PLOMS Journal of Artificial Intelligence (PLOMS AI)

PLOMS AI

Research Paper

Improving Intrusion Detection Systems Based on PCA and Hadoop Platform

Hassan I. Ahmed[1,*], Abdurrahman A. Nasr[2], Salah M. Abdel-Mageid[2,3], and Heba K. Aslan[1]

[1]Informatics Department, Electronics Research Institute, Cairo, Egypt
[2]Systems and Computers Engineering Department, Faculty of Engineering, Al-Azhar University, Cairo, Egypt
[3]Computer Engineering Department, College of Computer Science and Engineering, Taibah University, KSA
* Correspondence: hassanibrsayed@gmail.com
Received: 01-06-2021; Accepted: 20-06-2021; Published: 22-06-2021

Abstract: Intrusion detection system is developed for various types of attacks, and it was improved by using machine learning. However, it still suffers from more challenges when dealing with huge amount of data (i.e., Big Data) produced by networks traffics. Also, with the advances in cloud computing and its new technologies, big data analysis made possible. In this paper, this problem of intrusion detection in big data is investigated in detail. Also, suggestions for intrusion detection system design based on Big Data analytic techniques such as MapReduce and Hadoop platform are discussed. This is work in progress and the implementation and results will be published in future articles.

Keywords: Intrusion Detection System (IDS), Big Data Analytics, MapReduce, Feature Selection

I. Introduction

Big Data analytics can be leveraged to improve information security and situational awareness [1, 2, 3]. For example, Big Data analytics can be employed to analyze financial transactions, log files, and network traffic to identify anomalies and suspicious activities, and to correlate multiple sources of information into a coherent view. The difference between traditional analytics and Big Data analytics is that Big Data analytics addresses new contributions in storage, processing, and analysis of data such as:

- Decreasing cost of the storage.
- The flexibility and cost-reduction through using data centers and cloud computing for computation and storage.
- The development of new frameworks such as Hadoop, which allow users to take advantage of distributed computing systems in storing large quantities of data through flexible parallel processing.

Big Data tools such as the Hadoop ecosystem and NoSQL databases provide the technology to increase the processing speed of complex queries and analytics. Batch processing and stream processing are two groups of data processing which branched from Big Data technologies. A stream is a sequence of unbounded tuples of the form (a_1, a_2, a_3, ……., a_n, t) generated continuously in time, where (a_i) denotes an attribute and (t) denotes the time [1]. The Hadoop is considered as one of the well-known technologies for batch processing. The Hadoop framework with the Hadoop Distributed File System (HDFS) saves a method for storing large files. In addition, the MapReduce programming model, which is tailored for frequently occurring large-scale data processing problems, can be distributed and parallelized.

II. Intrusion Detection System and Big Data Challenge

IDS is a collection of methods, gadgets and resources that can detect, identify, classify and report intrusions. Intrusion is an unauthorized (unwanted) activity in a network which can be passive or active. Passive attacks can do information gathering or eavesdropping, active attacks can do harmful packet forwarding, packet dropping or hole attacks. Networks intrusion can take a them such as attempted break-in which means an attempt has an unauthorized access to the network, masquerade which means an attacker uses a fake identity to gain unauthorized access to the network, penetration which means the acquisition of unauthorized access to the network, leakage which means an undesirable information flow from the network, DoS -preventing the network resources or network services to the other users and malicious use which means harming the network resources deliberately [2].

III. Improvements in IDS

Nowadays, the data transfer rate through network expansion and the unpredictability in Internet usage have increased anomaly problems[3]. As a result, there is a necessity to develop more stable, effective, and self-monitoring schemes. This will decrease the catastrophic defeats of susceptible systems. Detection stability and precision are couple key indicators that are used to evaluate IDS. Improvements in IDS detection stability and precision have been reported in literature [4][5]. Research work focused on using statistical methods also rule-based expert systems. However, the results from the application of both methods were not accurate, particularly when data became Big Data.

