

Intrusion Detection Model using Machine Learning Algorithm on Big Data Environment

K.Ashok¹ Ms.S.Anthony Mariya Kumari²

¹Student ²Assistant Professor

^{1,2}Department of Computer Applications

^{1,2}KMM Institute of PG Studies, Tirupati, India

Abstract— Recently, the huge amounts of data and its incremental increase have changed the importance of information security and the data analysis systems for a Big Data. Intrusion detection system (IDS) is a system that monitors and also analyzes of the data to detect any intrusion in the system or the network. High volume, variety and also high speed of the data generated in the network have made the data analysis process to detect attacks by traditional techniques very difficult. Big Data techniques are to be used in IDS to deal with the Big Data for accurate and efficient data analysis process. This paper introduced Spark- Chi- SVM model for a intrusion detection. In this model, we have to be used ChiSq Selector for feature selection, and built an intrusion detection model by using support vector machine (SVM) classifier on Apache Spark Big Data platform. We used KDD99 to train and test the model.

Key words: Intrusion detection, Big Data, Apache Spark, Support vector machine (SVM), ChiSq Selector

I. INTRODUCTION

Big Data is the data that are to be difficult to store, manage, and to analyze using traditional database and also the software techniques. Big Data includes high volume and velocity, and also variety of the data that needs for new techniques to deal with it. Intrusion detection system (IDS) is the hardware or software monitor that analyzes data to detect any attack toward a system or a network. Traditional intrusion detection system techniques make the system more complex and less efficient when dealing with Big Data, because its analysis proper- ties process is complex and to take a long time. The long time it takes to be analyze the data makes the system prone to harms for some period of time before getting any alert. Therefore, using this Big Data tools and techniques to analyze and to store data in intrusion detection system can reduce computation and also training time.

The IDS has three methods for detecting attacks; Signature-based detection, Anomaly based detection, and Hybrid-based detection. The signature-based detection is to be designed to detect known attacks by using signatures of those attacks. It is an effective method of the detecting known attacks that are preloaded in the IDS database. Therefore, it is often considered to be much more accurate at identifying an intrusion attempt of the known attack. However, new types of attack cannot be detected as its signature is not presented; the databases are frequently updated in order to increase their effectiveness of detections. To overcome this problem Anomaly-based detection that compares current user activities against predefined profiles is used to detect abnormal behaviors that might be intrusions.

Anomaly-based detection is effective against unknown attacks or the zero-day attacks without any updates to the system. However, this method usually has high false

positive rates. Hybrid-based detection is a combination of the two or more methods of intrusion detection in order to overcome the disadvantages in the single method used and to obtain the advantages of the two or more methods that are used. Many researches proposed machine learning algorithm for the intrusion detection to be reduce false positive rates and to produce accurate IDS. However, to deal with the Big Data, the machine learning traditional techniques take a long time in learning and to classifying data. Using Big Data techniques and machine learning for IDS can solve many challenges such as speed and computational time and develop accurate IDS.

The objective of this paper is to be introduce the Spark Big Data techniques that deal with Big Data in the IDS in order to reduce computation time and achieve effective classification. For this purpose, we propose an IDS classification method named Spark-Chi-SVM. Firstly, a preprocessing method is used to convert the categorical data to a numerical data and then the dataset is a standardization for the purpose of improving the classification efficiency. Secondly, ChiSq Selector method is used to reduce the dimensionality on the dataset in order to further improve the classification efficiency and also to reduce of computation time for the following step. Thirdly, SVM is used for the data classification. More specifically, we use SVM With SGD in order to solve the optimization, in addition, we introduce comparison between SVM classifier and Logistic Regression classifier on the Apache Spark Big Data platform based on the area under curve (AUROC), Area Under Precision-Recall curve (AUPR) and time metrics.

The KDDCUP99 are to be tested in this study. The rest of this work is organized as follows: A review of relevant works is to conduct in “Related works” section. In “Methods” section, we introduced the proposed method. Also, each step in this method is described. Results and experiment settings are mentioned in “Result and discussion” section. Finally, we conclude our work and describe the future work in “Conclusion” section.

II. RELATIVE STUDY

A. *A big data framework for intrusion detection in smart grids using Apache Spark. In: Inter- national conference on advances in computing, communications and informatics*

Technological advancement enables the need of the internet everywhere. The power industry is not an exception in the technological advancement which makes everything smarter. Smart grid is the advanced version of the traditional grid, which makes the system is more efficient and also self-healing. Synchrophasor is a device used in smart grids to measure the values of electric waves, voltages and current. The phase or to measurement unit produces immense volume of the current and voltage data that is used to monitor and control the performance of the grid. These data are huge in

size and also vulnerable to attacks. Intrusion Detection is a common technique for finding the intrusions in the system.

