International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 7 (2018) pp. 5175-5178

Classifying the Probe attacks Using Machine Learning Techniques in R and Hadoop

B. V. Praveen
Department of I.T.
VRSEC, Vijayawada, India

P. Rama Devi
Assistant Professor, Department of I.T.
VRSEC, Vijayawada, India

D. Mahita
Department of I.T.
VRSEC, Vijayawada, India

A. Sudheshna
Department of I.T.
VRSEC,Vijayawada,India

Abstract

Attacks on computers are increasing day by day in different way and intrusion detection techniques that overcomes those attacks came into existence. Popular machine learning techniques like SVM, Decision Tree, KNN and Naïve Bayes are used to classify the probe attacks. KDDcup99 data set is taken for analysis purpose. Principal component analysis is used for attributes reduction. The performance of classifiers is compared for accuracy.

Keywords: Decision Tree, PCA, Intrusion, Probe attacks.

INTRODUCTION

When an information security, integrity and confidentiality are compromised due to set of actions which are either digital or physical then, those actions are said to be Intrusion. Software which detects this type of actions that acts on computer systems is called Intrusion Detection System. Majorly, there are 4 types of intrusion detection systems such as,

- Network Intrusion Detection System,
- Host Intrusion Detection System,
- Signature-based Intrusion Detection System,
- Anomaly-based Intrusion Detection System [2].

These intrusion detection [6] systems complement with firewall security by detecting attacks if someone tries to break in through firewall security and tries to have access on system and alerts the system administrator in case of breach in security [1]. But they can't detect some attacks which have no signature [4] and also it is difficult to configure. For that purpose we want to use machine learning techniques to detect the attacks [3].

R-Programming:

R is a language used for statistical computing and analysis and also describing using graphics. R is available as free software under GNU General Public License. It can run on wide variety of platforms like Windows, Linux, MacOS. R is used for data manipulation, analysis, and also displaying those using graphics. R also has a feature to connect to Hadoop and load the data from it and perform analysis on the data present in Hadoop file system. Due to its high performance, storage facility and effective work its usage is increasing day by day in performing different activities on data.

Hadoop Distributed File Structure:

Hadoop[9] is provided by Apache to process and analyze on the 3 v's of data i.e., velocity, variety and volume. It is written in Java and it is an open source framework. There are different modules in Hadoop but we use one module of it called as HDFS (Hadoop Distributed File System)[7]. By using Hadoop, we can easily store and access the large amount of data in a less time and can also perform analysis on the data in it.

In Hadoop we have a data ware house system called as Hive that runs SQL queries on it. These Hive Query Language is converted into Map Reduce jobs in backend.

Next, we discuss about Data set description, storing the data, Integrating r and Hadoop, processing and building a classification model using r and discuss about results. Finally we will conclude and mention the future work.

Data Set Description:

KDD CUP 99 data set is used mainly to analyze the different attacks. It consists of nearly 4,900,000 samples with 41 features and each sample is classified as either normal or attack. This largest dataset is called 'whole KDD'. This samples consists different type of attacks and they are shown in below table.

Table I: Different type of attacks present in KDD cup 99 Dataset

Attack Category	Attack Name
Probe Attacks	Satan, nmap, portsweep, ipsweep
Remote to local Attacks	Warezclient, spy, phf, imap, guess_passwd, ftp_write
Denial of service Attacks	Neptune, pod, land, smurf, teardrop
User to root Attacks	Buffer_overflow, perl, rootkit

Here, we have taken data 10% KDD and classify the probe attacks using some machine learning techniques.

Table II: The attributes list of KDD CUP99 data set

Feature Index	***Feature Name***
1	*Duration*
2	*protocol_type*
3	*Sevice*
4	*Flag*
5	*Scr-bytes*
6	*dst_bytes*
7	*Land*
8	*wrong_fragment*
9	*Urgent*
10	*Hot*
11	num_failed_logins
12	logged_in
13	num_compromised
14	root_shell
15	su_attempted
16	num_root
17	num_file_creations
18	*num_shells*
19	*num_access_files*
20	*num_outbound_cmds*
21	*is_host_login*
22	*is_guest_login*
23	*Count*
24	*srv_count*
25	*serror_rate*
26	*srv_serror_rate*
27	*rerror_rate*
28	*srv_rerror_rate*
29	*same_srv_rate*
30	*diff_srv_rate*
31	*srv_diff_host_rate*
32	*dst_host_count*
33	*dst_host_srv_count*
34	*dst_host_same_srv_rate*
35	*dst_host_diff_srv_rate*
36	*dst_host_same_src_port_rate*
37	*dst_host_srv_diff_host_rate*
38	*dst_host_serror_rate*
39	*dst_host_srv_serror_rate*
40	*dst_host_rerror_rate*
41	*dst_host_srv_rerror_rate*

In this dataset the probe attacks of different typesare considered as “True” and all the remaining attacks are considered as “False”. It is done to classify the probe attacks among the whole dataset and if it consists of attacks that will fall into different category. The probe attacks and their description is given below:

IPsweep: used to determine the hosts listening on a network.

