International Journal of Innovations in Engineering and Technology (IJIET)
http://dx.doi.org/10.21172/ijiet.81.044

Classification of Attacks in Data Mining

Bhavneet Kaur
Department of Computer Science and Engineering
GTBIT, New Delhi, Delhi, India

Abstract- Intrusion Detection and data mining are the major part of almost every application. With a diverse expansion of internet use, there are millions of intrusions that need to be detected and to get rid of them. Intrusion Detection System (IDS) is an effective tool that helps to prevent unauthorized access to system resources by various mechanisms and one among them is using machine learning tool called weka. It is an open source environment available for machine learning and data mining that could be used for classification of attacks. Collection of machine learning algorithms for data mining task written in java used for classification, clustering, association etc. This paper emphasis on performance evaluation and comparison of various classifiers on the basis of different types of attacks The paper is set to make comparative analyses among machine learning algorithm zeror, oner, naive, J48.

Keywords – Classifiers, Oner, Zeror, Naive Bayes , J48.

I. INTRODUCTION

An intrusion detection system (IDS) is a type of security software designed to alert administrators when someone or something is trying to compromise information system through malicious activities or through security policy violations. An IDS works by monitoring system activity through examining vulnerabilities in the system, the integrity of files and conducting an analysis of patterns based on already known attacks. The IDS can be classified into two categories .a) A host based intrusion detection System that monitors the activities associated with an individual host. It would only access the data coming in or going out from other host system. b) Network intrusion detection system on the other hand would monitor complete network traffic.

There are two general categories of Intrusion detection system: Anomaly based intrusion detection and signature based intrusion detection. Anomaly IDS is quiet complex. In real life it is just like security guard who protects people from theft, robbery or any illegal activity. In technical terms it collects all the data packets that are coming towards the network. Then it distinguishes the packets as legal & illegal allows only genuine raffic to pass through the network. At the initial state Anomaly detection system monitors the functioning of the network and makes a pattern of its working. When a packets tries to enter the network its behavior get analyzed, and if it produces anomalous behavior, it will be considered as attack and will not be allowed to enter the network. But anomaly detection cost high processing overheads. Whereas Signature based IDS makes use of database to identify the type of attack. The database consist of type of attack and all the information like its nature, its evidence etc. Whenever the agent collects the data first it compared with the list of attacks stored in the database, and only if the behavior matches, data will not be allowed to enter the network. But it has a disadvantage that the database must be consistently maintained. This can lead to building a huge database which takes up a lot of space .The database need to be specific so that variations on known attacks are not missed [1].

II. INTRODUCTION TO KDD99 DATASET

The term Knowledge Discovery in Database is a vast process of retrieving useful information from large set of data and emphasize on high level applications of data mining. It is of interest to researchers in machine learning, pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition for expert systems, and data visualization [2].

The 1998 DARPA Intrusion Detection Evaluation Program was prepared and driven by MIT Lincoln Labs. The goal was to examine and explore different ways of intrusion detection. A standard set of data was estimated, which includes a variety of intrusions simulated in a military network environment .The KDD'99 dataset is a part of

DARPA dataset which is prepared by Sal Stofo and Wenke Lee. The KDD dataset was acquired from raw tcpdump data for a period of nine weeks. It is formed by including large number of network traffic activities that include both general and malicious connections. [3]

A. Types of Attacks

*DOS Attacks:*A DOS which also refers as distributed denial of service is a method in which users working on a particular system failed to use the resources and couldn't complete their work. Eventually the resource sits idle and the system slows down, may leads to the system crash.

User to Root Attack: In this type of attack the sniffer or attacker tries to log in to user account and if successful then tries to gain the access to administrator or super user rights and exploits the system.

Probing: In this attack, the attacker unlike the above two attacks examines the network of computers and examines it to find the delicate points from where attacker an easily enters the network and exploit the network by malicious activity.

Remote to User Attack (R2U): In this type of attack, an attacker first enters the network illegally and then sends malicious packets over the network to exploit the system or at times to gain access to secret information [4].

Table -1 Attack types and their category

Category	Attack Type
DOS Attack	apache, back, land, mailbomb, neptune, pod,processtable, smurf, teardrop,udpstorm
U2R Attack	apache, back, land, mailbomb, neptune, pod,processtable, smurf, teardrop,udpstorm
R2L Attack	ftp_write, guess_password, imap, multihop,named, phf, sendmail,snmpgetattack, snmpguess, warezmaster, worm, xlock, xsnoop, httptunnel
probe	ipsweep, mscan, nmap, portsweep, saint, satan

III .STRUCTURE OF KDD99 DATASET

In case KDD dataset consist of 41 attributes features. Among them, 1-9 used to represent the basic characteristics of a packet, 10-22 reflect content features, 23-31 are for traffic features and 32-41 for host based features. [5]. They are basically grouped into three categories: basic features of individual connection, content features within a connection, & traffic features which are computed by two seconds time window.

