International Journal of Current Trends in Engineering & Technology
Received 4 February 2019, Received in revised form, 1 March 2019, Accepted 18 March 2019, Available online 7 April 2019
www.ijctet.org, ISSN: 2395-3152
Volume: 05, Issue: 02 (March- April, 2019)

Signature Based Intrusion Detection Systems by Using Genetic and Random Forest Algorithm

JYOTSNA PUROHIT*, JAYSHREE BOADHH, AKRATI SHRIVASTAVA
Department of Computer Science and Engineering, Mittal group of Institution, India
*Jyotsnapurohit28@gmail.com
+91-7222909350

Abstract: *To improve network security different steps has been taken as size and importance of the network has increases day by day. Then chances of a network attacks increase Network is mainly attacked by some intrusions that are identified by network intrusion detection system. This paper works to develop an intrusion detection system which utilizes the identity and signature of the intrusion for identifying different kinds of intrusions. Whole work was divided into two modules first was feature selection by using genetic algorithm where good set of feature were select. Here random forest tree algorithm was used for finding the patterns in the input data. In this work use of Gini index was done for the decision tree construction in recursive manner. Experiment was done on DSL-KDD dataset which was real. Comparison was done with latest RNN model in [11]. Result obtained after analyzing this system is quite good enough that precision, recall and accuracy values were improved.*

Index Terms- Clustering, Gini-Index, Intrusion Detection, Random Forest, Pattern generation.

I. Introduction

Providing network security for different web services on the internet, different network infrastructures, communications network many steps has been taken like encryption, firewall, and virtual private network etc. network Intrusion detection system is a major step among those. Intrusion detection field emerges from last few years and developed a lot which utilizes the collected information from different type of intrusion attacks and on the basis of those different commercial and open source software products come into existence to harden your network to improve network security of the different communication, service providing networks. As the number of network users and machine are increasing day by day to provide different kind of services and easiness for the smoothness of the world. But some unauthorized users or activities from different types of attackers which may internal attackers or external attackers in order to harm the running system, which are known as hackers or intruders, come into existence. The main motive of such kind of hacker and intruders is to bring down bulky networks and web services. Due to increase in interest of network security of different types of attacks, many researchers have involved their interest in their field and wide variety of protocols as well as algorithm has been developed by them, In order to provide secure services to the end users. Among different type of attack intrusions is a type of attack that develop a commercial interest. Intrusion detection system is introduced for the protection from intrusion attacks. From the above discussion we can conclude the main aim of the network Intrusion detection system is to detect all possible intrusion which perform malicious activity, computer attack, spread of viruses, computer misuse, etc. so a network intrusion detection system not only analyses different data packets but also monitor them that travel over the internet for such kind of malicious activity. So the smooth running of overall network different server has to settle on the whole network which act as network intrusion detection system that monitor all the packets movements and identify their behavior with the malicious activities. One more kind of network Intrusion detection system is developed that can be installed in a centralized server which also work in the similar fashion of analyzing and monitoring the different packet data units for their network intrusion behavior. Network Intrusion detection system can be developed by two different approaches which can be named as signature based and anomaly based. In case of signature-based Network Intrusion detection system it develops a collection of security threat signature. So according to the profile of each threat the data stream of different packets in the network are identified and the most matching profile is assigned to that particular packets. If the profile is malicious then that data packet comes under intrusion and it has to remove from the network in order to stop his unfair activities.

II. Related Work

Yogitha et. al. [1] Offered interruption discovery framework with Support Vector Machine (SVM). Affirmation is finished by coordinating explores on NSL-KDD Cup'99 data collection which is reformer type of KDD Cup'99 data index. By utilizing this NSLKDD Cup'99 data collection they have condensed wide time obligatory to shape SVM exemplary by achievement proper pre-training on data collection. In this association SVM made clustering of data. By obligation

International Journal of Current Trends in Engineering & Technology
Received 4 February 2019, Received in revised form, 1 March 2019, Accepted 18 March 2019, Available online 7 April 2019
www.ijctet.org, ISSN: 2395-3152
Volume: 05, Issue: 02 (March- April, 2019)

appropriate part accumulation assault location rate is opened up and false positive rate (FPT) is lessened. In this proposed work author has utilized Gaussian Circular Basis.

