International Journal of Current Science Research
www.drbgrpublications.in

Machine Learning Based on Classification for Detection of DDoS Attacks in Cloud Computing

A.Nazreen Banu

Assistant Professor, Vel Tech Multi Tech Dr.Rangarajan Dr.Sagunthala Engineering College, Chennai, India.
*Corresponding author E-mail id: Nazreenbanu8@gmail.com

Abstract

DDoS (Distributed Denial of Service) is a network security assault, and attackers have now infiltrated practically every technology, including cloud computing, IoT, and edge computing, to strengthen themselves. According to DDoS behaviour, the attacker consumes all available resources such as memory, CPU, and perhaps the entire network in order to bring down the victim's system or server. Though several defensive mechanisms have been developed, they are ineffective since attackers are educated by the newly accessible automated assaulting tools. As a result, we suggested a machine learning strategy based on categorization for detecting DDoS attacks in cloud computing. The mechanism can identify a DDoS assault with accuracy using three classification machine learning algorithms: K Nearest Neighbor, Random Forest, and Naive Bayes.

Keywords: *DDoS, Machine learning, Cloud computing*

Introduction

Instead of employing a single infected system, distributed denial of service (DDoS) assaults employed a collection of compromised machines to attack the target workstation. Attackers infiltrate innocent devices, often known as bots or zombies, and direct them to launch an assault on the victim's computer. Today, we are confronted with a plethora of assaults that deplete available network infrastructures and compromise security features such as confidentiality, integrity, and availability. DDoS attacks, in general, occur at the network and application layers by draining network and server services, respectively. In our article, we focused on DDoS assaults that are centered on the network layer and originate from virtual machines in the cloud. According to McAfee Lab research, it has a significant global influence. DDoS attacks have the ability to destroy, impair, or render inaccessible a single website, or a group of servers, such as email or http servers. To initiate the assault on commodity communities' cloud, the attackers used a fraudulent payment card to rent virtual machines.

Related Works

There are several ways that are based on a variety of concepts and algorithms. Though some of them are quite accurate, they fail because they increase the computing time required to identify a DDoS assault. Let's take a look at what's already been done. The net flow protocol was addressed by the authors in Somani *et al*., 2017. This protocol has a very short data retrieval time; however, it only supports Cisco network devices like Cisco routers. Machine learning-based defensive methods have been proposed by the authors. The authors created an Intrusion Detection System that worked for K Nearest Neighbor (KNN), MLP, and Decision Tree, among other machine learning techniques (DT). There are other issues, such as the fact that MLP requires a large quantity of data to train and that the sophisticated method increases processing time. Because DT cannot manage noise, the authors decided that KNN is the ideal algorithm because it can handle noise and give excellent accuracy. The authors employed the arrival rate for various packets, as well as the DBSCAN method for clustering the data, therefore this study work's disadvantage is that it cannot execute over live collected data. The authors Chung *et al*., 2014 shown how several machine learning classification algorithms function and how well they perform. The authors argue that Random Forest (RM) is superior to KNN in terms of accuracy [Breitenbacher *et al*., 2018], however they do not account for all types of DDoS attacks. As a result of our review of several research papers, we discovered that RM, naive bayes, and KNN are all viable options for use with our suggested algorithms.

TCP SYN Flood attacks—The spoofing of IP addresses is known as a TCP SYN Flood attack. Because it is dependent on the 3-way handshake protocol, this attack is more susceptible.

• PING Flood Attacks—PING attacks rely on ICMP request packets. The connection slows down when the PING attack targets the system, and end-user reply request packets cannot be transmitted.

• UDP Flood Attacks—Once the threshold limit is reached, the target system is unable to process permitted connections. Other packet requests are dropped once the servers approach their threshold limitations.

