IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 12, December 2017

Survey of Random Forest Based Network Anomaly Detection Systems

Rashmi H Roplekar[1], Prof. N. V. Buradkar[2]

ME Student, Department Of Information Technology, PICT, Pune, India [1]

Department Of Information Technology, PICT, Pune, India [2]

Abstract: Network intrusion poses a serious threat to the security of financial and all other systems. The main objective of any online security system is to provide protection against malicious intentions of a user. The techniques used by intruders are bound to change and every day new methods of attacks on the network are being faced by all the systems on the net. One method to gain reliable security against unknown intrusions is to use Anomaly Detection Systems. Many existing intrusion detection systems are Rules Based, which have limitations when new intrusions appear. The proposed work intends to provide a system which detects network anomalies using machine learning (ML). The proposed system intends to improve the accuracy of anomaly detection as compared to the existing systems.

Keywords: Machine Learning, Intrusion Detection, Anomaly Detection.

I. INTRODUCTION

Machine Learning is branch of Artificial Intelligence which uses existing data to train the machine to identify and classify new data coming in a system. Machine learning is being applied in many fields like Natural Language Processing, Medical Diagnosis, Search Engines, Financial Market Analysis, Computer Vision (Object Recognition) Etc. With the growth of network based services and important information on network, network security is getting more and more crucial. Even though there is wide variety of security technologies like encryption, access control and intrusion detection to protect a network, there are still many undetected attacks and intrusions on the network. Anomaly detection technique identifies attacks based on a significant deviation from the normal network traffic patterns. Anomaly detection can prevent new attacks on the network. Network anomaly detection systems use various techniques to capture and analyze network traffic. It includes net flow analysis, analysis of various logs in the network, like applications' logs, firewall logs and different registry entries. Network attacks like lateral movement and escalation of privileges are more subtle and difficult to identify. To solve this problem, Anomaly detection proves helpful as it works on allowing only the normal traffic to enter the network. Many Intrusion Detection Systems which are rule based highly rely on the rules already identified. Looking at the amount of huge network traffic, it is difficult to code each rule in the system. It is also cumbersome to identify new attacks, code new rules for them and modify the existing system each time.

Machine learning is the technique which can be appropriately applied to anomaly detection for analyzing new patterns. To increase the quality of Anomaly Detection system, it needs to have low false positive detection rate. Machine learning will help reduce the false positive rate in anomaly detection by learning new anomalies and classifying incoming traffic between normal traffic and anomalous traffic with more accuracy. There exist different machine learning classification techniques. Some use statistical methods, some use probabilistic approach and some use Deep Learning approaches. There is a lack of inherent superiority in these techniques. This poses a challenge to select a suitable classification technique for the purpose of Network Anomaly Detection. Rest of the paper is organised as-section II is literature survey discusses about the previous study/ research made by some researchers and is helpful for understanding the history of the topic. Literature survey tells the history and root of problem statement how the problem statement has formed. Section III gives proposed system architecture and it explains components of the system. It also tells how these components are inter-connected and how the communicate with each other. Section IV concludes the study.

II. LITERATURE SURVEY

Novi Quadrianto and Zoubin Ghahramani [1] has proposed a novel technique for generating Random Forest by calculating a weighted group of predictive probabilities and then taking random samples of many trees from earlier distributions. This technique uses power likelihood instead of likelihood, which decides space spanned by the combination of trees. It is called safe – Bayesian because even though the underlying probabilistic model is wrong, it gives good predictive performance. They have proved using nine different datasets that the proposed technique of Safe – Bayesian Random Forest gives better results than classification algorithms like KNN (K nearest Neighbors),

SVM(Support Vector Machine), RF:gini (Random Forest learned with Gini impurity) , RF:Ent. (Random Forest learned with Information Gain), MCMC (Markov Chain Monte Carlo of Bayesian Decision Tree) and BART (a Bayesian version of boosted Decision Tree).

JiSoo am, Yangchi Chen, Melba M. Crawford, and Joydeep Ghosh [2] have proposed two approaches based on Random Forest to achieve improved generalization in the analysis of hyper spectral data, when the volume of training data is small. The new classifier is suggested with (1) Bagging of training samples and (2) Adaptive random subspace feature selection within binary hierarchical classifier. This causes the number of features selected at each node to be dependent on quantity of relevant training data. In this paper, they have discussed a random forest of binary classifiers for increasing diversity of hierarchical classifiers. The results show a comparison between RF-BHC and RF-CART methods and the RF-BHC methods are proved to be superior.

Dwarikanath Mahapatra [6] has proposed a method to use Random Forest for improved image segmentation. In this paper, he has discussed how Random Forest learns discriminative features and provides with quantification of importance of these features. This quantified importance of features is used to design a strategy to select features effectively for classification. Different image features like intensity, curvatures etc. are combined differently based on the cost function. In this paper, he has tested this segmentation method on many medical images and the results show improved segmentation. The paper suggests an effective technique for improved image segmentation by exploiting the knowledge from the training process of the RF classifier.

