International Journal of Psychosocial Rehabilitation, Vol. 24, Issue 03, 2020
ISSN: 1475-7192

Detection of Malicious Intrusion Using R Based Random Forest Algorithm in Machine Learning

[1]Vinotha R, [2]Seethamani.P

ABSTRACT

The need to secure networks has increased as the number of people connecting to the network are increasing rapidly and using networks for storing or accessing critical information.An assessed and compared various machine learning algorithm and then propose a system based on the best performing algorithm.A system is an intrusion prediction system with low error rate and can be implemented in real world. The dataset utilized in the project comprises of a lot of information to assess, remembering an assortment of reproduced intrusion for a military system condition.The dataset contains mainly the normal state, DDoS and some other attacks. The system will not only predict a malicious network, but will also indicate exactly what type of attack the network is on.

Keywords: *learning, decision table, random forest, network intrusion.*

I. INTRODUCTION

With the far reaching utilization of the Internet and expanded access to online substance, cybercrime is happening at an expanding rate [1-2].Intrusion detection is the first step in preventing security attacks. Accordingly, security arrangements like the firewall, the Intrusion Detection System(IDS), the uniform risk model (UTM) and the intrusion prevention System(IPS) get a great deal of consideration in examines. IDS recognizes assaults from different frameworks and system sources by gathering data, at that point breaks down the data to identify any security ruptures [3]. System based IDS investigates information bundles that disregard a system and this examination is acted in two different ways. As of not long ago, abnormality based identification has falled a long ways behind mark based location, and in this way peculiarity based discovery stays a significant region of research [4-5].

The difficulties of irregularity based interruption discovery are that it faces another assault for which there is no earlier information to distinguish the inconsistency. Therefore, the framework should by one way or another have the insight to isolate which traffic is innocuous and which is destructive or unusual and thus AI procedures have been investigated by analysts as of late [6]. IDS isn't a response to all security issues, in any case. For instance, IDS can't

[1] Department of Information Technology, M.Kumarasamy College Of Engineering
[2] Department of Information Technology, M.Kumarasamy College Of Engineering

Received: 03 Dec 2019 | Revised: 21 Jan 2020 | Accepted: 18 Feb 2020

make up for feeble recognizable proof and confirmation components or if there is a shortcoming in arrange conventions.

With the headway in the innovation, a great many individuals are currently associated with one another through one or other type of system where they share bunches of significant information. Subsequently the need of security to defend information trustworthiness and privacy is expanded quickly. In spite of the fact that exertion have been made to make sure about information transmission and yet, assault method for rupturing the system kept on developing. Accordingly it prompts the need of such a framework which can adjust with this consistently changing assault strategies. Right now, have purposed a framework which depends on machine learning.The aim is to find the based suitable machine learning algorithm which can predict the type of network attack with highest accuracy and then develop a system which uses this algorithm to detect network intrusion. The algorithms which we have compared are Naïve Bayes, Decision Table, K Nearest Neighbor, Random Forest and Adaboost. The dataset used for training the model is KDD 99 dataset. The reason why we have used machine learning is the flexibility that it provides to the system for example, if any new type of attack is developed in future the system can be trained for predicting that attack. There are a few types of intrusion detection system out of which this is a knowledge-based IDS which is also known as the anomaly-based system. It registers the anomalies and in future predicts such malicious network to send out an alert. This way the network can disconnect to the such a connection and then have only secured connections.

The guarantee and commitment of machine learning to date are interesting. There are numerous genuine applications that are utilize today offered by machine learning. Machine learning seems to overwhelm the world in the coming days. Consequently, estimated that the test of recognizing new or zero-day assaults that associations with innovation today face can be overwhelmed by utilizing machine learning procedures. Here, built up an administered machine learning model that can arrange undetectable system traffic dependent on what has been gained from the traffic seen.

II. LITERATURE SURVEY

The intrusion detection system is classified into two types namely Network-based IDS and Host-b ased IDS. The latter monitors activities of inspected packets and resources that are being utilized by the programs. In the event of a network change, the user receives a network alert. HIDS is integrated into the IT architecture to protect information against the intruder. On the other hand, the attribute function of the target system is NIDS. Use anti-threading software to check incoming and outgoing texts. It has a signature classification that helps identify anomalies by comparing log files and previous signatures. In [2], the authors proposed a ML-based Intrusion detection system using a deep neural network. Neural networks consisting of four hidden layers and 100 hidden units was used for the intrusion detection system. They used non-linear ReLU as the activation function for the hidden layer neurons to enhance the model's performance. They adopt stochastic optimization method for learning in DNN. For the training and testing of their model they used KDD CUP 99 dataset. They were able to reach the accuracy of 99% for all the cases. In [3], they proposed a NIDS (Network Intrusion Detection System), which is based on the feature selection

method, the addition of recursive features (RFA) and the Bigram technique. They tested the model on the ISCX 2012 data set. In addition, they proposed a bigram procedure to encode the highlights of payload strings into a useful representation that can be used in feature selection. They also proposed another evaluation measure called a combination of precision, detection rate, and false alarm rate, which helps examine the modified tables and choose the best one.