In addition, some improvements in the area of detection accuracy for IDS design have been proposed by using machine-learning methods including Support Vector Machines (SVM), Linear Genetic Programming (LGP), and Fuzzy Inference Systems (FISs) [6][7][8]. One of the most prevalent techniques is the Neural Network (NN), which has successfully been used to resolve complex practical problems in IDS. The emergence of Big Data has pointed to the major drawbacks in the traditional IDS:

1. Inability to store and retain voluminous data.
2. Inability to perform analytics of complex queries and structured data sets.
3. Inability to analyze and manage numerous unstructured data sets.
4. Inability to create cluster computing infrastructures.

The bottlenecks in traditional IDS for Big Data arose primarily due to the change in environment, system, and input data. The proposed solutions are presented in Figure 1.

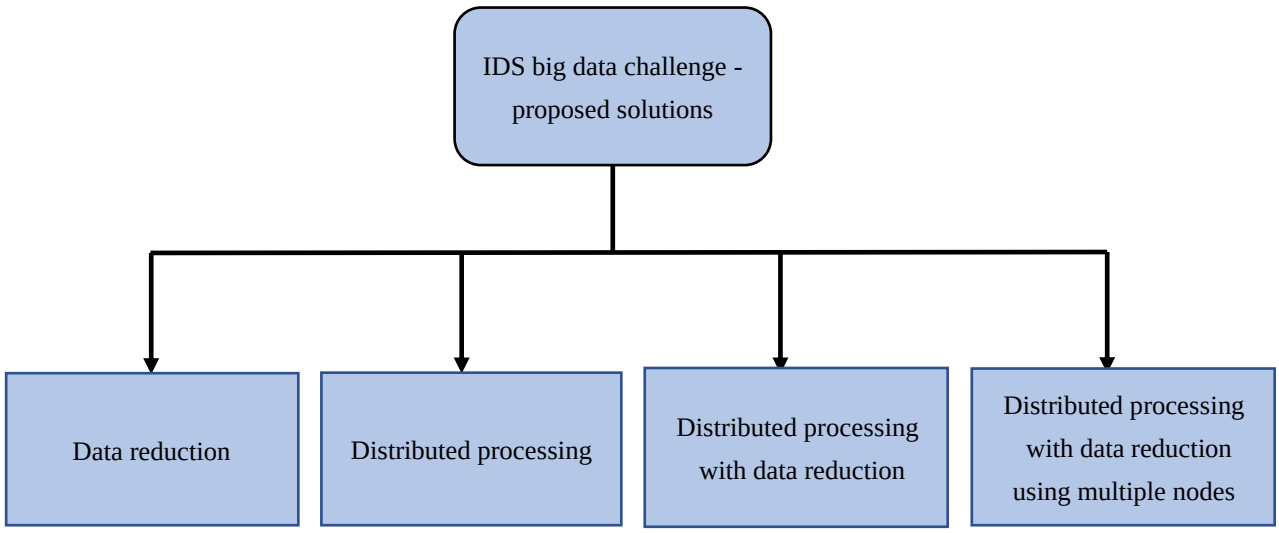

Figure 1: Improvements in IDS.

Data reduction relates to the amount of data which the IDS needs to process in order to specify the traffic behavior. This is also known as Feature Selection, which is used to eliminate redundant data in intrusion detection [9]. In this study [9], using all the features of collected data does not necessarily guarantee the best

performance of the IDS. For example, the computational cost and the error rate of the system may increase. Increasing the number of selected features also increases the need for computation power and vice versa. Due to correlations between features, some selected features can perform similar or better results in corresponding to using all the features of a dataset.

The feature selection techniques include Random Forest-Forward Selection Ranking "RF-FSR" furthermore Random Forest-Backward Elimination Ranking "RF-BER". Machine learning-based (ML-based) feature selection for the detection of network intrusions consists of four consecutive steps - dataset selection and preprocessing, feature selection, model selection, and evaluation.

