Machine Learning Techniques for Anomaly Detection in High-Speed Big Data Networks

Vanita Dnyandev Jadhav

SVERI's College of Engineering, Pandharpur, Solapur University,

Maharashtra, India.

vdjadhav@coe.sveri.ac.in

Abstract: *Network has brought convenience to the world by allowing fast transformation of data, but it also exposes a number of vulnerabilities. With anomaly detection systems, the outliers of packets can be detected and computers are prevented from attacks. Some anomaly detection systems found in literature are based on data mining methods. Recently, as machine learning becomes a popular area of research, we propose using machine learning as the model for anomaly detection in this project , Anomaly detection systems, also known as intrusion detection systems (IDSs), continuously monitor network traffic aiming to identify malicious actions. Extensive research has been conducted to build efficient IDSs emphasizing two essential characteristics. The first is concerned with finding optimal feature selection, while another deals with employing robust classification schemes. However, the advent of big data concepts in anomaly detection domain and the appearance of sophisticated network attacks in the modern era require some fundamental methodological revisions to develop IDSs. Therefore, we first identify two more significant characteristics in addition to the ones mentioned above. These refer to the need for employing specialized big data processing frameworks and utilizing appropriate datasets for validating system's performance, which is largely overlooked in existing studies.*

Keywords: *anomaly detection, big data networks etc.*

1. INTRODUCTION

The Internet is evolving and it has revolutionized the world since the World Wide Web was invented. The usage of the Internet has become necessary in various areas. Through

the Internet, we are able to gain access to remote hosts, retrieve data and operate on the hosts. This simplifies our day-to-day life, but without appropriate security measures, it is likely that the systems would be compromised, causing individuals and companies suffering from great loss. Intruders may gain unauthorized privileges, or simply overload the server to make it unavailable. Both of these may incur great loss for the system owners.

In order to protect the computers from being hacked, intrusion detection systems (IDS) can be installed. Some common open source IDS are Snort [1] and Suricata [2]. With IDS installed, whenever a system encounters unauthorized access, it can respond by refusing such access request. Moreover, it can generate alerts for human to inspect if there is any system defect. Anomaly detection is a significant issue in computer networks. The advances in high-speed big data networks and the concomitant rise in network attacks require fundamental methodological revisions to develop efficient NIDSs. Based on this need, we first identified four vital characteristics that form the basis to devise a comprehensive network anomaly detection system. The first two specifically deal with implementing machine learning concepts and greatly affect the performance of the system, namely feature selection and the employment of classification schemes. The other two characteristics combat the challenges introduced by large-scale networks and sophisticated network attacks, namely utilizing specialized big data computing engines and obtaining contemporary workloads to conduct performance evaluations of the proposed systems. Building on the principles and issues analyzed in this paper, we proposed an efficient intrusion detection system that comprehensively follows the identified characteristics with an emphasis to handle big data problems in high speed networks. This is because anomaly detection in big data networks is particularly crucial due to the volume, velocity, variety, and veracity of the datasets. Thus, the proposed system incorporated a powerful BSP computing engine, which is capable of handling large volume of network traffic in real-time environments.

2.REVIEW OF LITERATURE SURVEY

Intrusion detection systems have been developed for over three decades, whereas the notion of big data trend in network security is relatively new. In order to establish a better understanding, in this section, we review some important concepts of network intrusion detection, related work and associated challenges that form the basis for motivation behind utilizing BSP-based machine learning computing in this area

The possibility of automatic intrusion detection was first put forward by James Anderson in 1980 in his classic paper, which states that a certain class of intruders or masqueraders who usually operate with stolen identities could probably be detected by their departures from the set norm for the original user. The proposed idea was basically to monitor security threats to audit trails. Afterward, the notion of checking all activities against a set security policy was introduced. Since Anderson’s paper, numerous theories, methods, and hardware and software frameworks have been presented in

the literature as well as in the form of commercial products. Consequently, security mechanisms such as firewalls, access control, and cryptography are now available and function as the first line of security defense with challenges involved from ever-evolving intrusion skills and techniques

[18]. A firewall mainly protects the network resources by allowing and disallowing certain types of access on the basis of a configured security policy. An access control is usually deployed for authentication purposes, whereas cryptography is used to achieve secure communication. These traditional defensive techniques have several limitations in fully protecting networks and systems from increasingly sophisticated attacks and malware. Moreover, most systems built on such techniques suffer from high false positive and false negative detection rates and also lack the ability to continuously adapt with the changing malicious behaviors . IDSs overcome certain limitations and provide a better security solution by protecting the network from both internal and external attacks.

