A DDOS PREVENTION SYSTEM DESIGNED USING MACHINE LEARNING FOR CLOUD COMPUTING ENVIRONMENT

P. G. Om. Prakash
K L University, Vijayawada, India

K. Sasirekha
Jain University, Kanakapura, India

Daniel Vistro
Asia Pacific University, Malaysia

ABSTRACT

The cloud computing is the new emerging solution for the IT industry to grow their business with the resources available on the pay as you go policy but what if the server goes down due to the cyber attacks it is a huge loss for the enterprise as each minute is valuable and also if the data is lost or misused then the consequences because of that is a major concern for the cloud service provider, Hence in this paper we proposed a design which when implemented on the real world will be able to solve the DDos attack in cloud environment as per the design we have trained the three machine learning classification models such as random forest, support vector machine and third one logistic regression which are trained on the real world dataset named CICDDoS2019 from the Canadian institute for cyber security so that based on the traffic rate the proposed algorithm will select the best classifier from the three to prevent the attack.

Keywords: Distributed Denial of Service, Logistic Regression, Confusion matrix, Domain Name System, Platform as a Service, Infrastructure as a service, software as a service, support vector machine, Random forest.


http://www.iaeme.com/IJM/index.asp
1. INTRODUCTION
In cloud environment virtualization is the process of running a virtual server rather than a physical server. The security continues to be the major concern for the virtualization in cloud. Many types of cyber attacks can affect the server but if one server or virtual machine goes down because of the attack the other server connected to it must also not get affected. The types of attacks that are possible on the server are man in the cloud, malware injection, side channel attacks, Denial of service attacks, insider attacks. Among the different characteristics of cloud computing the availability of resources is crucial for the cloud providers to perform well in the market. The denial of service attack is one where an intruder tends to send the server with huge amount of request that have an invalid return addresses. Distributed Denial of service technique is a major threat to most of the cloud computing resources since it results huge of amount of revenue loss for the company. Table-1 shows some of the DDos attacks that affected the victim in the recent years. Nowadays DDos attack is so easy to be implemented in the servers so it has increased the interest to provide solution.

DDos attacks can be performed in many types so it is not possible for the common algorithms to detect them and prevent them in advance. That is why many classification algorithms have been proposed to detect the attacks. But individually when the algorithms are performed they cannot provide results with high accuracy that is why in this paper we have designed a method that uses classification algorithms based on the amount of traffic and the accuracy accordingly the algorithm is chosen.

Here we take logistic regression, support vector machines and random forest as the classification algorithms and we train these algorithms on the CICDDoS2019 dataset and the accuracy of the algorithms is compared.

2. BRIEF INTRODUCTION OF DDOS
The denial of service attack is an attack to prevent the user from accessing the server. It is a large scale attack in a coordinated fashion. The DDos attack is extremely simple to initiate an attack but very difficult to protect our resources from it. For an effective attack the intruder often establishes a network of computers connected together or separately to send traffic which is called as botnet. The attacks can be divided into two categories based on bandwidth and resources.

The attacks which happen in the network level such as on TCP, UDP, ICMP, and DNS are the attacker tries to get access by exploiting the user’s bandwidth. The Application level attacks tries to exploit the server resources.

2.1. Various Techniques in Ddos
**UDP FLOOD ATTACK:** In this attack the attacker overwhelms random ports on the targeted hosts with the IP containing UDP datagram. The receiving server checks the packet if the application is not found then it sends “DESTINATION UNREACHABLE”. If more such packets are sent continuously then the server becomes unavailable for other users. In UDP the handshaking property is not there like TCP which makes the attack easy. Some UDP attacks can take in the form of DNS amplification attack or alphabet soup attacks also. Domain Name System (DNS) amplification is the process of making use of the vulnerabilities present in the DNS to increase the small queries which are received to larger payloads.

**PING FLOOD ATTACK:** In this attack the attacker takes down the server by overwhelming the server with ICMP echo requests, also known as pings. The ping request are used to check the connectivity status between the two systems measuring the round trip time that is the time in which the packet is sent and the packet is received. But in the attack they are sent in large
numbers to the targeted server to overload the server so that they won’t be able to handle the request from the legitimate users.

**SLOWLoris:** It is the highly used attack making the targeted server to hold for the connection as long as possible. In this the attacker sends more HTTP requests to the targeted server but will never complete the request. When such type of request are sent in large number the server will wait for the HTTP request to complete not able to give access to other users.

There are many other techniques also other than this but we have listed some of the most commonly used techniques.

