International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 5| ISSN : 2456-3307

Random Forest Algorithm in Intrusion Detection System : A Survey

Kritika Singh*1, Bharti Nagpal2

*1 Ambedkar Institute of Advanced Communication Technologies and Research, Department of Computer Science & Engineering, New Delhi, India
kritikasingh260@gmail.com1

2 Ambedkar Institute of Advanced Communication Technologies and Research, Department of Computer Science & Engineering, New Delhi, India
Bharti_553@yahoo.com2

ABSTRACT

A randomized forest algorithm is based on the classification algorithm under supervision. In this algorithm, the forest is created randomly. The more the number of trees is present, the more accurate result they produced. It is important to note that decision-making using the gain or gain approach is not the same as creating a random forest. This paper presents a survey of Random Forest and other data mining techniques used in Intrusion Detection System.

Keywords : Random Forest, Intrusion Detection System, NIDS, HIDS.

I. INTRODUCTION

A random forest algorithm works on the decision tree concept. It is a classification algorithm under supervision. Generally, the strength of the forest depends on the number of trees on it, the more trees gives more accurate the result. If a greater number of decision trees are created to create the forest, in this case the same apache is not used to generate the resolution with the gini information or index. There is a difference between the Random Forest algorithm and the Decision Tree algorithm. In Random Forest, the root node is found and the feature nodes are randomly broken. "Random" refers mainly to two processes:

- Trees grow with random observations.
- For each node, random variables are selected.

A Random Forest algorithm is widely used for classification. It helps to improve the predictive model. The advantage of using this algorithm is that the results are generated in less time because of the limited assumptions associated with them.

1.1 Intrusion Detection System

An Intrusion Detection System is any device or an application that monitors a host system or a network for the intrusions or any malicious activity. It functions as an alarm system that can report illegal activities when detected. IDS were discovered by James Anderson et al. [1]. The accuracy of IDS depends on discovery rate. When performance is high for IDS, detection accuracy is also high. The types of Intrusion Detection Techniques are:

- Misuse Detection- In misuse detection, the previous intrusions are used to detect the current unknown attacks [2]. Although, it produces a false positive rate, new attacks are not successfully detected.
- Anomaly Detection. If any major deviations are identified from normal activities then it is detected by anomaly detection. Anomaly

detection usually produces a high false positive rate but unknown attacks can be discovered.

Several hybrid approaches proposedbyvarious researchers has been presented in [3], [4], [5].

II. LITERATURE SURVEY

Many researchers have proposed different ways to use data mining techniques in intrusion detection system. Among them the network intrusion detection system is most popular. In 2010 a new decision tree approach was proposed by Manikandan R. et al. [6] for IDS. The focus is on removing the technical issues and the performance enhancement. Eskin et al. [7] applied three different approaches in the unsupervised anomaly detection. These approaches are Cluster-based estimation, K-Nearest Neighbor and SVM. In [8], two data mining techniques are used in misuse and anomaly detection. KDD'99 dataset was used. Misuse detection part is handled by the random forest algorithm and to cluster the data weighted k-means clustering algorithm is used. The patterns are built automatically by using random forest algorithm. In misuse detection framework, intrusion patterns are generated in the offline phase. In anomaly detection techniques noteworthy deviation of a system from normal behaviour is checked. In [9] the random forests algorithm is applied in misuse, anomaly, and hybrid intrusion detection. The issues of the rule-based frameworks were managed by using the Random Forest algorithm to discover new patterns for intrusions. Random Forest algorithm uses training data to build the patterns automatically rather than coding rules manually. In [10], Random Forest model for Intrusion Detection Systems (IDS) was introduced. The aim was to improve the performance of intrusion detection by reducing the input features. The results showed that reduced features produced more accurate result than the Random Forest classification results with all features. Also, time required for processing lesser features with RF will be much less than the processing time of RF with more number of features. In [11] SVM and random forest along with random projection was used for IDS. NSL – KDD dataset was used for this approach. The classification techniques SVM; random forest along with random projection is used. It was concluded that random forest along with random projection yielded the best output than support vector machine along with random projection. In [12] four types of attack: DOS, probe, U2R and R2L were successfully detected by the Random Forest. Any redundant and irrelevant features are removed by applying feature selection on the dataset. The problems of information gain are overcome by the symmetrical uncertainty of attributes. Otherresearchers [13],[14] apply clustering approaches in unsupervisedIDSs. Supervised anomaly detection has been studied extensivelysuch as fuzzy data mining and genetic algorithms [15],neural networks [16], [17], and SVM [18].

Table 1 summarizes several approaches introduced by various researchers.

III. RANDOM FOREST ALGORITHM

In random forests, predictions of multiple decision trees are combined and the final result is based on majority voting. When a decision tree is constructed, the data set must be divided into subtrees, supported by the best combination of variables. However, it is not very easy to find the right combination of variables [19]. A random forest is created from an aggregate result of this forest of assembled trees. Random forest provides better results than individual decision trees. The random forest algorithm is used for both classification and regression. Random forest also handles lost value errors. The Random Forest algorithm works in two stages, first the random forest is created and in the second random forest classifier prediction is made.

