EasyChair Preprint

№ 9235

Prediction of DDoS Attacks Using DecisionTree Technique

A Venkata Dineshreddy, Chandana Vamshi Krishna and
V Lavanya

November 3, 2022

Prediction of DDoS Attacks Using DecisionTree Technique

A.Venkata Dinesh Reddy
Department of Networking and Communication
SRM Institute Of science And Technology
Kattankulathur, Chennai, India
ad3659@srmist.edu.in

Ch.Vamshikrishna
Department of Networking and Communication
SRM Institute Of science And Technology
Kattankulathur, Chennai, India
cv6695@srmist.edu.in

V.Lavanya
Department of Networking and Communication
SRM Institute Of science And Technology
Kattankulathur, Chennai, India
lavanyav@srmist.edu.in

Abstract:-

As a result of cloud computing in previous years, the fields of computing and information technology have been subjected to a massive change in industry. Customers have been given the capacity, as a result of this capability, to rent virtual resources and utilize a wide variety of on-demand services at the most cost-effective prices. The use of cloud computing comes with a lot of advantages; yet, it is also predictable for a variety of threats. One of these threats is a distributed denial of service attack, also known as a DDoS attack, which is considered to be one of the most deadly varieties of cyberattack. In this essay, we will examine the prediction of distributed denial of service attacks using decision tree algorithm more accurately. A subset of data collected by the Canadian Institute for Cybersecurity is used. The dataset contains attacks that can be executed using TCP and UDP-based protocols has been taken, The project aims to extract knowledge from data for the classification of DDoS attacks and predict the DDoS attacks using decision tree technique.

Keywords: DDoS attacks, machine learning, decision tree, predictive analytics, threats, security

I. INTRODUCTION

For the purpose of a distributed denial of service attack, often known as a DDoS, a network of computers and other internet resources, sometimes referred to as a botnet, are utilized in order to bombard the website in question with activity that does not originate from real users [1]. An attempt to breach your security perimeter is not the goal of an attack that uses distributed denial of service [2]. Instead, they want to make your website and servers inaccessible to individuals who are authorized to use them in order to achieve their purpose [3].The DDoS attack can be used to disrupt security appliances and breach the target's security perimeter, as well as as a smokescreen for other malicious operations[4].It can also be used to bring down security appliances and bring down the target's security perimeter[5]. The availability of services in cloud has been one of the primary concern for cloud service providers when hosting a variety of cloud related information technology services and helps in managing a different internet resources[6]. The internet's vulnerability, the distributed nature of cloud computing, and the various security issues associated with cloud computing service models, and the primary characteristics of the cloud all contribute to the susceptibility of the cloud to security threats associated with the availability of cloud services[7].

Auto-scaling mechanism which is available in cloud platforms has been provided to us by making some defence for adding or deleting machines in response to varying load, as well as for making some factors like size and upper CPU utilisation criterion levels [8]. Using the other studies also plays a major role in detecting these DDoS attacks and there are many protocols have been implemented, which helps for improving the level of performance in cloud computing platforms and where the DDoS attacks happen through application layers or with the help of protocols hitting the application which helps in leading to DDoS attack techniques which have evolved rapidly [9]. These attacks are especially common against networks and websites of organisations, businesses that have been chosen for their importance of accessibility [10]. It is impossible to overestimate the importance of multiple layered barriers and collaboration [11].

A DDoS attack is carried out by breaking into computers and using them to attack an Internet-connected network [12]. Hundreds or thousands of computer frameworks can be turned into zombies and helps in attacking another framework or site through the Internet [13]. A successful DDoS attack usually conceals a communication via the handler/zombie with the attack (for example we can take the channel is encrypted) [14].Basically DDoS attacks will happen from multiple locations where an attacker makes the traffic busy ,so that the user cannot use the application[15].Even at sometimes there is a leakage of important information which may goes into the hands of an attacker[16].DDoS handlers are dispersed geographically across multiple networks[17]. As a result, reaching them may be time-consuming and difficult [18]. If the designer and implementer of the attack are not publicly disclosed, no other person or organisation is aware that the attack is extremely dangerous [19]. As a result, there is no awareness of an attack and no preparations to stop it [20]. Attacks are not detected unless the author publishes them or a third party accidentally identifies them. Here we have taken a dataset which is experimentally proved by some experienced people which they have provided us with all the types of DDoS related attacks and we have to classify and predict those attacks [21][22].

