Classification Based Machine Learning for Detection of DDoS attack in Cloud Computing

Anupama Mishra
Swami Rama Himalayan University
Uttarakhand, India
Email: tiwari.anupama@gmail.com

B. B. Gupta*
National Institute of Technology, Kurukshetra
Haryana, India & Asia University, Taiwan
Email: *gupta.brij@gmail.com
{* corresponding author }

Dragan Peraković
University of Zagreb, Croatia
Email: dperakovic@fpz.unizg.hr

Francisco José García Peñalvo
University of Salamanca, Spain
Email: fgarcia@usal.es

Ching-Hsien Hsu
Department of Computer Science and Information Engineering
Asia University & National Chung Cheng University, Taiwan
Email: robertchh@gmail.com

2021 IEEE International Conference on Consumer Electronics (ICCE) | 978-1-7281-9766-1/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICCE50685.2021.9427665

***Abstract*—Distributed Denial of service attack(DDoS)is a network security attack and now the attackers intruded into almost every technology such as cloud computing, IoT, and edge computing to make themselves stronger. As per the behaviour of DDoS, all the available resources like memory, cpu or may be the entire network are consumed by the attacker in order to shutdown the victim's machine or server. Though, the plenty of defensive mechanism are proposed, but they are not efficient as the attackers get themselves trained by the newly available automated attacking tools. Therefore, we proposed a classification based machine learning approach for detection of DDoS attack in cloud computing. With the help of three classification machine learning algorithms K Nearest Neighbor, Random Forest and Naive Bayes, the mechanism can detect a DDoS attack with the accuracy of 99.76%.**

I. Introduction

The distributed denial of service (DDoS) attacks used a group of compromised machines to attack on the victim machine, rather using a single machine. Attackers compromise the innocent machines which are also known as bots or zombies, and command them to launch the attack on victim's machine.Today, we are facing so many attacks which consumes the available network infrastructures and break the security such as confidentiality, integrity and availability[1].

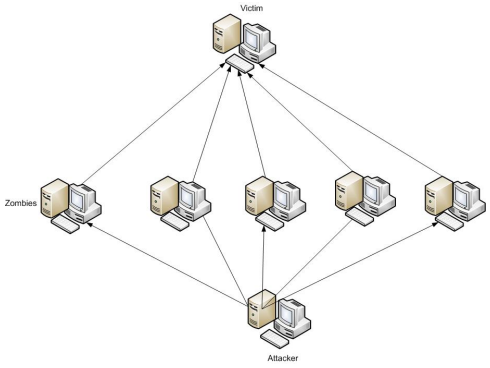

Fig. 1. Presents DDoS Attack.

In general, DDoS attack happens on network layer and application layer by exhausting services of network and servers respectively.In our paper, we focused on the DDoS attacks based on network layer which has its sources from virtual machines in the cloud environment. As per the reports of McAfee Lab, it has a huge impact around the world. The DDoS attacks are able to destroy or degrade or make unavailable a single page, or a lots of servers together, such as email servers, or http servers. The attackers used a fake credit card to rent Virtual machines to launch the attack in commodity communities cloud[2].

II. Related work

There are so many existing approaches which are based on many different concepts and algorithm. Though some of them are very good in terms of accuracy but fails as they are increasing the computation time to detect the DDoS attack. Lets have a look on the already existing work.
In[3], the authors have discussed about the netflow protocol. The data retrieval time from this protocol is very less but at the same time it only support the Cisco network devices like Cisco routers. The authors have proposed defensive mechanisms based on machine learning.[4]. The Authors,[5][6] have developed an Intrusion Detection System, which worked for different machine learning algorithms are K Nearest Neighbor(KNN), MLP and Decision Tree(DT). There are many problems like MLP required a huge amount of data to train and due to complex algorithm, it increases processing time. DT can not deal with noise, so the authors concluded KNN deals as best algorithm since it can handle the noise and provide a good accuracy. The Authors [7] used arriving rate for various packet along with the algorithm DBSCAN for clustering the data, so the limitation of this research work is, it can not perform over live captured data. The authors [8] presented the working and performance of various machine learning classification algorithms. The Authors discussed that Random Forest(RM) is good than KNN as it gives better accuracy [9] but they have not considered the all type of ddos attack. In

[11],[13] and [14], authors experienced that naive bayes better over RM and KNN due to its the precision. So at the end of the survey of many research work, we found that RM, naive bayes and KNN can be the good choice to be performed on our proposed algorithms.