The question of versatility in the field of new intrusion detection system and interference detection has been addressed [4]. The proposed IDS is a versatile arrangement that will give us the ability to identify mentioned and new occupations, as well as be renewed in a financially interested manner by the new contribution of human experts. [5], This is a statistical evaluation and analysis of the tagged flow-based CIDDS-001 data set used to evaluate anomoly-based network intrusion detection systems (NIDS). Basically, they used two techniques, the most significant k-clustering and the nearest k-class classification, to measure the degree of complexity of the prominent data. Based on the assessment, they concluded that both the closest k-mean classifications of the cluster performed well on the CIDDS-001 dataset of the outstanding data used.Based on anomalies,the dataset can be used to evaluate network IDS.

The IDS is based on anomaly detection method[6]. In such technique, a system tries to estimate the 'normal' state of the network and generates an alert when any activities deviate from this 'normal' state. The main advantage of the anomaly-based system is that it can detect new intrusion events. They classify detection techniques into three categories: statistical, knowledge-based, and machine-based.In statistical based technique, a random viewpoint is used to represent the behavior of the system. While knowledge based technique, utilize the available system data to capture the behavior of system.Finally, the machine learning-based technique uses a clear or implicit model to allow categorization of the analytical model. In [7], different machine learning techniques can lead to higher detection rates, lower false alarm rates, reasonable calculations, and communication costs of detecting interference. In this article, Mahdi Zamani and Mahnush Movahedi studied some of these techniques and patterns to compare all their performances. They divide the diagrams into methods based on classical computer intelligence (CI) and artificial intelligence (AI). They explain how various aspects of CI techniques can be used to create modern and effective IDSs.

In [8], first, network attacks are identified and the performance of the algorithms is compared. Dimension reduction focuses on using the data in the obtained KDT 99 file to select the properties used to identify the type of attack. Dimension reduction is carried out first according to the best first search method in attributes 41 to 14 and 7, then the two-class algorithm is used.

In [9], Mohammad Saiful Islam Mamun and A.F.M. Sultanul Kabir proposed an intrusion detection system based on a hierarchical architectural design that meets the current restrictions and requirements of the ad hoc wireless sensor network. In their proposed intrusion detection system architecture, they followed a clustering mechanism to build a four-tier hierarchical network that increases network scalability to a large geographic area and uses both detection techniques. anomalies and abuse for intruder detection. They introduced the intrusion response with the GSM cell concept, as well as a policy-based detection mechanism for the intrusion detection architecture.

Received: 03 Dec 2019 | Revised: 21 Jan 2020 | Accepted: 18 Feb 2020

III. PROPOSED WORK

The dataset being used is the KDD99 dataset for network intrusion. It's a famous dataset being used by many researchers for the purpose of intrusion detection applying various learnings. The dataset contains many attack types like the DOS, U2R, R2L, Probe and normal (no attack).

Table 1.KDD99 Dataset

Categories of Attack	Attack name	Number of instances
DOS	SMURF	2807886
	NEPTUNE	1072017
	Back	2203
	POD	264
	Teardrop	979
U2R	Buffer overflow	30
	Load Module	9
	PERL	3
	Rootkit	10
R2L	FTP Write	8
	Guess Passwd	53
	IMAP	12
	MulitHop	7
	PHF	4
	SPY	2
	Warez client	1020
	Warez Master	20
PROBE	IPSWEEP	12481
	NMAP	2316
	PORTSWEEP	10413
	SATAN	15892
normal		972781

The dataset taken from the Kdd99 is a huge dataset and the one that used in research is under the folder corrected. The goal is to not only to find the best algorithm suited for the intrusion detection but also to implement it using the programming language R. The first process of applying learning is to pre-process the data. First,convert the files to CSV format. Then we have to remove the redundant rows from the dataset. Then next step is to see whether there are any missing values and then to remove those corresponding rows too. The next process is to use this dataset and put it across various machine-learning algorithms that might give good results by correctly classifying the instances. The tool that used is the Weka. Weka is an open source Java platform for processing, classifying, clustering and visualization. It is considered as one of the better data mining tools and therefore it can be used. Steps involved in using Weka are

1. Importing the dataset
2. Classifying and choosing the algorithm.
3. Using the 10 Fold method
4. Again testing using the Percentage Split (70%)
5. Checking the Correctly Classified Instance Percentage.