Data mining techniques and machine learning algorithms can be applied to intrusion detection systems, and these techniques have been extensively studied in the past decade. Clustering and classification are some techniques used in IDS. Münz, Li & Carle proposed an anomaly detection system using k-means algorithm which combines both classification and outlier detection [5]. In [6, 7], the authors combined the k-means clustering with naïve Bayes classification. In [8], the authors further utilized the result

from k-means clustering as new features for naïve Bayes classifier. In [9], naïve Bayes classifier is combined with decision tree algorithms. The above mentioned methods operate on network features only, namely, the connection records. To take payload data from packets into consideration, we need some other techniques. PAYL is a histogram-based classification method that takes payload data as input. It builds a histogram from the input, with frequency of each byte pattern being a bin (see Fig. 1Fig.), and compares the histogram built from the data with baseline [10]. Deep learning techniques can also be implemented to detect anomalies. Wang [11] built a system that can classify TCP packets according to their application layer protocols, and suggested that misclassified packets may be anomalous.

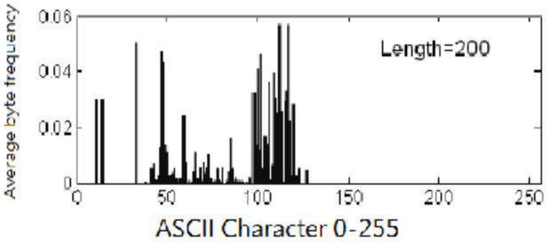

FIG. 1. Example of byte distribution for a 200-byte packet.

3.OBJECTIVES

The ultimate goal for this project is to build an anomaly detection system using Machine learning, and then study the effectiveness of different models and learning techniques.In order to achieve the goal following objectives are decided : -

I. Data Preprocessing.

II. feature ranking and selection using information gain (IG) and automated branch-and bound (ABB) algorithms, respectively.

III. Implementation of logistic regression (LR) and eXtreme gradient boosting (XGBoost) techniques for classifying network traffic.

IV. Employing an emerging and powerful big data computing framework based on bulk synchronous parallel (BSP) processing.

V. Attack recognition.

4.METHODOLOGY

Using machine learning for anomaly detection helps in enhancing the speed of detection. Implementing machine learning algorithms will provide companies with a simple yet effective approach for detecting and classifying these anomalies. Machine learning algorithms have the ability to learn from data and make predictions based on that data. Machine learning for anomaly detection includes techniques that provide a promising alternative for detection and classification of anomalies based on an initially large set of features. Machine learning techniques that can enable effective anomaly detection : Supervised Machine Learning for Anomaly Detection

This method requires a labeled training set that contains both normal and anomalous samples for constructing the predictive model. Theoretically, supervised methods are believed to provide better detection rate than unsupervised methods. The most common supervised algorithms are supervised neural networks, parameterization of training model,
learning, k-nearest neighbors, Bayesian networks and decision trees. K-nearest neighbor (k-NN) is one of the most conventional nonparametric techniques that are used in supervised learning for anomaly detection. It calculates the approximate distances between different points on the input vectors and then assigns the unlabeled point to the class of its K-nearest neighbors. The Bayesian network is another popular model that can encode probabilistic relationships among variables interest. This technique is generally used for anomaly detection in combination with statistical schemes. These supervised techniques have several advantages, including the capability of encoding interdependencies between variables and of predicting events, along with the ability to incorporate both prior knowledge and data.

Unsupervised Machine Learning for Anomaly Detection

These techniques do not require training data. They are based on two basic assumptions. First, they presume that most of the network connections are normal traffic and only a small amount of percentage is abnormal. Second, they anticipate that malicious traffic is statistically different from normal traffic. Based on these two assumptions, data groups of similar instances that appear frequently are assumed to be normal traffic and those data groups that are infrequent are considered to be malicious. The most common unsupervised algorithms are self-organizing maps (SOM), K-means, C-means, expectation-maximization meta-algorithm (EM), adaptive resonance theory (ART), and one-class support vector machine. One popular technique is the self-organizing map (SOM). The main objective of the SOM is to reduce the dimension of data visualization. Anomaly detection can effectively help in catching the fraud, discovering strange activity in large and complex Big Data sets. This can prove to be useful in areas such as banking security, natural sciences, medicine, and marketing, which are prone to malicious activities. With the machine, a learning organization can intensify search and increase effectiveness of their digital business initiatives.