Emerging Need for DDoS Attack Detection in Cloud Environments

The subject of DDoS attack prevention, detection, and mitigation has gained a lot of traction in the Cloud computing world. Researchers have given the problem of DDoS attack detection the highest priority among these three difficulties. Researchers from all around the world have been experimenting with new methodologies and ways to identify DDoS assaults. Despite the existence of different contributions that addressed strategies and techniques for preventing DDoS attacks, the deployment of the current ways could not, sadly, resist DDoS assaults impacting Cloud settings even today. In reality, the number of assaults, as well as the size of the attacks, has been steadily growing over time. One of the most prominent causes is that there is no agreement among multiple end points in a dispersed internet network because global collaboration is impossible to impose.

According to statistics from Amazon Web Services, the largest DDoS assault to date occurred in the month of February in the year 2020. This attack's peak incoming traffic is estimated to reach 2.3 Tbps. Hijacked CLDAP webservers (Connection-less Lightweight Directory Access Protocol webservers), an alternative to LDAP and a protocol for processing user directories, were exploited by the attackers in this attack. The 1.3 Tbps DDoS assault, which targeted GitHub and sent 126.9 million traffic packets per second, was the second greatest DDoS attack before this 2.3 Tbps DDoS strike in February of 2020.

DDoS Attack Detection Framework Using Multiple Linear Regression

The suggested method's main goal is to create a machine learning model based on multiple linear regression analysis, as well as to perform data visualization using residual plots and fit charts. The goal of the suggested technique is to see if multiple linear regression analysis can be applied to the CICIDS 2017 dataset, which is a widely used benchmark dataset in some of the most recent research projects. The goal is to use the feature selection approach to identify the relevant features that will help the prediction model give better results. We employed the Information Gain approach method to conduct out feature selection in the current approach. In numerous data mining-based applications, the Information Gain technique is a frequently used paradigm. The behavior of the chosen and retained essential characteristics of the CICIDS 2017 dataset is analyzed by assessing the fit charts and residual plots, and the selected features are then evaluated for multiple linear regression analysis.

Experiment and Result

The experimental dataset comprised of five-day log recordings in csv format from Monday to Friday. We used the log file from Friday afternoon for experiment analysis, which included two class labels.

The qualities at dimensions 1, 5, 6, 7, 9, 11, 13, 35, 36, 53, 54, 55, 56, 64, 66, and 67 are preserved after the initial analysis is completed. As a result, the analysis is carried out using the reduced dimensionality log file with the 16 attributes listed above. For the linear regression model, the mean absolute percentage error is 0.2621. As a result, the multiple linear regression model's percentage accuracy is calculated to be 73.79 percent, or 0.7379.

Table1: P value and confidence intervals – ANNOVA for CICIDS2017 DATASET

ANOVA	Df	SS	Ms	F	Significance f
Regression	16	29778.18946	1861.136841	23832.54328	0
Residual	225733	25640.72297	0.113588722		
Total	225749	55418.91243			

