ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940www.ijsem.org

Vol 9, Issuse.1Jan 2021

Examination of Port Scan Detection Algorithms Utilizing Deep Learning and Support Vector Machines

DR SATYAJIT NAYAK

Abstract—

Recent advances in computing and communication have made significant and far-reaching improvements over their predecessors. The employment of modern technology has many positive effects on people's lives, businesses, and governments; nevertheless, it also has some negative effects. Concerns have been raised about a variety of issues, including the availability of knowledge, the privacy of sensitive data, and the safety of data storage systems. As a result of these factors, cyber terrorism has emerged as one of the most pressing concerns of our day. Many other types of groups, including criminal organizations, professionals, and cyber activists, are now capable of posing a danger to national security via acts of cyber terror that have already caused significant disruptions to people and institutions. To that end, Intrusion Detection Systems (IDS) have been designed to safeguard networks against malicious software. Based on the latest CICIDS2017 dataset, this research employed deep learning and support vector machine (SVM) algorithms to identify port scan attempts, with accuracy rates of 97.80% and 69.79%, respectively.

INTRODUCTION

Computer crimes continue to increase over the years. They are not only restricted to insignificant acts such as estimating the login credentials of a system but also they are much more dangerous. 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

ASSOCIATE PROFESSOR, Mtech,Ph.D
Department of CSE
Gandhi Institute for Technology,Bhubaneswar.

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.

introduced in Section 4. Section 5 provided conclusion and future works.

LITERATURE REVIEW

Information security concepts consist of human, period, methodology, knowledge, system and technology as is shown in Figure 1. Confidentiality, integrity, and accessibility have to be provided by a secure system. First, the confidentiality of the information means allowing access only to the person who needs to access that information. Second, the integrity of the information is ensuring that the information is protected without distortion and the original structure is intact. Finally, the accessibility of information is the ability to access and use information at the desired time.

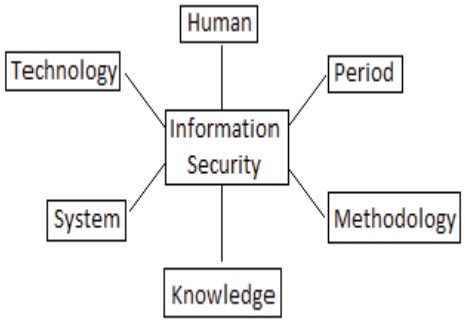

Fig. 1. Information security concepts [3].

In this work, deep learning and SVM machine learning algorithms were applied to create IDS models to detect port scan attempts. The models were presented comparatively. We categorized other parts of the paper as follows: a literature review was presented in Section 2. Section 3 presented an explanation of used material and methods. Experimental results of the classification algorithms and performance measurements were

As is signified by Stanford et al, there has been astonishingly limited work on the issue of detecting port scans [4].

Robertson et al. used a threshold method to detect the failed connection attempts [5]. Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA) were applied by Ibrahim and Outdone to identify the intrusion with NSL-KDD dataset [6]. Comparative consequences of KDD99 and UNSW-NB15 datasets analyzing network behaviors were showed by Mustafa and Slay [7]. Laying et al. detected and classified malicious patterns in network traffic based on the KDD99 dataset [8]. Naive Bays and Principal Component Analysis (PCA) were

Used with the KDD99 dataset by Alanson and Lumet [9]. Similarly, PCA, SVM, and KDD99 were used Chitin and Rabbinic for IDS [10]. In Aljawarneh et al.'s paper, their analysis and experiments were produced based on the NSLKDD dataset for their IDS model [11]. Literature studies show that KDD99 dataset is always used for IDS [6]–[10]. There are 41 features in KDD99 and it was developed in 1999. For this reason, KDD99 is old and does not provide any information about up-to-date new attack types such as zero days exploits etc. Therefore we used an up-to-date and new CICIDS2017 dataset [12] in our study. There are different but limited studies based on the

CICIDS2017 dataset. Some of them were discussed here. D. Aksum et al. showed performances of various machine learning algorithms detecting Dodos attacks based on the CICIDS2017 dataset in their previous work [13]. They did not apply all dataset and used limited data 26.167 Dodos and 26.805 benign samples from the dataset in their study. Moreover, they used the Fisher score feature selection algorithm to select the best features. Therefore, their previous SVM models reached a very high accuracy result. However, they were planning to apply deep learning algorithm as a feature work to detect Dodos attacks. N. Mari et al. proposed a distributed study to discover abnormal activity in a large scale network [14]. In another study, Resend et al. used genetic algorithms to detect intrusions on the CICIDS2017 dataset [15].

