Regression algorithms for efficient detection and prediction of DDoS attacks

Gudipudi.Dayanandam
Research Scholar, Dept of CSE,
ANUCET, Guntur, INDIA,
gdayanandam@gmail.com

Dr.E.Srinivasa Reddy
Principal, ANUCET,
Guntur, INDIA,
esreddy67@gmail.com

Dr.Dasari.Bujji Babu
Professor & HOD,
Dept of MCA,QISCET,
Ongole,INDIA,
bujjibict@gmail.com

Abstract— **In the ICT era the need of depth investigation and analysis is required on network traffic. The analysis should focus on detecting DDoS attacks. In the 21st century the use of communication or transactions are completely doing through online, the political activists, and international cyber terrorists are choosing the DDoS attacks as a powerful weapon for their illegal an un ethical activities. It is impossible to the human being to identify all these unethical activities, hence the need of machine based algorithms are required. In this paper we used GLM, GBM, NN, RF regression algorithms for detection and prediction of DDoS attacks, and also proved that by using regression algorithms we observed more accurate result than using KNN SVM algorithm.**

*Keywords— **Distributed Denial of Service (DDoS), Machine learning, smurf attacks, Random Forest (RF), Neural Networks (NN), Generalized Linear Models (GLM), and Stochastic Gradient Boosting (GBM)***

I. Introduction

An attack is nothing but a violation of security policy of systems. There can be passive attack which aims to learn or make use of information from the system but does not affect the system resources. Active attack which contains of original message are modified by attacker.

DoS attacks:

DoS is nothing but denial of service. In DoS attack, the attacker makes the resource unavailable to the legitimate users in the computer network by making the resource busy. DoS attack when deployed form various systems globally distributed then it is known as Distributed Denial of Service (DDoS).

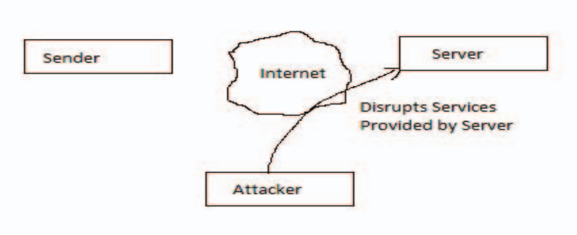

Fig.1. Denial of service attack

DoS attacks are of threat to internet they decrease the quality of internet service. The main objective of DoS attack is to disrupt the service to authorized persons, to damage the reputation, to get the financial gain etc. DoS attack can be performing in two ways. Such one way is attacker smash the services and in another way attacker flood the traffic to consume the resources, in both ways servers are not available to authorized clients because all of the target critical resources are busy to handle the attack traffic.

Aims of DoS attacks are:

- To consume its bandwidth by sending large volume traffic.
- To consume its resources by sending huge number of requests.
- To flood packets to overload the networks.

Distributed Denial of Service (DDoS) attacks:

DDoS attack is one in which multiple number of computers attack a single target system. The flood of incoming messages to the target system essentially forces it to shutdown, there by denying service to the system to legitimate users.

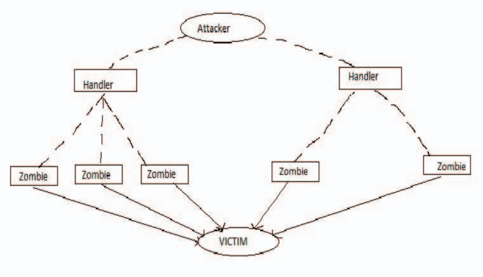

Fig.2. Distributed Denial of Service attacks [3]

A Computer under the control of an intruder is known as "Zombie" or "Bot". a group of zombies are called zombie army. A group of bots is called "Botnet"

IP Spoofing:

IP spoofing is a technique used to gain unauthorized access to computers where by the intruders sends messages to a computer with an IP address indicating that the message is coming from a trusted host.