	coefficients	Standard Error	t stat	p-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
intercept	1.4801631	1.21497E -4	1.2182773E -5	0	1.47781E 0	1.482544E 0	1.47781E 0	1.482544E 0
Destination Port	-7.9586E-06	5.07513E -08	- 1.568157E2	0	- 8.05807E-o6	-7.85913E-06	-8.0580E-06	-7.85913E-06
Total Length of FWD Packets	0	0	6.5535E4	#NUMI	0	0	0	0
Total length of BWD Packets	3.0984E-06	6.49544E -08	4.77018E1	#NUMI	2.9711E-06	3.2257E-06	2.9711E-06	3.2257E-06
FWD packet length MAX	8.6460E-06	1.7102E-06	7.3885E0	1.4899E -13	6.3524E-06	1.0939E-05	6.3524E-06	1.09396E-05
FWD Packet length Mean	0	0	6.553E4	#NUMI	0	0	0	0
BWD Packet length Max	-5.9706E-4	5.71258E -06	-1.04517E2	#NUMI	1.49401E-06	3.2256E-07	1.49401E-06	3.2256E-07
BWD Packet Length Mean	7.066E-05	3.26681E -06	2.16188E1	1.523E-103	6.4247E-05	7.70585E-05	6.4247E-05	7.70585E-05
FWD Header Length	0	0	6.553E5	#NUMI	0	0	0	0
BWD Header Length	-5.97066E-4	5.7125E-06	4.0913E0	#NUMI	-6.082E-4	-5.858E-4	-6.082E-4	-5.858E-4
Average Packet Size	2.8167E-4	4.077E-06	2.6188E1	0	2.738E-4	2.898E-4	2.738E-4	2.898E-4
AVG FWD segment Size	-1.4937E-4	4.891E-06	-3.5535E5	2.080E-204	-1.5894-4	-1.3979E-4	-1.5894-4	-1.3979E-4
AVG BWD Segment Size FWD Header Length	0 4.1925E-4	0 7.0436E-06	6.555E4 5.9249E1	#NUMI #NUMI	0 4.054E-4	0 4.3305E-4	0 4.054E-4	0 4.3305E-4
Subflow FWD Bytes	-3.81369E-06	4.8253E-07	-7.9034E0	2.724E-15	-4.758E-07	-2.867E-06	-4.758E-07	-2.867E-06
Subflow BWD Bytes Init_Win_bytes_forward	0 -1.2476E-05	0 9.9195E-08	6.553E4 -1.2577E2	#NUMI #NUMI	0 -1.267E-05	0 -1.228E-05	0 -1.267E-05	0 -1.228E-05