.

Table -II THE 41 FEATURES IN KDD'99

S.No	Features	S.No	Features
1	Duration	22	Is_guest_login
2	Protocol Type	23	Count

3	Service	24	Serror_rate
4	Src_byte	25	Rerror_rate
5	Dst_byte	26	Same_srv_rate
6	Flag	27	Diff_srv_rate
7	Land	28	Srv_count
8	Wrong_fragment	29	Srv_serror_rate
9	Urgent	30	Srv_rerror_rate
10	Hot	31	Srv_diff_host_rate
11	Num_failed_logins	32	Dst_host_count
12	Logged_in	33	Dst_host_srv_count
13	Num_compromised	34	Dst_host_same_srv_count
14	Root_shell	35	Dst_host_diff_srv_count
15	Su_attempted	36	Dst_host_same_src_port_rate
16	Num_root	37	Dst_host_srv_diff_host_rate
17	Num_file_creations	38	Dst_host_serror_rate
18	Num_shells	39	Dst_host_srv_serror_rate
19	Num_access_shells	40	Dst_host_rerror_rate
20	Num_outbound_cmd	41	Dst_host_srv_rerror_rate
21	Is_hot_login		

IV.DATA MINING CLASSIFICATION METHODS

A. ZeroR classifier: It is the basic method for classifying different data sets. This technique gives the result in favour of common class. For example Consider the "Weather" data set. It consist of four attributes outlook, temperature, humidity and windy with their corresponding values and a class play which can have any values among yes or no. Now suppose that dataset "Weather" consist of 10 instances among which 6 instances shows that child can play and 4 instances shows that child cannot play .So when ZeroR classifer runs it results in "child can play". It means that zeror results in favour of majority without predicting about other outcomes. It is low level classifier and cannot be therefore used on complex data set.

B. OneR classifier: OneR, short for "One Rule", is better then ZeroR classifier.It is a simple, yet accurate, classification algorithm that generates a rule for each instance in the data set and selects the rule with the minimum error rate..

C. NaiveBayes classifier: The naive Bayes model is a heavily simplified Bayesian probability modelthat calculates a set of probabilities by counting the frequency and combinations of values in a given data set. The algorithm uses Bayes theorem and assumes all attributes to be independent given the value of the class variable [6]. Thus it is completely opposite strategy when compared with OneR classifier .Here all attributes are equally important unlike OneR classifier but attributes are statistically independent i.e. knowing the value of one attribute does not provide any information about the value of another attribute.

D. Decision tree algorithm J48: J48 classifier is a simple C4.5 decision tree for classification. It creates a binary tree. It is an open source Java implementation of the C4.5 algorithm in the Weka data mining tool. C4.5 is a program that creates a decision tree based on a set of labelled input data. This algorithm was developed by Ross Quinlan. It is the most useful approach in classification problem. With this technique, a tree is constructed to model the classification process. Once the tree is built, it is applied to each tuple in the database and results in classification for that tuple [7].

V.STANDARD TERMS

a. True Positive Rate: If the outcome from the prediction is same as that of actual value, then it is called as true positive. Tp= Tp/Tp+ FN.

b. False Positive Rate: If the outcome from the prediction is different from the actual value, then it is called as false positive. Fp= Fp/FP+TN.

c. Precision: It is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. It is usually expressed as apercentage.TP/TP+FP.

d. Recall: It is the ratio of the number of relevant records retrieved to the total number of relevant records in the database.

Table- III TRUE POSITIVE RATE OF ALL THE ATTACKS

Attacks/Classifier	ZeroR	OneR	NaiveBayes	J48
back	0	0.990	0.973	0.997
teardrop	0	0.990	0.997	1.000
Neptune	0	1.000	0.998	1.000
land	0	0.000	0.765	0.941
smurf	0	0.997	1.000	1.000
pod	0	0.980	0.971	0.980
loadmodule	0	0.000	0.25	0.000
rootkit	0	0.000	0.571	0.000
bufferoverflow	0	0.000	0	0.583
perl	0	0.000	0.917	0.000
phf	0	0.000	0	1.000
ftpwrite	0	0.000	0.25	0.000
multihop	0	0.000	0.571	0.286
warezmaster	0	0.000	0.8	0.800
guess_passwd	0	0.849	0.943	0.943
satan	0	0.106	0.967	0.986
ipsweep	0	0.891	0.99	0.984
nmap	0	0.000	0.502	0.961
imap	0	0.000	0.917	0.250
portsweep	0	0.000	0.909	0.977

From table above it can be seen that zeroR classifier is a weak classifier. It cannot be used for determining the accuracy .It relies only on single value discarding the rest. Except the smurf attack it could not detect other attacks.

So it is the least used classifier for determining intrusions. OneR classifier on the other hand works better in detecting attacks as compared to zeroR classifier .It sets rule for each of the attributes in dataset and selects that rule which has the lowest error rate unlike the zerorR.