### 3. LITERATURE SURVEY

Amjad Alsirhani, Srinivas Sampalli, Peter Bodorik [1] proposed a DDoS Detection System: They used three Classification Algorithms that are controlled by the Fuzzy Logic System in Apache Spark and compared the results.

Shi Dong, Khushnood Abbas, and Raj jai [2] did a survey on the various attacks which can be deployed in the SDN and cloud environment and provided some of the solutions for that attack.

Qiao Yan, F. Richard Yu [3] did a survey on the denial of service attacks on the SDN and cloud environment and provided solutions using machine learning and they claim that machine learning algorithms provide better solutions than the general algorithm.

#### Table 1 The various attacks which took place in the past decade

<table>
<thead>
<tr>
<th>Vendor</th>
<th>Year</th>
<th>Traffic Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Github</td>
<td>2018</td>
<td>The traffic was traced around thousands of different autonomous systems. The traffic hit was around 1.35 terabits which resulted in huge amount of loss.</td>
</tr>
<tr>
<td>Occupy Central, Hong Kong</td>
<td>2014</td>
<td>It was carried out by five different botnets. This resulted in the traffic levels of up to 500 gigabits.</td>
</tr>
<tr>
<td>CloudFare</td>
<td>2014</td>
<td>This attacker used a technique called reflection to send traffic levels up to 400 gbps. This is launched because of the vulnerability in the network time protocol (NTP).</td>
</tr>
<tr>
<td>Spamhaus</td>
<td>2013</td>
<td>It made the email services and the website to go offline for hours because of the traffic rate scaling to 300 gigabits.</td>
</tr>
<tr>
<td>Netscout</td>
<td>2018</td>
<td>It got affected with a rate of 1.7Tbps.</td>
</tr>
<tr>
<td>Imperva</td>
<td>2019</td>
<td>One of the clients of Imperva was able to get affected with a peak scale of traffic up to 580 million.</td>
</tr>
</tbody>
</table>

Da yin, Lianming zhang [4] proposed an idea to handle the ddos attack on the SDN and the internet of things where by using cosine similarity of vectors in the message rate at boundary iot devices they claim to avoid the ddos attack.

Preeti Daffu, Amanpreet Kaur [5] they avoid the DDoS attack by filtering out the packet which has large data overhead thus preventing the DDOS attack.

Karthik Srinivasan [6] performed a survey on the various techniques that are used to perform the ddos attacks in the cloud. Even though many methods have been provided using the existing algorithms to prevent DDoS attack in the cloud but still since the DDoS attacks are becoming sophisticated it is not possible for the general algorithms to provide solution that is
why in this paper we have designed a framework to overcome this issue and tried to implement the dataset on the three machine learning algorithms to avoid the DDoS attack. This is a supervised learning model.

3.1. Why DDoS attack in cloud environment is increasing

![Figure 1 Increase in the DDoS attack scale over the years](image)

We can see from the above figure that the magnitude of DDoS attacks in the past years from 2003 to 2016 it is increasing in large scale and in 2020 with the increase in use of cloud computing it is expected to cross 1300 GB/sec. Since the cloud infrastructure is actually a multi-tenant environment an attack attempted against a single Virtual machine is actually a major attack against all the users connected in the same environment. DDoS attack using virtualization is when one virtual machine takes all the available resources such that the hypervisor won’t be able to allocate more resources to other users.

Everyone is using mobile phones today with increased amount of memory capacity and everyone trying to access the cloud attackers can easily exploit using the vulnerabilities present in the mobile phone since the security in most of the mobile phone users is less.

Nowadays the DDoS attacks are becoming more sophisticated that the regular paradigms which are used to protect the servers won’t work in the cloud computing environment that the cloud providers have to have high level of security to make sure that the cloud servers are able to provide services to the users.

4. PROPOSED DESIGN

The proposed DDoS system is shown in figure-3. That is composed of the following steps the N-classification algorithms in our paper we have chosen three classification algorithms. We use the CICDDoS2019 dataset. Before training we analyze the dataset we perform preprocessing on the dataset so that we get better results and then we perform classification. After performing classification on all the algorithms the best accuracy is choose based on the traffic and the accuracy of the classifier using genetic algorithm. The classification is performed to identify the normal packets and the abnormal packets. The three classification algorithms taken are the logistic regression support vector machine and random forest. The results are evaluated based on the confusion matrix. Only one of the algorithms will be selected based on the traffic at the time using algorithm.

5. WORK FLOW

The system workflow is given in the diagram above. The steps to train individual classifier is given in the figure-2.