II. Types of DDoS Attacks

TCP SYN flood attacks:

TCP SYN flood attack is one type of DDoS attack that spoofs the IP address. It exploits the part of the normal TCP 3-way

978-1-5386-1144-9$31.00 ©2017 IEEE

handshake to consume resources on the targeted server and render it unresponsive. Server responds to spoofed IP address client, but it never receives response. Final step in 3-way handshake should not complete, so it goes to blocked state. This process continues and overflow the system buffer so that the server could not respond to authorized users.

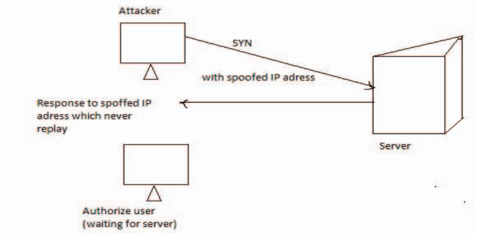
Attacker
SYN
with spoofed IP adress
Response to spoffed IP adress which never replay
Server
Authorize user (waiting for server)

Fig.3. TCP SYN flood Attack [4]

PING flood attacks:

Ping flood attack is one in which attacker sends a continuous series of ICMP echo request ping packets to a target host on network. The target host replies with ICMP replay packets that is ICMP echo replay. The network becomes slow and authorized user can not connect to server due to continuous request and reply packets.

UDP flood attacks:

UDP flood attacks occur when attacker sends huge number of UDP packets to target system to slowdown the target system. Target system cannot handle authorize connections after UDP threshold reaches, the server then rejects other request of UDP packets.

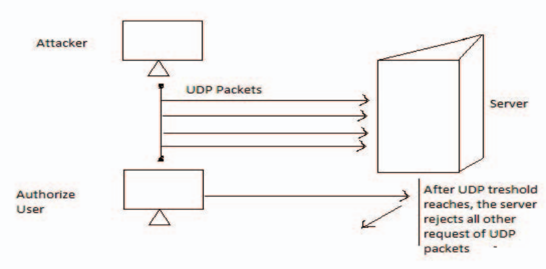
Attacker
UDP Packets
Server
Authorize User
After UDP treshold reaches, the server rejects all other request of UDP packets

Fig.4. UDP Flood Attack [4]

The Smurf Attacks:

Smurf attack is a type of DDOS attack [2] where a system is flooded with spoofed ping messages. Victim's IP address received huge number of ICMP requests, which leads to significant amount of traffic on the victim's network, resulting in consumption of bandwidth. This can slow down the victim's computer to the point where it becomes impossible to work on.

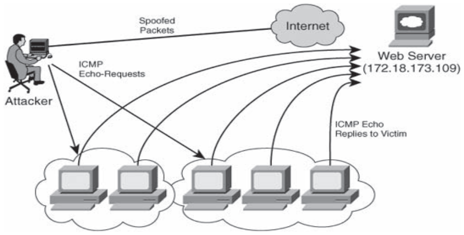

Fig.5. Smurf Attack

III. DDOS COUNTERMEASURES

Attack prevention system:

DoS attack prevention system is used before the actual attack done by doing penetration testing. Penetration testing is used to test the internal vulnerabilities of the server and apply patches to any identified vulnerabilities.

IP Trace back: IP trace back methods are used to stop the attacks once we detect the attacks. Once we trace back the correct path of the attack, we isolated the systems that are in attack path from the attacker and server system.

Detection System:

To minimize the impact of DDOS attack, we use detection mechanism during the attack. It detects unknown behavior of that packet such that we can take necessary steps to stop the attack. Detection methods can be classified into four categories [5]. They are

1. Statistical based Methods
2. Knowledge based Methods
3. Soft computing Methods
4. Machine Learning Methods

Statistical based methods involve the collection of unauthorized user's data over a period of time. Then statistical tests are applied to observed behavior to determine with a high level of confidence whether that behavior is not unauthorized behavior.

Knowledge based methods are also called rule based methods. These methods are involving to observing events and applying a set of rules to take decision regarding whether a given pattern of activity is suspicious.