Table 1 Survey of various approaches

Technique	Authors	Description
C4.5 compared with SVM	Mohamadreza Ektefa et al.	Two techniques C4.5 and support vector machine (SVM) were used for the IDS This paper concludes that C4.5 algorithms gives better results in detecting network intrusions and false alarm rates than SVM does [21].
C4.5 and BayesNet for IDS	Hai Nguyen et al.	In this paper C4.5 and BayesNet were applied for intrusion detection on KDD CUP'99 Dataset [22].
Random forest to detect attacks like DOS, probe, U2R and R2L	Farnaaz et al.	Any redundant and irrelevant features are removed by applying feature selection on the dataset. The problems of information gain are overcome by the symmetrical uncertainty of attributes. NSL KDD data set was used for this approach [12].
Anomaly intrusion detection.	Eskin et al.	This paper applies three different approaches in the unsupervised anomaly detection. These approaches are Cluster-based estimation, K-Nearest Neighbor and SVM [7].
Ensemble boosted decision tree for IDS.	Manikandan R. et al.	A new decision tree approach is proposed for IDS. The focus is on removing the technical issues and the performance [6].

Pseudocode of creation of Random Forest[20]

```
1. The "m" characteristics of "m" are randomly selected where k << m
2. Node "d" is calculated using the best division point between the "K" characteristics.
3. The best divided approach is used to divide the node into secondary nodes.
4. Repeat 1 to 3 steps until the "l" number of nodes is reached.
5. Steps 1 to 4 are repeated for "n" number of times to create "n" the number of trees that a forest builds.
```

At the beginning of the random forest algorithm, the "k" characteristics are randomly selected from the total "m" characteristics. The child nodes are calculated in the next stage. This is done using the best split approach. Finally, repeat 1 to 4 stages to create "n" trees created at random. A random forest is formed from these trees.

<u>Pseudocode of random forest prediction</u>:

Random forest algorithm uses the prediction pseudocode to make a prediction. It is summarized as:

```
1. The result is predicted by taking the characteristics of the test and using the rules of each decision tree created at random.
2. The votes are calculated for each planned objective.
3. Most voted predicted target is considered as the final prediction.
```

3.1 Random Forest in Intrusion detection System

Random forest method works on the rule divide – and – conquer scheme which is used in the classification mission. As it is an ensemble process, it amalgamates a group of fragile learner to produce well-built leaner which can categorize the data precisely. The bagging scheme and random selection of features are united in it. N number or tresses are produced in random forests. Each tree represents regular and dissimilar malicious classes. A large number of datasets are easily managed by a random forest algorithm. However, there are several challenges in Intrusion Detection System. They are summarized as:

- The intrusion detection rate is improved by the feature selection. The features from raw network traffic data must be constructed by the IDS but a lot of computation is involved in it.
- Imbalanced intrusion is also an issue. Error rate is usually reduced by most of the data mining algorithm, which however leads to rising error rate of minority intrusions which are worse than majority attacks.

3.2 Working of Random forest in Network IDS

The working of misuse detection is shown in Figure 1. Data mining techniques are used to build patterns for network intrusion detection. The framework works in two phases [9]:

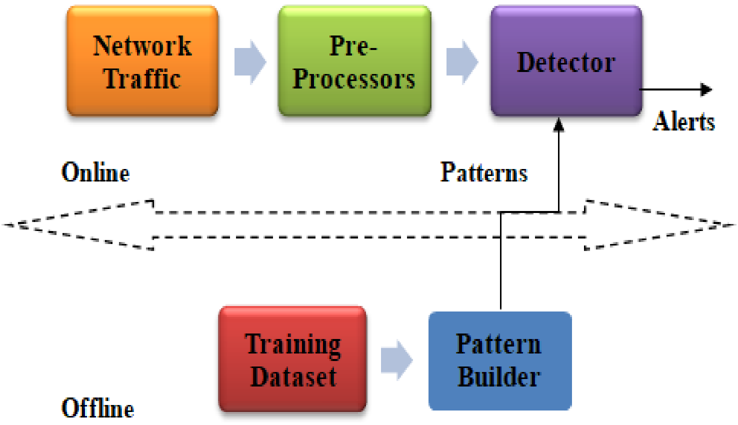

Figure 1 Random forest in Network IDS [9]

- Offline phase: In this phase the patterns of intrusion are built by the system. The training dataset is passed through the pattern builder. This module builds the patterns that are useful for detecting intrusion. Feature selection algorithm and parameter building with random forest algorithm is employed in this module which handles the imbalanced intrusions and. After the patterns are mined, they are sent as an input to the detector module.
- Online phase. In this phase the intrusions are detected. The packets are captured from the network traffic. The pre-processors generate features for each connection captured from the network traffic. The detector module classifies whether the connection is normal traffic or an intrusion. It uses the patterns that are built in the offline phase. Finally, an alert is raised by the system upon identifying intrusion detection.