Choosing the right dataset for model validation and evaluation is an important task [9]. The preprocessing operation is done through a number of steps such as:

- Removal of redundant records: Deleting repetitive records improves the accuracy of the model and reduces computation cost. For instance, 78% of KDD-99 records are repeated records.
- Normalization of data: This involves condensing data to the scope of the model without losing the effect of the data on the model.
- Discretization of data: In the KDD-99 dataset, there are many types of variables. Some are continuous while others are discrete. To make computation easier, it is important to discretize continuous variables.
- Balancing data: In the KDD-99 dataset, for example, each type of attack has many connections. Denial of Service (DoS) attack type has 391,458 connections whereas User-to-Root (U2R) attack has only 52 connections. Unbalanced dataset causes biasing in the learning process of the models.

Feature selection method uses Support Vector Decision Function (SVDF) to determine different feature weights [9]. Feature correlations, which involve the dependence of features on one other, are determined either by Forward Selection Ranking (FSR) or Backward Elimination Ranking (BER) algorithms. Random Forest (RF) has been used for model selection [9]. RF consists of decision trees, each of which gives a specific classification of input data. Voting then takes place and RF determines the final classification decision based on the voting result. Several experimental settings have been considered for the evaluation of the system's performance [9].

Overall, the shortcomings of this solution are listed below [9]:

1. It is based on features and if the dataset changes, the system must be rebuilt.
2. It cannot be applied in a real-time operation mode since it is based on a number of features. These features cannot be easily reassigned.
3. Data cannot be divided into blocks and distributed processing is not possible to accelerate the processing operation.
4. It has tried to solve the IDS Big Data problem and did not use Big Data analytics technology.

In another study [10], the authors suggest feature selection (FS) based on Ant Colony Optimization (ACO) for IDS optimization. The ACO algorithm depends on a computational model inspired by real ant colonies and the method they function. The FS process identifies which features are more discriminative than the others. This has the benefit of generally improving system performance by eliminating irrelevant features. The shortcomings of this method include [10]:

1. It is based on a computation algorithm to choose the best and most affected feature for IDS output. This presents a difficulty when running voluminous data in real-time mode.
2. It is time consuming for large data sets.
3. There is no parallelism in its operation.
4. It did not use any Big Data processing technology.

A new model was proposed for the detection of unknown attacks based on Big Data analytics while extracting information from various sources [11][12]. Architecture based on Big Data for large-scale security monitoring has also been proposed [13]. The major drawbacks in these studies are that the systems did not consider accuracy and efficiency and only limited analysis was performed. Although these systems propose some theoretical models, frameworks, and architecture, they lack practical implementation.

Another model has been proposed to address the problem of Big Data and IDS. This model is based on distributed processing with data reduction techniques such as feature selection. Distributed processing is verified by using Big Data analytics technology such as Hadoop and MapReduce. Hadoop has been implemented in a real-time intrusion detection system for ultra-high-speed big data environment [14]. IDS based

on Hadoop implementation and feature selection was suggested as well as architecture with four layers for IDS operations. In the study[14], the network traffic is processed in parallel based on MapReduce method and Hadoop Distributed File System (HDFS).

Feature selection techniques such as forward selection ranking (FSR) and backward elimination ranking (BER) were implemented to select the best and most weighted features. Machine learning (ML) algorithms such as naïve Bayes, support vector machine, random forest, J48, and REPTree were used for decision making in IDS. The authors suggested that IDS is able to store decisions about an appropriate flow in its list for In-Memory intrusions, that are employed by the filtration server for detecting the intruder's traffic. Shortcomings of this investigation are:

1. The system is implemented on a single node Hadoop, which reduces the benefits of using Hadoop. It also does not provide distributed processing.
2. The system depends on only nine out of 41 features in the standard IDS evolution data set. This contradicts with real-time operation.
3. How to preprocess voluminous data in Hadoop is not covered.
4. The data processing time and its improvement based on MapReduce or other methods are not discussed.