Soft computing methods involve classification and IP trace back schemes. Classification lead to an attack mitigation scheme that rate limits or filters the malicious packets. IP trace back scheme is concurrently performed while attack mitigation takes place. Using this, we will find true source of the packets and avoid IP spoofing attacks.

Machine Learning Techniques:

To protect the network efficiently, we use Intrusion detection system(IDS) along with other types of security mechanisms. There are two types of IDS: signature based and anomaly – based detection.

A signature-based IDS is used to detect known attacks.

The anomaly – based detection IDS is used to detect unknown attacks [11].

Signatures based IDS rely on humans to create, test and deploy the signatures. Thus, it may take hours or days to generate new signature for an attack, which can be too long. When dealing with rapid attacks. Never the less, in order to offer a human-independent solution to the above-mentioned problem, anomaly based IDS based on machine learning techniques have the ability to implement a system that can learn from previous data and give the decision for unseen data.

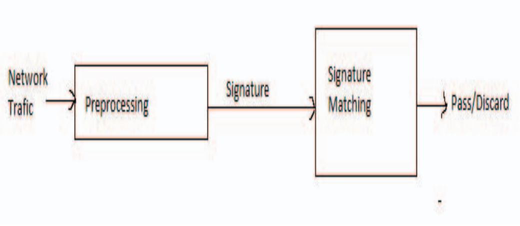

Fig.6. Signature based detection[6]

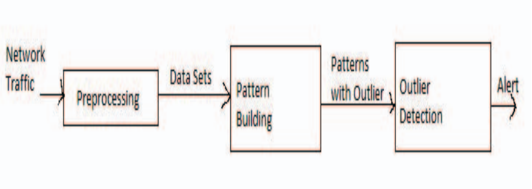

Fig.7. Anomaly Based Detection [6]

Machine learning methods use advanced statistical techniques for handling classification and regression techniques with multiple target and independent variables.

Classification:

It is technique used to arrive at schematic that shows the organization data starting with a precursor variables. The dependent variables are what classify the data in to group. The classification tree starts with independent variable, which branches out in to two groups as determined by the existing dependent variables. It is meant to elucidate the responses in the form of categorization brought about by the dependent variables.

Regression:

Regression is a prediction method that is based on an assumed or known numerical output values. These output value is the result of series of recursive partitioning, with every step having one of numerical value and another group of dependent variables which branch out to another pair such as this. The regression tree starts one or more precursor variables and terminates with one final output variables, and terminates with one final output variables. The dependent variables are either continuous or discrete numerical variables.

Decision Tree:

Decision tree algorithm is normally used to solve the classification problem. Here, the dataset is learnt and modelled. If we want to verify new data, we can give new data to decision tree as input, and rules in decision tree will decide whether it is attack data or not. Decision tree algorithms are useful in detecting DDoS attacks.

Artificial Neural Networks:

Artificial neural networks are transform a set of inputs to a set of target outputs using processing elements which functionality is similar to biological nervous systems, such as brain, process information.

K-Means Clustering:

K-means clustering (Mac Queen, 1967) is used to partition the dataset into k-groups. Initially It creates k centers and then interactively refining them as follows.

1. Each instance d_i is assigned to its closest cluster center.
2. Each cluster center c_j is updated to be the mean of its constituent instance. The algorithm converges when there is no further charge in assignment of instances to cluster.

IV. RELATED WORK

Ahmad Rizaain Yusof et al[1] proposed that "Fuzzy c-means clustering gives better classification and it is fast compared to the other algorithms."

Gavrilis et al [7] used "RBF-NN detector which uses nine packet parameters and the frequencies of these parameters are estimated. Based on the estimated frequencies, RBF-NN classifies traffic into attack or normal class."

Ming-Yang Su et.al [8] proposed "a hybrid technique for feature selection and weighting with genetic algorithm and KNN (k-nearest-neighbor)."

Suresh [9] evaluated "the Fuzzy c-means clustering on DDoS attacks with machine learning technique and obtained better classification."

