Semi Supervised Machine Learning Approach for DDoS Detection

[1]Dulapalli Pavani Rani, [2]D S Ramkiran
[1,2]Dept. of CSE, KIET, Kakinada, AP, India

Abstract
In the era of internet and online connectedness, where data is the most valuable asset, it is ever important for an organization to protect itself and it's assets from various security threats. One of these threats is a Distributed Denial of Service (DDoS) attack that can cut off the network service by overwhelming the targeted server or network by flooding it with superfluous requests in an attempt to overload the server to prevent legitimate requests from being fulfilled. DDoS attacks utilize multiple compromised systems as sources of internet traffic to increase their effectiveness. What makes DDoS attacks more lethal is that fighting them requires differentiating legitimate requests from illegitimate ones. A site or service unexpectedly being sluggish or inaccessible is the most obvious symptom of a DDoS attack. But since a number of causes like legitimate spike in network traffic can create similar issues, further investigation is necessary.

Keywords
DDoS Detection; Intrusion Detection; Machine Learning; Distributed Denial of Service

I. Introduction

Distributed Denial of Service (DDoS) attacks have been one of the most prominent attacks over the last decade. Distributed denial of service (DDoS) attack is an effort to make an online service unavailable to legitimate users by overwhelming it with traffic from multiple sources. These sources are generally computers that are infected and used as a bot in botnets. DDoS attacks are costly to an organization or company as they prevent legitimate users from accessing the resources. (Times New Roman, 10)
Consequently, it is important to propose an effective method for detecting DDoS attacks from massive data traffics. The existing methods, however, do have limitations, such as the need for large labeled dataset for supervised learning methods, and the relatively low accuracy and high false positive rate for unsupervised learning algorithms. In order to combat these issues, this paper presents a hybrid approach that uses a combination of five different classifiers - Naive Bayes, SVM, KNN, Fuzzy c-means and Random Forest. Most of the research has been done using older datasets like KDD99 , NSL KDD and DARPA. The models that are trained on these older datasets are found to be less accurate and inconsistent. Therefore, we are using a new and improved dataset - CICDDoS2019. (Times New Roman, 10)

II. Literature Review

Wu Zhijun, Xu Qing, Wang Jingie, YueMeng, Liu Liang [2] proposes a multi feature DDoS attack detection based on FM and uses a combination of Support Vector Machine (SVM) and Self-Organizing Mapping (SOM) to detect DDoS. This works for special data instead of general prediction tasks.

Shi Dong, MudarSarem [3] used an improved K-Nearest Neighbour (KNN) and four features(flow length, flow size, flow ratio) to detect DDoS attacks. Uses a combination of DDADA and DDAML algorithms but some further research is needed.

Sabah Alzahrani, Liang Hong [5] had a signature based artificial neural network(ANN). It has a signature based approach where if the attack has a known signature then a predefined approach is followed and if not then an anomaly detection distributed neural network will be used to detect the unknown DDoS attack.

Saikat Das and team [6] uses an ensemble of different techniques like Artificial Neural Network (ANN), Support Vector Machine (SVM), Multilayer Perceptron (MLP), Na¨ıve Bayes (NB), Multivariate Adaptive Regression Splines (MARS), K Nearest Neighbor (KNN).

SumanNandi, SantanuPhadikar, KoushikMajumder [7] uses combination of Naive Bayes, Bayes Net, Decision Table, J48 and Random Forest and a five feature selection method like information gain, gain ratio, chisquared, reliefF, and symmetrical uncertainty.

Petr Blazek, Tomas Gerlich, and ZdenekMartinasek [8] uses forward Neural Networks (NN) as an anomaly detection method and a signature based intrusion detection system Suricata which is an open source threat detection engine and has been owned and implemented by Open Information Security Foundation.

Wenwen Sun, Yi Li, Shaopeng Guan [9] first uses the entropy to detect whether the flow is abnormal or not. The BLSTM-RNN neural network algorithm is used to train the data set, and the model is used to detect DDoS attacks on real time traffic.

