Implementation of a hybrid machine learning technique for network intrusion detection

MAXWELL EICHIE, University of Cincinnati, United States

ABSTRACT

The current rapid growth in the computer and internet development has ushered in numerous cybersecurity challenges which are constantly evolving with time. The current cybersecurity solutions are no longer optimal in tackling these emerging cyber threats and attacks. This paper proposes the creation of a cybersecurity dataset to be used for a hybrid machine learning (ML) approach of supervised and unsupervised learning for an effective intrusion detection system. The proposed model entails a five-stage process which starts at the setup of a simulated network environment of network attacks to generate a dataset which feeds into the data normalization stage and then to data dimension reduction stage using the principal component analysis as a feature extraction method after which the data of reduced dimension is clustered using the k-Means method to bring about a new data set with fewer features. This new dataset is afterward classified using the enhanced support vector machine (ESVM). The proposed model is expected to provide a high-quality dataset and an efficient intrusion detection system in terms of intrusion detection accuracy of 99.5%, short train time of 5 seconds and a low false-positive rate of 0.4%.

Additional Key Words and Phrases: Intrusion detection, cybersecurity, machine learning

ACM Reference Format: Maxwell Eichie. 2020. Implementation of a hybrid machine learning technique for network intrusion detection. In *IT Research Symposium '20: School of Information Technology IT Research Symposium, April 14, 2020, Cincinnati, OH*, USA, 6 pages.

1 Introduction

Machine learning is a promising solution to address the security challenges in the technology field today. It is a type of artificial intelligence that makes use of different learning algorithms to train devices without explicit programming. Its use of mathematical models, data sets, dynamic (regular and irregular) data behavioral patterns and learning algorithms that require no human intervention, makes it applicable for the defense against new threats. A few challenges have been identified which are limitation in computational resources and the need for new data sets required for learning [1]. Machine learning has been applied in recent times to solve cybersecurity challenges, most notable of these challenges are software application, system and network vulnerabilities using network security systems in the form of firewalls, antivirus software and network intrusion detection systems (NIDS). An IDS is a security system designed to detect attacks in the form of malicious activities by identifying the intrusions and prevent host systems from getting compromised in a network environment[2]. Three types of intrusion detection techniques are available which are signature-based also called misuse-based, anomaly-based and hybrid[3]. The signature-based technique uses the predefined set of signatures or rules to detect known attacks. Variation from the known attack would result in no detection of an attack and as such the system is unable to detect new and zero-day attacks. A new signature must be implemented for every new attack which makes this technique time and resource-

IT Research Symposium '20, April 14, 2020, Cincinnati, OH

intensive. The anomaly-based technique studies aggregate data relating to the regular behavior of statistical analysis, any deviation from this pattern of behavior indicates an anomaly. It can detect systems, users and applications in a network over a long period and create a pattern of a zero-day threat since behavioral patterns are particular to individual systems and it is very difficult for a threat to model the behavioral pattern that conforms to the system pattern. It records a very high alarm rate than the signature-based because of the strict non-correlation framework built into it. The hybrid detection technique combines the signature-based and anomaly-based technique for intrusion detection to reduce the false alarm rate, intrusion rate and accuracy of intrusion detection[4]. In this paper, we propose a way to generate a new data set and apply the hybrid machine learning approach for intrusion detection.

The rest of this paper is organized as follows: Section 2 focuses on research background, section 3 discusses gives a summary of related work for ML in cybersecurity, section 4 discusses the research question, section 5 discusses the methodology and sections 6 discusses the expected result and conclusion.

2 Background

2.1 Machine Learning and Techniques

Machine learning algorithms have broad categorizations into supervised, unsupervised and semi-supervised learning[5]. Supervised learning makes known inputs and its corresponding outputs are available for learning which helps the machine to identify the output for other similar inputs. Most supervised learning algorithms which include K-Nearest neighbor (KNN), Support Vector Machine (SVM), Neural Networks and Bayesian have been proposed and used to develop intuitive security frameworks. For Unsupervised learning, no outputs are given, only inputs are used in learning and based on these inputs, the system classifies the dataset into clusters after which a new input can be classified into the right group. Algorithms like Principal Component Analysis (PCA) and k-means Clustering have been used to develop brilliant security frameworks. The semi-supervised learning is the case in which a few inputs are known with most of the inputs being unknown. It is a learning model developed with a small amount of training data with known labels and a significant amount of data with no labels[6]. The Hybrid approach makes use of both a supervised learning and an unsupervised learning algorithm.

