International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-4, November 2019

Random Forest based Hybrid Model for Intrusion Detection System

Harshal N. Datir, Pradip M. Jawandhiya

Abstract- Malicious threats are better known by their work of damages. This damages are not just limited to the system, but it might lead to significant information damage too. Along with this, threats are also responsible for financial loss. As technology increases, Types and attacks of threats also increases. Though the research community investigated a number of cyber attack prevention models it is challenging to detect the threat and preventing them from data, for the industries. Detection of the attacks with IDS is common and popular in organizations . Now a days data mining and hybrid approaches are getting priority combine with IDS in the area of anomalies and attack detection. In this paper, we focus on the designing a tool based on signature approach and the random forest algorithm for intrusion detection that offers data security and protection. Both algorithm works individually for IDS system but signature base algorithm have some limitations of known database requirement.

In our research paper, we proposed a Hybrid intrusion detection model which allows us to double filtration of the intrusions in the application with implementation of combine signature and behavior based algorithm in one system. This paper addresses the various kinds of feature and the behavior of the threat and their different functioning further intrusion detection hybrid model is the extension for the simple individual model who work on either behavior or on signature.

Keywords- Malware ; behavior and signature ; Packet-data; Network data security ; Detector.

I. INTRODUCTION

In the era of www(world wide web), within a fraction of the second attacker send the malicious function and give the financial as well as data loss to the industries, banking and the organizations without keeping the track of footprints. deploying highly effective IDS system is an extremely challenging task although most of the existing system provides security against the threats, however, research is going on a hybrid approach to obtain the best data security in packet data as the event of network. We contribute in research by implementing a hybrid approach of intrusion detection in the network data.

There are different approaches had been used in the network management and data management field to find-out the behavior and signature of the malicious functions in the data.

Though we select the random forest algorithm for our approach because of its strength like it run efficiently on large dataset with more accuracy.

In this approach we will filter and process the overall application for the intrusion testing through the detection of signature in the first test, this gives a first result of filtered IDS and in next stage classification using the random forest tree algorithm where we have new database which is miss by the signature based algorithm in first testing. Furthermore, we will see the both approaches. In paper [1][2][3] researchers introduces data mining classisication approach, fast approach and effective approach towards the IDS system. Each method has its own feature and hence only any individual method is not enough for the detection of intrusion present in data packets on the network or data in the system. This is the reason behind our motivation to work for hybrid method.

In the next section of the literature review, we will see some of them.classification approach using the random forest algorithm is as shown in figure 1.

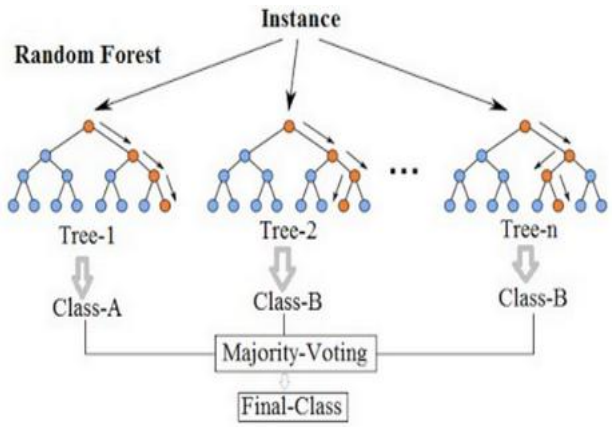

Fig. 1: Generalized Random Forest

II. LITERATURE REVIEW

In paper [1], the data-mining, classification approach was used to detect the malicious behavior. In this case, for training purpose different classification algorithm like Naïve-Bayse, BayseNet,IB1 and the regression algorithms.The Result of this research shows that the regression classification method is best for the resultant. The Results of this research sufficiently said that implemented method is good for malware detection. However the drawback of this method is that it is not compatible with real behavioral detection of packet data.

Manuscript published on November 30, 2019.

* Correspondence Author

Harshal N. Datir*, Research Scholar, Department of Information Technology, SGB Amravati University, Amravati,Maharashtra, (India). Email: harshaldatir809@gmail.com

Pradip M. Jawandhiya, Professor, Department of Computer Science & Engineering, PLIT, Buldhana, Maharshtra, (India).

Email: pmjawandhiya@rediffmail.com

Similar to this, in paper [2,3] data mining approach is used for malware detection in android platform.