II. LITERATURE SURVEY

In [1],the primary focus of this research will be on the concerns, problems, and threats associated with cloud computing security, as well as potential solutions to these problems. Concerns regarding the safety of sensitive data have been brought to light as a direct result of the rapid growth of cloud computing. Because of these worries, the development of cloud computing is slowed down, and a solution is needed because security has become the most critical challenge connected with cloud computing.

In [2], the major goal, which is to prevent a distributed denial of service (DDoS) assault, is discussed. This goal is to filter out the data packets that have a significant data overhead. In addition, the average Mean Time to Security Failure (MTSF), has been calculated in order for an alternative dynamic plan.

In [3], the outcomes of this study indicate that certain varieties of DDoS attacks are not capable of being halted, detected, or mitigated in any way. These are the ones which were started from a original IP address and have a signature that is not currently stored in a database. Additionally, they are the ones that have a original IP address. In addition, the defence system might not be able to differentiate between the increased traffic that is originated by an attack and the vast increase in traffic which is caused by actual heavy traffic. This makes it difficult to know when extra resources should be provided. In addition, Others are given by unaffiliated third parties, while other defensive measures are implemented off-site, which could reduce the effectiveness of the entire system.

In [4], the authors present a collection of machine learning algorithms that can detect and identify MITM assaults that are carried out on a wireless communication network. These methods are described in this section. In addition, we evaluate and validate our methodology based on the performance indicators, and we compare the performance results with those achieved using various different machine learning approaches. This allows us to determine which method produces the best results.

In [5], This paper provides an examination of the adversarial techniques, tactics, and procedures, which are abbreviated as TTPs. The analysis is based on 549 honeypots that were distributed across 5 clouds and rallied to take part in 13,479 separate attacks. They discovered that adversaries actively test for plausibility, packet loss, and amplification benefits of these servers when they use a traffic shaping approach to avoid meaningful participation in DDoS activities while allowing short bursts of adversarial testing. They also show evidence of a "memory" of previously exploited servers among attackers. This is the case even when we allow short bursts of adversarial testing. When we implement a traffic shaping strategy to stop significant involvement in DDoS attacks, we uncovered this fact as a result of our efforts. In actual practice, we show that even for very ordinary amplification attacks, adversaries exhibit distinct variations in the manner that they carry out their operations, and this holds true even when they consider attacks that are pretty similar.

In [6] they present a comprehensive security analysis of CC-enabled Internet of Things (IoT) devices after analyzing the research field of the current state. Finally, prospective implementation and consideration topics, as well as future study effort and discussion topics, are provided in order to engage in discussion about open issues.

In [7], there is a presentation of a mathematical model for distributed denial-of-service attacks in the article that describes this study. This study makes a model suggestion. Machine learning algorithms such as Logistic Regression and Naive Bayes are frequently used in the process of identifying assaults and regular events. In this context, the CAIDA 2007 Dataset is used to conduct research. The dataset serves as a primary teaching resource as well as a critical resource for evaluating machine learning methods.

III.METHODOLOGY

A. System model

1) Data collection:

The process of gathering a certain amount of information in a systematic manner is known as data collection. This aids in the resolution of numerous questions, hypotheses, and outcomes. It is a process of gathering and estimating data on specific factors in a structured framework, which then allows you to ask pertinent questions and evaluate outcomes. Researchers in every field related study, including theoritically and sociology, humanities, and business, includes information gathering. While strategies may be different by discipline, the emphasis on ensuring precise and legal selection remains the same. It has been attempted on Kaggle for various datasets that would suit our project objective. After reviewing numerous datasets, the Canadian Institute for Cybersecurity selected a subset of data, This dataset contains the experimental analysis of tcp and udp floods which contains different types of attacks.