2.2 Support Vector Machine

Support Vector Machine (SVM) is one of the most used, robust and accurate methods in machine learning. It consists of a support vector classification (SVC) and a support vector regression (SVR). It is a binary classification technique that classifies input instances into two classes. It also supports the multi-class classification [7]. SVM maps the input vector into a higher dimensional feature space. There are two SVM methods which are majorly used, the soft margin SVM and the one-class SVM. A better model for SVM is the enhanced SVM which inherits the advantages of both SVM approaches in terms of faster processing performance, a higher detection rate, unsupervised learning feature of one-class SVM and unlabeled capability.

2.3 K – Means

K-means is one of the well-known data mining clustering algorithms based on centroids and has been used to detect abnormal network user behavior in network traffic. This algorithm positions K centroids in an N-dimensional plane and using distance measurements, it repositions the centroids iteratively in such a way that the entries of the dataset are absorbed by the closest centroid, classifying the entries. This algorithm is scalable to large volume datasets[8].

2.4 Confusion Matrix

The confusion matrix is a performance metric used in the area of ML [9]. The confusion matrix is a table that describes the classification of experimental results in detail as being correct or incorrectly classified. The metric makes use of four key criteria being true positives (TP), true negatives (TN), false negatives (FN) and false positives (FP). For binary classification, a 2 * 2 matrix is used and for n classification, an n * n matrix is used. For binary classifications, it is divided into four categories:
True Positive (TP): Positive samples correctly classified by the model
False Negative (FN): A positive sample that is misclassified by the model
False Positive (FP): A negative sample that is misclassified by the model
True Negative (TN): Negative samples correctly classified by the model

2.5 Feature Selection Technique

Principal component analysis (PCA) is a technique that is used for classification and compression of data set dimensionality by extracting a new feature set that's smaller than the initial one. The new extracted feature set includes most of the sample data information. PCA helps to identify the patterns in data in a way that highlights their similarity and differences by the feature reduction process. The new features called principal components (PCs) are ordered by the amount of total information retained and they are uncorrelated[10].

2.6 Network Security Data Set

Data forms a fundamental component of systems and network security. A dataset is required to perform relevant evaluations and detection techniques for machine learning. Network security data is extracted directly from enterprise network security log events and user/system activity using special software tools or appliances to capture network packets.

2.7 Scapy

Scapy is a program written in python to manipulate network packets. It can be used to send, dissect, sniff and forge network packets[11]. It is used to construct and decrypt packets of a variety of protocols, send and capture these packets, match packet requests, replies and numerous other capabilities. Other tasks including tracerouting, scanning, unit testing, probing, attacking and discovery of networks can be done by Scapy. It also allows injection of packets into the network, sends invalid frames, does ARP cache, poisoning, VLAN hopping can add value to any field, stack them as required and send the packets. It comes as an interpreter and not a shell command program.

2.8 Wireshark

Wireshark happens to be the most popular available free sniffing tool. It gains credit for this because of its simple and graphical interface as well as powerful capturing and filtering options. It can scan ethernet, Wi-Fi, Bluetooth or any sort of network even in monitor mode[12]. It is globally used in the industry as a sniffer because of its ability to go up to the depths of bits of packets. It also separates the different network protocol packets by highlighting protocols with different color schemes like green for HTTP, blue for DNS amidst others.

3 Related Work

Much work has been done in the area of intrusion detection in computer networks using hybrid ML techniques and the multi-level hybrid model using existing public datasets. The most recent works are discussed below. Guo C, et al, proposed a hybrid learning model for intrusion detection using distance sum-based support vector machine (DSSVM)[13]. The model used the sum of the distances between each data sample, cluster centers and the relationship of each data sample to several samples in the KDD 99 training data set. An SVM classifier is trained using the new distance sum-based feature

vectors. The results achieve competitive detection accuracy and high efficiency with the light challenge of additional computational cost is needed to transform the original data features.

Shon T, et al proposed a new hybrid approach to intrusion detection using enhanced SVM which combines the best features of one-class SVM and soft margin SVM[14]. In addition to this algorithm, a self-organized feature map (SOFM) is used to create a profile for normal packets in the DARPA dataset. Passive TCP/IP fingerprinting (PTF) as a packet filtering scheme was used to reject incomplete network traffic and the Genetic Algorithm as a feature extraction scheme to extract optimized information from internet packets. The results showed an over 87.74% detection accuracy

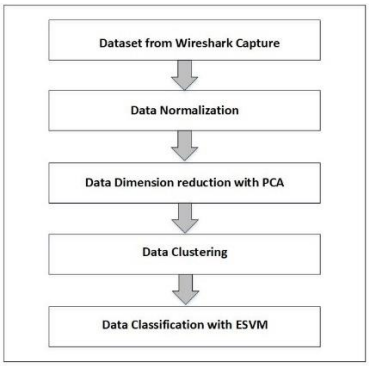
S. Varuna, et al, proposed a hybrid model that combines K-means clustering and naïve Bayes classifications[15]. The model made use of the relational distances between data samples to many centroids found by a clustering algorithm to form five clusters with four of the clusters representing the four different types of attack and one for normal traffic. The five distances from the five clusters to the centroid were considered the features and these features were given to naïve Bayes classifier for training and test. Results showed improved efficiency and amidst the computational overhead of the system.