Roshni Mary Thomas, Divya James [10] uses a traffic monitoring method iftop in the server to check the traffic for a specific amount of time. iftop is a traffic monitoring tool to find the bandwidth of incoming packets along with the address.

Swati Sahu, Amit Verma [12] detects the type of DDoS attack then the traffic is categories into normal or malicious. Again a filtering is done to categories it as either suspicious or normal. If suspicious then passed to a honey pot. It is set up on a server level and not on the subscriber level. The Honey Pot assumes that the attack must be detectable using a signature based detection tool.

RuchiVishwakarma and Ankit Kumar Jain [13] use a honey pot to intentionally lure in attackers with the purpose to capture the malware properties, the signature the style of invading and captures the whole information inside of a log file. A detection framework is used to predict the abnormal activities based on the log files generated in the Honey Pot.

Yuze SU and team [14] used a phase space reconstruction technique to denoise the original traffic. RBF neural network is used to train the network traffic sample and predict the future incoming network traffic for DDoS attacks.

S. ShanmugaPriya, M. Sivaram, D. Yuvaraj, A. Jayanthiladevi [15] uses three classification algorithms that are KNN, RF, NB to classify DDoS packets from normal packets into two features which are delta time and packet size.

Obaid Rahman and [16] uses a combination of J48, RF (Random Forest), SVM (Support Vector Machine), KNN (K Nearest Neighbour) to detect and block the DDoS attack in the SDN network.

Gaganjot Kaur, Prinima Gupta [17] makes the use of Bayesian Network, Wavelets, SVM (Support Vector Machine) and KNN (K-Nearest Neighbour). It has parameters like packet flow, time duration, accuracy and precision rate which is applied on the KNN dataset.

V. Deepa, K. MuthamilSudar, P. Deepalakshmi [18] uses a hybrid approach combining both SVM(Support Vector Machine) and SOM(Self Organized Map) or using them standalone.

Shuang Wei, Shuaifu Dai, Xinfeng Wu, Xinhui Han [19], it uses a two stage hierarchical architecture. In the first stage it inputs the flow information gathered controller to do clustering. DDoS attacks are detected using KL distance between the real time flow distribution with the distribution.

III. Methodology

The purpose of our research is to build an accurate DDoS detection model with a low false positive rate . Here, we propose a model based on an ensemble of five machine learning classifiers . The selected classifiers are Na¨ıveBayes, SVM, KNN, Fuzzy c-means and Random Forest. In our model all five classifiers work independently and build a different model of the data. The outputs of the five classifiers are combined by a majority voting method to obtain a final result of the model. CIC-DDoS2019 dataset is used to train the model. This dataset has 80 features. The feature set for the training data set must be reduced and is done using a feature selection algorithm that combines Information Gain , Gain Ratio , Chi-squared and ReliefF.

Modules

The Proposed Model has three different modules:

- Data Preprocessing
- Data Classification
- Ensemble with Majority Voting

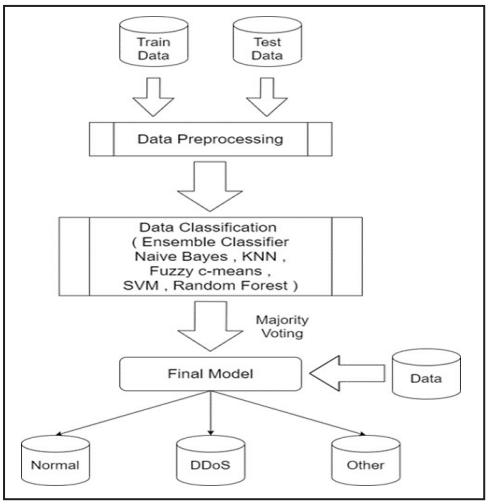

Figure 1.1 Architecture Diagram

A. Data Preprocessing

1. Dataset

The Dataset used for training and testing the model is "DDoS Evaluation Dataset (CIC-DDoS2019)" . The dataset was chosen based on earlier research . It fixes various issues that are in NSL KDD , DARPA 99 and CIADA 2007. The datasets have a total of 80 features. The dataset contains network traffic from 12 different DDoS attacks including NTP, DNS, LDAP, MSSQL, NetBIOS, SNMP, SSDP, UDP, UDP-Lag, WebDDoS, SYN PortScan, and TFTP. The dataset has a total of 80 different network traffic features.