A.R. Jakhale, et. al [2] In this work the author portrays an anomaly discovery framework and its two stages particularly training and testing. The slipping window and bunching is accustomed to nursing the network movement by mining the repetitive examples utilizing calculations. The calculations are so genuine and utilized as a part of constant observing. The normal multi-design catching calculation has high location rate. At long last, increase the identification rate and reduced the false alert rate.

Jiefei, Lobo and Russo [3] explores the event of Multi-way steered attack where an assault is divided and sent over different courses to endeavor to trick an IDS framework. This is influenced conceivable due to multi way TCP (MPTCP) which enables transmissions to course finished numerous ways between a source and target

Barolli et al [4] researches the utilization of IDS utilizing neural network for giving IDS arrangement in a Tor (The Onion Router) organize. Tests did utilize a Tor server and customer with back engendering NN to reproduce exchanges over the Tor organize while catching for examination. The framework proposed is a prepared ANN with information caught from Wireshark, at that point the server and customer information are analyzed, contrasts will recognize an interruption or misuse. The outcomes from testing were fruitful in giving viable exactness when assessed in the test condition.

Chuan Long [5] In this paper, author investigate how to display an interruption recognition framework in light of profound learning, and this work propose a profound learning approach for intrusion identification utilizing recurrent neural networks (RNN-IDS). Additionally, this work examines the execution of the model in paired classification and multiclass classification, and the quantity of neurons and distinctive learning rate impacts on the execution of the proposed display. This work contrasts it and those of J48, artificial neural network, arbitrary woodland, bolster vector machine, and other machine learning strategies proposed by past analysts on the benchmark data index.

III. Proposed Solution

Whole work was done by two step first was feature selection by using genetic algorithm where good set of feature were select and second was Random Forest Gini Index Based Intrusion Detection which was further divide into two stage, first is training shown in fig 1 and 2, second is testing. In order to make the analysis better feature vector of each class is prepare for training the neural network of the current updated dataset sessions.

Dataset Preprocessing Module In order to increase the efficiency of the work dataset should be pre-process as the preprocessing of Raw Dataset Instead of direct input of raw dataset to selected classifier, raw dataset is preprocessed in different ways to overcome different issues like training overhead, classifier confusion, false alarms and detection rate ratios. Separating feature space from one another is very necessary and arrange in vector. Let us consider single vector Ds of the dataset and n number of events load in the Vs vector.

{0,tcp,ftp_data,SF,491,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,2,0.00,0.00,0.00,0.00,1.00,0.00,0.00,150,25,0.17,0.03,0.17,0.00,0.00,0.00,0.05,0.00,normal,20}

In above vector presence of comma ‘,’ and d**iscarding symbolic characters that are of** three kind s of symbolic features (tcp, ftp_data and SF etc.) in feature space of 41 features. As symbolic values are not of interest to our research, these three feature vectors are discarded to get the feature space this is shown in fig 1.

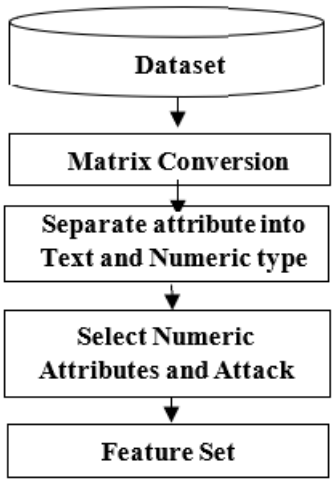

Fig. 1 Dataset Preprocessing and Feature Selection

So after the preprocessing the obtain vector is where all element is required for dataset analysis.