Jiong Zhang, Mohammad Zulkernine, and Anwar Haque [3] have proposed a Network Intrusion Detection system based on Random Forest Algorithm. Under Intrusion Detection, they have discussed Misuse detection, Anomaly detection and Hybrid frameworks using data mining by Random Forest. In misuse detection, intrusion patterns are built by Random Forest using training data. Misuse is detected by matching the network traffic with the detected patterns. New intrusions are detected in Anomaly detection by the outlier detection technique of Random Forest. For this, patterns of network services are built by Random Forest and then outliers related to these patterns are recognized. The results achieved using KDD99 dataset show that all the three proposed systems show improvements like better detection rate, lower false positive rate and improved overall performance.

Taeshik Shon, Jongsub Moon [4] have suggested a network anomaly detection system using machine learning techniques. They have proposed a combination of soft-margin SVM (Supervised) and one class SVM (Unsupervised) which gives benefits of both like unsupervised learning and low false alarm rate of supervised learning. They have also implemented additional techniques to improve the performance like using Self-Organized feature maps (SOFM) for unsupervised learning of SVM. Second technique is Passive TCP/IP fingerprint scheme for filtering incomplete network traffic which violates TCP/IP standard. The third technique is to use Genetic Algorithm for selecting features to extract optimized information from raw internet packets. The forth technique is to use temporal relationship of packet flow during data processing. All these techniques together prove to provide better performance when compared with real world IDS systems.

Jerry Murphree [8] has analyzed different machine learning techniques for detecting anomalies in large systems in this paper. Large systems include diagnostic and prognostic healthcare systems with multiple devices and network traffic monitoring systems. The data in such systems is having high dimensionality. To detect anomalies in such high dimension data, different machine learning techniques, supervised and unsupervised are studied to select a properly suitable anomaly detector. Different anomaly detection approaches are also discussed, like discretizing continuous range data and then label each discretized section as normal or anomalous. Finally this paper proposes the use of Neural Networks as the best machine learning technique for Anomaly detection in large systems.

Francesco Palmieri and Ugo Fiore [7] have discussed non-linear analysis method of Network Anomaly Detection in this paper. This work focuses on non stationary properties and “hidden” recurrence patterns appearing in the aggregated traffic flow. They have used recurrence quantification analysis, a non linear technique popularly used in hidden dynamics and time correlations of statistical time series. The paper discusses in depth the techniques of Recurrence Quantification Analysis and SVM for event classification. This paper concludes that non-linear techniques like RAQ can prove helpful in getting knowledge about hidden statistical nature of network traffic and when used with SVM machine learning, can be reliably used for Network Anomaly detection.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 12, December 2017

Title	Publication & Year	Author	Merit	Demerit
A Very Simple Safe-Bayesian Random Forest	IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 99, NO. 99, SEPTEMBER 2009	Novi Quadrianto and Zoubin Ghahramani	Uses earlier distribution weight age of predictive probabilities and then decides the power likelihood that decides the space spanned by the decision trees.	-
Investigation of the Random Forest Framework for Classification of Hyper spectral Data	IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, VOL. 43, NO. 3, MARCH 2005	JiSoo Ham, Yangchi Chen, Melba M. Crawford, and Joydeep Ghosh	It presents the capabilities of improving generalization in analysis of hyper spectral data when volume of training data is small in size.	-
Random-Forests-Based Network Intrusion Detection Systems	IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS, VOL. 38, NO. 5, SEPTEMBER 2008	Jiong Zhang, Mohammad Zulkernine, and Anwar Haque	Presents the use of Random forest in misuse detection and Anomaly Detection sections of network Intrusion Detection system.	Some intrusions with high degree of similarity cannot be detected as Anomalies by the outlier detection algorithm.
A hybrid machine learning approach to network anomaly detection	Information Sciences 177 (2007) 3799–3821	Taeshik Shon , Jongsub Moon	It proposes a hybrid SVM for network anomaly detection. It also uses additional techniques like SOFM, TCP/IP fingerprints, Genetic Algorithms etc. to improve the accuracy of Anomaly detection.	Needs to construct more realistic profiles of normal packets to better classify novel attacks.

III. PROPOSED SYSTEM ARCHITECTURE

Every intrusion detection system aims to provide zero day attack protection to the environment it protects. Outlier or Anomaly detection is the technique which has the potential to address this issue. In the last decade, researchers have developed systems which are able to provide good quality Anomaly Detection. Researchers are able to achieve good detection accuracy in the same, but the False Positive rate of these systems is still a major problem. Very less research has been done on the topic of Anomaly Detection using Random Forest algorithm. So the proposed system tries to reduce the false positive rate of the Anomaly detection by refining the tuning of parameters in Random Forest algorithm.

