Performance Analysis of Semi-Supervised Machine Learning Approach for DDoS Detection

Mehnaz Anjum[1], Dr. Shreedhara K S[2]
[1]*Computer Science and Engineering, UBDTCE, Davanagere-577001, Karnataka, India*
[2]*Professor Computer Science and Engineering, UBDTCE, Davanagere-577001, Karnataka, India*

***Abstract*- Distributed daniel of service (DDoS) is one of the cyber-attack, which remains as a major attack on internet for past many years. DDoS detection based on Machine Learning techniques such as, Supervised and Unsupervised techniques has been already implemented which has some drawbacks like low detection accuracy and high false positive rates. In this paper, DDoS detection based on Semi-Supervised Machine learning technique is presented which is the combination of both supervised and unsupervised techniques that provides better results compared to the existing approaches. Unsupervised part consists of some estimation steps including clustering which reduces the false positive rates and increases the accuracy by reducing irrelevant data. In supervised part Random forest algorithm is used to accurately classify the DDoS attack data and it also reduces the false positive rate of unsupervised part.**

Index terms- Semi-Supervised, Clustering, Random forest

I. INTRODUCTION

The science of modeling the Computers to learn and make like human to solve the problems correctly is known as Machine Learning. The important focus of machine learning is to present the algorithms that can be trained to perform the task. Python is the wonderful programming language for ML. Python consists of almost all machine learning built-in algorithms, which is easy to access those algorithms for particular task by downloading the packages required. This makes the code smaller, easier and helps to provide better results.

In this approach, initially the network traffic data is read. Average entropy is estimated for the data and then by applying the clustering algorithm 3 clusters are formed. By making use of average entropy of data, the information gain ratio is calculated. Anomalous cluster is formed by combining the clusters which having highest gain ratio value. Random forest algorithm is applied for the obtained anomalous cluster to accurately classify the data and detect the DDoS attack.

For better evaluation of performance of proposed approach, NSL-KDD network traffic dataset is used. The NSL-KDD dataset contains the different types of attack data. It has 42 features which can be classified into some groups. This dataset contains a total number of 125973 records in training set and 22554 records in testing set. In this paper section II shows the background work, section III shows methodology, section IV shows results and section V shows conclusion.

II. BACKGROUND WORK

The attacks were classified into four main categories by the group of research workers V. Jaiganesh, Dr. P. Sumathi, S. Mangayarkarasi [1], and the categories are named as Dos, Probe, U2R and for this purpose BPN and machine learning techniques were used. For the attacks of different kinds they have worked on the detection rates and for the detection of DoS attacks the BPN usage rate is 78.15 percent.

Making use of BPN model the data was trained with eight different types of attack by Changjun Han, Yi Lv and Dan Yang, Yu Hao in [2]. In this for the training the total of 1325 connections were used as well and also for testing the used samples were 1245. The outcomes which are obtained indicate that there is 80.5 percent rate of detection, the rate of 7.4 percent is the false alarm and only 11.3 percent is the omission rate. The first data was trained by some research workers named Sufyan T. Faraj and et al. in [3], and they even used the BPN for the classification of normal and abnormal events. The different scenarios were considered for the rate of detection and false positive rate. For the test set of about 90

percent the detection of normal and abnormal events was made and the approximate probe is 60 to 80 percent for the classification of DoS U2R and R2L.

A model was trained for DoS, U2R and Probe for the BPN neural network and the classes of normal attack by Mukhopadhyay and etl al [4], in this case the rate of success for the system was as much as 73.9 percent for the latest testing sets while for level 1 test set equal to 95.6 percent was the obtained rate. For the detection of anomaly the Hua TANG and Zhuolin CAO was used in the MLP neural network. For DoS, U2R, Probe as well as U2L the accuracy was compared and it with the neural network accuracy and it was found that the accuracy was better as compared to SVM. The neural network ensemble approach was also used by Vladimir Bukhtoyarovf and Eugene Semenkin. On the classification of probe attacks their work was mainly focussed and in this way making use of joint usage of the networks which are trained and neural the probes are classified. The detection rate was found to 99.87 percent for the probing attacks but training time and large amount is needed which is needed by the IDS issues for the time of training.