Pv[] ←Pre-Process (Vs)

International Journal of Current Trends in Engineering & Technology
Received 4 February 2019, Received in revised form, 1 March 2019, Accepted 18 March 2019, Available online 7 April 2019
www.ijctet.org, ISSN: 2395-3152
Volume: 05, Issue: 02 (March- April, 2019)

{491,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,2,0.00,0.00,0.00,0.00,1.00,0.00,0.00,150,25,0.17,0.03,0.17,0.00,0.00,0.00,0.05,0.00,normal}

Features Selection

Feature selection is an important factor in Network intrusion detection system for initial training. Since, the large numbers of features that can be monitored taking into account the large variety of possible values especially for continuous feature even for a small network. So genetic algorithm was used for finding the best set of feature set for the Random forest tree training.

Generate Population

Here assume some cluster set that are the combination of different feature of the 42 available features in a session. This is generated by the random function which select fix number of cluster for the centroid.

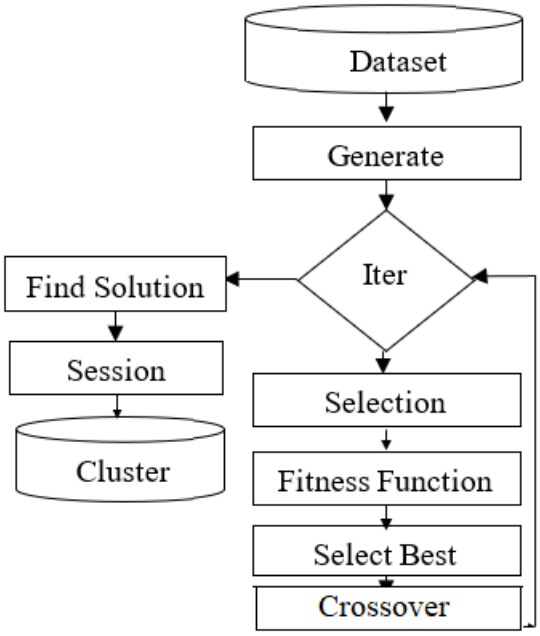

Fig. 1 Proposed work first module.

This can be understood as let the number of class of intrusion n and number of chromosome are N then one of the possible solution is {C1, C2, ….Cn}. In the similar fashion other possible solutions are prepared.

Fitness Function

Fitness of the solutions are calculated on the basis of feature obtained from the dataset after preprocessing. In case of feature distance between the session from the centroid are obtained by calculating the difference values between the sessions. Similarity is obtained by the distance function.

Ellucian Distance: This can be understood as Let X be a session matrix and Y be the other session matrix. Then distance between them is calculated by:

$$D = \sqrt{sum((X - Y)^2)}$$

Here both X, Y are vectors of elements from dataset feature. Base on the minimum distance value between session and dataset output is matching with target class of attack. This is considered as final rank of the probable solution.

Select Best solution

Main motive of this step is to find best solution from the generated population. Here each possible solution is evaluated for finding the distance from each centroid so that image closer to the centroid are cluster together. Then calculate the fitness value which give overall rank of the possible solution.

Cross-Over

Top possible solution after sorting will act as the best for other possible solutions. Now selected solution will modify other possible solution by replacing fix number of centroid as present in best solution. By this all possible solution will learn from best solution. This difference modifies the existing solution according to the following expression

$$X_{new,i} = \text{Crossover}\ (X_{best,i},\ X_{,i})$$

Accept Xnew if it gives a better function value. Once this cycle is over then check for the maximum iteration for the genetic if iteration not reach to the maximum value then GOTO step of finding fitness function and rest of steps in regular fashion of genetic algorithm stop learning and the best solution from the available population is consider as the final centroid of the work. Now session is cluster as per centroid.