2)**Data Visualization:**

The pictorial or graphical representation of information is known as data visualisation. It allows you to understand difficult concepts and recognise new patterns. Many organisations regard data visualisation as a cutting-edge technology used. It entails the creation and investigates the visual representation of information. Information representation employs measurable illustrations, plots, data designs, and other apparatuses to convey data clearly and effectively. Customers benefit from effective visualisation when it comes to separating and reasoning about data and verification. It gradually makes complex data more understandable and provides us a choice to move further. Customers may have explicit logical endeavours, such as making assessments or obtaining causality; similarly, the reasonable structure standard (i.e., indicating examinations or demonstrating causality) follows the undertaking. Some regard it as a component of specific estimations, while others regard it as a tool for grounded theory development. Extensive amounts of data generated by certain amount of websites and an increasing number of sensors on the planet are referred to as "enormous data" or the "Web of things." Dealing with, analysing, and transmitting this data presents good and orderly challenges for data representation. This test is addressed by the field of data science and experts known as data scientists. we will analyse the data and choose what has to be done further.

3)**Data Pre-Processing:**

It is the transformation of data before it is fed into the algorithm. It converts unclean data into a clean dataset. It

is a method which analyses the content and splits the data according to our requirement as train and test. Data pre processing is an information mining procedure that is used to transform some certain amount of data into a useful and productive format, so that we can move further.

4)Data Cleaning:

It is a way of process which helps in identifying and removing errors from data in order to increase its value. Data cleaning is basically means removing of null and duplicate values. It is a method for identifying and removing all the unwanted information and helps in reducing the dataset size. It locates the missing information and helps in replacing the jumbled information. The data is altered to ensure that it is accurate and correct. Information cleaning is the process of identifying and revising incorrect records in a record set, table, or database. It is the process of recognising insufficient information and then replacing the jumbled information. The data is cleaned by removing all the unnecessary values. Its goal is to predict a dataset. The main aim of data cleaning is to detect and remove errors so that dynamic information estimation can be built. The primary emphasis should be on detecting the necessary characteristics and discovering interfaces between various information ancient rarities, for example, patterns and records.

5)Algorithms used:

Decision tree algorithm has been used by performing the information gain and evaluation gain on principal component analysis which is basically belongs to the feature selection part Then stratified k-fold technique has been done and the best entropy has been taken and shown the decision tree learner.

6)Explanation:

A subset of data collected by the Canadian Institute for Cybersecurity is used. The dataset contains attacks that can be executed using TCP and UDP-based protocols has been taken. Then we have used some libraries like Python, Scikit-learn, Pandas ,Matplotlib, Numpy etc to perform the experiment. Then the data has been visualized and analysed and removed all the duplicate and null columns to get better accuracy. Then we have trained the data and performed feature selection process.Then after mutual information gain and information gain has been performed to calculate the same quantity if applied to the same data or not, then with the stratified k-fold technique, we have tested the model's ability to predict new data that was not used in estimating and the principal component analysis is used to reduce the dimensions or size of the data to help out in figuring which has the most impact on the target variable. Then the decision tree algorithm has been used for determining the accuracy which is F-score by performing decision tree learner which helps in classifying and predicting the overall DDoS Attacks with good accuracy.

IV. FLOW OF ACTIVITY

A. IMPLEMENATATION:

This work will going to predict the DDoS attacks more accurately. At first, we have to import the dataset. This information includes important details such as the protocol, flow time, total forward packets, and total length of forward packets, and other things. The data must then be pre-elaborated, and atlast, we must eliminate columns that aren't being used. In addition, we will need to carry out a future selection process that involves mutual information and information gain. After that, we need to carry out a principle component analysis and a stratified K-fold CV. Then by using Decision tree learner we have to train the dataset and by using decision tree prediction, it gives us the output. Finally, we will determine the precision, recall, F1-score, and support by employing a decision tree K-fold configuration based on PCA. At last, we depict the confusion matrix into action.