The values are observed and tabulated shown in the Results below. It can clearly be seen from the above give values that the best algorithm that can be used for the network intrusion detection is the Random Forest. Random Forest algorithm is a classification algorithm based on ensemble learning. It works by building multiple decision trees

at training and the developed decision trees forms the output function. The algorithm is elected and now the implementation using the R programming language is to be done. The drawback of this intensive and the accurate algorithm in this case is that the computation time is very high. To lessen the computational time,use feature selection algorithm. The feature selection algorithm used is the InfoGainAttribute using the Weka tool. Information gain is a feature selection method uses entropy of the class variable and then assess the feature. Using this method,expect no considerable drop in accuracy in terms of the percentage of correctly classified classes and a great reduction in time taken to detect intrusion in the network. This makes the system for intrusion detection more efficient. The attributes elected are used in the R program to predict the error rate and in future to predict if a network is bad or normal. This program when developed fully will act as a filter to determine if a network is secure and will continuously learn from its own series of data making it better and stronger with each type of attack.

IV. IMPLEMENTATION

4.1 Modules

The project mainly consists of 2 main modules: i) The Algorithm testing using WEKA ii) Feature/Attribute selection using the InfoGain Algorithm iii) The implementation on the Algorithm using R.

```
> training <- data1[inTrain,]
> testing <- data1[-inTrain,]
> dim <-nrow (training)
> dim(training)
[1] 38653     6
> #data2 <- data.frame(SrvRerrorRate=0,RerrorRate=0,Flag="SF",DstHostRerrorRate=0,LoggedIn=0,ProtocolType="udp")
> output.forest <- randomForest(Attack ~ ., data = training)
> print(output.forest)

Call:
 randomForest(formula = Attack ~ ., data = training)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 1.17%
```

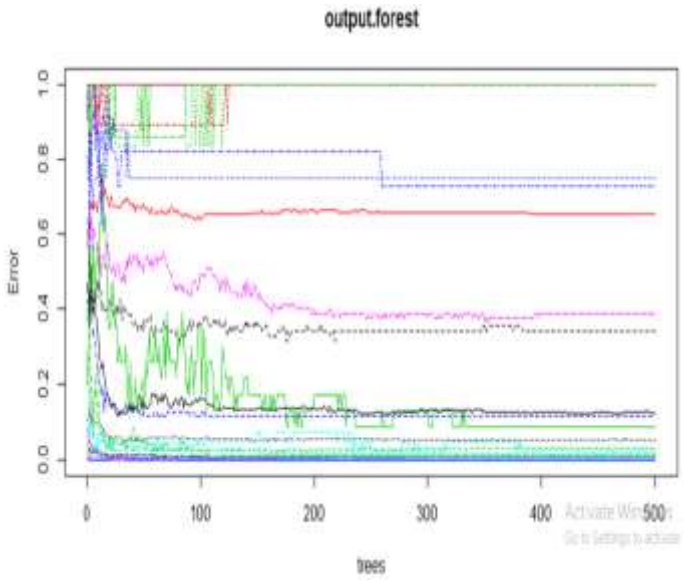

Figure 1:R Output

V. CONCLUSION

The Random Forest, KNN and the Decision Table give high accuracy but the accuracy achieved in case of the Random Forest is the highest. Therefore, to further project choose the Random Forest algorithm, use it practically to find out the error rate using the R Programming language.

Table2:classification based on ML algorithm

	Cross Validation (10 Folds)		Percentage Split (70%)	
	Correctly Classified Instances	Incorrectly Classified Instances	Correctly Classified Instances	Incorrectly Classified Instances
Naïve Bayes	58866 (76.16%)	18425 (23.84%)	17987 (77.57%)	5200(22.43%)
Decision Table	76141 (98.51%)	1150 (1.49%)	22868 (98.51%)	319 (1.36%)
KNN	76474 (98.94%)	817 (1.06%)	22944 (98.95%)	243 (1.05%)
Random Forest	76829 (99.40%)	462 (0.60%)	23040 (99.37%)	147 (0.63%)
AdaBoost M1	68113 (88.13%)	9178 (11.87%)	20441 (88.16%)	2746 (11.84%)