The proposed system architecture consists of following components:

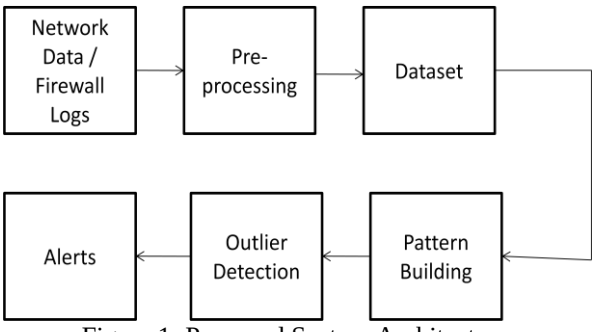

Figure 1: Proposed System Architecture

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 12, December 2017

A. Network Traffic (data from net flow packets or firewall logs)
B. Pre-processing (Attribute selection algorithm)
C. Dataset is the output of pre processing stage
D. Pattern Building (Detects network patterns based on services in the data)
E. Outlier Detection (Outliers are detected from the input traffic data based on the built patterns)
F. Alerts (On detection of Anomalous traffic, alerts are shown to the system administrator)

A. Network Data
Network traffic is collected from different sources like Net flow Packets or Firewall logs. The system accepts this data in the form of files.

B. Pre-processing
In this section, a dataset is constructed by using appropriate fields from the input traffic data. This task of feature selection is performed using Python libraries and files.

C. Dataset
It is the outcome of the previous section; Dataset is in the form of flat files.

D. Pattern Building
In this section, we build the patterns of the input network traffic data. Network traffic can be categorised by services. Each network service has its own pattern. Patterns are built using Random Forest algorithm.
As it is a supervised algorithm, it needs labelled dataset. We also need to optimise the parameters like number of features and number of trees related to the Random Forest algorithm.

E. Outlier Detection
Outlier is detected by finding an activity that deviates significantly from the other activities in the same network service. Random Forest uses proximities to find outliers whose proximity to all other cases in the entire dataset is very small. Outlierness also has a degree.

F. Alerts
After detection of outliers, an alert is displayed to the administrator. Alerts can be displayed in different ways like in local applications or on web pages.

IV. CONCLUSION

Aim of any Intrusion Detection system is to provide protection to all types of attacks including the Zero day attacks. Anomaly detection is the technique that promises to provide protection against zero day attacks. This paper focuses on how the accuracy of Anomaly detection can be improved, Machine Learning techniques are suitable for handling new types of attacks, and hence this paper proposes the use of Machine Learning for Network Anomaly Detection. In this study, Random Forest algorithm of machine learning is used because of its accuracy and ability to provide proximity functionality. Proximity function is used to detect outliers in the input data. This study can be used as a guide to implement a Network Anomaly detection system. The scope of this study is not only to increase accuracy of Anomaly detection, but also to reduce the false positive rate in the system. However, it can be concluded that this study can prove helpful in improving overall functioning of Network Anomaly Detection systems.

REFERENCES

[1] Quadrianto, N. and Ghahramani, Z. (2015). A Very Simple Safe-Bayesian Random Forest. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(6), pp.1297-1303.
[2] Ham, J., Yangchi Chen, Crawford, M. and Ghosh, J. (2005). Investigation of the random forest framework for classification of hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, 43(3), pp.492-501.
[3] Jiong Zhang, Zulkernine, M. and Haque, A. (2008). Random-Forests-Based Network Intrusion Detection Systems. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(5), pp.649-659.
[4] Shon, T. and Moon, J. (2007). A hybrid machine learning approach to network anomaly detection. *Information Sciences*, 177(18), pp.3799-3821.
[5] Callegari, C., Giordano, S. and Pagano, M. (2017). An information-theoretic method for the detection of anomalies in network traffic. *Computers & Security*, 70, pp.351-365.
[6] Mahapatra, D. (2014). Analyzing Training Information From Random Forests for Improved Image Segmentation. *IEEE Transactions on Image Processing*, 23(4), pp.1504-1512.
[7] Palmieri, F. and Fiore, U. (2010). Network anomaly detection through nonlinear analysis. *Computers & Security*, 29(7), pp.737-755.
[8] Murphree, J. (2016). Machine learning anomaly detection in large systems. *2016 IEEE AUTOTESTCON*.
[9] Casas, P., D'Alconzo, A., Fiadino, P. and Callegari, C. (2016). Detecting and diagnosing anomalies in cellular networks using Random Neural Networks. *2016 International Wireless Communications and Mobile Computing Conference (IWCMC)*.
[10] Hajji, H. (2005). Statistical Analysis of Network Traffic for Adaptive Faults Detection. *IEEE Transactions on Neural Networks*, 16(5), pp.1053-1063.
[11] Meneganti, M., Saviello, F. and Tagliaferri, R. (1998). Fuzzy neural networks for classification and detection of anomalies. *IEEE Transactions on Neural Networks*, 9(5), pp.848-861.
[12] Rajpal, R., Kaur, S. and Kaur, R. (2016). Improving detection rate using misuse detection and machine learning. *2016 SAI Computing Conference (SAI)*.