V. PROPOSED WORK

In this section, we look step by step procedure of our proposed work.

Read Dataset:

The first step in machine learning algorithm is read dataset. The standard format for machine learning is csv file.

Data preparation process:

Data preparation algorithms learn from data. We need to give right data for the problem we want to solve. Data selection, Data Preprocessing and Data transformation are three steps used in data preparation process.

Data selection is used to identify missing values and what data to be removed. Data Preprocessing organizes our selected data by cleaning, formatting and sampling from it. Data Transformation transforms preprocessed data ready for machine learning by engineering features using scaling, attribute decomposition and attribute aggregation.

Data splitting:

We can divide the data into two types trained data and test data. Trained data is used to build the model and test data is used to validate the model

Feature selection and Extraction:

Feature selection is the process of selecting a subset of relevant features from all features for constructing model. Wrapper methods and filter methods are used in feature selection. Recursive Feature elimination, which is a wrapper method to find the best subset of features to use for modeling.

Recursive Feature Elimination

The Recursive Feature Elimination (or RFE) works by recursively removing attributes and building a model on those attributes that remain. To predicting the target attribute, RFE identifies correlated attributes.

Apply model:

We apply different models for our data to analyze which model get good results.

Neural networks:
Neural Network enables Neural Network to learn non-linear relationship between input x and output z.
The error is propagating from the output layer to input layer to adjust the weight in backward error propagation

Random Forest:
Random Forest is used to create large number of decision trees and then each observation is given input to the decision tree. Maximum of the observations is considered for final output.

Generalized linear models:
In Generalized linear models, we use extensions of traditional regression models that allow the mean to depend on the explanatory variables through a link function, and the response variable belong to any the exponential family.

Stochastic Gradient Boosting:
Gradient boosting algorithms are performing better than random forest algorithms. We need to tune gradient boosting algorithms where random forest is free from tuning.

Evaluation: We can evaluate the algorithms by giving test dataset as input to classifier and finally find the accuracy of the model.
Performance of a model can be done by confusion matrix [7][10], ROC Curves and F-measures. Here we predict our model using both confusion matrix and F-measures.
Confusion matrix for our test data:

TABLE I. Confusion matrix [10]

n=999	Predicted:0	Predicted:1
Actual:0	428	0
Actual:1	0	571

From the above table we observe that there are two possible predicted classes:”1” for Yes and “0” for No. If we were predicting the presence of an attack we represent with 1 otherwise we represent with 0. For example,”1” would mean they have the smurf attack, and “0” would mean they don’t have the smurf attack.
Out of 999 test samples, the classifier predicted “1” 571 times and “0” 428 times. In reality, 571 attacks are correctly identified as smurf, and 428 attacks are not smurf.

Table1 can be presented with respect to True Positive(TP) and True Negative(TN) as shown in the following table.

TABLE II. Confusion matrix with TP and TN, Where N is Total packets.

N=999	Predicted:0	Predicted:1
Actual:0	TN=428	FP=0
Actual:1	FN=0	TP=571

Classification Accuracy of the model can be computed with the following formula
Accuracy=(TP+TN)/N.
Misclassification rate=(FP+FN)/N, which is also called Error rate. Error rate=1-Accuracy
Sensitivity or Recall: When its actually 1, how often does it predict 1, so Sensitivity=TP/(Actual 1)
False positive rate: When its actually 0, how often does it predict 1. So False Positive Rate=FP/(Actual 0)
Specificity: when its actually 0, how often does it predict 0?
Specificity=TN/(Actual 0) or 1-False Positive Rate
Precision: When it predicts 1, how often is it correct?
Precision=TP/predicted 1
Prevalance: How often does the 1 condition actually occur in our sample?
Prevalance=actual 1/N;
F-Score is a weighted average of the true positive rate(Recall) and Precision.
So F-Score=2*(precision*recall)/(precision+recall)

VI. COMPUTATIONAL RESULTS

Data set:
The KDD99 dataset is used in the experiments as the attack component. This dataset is used for predicting model capable of distinguishing between smurf attacks, other types of attacks and normal attacks. This dataset consists of 42 features.