In a survey of malware detection technique by I Nwokedi and M Aditya of Purdue University different approaches and methods related to malware detection can be seen. The Aim of the surveyor was clear to provide comprehensive bibliography and successfully they gave their best contribution in the research area [4].They took survey on 45 techniques of malware detection and the classification and provide a comparison among them. They also

gave the reason for why only signature based detection is not sufficient. In research, they found that a base signature is not efficient for peer to peer environment.

Another major contribution of Phyu Thi Htun is noticeable. He has proposed an intrusion detection system which is able to detect 'normal or attacks status' for the research on Intrusion detection he combined the two methods, Machine learning and pattern Recognition. And the results were obtained using the KDD Cup 1999 dataset[5] .The advantage of his model was better accuracy and faster reduction of features. His method had been compared with Random forest,K-neighbor,Naïve Bayes approaches.

Paper [6] and [7] shows the use of the random forest algorithm as a classifier. In the research area reason behind the selection of this algorithm is its accuracy and speed. In paper [6] recognition of the sign language is done with the help of the random forest algorithm and the sencors input.along with this another researcher in [7],M.Belouch and others has been evaluated performance of intrusion detection based on machine learning using apache spark. We can find comparisons of different method like SVM, Neive Bayes,Decision tree with the random forest method. From the result of the paper it is clear that the 'Random forest is best among all.' It gives the accuracy almost 97%.

In [8] M.Hasan et.al; developed a another system of ' Feature selection for intrusion detection using random forest' This system can collect and analyze the information of different areas within the computer. In the result they had proven that Random forest algorithm can select most important and relevant feature useful for classification.Researchers used a permutation importance measure .The Advantage of this model is applicable to maximize the performance rate and to minimize the false positive rate.

P.Agrawal and S.Sharma [9], analyzed the KDD cup dataset attributes.They have been used a random tree for the classification purpose and WEKA tool was selected for the simulation. In the paper, we can observe that 41 attributes are captured from the KDD dataset 15 variants of dataset were generated by forming all combinations of four classes(Basic, Content, Traffic and host).

H.Bahram and N.Nima[10], designed a Intrusion detection stystem using the fuzzy clustering and ABC algorithm. As like other IDS, it has trainning validation and testing phase. In this case attacks were classified into four groups-1) Dannial of Service(DoS),2)Probe(PRB),3)Remote to Local (R2L) and4) User to Root(U2R).

Some researchers have taken a survey over various system of IDS , one of them is [4], taken and refer for our study of IDS system , by which we have enough able to uderstand the technologies of IDS.

This paper[13] proposes a novel intrusion detection system (IDS) that combines different classifier approaches which are based on decision tree and rules-based concepts, namely, REP Tree, JRip algorithm and Forest PA. Specifically, the first and second method take as inputs features of the data set, and classify the network traffic as Attack/Benign.

From the brief survey, we can point-out the importance of intrusion detection system(IDS) . The IDS system can be used to analyze the type of attacks and with this information we can protect our data by changing the security system or by changing the control system. That is IDS has an ability to identify the security incidence.

IDS also helpful for regulatory compliance. IDS is more powerful by its sensors for network host and devices. It can also inspect the data within the network packet and also from the OS.

In the next section we will have a look at the random forest algorithm. First, we should know the function of Random forest algorithm of classification and feature extraction.

III. SYSTEM ARCHITECTURE

Instrusion Detection System(IDS) can be classified in to following groups according to there functions.

- Network base
-Host base
-Application based

Our aim is to implement an IDS system with maximum accuracy and low false rate. Following figure 2 shows the proposed system architecture .

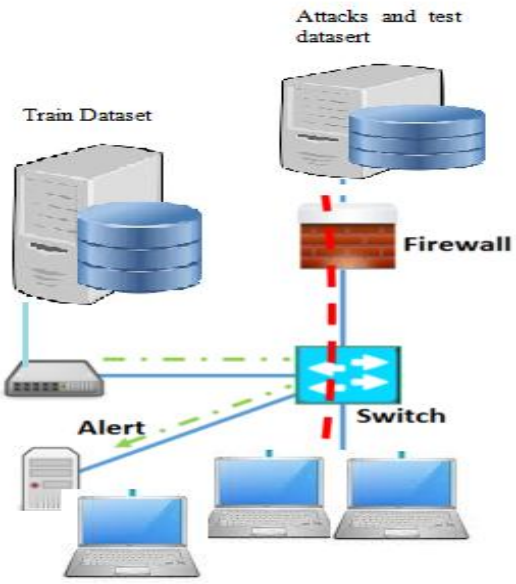

Fig. 2:Proposed System Architecture

The figure 2 gives the brief idea of our system where KDD-cup dataset is splited in two parts. Three forth part is used for trainning and remaning one third part is OOB(out of bag). Trained dataset and test dataset both have to follow the phase of feature extraction.once the feature get extracted and classification get completed system is able to warn us about the appearance of intrusion.

