Advances in Parallel Computing Algorithms, Tools and Paradigms
D.J. Hemanth et al. (Eds.)

doi:10.3233/APC220092

Host Intrusion Detection System Using Novel Predefined Signature Patterns by Comparing Random Forest over Decision Tree Algorithm

P.Sai Chowdary[a] and D.Vinod[b,1]
[a] *Research Scholar, Department of CSE, Saveetha School of Engineering,*
[b]*Professor, Department of CSE, Saveetha School of Engineering,*
[a,b]*Saveetha Institute of Medical and Technical Sciences,*
[a,b]*Saveetha University, Tamil Nadu, India*

Abstract: The objective of the work is to identify anomaly intrusion detection in a user network environment using information accessing and retrieval from homogeneous network databases. Here machine learning algorithms namely Random Forest and Decision Tree are been used to categorize passive and active attack. To accomplish focused accuracy, sample size of n= 10 in Random Forest and n= 10 in Decision Tree was repeated for 20 intervals for well-organized and precise investigation on categorized images with G power in 80% and threshold 0.05%, CI 95% mean and standard deviation. The existing works proves the sequential implemented in focus to intrusion detection, while comparing Random Forest and Decision Tree has classified and predicted the values from the network intrusion to generate accuracy with Random Forest has higher accuracy (76.37%) compared to Decision Tree accuracy (71.57%) with a significance of P<0.001 (2-tailed). Prediction in identifying anomaly intrusion detection systems shows that Random Forest has higher accuracy over Decision Tree.

Keywords:Network Data, Machine Learning, Decision Tree, Random Forest, Novel Predefined Signature Patterns, KDD-CUP99.

1. Introduction

Motive of research is focused towards predicting anomaly detection in a network environment based on numerical and categorical data to predict and improve accuracy for prediction of attacks using novel predefined signature patterns. Efficiency and importance of predicting attacks are framed to be compromised when an intrusion occurs in any protected network [1]. It is found to be important in today's world since intrusion plays a sequential role in trusted networks with a supervised intrusion detection method. A novel proficiency-based detection with effective safeguards are incorporated using Random Forest and Decision Tree algorithm for intrusion detection [2]. Random Forest and Decision Tree procedure is a frequently used machine learning

[1]D.Vinod Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Tamil Nadu, India. E.MailVinodd.sse@saveetha.com

and statistics withdrawal method, thus effective in deception discovery, configuration appreciation and outlier recognition [3]. The serious issue is focused towards normal data to train and give a new piece of test data to find the exact location of intrusion [4]. Novel proficiency on identifying anomaly intrusion are implemented and used in railtel, tele service provider and server-based communication. This happens principally since formerly unseen classification behaviors are also known as variances, and hence flagged as potential intrusions [5]. In order to sequence the experimental study Random Foresthas a good application in intrusion detection, but its performance needs to be further improved.

In the secured network environment, the research papers on anomaly detection includes 1200 journals from IEEE Xplore digital library, 612 articles from ScienceDirect, 863 articles from google scholar and 706 articles from Springer. The Most cited article related to anomaly detection has found the proposed strategy in [6] thus minimizing large amounts of classification [7] workload for domain experts and reducing the scale of training data set. (b) consequently, reduce computational cost of Random Forest and Decision Tree feature selection method is proposed to select the most necessary and important features, therefore, greatly reduce computational cost and avoid curse of dimensionality effectively [8]. Experimental results and performance analyses show that, proposed algorithm is better than general clustering classifications for real-time network intrusion detection.

The procedures which were used previously have proved that it has produced a minimum rate of accurateness and finding rate in anomaly detection. Random Forest was improved using optimization algorithms while comparing decision trees [9]. The aim of the work is recognizing anomaly intrusion detection with signalretrieving and information retrieval on the database.

2. Materials and Methods

The research work was organized in machine learning lab Saveetha School of Engineering, SIMATS. Dual groups of classifier algorithms are used to classify normal and abnormal intrusion detection. The quantity of clusters recognized for the training is 2. Pre-test values have been analyzed and prepared by using clinical.com by having G power of 80% and threshold 0.05%, CI 95% mean and standard deviation [10]. Group 1 is Random Forest and group 2 is Decision Tree. Sample size has been calculated and it is identified that 10 samples/ group in total 20 samples.

