IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 21, Issue 3, Ser. IV (May - June 2019), PP 42-46
www.iosrjournals.org

Survey on Detecting Port Scan Attempts with Combined Analysis of Support Vector Machine and Deep Learning Algorithms

Mrs. Manasee Kurkure
Assistant Professor PCCOER, Ravet
Corresponding Author: Mrs. Manasee Kurkure

***Abstract**: Compared to the past security of networked systems has become a critical universal issue that influences individuals, enterprises and governments. The rate of attacks against networked systems has increased melodramatically, and the strategies used by the attackers are continuing to evolve. For example, the privacy of important information, security of stored data platforms, availability of knowledge etc. Depending on these problems, cyber terrorism is one of the most important issues in today's world. Cyber terror, which caused a lot of problems to individuals and institutions, has reached a level that could threaten public and country security by various groups such as criminal organizations, professional persons and cyber activists. Intrusion detection is one of the solutions against these attacks. A free and effective approach for designing Intrusion Detection Systems (IDS) is Machine Learning. In this study, deep learning and support vector machine (SVM) algorithms were used to detect port scan attempts based on the new CICIDS2017 dataset.*

***Keywords**: IDS, SVM, CICIDS2017, Cyber Terror, Deep Learning*

Date of Submission: 01-06-2019 Date of acceptance: 17-06-2019

I. Introduction

Network Intrusion Detection System (IDS) is a software-based application or a hardware device that is used to identify malicious behavior in the network [1,2]. Based on the detection technique, intrusion detection is classified into anomaly-based and signature-based. IDS developers employ various techniques for intrusion detection. Information security is the process of protecting information from unauthorized access, usage, disclosure, destruction, modification or damage. The terms ”Information security”, ”computer security” and ”information insurance” are often used interchangeably. These areas are related to each other and have common goals to provide availability, confidentiality, and integrity of information. Studies show that the first step of an attack is discovery [1]. Reconnaissance is made in order to get information about the system in this stage. Finding a list of open ports in a system provides very critical information for an attacker. For this reason, there are a lot of tools to identify open ports [2] such as antivirus and IDS.

One of these techniques is based on machine learning. Machine learning (ML) techniques can predict and detect threats before they result in major security incidents [3]. Classifying instances into two classes is called binary classification. On the other hand, multi-class classification refers to classifying instances into three or more classes. In this research, we adopt both classifications Information security is the process of protecting information from unauthorized access, usage, disclosure, destruction, modification or damage. The terms ”Information security”, ”computer security” and ”information insurance” are often used interchangeably.

These areas are related to each other and have common goals to provide availability, confidentiality, and integrity of information. Studies show that the first step of an attack is discovery [1]. Reconnaissance is made in order to get information about the system in this stage. Finding a list of open ports in a system provides very critical information for an attacker. For this reason, there are a lot of tools to identify open ports [2] such as antivirus and IDS.

II. Litrature Review

Sharafaldin et al. [4] used a Random Forest Regressor to determine the best set of features to detect each attack family. The authors examined the performance of these features with different algorithms that included K-Nearest Neighbor (KNN), Adaboost, Multi-Layer Perceptron (MLP), Naïve Bayes, Random Forest (RF), Iterative Dichotomiser 3 (ID3) and Quadratic Discriminant Analysis (QDA). The highest precision value was 0.98 with RF and ID3 [4]. The execution time (time to build the model) was 74.39 s. This is while the execution time for our proposed system using Random Forest is 21.52 s with a comparable processor.

Furthermore, our proposed intrusion detection system targets a combined detection process of all the attack families.

D. Aksu, S. U¨ stebay, M. A. Aydin, and T. Atmaca[09], There are different but limited studies based on the CICIDS2017 dataset. Some of them were discussed here. D.Aksu et al. showed performances of various machine learning algorithms detecting DDoS attacks based on the CICIDS2017 dataset in their previous work [13].The authors of [13] applied the Multi-Layer Perceptron (MLP) classifier algorithm and a Convolutional Neural Network (CNN) classifier that used the Packet CAPture (PCAP) file of CICIDS2017. The authors selected specified network packet header features for the purpose of their study. Conversely, in our paper, we used the corresponding profiles and the labeled flows for machine and deep learning purposes. According to [13], the results demonstrated that the payload classification algorithm was judged to be inferior to MLP. However, it showed significant ability to distinguish network intrusion from benign traffic with an average true positive rate of 94.5% and an average false positive rate of 4.68%.