Feature Extraction:
By using recursive feature elimination, we can extract 4 important features out of 42 features which are shown in the following table.

TABLE III. Feature extraction

Feature Index	Feature name	Description	type
1	service.ecr_i	Network service on the destination	factor
2	srv_count	Number of connections to the same service as the current connection in the past two seconds	int
3	Src_byte	Number of data bytes from source to destination	int
4	Count	Number of connection to the same host as the current connection in the past two seconds	int

So, by using feature extraction we can get the features which are more correlated to our target feature which will give more accuracy.

Classification results:

TABLE IV. Classification Results

Method Used	Correct Classification	Detection Time(In Sec)
GBM	1.00000	0.17
GLM	1.00000	0.18
Random Forest	1.00000	0.16
Neural Networks	1.00000	0.16

TABLE V. F-measure details of classifiers

Methods	True Positive	False Positive	True Negative	False Negative	F-Measure
GBM	571	0	428	0	1.000
GLM	571	0	428	0	1.000
Random Forest	571	0	428	0	1.000
Neural Networks	571	0	428	0	1.000

VII. CONCLUSION

With the help of machine learning algorithms the researchers evaluated KDD99 dataset to predict DDoS attacks. Four important features are extracted out of 42 features using recursive feature elimination to improve the accuracy of the model. Experimental results show that GBM, GLM, Random Forest & Neural Networks are well performing algorithms for classification.

VIII. ACKNOWLEDGEMENT

The authors would like to thank anonymous reviewers for their constructive comments and valuable suggestions. The authors wish to thank TunedIT for providing KDD dataset for completion of this research. Last but not least, the authors would like to thank to all the authors of the reference papers.

IX. REFERENCES

[1] Ahmad Riza'ain Yusof,Nur Izura Udzir and AliSelmat,"An evaluation on KNN-SVM Algorithm for Detection and prediction of DDOS Attack"

[2] NiharikaSharma, Amit Mahajan, Vibhakar Mansotra"Machine Learning Techniques Used in Detection of DOS Attacks: A Literature Review"International Journal of Advanced Research in Computer Science and Software Engineering,March - 2016

[3] C Balsrengadurali and Dr.S Saraswathi,‟ Fuzzy Based Detection and Prediction of DDoS Attacks in IEEE 802.15.4 Low Rate Wireless Personal Area Network,‟ IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 6, No 1, November 2013

[4] Mangesh D. Salunke ,Prof. Ruhi Kabra,‟ Denial-of-Service Attack Detection „International Journal of Innovative Research in Advanced Engineering (IJIRAE) „„Volume 1 Issue 11 (November 2014

[5] Monowar H. Bhuyan , H. J. Kashyap , D. K. Bhattacharyya and J. K. Kalita,"Detecting Distributed Denial of Service Attacks: Methods, Tools and Future Directions", The computer Journal, 57 (4), 537-556

[6] Jayveer Singh, Manisha J. Nene,‟ A Survey on Machine Learning Techniques forIntrusion Detection Systems,‟ International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 11, November 2013

[7] Gavrilis, D., & Dermatas, E. (2005). Real-time detection of distributed denial-of-service attacks using RBF networks and statistical features. Computer Networks, 48(2),235–245. doi:10.1016/j.comnet.2004.08.014

[8] Lee, K., Kim, J., Kwon, K.H., Han, Y., Kim, S.: DDoS Attack Detection Method usingCluster Analysis. Expert Systems with Applications 34, 1659–1665 (2008)

[9] Livadas, C., Walsh, R., Lapsley, D., & Strayer, W. T. (2006). Usilng Machine LearningTechnliques to Identify Botnet Traffic. Local Computer Networks, Proceedings2006 31st IEEE Conference on, 967–974. doi:10.1109/LCN.2006.322210

[10] http://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/amp/

[11] http://flylib.com/books/en/2.464.1.18/1/