1) Accuracy:

Good accuracy is actually how properly our version predicts the proper class or labels. If our dataset is reasonably balanced and all classes are similarly important, this ought to be our baseline metric to measure our version's performance.

$$\textbf{Accuracy} = \frac{\textbf{True Positives+True Negatives}}{\textbf{All Samples}}$$

2) Precision:

In easy terms, Precision is the ratio of what our version expected efficiently to what our version expected. For every category/class, there may be one precision value. We focus on accuracy when our predictions need to be correct, Idealistically, we want to make sure that our model is correct when it predicts a label.

$$\textbf{Precision} = \frac{\textbf{Total Positives}}{\textbf{Total Predicted Positives}}$$

3) Recall:

In easy terms, Recall is the ratio of what our version expected efficiently to what the real labels are. Similar to precision, for every category/class, there may be one keep in mind value.

$$\textbf{Precision} = \frac{\textbf{Total Positives}}{\textbf{Total Actual Positives}}$$

4)confusion matrix:

A Confusion matrix gives us the overall view of the analysis with the diagram depicting the true positive values ,true negative values which means contradicting true

values and false positive and false negative values belongs to the experiment will be shown.

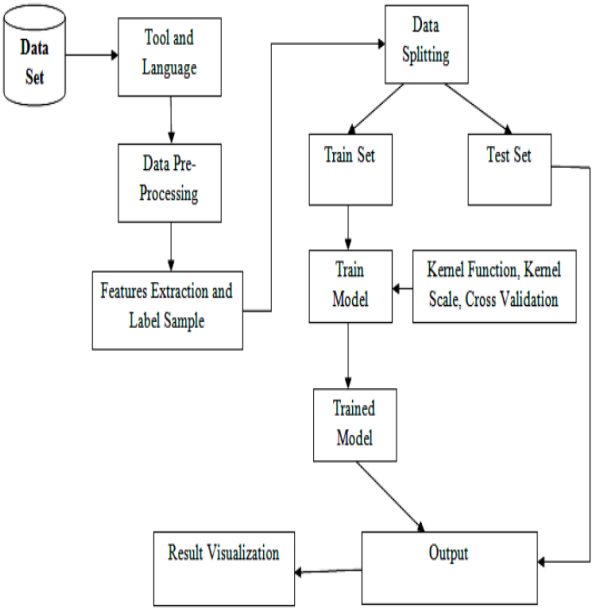

Fig 1: Data flow chart diagram

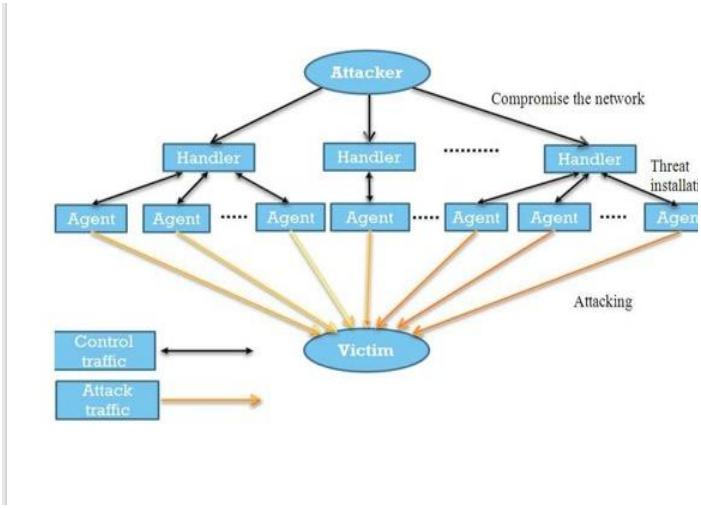

Fig 2: Architecture Diagram

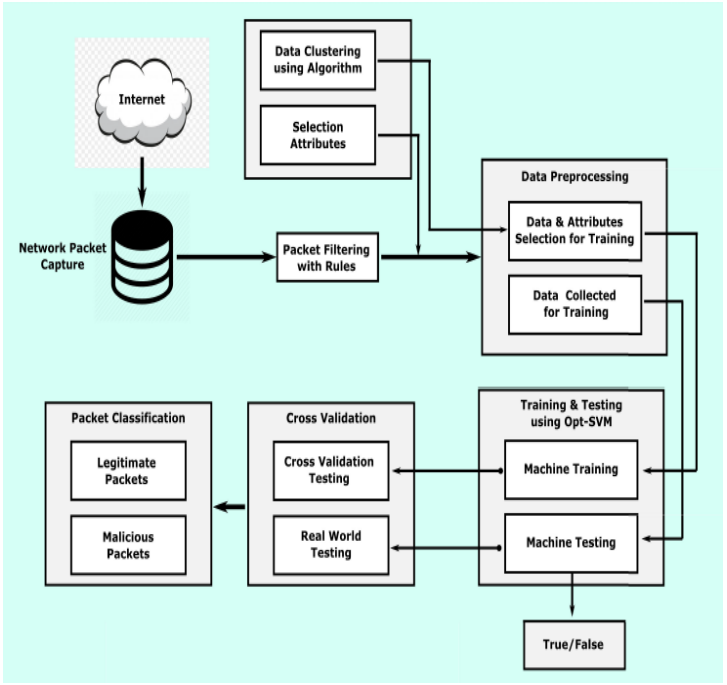