Random Forest

The random forests [7] is an ensemble of unpruned classification or regression trees. Random forest generates many classification trees. Each tree is constructed by a different bootstrap sample from the original data using a tree classification algorithm. After the forest is formed, a new object that needs to be classified is put down each of the tree in the forest for classification. Each tree gives a vote that

International Journal of Current Trends in Engineering & Technology
Received 4 February 2019, Received in revised form, 1 March 2019, Accepted 18 March 2019, Available online 7 April 2019
www.ijctet.org, ISSN: 2395-3152
Volume: 05, Issue: 02 (March- April, 2019)

indicates the tree's decision about the class of the object. The forest chooses the class with the most votes for the object.

Feature Selection Randomly: As random forest has n number of tree and their feature set is different. In this work for each tree in forest random features were select. This can set of column like [5, 10, 12, 17, 19, 20, 27, 28, 29, 33], OR [2, 6, 8, 11, 15, 22, 17, 18, 29, 33], etc. Based on this feature set random tree was built in next step.

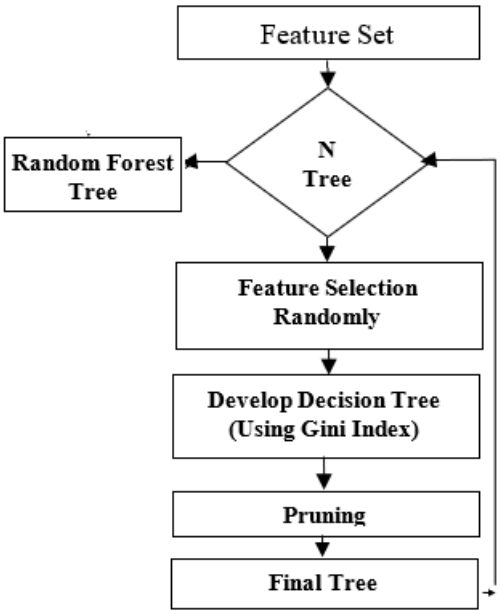

Fig. 2 Block diagram of proposed work.

Develop Decision Tree: The process of tree building begins by splitting the root node into two child nodes. CART computes the best split by considering all probable splits for each independent or explanatory variable. The best split is obtained when the impurity function, which exists between the parent node and two child nodes, is minimized. The best split equation is given as:

$$GI = \sum_{C\epsilon\{Y,N\}} -P_{c,L} log(P_{c,L}) + \sum_{C\epsilon\{Y,N\}} -P_{c,R} log(P_{c,R})$$

Where $P_{C,L}$ is proportion of total number of elements move towards Left side of tree to the total number of elements in the input dataset. In similar fashion $P_{C,R}$ is proportion of total number of elements move towards Right side of tree to the total number of elements in the input dataset. Where C is number of class element which need to be finally classified. In this way one value of the Gini index obtain for the feature set column value. In similar fashion other values of the Gini index were obtained from the other set of feature column. At last the highest gain or Gini Index value is consider as the final node value for the partition.

Pruning: For a complex or larger tree grown on the initial step of CART, though the prediction of data is described correctly, the prediction accuracy of the tree is low for new samples. Therefore, there is a need to build a tree with better accuracy and predictive ability. Pruning develops an optimal tree, by shedding off the branches of the large tree. The pruning procedure develops a sequence of smaller trees and computes cost complexity for each tree. Based on the cost-complexity parameter, the pruning procedure determines the optimal tree with high accuracy. The cost-complexity parameter R is set forth as a linear combination of tree complexity and cost associated with the tree. Complexity is given by the following equation:

$$Cn = \frac{Misclassified\_Elements}{Total\_Elements}$$

Where n is number of node in a tree and elements are number of session classified by the node. Misclassified means elements (session) which are incorrectly classified in the tree.

Testing of Random Forest

As for testing the trained network dataset is again required with different vector, of different or may be of same pattern of the classes. Here it also needs to make the feature vector of all the vector for testing from the trained random forest pattern, but only numeric feature is collected in the Fv then as per training the values of the network is obtained that the input vector is belong to which class. Here feature is pass as per random tree feature set. Each tree gives its own output and majority of tree output is consider as final class of the input session. It may be normal or intrusion.