4 Problem Statement

After review of the related works and trends, four key factors of an effective network intrusion detections system stand out in order of priority which is the type of training dataset, attack detection and precision rate, data training time and the algorithm's ability to handle a large volume of data. Existing works of literature have a trade-off on these factors and make use of the existing outdated public data set amidst evolving cyber-attacks. The most used datasets are the KDD Cup 99, NSL-KDD, ADFA and DARPA datasets. The current machine learning hybrid approaches using existing datasets do not stand a high chance of efficiency in intrusion detection in enterprise environments. There is the need to generate new datasets inclusive of recent attacks that would be used for training and testing.

5 Proposed Approach

The proposed solution is the generation of a new data set that would be used in a hybrid machine learning approach comprising of the Principal component analysis, K-means and, the enhanced support vector machine. Since the reduction of training time, as well as data volume, are also important factors in intrusion, we propose an experimental method to achieve this by setting up a lab environment in a staging process. Figure 1 shows the proposed intrusion detection system.



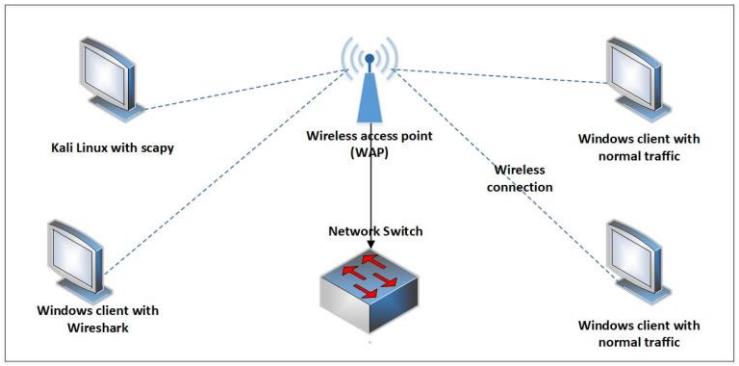


Figure 1: Proposed ML Approach **Figure 2: Simulated Local Area network of attack.**

Generation and description of Dataset: The proposed approach entails generating a training and test data set that would be used for unsupervised and supervised learning. This would be achieved by setting up a local network in a lab environment as seen in figure 2. The network would consist of a network switch, two windows machine that would generate normal traffic, a kali Linux machine with

Scapy installed to craft packets with different types of attack and a windows client machine with Wireshark installed to sniff packets on the network. All the machines would be connected to the network via a wireless connection. The wireless network ensures that all devices are all on the same collision domain. That way, Wireshark can sniff all the packets on the network across all devices. Five attacks including ARP poisoning, smurf, overlapping fragments, ping of death and TCP hijacking consisting of one thousand packets each are to be generated on the training dataset along with four thousand normal packets while the test dataset would have 200 data samples of all the training data attack and two new attacks consisting of buffer overflow and port sweep alongside two thousand normal packets bringing the total training dataset to 9,000 and test dataset to 3,400.

Data Normalization: A major requirement in the proposed methodology is data normalization which involves taking the dataset as input and producing the normalized data at the corresponding output. The proposed interval for normalization would be between zero and one. Data normalization would spread the data according to its highest possible value and then map them into a 0,1 interval. This phase would make it suitable for dimension reduction and classification.

Data Dimension Reduction: This feature extraction method is the third phase of the proposed method and It would be accomplished by using the principal component analysis method. The input of this section is the normalized data and its output would be the dimensionally reduced data with the corresponding scatter matrix. This matrix will be used to reduce the dimension of the test data. The PCA works by combining values of existing features in such a way that the obtained features include all initial features. It maps the data into space with less dimension.

Data Clustering: At this stage, the data of the reduced dimension is partitioned into the mean of the data sets making K clusters which are the new data features. We propose the data set be clustered into five features. At this stage, the original dataset is transformed into a new dataset.