III. Machine learning classification algorithms

A. Random Forest

Random Forest[15], a supervised machine learning algorithm is majorly recommended due to its much better results in comparisons to other machine learning algorithm. As a merit of it, it is used to solve the problem of classification and regression both. The algorithm works on multitude of decision tree under the process of training and generate individual trees for classification. To train the algorithm, a bootstrap aggregation mechanism is applied A training set $X = x1, x2, ...xn$ with $Y = y1, y2, ...yn$ bagging N times by randomly selected sample and applies trees to it. For $n = 1, 2, 3, ...N$ Replacement Training Sample ,$X, Y = (Xn)(Yn)$ Training classification of tree of fn on Xn, Yn
Predictions from every individual regression trees on x‘
$\widehat{f} = \frac{1}{N}\Sigma_{n=1}^{N} fn(x‘)$
Estimating of uncertainty, as the standard deviation can be taken for the prediction of all individual regression trees on$x$‘
$\sigma = \sqrt{\frac{\Sigma_{n=1}^{N}(fn(x‘)-\widehat{f})^2}{N-1}}$

B. K Nearest Neighbor

The K Nearest Neighbor(KNN)[16] is supervised machine learning algorithms which are able to solve the problems based on classifications. In the method the same kind of objects are placed near by to each other. To make this algorithm work efficiently, one has to be care full to choose the value of K. As the stability of the model depends on it, if K decreases, the model‘s stability increases. Also the increases of K value support the model stability but if it goes out of a certain fixed limit then, the error also increases. So after selecting the optimum value of k, the process of load, initiate and acquire result will be started and from the resultant output we can take mean of the data as output.

C. Naive Bayes

The Naive Bayes[10], a supervised machine learning algorithm works the probabilistic models. In this model, For every class, probability is calculated and determine their classification, which again used for a new class to predict the values. Here, x is a instance of a problem which has to be classified. A vector can represent it by $x = x1, x2, ...xn$ where n represents independent variables, and assigned to instance probabilities $p(cx/(x1, x2, ...xn))$ For each of K possible outcomes or classes ck

$p(ck|x) = \frac{p(ck)p(X|ck)}{p(X)}$

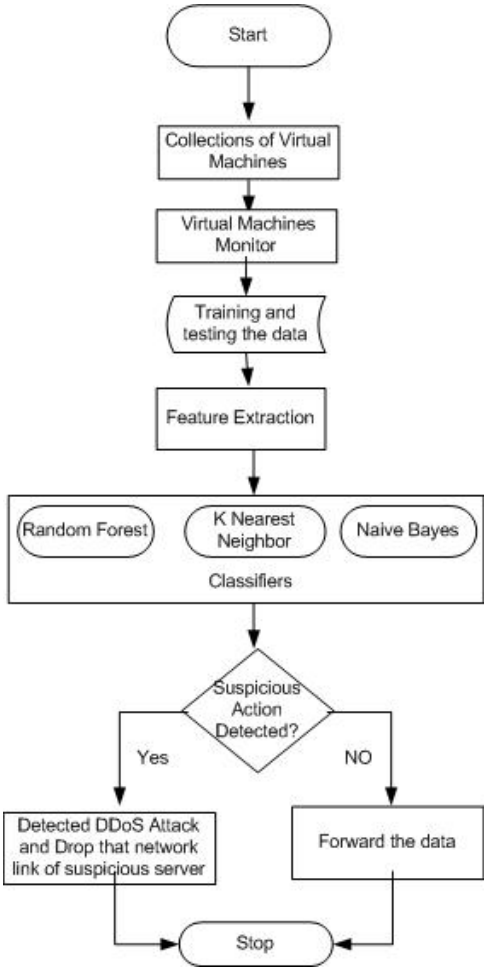

Fig. 2. Proposed Approach.