The auther E. Biglar Beigi, H. Hadian Jazi,Machine [14] learning techniques have the ability to learn the normal and anomalous patterns automatically by training a dataset to predict an anomaly in network traffic. One important characteristic defining the effectiveness of machine learning techniques is the features extracted from raw data for classification and detection. Features are the important information extracted from raw data. The underlying factor in selecting the best features lies in a trade-off between detection accuracy and false alarm rates. The use of all features on the other hand will lead to a significant overhead and thus reducing the risk of removing important features. Although the importance of feature selection cannot be overlooked, intuitive understanding of the problem is mostly used in the selection of features [16].

The authors in [14] proposed a denial of service intrusion detection system that used the Fisher Score algorithm for features selection and Support Vector Machine (SVM), K-Nearest Neighbor (KNN) and Decision Tree (DT) as the classification algorithm. Their IDS achieved 99.7%, 57.76% and 99% success rates using SVM, KNN and DT, respectively. In contrast, our research proposes an IDS to detect all types of attacks embedded in CICIDS2017, and as shown in the confusion matrix results, achieves 100% accuracy for DDoS attacks using (PCA RF)Mc10 with UDBB.The authors in [15] used a distributed Deep Belief Network (DBN) as the the dimensionality reduction approach. The obtained features were then fed to a multi-layer ensemble SVM. The ensemble SVM was accomplished in an iterative reduce paradigm based on Spark (which is a general distributed in-memory computing framework developed at AMP Lab, UC Berkeley), to serve as a Real Time Cluster Computing Framework that can be used in big data analysis [16]. Their proposed approach achieved an F-measure value equal to 0.921.

III. Methods

1.1 CICIDS2017 Dataset

The CICIDS2017 dataset is used in our study. The dataset is developed by the Canadian Institute for Cyber Security and includes various common attack types. The CICIDS2017 dataset consists of realistic background traffic that represents the network events

produced by the abstract behavior of a total of 25 users. The users' profiles were determined to include specific protocols such as HTTP, HTTPS, FTP, SSH and email protocols. The developers used statistical metrics such as minimum, maximum, mean and standard deviation to encapsulate the network events into a set of certain features which include:

1. The distribution of the packet size
2. The number of packets per flow
3. The size of the payload
4. The request time distribution of the protocols
5. Certain patterns in the payload

Moreover, CICIDS2017 covers various attack scenarios that represent common attack families.
The attacks include Brute Force Attack, Heart Bleed Attack, Botnet, DoS Attack, Distributed DoS (DDoS) Attack , Web Attack, and Infiltration Attack.

1.2 SUPPORT VECTOR MACHINE

The SVM is already known as the best learning algorithm for binary classification. The SVM, originally a type of pattern classifier based on a statistical learning technique for classification and regression with a variety of kernel functions, has been successfully applied to a number of pattern recognition applications. Recently, it has also been applied to information security for intrusion detection. Support Vector Machine has become one of the popular techniques for anomaly intrusion detection due to their good generalization nature and the ability to overcome the curse of dimensionality .Another positive aspect of SVM is that it is useful for finding a global minimum of the actual risk using structural risk minimization, since it can generalize well with kernel tricks even in high-dimensional spaces under little training sample conditions. The SVM can select

appropriate setup parameters because it does not depend on traditional empirical risk such as neural networks [13]. One of the main advantage of using SVM for IDS is its speed, as the capability of detecting intrusions in real-time is very important. SVMs can learn a larger set of patterns and be able to scale better, because the classification complexity does not depend on the dimensionality of the feature space. SVMs also have the ability to update the training patterns dynamically whenever there is a new pattern during classification [14].

1.2.1 Limitation of Support Vector Machine in IDS

SVM is basically supervised machine learning method designed for binary classification. Using SVM in IDS domain has some limitation.SVM being a supervised machine learning method requires labelled information

for efficient learning. Pre existing knowledge is required for classification which may not be available all the time[13].SVM has the intrinsic structural limitation of the binary classifier i.e. it can only handle binary-class classification whereas intrusion detection requires multi-class classification [14].Although there are some improvements, the number of dimensions still affects the performance of SVM-based classifier [16]. SVM treats every feature of data equally. In real intrusion detection datasets, many features are redundant or less important .It would be better if feature weights during SVM training are considered [16]. Training of SVM is time-consuming for IDS domain and requires large dataset storage.Thus SVM is computationally expensive for resource-limited ad hoc network[12].Moreover SVM requires the processing of raw features for classification which increases the architecture complexity and decreases the accuracy of detecting intrusion

1.3 DEEP LEARNING

Deep learning is an improved machine learning technique for feature extraction, perception and learning of machines. Deep learning algorithms performs their operations using multiple consecutive layers. The layers are interlinked and each layer receives the output of the previous layer as input. It is a great advantage to use efficient algorithms for extracting hierarchical features that best represent data rather than manual features in deep learning methods [7], [8]. There are many application areas for Deep Learning, which covers such as Image Processing, Natural Language Processing, biomedical, Customer Relationship Management automation, Vehicle autonomous systems and others.