This IDS system will able to detect the signature as well as to detect the behavioral activities which make it Hybrid.The advantage of both method is to improve the performance by determining the malicious behavior as well as signature also.

If the signature doesnot match it will check by behavior and hence security is improved in this system.The signature base detection is also known as misuse detection method. This method is based on the pattern matching algorithm . In this case it is applicable to compare the data/ network packets against known attacks signature knowlwdge base.

If the signature get match it will move to next step or the connections cut down from the source IP, so packet is droped and alarm notifies the administrator.

Once the signature is match ,control will pass to the detection and if any mismatch occurs then terminate the process.

step 1:In the first step the unprocessed data is taken and data set process for feature extraction pre-processing techniques. A number of data sets are trained and rest are used for testing in the step 2, after receiving the dataset it is processes for training data. Next is process for classification , in our Step 4:Data set processed is sent to different classifiers and accuracy is calculated. Data sets are given all the five classifiers approach we have chosen a random forest algorithm in which number of sub trees are used for finding the classes. Final stage of outcome is based on the voting amoung the classes for the best and accuracy is obtain by calculating the different dataset . In our approach based on the test data input best result obtain with maximum accuracy and low false rate.

When we look at the malware , malware malicious function writers do not create new malwares always, Instead of this they prefer to recycling the existing malwares by addition and modification of new signatures. Signature based IDS used a database of known vulnerabilities or known attack pattern.Due to this non existence of new signature in knowledge base of antivirus , it does not get detected easily and may cause data and financial loss. To detect such hidden malwares behavior analysis is helpful.Thus the main function of IDS is to examine the vulnerabilities in the system. However, IDS can only detect the attack but can not prevent them and hence unblocked attacks can consume the resources on the system and ultimately causes it to crash. While configuring the IDS, if we instruct the firewall to deny all the traffic from "attacking" networks, the hackers can easily exploit it.

For the intrusion detection system we can say that , IDS is a software application which can only monitor the system network.Security engineers put the best ways of development and security about the building of IDS systems. IDS systems are compatible to false positive rate and false negative rate.False positive is a alert for harmless activities while false negative is when an attacker is actively attacking the network and IDS triggers when an event exceeds the threshold.

Positive rate or sencitivity is used to measure the proportion of positive that correctly identified. We can measure the sencitivity by following formula-

$$\text{Sencitivity} = \frac{TP}{TP + FN}$$

Where,

FN is the false negative –represent the number of misclassified set.

Similarly we can measure true negative that is specificity by using following formula-

$$\text{Specificity} = \frac{TN}{TN + FP}$$

Where, the TN is the true negative that is record classified as normal.

A. Random forest algorithm flowchart

This framework gives good view of random forest algorithm and its working.As we discussed about RFA (random forest algorithm) in the previous section- It can deal with the huge dataset , And hence we have a large scope to running as many trees as we would like to run . In this algorithm,there is no need of cross validation.and it has two methods for replacing the missing number.

Random forest is used as a main classification framework. As shown in figure 1 random forest consist of many tree structured classifiers. In the proposed method decision tree and feature extraction is adopted as the base tree structure of the random forest.All the other tree have the same structure in the forest tree and random vector is used to generate the bootstrap sample as training set. At the end on the basis of voting most popular class get selected and classified. Figure 3 shows the generalized flow of the random forest algorithm.