MATERIAL AND METHODS

The CICIDS2017 dataset and deep learning and SVM algorithms are explained respectively in this section.

A. 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. In this study, we focused on port scan attempts. There are 286467 records consisting 127537 benign and 158930 port scan attempts and each record has 85 features such as source IP, source port, destination port, flow duration, total fwd packets, total backward packets etc. A part of the records is as shown in Table I.

When creating the dataset, Attack-Network and Victim- Network, completely were separated two networks, were designed and implemented by Sharafaldin H. et al [12]. They collected data from July 3, 2017, to July 7, 2017, for the dataset.

B. SVM

Statistical learning and convex optimization, based on the principle of structural risk minimization, form the basis of Support Vector Machine (SVM) algorithms. Vapid et al developed SVM as a solution to different problems [16]. For example, it can be used in many different areas such as learning, pattern recognition, regression, classification, and analysis.

TABLE I

A SAMPLE SET OF RECORDS FROM DATASET [12]

Source IP	Source Port		Flow Duration	Total Fwd Packets
192.168.10.12	35396		1266342	41
192.168.10.16	60058		1319353	41
192.168.10.12	35396		160	1
192.168.10.12	35398		1303488	41
192.168.10.50	22		77	1
192.168.10.16	60058		244	1
192.168.10.16	60060		1307239	41
192.168.10.50	22	--	82	1
192.168.10.12	35398		171	1
192.168.10.16	60060		210	1
192.168.10.50	22		75	1
192.168.10.50	22		77	1
192.168.10.14	53235		2	2
192.168.10.14	53235		27701	15
192.168.10.14	53234		152547	19
192.168.10.50	52520		4	3

SVM is a supervised learning method because it uses tagged data in a dataset as an input. The number of output classes changes depending on the dataset. For example, two classes of output data are generated when a dataset of two classes is given as the input. Therefore, the samples given as the input are categorized according to these classes. During the training process, a model is created according to the input dataset and classification is performed by using the model.

C. Deep Learning

Deep Learning algorithms allow to extract features automatically from a given dataset and they consist of a sequential layer architecture. Applying non-linear transformation functions to the sequential layer structure constitute the basis of deep learning algorithms. Increasing the number of layers will

increase the complexity of nonlinear transformations to be constructed. Deep learning algorithms learn the abstract hidden properties of the data obtained in the last layer from its abstract representations acquired at multiple levels. Therefore, the abstract properties of the final layer's output are obtained by introducing the data into a high-level non-linear function.

D. Methodology

The SVM and deep learning algorithms were used to detect port scan attempts based on the CICIDS2017 dataset. The flowchart of the proposed method was presented in figure 2. First of all, 286.467 records which consist of 158.930 port scan attempts and 127.537 benign behaviors are taken from the dataset and then these records were normalized. After normalization samples were split into two as a 67% training data and 33% testing data. In addition, the SVM and deep learning IDS models were created based on the training data. Finally, the models were tested with test data and the performance of models was calculated comparatively. In addition, the deep learning IDS model consist of 7 hidden layers and each layer include the different number of neurons such as 100, 150, 70, 40 and 6 respectively. The rely was selected and used as an activation function in the model. Depending on the number of neurons and hidden layer model performances were changed. In this paper, we selected optimum numbers based on the model's accuracy. On the other hand, we did not apply any feature selection algorithm for SVM and we used all features. As a future work, we are going to use different artificial intelligence approaches to define select this optimum values.