IV. Proposed System

As depicted in Figure 1, a proposed architecture consists of four components: CICIDS2017; data pre-processing module, DM method and security responses, The proposed algorithm as elaborated below.

1. Select the CICIDS2017 dataset
2. Preprocessing the normalized dataset
3. Split this dataset in to two phases
4. Creating the models of IDS and SVM.
5. Evaluate the Performance

In normalization, nonnumeric label features were converted into numeric forms. In addition, unrelated features such as Timestamp and some samples that have NaN, infinity and empty values were removed to detect port scan attempts by using the training data.

Consequently, the performances of the models were calculated. True Positive (TP), True Negative (TN), False Positive(FP) and False Negative (FN) statistics (Table II) are used for evaluation of model performances.

Table II can be explained in below items.

• TN : Actual Benign is classified as Benign.
• FP : Actual Benign is classified as Port Scan.
• FN : Actual Port Scan is classified as Benign.

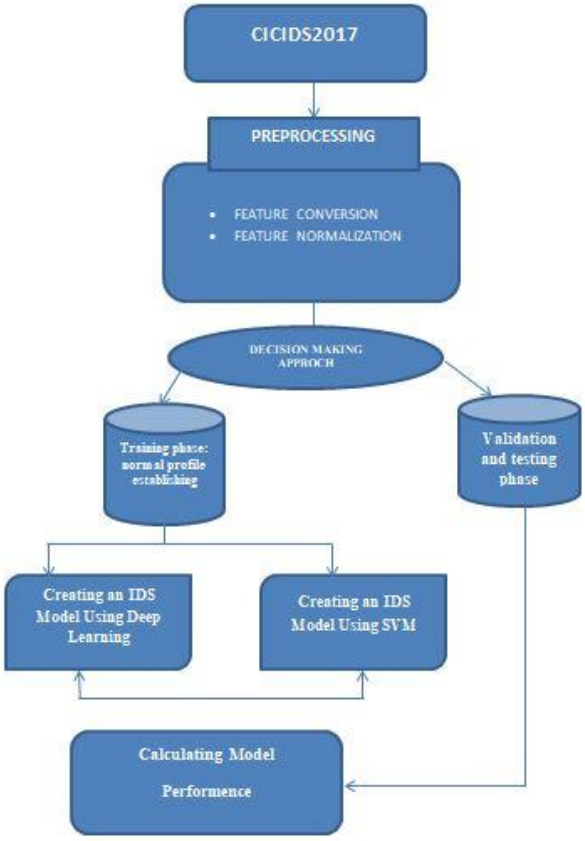

Fig 4.1. Proposed Architecture

V. Performance Evaluation Metrics

This study used various performance metrics to evaluate the performance of the proposed system, including False Alarm Rate (FAR), F-Measure [43], Detection Rate (DR), and Accuracy (Acc) as well as the processing time. The definitions of these metrics are provided below. The metrics are a function of True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN). (1) False Alarm Rate (FAR) is a common term which encompasses the number of normal instances incorrectly classified by the classifier as an attack, and can be estimated through Equation (12).

$$FAR = \frac{FP}{TN+FP}$$

(2) Accuracy (Acc) is defined as the ability measure of the classifier to correctly classify an object aseither normal or attack. The Accuracy can be defined using Equation (13).

$$Acc = \frac{TP+TN}{TP+TN+FP+FN}$$

(3) The F-measure (F-M) is a score of a classifier's accuracy and is defined as the weighted harmonic mean of the Precision and Recall measures of the classifier. F-Measure is calculated using Equation

$$F - Measure = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

(4) Precision represents the number of positive predictions divided by the total number of positive class values predicted. It is considered as a measure for the classifier exactness. A low value indicates large mumber of false positives, The precision is calculated using equation.

$$Precision = \frac{TP}{TP+FP}$$

(5) Recall is the number of True Positives divided by the number of True Positives and the number of False Negatives. Recall is considered as a measure of a classifier completeness such that a low value of recall realizes many False Negatives [44]. Recall is estimated through Equation

$$Recall = \frac{TP}{TP+FN}$$

VI. Conclusion

In this paper, we survey on performance measurements of support vector machine and deep learning algorithms based on up-to-date CICIDS2017 dataset were presented comparatively. Our proposed algorithm will show that the deep learning algorithm performed significantly better results than SVM.