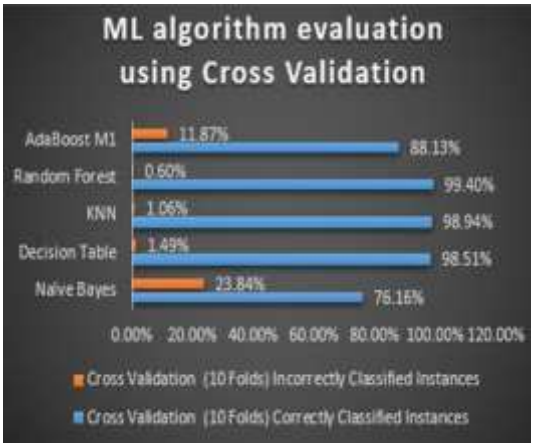

Figure 2:ML algorithm evaluation using Cross Validation

The graph in Figure 1 shows that the accuracy for the Random Forest algorithm is 99.40%, the highest. It has been evaluated using the 10 Fold Cross Validation method. In this method the dataset for training is divided into 10 parts and for each part the other 9 parts acts as the training set and the 1 part as the testing set. Average of all such sets gives the 10 Fold validation.

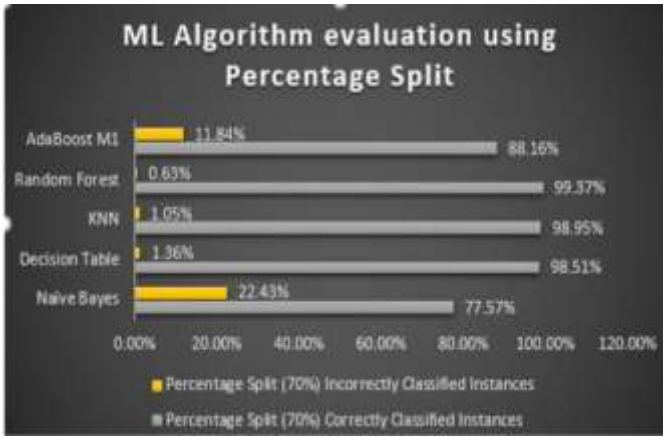

Figure 3: ML algorithm evaluation using Percentage Split

The graph in Figure 2 shows that the accuracy for the Random Forest algorithm is 99.37%, the highest. It has been evaluated using the Percentage split (70) method. In this method the dataset for training is divided 70 percent for training purpose and the other 30 percent for the testing purpose. Before directly using all the 41 attributes onto the R do feature selection to select the attributes using the InfoGain method to reduce the time of computation whereas the accuracy is not much compromised. The attributes elected are

- SrcBytes - number of bytes of data from source to destination
- DstBytes - number of bytes of data from destination to source
- DstHostSameSrvRate – Destination host same server rate.

• Number - The no. of connections with a similar host of the present association over the most recent two seconds

• DstHostDiffSrvRate - Destination host different server rate.

The number of attributes reduce from 41 to 5 using the InfoGain method which makes it much faster in terms of execution time. The R language is then used to see the OOB Estimate of Error that is retrieved after using the Random Forest machine-learning algorithm. The OOB Estimate of Error was found to be 1.46% when the half of dataset was used in training and the other half in testing the algorithm. For each of the testing dataset need to predict if the connection setup is normal or malicious and even pin point to the particular type of attack (example – Neptune, smurf, Saturn, teardrop, rootkit etc.) This means that we can be 98.54% sure of the prediction value shown by the algorithm running on R.

REFERENCES

[1] Review on Anomaly based Network Intrusion Detection System Rafath Samrin Computer Science and Engineering ISL Engineering College Hyderabad, India D Vasumathi Computer Science and Engineering JNTUH Hyderabad, India.

[2] Method of Intrusion Detection using Deep Neural Network by Jin Kim, Nara Shin, Seung Yeon Jo and Sang Hyun Kim.

[3] Network intrusion detection system based on recursive feature addition and bigram technique.

[4] Adaptive and online network intrusion detection system using clustering and Extreme Learning Machines.

[5] Statistical analysis of CIDDS-001 dataset for Network Intrusion Detection Systems using Distance-based Machine Learning.

[6]Anomaly-based network intrusion detection: Techniques, systems and challenges by P. Garcı´a-Teodoroa, J. Dı´az-Verdejoa, G. Macia´-Ferna´ndeza, E. Va´zquezb.

[7] Machine Learning Techniques for Intrusion Detection Mahdi Zamani and Mahnush Movahedi.

[8] Network Intrusion Detection System Using Reduced Dimensionality

[9] Hierarchical Design Based Intrusion Detection System For Wireless Ad Hoc Sensor Network.

Received: 03 Dec 2019 | Revised: 21 Jan 2020 | Accepted: 18 Feb 2020