In this paper, a big data framework is designed using various machine learning techniques, and intrusions are detected based on the classifications applied on the synchrophasor dataset. In this approach various machine learning techniques like a deep neural networks, support vector machines, random forest, decision trees and also naive bayes classifications are done for the synchrophasor dataset and the results are compared using metrics of the accuracy, recall, false rate, specificity, and to prediction time.

Feature selection and dimensionality reduction algorithms are used to reduce the prediction time taken by the proposed approach. This paper uses apache spark as a platform which is the suitable for the implementation of Intrusion Detection system in smart grids using big data analytics.

B. A hybrid scheme based on Big Data analytics using intrusion detection system

Network security plays a key role for many organizations. Host based and also network based Intrusion Detection Systems are available in the market depending upon the detection technology used by them. The objective of this research paper is maintaining security across the heterogeneous data from homogeneous sources and also correlating the heterogeneous data from the different sources using hybrid strategy. Methods/Statistical Analysis: A real time detection Intrusion Prevention Systems (IPS), prevents security intrusions by gathering and to composing with technologies. Findings: Heterogeneous data from the different sources has been collected from the KDD Cup Dataset and segregated into learning phase and detection phase.

In the learning phase, known attacks will be identified. Similarly detection phase also will be consider the same. Applications/Improvements: The proposed system specifies a set of rules and high DoS, R2L, U2R, Probe. One may attempt to get good results by improving the efficiency and to reducing the complexity present in the model. In future several reduction techniques may be studied to get the more features.

C. Intrusion detection and big heterogeneous data: a survey

Intrusion Detection has been heavily studied in both industry and the academia, but cyber security analysts still desire much more alert accuracy and overall threat analysis in order to secure their systems within cyberspace. Improvements to Intrusion Detection could be achieved by embracing a more comprehensive approach in monitoring the security events from many the different heterogeneous sources. Correlating security events from heterogeneous sources can grant a more holistic view and the greater situational awareness of cyber threats. One of the problem with this approach is that currently, even a single event source (e.g., network traffic) can experience Big Data challenges when considered alone. Attempts to use more heterogeneous data sources pose an even greater Big Data challenge. Big Data technologies for the Intrusion Detection can help solve these Big Heterogeneous Data challenges. In this paper, we review the scope of the works considering the problem of

heterogeneous data and in the particular Big Heterogeneous Data. We discuss the specific issues of Data Fusion, Heterogeneous Intrusion Detection Architectures, and the Security Information and Event Management (SIEM) systems, as well as presenting areas where more research opportunities exist. Overall, both cyber threat analysis and cyber intelligence could be enhanced by correlating security events across many diverse heterogeneous sources.

III. PROPOSED ALGORITHM

In the proposed method, the KDD Cup 1999 is used for a training and testing. In this proposed method the authors didn't use feature for selection technique to be select the related features. Peng et al. Proposed a clustering method for IDS based on the Mini Batch K-means combined with principal component analysis (PCA).