IV. EXPERIMENT AND RESULTS

Data Set For the evaluation of the whole work the dataset is NSL KDD [12] about which previous section has already explained and the collection of the all evaluating vectors look like. Where numeric terms are used for feature learning and at the end of each vector it has the corresponding class. The pre-processing step and its requirement have been already explained.

International Journal of Current Trends in Engineering & Technology
Received 4 February 2019, Received in revised form, 1 March 2019, Accepted 18 March 2019, Available online 7 April 2019
www.ijctet.org, ISSN: 2395-3152
Volume: 05, Issue: 02 (March- April, 2019)

Evaluation Parameter

To test our result this work use following measures the accuracy of the, that is to say Precision, Recall and F-score. These parameters are depending on the TP, TN, FP and FN.

$$Precision = \frac{True\_Positive}{True\_Positive + False\_Positive}$$

$$Recall = \frac{True\_Positive}{True\_Positive + False\_Negative}$$

$$F\_Score = \frac{2 * Precision * Recall}{Precision + Recall}$$

In order to make the better evaluation for this work one more parameter has introduced that is accuracy of the class of the intrusion. Accuracy of the work is calculated by:

Accuracy = (true positives +false negatives)/(Total_Normal + Total_Intrusion)

Table 1. Precision value comparison of RNN and RFGIID at different Dataset Sizes.

Data-Set Size	Precision Value Comparison	
	RNN	RFGIID
3000	0.879694	1
6000	0.877468	0.998433
8000	0.874707	0.99718

From Table1, it is obtained that with the increase in dataset size precision value rate increase. As number of patterns are more in the dataset so results are more accurate. Here it was shown that use of random forest tree with Gini index increase the precision value. From table 2 it is obtained that with the increase in dataset size recall value rate increase. As number of patterns are more in the dataset so results are more accurate. Here it was shown that use of random forest tree with Gini index increase the recall value.

Table 2 Recall value comparison of RNN and RFGIID at different Dataset Sizes.

Data-Set Size	Recall Value Comparison	
	RNN	RFGIID
3000	0.978754	0.994934
6000	0.977995	0.986989
8000	0.976427	0.9858

. Table3 F-Measure value comparison of RNN and RFGIID at different Dataset Sizes.

Data-Set Size	F-Measure Value Comparison	
	RNN	RFGIID
3000	0.926584	0.9974
6000	0.925	0.992678
8000	0.922	0.99148

From table 3 it is obtained that use of random forest tree with Gini index in proposed work has high F-measure value as compared to previous work. Here it was shown that use of new approach of neural network training reduce the execution time as compared to RNN used in previous method.

Table 4. Execution time value comparison of RNN and RFGIID at different Dataset Sizes

Data-Set Size	Training Time (second)	
	RNN	FGIID
3000	19.8547	18.6253
6000	40.2612	23.6892
8000	54.0972	

From table 4 it is obtained that with the increase in dataset size execution time value increase. Here it was shown that use of new approach of random forest tree with Gini index for training reduce the execution time as compared to RNN used in previous method.

Table 5 Execution time value comparison of RNN and RFGIID at different Dataset Sizes.

Data-Set Size	Testing Time (second)	
	RNN	RFGIID
3000	38.7396	22.5932
6000	59.391	24.6259
8000	72.1846	27.2998

Table 6. Execution time value comparison of RNN and RFGIID at different Dataset Sizes.

