and a case study. in Big Data (Big Data), 2016 IEEE International Conference on. 2016. IEEE.

[10]. Celebi, M.E., H.A. Kingravi, and P.A. Vela, A comparative study of efficient initialization methods for the k-means clustering algorithm. Expert systems with applications, 2013. 40(1): p. 200-210.

[11]. Münz, G., S. Li, and G. Carle. Traffic anomaly detection using k-means clustering. in GI/ITG Workshop MMBnet. 2007.

[12]. Al-Yaseen, W.L., Z.A. Othman, and M.Z.A. Nazri, Multi-level hybrid support vector machine and extreme learning machine based on modified K-means for intrusion detection system. Expert Systems with Applications, 2017. 67: p. 296-303.

[13]. Nian, K., et al., Auto insurance fraud detection using unsupervised spectral ranking for anomaly. The Journal of Finance and Data Science, 2016. 2(1): p. 58-75.

[14]. Stripling, E., et al., Isolation-based conditional anomaly detection on mixed-attribute data to uncover workers' compensation fraud. Decision Support Systems, 2018.

[15]. Puggini, L. and S. McLoone, An enhanced variable selection and Isolation Forest based methodology for anomaly detection with OES data. Engineering Applications of Artificial Intelligence, 2018. 67: p. 126-135.

[16]. Ding, Z. and M. Fei, An anomaly detection approach based on isolation forest algorithm for streaming data using sliding window. IFAC Proceedings Volumes, 2013. 46(20): p. 12-17.

[17]. Abdullah, J. and N. Chanderan, Hierarchical Density-based Clustering of Malware Behaviour. Journal of Telecommunication, Electronic and Computer Engineering (JTEC), 2017. 9(2-10): p. 159-164.