The work has been carried out with 702 records which were taken from a kaggle data set. The accuracy in predicting the attacks was initiated by two different groups. Totally 10 iterations were analyzed and performed on each group to accomplish maximum accuracy. The dataset contains 702 instances and 22 features. Here the data is from the Kaggle website. The data is from the year 2015-2017.

Pseudo Code

Input : Training and Test Data

Output: Accuracy

Start: 1.Standardize the data set

2. Aimed at Both C,Y

3. Authenticate by means of one- out

4. Store the success rate RT

5. Compose the average success rate T
6. Modernize the greatest of C and Y if required
7. Select C,Y with best typical accomplishment rate and do step 2
8. End

Decision Trees are often used for non-parametric classification in finding accuracy with regression procedures. Every active attack is sometimes directed among one or multiple sessions, and each session contains many processes. Since the Decision Tree classifier technique monitors the execution of every method, it's extremely probable that associate attacks are often detected whereas it's operational. However, the associate attacker will avoid being detected by not having the method exit. Therefore, there's a necessity for effective classification through a process's execution that may be for future work. Non-parametric technique refers to a way that doesn't assume any distribution.

Pseudo Code

1. Start>>
2. Dataset with S* variables
3. Find the optimal values for the turning parameters of the Decision Tree model
4. Data preprocessing remove punctuations,stopwords from data
5. D = Remove
6. Word embedding_tokenize each message
7. S1 = X_train and Y_train //dataset training
8. S2 = predict(X_test) // testing layer
9. S3 = S2
10. Return input S3

2.1. Statistical Analysis

Study was organized using IBM SPSS version 21. It is an arithmetical software tool used for records analysis. Independent sample t-test calculation for analyzing equal variance, standard error, and levene's test are evaluated. The dependent variables are cyber variants, anomalies and malicious nodes.

3. Results

In (Table 1) Intrusion detection protocol validation with past and primary change of data type with respect to checkpoints and network environment with an initial starting value= primary. In Table 2 shows statistical calculations such as mean, standard deviation and standard error mean for Random Forest and Decision Tree algorithm. It can be seen that deviation for t-test is far lesser than the comparison algorithm. Moreover, the accuracy value of Random Forest is around 76.37 while loss is around 20.80, which seems to be superior to the Decision Tree classifier.

Table 1. Intrusion detection protocol validation with past and primary change of data type with respect to checkpoints and network environment with an initial starting value= primary.

S.No	Attribute	DataType	Value	Description
1	Data	String	Primary	States the type of data
2	Value	int	2017	Stores in integer format
3	flags	char	Sf	checkpoints
4	Src_bytes	String	data	Source data

Table 2: Network intrusion detection description with past and primary variety of intrusions in a network environment. Protocol (TCP) and Duration set to 0.

Duration	Protocol	Service	Flag	Src_bytes
1	Tcp/ip	Ftp data	SF-1	489
1	Udp/ip	other	SF-2	157
1	Tcp/ip	Private	SO-1	1
1	Tcp/ip	http	SF-3	321

In (Table 3) it was observed that the Levens test for equality of variance and its significance for Random Forest and standard error difference and confidence interval are lower than Decision Tree.

Table 3: Group Statistics of Random Forest with Decision Tree by grouping iterations with Sample size 10, Mean = 76.37. Here it specifies Equal variances with and without assuming a T-Test Score of two groups with each sample size of 10.

	Group	N	Mean	Std.Deviation	Std. Error Mean
Accuracy	Random Forest	10	76.370	1.33087	.42086
	Decision Tree	10	71.57	2.11097	.66755
Precision	Random Forest	10	74.1700	1.4461	.45730
	Decision Tree	10	72.375	2.4546	.77621

In (Table 4) Independent Sample values have been validated with respect to Accuracy and Precision. It proves that the Decision Tree and Random Forest are significantly different from each other.

Table 4: Independent Sample Test of (Calculate P-value = 0.001 and Significant value = 0.001 (2-tailed), Mean Difference= 6.986 and confidence interval = (0.3369- 0.2656). Decision Tree and Random Forest are significantly different from each other.