IV. Proposed Approach

Here, In the algorithm, we join all the features vectors of some of interest to generate a single long feature vector and forward it to the Machine Learning engine. Then the engine used a PreTrained learning module to detect the DDoS attacks. In the proposed system, we have also used Online module in the background, which is used to classifying the feature vector. If a predefined number p of these modules classify this feature vector F as an authenticated or suspicious vector, then it is used to update the main module.
Table 1 shows the algorithm for detection of DDoS attack. The used symbols are:
n:the number of statistical features
p:the number of servers
qi:the number of Virtual machine on every i server
SFij: Join all features vectors for all servers

V. Implementation and Performance Evaluation

For experimental set up, we have used three servers and the virtual machines are running on every server.All the servers has 64 GB memory and all of them are running Ubuntu 18.04

TABLE I
ALGORITHM FOR FEATURE GENERATION AND CLASSIFICATION

Feature selection form virtual machine monitor $F1, F2...Fn$
for collection of VM Servers $VMSij$, $i = 1, 2, 3, ...p$ and
$j = 1, 2, 3, ...qi$ do
$\forall$i virtual machine monitor reports for the statistical
features SFij of VMSij where $SFij = F1ij, F2ij, F3ij, ...Fnij$
forward the extracted features to the classifiers
if action is suspected based on predefined rules then
attack detected and the services from that server is disconnected
otherwise
packet is normally forwarded
end
end

TABLE II
RESULTS FOR DETECTION OF DDOS ATTACK USING DISTINGUISH MACHINE LEARNING ALGORITHMS

Algorithms	Acc.(%)	FP(%)	FN(%)	Prec.(%)	Recall(%)	F1-Score
Naive Bayes	93.58	0.00	6.01	100.00	92.37	0.9302
KNN	97.69	0.00	6.49	99.29	91.70	0.9609
Random Forest	99.68	0.89	6.44	99.69	94.51	0.9660

operating system. Here, to evaluate our performance, we used several metrics as Accuracy which presents the percentage of actual detection over tested samples. False Positive (FP) and False Negative (FN) presents false alarms, respectively. The Precision metric presents that how much portion of alarms are actually true alarms. Recall metric shows the detected portion of attacks.The F1 score metric is used to make a balance between FP and FN. As the higher F1 score increases, the performance of that algorithm is also increases.Table 2 and table 3 shows all the results. The Precision, Recall and F1 score can be defined as:

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

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

$$F1 = 2 * \frac{Precision \times Recall}{Precision + Recall}$$

In supervised algorithms, we almost achieve 99% accuracy and F1 scores as approx. 0.96. Form the table 2, One can see that the Random Forest is the best, with having 99.68% accuracy, precision 99.69%, 0.9660 F1-score and with best recall 94.51, that helps to detects the most of the attacks. Also table 3, which gives the combined results of three vms shows that Random Forest gives the best result 99.58% on running our algorithm over it. The results can be verified by the figure 4. Figure 3 depicts the tcpsyn attack scenario which works on three handshaking concept. it shows the attack and non attack ration as per the time that how much chaos it can create.

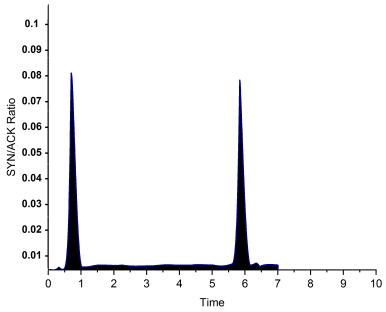

Fig. 3. TCP SYN attck.

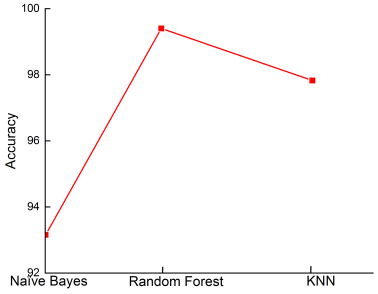

Fig. 4. ROC Curve.