Nmap: It is used for network scans.

Satan: used for remotely analyzing security of networks.

Port sweep:used to find weak access points to break into a computer system.

Storing data in Hive table:

Firstly, we need to create a table in hive with same attribute names present in the dataset[8]

- Create table [table name] ([column name] [data type]) terminated by ‘\t’ stored as text file;

  then write a command to load the dataset from HDFS to hive table

- Load data local inpath “[path of the file]” into table [table name];

  By executing above 2 commands we can import our data into hive table.

Integrating R & Hadoop:

We can connect Rstudio and Hadoop using 3 libraries[8]. They are

1. **rjava**
2. **RJDBC**
3. **RDBI**

Install these 3 libraries in R studio[7] and load them.

To access the data in R studio from hive table, we have to start hive service and then connect the hive and R by writing required queries.

The required code is written in Rstudio then it is connected and we can access the data present in hive table in Rstudio.

Processing and building classification model using R:

Now we will describe how we find accuracy of each model using R

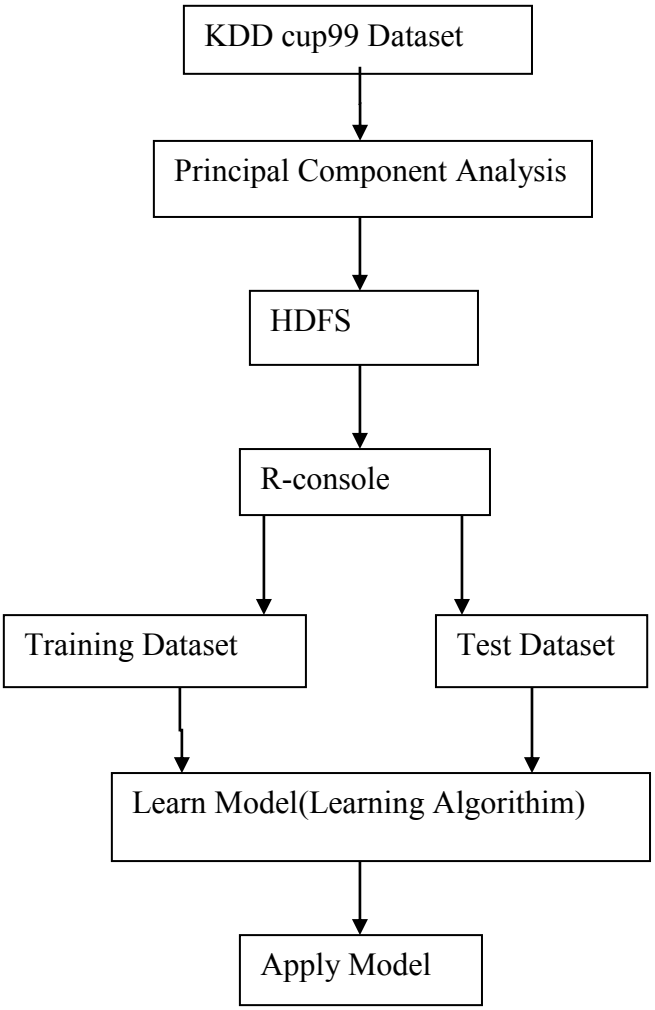


Fig I: Architecture of Proposed Model

It is done it in 3 steps.

In first step, we applied principal component analysis on the data which is used to select the attributes that are required in classification. PCA uses orthogonal transformation to convert possibly correlated variables into set of linearly uncorrelated variables called principal components. For that purpose we use Rapid Miner tool and select the attributes with positive threshold value sorted in descending order.By using PCA we got 16 attributes that are required in classifying data.

Table III: Attributes selected using PCA

Dst_host_diff_srv_rate	Protocol_type
Rerror_rate	Dst_host_count
Dst_host_rerror_rate	Serror_rate
Dst_host_srv_rerror_rate	Dst_host_serror_rate
Srv_rerror_rate	Dst_host_srv_serror_rate
Diff_srv_rate	Srv_serror_rate
Service	Duration
Flag	Count

In second step, we uploaded the dataset with 16 attributes into hive database and connected R with hive to retrieve the dataset.