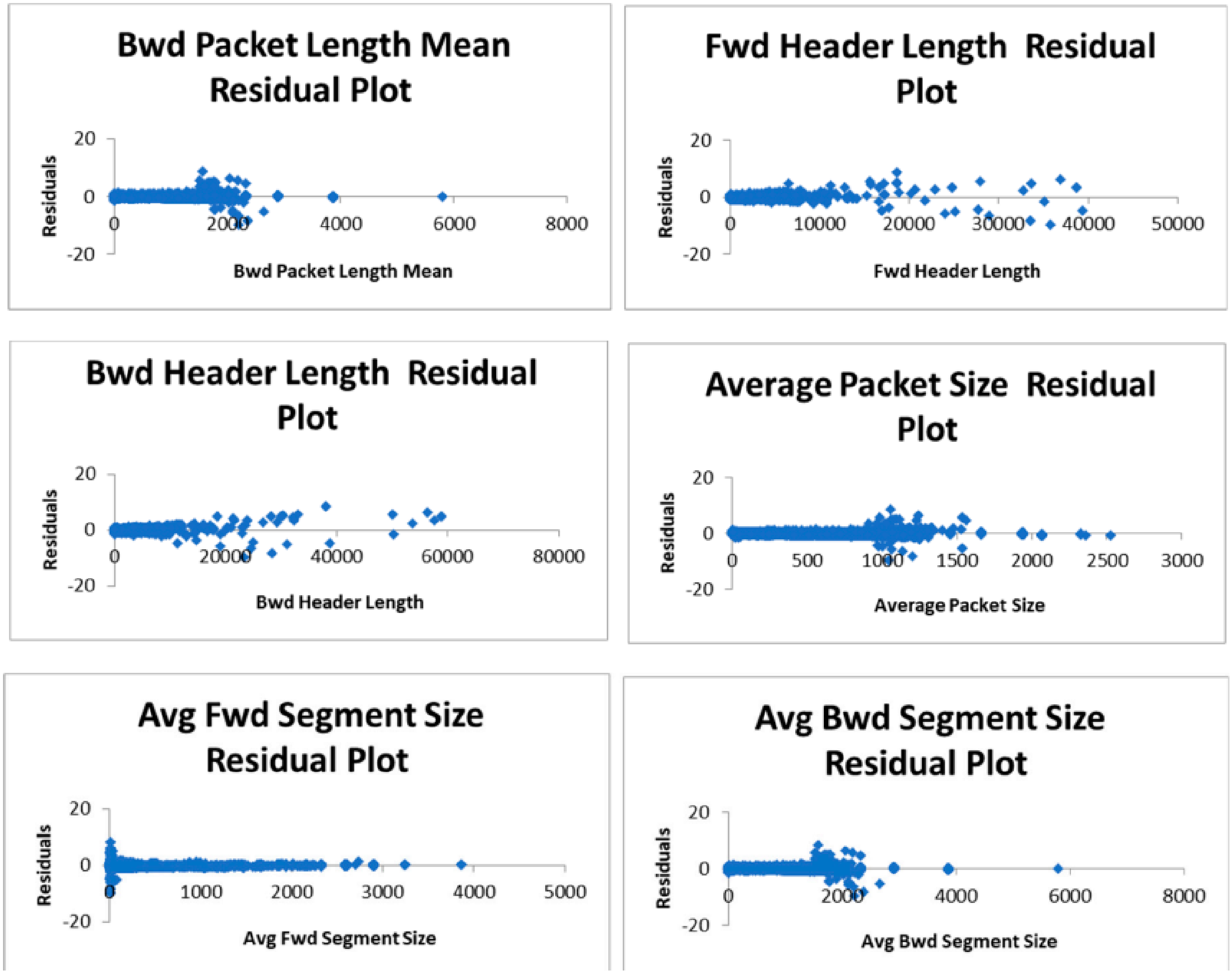

Figure 2: Residual plots log of CICIDS 2017 dataset

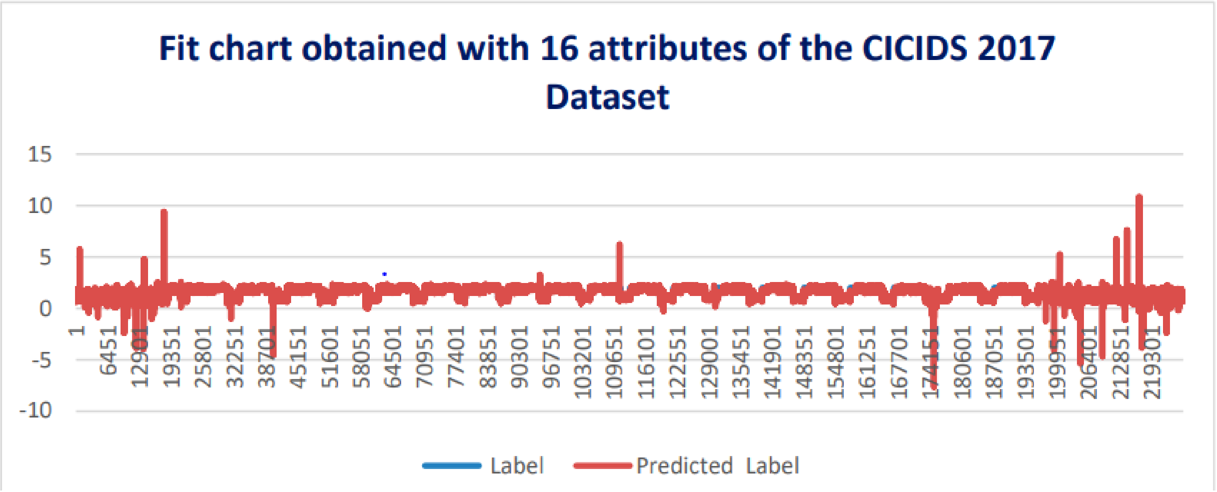

Figure 3: Fit chart obtained with CICIDS data set

The fit chart obtained by performing multiple regression analysis on the CICIDS 2017 dataset with the top 16 attributes obtained through information gain-based feature selection method is better than the fit chart obtained by considering six attributes of the CICIDS 2017 dataset, as shown in the experiment result analysis. This is due to the fact that the later fit takes into account six dataset properties.