		Levene' s Test For					**T-Test For Equality**	**Of Means**	**95% ConfidenceI nterval of**	**The Difference**
		F	Sig	t	gf	Sig.(2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Accuracy	Equal Variances assumed	0.741	0.772	20.733	18	<0.001	6.98600	0.33695	6.27810	7.69390
	Equal Variances not assumed			20.733	17.957	<0.001	6.98600	0.33695	6.27797	7.69403
Precision	EEqual Variances assumed	1.377	0.022	4.867	18	<0.001	1.29300	0.26565	0.73489	1.85111
	Equal Variances not assumed			4.867	12.958	<0.001	1.29300	0.26565	0.734891	1.86709

Mean accuracy and mean loss graph depicted in (Fig. 1) shows that Random Forest seems to appear better for given dataset of kddcup99 in detecting anomaly attacks. Hence this proves Random Forest and Decision Tree are separated with each other.

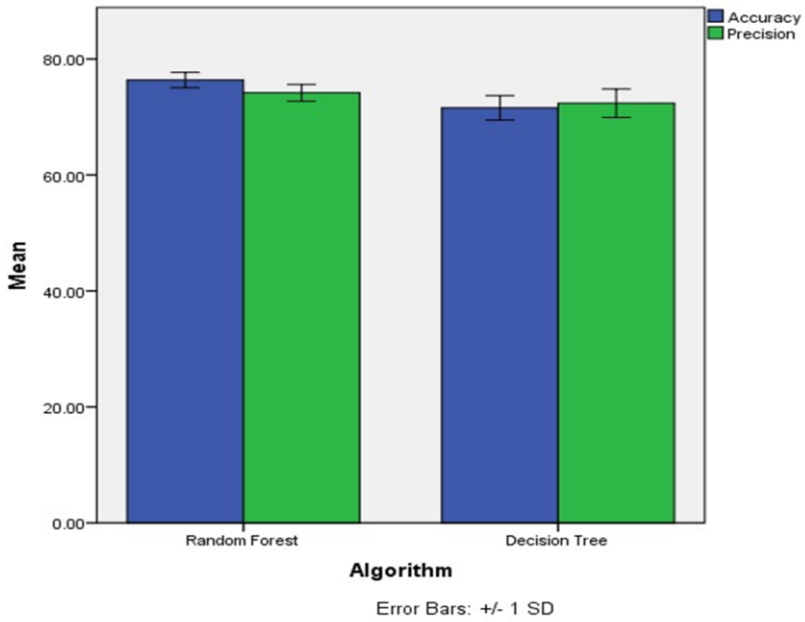

Figure 1. Comparison of Random Forest over Decision Tree in terms of mean accuracy. Bar graph is strategized using group id as X-axis Random Forest vs Decision Tree, Y-Axis exhibiting the inaccuracy bars by mean accuracy of detection +/- 1 SD.

4. Discussion

In this study SVM has higher accuracy than Random Forest which has independent samples t-test. The proposed system provides a better anomaly detection technique using a Random Forest with a normal count with over 76% accuracy, this appears better in comparison to Decision Tree. A systematic review on spam detection techniques presented around 76 studies, analysis of various papers shows that Random Forest is the most used algorithm for data classification. Since most comments are related to text, where algorithm provides the best accuracy for finding anomaly attacks.

The factors that affect intrusion predictions are HIDS network traffic, subnet, in-bond and out-bound NIDS based traffic [11, 12]. Intrusion detection approach proves that safeguards are enabled in live networks to defend the anomaly access to prevent intruders, while a system that analyzes incoming network traffic in NIDS [13,14]. It is also possible to classify IDS by detection approach. The abnormal attacks which may lead to abnormal intrusion prediction with respect to various scenarios [15]. It was observed that no opposition findings have been found against the present research work.

Hence the investigation results on intrusion detection systems are have received more number of responses using ML, which produce better performance in both experimental and statistical analysis, but it has some limitations to the proposed work. Future work of research is to use variants of Random Forest with polynomial kernel and Decision Tree [16]. NIDS has a subnet connectivity from source end then the accuracy evolution goes down due to enormous replay from malicious nodes. The accuracy level of predicting attacks can still be improved by tuning signature patterns.

