A Novel Approach for DDoS Attack Detection Using Big Data and Machine Learning

Document Version
Publisher's PDF, also known as Version of record

A Novel Approach for DDoS Attack Detection Using Big Data and Machine Learning

Akshat Gaurav*[1], Zhili Zhou[2], Kwok Tai Chui*[3], Francesco COLACE[4],
Priyanka Chaurasia[5] and Ching-Hsien Hsu*[6]

[1]Ronin Institute, Montclair, New Jersey 07043, U.S.

[2]School Engineering Research Center of Digital Forensics, Ministry of Education, School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing 210044, China.

[3] Hong Kong Metropolitan University (HKMU) , Hong Kong

[4]University of Salerno, Italy

[5]Ulster University, Magee campus, Londonderry, UK

[6] Department of Computer Science and Information Engineering, Asia University, Taiwan
& Department of Computer Science and Information Engineering, National Chung Cheng University, Taiwan
& Department of Medical Research, China Medical University Hospital, China Medical University, Taiwan
*Corresponding Author

Abstract

Due to the developemt in the latest digital technologies, internet service use has surged recently. In order for these online businesses to succeed, they must be able to consistently and effectively supply their services. As a result of the DDoS assault, online sources are impacted in terms of both their availability and their computational capacity. DDoS attacks are useful for cyber-attackers since there is no effective techniqque for the identification of them. In recent years, researchers have been experimenting with duffernet latest techniques like machine learning (ML) approaches to see whether they can build effective methods for detecting DDoS assaults.Machine learning and big data are used to identify DDoS assaults in this research paper.

Keywords

DDoS attack, Machine learning, Big data

1. Introduction

When a huge number of malicious computers assault the victim's resources in a coordinated fashion, it is known as a DDoS attack. Assault programmes such as Slowloris, GoldenEye, and others make it easy for anybody to launch a DDoS attack on a target and wreak havoc on their resources or make their bandwidth inaccessible to others [1]. DDoS assaults come in a variety of forms, making it difficult for the detection filter to keep up [2]. When an attacker sends a high number of SYN packets to the victim's end in order to overwhelm the connection table, this is known as TCP flooding. There are also UDP and HTTP flooding attacks that use

Figure 1: DDoS attack architecture

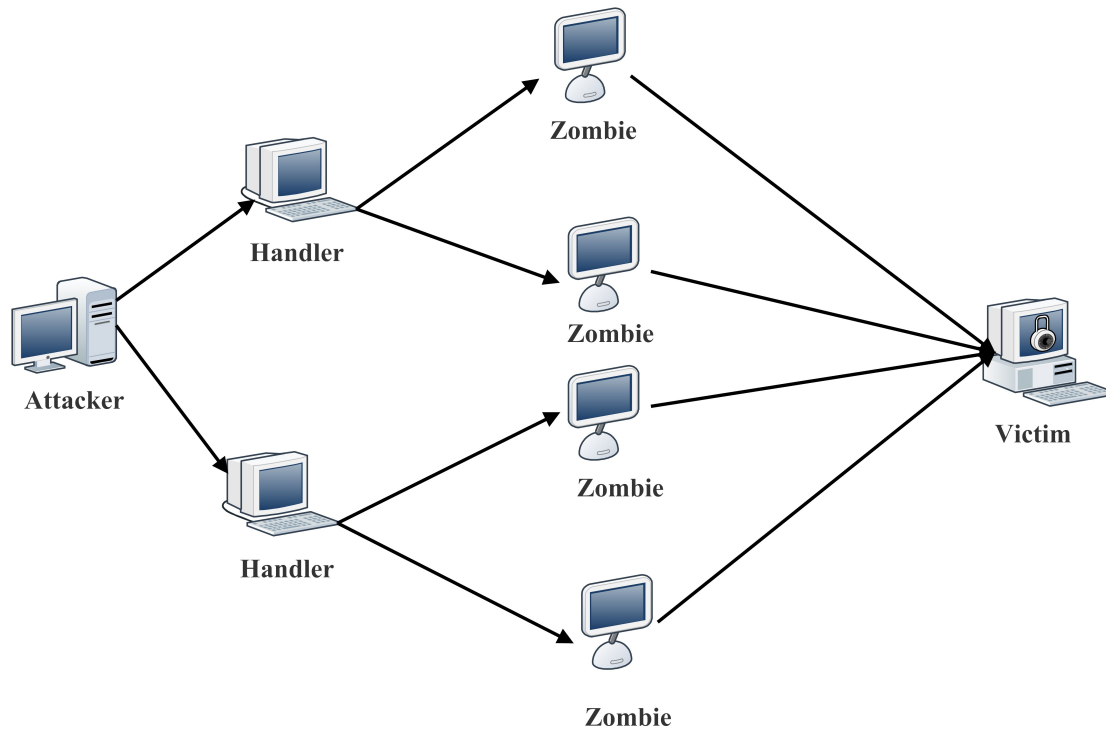the victim's bandwidth and prevent legitimate users from accessing it. Detection of DDoS attacks may be broken down into three categories: preventive methods, defensive methods, and traceback methods. DDoS attacks may be mitigated using a variety of approaches, including load balancing, honeypots, and more. In DDoS attack traceback procedures, a variety of methods are used to locate the assault's origin. DDoS defensive techniques must be developed that not only identify the specific DDoS assault, but also take suitable countermeasures against it. It is possible to distinguish a DDoS assault from flash crowd traffic by using a decent DDoS protection strategy. As many legal people try to access an online resource, this traffic is known as flash crowd. Because of this, an effective DDoS detection algorithm must be able to distinguish the flash crowd from the DDoS assault traffic and not discard it as part of the attack traffic. There are several machine learning algorithms being used to identify DDoS assaults and flash crowds because of recent advancements in the area of machine learning. With the use of machine learning algorithms and attack patterns, it's feasible to train security filters to block new forms of threats. The supervised learning approach and the unsupervised learning approach are two of the strategies available in machine learning for detecting aberrant traffic. Unsupervised learning, on the other hand, relies on labelled data sets that are difficult to get, whereas supervised learning relies on data sets that are labelled.