2. Data Pre-Processing

Most machine learning models work with only numeric values. In this phase , we will convert non numeric values to numeric . The dataset contains 80 network network traffic feature . The features used to train the model must be reduced . Using more features can lead to low accuracy while using less features can lead to high false positive rate. The number of feature must be balanced to get a model with high accuracy but also low false positive rate . The selection of the feature set is done using a Feature Selection Method that combines Information Gain, Chisquare, Gain Ratio, and ReliefF.

i. Information Gain

This technique is used to determine the relevant features present in the dataset based on the information gain theory. Information gain is calculated by comparing the entropy of the dataset before and after the transformation

ii. Gain Ratio

The big downside to Information gain is that it is biased towards attributes with bigger values. The gain ratio is a changed information gain that normalizes the outcome of information gain.

iii. Chi – Squared

The chi-squared test is a predictive test that is used to calculate self-reliability between two attributes. IT calculates the difference between expected and the observed value.

iv. ReliefF

The significance of the feature is calculated by recognizing the difference between closest neighbours. It calculates a feature score for every feature and the features can be selected based on this score.

B. Data Classification

The model classifies data one by one with individual classifiers . These classifiers runs parallel and build a different model based on the training dataset. There are several classifiers in machine learning that can be used but we will use five classifiers: Na¨ıveBayes , SVM , KNN , Fuzzy c-means and Random Forest.

1. Naive Bayes

It is a probabilistic learning model that is used for classification in Machine Learning . It is based on the Bayes theorem. It is fast and easy to implement but it needs the predictors to be independent.

2. KNN

KNN stands for k-nearest neighbours. It is a non parametric algorithm based on supervised learning technique. It stores all

the available data and classifies a new data point based on their similarity.

3. SVM

Support Vector Machine (SVM) is a supervised machine learning model and it uses classification and regression. It categorizes data in different classes based on the labelled training data.

4. Fuzzy c-means

It is similar to the k-means algorithm that it divides data in clusters based on their similarity but unlike k-means data points can belong to multiple clusters.

5. Random Forest

It is made up of several independent decision trees, which are independently trained on a random subset of data from the labeled dataset. Random Forest works well because a large number of relatively independent trees will perform better than any of the individual models.

C. Ensemble with Majority Voting

The ensemble consists of five different classifiers that work parallel and each classifier builds a different model based on the training dataset. Majority Voting is an ensemble machine learning algorithm that combines predictions from multiple classifiers or models . The predictions for each label are summed and the label that has the majority vote is predicted.

IV. Conclusion

In this paper to deal with the problems or issues arising for supervised and unsupervised learning methods a hybrid approach is proposed that uses five different classifiers and is applied on the data set KDD99, NSL KDD and DARPA through which we also found out that these data sets are less accurate and inconsistent, therefore we are using CICDDos2019 data set.

V. Future Work

In future, better and bigger datasets will be used to test and train the model to achieve better results. In addition, using a combination of signature-based detection and current model will be able to improve the accuracy

References

[1] Santos And M. Nogueira, "A Distributed Architecture For Ddos Prediction And Bot Detection" IEEE Access, Vol. 8, Pp. 159756-159772, 2020, Doi: 10.1109/Access.2020.3020507.

[2] W. Zhijun, X. Qing, W. Jingjie, Y. Meng and L. Liang, "Low-Rate DDoS Attack Detection Based on Factorization Machine in Software Defined Network" in IEEE Access, vol. 8, pp. 17404-17418, 2020, doi: 10.1109/ACCESS.2020.2967478.

[3] S. Dong and M. Sarem, "DDoS Attack Detection Method Based on Improved KNN With the Degree of DDoS Attack in Software-Defined Networks" in IEEE Access, vol. 8, pp. 5039-5048, 2020, doi: 10.1109/ACCESS.2019.2963077.