In third step, we will write the code in R and separate the training dataset (i.e.; 80% samples in dataset) and test dataset (i.e.; remaining 20% samples in dataset).

In fourth step, we let the machine to learn different classifier algorithms using training dataset such as SVM, Decision tree, KNN, Naïve Bayes. After learning on the training dataset It will try to predicton the test dataset. After prediction we have to calculate the True Positive, True Negative, False Positive, False Negative values to know how accurately the models can predict the probe attacks on test dataset.

True Positive:Number of positive samples correctly predicted by classification model.

True Negative:Number of negative samples correctly predicted by classification model.

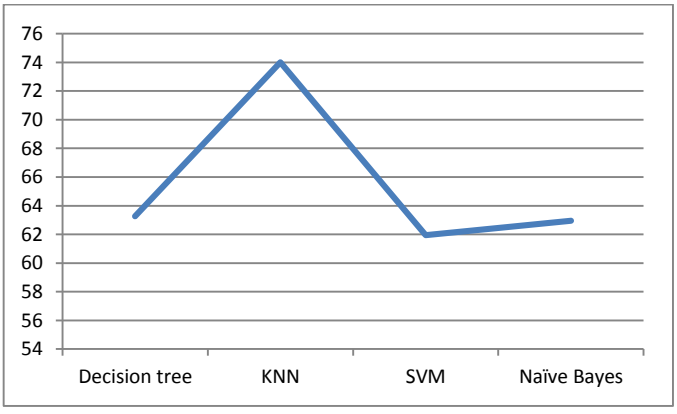
False Positive: Number of negative samples wrongly predicted as positive by classification model.

False Negative: Number of positive samples wrongly predicted as negative by classification model.

Accuracy: Accuracy is defined as ratio between numbers of correct predictions to the total number of predictions made by classification model.

Accuracy = (TP+TN) / (total no. of predictions).

The below figure shows performance of each classifier. X-axis represents classifier and y-axis represents accuracy of each classifier.



CONCLUSION:

Here, to determine the probe attacks we have applied various machine learning techniques i.e., Decision Tree, KNN, SVM, Naïve Bayes on KDDcup99 dataset. Among them KNN classifier gives more accuracy than remaining classifiers .The performances of all classifiers are clearly shown in the above graph. Later we try to improve the accuracy of these algorithms and also use different genetic algorithms, Fuzzy[5]

logic algorithms to improve accuracy in predicting the attack along with loading more amounts of data in Hadoop.

REFERENCES

Journal:

[1] Bharath Kumar, Gowru. And Ramadevi, Polagani. "Efficient Privacy Protection in Social Networks". International Journal of Science and Research (IJSR).Volume 3 Issue 10, October 2014.

[2] Garcia- Teodoro, P. Macia , G and Vazquez, E. "Anomaly-based network intrusion detection" . Computers and Security. Volume 28 Issue 1-2,February,2009. Pages 18-28.

[3] Lakshmi, T.V.N and Babu, V.K. " Detection of User to Root Attacks using Machine Learning Techniques". International Journal of Advanced Engineering and Global Technology (IJAEGT). Vol. 3, Issue 3, Mar 2015.

[4] Nguyen, T.T.T. & Armitage, Grenville. (2008). Grenville, A.: A Survey of Techniques for Internet Traffic Classification using Machine Learning. IEEE Communications Surveys & Tutorials 10(4), 56-76. Communications Surveys & Tutorials, IEEE. 10. 56 - 76. 10.1109/SURV.2008.080406.

[5] Shanmugavadivu, R. and Nagarajan, N. "Network Intrusion Detection System using Fuzzy Logic". Indian Journal of Computer Science and Engineering .Vol. 2 No. 1, pp-101-111,2011.

[6] Vinchurkar, D.P. and Reshamwala, A. "A Review of Intrusion Detection System using Neural Network and Machine Learning Technique". International Journal of Engineering Science and Innovative Technology (IJESIT). Vol 1, Issue 2, Nov 2012.

[7] Mathod T., E.Bhagheri, Wei Lu, and A.A. Ghorbani,"A Detailed Analysis of the KDD CUP 99 Data Set", p.2, 2009.

Book:

[8] Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data" . EMC Education Services, January 2015.

[9] Vignesh Prajapati. "Big Data Analytics with R and Hadoop". packt publishing.2013 .

[10] Chris Eaton, Dirk DeRoos, Tom Deutsch, George Lapis, Paul Zikopoulos, "Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data". McGrawHill Publishing. 2012.