International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue V, May 2019- Available at www.ijraset.com

Real-time System for Detection of DOS attack using Data Mining Algorithms

Parakh Shah[1], Harsh Bhayani[2], Archana Gupta[3]
[1, 2, 3]*Department of Computer Engineering, K.J.S.C.E, Mumbai, India*

Abstract: ***There is a marked increase in transactional services such as online shopping, online trading etc. provided on the internet.[6] Along with this growth, there is an increase in the frequency of malicious attacks attempting to breach the websites' integrity. The Denial of Service (DOS) is one type of malicious attack on the internet world which experienced infamy during the late 1990s and it's still a cause of a problem for network security officials today. DOS attacks are rapidly developing a threat to today's Internet. This paper explores the concept of detection of DOS attacks using various data mining algorithms such as Random Forest, KNN and SVM. It presents a comprehensive survey of DOS attacks and detection method algorithms. In this paper- open issues, research challenges and possible solutions in this area are also highlighted. NSL-KDD Cup '99 dataset[8] is used for applying Data Mining algorithms and testing. This paper aims at developing a system to detect DOS attacks on a system in real time and most accurately distinguish between legitimate and malicious network traffic.***
Keywords- DOS attack, Random Forest, KNN, SVM, NSL-KDD dataset.

I. INTRODUCTION

A Denial-of-Service (DOS) attack is an attack intended to shut down or make inaccessible a network or a network in the machine. The DOS attack is used to tie up a website's resources so that users who need to access the site cannot do so. During a DOS attack, the target is either flooded with a high number of traffic packets or by sending the target information that triggers a crash. Hence the DOS attack is classified in two forms- flooding attacks and crash attacks.[1] This paper takes into consideration the flooding form of the DOS attacks.

Flooding attacks are executed using different methods.

A. SYN Flood Method

In SYN Flood method the client sends a SYN/TCP package to the server and the server sends a response with ACK package and then client continually sends many of SYN/TCP packages.[2] In this case, the server is not able to respond to all of the sent packages

B. UDP Flood Method

This attack randomly selects ports and sends many UDP packages to the server.[2]

C. HTTP Flood Method

This attack is a method which hackers use for damaging the server.[2] The hacker creates disorder in services of the server by sending many HTTP requests to the server

II. BASIC CONCEPTS

In section 2.1, an in-depth analysis of the KDD Cup '99 dataset is done where we classified the features into different groups. Further, different types of attacks were classified into 4 categories as mentioned in 2.2. For the proposed system 3 different classification algorithms were applied on the KDD Cup '99 dataset, thus, 2.3 explains the working of the three algorithms and accuracy of the three algorithms.

A. Analysis of the KDD Cup '99 Dataset

The KDD Cup '99 dataset is the most preferred intrusion detection dataset available. The dataset contains 41 input features, 34 continuous-valued, and 7 discrete-valued, further divided into *basic features* and *high-level features.* The data classified as *basic features* is directly obtained from the header of IP packets of the protocol such as TCP/UDP/ICMP. The high-level features are divided into typical features such as the number of failed login attempts, number of root access, number of file creation operations, etc. and into *time-based* and *connection-based derived features* to check the connections with a 2 second time window.[3]

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue V, May 2019- Available at www.ijraset.com

B. Attacks in KDD Cup '99 dataset

Only 10% of the KDD Cup '99 dataset was used for the system which contains 4,94,201 records. The dataset contains 22 different types of attack which is further grouped into 4 categories - 'Probe', 'U2R', 'R2L', 'DOS' (as shown in the table 1 below).