We will be using not only port scan attempts but also other attack types with machine learning and deep learning algorithms, and detect scams.

References

[1]. P. A. A. Resende and A. C. Drummond, "Adaptive anomaly-based intrusion detection system using genetic algorithm and profiling," *Security and Privacy*, vol. 1, no. 4, p. e36, 2018.

[2]. Vijayan and, R.; Devaraj, D.; Kannapiran, B. Intrusion detection system for wireless mesh network using multiple support vector machine classifiers with genetic-algorithm-based feature selection. Comput. Secur. **2018**, 77, 304–314.

[3]. Sharafaldin, I.; Lashkari, A.H.; Ghorbani, A.A. Toward generating a new intrusion detection dataset and intrusion traffic characterization. In Proceedings of the Fourth International Conference on Information Systems Security and Privacy, ICISSP, Funchal, Madeira, Portugal, 22–24 January 2018.

[4]. Lee, C.H.; Su, Y.Y.; Lin, Y.C.; Lee, S.J. Machine learning based network intrusion detection. In Proceedings of the 2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA), Beijing, China, 8–11 September 2017; pp. 79–83.

[5]. Abdulhammed, R.; Faezipour, M.; Elleithy, K. Intrusion Detection in Self organizing Network: A Survey. In Intrusion Detection and Prevention for Mobile Ecosystems; Kambourakis, G., Shabtai, A., Kolias, C., Damopoulos, D., Eds.; CRC Press Taylor & Francis Group: New York, NY, USA, 2017; Chapter 13, pp. 393–449.

[6]. K. Ibrahimi and M. Ouaddane, "Management of intrusion detection systems based-kdd99: Analysis with lda and pca," in *Wireless Networks and Mobile Communications (WINCOM), 2017 International Conference on*. IEEE, 2017, pp. 1–6.

[7]. N. Moustafa and J. Slay, "The significant features of the unsw-nb15 and the kdd99 data sets for network intrusion detection systems,"in *Building Analysis Datasets and Gathering Experience Returns for Security (BADGERS), 2015 4th International Workshop on*. IEEE, 2015, pp. 25–31.

[8]. L. Sun, T. Anthony, H. Z. Xia, J. Chen, X. Huang, and Y. Zhang,"Detection and classification of malicious patterns in network traffic using benford's law," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2017*. IEEE, 2017, pp. 864–872.

[9]. S. M. Almansob and S. S. Lomte, "Addressing challenges for intrusion detection system using naive bayes and pca algorithm," in *Convergence in Technology (I2CT), 2017 2nd International Conference for*. IEEE 2017, pp. 565–568.

[10]. M. C. Raja and M. M. A. Rabbani, "Combined analysis of support vector machine and principle component analysis for ids," in *IEEE International Conference on Communication and Electronics Systems*, 2016, pp. 1–5.

[11]. S. Aljawarneh, M. Aldwairi, and M. B. Yassein, "Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model," *Journal of Computational Science*, vol. 25, pp. 152–160, 2018.

[12]. I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization." in *ICISSP*, 2018, pp. 108–116.

[13]. D. Aksu, S. U¨ stebay, M. A. Aydin, and T. Atmaca, "Intrusion detection with comparative analysis of supervised learning techniques and fisher score feature selection algorithm," in *International Symposium on Computer and Information Sciences*. Springer, 2018, pp. 141–149.

[14]. N. Marir, H. Wang, G. Feng, B. Li, and M. Jia, "Distributed abnormal behavior detection approach based on deep belief network and ensemblesvm using spark," *IEEE Access*, 2018.

[15]. P. A. A. Resende and A. C. Drummond, "Adaptive anomaly-based intrusion detection system using genetic algorithm and profiling," *Security and Privacy*, vol. 1, no. 4, p. e 36, 2018.

[16]. C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*,vol. 20, no. 3, pp. 273–297, 1995. [12] J.F Joseph,A. Das,B.C. Seet, (2011) Cross-Layer Detection of Sinking Behavior in Wireless Ad Hoc Networks Using SVM computing, Vol. 8, No. 2, Marh April 2011 and FDA. IEEE Transaction on dependable and secure computing,

IOSR Journal of Computer Engineering (IOSR-JCE) is UGC approved Journal with Sl. No. 5019, Journal no. 49102.

Mrs. Manasee Kurkure. " Survey on Detecting Port Scan Attempts with Combined Analysis of Support Vector Machine and Deep Learning Algorithms" IOSR Journal of Computer Engineering (IOSR-JCE) 21.3 (2019): 42-46.