In this paper, we proposed a big data-based method for the detection of DDoS attacks and flash crowds. The contributions of this paper are as follows:

- We used the dataset generated by OMNET++ for training and testing the machine learning model.
- Our processing performance has been boosted thanks to Apache Spark, which we utilise for data processing.
- The performance of our suggested model is evaluated using standard statical metrics.

2. Related Work

An adaptive density-based clustering technique (ADBSCAN) developed by Li et al. [3] is based on closest neighbour graphs. KNN density estimation and distributional assumptions are used in the proposed method to quickly identify various density clusters. Samples in dense areas are found using the KNN estimator, and then the statistical technique is used to determine the clusters' densities, K-means clustering was developed by Gu et al. [4] for the identification of DDoS attacks. Handloop-based feature selection is utilised to detect DDoS assault characteristics properly in the proposed methodology. In K-means clustering, these characteristics are utilised to distinguish regular traffic from malicious traffic.

Zombies are counted using an artificial neural network (ANN) in the proposed technique by the authors in [5]. Low frequency attacks can be accurately predicted with this strategy since it doesn't rely on attack frequency. NS-2, a network simulator for Linux, is used to produce the training data for feed forward neural networks. MSE is used to compare the estimate performance of various feed forward networks. The network's ability to anticipate the number of zombies engaged in a DDoS assault with extremely low test error is encouraging.

As part of the proposed method in [6] for DDoS detection, the authors offer a unique architecture that monitors traffic changes inside the ISP Domain and then classifies the network flows that convey attack traffic. Detection of DDoS assaults relies on two statistical metrics: volume and flow. The precision of threshold value choices has a significant impact on the effectiveness of a system for detecting and characterising anomalies. When threshold values are set too high or too low, a great many tests will return erroneous results. Six-Sigma and variable tolerance factor approaches are employed in the proposed strategy in order to properly and dynamically establish threshold values for a wide range of statistical measures. It is used as a testbed to evaluate the efficacy of the suggested technique on a Linux platform. There are several assault scenarios, each with a varied quantity and attack strength of zombie machines. Authors in [7] represents the imapct of DDoS attack on IoT devices. Also, in [8] author proposed captch method for the identification of DDoS attack.

It is difficult to discern legal traffic from attack traffic during DDoS attack. Wireless networks are especially vulnerable to a DDoS assault because of the nature of ad hoc networks. Rather of allowing the DDoS attack to occur and then taking the required actions to deal with it, it is preferable to prevent it from happening in the first place. The author in [9] address how MANET might be damaged by DDoS assaults in their article. Besides this, an unique DDoS mitigation strategy is suggested for MANETs.