Classification of attacks	Attack name
DOS	smurf, land, pod, teardrop, back
R2L	ftp_write, guess_passwd, imap, multihop,phf, spy, warezmaster, warezclient
U2R	perl, buffer_overflow, rootkit, loadmodule
Probe	ipsweep, nmap, satan, portsweep

Table 1: Classification of attacks [4]

III. CLASSIFICATION ALGORITHMS

A. Random Forest

The random forest algorithm is a supervised learning algorithm generally used for both classification and regression based models. It builds multiple decision trees during training and outputs the mode of classes (classification) or the mean prediction (regression) of the individual trees. The different decision trees are trained on the training dataset KDD Cup '99, but the training data used for learning for each tree is different using Bagging.

B. SVM

Support Vector Machine (SVM) is a supervised learning algorithm used mainly for classification problems. In SVM the attributes of network packets are represented as data items which are plotted as points in an n-dimensional space and the algorithm classifies the data points into different classes and outputs a hyperplane that acts as a separator between the classes.[5]

C. KNN

The K-Nearest Neighbours (KNN) algorithm is a supervised learning algorithm that is used for both classification and regression based problems. KNN is based on feature similarity. The attributes of network packets are plotted in a n-dimensional space in the form of data points. The k (number) data points closely grouped together are classified into one class. Each data point from the training dataset is classified based on the majority of its k-neighbours, i.e. the data point is assigned to that class where it is closest to k-nearest neighbours of that class.

IV. PROPOSED SYSTEM

Proposed system for the prediction of DOS attack is shown in Fig 1. The proposed system figure lays down the steps to be followed for designing the system. The dataset to be tested will first be preprocessed for any anomalies and any data entries that would hinder the accuracy of the classifier. The classifier works on this output and gives its result based on 4 categories namely DOS, Probe, U2R, R2L. If a DOS attack is detected, the system notifies the user.

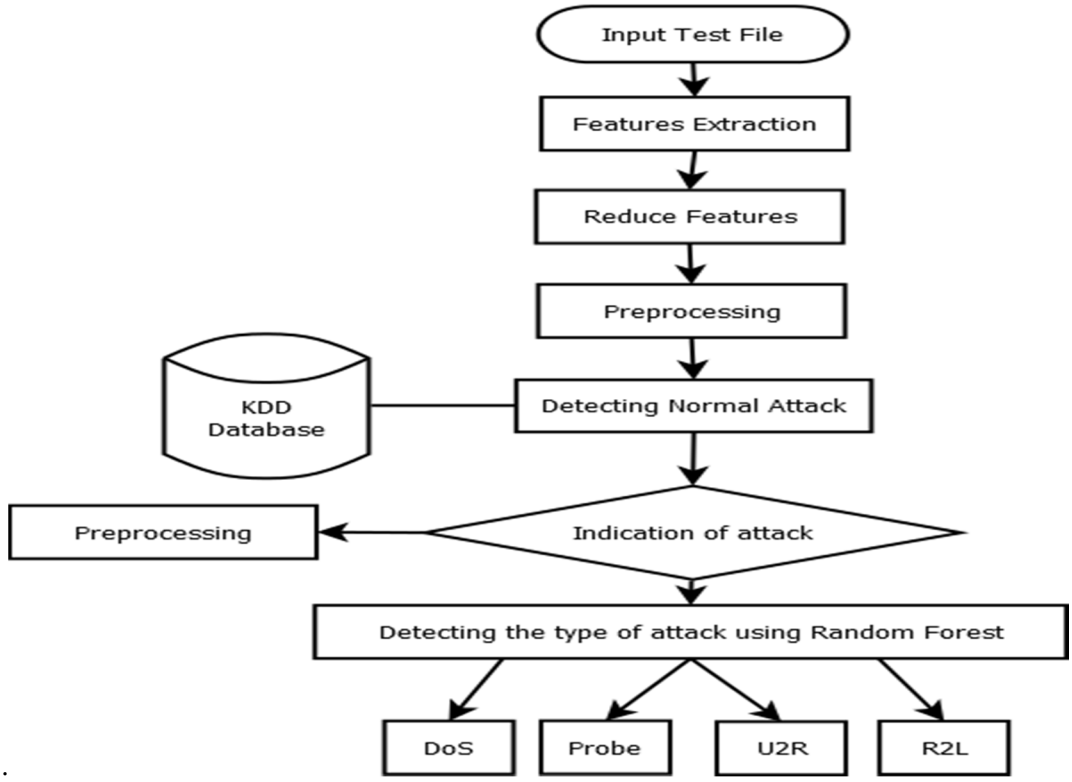

Fig 1. Proposed system for the prediction of DOS attack [4]