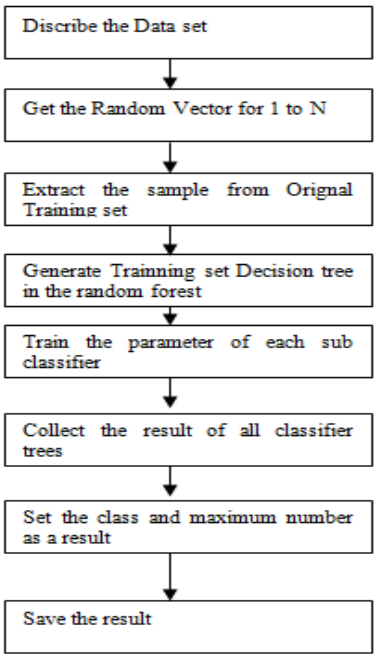

Fig. 3: Flow of Random Forest

The function and accuracy depend on the training dataset. When the training set for the current tree is down by sampling the replacement ,about 1/3 of cases are the left-out of the sample, this left-out samples are said to be Out Of Bag (OOB) dataset and it will use in the classification phase for unbiased estimation error. It is also used to estimate the variable importance also. Two common variable importance measure are- Gini imortance index and Permutation importance index. A Specific feature is used to split in the Gini index reduction. In paper[8], is is explain the variable importance. According to paper, The overall variable importance for a feature in the forest is defined as the summation or the average of its importance value among all trees in the forest.

In our approach aim of random forest algorithm is feature ranking, Selection and the filtering of the pre-processing state.

B. Double Filtration scheme

In our approach aim of random forest algorithm is feature ranking, Selection and the filtering of the pre-processing state.

Redundant and irrelevant features in data have caused a long-term problem in network traffic classification. These features not only slow down the process of classification but also prevent a classifier from making accurate decisions In first stage of our work, we propose a signatature detection and matching method to block the intrusions if foud. However, some unknown intrusions may pass the through the algorithm and can attack on system data sequrity hence it is not enough for us with known information based algorithm that analytically selects the optimal IDS with known databse. Hence we have added a method with can take the input intrusions and can create there signature also . This new signature are then get store in to the database.This new signature are created by RF algorithm. Hence if in next flow of first stage it get detected and protect the system. In such way two stage filter is applied and that's the teason to increase in sequrity as well as database improvement.Its effectiveness is evaluated in the cases of network intrusion detection. Although selecting important and relevant features from input data builds a fast and accurate IDS and conversely, selecting irrelevant features cause enormous problems in network traffic classification. The successful combination of feature selection and intrusion detection improves the detection and accuracy rates, decrease the false positive rate and the complexity of computation in the system

C. KDD-Cup dataset and Proposed Model

KDD –cup dataset is a standard dataset and it is a subset of the Defense advanced Reserch Project Agency (DARPA) prepared by Sal Stofo and Wnke Lee[11]. This dataset is available and used for research on 'Intrusion detection systems'.

This approach is a clear extension to the Hybrid intrusion detection system [3]. It will monitor the behavior as well as detect the signature. If any attack is found, automatically drops the packet,otherwise it will process the packet data by considering the packet data is normal.

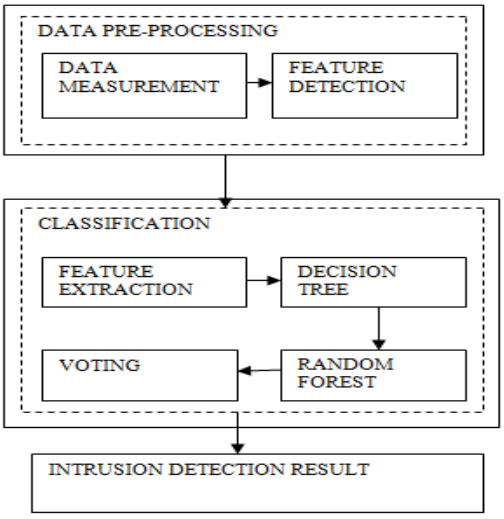

Fig. 4: Block Diagram -Intrusion Detection Based on Random Forest

In this approach database is divided into parts as,for AI tools, Training dataset and testing data-sets. In the training phase label and unlabeled data will go for training and will save the training results. In the classifier phase, both data gives the result of the majority voting system. Those data which do not get the majority will drop in the process.

In this model we are using the KDD-cup dataset. This input is processed for the feature detection. The aim of this step is to obtain the specified features from all the data.In the feature extraction process efficiently describe sets of features are extracted to match our requirement of the base classifier in the random forest.

Table I: Algorithm and Accuracy Rate

Algorithm	Accuracy
Logistic regration	91.56
SVM	78.84
Naïve Bayes	90.68
Random forest	88.65
GB	91.13

Table 1 is related to the different algorithms used in past research for building IDS tools of intrusion detections. Mostly signature base and behavior based classifer are designed for the classification purpose. Above table is taken from[12] where the cyber security was aim to design the IDS.

D. Contribution In Research:

This paper contribute in research area by-

our approach is to implement a 'Hybrid Detector for Intrusion data'. This system model is fully automated with classifier and features detector. Proposed model is able to build several decision tree and integrate them together to get more accurate and stable result.