1. The dataset from Canadian institute for cyber security is taken for the training
2. The preprocessing is done on the dataset.
   - Normalization is performed
   - Various attributes are analyzed
   - Outliers are checked
3. Feature Extraction is done.
4. Using this dataset the algorithms are trained one by one to perform classification on the real data.
5. To train the algorithms we use 80 percent of the data and for the testing we use 20 percent of the data.
6. The accuracy of the algorithms is checked by using confusion matrix.
7. When the real data is given using algorithm the best classifier is selected based on the traffic rate.

Now the model is ready to be deployed for future data. In this paper the results up to the classification and training are given. The deployment and the results on the cloud based on the traffic are discussed on the future work.

![Figure 2 Process of Training the Classifiers](image)

**Figure 2** Process of Training the Classifiers

![Figure 3 Proposed work design](image)

**Figure 3** Proposed work design

*Classification Algorithms:* There are many classification algorithms in machine learning. But we choose logistic, support vector and random forest because of the accuracy and if the arrival of the traffic is linear then logistic regression can provide better accuracy if the arrival of the traffic is non-linear then support vector machine provides better accuracy. Random forest trees can be used for both and also some time when there is many user requests then random forest provides quicker results. In the next section we briefly describe about the different types of algorithm.

**5.1. Logistic Regression:**

The logistic function is otherwise called as the sigmoid function. This function is used to provide any values between 0 and 1. The function is given by
The ‘e’ here is the exponential function which makes the value of the function to be in the range. The cost function for the logistic regression can be obtained from the linear regression cost function as

\[ J(\theta^*) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x(i)) - y(i))^2 \]

The above function is the cost function for the linear equation which can be rewritten as

\[ J(\theta^*) = \frac{1}{m} \sum_{i=1}^{m} 1/2 (h_\theta(x(i)) - y(i))^2 \]

By writing this function in general way

\[ \text{Cost}(h_\theta(x(i)), y(i)) = \frac{1}{2} (h_\theta(x(i)) - y(i))^2 \]

From this the cost function can be defined as:

\[ \text{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y=1 \\ -\log(1-h_\theta(x)) & \text{if } y=0 \end{cases} \]

There are various methods to perform the logistic regression

**Newton method**: It is similar to the gradient descent method but it uses a better quadratic minimization. It is based on the iterative equation solver. For each iteration one approximates \( f(x) \) by a quadratic function around where \( f(x) \) is the function or the output and then takes steps towards the maximum or minimum.

**Limited memory Broyden Fletcher algorithm**: It is an algorithm which inverses the hessian matrix. Here limited memory means it stores only a few vectors representing the approximation.

**Stochastic Average Gradient**: This is good when the dataset is very large as it converges very fast when compared to the stochastic method by making use of the previous memory values.

**Random Forest Algorithm**: Random forest is the combination of different trees such that each tree is based on some random value. The same distribution is applied to all the trees in the forest [18]. The algorithm follows the following steps 1) it starts from selecting random samples from the dataset. 2) Then it will generate decision tress for every sample and get the results from each of the tree. 3) The weight will be assigned for each tree. 4) The result or the tree which has highest weight will be the final result. In trees there are two methods for performing classification.

**Entropy**: It is based on the uncertainty in the data. The equation can be given by

\[ E = \sum_{x=1}^{c} -fx\log(fx) \]

**Gini impurity**: It is calculated by the entropy obtained before splitting of the set and after splitting the set.

\[ G = \sum_{x=1}^{c} fx(1 - fx) \]

**Support vector machine**: It works similar to the linear regression but in this case it classifies using a hyperplane. The distance between its extreme point and the hyperplane should be maximum. That is the hyper plane for which the margin is maximum it is selected. It also has
the kernel function when the dimensions are greater using the kernel function it transforms it into smaller number of dimensions to solve the problem.

5.2. Dataset
The dataset used is CICDDoS2019 which contains infected systems information and till date common DDos and the most recent attacks such as portMap, LDAP, NetBIOS, MSSQL, UDP-Lag, UDP, NTP, DNS, SYN, and SNMP. We are taking 22547 samples recorded per day in various servers such as windows 7, windows Vista windows 8.1 and windows 10. There are 72 columns in the dataset. For this work we consider only 18 columns. The data is obtained from Canadian institute for cyber security which is a real world dataset. The sample of the dataset is shown in the figure-4 below. The sample of 225745 is taken for the process.