Fig 3: Overview of the implementation

V. RESULTS AND DISCUSSION

This model has successfully predicted and classified DDoS Attacks. Accuracy of our proposed approach is greater than 90% which is good. One should notice that the accuracy depicts the f1 score and the confusion matrix has been described between true label and predicted label. The precision, recall and f1 score values has been shown below.

	precision	recall	f1-score	support
0	0.99668	1.00000	0.99834	300
1	0.96602	0.99500	0.98030	200
2	1.00000	0.99500	0.99749	200
3	0.98974	0.96500	0.97722	200
4	1.00000	0.99000	0.99497	100
accuracy		0.99000	1.000	
macro avg	0.99049	0.98900	0.98966	1.000
weighted avg	0.99016	0.99000	0.99000	1.000

Fig 4: Classification report of decision tree

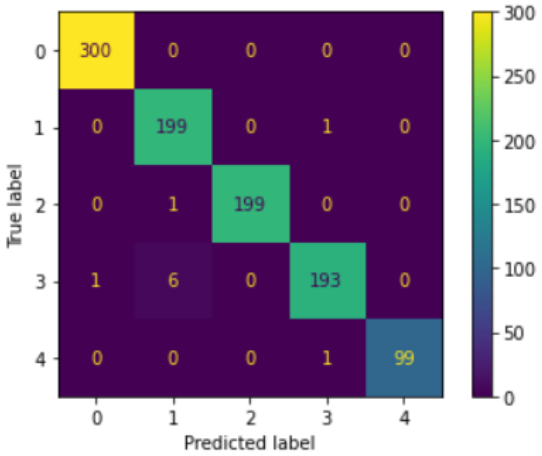

Fig 5: Confusion Matrix

VI. CONCLUSUION AND FUTURE SCOPE

A subset of data collected by the candian institute for cybersecurity in 2019.The dataset contains attacks which can be executed using TCP and UDP based protocols and we have successfully predicted the DDoS attack using machine learning algorithm and a depicted confusion matrix has been shown.
Looking forward in future, one can try to make some faster way of prediction of DDoS attacks as well as one has to work on to produce better results with in less time, so that it can produce better outcomes. It is critical to progress from unsupervised to supervised learning for unlabeled and labelled datasets. We will also have to look into how unlabelled datasets will affect DDoS attack detection.