It is difficult to design a specific security mechanism for IoT, as this environment is heterogeneous, fragmented and not supportive of interoperability. Some solutions for the enhancement of IoT security have been developed. These methods were applied for data confidentiality and authentication, access control toward IoT, and privacy among users and things. However, in spite of these mechanisms, IoT networks still suffer from attacks. IDS could be used in IoT to ensure security in the face of a variety of attacks.

With regards to the development of IDS technology for traditional networks, current solutions are inefficient for IoT, as they are not flexible enough toward the complicated and heterogeneous IoT network. Characteristics like devices with restrained resources, network architecture, particular protocol stacks, and standards describe the necessity for the improvement of IDS for IoT [15].

IV. Proposed model "big data analytics with data reduction technique"

Figure 2 shows the proposed architecture of intrusion detection system based on Hadoop platform. Hadoop platform proves the distributed processing multiple nodes to detect attacks in real time.

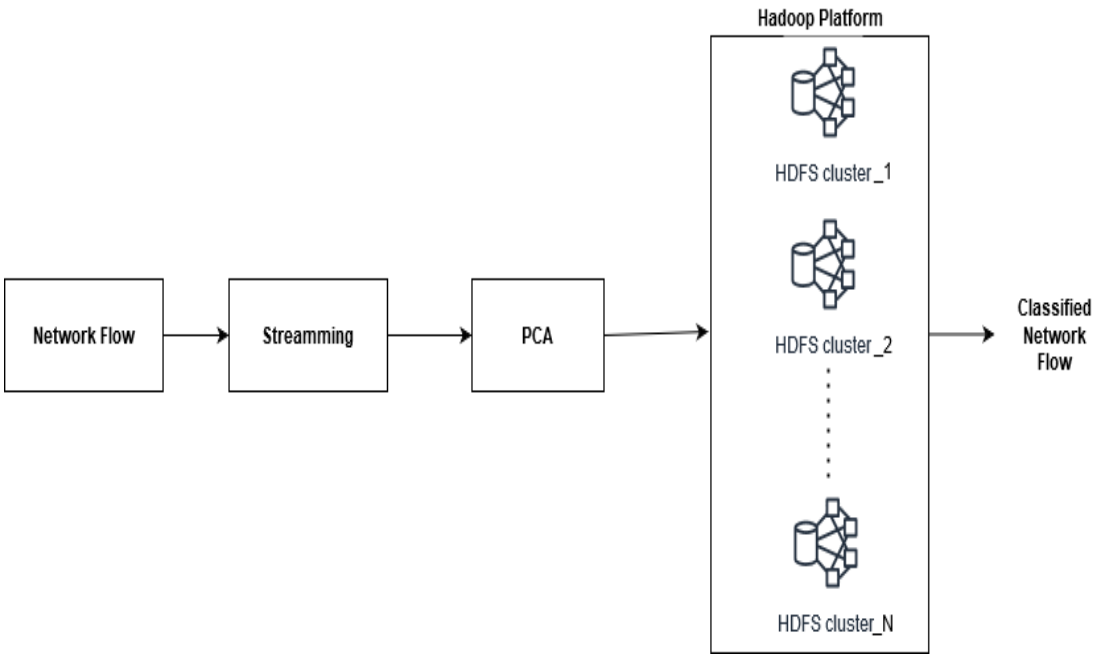

Figure 2: Proposed model: Distributed processing with features reduction intrusion detection model.

Hadoop platform contains HDFS clusters which divides the data into some blocks and process them through MapReduce method and one of machine learning algorithm to classify the network traffic. Principle Component Analysis (PCA) used for reducing the number of features that enter to the Hadoop platform. Principal component analysis (PCA) is a technique for reducing the dimensionality of such datasets, increasing interpretability but at the same time minimizing information loss. It does so by creating new uncorrelated variables that successively maximize variance.