Naive Bayes classifier follows the probabilistic method (Bayes theorem).From above it can be seen that naive works well in determining Denial of service attacks(back, land , neptune , pod, smurf, teardrop,) as compare to other attacks. J48 classifier on the other hand builds a decision tree and works on each node of the tree.

VI. ENSEMBLE LEARNING

Using a single classifier on a data set may not give the accurate results. One of the ways to overcome this problem is ensemble learning i.e. not to choose best-performing learning scheme for the dataset (using cross-validation) but to use them all and combine the results. In Weka we can select multiple classifiers to be used in together for the analysis of dataset. The package used is weka.classifier.meta.vote.

A. *Majority Voting*

If you select majority voting as the combination rule, then each of these classifiers will predict a nominal class label for a test sample. The label which was predicted the most will be taken as the output of the vote classifier Consider a real life example, when there is a need to make a critical decision several suggestions are taken from number of experts instead of relying only on the judgment of one expert.[8]

In data mining also the model generated by a classifier can be regarded as an expert and if we combine several models then that would create an ensemble. But ensemble does have a disadvantage also: it get hard to analyze several models and their result at the same time.

Table-IV TRUE POSITIVE RATE USING MAJORITYVOTING

Attacks	***True Positive Rate***
back	0.993
teardrop	1.000
neptune	1.000
land	0.941
smurf	1.000
pod	0.980
loadmodule	0.000
rootkit	0.000
bufferflow	0.667
perl	0.000
phf	0.667
ftp_write	0.000
multihop	0.000
Warezmaster	0.800
Guess_password	0.943
satan	0.992

Attacks	*True Positive Rate*
ipsweep	0.997
nmap	0.502
imap	0.667
portsweep	0.901

VII CONCLUSION

This paper presents a study and comparative analysis between various machine learning classifiers such as zeror, oner, naive bayes, J48. And among them it can be concluded that Naive Bayes and Decision Tree are surely the best algorithms for the better results as ZeroR and OneR algorithms are used just to give a start-up, It also explains briefly the types of attacks such as DOS, U2R, R2L, Probes.
Each classifier has its own merits that help in improving the accuracy of classifier and to achieve efficiency in detecting various attacks and their types. Also to overcome the demerits of each classifier a technique called as majority voting has been explained. Future work includes testing more attacks with more classifiers and how it works in other real time environment.

VIII FUTURE SCOPE

These intrusion detection techniques provides a different outlook into security and network issues that an organization face .Weka the data mining tool, is a collection of machine learning algorithm that can be implemented in every real-life data mining application. It can also be used in other fields such as medical, education, marketing etc and can also be used to make comparison among other machine learning algorithms.

REFERENCES

[1] Deepika P Vinchurkar, AlpaReshamwala (2012) “A Review of Intrusion Detection System Using Neural Network and Machine Learning Technique”, International Journal of Engineering Science and Innovative Technology (IJESIT) Volume 1, Issue 2, November 2012.
[2] Mr.Kamlesh Lahre, Mr.Tarun dhar Diwan, Suresh Kumar, Kashyap , PoojaAgrawal “Analyze Different approaches for IDS using KDD 99 Data Set”, International Journal on Recent and Innovation Trends in Computing and Communication ISSN 2321 – 8169, Volume: 1, Issue: 8 , August 2013.
[3] Md.Al Mehedi Hasan, Mohammed Nasser, Biprodip Pal1, Shamim Ahmad “Support Vector Machine and Random Forest Modeling for Intrusion Detection System (IDS)”, Journal of Intelligent Learning Systems and Applications, 2014, 6, 45-52 Published Online February 2014.
[4] Subaira A. S. PG Scholar Dr N. G. P. Institute of Technology Coimbatore, India Anitha P. Assistant Professor Dr.N. G. P. Institute of Technology Coimbatore, India ” International Journal of Computer Science and Business Informatics”, ISSN: 1694-2108 | Vol. 6, No. 1. OCTOBER 2013
[5] A.M.Chandrashekhar , K. Raghuveer “Improvising an Intrusion Detection Precision of ANN Based Hybrid NIDS by incorporating Various Data Normalization Techniques - A Performance Appraisal”, IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 2, Issue 2, Apr-May, 2014 ISSN: 2320 – 8791
[6] Trilok Chand Sharma, Manoj Jain “WEKA Approach for Comparative Study of Classification Algorithm”, International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 4, April 2013
[7] Tina R. Patil, Mrs. S. S. Sherekar Sant Gadgebaba Amravati University, Amravati, ” Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification”, International Journal of Computer Science and Applications Vol. 6, No.2, Apr 2013 ISSN: 0974-1011.
[8] lanH. Witten, Eibe Frank, Mark A. Hall. “Data mining, practical machine learning tools and techniques”, Third Edition.