Section 3.1 explains the analysis using R programming that highlighted key parameters of the data packets that are active during DOS attack. [citation of 13 parameters] Comparing the accuracy of the three algorithms chosen, it was noted that the random forest algorithms' results gave the highest accuracy.

A. Pre-work

For the analysis of the dataset and finding the crucial patterns of how and with what features or attributes does the DOS attack occur, R programming was used to plot various combinations of parameters and obtained the following observation plots. Using these plots and analysis, parameters were selected and filtered to predict DOS attack using the classification algorithm applied to the KDD Cup '99 dataset.

1) Observations During Data Analysis[7]

a) Flag is a strong predictor for DOS attack. For flag= "REJ" and flag="S0", it is mainly DOS. Fig 2 shows the observations based on the considered dataset.

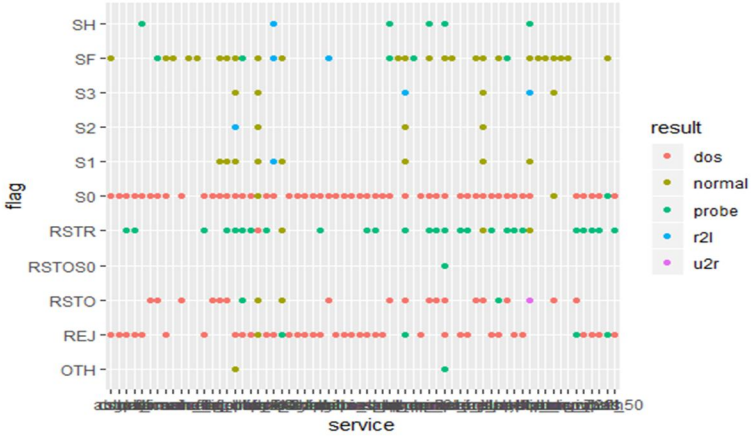

Fig 2. Flag Observations

b) For protocol-type="TCP" has greater possibility for DOS attack as induced by the study of data as shown in Fig 3. It is also a strong predictor of DOS.