[4] D.Yin,L. Zhang and K. Yang, "A DDoS Attack Detection and Mitigation With Software-Defined Internet of Things Framework" in IEEE Access, vol. 6, pp. 24694-24705, 2018, doi: 10.1109/ACCESS.2018.2831284.

[5] S. Alzahrani and L. Hong, "Detection of Distributed Denial of Service (DDoS) Attacks Using Artificial Intelligence on Cloud" 2018 IEEE World Congress on Services (SERVICES), San Francisco, CA, 2018, pp. 35-36, doi: 10.1109/SERVICES.2018.00031.

[6] S. Das, A. M. Mahfouz, D. Venugopal and S. Shiva, "DDoS Intrusion Detection Through Machine Learning Ensemble " 2019 IEEE 19th International Conference on Software Quality, Reliability and Security Companion (QRSC), Sofia, Bulgaria, 2019, pp. 471-477, doi: 10.1109/QRSC.2019.00090.

[7] S. Nandi, S. Phadikar and K. Majumder, "Detection of DDoS Attack and Classification Using a Hybrid Approach" 2020 Third ISEA Conference on Security and Privacy (ISEA-ISAP), Guwahati, India, 2020, pp. 41-47, doi: 10.1109/ISEA-ISAP49340.2020.234999.

[8] P. Blazek, T. Gerlich and Z. Martinasek, "Scalable DDoS Mitigation System" 2019 42nd International Conference on Telecommunications and Signal Processing (TSP), Budapest, Hungary, 2019, pp. 617-620, doi: 10.1109/TSP.2019.8768869.

[9] W. Sun, Y. Li and S. Guan, "An Improved Method of DDoS Attack Detection for Controller of SDN" in 2019 IEEE 2nd International Conference on Computer and Communication Engineering Technology (CCET), Beijing, China, 2019, pp. 249-253, doi: 10.1109/CCET48361.2019.8989356.

[10] R. M. Thomas and D. James, "DDOS detection and denial using third party application in SDN" 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), Chennai, 2017, pp. 3892-3897, doi: 10.1109/ICECDS.2017.8390193.

[11] D. Erhan and E. Anarim, "Hybrid DDoS Detection Framework Using Matching Pursuit Algorithm," in IEEE Access, vol. 8, pp. 118912-118923, 2020, doi: 10.1109/ACCESS.2020.3005781.

[12] S. Sahu and A. Verma, "DDoS attack detection in ISP domain using machine learning," 2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA), Pune, India, 2019, pp. 1-4, doi: 10.1109/ICCUBEA47591.2019.9128624.

[13] R. Vishwakarma and A. K. Jain, "A Honeypot with Machine Learning based Detection Framework for defending IoT based Botnet DDoS Attacks," 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2019, pp. 1019-1024, doi: 10.1109/ICOEI.2019.8862720.

[14] Y. Su, X. Meng, Q. Mengand X. Han, "DDoS Attack Detection Algorithm Based on Hybrid Traffic Prediction Model," 2018 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), Qingdao, 2018, pp. 1-5, doi: 10.1109/ICSPCC.2018.8567771.

[15] S. S. Priya, M.Sivaram, D. Yuvaraj and A. Jayanthiladevi, "Machine Learning based DDOS Detection," 2020 International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, 2020, pp. 234-237, doi: 10.1109/ESCI48226.2020.9167642.

[16] O. Rahman, M. A. G. Quraishi and C. Lung, "DDoS Attacks Detection and Mitigation in SDN Using Machine Learning," 2019 IEEE World Congress on Services (SERVICES), Milan, Italy, 2019, pp. 184-189, doi: 10.1109/SERVICES.2019.00051.

[17] G. Kaur and P. Gupta, "Hybrid Approach for detecting DDOS Attacks in Software Defined Networks" 2019 Twelfth International Conference on Contemporary Computing (IC3), Noida, India, 2019, pp. 1-6, doi: 10.1109/

IC3.2019.8844944.

[18] V. Deepa, K. M. Sudar and P. Deepalakshmi, "Detection of DDoS Attack on SDN Control plane using Hybrid Machine Learning Techniques," 2018 International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2018, pp. 299-303, doi: 10.1109/ICSSIT.2018.8748836.