V. Conclusions

Network's security faces Big Data challenges more often as time progresses and especially more so for larger private and government organizations. These trends of Big Data challenges will continue as a multitude of more heterogeneous sources are analyzed. The goal of Hadoop platform in intrusion detection system is to obtain actionable intelligence in real time. Although Big Data analytics have significant promise, there are a number of challenges that must be overcome to realize its true potential. PCA can be used for reduction the amount of data which are analyzed by the IDS.

VI. References

[1]. Kamburugamuve, S., Fox, G., Leake, D., & Qiu, J. (2013). Survey of distributed stream processing for large stream sources. Grids Ucs Indiana Edu, 2, 1-16..

[2]. Butun, I., Morgera, S. D., & Sankar, R. (2013). A survey of intrusion detection systems in wireless sensor networks. IEEE communications surveys & tutorials, 16(1), 266-282.

[3]. de Sá Silva, L., dos Santos, A. C. F., Mancilha, T. D., da Silva, J. D. S., & Montes, A. (2008). Detecting attack signatures in the real network traffic with ANNIDA. Expert Systems with Applications, 34(4), 2326-2333.

[4]. Patcha, A., & Park, J. M. (2007). An overview of anomaly detection techniques: Existing solutions and latest technological trends. Computer networks, 51(12), 3448-3470.

[5]. Manikopoulos, C., & Papavassiliou, S. (2002). Network intrusion and fault detection: a statistical anomaly approach. IEEE Communications Magazine, 40(10), 76-82.

[6]. Mukkamala, S., Sung, A. H., & Abraham, A. (2004, May). Modeling intrusion detection systems using linear genetic programming approach. In International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (pp. 633-642). Springer, Berlin, Heidelberg.

[7]. Mukkamala, S., Sung, A. H., & Abraham, A. (2003). Intrusion detection using ensemble of soft computing paradigms. In Intelligent systems design and applications (pp. 239-248). Springer, Berlin, Heidelberg.

[8]. Chavan, S., Shah, K., Dave, N., Mukherjee, S., Abraham, A., & Sanyal, S. (2004, April). Adaptive neuro-fuzzy intrusion detection systems. In International Conference on Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004. (Vol. 1, pp. 70-74). IEEE.

[9]. Al-Jarrah, O. Y., Siddiqui, A., Elsalamouny, M., Yoo, P. D., Muhaidat, S., & Kim, K. (2014, June). Machine-learning-based feature selection techniques for large-scale network intrusion detection. In 2014 IEEE 34th international conference on distributed computing systems workshops (ICDCSW) (pp. 177-181). IEEE.

[10]. Aghdam, M. H., & Kabiri, P. (2016). Feature selection for intrusion detection system using ant colony optimization. Int. J. Netw. Secur., 18(3), 420-432.

[11]. Ahn, S. H., Kim, N. U., & Chung, T. M. (2014, February). Big data analysis system concept for detecting unknown attacks. In 16th International Conference on Advanced Communication Technology (pp. 269-272). IEEE.

[12]. Jangla, G., & Amne, D. (2015). Big Data System for Unauthorized Attacks Detection. Int. J. Sci. Eng. Technol. Res, 4(55), 11800-11803.

[13]. Marchal, S., Jiang, X., State, R., & Engel, T. (2014, June). A big data architecture for large scale security monitoring. In 2014 IEEE International Congress on Big Data (pp. 56-63). IEEE.

[14]. Rathore, M. M., Ahmad, A., & Paul, A. (2016). Real time intrusion detection system for ultra-high-speed big data environments. The Journal of Supercomputing, 72(9), 3489-3510.
Ahmed, H. I., Nasr, A. A., Abdel-Mageid, S. M., & Aslan, H. K. (2021). DADEM: Distributed Attack Detection Model Based on Big Data Analytics for the Enhancement of the Security of Internet of Things (IoT). International Journal of Ambient Computing and Intelligence (IJACI), 12(1), 114-139.