Data Classification: This is stage five of the proposed method were data classification would be done using the enhanced SVM on the new dataset We propose an enhanced SVM method which has the added capability of one-class SVM and the soft margin SVM. The proposed SVM approach will have a higher detection rate and faster processing performance than the soft-margin SVM but also possess the unsupervised learning feature of one-class SVM with an added computational overhead.

Implementation: The proposed classification method would be implemented on a computer with a very high computational resource with an expected capacity of 32 GB RAM, and i9 CPU.

6 Conclusion

With the expectation that the proposed method would stand out on three major factors that drive the efficiency of intrusion detection systems which are detection and accuracy, training time, and false-positive rate, the expected results are a 99.5% detection accuracy, 5seconds training time and 0.4% false-positive rate. This research proposes the creation of a new dataset to be used for a hybrid machine learning approach with principal component analysis and the enhanced SVM that incorporates both the unsupervised soft-margin SVM and the supervised one-class SVM for intrusion detection. We proposed the use of PCA because of its perceived benefits for defined feature extraction, after which k-mean is used to cluster the data and finally the ESVM is chosen because of its ability to handle extremely data samples which are prevalent in real-world situations.

REFERENCES

[1] Mamdouh, Marwa, Mohamed AI Elrukhsi, and Ahmed Khattab. "Securing the internet of things and wireless sensor networks via machine learning: A survey." In 2018 International Conference on Computer and Applications (ICCA), pp. 215-218. IEEE, 2018.

[2] Kim, D., Shin, D., & Shin, D. (2018, August). Unauthorized Access Point Detection Using Machine Learning Algorithms for Information Protection. In 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE) (pp. 1876-1878). IEEE.

[3] Milenkoski, Aleksandar, Marco Vieira, Samuel Kounev, Alberto Avritzer, and Bryan D. Payne. "Evaluating computer intrusion detection systems: A survey of common practices." ACM Computing Surveys (CSUR) 48, no. 1 (2015): 1-41.

[4] Lou, Yan, and Jeffrey JP Tsai. "A framework for extrusion detection using machine learning." In 2008 11th IEEE International Symposium on Object and Component-Oriented Real-Time Distributed Computing (ISORC), pp. 83-88. IEEE, 2008.

[5] Xin, Yang, Lingshuang Kong, Zhi Liu, Yuling Chen, Yanmiao Li, Hongliang Zhu, Mingcheng Gao, Haixia Hou, and Chunhua Wang. "Machine learning and deep learning methods for cybersecurity." IEEE Access 6 (2018): 35365-35381.

[6] Hu, Wenjie, Yihua Liao, and V. Rao Vemuri. "Robust Support Vector Machines for Anomaly Detection in Computer Security." In ICMLA, pp. 168-174. 2003.

[7] Modi, Chirag N., and Kamatchi Acha. "Virtualization layer security challenges and intrusion detection/prevention systems in cloud computing: a comprehensive review." the Journal of Supercomputing 73, no. 3 (2017): 1192-1234.

[8] Prasad, M. S., A. Vinay Babu, and Mr K. Babu Rao. "An intrusion detection system architecture based on neural networks and genetic algorithms." International Journal of Computer Science and Management Research 2 (2013): 1344-1361.

[9] Bhattacharyya, Dhruba Kumar, and Jugal Kumar Kalita. Network anomaly detection: A machine learning perspective. Crc Press, 2013.

[10] P. Biondi, "Scapy, a powerful interactive packet manipulation program," Available at http://www.secdev.org/projects/scapy/

[11] Rohith, R., Minal Moharir, and G. Shobha. "SCAPY-A powerful interactive packet manipulation program." In 2018 International Conference on Networking, Embedded and Wireless Systems (ICNEWS), pp. 1-5. IEEE, 2018.

[12] Guo, Chun, Yajian Zhou, Yuan Ping, Zhongkun Zhang, Guole Liu, and Yixian Yang. "A distance sum-based hybrid method for intrusion detection." Applied intelligence 40, no. 1 (2014): 178-188.

[13] Shon, Taeshik, and Jongsub Moon. "A hybrid machine learning approach to network anomaly detection." Information Sciences 177, no. 18 (2007): 3799-3821.

[14] Meinel, Christoph, Mohammad Ghasemzadeh, and HamidReza Hemati. "A hybrid machine learning method for intrusion detection." International Journal of Engineering 29, no. 9 (2016): 1242-1246.

[15] Goyal, Piyush, and Anurag Goyal. "Comparative study of two most popular packet sniffing tools-Tcpdump and Wireshark." In 2017 9th International Conference on Computational Intelligence and Communication Networks (CICN), pp. 77-81. IEEE, 2017.