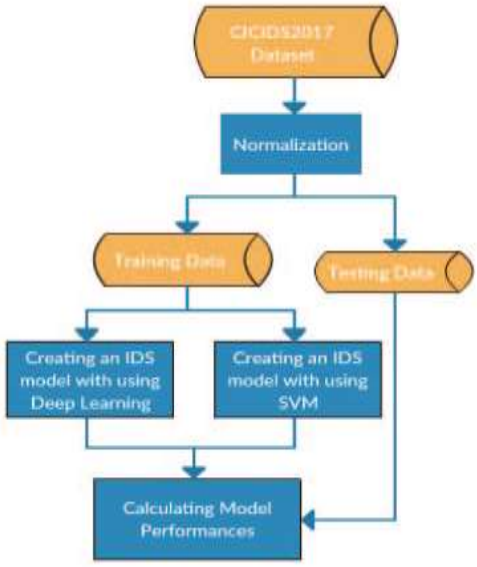

Fig. 2. Flowchart of the method.

As is shown in figure 2, main steps of the algorithm are presented in below.

1) Normalize the dataset.

2) Split the normalized dataset into two as training and testing.

3) Create IDS models with using SVM and deep learning algorithms.

4) Evaluate the models' performances. 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. Furthermore, we rescaled all observed values of features to have a length of 1. As a second step, the normalized dataset was split into 67% training and 33% testing. In the third step, the IDS models were trained and generated 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.

TABLE II

CONFUSION MATRIX

Actual Class\Predicted Class	Normal (Benign)	Anomaly (Port Scan)
Normal (Benign)	TN	FP
Anomaly (Port Scan)	FN	TP

• TP: Actual Port Scan is classified as Port Scan. Accuracy, recall, precision and f1 score performance metrics are calculated using the statistics of the confusion matrix (Table III).

TABLE III

PERFORMANCE METRICS [17]

Measure	Formula
Accuracy	(TP+TN) / (TP+FP+FN+TN)
Recall	TP / (TP+FN)
Precision	TP/(TP+FP)
F1 score	2TP / (2TP+FP+FN)

The ratio of correctly predicted observations is accuracy, while precision means a ratio of correct positive observations. The recall is a proportion of correctly predicted positive events. F1 score signifies the weighted average of precision and recall.

EXPERIMENTAL RESULTS

The personal computer which has Intel(R) Core(TM) i7- 5700HQ CPU @2.70 GHz, 16 GB Ram capacity was used for experiments. We used the CPU; however, we are considering applying GPU as a future work. 286.096 records, which were taken from the normalized dataset, were divided into two sets with 67% training and 33% testing ratios such as 191684 samples for training and 94412 samples for testing. The deep learning model was trained in 30 Epochs and performance measurement of the SVM and deep learning models presented in Table IV.

TABLE IV

PERFORMANCE METRICS OF USED CLASSIFICATION TECHNIQUES BASED ON CICIDS2017 DATASET.

Method	Accuracy	Precision	Recall	F1 score
Deep Learning	0.9780	0.99	0.99	0.99
SVM	0.6979	0.80	0.70	0.65

Table IV shows the accuracy, recall, precision and F1 score rates of the IDS models which were developed by using deep learning and SVM. Deep learning achieved a higher success than SVM.

CONCLUSION AND FUTURE WORKS

In this research, we give a comparison of support vector machine and deep learning algorithms' performance metrics using the most recent CICIDS2017 dataset. The findings demonstrate that the deep learning algorithm outperformed the SVM by a wide margin. Based on this information, we want to leverage machine learning and deep learning techniques, Apache Hadoop and Spark technology, and various attack types beyond port scan efforts.

REFERENCES

[1] K. Graves, Chef: Official certified ethical hacker review guide: Exam 312-50. John Wiley & Sons, 2007.

[2] R. Christopher, "Port scanning techniques and the defense against them," SANS Institute, 2001.

[3] M. Bayar, R. Das, , and ˙I. Karado˘gan, "Bilge g¨uvenli˘gi sistemlerinde kullanılan arac, ların incelenmesi," in 1st International Symposium on Digital Forensics and Security (ISDFS13), 2013, pp. 231–239.

[4] S. Staniford, J. A. Hoagland, and J. M. McAlerney, "Practical automated detection of stealthy portscans," Journal of Computer Security, vol. 10, no. 1-2, pp. 105–136, 2002.

[5] S. Robertson, E. V. Siegel, M. Miller, and S. J. Stolfo, "Surveillance detection in high bandwidth environments," in DARPA Information Survivability Conference and Exposition, 2003. Proceedings, vol. 1. IEEE, 2003, pp. 130–138.

[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 IEEEInternational Conference on Communication and Electronics Systems, 2016, pp. 1–5.