DDoS attacks may be detected using Gu et al [10] semi-supervised K-means clustering. DDoS attacks may be identified using three primary elements retrieved from the datasets in the suggested technique. The k-means clustering procedure is accurate because the extracted

features are utilised to label samples in the data sets. Using a semi-supervised technique, the clustering algorithm has a high convergence rate.

Authors in [11] proposed DDoS attack detction technique for healthcae services. Also, the authors propsoed DDoS detection techniue in cloud environment[12] and VANET envieonment [13].

Using density-based semi-supervised clustering, Gertrudes et al [14] suggested a new approach. There are many semi-supervised clustering techniques that are used in the suggested method. The author depicts the link between graph-based techniques and density-based approaches as well. No re-computation and ordered-dependencies are present in the proposed framework compared to prior semi-supervised techniques.

3. Proposed approach

This section of the article discusses the solution we offer. We've put the plan into action on the routers. For each time period ∇ t, routers extract the incoming traffic characteristics. Afterwards, the obtained attributes are fed into the aforementioned machine learning model. This classification is made by the machine learning algorithms as soon as a packet is found to be malicious. Finally, all malicious packets are discarded by the router.

3.1. Preprocessing phase

For our suggested technique, the first step in developing it is to pick a training data set. Few datasets are available for testing, however, since datasets comprise personal and secret information about users, and revealing it publicly would violate privacy restrictions. For the dataset creation, we utilised the ONMET++ programme.

With Apache Spark, we can preprocess large datasets quickly and easily. Apache Spark is a memory-based distributed computing system that uses RDDs as its primary data structure. The RDD is a distributed, immutable storage system that may be built in several stages. An RDD is divided into various sections. The number of the divisions determines the granularity of RDD calculation, with each RDD partition being computed in a separate job, as shown Figure 2.

4. Result and Discussion

A 64-bit Intel processor with 16GB of RAM and Keras and TensorFlow are used to reproduce the machine learning model. A discrete event simulator, OMNET++, is used to simulate DDoS attacks as well as normal traffic for our proposed technique. The entropy value is calculated for the chosen time period by running OMNET++ for 100 seconds. Machine learning techniques are explained in detail below after preprocessing the entropy values provided by OMNET++. As many null values are included in the dataset, they are omitted throughout the data preparation process. Last but not least, the data set is divided into two sections: training and testing, so that the proposed approach can be thoroughly examined. We used traditional statistical methods to assess the performance of our machine learning model. Figure 3, represents the performance of our proposed approach.
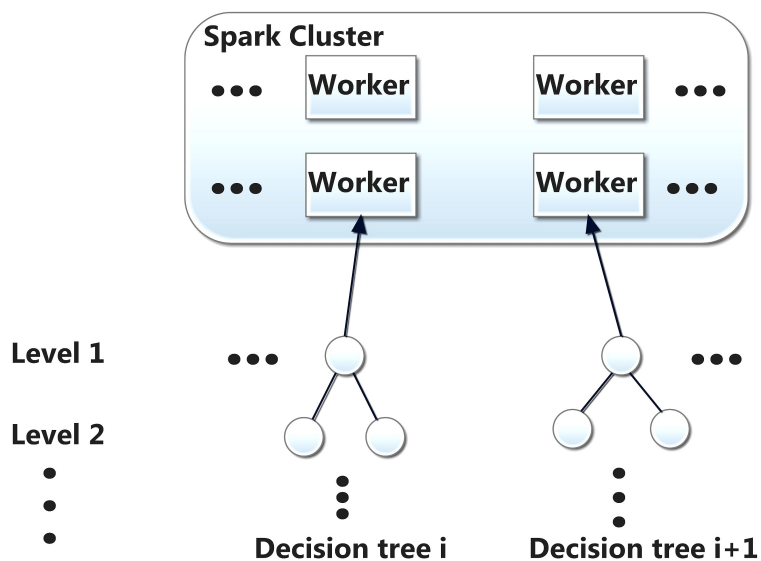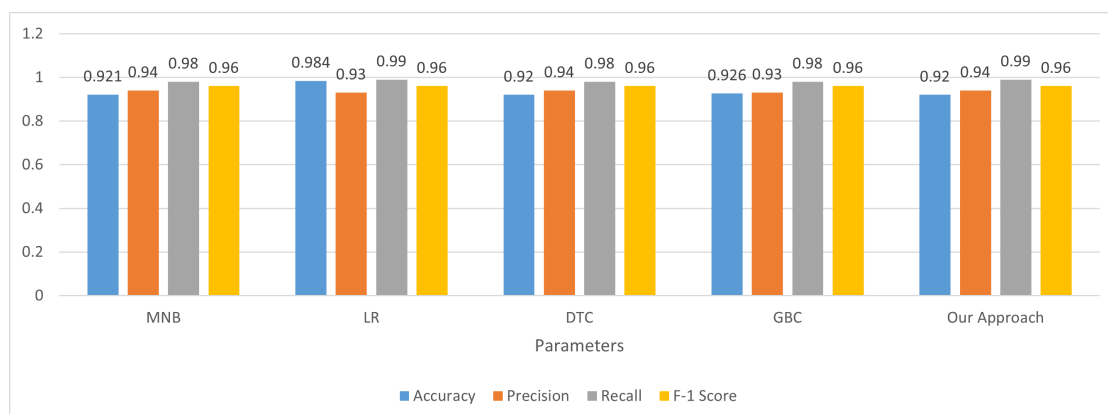
Figure 2: Apache Spark