5. Conclusion

The outcome of the study shows real time traffic with fabricated responses have more value in detecting intrusion in various substantial environments and it was proven in the work established, by comparing Random Forest has higher accuracy (76.23) over Decision Tree (71.25). Continuing with this the same work can be enhanced in future to overcome time-based intrusion detection systems with respect to various network scenarios.

Reference

[1] Bohara, Binita, Jay Bhuyan, Fan Wu, and Junhua Ding. 2020. "A SURVEY ON THE USE OF DATA CLUSTERING FOR INTRUSION DETECTION SYSTEM IN CYBERSECURITY." International Journal of Network Security & Its Applications 12 (1): 1–18.

[2] Bray, Rory, Daniel Cid, and Andrew Hay. 2008. OSSEC Host-Based Intrusion Detection Guide. Syngress.

[3] Dongre, Snehlata S., and Kapil K. Wankhade. 2012. "Intrusion Detection System Using New Ensemble Boosting Approach." International Journal of Modeling and Optimization. https://doi.org/10.7763/ijmo.2012.v2.168.

[4] Harale, Nareshkumar D., and B. B. Meshram. 2016. "Data Mining Techniques for Network Intrusion Detection and Prevention Systems." International Journal of Innovative Research in Computer Science & Technology. https://doi.org/10.21276/ijircst.2016.4.6.3.

[5] Kallel, Slim, Frédéric Cuppens, Nora Cuppens-Boulahia, and Ahmed HadjKacem. 2020. Risks and Security of Internet and Systems: 14th International Conference, CRiSIS 2019, Hammamet, Tunisia, October 29–31, 2019, Proceedings. Springer Nature.

[6] Kumar, Sunil, and Kamlesh Dutta. 2016. "Intrusion Detection in Mobile Ad Hoc Networks: Techniques, Systems, and Future Challenges." Security and Communication Networks. https://doi.org/10.1002/sec.1484.

[7] Li, Yang, and Li Guo. 2007. "An Active Learning Based TCM-KNN Algorithm for Supervised Network Intrusion Detection." Computers & Security. https://doi.org/10.1016/j.cose.2007.10.002.

[8] Mehedi, Sk Tanzir, Adnan Anwar, Ziaur Rahman, and Kawsar Ahmed. 2021. "Deep Transfer Learning Based Intrusion Detection System for Electric Vehicular Networks." Sensors 21 (14). https://doi.org/10.3390/s21144736.

[9] Niemiec, Marcin, Rafał Kościej, and Bartłomiej Gdowski. 2021. "Multivariable Heuristic Approach to Intrusion Detection in Network Environments." Entropy 23 (6). https://doi.org/10.3390/e23060776.

[10] Pasumpon Pandian, A., Ram Palanisamy, and KlimisNtalianis. 2020. Proceeding of the International Conference on Computer Networks, Big Data and IoT (ICCBI - 2019). Springer Nature.

[11] Rash, Michael, Angela Orebaugh, and Graham Clark. 2005. Intrusion Prevention and Active Response: Deploying Network and Host IPS. Elsevier.

[12] Rubika, J., A. Sumathi Felicita, and V. Sivambiga. 2015. "Gonial Angle as an Indicator for the Prediction of Growth Pattern." World Journal of Dentistry 6 (3): 161–63.

[13] Sahu, Santosh Kumar, Durga Prasad Mohapatra, Jitendra Kumar Rout, Kshira Sagar Sahoo, and Ashish Kr Luhach. 2021. "An Ensemble-Based Scalable Approach for Intrusion Detection Using Big Data Framework." Big Data, July. https://doi.org/10.1089/big.2020.0201.

[14] Sampath, Nithya. 2019. "Intrusion Detection in Software Defined Networking Using Snort and Mirroring." International Journal of Psychosocial Rehabilitation. https://doi.org/10.37200/ijpr/v23i4/pr190501.

[15] Schmid, Matthew. 2002. "Computing Platform Coverage via Light Host-Based Intrusion Detection." https://doi.org/10.21236/ada402859.

[16] Skrobanek, Pawel, and Marek Wo. 2011. "Analysis of Timing Requirements for Intrusion Detection and Prevention Using Fault Tree with Time Dependencies." Intrusion Detection Systems. https://doi.org/10.5772/15609.