Conclusion

As DDOS attack detection has grown increasingly widespread in a distributed setting such as the Cloud, it is critical to identify assaults that cause Cloud service unavailability. Machine learning models may be used to train and test attack detection datasets in order to detect such assaults. Alternatively, we can utilise regression analysis by employing one of its most important versions, multiple linear regression analysis. The goal of this work is to create a machine learning model that uses information gain and regression analysis to create an ensemble of feature selection. The dataset used in the experimental investigation was the well-known CICIDS 2017 dataset. This study has therefore opened the road to demonstrate the relevance of regression analysis in the construction of an ML model, as well as certain key visualisations such as residual plots and fit charts, which demonstrate the model's importance and appropriateness for prediction. In this study, the analysis was confined to a one-day log file; however, in the future, this research might be expanded to include all five-day traffic log files and produce a consensus-based machine learning model.

References

Dayanandam, G.; Reddy, E.S.; Babu, D.B. Regression algorithms for efficient detection and prediction of DDoS attacks. In Proceedings of the 2017 3rd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), Tumkur, India, 21–23 December 2017; pp. 215–219.

Sharma, N.; Mahajan, A.; Mansotra, V. Machine Learning Techniques Used in Detection of DOS Attacks: A Literature Review. *Int. J. Adv. Res. Comput. Sci. Softw. Eng*. 2016, 6, 100.

Somani, G.; Gaur, M.S.; Sanghi, D.; Conti, M.; Buyya, R. DDoS attacks in cloud computing: Issues, taxonomy, and future directions. *Comput. Commun*. 2017, 107, 30–48.

Perera, P.; Tian, Y.-C.; Fidge, C.; Kelly, W. A Comparison of Supervised Machine Learning Algorithms for Classification of Communications Network Traffic. In International Conference on Neural Information Processing; Springer: Cham, Switzerland, 2017; pp. 445–454.

Zammit, D. A Machine Learning Based Approach for Intrusion Prevention Using Honeypot Interaction Patterns as Training Data. Bachelor’s Thesis, University of Malta, Msida, Malta, 2016.

Doshi, R.; Apthorpe, N.; Feamster, N. Machine Learning DDoS Detection for Consumer Internet of Things Devices. arXiv 2018, arXiv:1804.04159.

Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput*. 1997, 9, 1735–1780.

Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. In Proceedings of the NIPS 2014 Deep Learning and Representation Learning Workshop, Montreal, QC, Canada, 12 December 2014.

Breitenbacher, D.; Elovici, Y. N-BaIoT—Network-Based Detection of IoT Botnet Attacks Using Deep Autoencoders. *IEEE Pervasive Comput*. 2018, 17, 12–22.

Zekri, M.; El Kafhali, S.; Hanini, M.; Aboutabit, N. Mitigating Economic Denial of Sustainability Attacks to Secure Cloud Computing Environments. *Trans. Mach. Learn. Artif. Intell*. 2017, 5, 473–481.

Liao, Q.; Li, H.; Kang, S.; Liu, C. Application layer DDoS attack detection using cluster with label based on sparse vector decomposition and rhythm matching. *Secur. Commun. Netw*. 2015, 8, 3111–3120.

Xiao, P.; Qu, W.; Qi, H.; Li, Z. Detecting DDoS attacks against data center with correlation analysis. *Comput. Commun*. 2015, 67, 66–74.

Karimazad, R.; Faraahi, A. An anomaly-based method for ddos attacks detection using rbf neural networks. In Proceedings of the International Conference on Network and Electronics Engineering, Hong Kong, China, 25–27 November 2011; pp. 16–18.

Zhong, R.; Yue, G. Ddos detection system based on data mining. In Proceedings of the 2nd International Symposium on Networking and Network Security, Jinggangshan, China, 2–4 April 2010; pp. 2–4.

Wu, Y.-C.; Tseng, H.-R.; Yang, W.; Jan, R.-H. Ddos detection and traceback with decision tree and grey relational analysis. *Int. J. Ad Hoc Ubiquitous Comput*. 2011, 7, 121–136.