TABLE III
JOINT RESULTS FOR DETECTION OF DDOS ATTACK USING THREE VMS

Algorithms	Acc.(%)	FP(%)	FN(%)	Prec.(%)	Recall(%)	F1-Score
Naive Bayes	93.69	0.00	6.06	100.00	92.89	0.9319
KNN	97.89	0.00	7.00	99.19	91.81	0.9619
Random Forest	99.58	0.80	6.38	99.68	94.49	0.9659

VI. CONCLUSION

In our paper, we proposed a machine learning based system to detect and prevent DDoS attacks at server over the cloud. We used extraction of statistical features to perform our proposed approach. From the results, presented in table 2 and table 3, it can be said that the proposed approach can detect DDoS attacks with a high accuracy approx. (99.68%) and low false positives.

In our research work, we majorly focused on the supervised learning algorithm, so the unsupervised or reinforcement learning can be the research work for future.

REFERENCES

[1] Kaushik, Shweta, and Charu Gandhi. "Ensure Hierarchal Identity Based Data Security in Cloud Environment." International Journal of Cloud Applications and Computing (IJCAC) 9.4 (2019): 21-36.

[2] Tewari, Aakanksha, and B. B. Gupta. "Security, privacy and trust of different layers in Internet-of-Things (IoTs) framework." Future generation computer systems 108 (2020): 909-920.

[3] J. Hou, P. Fu, Z. Cao and A. Xu, "Machine Learning Based DDos Detection Through NetFlow Analysis," MILCOM 2018 - 2018 IEEE Military Communications Conference (MILCOM), Los Angeles, CA, 2018, pp. 1-6.
[4] 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.
[5] N. Pise and P. Kulkarni, "Algorithm selection for classification problems," 2016 SAI Computing Conference (SAI), London, 2016, pp. 203-211
[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 (QRS-C), Sofia, Bulgaria, 2019, pp. 471-477.
[7] U. Dincalp, M. S. Güzel, O. Sevine, E. Bostanci and I. Askerzade, "Anomaly Based Distributed Denial of Service Attack Detection and Prevention with Machine Learning," 2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Ankara, 2018, pp. 1-4.
[8] P. Khuphiran, P. Leelaprute, P. Uthayopas, K. Ichikawa and W. Watanakeesuntorn, "Performance Comparison of Machine Learning Models for DDoS Attacks Detection," 2018 22nd International Computer Science and Engineering Conference (ICSEC), Chiang Mai, Thailand, 2018, pp. 1-4.
[9] R. Doshi, N. Apthorpe and N. Feamster, "Machine Learning DDoS Detection for Consumer Internet of Things Devices," 2018 IEEE Security and Privacy Workshops (SPW), San Francisco, CA, 2018, pp. 29-35.
[10] Bhushan, Kriti, and Brij B. Gupta. "Distributed denial of service (DDoS) attack mitigation in software defined network (SDN)-based cloud computing environment." Journal of Ambient Intelligence and Humanized Computing 10.5 (2019): 1985-1997.
[11] Alsmirat, Mohammad A., et al. "Impact of digital fingerprint image quality on the fingerprint recognition accuracy." Multimedia Tools and Applications 78.3 (2019): 3649-3688.
[12] Dahiya, Amrita, and B. B. Gupta. "Multi attribute auction based incentivized solution against DDoS attacks." Computers & Security 92 (2020): 101763.
[13] Al-Qerem, Ahmad, et al. "IoT transaction processing through cooperative concurrency control on fog–cloud computing environment." Soft Computing 24.8 (2020): 5695-5711.
[14] K. Verma and A. K. Ghosh (eds.), Computational Intelligence: Theories,Applications and Future Directions—Volume I, Advances in Intelligent Systems and Computing 798 © Springer Nature Singapore Pte Ltd. 2019
[15] Dahiya, Amrita, and Brij B. Gupta. "A reputation score policy and Bayesian game theory based incentivised mechanism for DDoS attacks mitigation and cyber defense." Future Generation Computer Systems (2020).
[16] Al-Sharif, Ziad A., et al. "Live forensics of software attacks on cyber–physical systems." Future Generation Computer Systems 108 (2020): 1217-1229.