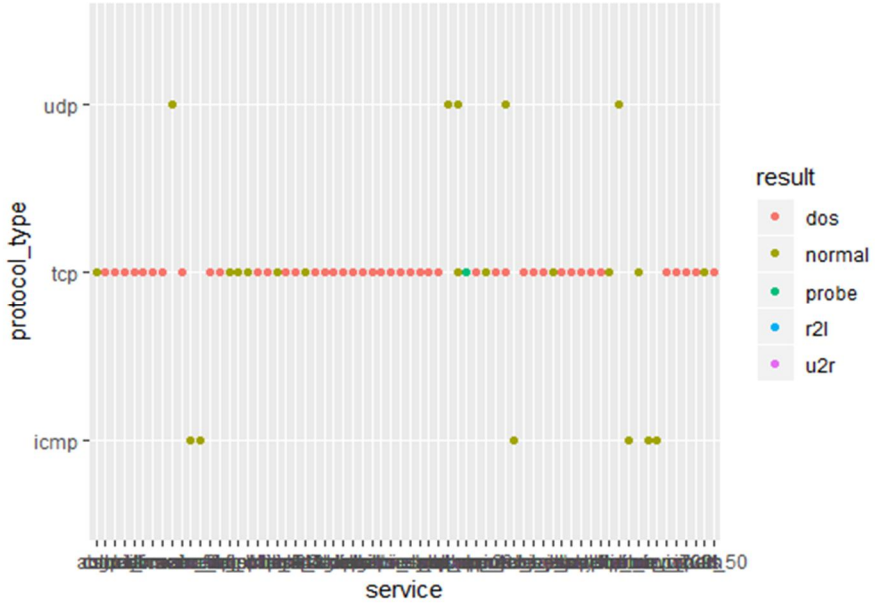

Fig 3. Study of Protocol type

c) For DOS attack serror_rate=1 and srv_serror_rate=0 or 1 as shown in Fig 4.

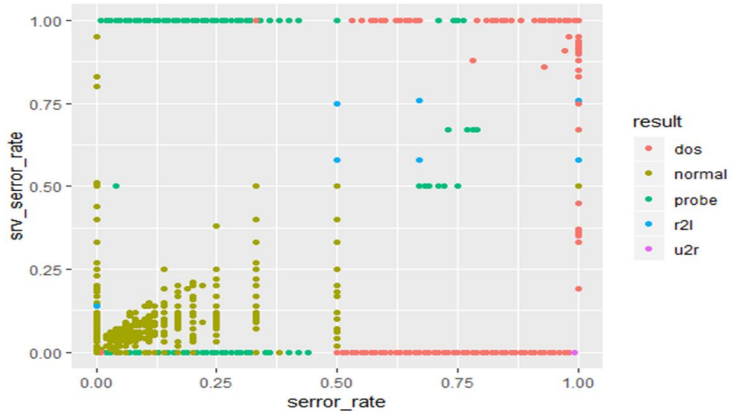

Fig 4. Study of serror_rate and srv_serror_rate

B. Feature Extraction

Due to the large size of the dataset, with more than 41 attributes monitored, it is extremely important to only consider the salient features for faster results and better accuracy. Based on the above mentioned observations of the dataset, feature extraction is done. Thus for the rest of the prediction process, some of the attributes will not be considered. The reduced attributes should maintain similar performance and accuracy with the classifier compared to the entire dataset. After experimentation 13 attributes were filtered as important.[3] Namely- protocol_type, service, count, srv_ count, serror_rate, srv_serror_rate, rerror_rate, srv_rerror_rate, same_srv_rate, dst_host_srv_count, dst_host_diff_srv_rate, dst_host_serror_rate, dst_host_srv_rerror_rate.

C. Data Preprocessing

To make the data suitable to implement any prediction algorithm, preprocessing is required. Data preprocessing is required to handle missing values, to normalize the values, to maintain consistency and many more reasons. In this work preprocessing is done in Language R using logistic regression to fill the missing values.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue V, May 2019- Available at www.ijraset.com

D. Detection of Attacks

For the detection of attacks three well known algorithms are implemented in this system. These algorithms are giving different accuracy, and perform well in different scenario. The performance and parameters used in the implementation of these algorithms are given below:

In Random Forest algorithm, the parameters chosen were n_estimators = 10, criterion=gini, min_samples_split = 2. This algorithm has high precision and recall. [1]

In SVM algorithm, the kernel can be linear or non-linear kernel. Here, rbf kernel which is a non-linear kernel mode is used. Parameters chosen were gamma = auto, degree = 3, kernel = rbf, C which is the regularization parameter is set to 1.

In KNN algorithm, parameters chosen were n_neighbors = 5, weights = uniform, leaf_size = 500, p which is the power parameter for minowski metric is set to 2.

These 3 classification algorithms were trained using the dataset and used for real-time DOS detection system.

V. REAL-TIME IMPLEMENTATION

The above given proposed system was designed and trained using KDD Cup '99 dataset. During the real time implementation of the system the major requirement is to collect the data as per dataset designed based on the real time data traffic in the network. So the implementation of the system starts from collecting the data based on the traffic in the network.