A. Application Architecture:

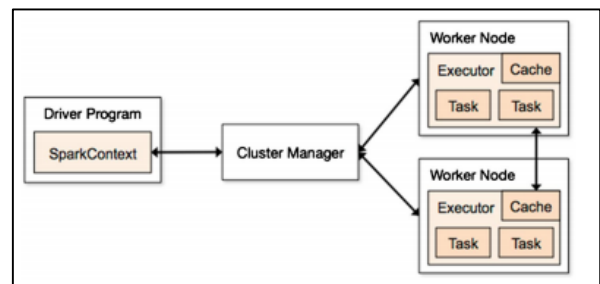


Fig. 1: Spark-architecture—official spark master/slave architecture

B. Algorithm:

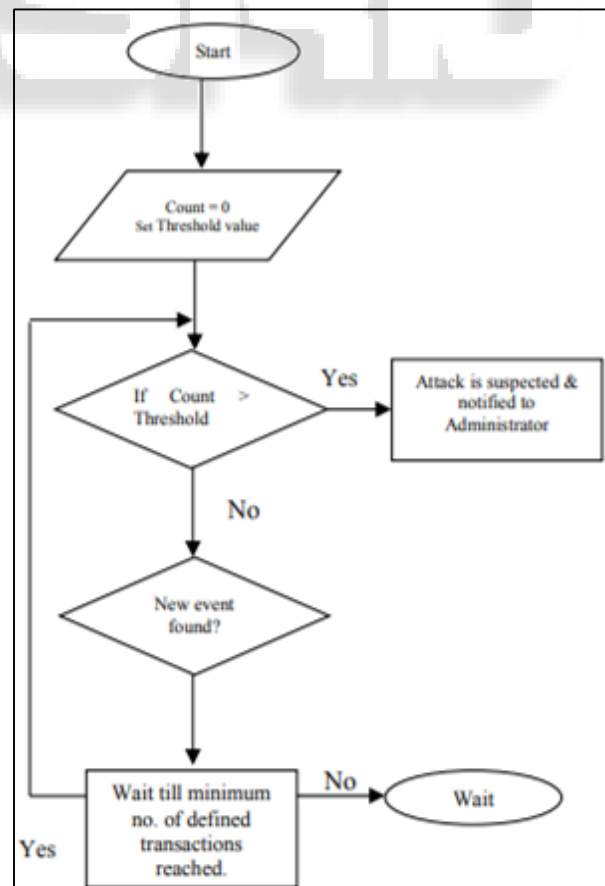
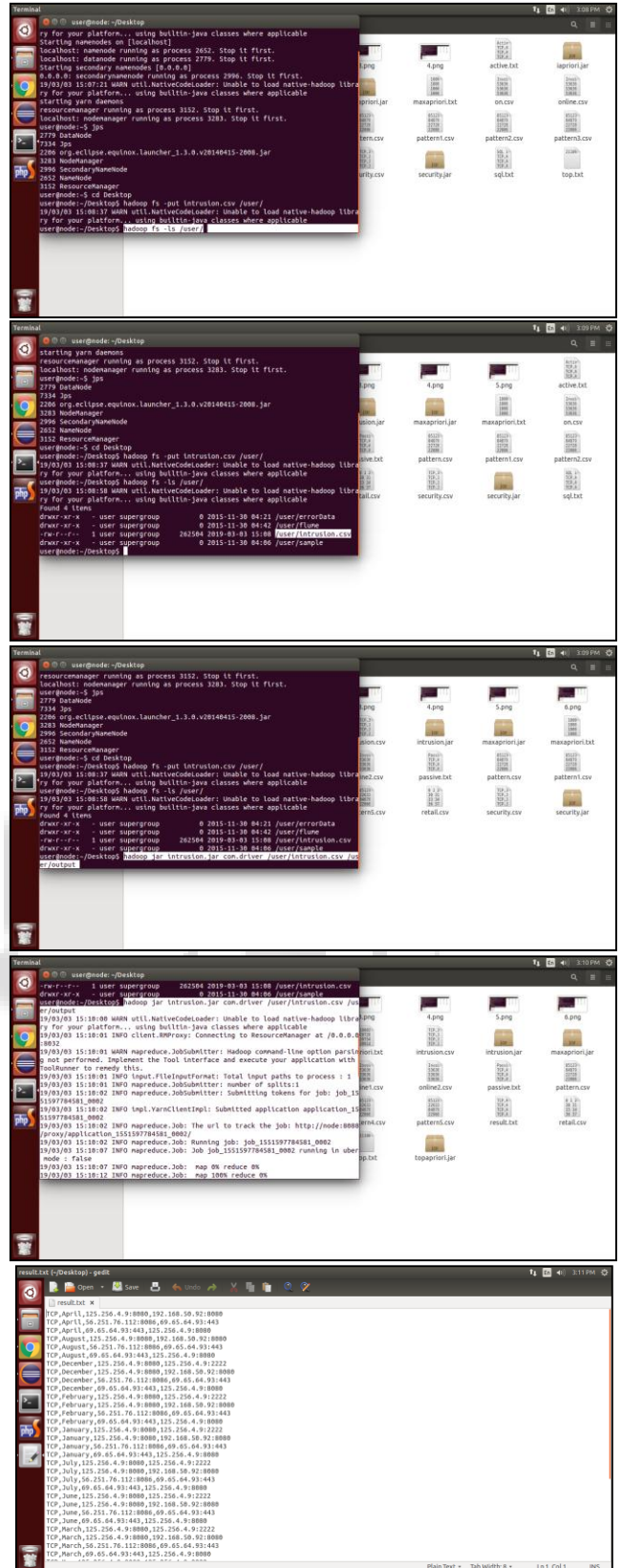
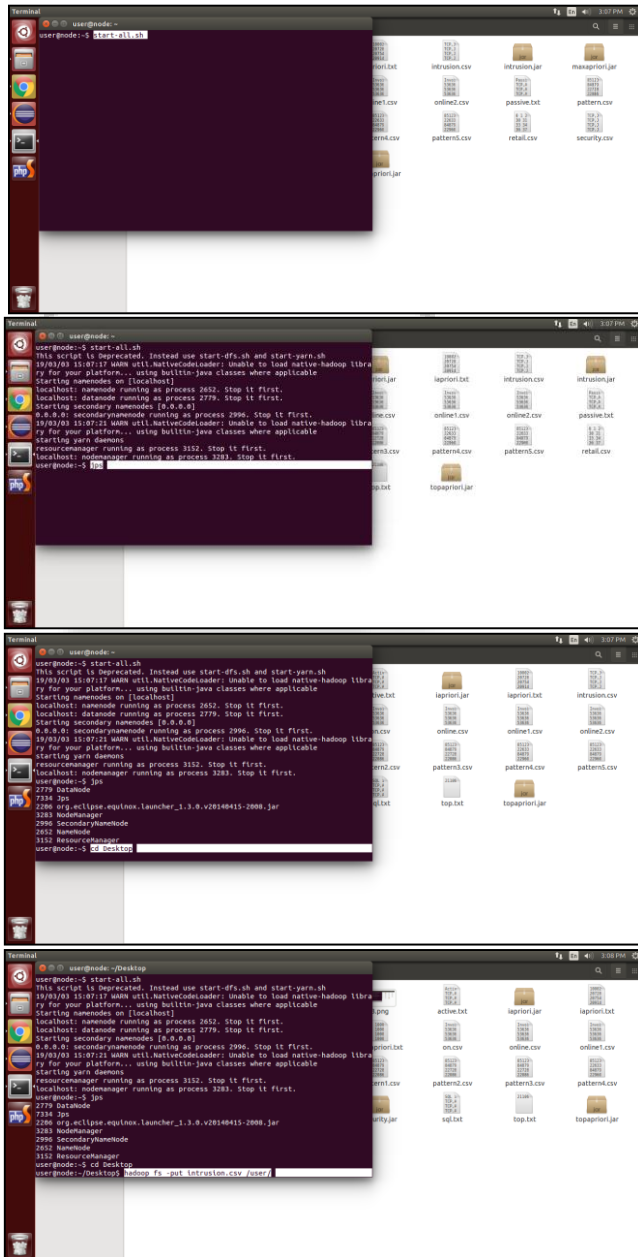
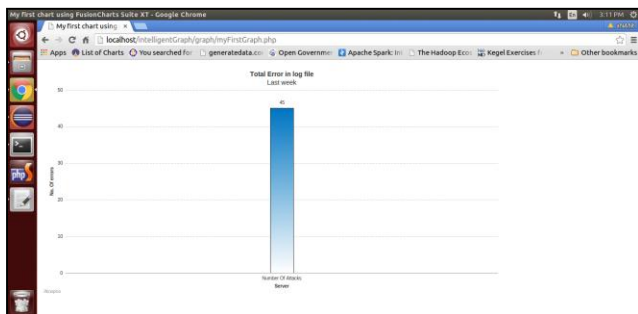