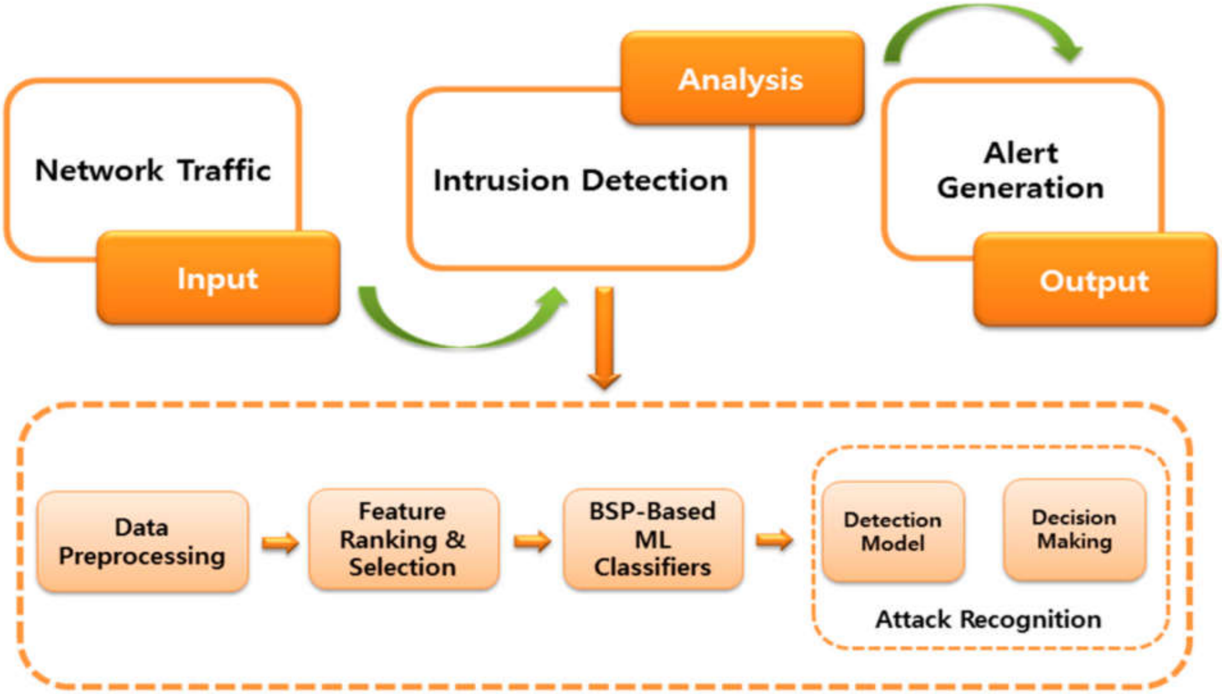

Figure. **The architecture of the proposed Intrusion detection system**

The architecture of the proposed framework is depicted in Figure. Basically it involves
three major phases: (i) input, which captures the network traffic and transmits it to the next phase; (ii) analysis, which is the most significant phase of the system where actual computation is performed and it consists of several components; and (iii) output, where alerts are being generated to identify malicious activities.

5. REFERENCES

[1] Kim, D.Y.; Jeong, Y.S.; Kim, S. Data-filtering system to avoid total data distortion in IoT networking. Symmetry 2019

.

[2] Harvinder Pal Singh Sasan and Meenakshi Sharma INTRUSION DETECTION USING FEATURE SELECTION AND MACHINE LEARNING ALGORITHM WITH MISUSE DETECTION 2016.\

[3] Cisco. (2016). *Snort*. Available: https://www.snort.org/

[4] Open Information Security Foundation. (2016). *Suricata*. Available: https://suricata-ids.org/

[5] A. Patcha and J.-M. Park, "An overview of anomaly detection techniques: Existing solutions and latest technological trends," *Computer networks,* vol. 51, pp. 3448-3470, 2007.

[6] S. Agrawal and J. Agrawal, "Survey on Anomaly Detection using Data Mining Techniques," *Procedia Computer Science,* vol. 60, pp. 708-713, 2015.

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

[8] R. Chitrakar and C. Huang, "Anomaly based Intrusion Detection using Hybrid Learning Approach of combining k-Medoids Clustering and Naïve Bayes Classification," in *2012 8th International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM)*, 2012, pp. 1-5.

[9] Z. Muda, W. Yassin, M. Sulaiman, and N. I. Udzir, "A K-Means and Naive Bayes learning approach for better intrusion detection," *Information Technology Journal,* vol. 10, pp. 648-655, 2011.

[10] S. Varuna and P. Natesan, "An integration of k-means clustering and naïve bayes classifier for Intrusion Detection," in *2015 3rd International Conference on Signal Processing, Communication and Networking (ICSCN)*, 2015, pp. 1-5.

[11] D. M. Farid, N. Harbi, and M. Z. Rahman, "Combining naive bayes and decision tree for adaptive intrusion detection," *arXiv preprint arXiv:1005.4496,* 2010.

[12] K. Wang and S. J. Stolfo, "Anomalous payload-based network intrusion detection," in *International Workshop on Recent Advances in Intrusion Detection*, 2004, pp. 203-222.

[13] Z. Wang, "The Applications of Deep Learning on Traffic Identification," presented at the Black Hat, USA, 2015.

[14] R. Lippmann, J. W. Haines, D. J. Fried, J. Korba, and K. Das, "The 1999 DARPA off-line intrusion detection evaluation," *Computer networks,* vol. 34, pp. 579-595, 2000.

[15] F. Chollet. (2016). *Keras Documentation*. Available: https://keras.io/