![Figure 4 dataset sample.](image)

6. DISCUSSION
We can see from the sample dataset figure that values in the dataset are not in common scale so it will result in lesser efficiency as the model will think the higher values are having more importance so we do preprocessing on the dataset.

Normalization: For preprocessing we use the technique called normalization. So the numeric values in the columns are changed to a common scale that is they are distributed normally.

The formula is

\[ x' = \frac{x - \text{mean}}{x_{\text{max}} - x_{\text{min}}} \]

![Figure 5 Heat map of Correlation](image)

Null values: The null values in different columns are checked. As it is a large dataset if more null values are present it results in the decrease in the efficiency of the algorithm.
Outliers: The outliers are not removed because in our problem outliers are important in analyzing the task. For input to the classifier we have given flow duration, forward packets, backward packets, total length of the packet, forward IAT total, Backward IAT. Training data and testing data are split in the ratio of 80 percent for training and 20 percent for testing.

The relation between the forwarded packets and the received packets is shown in the Figure-6 below.

![Figure 6 Scatter plot between the packets received and forwarded](image)

7. EVALUATION
The three algorithms results are given in the table-2 below.

<table>
<thead>
<tr>
<th>MODEL</th>
<th>SOLVING METHOD</th>
<th>ACCURACY SCORE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic regression</td>
<td>lbfgs</td>
<td>0.9057343462756651</td>
</tr>
<tr>
<td>Logistic regression</td>
<td>newton</td>
<td>0.9057343462756651</td>
</tr>
<tr>
<td>Random forest</td>
<td>gini</td>
<td>0.9771202019978295</td>
</tr>
<tr>
<td>Random Forest</td>
<td>entropy</td>
<td>0.977275244191455</td>
</tr>
<tr>
<td>Svm</td>
<td>linear</td>
<td>0.877907385038038</td>
</tr>
<tr>
<td>Svm</td>
<td>Poly(degree=3)</td>
<td>0.9432683628974</td>
</tr>
</tbody>
</table>

CONFUSION MATRIX: Confusion matrix is used for the visualization of the performance analysis of the algorithm. It is used to provide the results on the test data when we have the true value after training the model.

![Confusion Matrix](image)

Figure 7 The confusion matrix

Here the above figure shows the confusion matrix which is used to analyze the performance of the algorithms. The confusion matrix for the three algorithms is given in the table-3.

Accuracy: It is given by the formula given below. It is calculated by the total number of correct predictions to the overall data present in the dataset.
Accuracy = \( \frac{(TN+TN)}{(TP+FP+FN)} \)

**Precision:** It is calculated by taking the value of true positive dividing it by the sum of true positive values and false positive values.

\[ \text{Precision} = \frac{TP}{TP + FP} \]

**Recall:** Recall value is obtained by dividing the values of true positive along with the sum of true positive and false negative values.

\[ \text{Recall} = \frac{TP}{TP + FN} \]

Table 3 Confusion matrix for the algorithms

<table>
<thead>
<tr>
<th>MODEL</th>
<th>SOLVING METHOD</th>
<th>CONFUSION MATRIX</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>lbfgs</td>
<td>[[15231, 27], [4229, 25662]]</td>
</tr>
<tr>
<td>Logistic regression</td>
<td>newton</td>
<td>[15231, 27], [4229, 25662]</td>
</tr>
<tr>
<td>Random forest</td>
<td>gini</td>
<td>[[18764, 330] [696, 25359]]</td>
</tr>
<tr>
<td>Random forest</td>
<td>entropy</td>
<td>[[18764, 330] [696, 25359]]</td>
</tr>
</tbody>
</table>

8. CONCLUSION

As we can observe from the above table-2 and table -3 .Random forests performs better with this dataset with an accuracy of 97 percent. Even though svm performs better the time factor is important for large dataset it takes lot of time to get trained and in cloud computing as there are lot of servers involved and the request are more in number svm cannot perform well when the traffic is high.

9. FUTURE WORK

In this paper we designed, analyzed our algorithms and compared the results in further work we will deploy our design and also design a logic to select the classifier in the cloud computing environment to know how efficiently our design works for the real world data.

As a an author of this paper I don’t have any competing interest.

REFERENCE

[1] Amjad Alsirhani, Member, IEEE, Srinivas Sampalli, Member, IEEE and Peter Bodorik, Member, IEEE, “DDoS Detection System: Using a Set of Classification Algorithms Controlled by Fuzzy Logic System in Apache Spark” IEEE-2019


A DDOS Prevention System Designed Using Machine Learning For Cloud Computing Environment