Fig. 2: Flowchart for detecting Dos Attacks

- 1) Step 1: Start
- 2) Step 2: Let the Count=0, set the threshold value. The threshold value can be set based on the working environment.
- 3) Step 3: Check if the counts of matched rules have crossed the threshold value.
 - If true, intimate the administrator assuming as an attack.
 - If false, continue.
- 4) Step 4: Check whether new event is recorded in log file.
 - If no new event found, wait
 - If event_found, go to step 2

IV. RESULTS





V. CONCLUSION

In this paper, the researchers introduced the Spark-Chi-SVM model for intrusion detection that can deal with Big Data. The proposed model used Spark Big Data platform which can process and analyze the data with high speed. Big data have a high dimensionality that makes the classification process more complex and takes a long time. Therefore, in the proposed model, the researchers used ChiSq Selector to select related features and SVM with SGD to classify the data into normal or attack. The results of the experiment showed that the model has high performance and speed. In future work, the researchers can extend the model to a multi-classes model that could detect types of the attack.

REFERENCES

- [1] Tchakoucht TA, Ezziyyani M. Building a fast intrusion detection system for high-speed networks: probe and DoS attacks detection. *Procedia Comput Sci.* 2018; 127:521–30.
- [2] Zuech R, Khoshgoftaar TM, Wald R. Intrusion detection and big heterogeneous data: a survey. *J Big Data.* 2015; 2:3.
- [3] Sahasrabudde A, et al. Survey on intrusion detection system using data mining techniques. *Int Res J Eng Technol.* 2017; 4(5):1780–4.
- [4] Dali L, et al. A survey of intrusion detection system. In: 2nd world symposium on web applications and networking (WSWAN). Piscataway: IEEE; 2015. p. 1–6.
- [5] Scarfone K, Mell P. Guide to intrusion detection and prevention systems (idps). NIST Spec Publ. 2007; 2007(800):94.
- [6] Debar H. An introduction to intrusion detection systems. In: *Proceedings of Connect*, 2000. 2000.