IV. CONCLUSION

In Random Forest Algorithm the number of trees in the forest and the results from them are directly related, i.e. the more trees, the more accurate the result. It is important to note that, decision-making using the gain or gain approach is not the same as creating a random forest. This paper presented an overview of the random forest algorithm and a survey of various techniques proposed by several researchers has also been summarized.

V. REFERENCES

1. J. P. Anderson, Computer Security Threat Monitoring and Surveillance, Technical Report, James Anderson Report, Pennsylvania, (1980).
2. D. Barbara and S. Jajodia, Applications of Data Mining in Computer Security. Norwell, MA: Kluwer, 2002.
3. D. Anderson, T. Frivold, and A. Valdes, "Next-generation intrusion detection expert system (NIDES) A summary," SRI Int., Menlo Park, CA, Tech. Rep. SRI-CSL-95-07, May 1995.
4. D. Barbara, J. Couto, S. Jajodia, L. Popyack, and N. Wu, "ADAM: Detecting intrusions by data mining," in Proc. 2nd Annu. IEEE Workshop Inf. Assur. Secur., New York, Jun. 2001, pp. 11-16.
5. E. Tombini, H. Debar, L. Me, and M. Ducasse, "A serial combination of anomaly and misuse IDSes applied to HTTP traffic," in Proc. 20th Annual Computer Security. Appl. Conf., Tucson, AZ, Dec. 2004, pp. 428-437.
6. Manikandan R, Oviya P, and Hemalatha C, "A new data mining based network intrusion detection model," Journal of Computer Applications, vol.5, February 2012.
7. E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo, "A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data," in Applications of Data Mining in Computer Security. Norwell, MA: Kluwer, 2002.
8. Ndumiyana, David. "Data mining techniques in intrusion detection: tightening network security." Unspecified (2013).
9. Zhang, Jiong, Mohammad Zulkernine, and Anwar Haque. "Random-forests-based network intrusion detection systems." IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 38.5 (2008): 649-659.
10. Hasan, Md Al Mehedi,"Feature selection for intrusion detection using random forest." Journal of information security 7.03 (2016): 129.
11. Johnson, Susan Rose, and Anurag Jain. "An Improved Intrusion Detection System using Random Forest and Random Projection." Probe 2: U2R.
12. Farnaaz, Nabila, and M. A. Jabbar. "Random forest modeling for network intrusion detection system." Procedia Computer Science 89 (2016): 213-217.
13. K. Leung and C. Leckie, "Unsupervised anomaly detection in network intrusion detection using clusters," in Proc. 28th Australasian CS Conf., Newcastle, Australia, Jan. 2005, vol. 38, pp. 333-342.
14. R. Smith, A. Bivens, M. Embrechts, C. Palagiri, and B. Szymanski, "Clustering approaches for anomaly based intrusion detection," presented at the 1st Annu. Walter Lincoln Hawkins Graduate Res. Conf., New York, Oct. 2002.
15. S. Bridges and R. Vaughn, "Fuzzy data mining and genetic algorithms applied to intrusion detection," in Proc. Nat. Inf. Syst. Secur. Conf. (NISSC), Baltimore, MD, Oct. 2000, pp. 13-31.
16. A. Bivens, M. Embrechts, C. Palagiri, R. Smith, and B. Szymanski, "Network-based intrusion detection using neural networks," in Proc. Artif. Neural Netw. Eng., St. Louis, MO, Nov. 2002, vol. 12, pp. 527-535.
17. M. Ramadas, S. Ostermann, andB. Tjaden, "Detecting anomalous network traffic with self-organizing maps," in Proc. Recent Adv. Intrusion Detect. (RAID), Pittsburgh, PA, Sep. 2003, Lecture Notes in Computer Science, vol. 2820, pp. 36-54.
18. Q. Tran, H. Duan, and X. Li, "One-class support vector machine for anomaly network traffic detection," presented at the 2nd Netw. Res. Workshop 18th APAN, Cairns, Australia, Jul. 2004.
19. Deepanshu B., (2014). Random Forest Tutorial. Retrieved from

https://www.listendata.com/2014/11/random-forest-with-r.html

20. Saimadhu P., (2017, May 22). How Random Forest Algorithm works in Machine Learning. Retrieved from http://dataaspirant.com/2017/05/22/random-forest-algorithm-machine-learing/
21. Mohammadreza Ektefa, Sara Memar, Fatimah Sidi, Lilly Suriani Affendey,"Intrusion Detection Using Data Mining Techniques," Proceesings of IEEE International Conference on Information Retrieval & Knowledge Management , Exploring Invisible World, CAMP'10,2010,pp.200-203.
22. Hai Nguyen, Katrin Franke and Slobodan Petrovi'c, "Improving Effectiveness of Intrusion Detection by Correlation Feature Selection," International Conference on Availability, Reliability and Security, pp. 17-24, IEEE 2010