Unknown signatures of code also be detected by the model and will register new signature in the knowledgebase system for the further detection process. This is a extension to the previous work.

We have feature detection method and pattern matching algorithm combine to overcome the drawback of individual method of behavior and signature based models. This hybrid module maximize accuracy rate as compared to individuals.

IV. CONCLUSION:

This research paper proposed hybrid model to detect malicious activities based on pattern matching and behavior based analysis of network packets. The performance of model is evaluated using standard dataset and hybrid model gives better detection of novel malicious activities. Our main focus on providing a compatible solution for enterprises and MNC and overcome the limitation of existing models. This proposed IDS system will able to respond after detecting an attack, and the response can be either passive or active . passive responces will belongs to notification and the active response for the harmful attacks that should be block.

REFERENCES

1. N.Monire et.al; "A Data Mining Classification Approach for behavioral Malware Detection" Journal of Computer Networks and Communications, volume 2016.
2. C. Hongmei et.al; "A fast approach toward the android malware detection" Springer International Publishing Switzerland 2015
3. G.V.Nadiammai and M.Hemlatha "Effective approach toword intrusion detection system using datamining technique" Egyptian Informatics Journal,2014
4. I Nwokedi and M Aditya "Survey of Malware Detection Techniques" Purdue University ,2007.
5. P.T.Htun and K.T.Khaing 'Anomaly Intrusion Detection System Using Random Forests and K-nearest neighbor' International Journal of P2P Network Trends and Technology (IJPTT)-Vol-3 Issue 1 (Feb-2013)
6. R.Su and et al. "Random forest based recognition of isolated sign language subword using data from accelerometers and surface electromyographic sensors" Sensors 2016 – volume 16 issue(1).
7. M.Belouch and et al. "Performance Evaluation of Intrusion Detection Based on Machine Learning Using Apache Spark' Procedia computer science 127 (2018), Science direct.com.
8. M.Hasan and et al. "Feature Seletion for intrusion selection Using Random forest" Journal of information sequrity –April 2016
9. P.Agrawal and S.Sharma "Analysis of KDD dataset attributes- class wise for intrusion detection" Procedia computer Science- vol-57-2015
10. H.Bahram and N.Nima "Intrusion detection for cloud computing using Neural networks and artificial bee colony optimization algorithm" ICT express 5(2019
11. KDD's 99 dataset kdd.ics.uci.edu/databases
12. G.Gupta and K. Manish " A framework for fast and effieient cyber security network intrusion detection using appachi spark" Procedia computer Science- 96(2016)
13. Ahmim, Ahmed, Leandros Maglaras, Mohamed Amine Ferrag, Makhlouf Derdour, and Helge Janicke."A novel hierarchical intrusion detection system based on decision tree and rules-based models." In 2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS), pp. 228-233.IEEE, 2019.
14. Nanda, Nilesh B., and Ajay Parikh. "Hybrid Approach for Network Intrusion Detection System Using Random Forest Classifier and Rough Set Theory for Rules Generation." In International Conference on Advanced Informatics for Computing Research, pp. 274-287. Springer, Singapore, 2019.

AUTHORS PROFILE

Harshal N. Datir, He is Research Scholar, SGB Amravati University, Amravati (Maharashtra). Having completed Masters in Engineering (Information Technology). He is having more than four years of research experience. His area of interest is Machine Learning and Data Mining & Analytics. Working as a faculty in Department of Information Technology, Sipna College of Engineering and Technology, Amravati, Executive Member IETE, Amravati Centre, Amravati, He has around 24 publications in national and international journals. He is Lifetime Associate member of Institute of Electronics & Telecommunication Engineering (IETE). Life Member of Computer Society of India (CSI) and Indian Society of Technical Education (ISTE).

Pradip M. Jawandhiya, He is working as Principal & Profesor in Department of Computer Scienec & Engineering, Pankaj Laddhad Institute of Technology and Management Studies , Buldhana. (Maharashtra). Having completed Doctrate in Computer Science and Engineering from SGB, Amravati University, Amravati, MBA in Human Resource Management. He is member of board of studies, SGBAU, Amravati. He has more than 10 years of research experience and 23 years of teaching experience. He has presented and published many research papers in national and international journals. His area of specialization is MANET and Artificial Intelligence. He is life member of I.S.T.E. Computer Society of India (CSI) and Fellow of I.E.T.E., Member of IEEE, Member of IACSIT. He is vice-chairman of IETE Amravati Centre, Amravati.