Novel one class learning approach was suggested by Van Loi [4] for the anomaly detection of the network at the density estimation and auto encoders. The method was tested by the authors on the data sets of NSL KDD and quite satisfactory outcomes were obtained as well.

III. METHODOLOGY

Figure 1 represents the methodology diagram of the proposed approach. It is comprised of some components that are related to each other and work together to implement the system. There are four major steps in the approach suggested here; these include the traffic network entropy estimation, the co clustering, computation of information gain ratio and the classification of network traffic.

Network Traffic Data: In proposed work, NSL-KDD[18] network traffic data is used. The data set is known as NSL KDD which also proposes the solution of the inherent concerns for the data set of KDD 99. Though KDD newer version for the data set is there but still from some problems it is seen to suffer and hence for the existing real network it might not be a perfect representative since public data sets are lacking in the IDSs based networks, it can also be applied as per belief that some effective as well as benchmark data sets are applicable so that investigators can be helped in such a way that they can contrast various methods of intrusion detection.

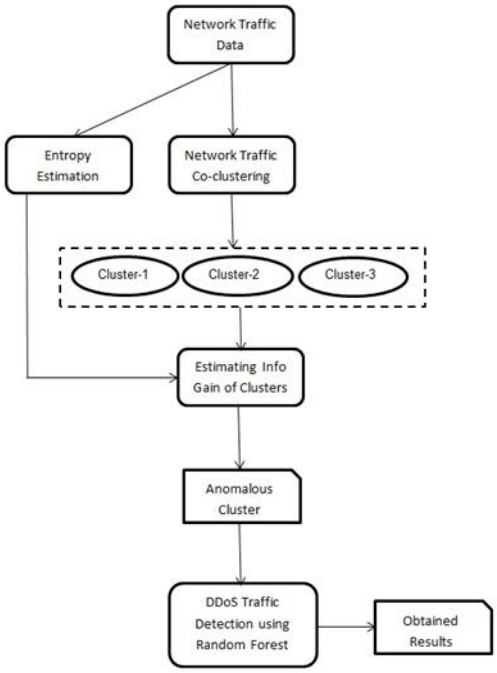

Figure 1: Methodology diagram

Entropy Estimation: In the NSL KDD dataset like cases, two file size distribution (FSD) features are used such as source bytes as well as destination bytes for which average entropy is estimated using entropy calculation algorithm. This allows reducing the high dimensionality of data.

Network Traffic Co-clustering: The traffic data in this next step is split into the 3 main clusters and for this purpose the co clustering algorithm is used such as spectral co clustering algorithm. Spectral co clustering algorithm is considered as simple and it produces the greater accuracy as compared to other clustering algorithms. The network traffic splitting goal is generally to lower the data amount which is supposed to be classified by the exclusion of normal data.

Estimating Info Gain of Clusters: The information estimation which is gained on the basis of the features of FSD allows the two cluster identification allowing more data about the attack of DDoS and cluster which comprises of the normal traffic only. Hence, for the clusters with high information ratio is considered as anomalous cluster.

DDoS Traffic Detection using Random Forest: The data present in the anomalous cluster is preprocessed for classification by taking care of missing data, encoding categorical data and feature scaling. The trees which are based on ensemble such as Random Forest are also utilized so as to overcome the problem of representation of the decision tree unvaried and to classify the attack data accurately. Therefore the trees which are based on ensemble are used extensively for the purpose of classification.