Figure 3: Performance of our propsoed approach

5. Conclusions and future work

In this research, we offer a strategy based on big data and machine learning for detecting DDoS attacks. In the suggested solution, we employed OMNET++ to create traffic data, followed by Apache Spark for data preparation, and ultimately, machine learning techniques to detect malicious packets. We use statistical techniques to determine the accuracy of our suggested methodology, and the findings indicate that our proposed strategy easily identifies and separates DDoS attacks from flash crowd traffic.

References

[1] A. Srivastava, B. Gupta, A. Tyagi, A. Sharma, A. Mishra, A recent survey on ddos attacks and defense mechanisms, in: International Conference on Parallel Distributed Computing Technologies and Applications, Springer, 2011, pp. 570–580.

[2] S. Tripathi, B. Gupta, A. Almomani, A. Mishra, S. Veluru, Hadoop based defense solution to handle distributed denial of service (ddos) attacks (2013).

[3] H. Li, X. Liu, T. Li, R. Gan, A novel density-based clustering algorithm using nearest neighbor graph, Pattern Recognition 102 (2020) 107206.

[4] Y. Gu, K. Li, Z. Guo, Y. Wang, Semi-supervised k-means ddos detection method using hybrid feature selection algorithm, IEEE Access 7 (2019) 64351–64365.

[5] B. B. Gupta, R. C. Joshi, M. Misra, Ann based scheme to predict number of zombies in a ddos attack., Int. J. Netw. Secur. 14 (2012) 61–70.

[6] B. B. Gupta, M. Misra, R. C. Joshi, An isp level solution to combat ddos attacks using combined statistical based approach, arXiv preprint arXiv:1203.2400 (2012).

[7] B. B. G. T A. Dahiya, How iot is making ddos attacks more dangerous? (2021).

[8] D. Singh, Captcha improvement: Security from ddos attack (2021).

[9] M. Chhabra, B. Gupta, A. Almomani, A novel solution to handle ddos attack in manet (2013).

[10] M. E. Ahmed, S. Ullah, H. Kim, Statistical application fingerprinting for ddos attack mitigation, IEEE Transactions on Information Forensics and Security 14 (2018) 1471–1484.

[11] Z. Zhou, A. Gaurav, B. Gupta, H. Hamdi, N. Nedjah, A statistical approach to secure health care services from ddos attacks during covid-19 pandemic, Neural Computing and Applications (2021) 1–14.

[12] A. Gaurav, B. Gupta, C.-H. Hsu, D. Peraković, F. J. G. PEÑALVO, Filtering of distributed denial of services (ddos) attacks in cloud computing environment, in: 2021 IEEE International Conference on Communications Workshops (ICC Workshops), IEEE, 2021, pp. 1–6.

[13] A. Gaurav, B. Gupta, F. J. G. Peñalvo, N. Nedjah, K. Psannis, Ddos attack detection in vehicular ad-hoc network (vanet) for 5g networks, in: Security and Privacy Preserving for IoT and 5G Networks, Springer, 2022, pp. 263–278.

[14] J. C. Gertrudes, A. Zimek, J. Sander, R. J. Campello, A unified framework of density-based clustering for semi-supervised classification, in: Proceedings of the 30th International Conference on Scientific and Statistical Database Management, 2018, pp. 1–12.