Data-Set Size	Accuracy Value Comparison	
	RNN	RFGIID
3000	0.927	0.99733
6000	0.92433	0.992167
8000	0.9227	0.9908

From table 5 it is obtained that with the increase in dataset size execution time value increase. Here it was shown that use

International Journal of Current Trends in Engineering & Technology
Received 4 February 2019, Received in revised form, 1 March 2019, Accepted 18 March 2019, Available online 7 April 2019
www.ijctet.org, ISSN: 2395-3152
Volume: 05, Issue: 02 (March- April, 2019)

of new approach of random forest tree with Gini index for training reduce the execution time as compared to RNN used in previous method. From table 6 it is obtained that with the increase in dataset size execution time value increase. Here it was shown that use of new approach of random forest tree with Gini index for training reduce the execution time as compared to RNN used in previous method.

CONCLUSION

Network security is one of the most important nonfunctional requirements in a system. Over the years, many software solutions have been developed to enhance network security and this paper provides an efficient system which has been a promising one for detecting intrusion of different kind where, one can get the detail of the class of attack as well. Here combination of genetic algorithm with Random forest increase the efficiency of work. Results shows that all type of attack are accurately identified by the system as the accuracy value is above 99%. In future it needs to be improved by putting data on the unsupervised network, so it automatically updates the new behavior of the intruder. One more issue remain in this work is to use dynamic adaptable technique for learning new type of attack.

REFERENCES

[1]. Yogita B. Bhavsar, Kalyani C. Waghmare "Intrusion Detection System Using Data Mining Technique: Support Vector Machine" 2013 International Journal of Emerging Technology and Advance Engineering volume 3, Issue 3, March 2013.

[2]. A.R. Jakhale, G.A. Patil, "Anomaly Detection System by Mining Frequent Pattern using Data Mining Algorithm from Network Flow", International Journal of Engineering Research and Technology, Vol. 3, No.1, January 2014, ISSN. 2278-0181.

[3]. Aljurayban, N.S.; Emam, A. (21-23 March 2015). Framework for Cloud Intrusion Detection System Service. Web Applications and Networking (WSWAN), 2015 2nd World Symposium on, p1-5

[4]. Barolli, Leonard; Elmazi, Donald; Ishitaki, Oda, Tetsuya; Taro; Yi Liu, Uchida, Kazunori. (24-27 March 2015). Application of Neural Networks for Intrusion Detection in Tor Networks. Advanced Information Networking and Applications Workshops (WAINA), 2015 IEEE 29th International Conference on, p67-72.

[5]. Koushal Kumar, Jaspreet Singh Batth "Network Intrusion Detection with Feature Selection Techniques using Machine-Learning Algorithms" International Journal of Computer Applications (0975 – 8887) Volume 150 – No.12, September 2016.

[6]. R. Karthik, Dr. S. Veni, Dr. B. L. Shivakumar "Improved Extreme Learning Machine (IELM) Classifier For Intrusion Detection System" International Journal of Engineering Trends and Technology (IJETT) – Volume-41 Number-2 - November 2016.

[7]. Premansu sekhara rath, manisha mohanty, silva acharya, monicaa ich "optimization of ids algorithms using data mining technique" International Journal of Industrial Electronics and Electrical Engineering, ISSN: 2347-6982 Volume-4, Issue-3, Mar.-2016

[8]. Mohammadreza Ektefa, Sara Memar, Fatimah Sidi, Lilly Suriani Affendey "Intrusion Detection Using Data Mining Techniques", 978-1-4244-5651-2/10/$26.00 ©2010 IEEE

[9]. YU-XIN MENG," The Practice on Using Machine Learning for Network Anomaly Intrusion Detection" Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong, 978-1-4577-0308-9/11/$26.00 ©2011 IEEE

[10]. Liu Hui, CAO Yonghui "Research Intrusion Detection Techniques from the Perspective of Machine Learning" 2010 Second International Conference on Multimedia and Information Technology 978-0-7695-4008-5/10 $26.00 © 2010 IEEE

[11]. Chuanlong Yin , Yuefei Zhu, Jinlong Fei, And Xinzheng He. "A Deep Learning Approach For Intrusion Detection Using Recurrent Neural Networks" current version November 7,2017 .Digital Object Identifier 10.1109/ACCESS.2017. 2762418