VII. REFERENCES

[1] Alrasheed, S.H., Adubaykhi, S.A. and El Khediri, S., 2022, March. Cloud Computing Security and Challenges: Issues, Threats, and Solutions. In 2022 5th Conference on Cloud and Internet of Things (CIoT) (pp. 166-172). IEEE.

[2] Daffu, P. and Kaur, A., 2016, October. Mitigation of DDoS attacks in cloud computing. In 2016 5th International Conference on Wireless Networks and Embedded Systems (WECON) (pp. 1-5). IEEE.

[3] Radain, D., Almalki, S., Alsaadi, H. and Salama, S., 2021, March. A Review on Defense Mechanisms Against Distributed Denial of Service (DDoS) Attacks on Cloud Computing. In 2021 International Conference of Women in Data Science at Taif University (WiDSTaif) (pp. 1-6). IEEE.

[4] Saed, M. and Aljuhani, A., 2022, January. Detection of Man in The Middle Attack using Machine learning. In 2022 2nd International Conference on Computing and Information Technology (ICCIT) (pp. 388-393). IEEE.

[5] Griffioen, H., Oosthoek, K., van der Knaap, P. and Doerr, C., 2021, November. Scan, Test, Execute: Adversarial Tactics in Amplification DDoS Attacks. In Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security (pp. 940-954).

[6] Tahirkheli, A.I., Shiraz, M., Hayat, B., Idrees, M., Sajid, A., Ullah, R., Ayub, N. and Kim, K.I., 2021. A survey on modern cloud computing security over smart city networks: Threats, vulnerabilities, consequences, countermeasures, and challenges. Electronics, 10(15), p.1811.

[7] Kumari, K. and Mrunalini, M., 2022. Detecting Denial of Service attacks using machine learning algorithms. Journal of Big Data, 9(1), pp.1-17.

[8] Cirillo, M., Di Mauro, M., Matta, V. and Tambasco, M., 2021. Botnet identification in DDoS attacks with multiple emulation dictionaries. IEEE Transactions on Information Forensics and Security, 16, pp.3554-3569.

[9] Yang, Y., Wei, X., Xu, R., Peng, L., Zhang, L. and Ge, L., 2020. Man-in-the-middle attack detection and localization based on cross-layer location consistency. IEEE Access, 8, pp.103860-103874.

[10] Alsaeedi, A., Bamasag, O. and Munshi, A., 2020, November. Real-Time DDoS flood Attack Monitoring and Detection (RT-AMD) Model for Cloud Computing. In The 4th International Conference on Future Networks and Distributed Systems (ICFNDS) (pp. 1-5).

[11] Doshi, K., Yilmaz, Y. and Uludag, S., 2021. Timely detection and mitigation of stealthy DDoS attacks via IoT networks. IEEE Transactions on Dependable and Secure Computing, 18(5), pp.2164-2176.

[12] Prakash, A., Satish, M., Bhargav, T.S.S. and Bhalaji, N., 2016. Detection and mitigation of denial of service attacks using stratified architecture. Procedia Computer Science, 87, pp.275-280.

[13] Mittal, M., Kumar, K. and Behal, S., 2022. Deep learning approaches for detecting DDoS attacks: a systematic review. Soft Computing, pp.1-37.

[14] Agarwal, A., Khari, M. and Singh, R., 2021. Detection of DDOS attack using deep learning model in cloud storage application. Wireless Personal Communications, pp.1-21.

[15] Iman Sharafaldin, Arash Habibi Lashkari, Saqib Hakak, and Ali A. Ghorbani, "Developing Realistic Distributed Denial of Service (DDoS) Attack Dataset and Taxonomy", IEEE 53rd International Carnahan Conference on Security Technology, Chennai, India, 2019.