A. Real Time Data Collection from the Network

In order to analyze network packets in real time, data of the network packets is needed. The target machine creates this dataset with the help of a linux terminal, tcpdump and bro library. Initially, the tcpdump library traces the network packets and generates a packet capture .pcap file which consists of real-time network data. However, this is not to predict a DOS attack. The .pcap file is converted to a text file using the bro library. The text file consists of 47 parameters which are the properties of the network packets. Using python code this text file was converted to a .csv file and a dataset similar to KDD Cup was obtained.

B. Description of DOS Attacking Tools

Low Ion Orbit Cannon(LOIC) is an open source tool which allows users to perform DOS and DDOS attacks using either the website url or the target machine's IP. It is platform independent as it works on Windows and Unix/Linux systems. Two separate machines have been used, a target machine and a source machine to perform a DOS attack. The tool is installed on the source machine and using the IP of the target machine a DOS attack was performed on the target machine.

C. Experimental Results

1) *Random Forest:* From the KDD Cup '99 dataset, we trained the algorithm with 30,000 tuples and 13 attributes that are prominent in packets of a DOS attack. Using Random Forest ensured that the data is not overfitted and the accuracy obtained was 99.91%.
2) *SVM:* The SVM algorithm was trained on the KDD Cup '99 dataset and classified the data points by separating the data points representing the attributes of normal traffic packets to the data points representing the attributes of the packets that were used for the DOS attack. The accuracy obtained using the SVM algorithm was 99.04%.
3) *KNN:* The KNN algorithm was trained on the KDD Cup '99 dataset. The accuracy obtained using the KNN algorithm was 99.10%.

VI. CONCLUSION AND FUTURE WORK

The paper relates to detection of real-time DOS attack using data mining techniques. To maintain security of the server and other online transactional services, it was important to secure it from DOS attacks. Maintaining the web integrity from such malicious attacks was the main aim of this paper. In this paper, we created a real-time system to detect the suspicious behaviour of the attacker by detecting the IP of the attacker using data mining algorithms. We conclude that even though Random Forest algorithm is faster than KNN algorithm, KNN algorithm gives better output than other algorithms for detection of DOS attack in real-time.

The future work will be detection other different types of attacks by the proposed model using the aforementioned three algorithms. The system can be extended to be a true Intrusion Detection System (IDS) for detection any kind of harmful attack on the users machine.

VII. ACKNOWLEDGEMENT

The authors would like to thank their project mentor Prof. Archana Gupta for guiding them throughout the research of the project.

REFERENCES

[1] Anand Keshri, Sukhpal Singh, Mayank Agarwal, and Sunit Kumar Nandi, “DoS Attacks Prevention Using IDS and Data Mining”, 978-1-5090-4291-3/16/$31.00, IEEE 2016.
[2] H. Jelodar and J. Aramideh, "Presenting a pattern for detection of denial of service attacks with web mining technique and fuzzy logic approach," *2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)*, Kanyakumari, 2014, pp. 156-160.
[3] Ralf C. Staudemeyer, Christian W. Omlin, “Extracting salient features for network intrusion detection using machine learning methods”, SACJ, Submission, 2014.
[4] Phyu Thi Htun and Kyaw Thet Khaing, “Detection Model for Denial-of-Service Attacks using Random Forest and k-Nearest Neighbors”, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Volume 2, No 5, May 2013.
[5] V. Justin, N. Marathe and N. Dongre, "Hybrid IDS using SVM classifier for detecting DoS attack in MANET application," *2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, Palladam, 2017, pp. 775-778.
[6] Ghafar A. Jaafar, Shahidan M. Abdullah, and Saifuladli Ismail, “Review of Recent Detection Methods for HTTP DDoS Attack”, Journal of Computer Networks and Communications Volume 2019, Article ID 1283472.
[7] R programming analysis - https://rpubs.com/Pratik/196063
[8] KDD Cup’99 dataset - http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html