IV. RESULTS

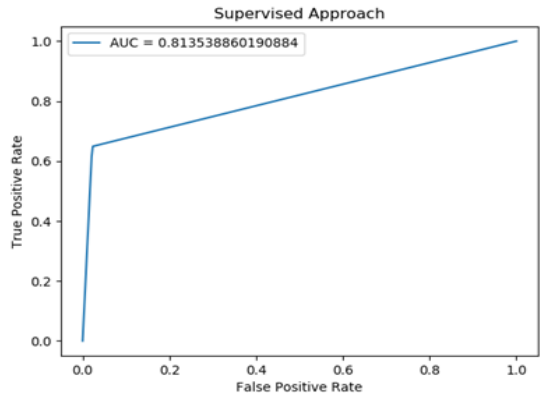

Figure 2: Performance measure of Supervised approach

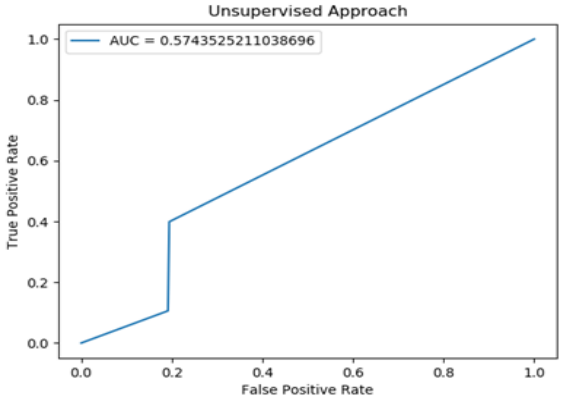

Figure 3: Performance measure of Unsupervised approach

The performance of Supervised and Unsupervised approach is represented in the graphs as shown in Figure 2 and Figure 3. Supervised approach provides 0.81 of accuracy where unsupervised approach provides 0.57. And also the false positive rates are high as observed from the graphs. The improved performance of the proposed approach is presented in the graph as shown in Figure 4. It gives the 0.93 accuracy and false positive rates are also reduced.

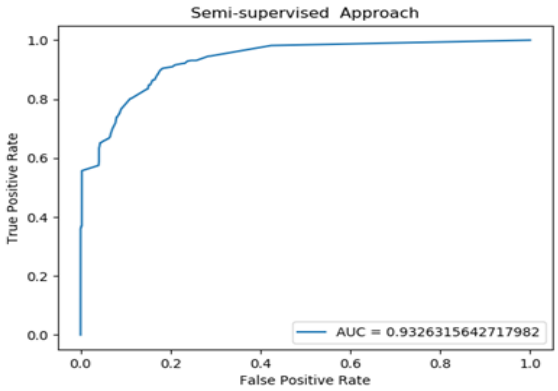

Figure 4: Performance measure of Semi-Supervised approach

V. CONCLUSION

The main purpose of this proposed approach is to improve the performance compared to the supervised and unsupervised techniques for attack detection. By adopting the steps mentioned in the methodology and implementing the same using python programming language, the satisfactory results are obtained for accuracy and false positive rates.

REFERENCES

[1] Jaiganesh V., Sumathi P. and MangayarkarasiS.,"An Analysis of Intrusion Detection System using Back Propagation NeuralNetwork",IEEE 2013 publication

[2] Han C., Yi Lv, Yang D., Hao Y., "An Intrusion Detection SystemBased on Neural Network",2011 International Conference onMechatronic Science, Electric Engineering and Computer, August 19-22, 2011, Jilin, China,IEEE Publication

[3] Faraj S, Al-Janabi and Saeed H, "A Neural Network Based AnomalyIntrusion Detection System",2011 Developments in E-systemsEngineering,DOI 10.1109/DeSE.2011.19,IEEE publication

[4] Mukhopadhyay I, Chakraborty M, Chakrabarti S, Chatterjee T, "BackPropagation Neural Network Approach to Intrusion DetectionSystem",2011 International Conference on Recent Trends inInformation Systems, IEEE publication

[5] Nicolau M, McDermott J et al (2016) A hybrid autoencoder anddensity estimation model for anomaly detection. In: International conference on parallel problem solving from nature. Springer, pp717–726