JCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-7, Special Issue-17, May 2019 E-ISSN: 2347-2693

NIDS using Random Forest and Random Tree

K. Mohanapriya[1*], M. Savitha Devi[2]

[1]Dept. of Computer Science, Government Arts College for Women-Krishnagiri,
[2]Dept. of Computer Science, Periyar University Constituent College of Arts & Science-Harur

Corresponding Author: kmpriya4@yahoo.com

Available online at: www.ijcseonline.org

***Abstract*—** Network Intrusion Detection Systems (NIDS) is the most important system in cyber security and it informs network administrators about policy violations. Identifying the network security violations and tells about where administrators to be improved. In Existing NIDS is designed to detect known network attacks. In this paper it is proposed to develop systematic methods for classifying intrusion detection. The key ideas are to use data mining techniques to discover network behaviour, anomalies and known Intrusions. Decision trees have been effectively used in NIDS but suffer from over sampling and the tree splitting being greedy locally. To overcome this some of the ensemble techniques like Random Forest, Random Trees and Ensemble Weak Learner Tree (EWL TREE) are used. Proposed technique reduces the number of trees required and also improves the precision and recall.

***Keywords*—**Intrusion Detection, Security, Intruder, Decision Tree, Ensemble Weak Learner.

I. Introduction

Internet and online electronic data process are more essential in the present day life. Intruders are main threat for the security of online data processing. Intruders are unauthorized users of the machines they attack and also may violate the permission privileges. Intrusion detection system is a system that monitors network to trace the intruders and report to the administrators.

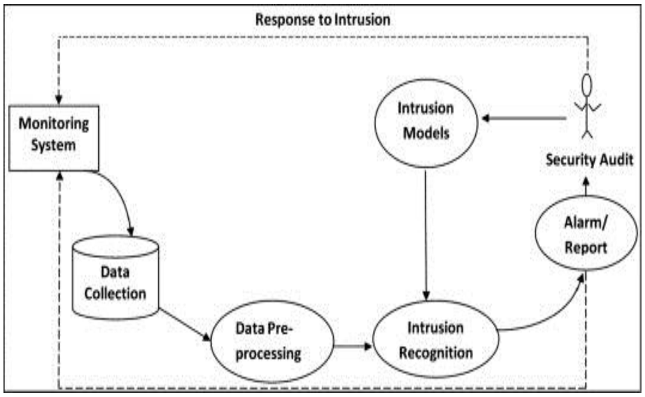

Fig. 1

A. Classification of Intrusion Detection System

IDSs can also be classified based on their deployment in real time.

A.1 1Host Based IDS (HIDS)

A HIDS monitors and analyzes the internals of a computing system rather than its external interfaces. A HIDS might detect internal activity such as which program accesses what resources and attempts illegitimate access.

A.2 Network Based IDS (NIDS)

An NIDS detects Intrusion in network data. Intrusions typically occur as anomalous patterns though certain techniques model the data in a sequential fashion and detect anomalous subsequence. The primary reason for these anomalies is attacks launched by outside attackers who want to gain unauthorized access to the network to steal information or to disrupt the network.

A Network Intrusion Detection System (NIDS) is important IDS that analyzes network traffic at all layers of the Open Systems Interconnection (OSI) model and makes decisions about the purpose of the traffic, analysing for suspicious activity. Most NIDSs are easy to deploy on a network and can often view traffic from many systems at once.

A.3 Misuse Based IDS (MIDS)

Misuse Based Intrusion Detection normally searches for known intrusive patterns but anomaly based Intrusion Detection tries to identify unusual patterns.

A.4 Anamoly Based IDS

An anomaly-based intrusion detection system detects both network and computer intrusions and misuse by monitoring system activity and classifying it as either normal or anomalous.

II. Existing System

In the Existing Intrusion Detection System does not fulfill the network administrator’s expectations. Decision Trees are one of the most widely used systems that can analyze data and identify significant characteristics in the network that indicate malicious activities.

The main function of Decision Tree in the previous system is as follows.

1. Decision Trees can be trained using the processed data and tools.
2. Running and analyzing the result of this data.
3. Creating rules to detect their tactics, techniques, and procedures.
4. Detect previously unknown Network Anomalies.
5. Identify and highlight malicious traffic.
6. Analysis of large sets of Intrusion Detection data.

After Decision Trees are built, they have the potential to reduce the amount of data required for analysis, help identify anomalous malicious activity, and provide analytic insight into the differences between malicious and benign network traffic.

III. Proposed System

The Proposed system is composed of the following actions.

1. Collect Network events.
2. Classify the Abnormal and Normal Behavioural Activities

A.1 Ensemble Weak Learner Tree

The ensemble idea in supervised learning has been investigated and suggests combining two linear regression models. The first linear regression model is fitted to the original data and the second linear model to the residuals. To partition the input space using two or more classifiers. An ensemble of similarly configured neural networks to improve the predictive performance of a single one. At the same time laid the foundations for the award winning AdaBoost algorithm by showing that a strong classifier in the Probably Approximately Correct (PAC) sense can be generated by combining “weak” classifiers (that is, simple classifiers whose classification performance is only slightly better than random classification). Ensemble methods can also be used for improving the quality and robustness of unsupervised tasks. Ensemble methods can be also used for improving the quality and robustness of clustering algorithms. Nevertheless, in this paper we focus on classifier ensembles. Given the potential usefulness of ensemble methods, it is not surprising that a vast number of methods are now available to researchers and practitioners

Ensemble method or any combination model train multiple learners to solve the classification or regression problems, not by simply ordinary learning approaches that can able to construct one learner from training data rather construct a set of learners and combine them.

Ensemble techniques like Random Forest, Random Trees have been proposed in literature which is NP complete. To overcome this optimization method is proposed to improve the selection process of the weak learners in an ensemble tree. The proposed technique Ensemble Weak Learner Tree (EWL TREE) not only reduces the number of trees required but also improves the Precision and Recall. The Ensemble Technique achieves high detection accuracy with less computational time, and minimum cost.

A.2 Random Forest

Machine learning approaches are currently used in the data mining domain, intention of defencing the random forest algorithm to generalize the intrusion detection by utilizing the ensemble machine learning technique. The decision tree is used as a common classifier in the RF and it generates multiple decision trees to form a forest. The process undergoes randomization at two stages; the bootstrap samples are randomly sampled as in bagging and then the individual base decision tree is generated by random selection of input attributes. The generalization error of Random Forest classifier is based on the strength of individual decision tree and correlation among base trees. The performance of the Random Forest classifier is at equivalence with the other ensemble techniques like bagging and boosting based on the accuracy measure. Random Forest has the ability to potentially run on large databases that have millenary of input variables and can deal with these variables without any deletion in the variable as stated by [Brieman 2001]. The RF estimates the intrusion search category by which the class are categorized related reference of unbalanced populated attribute reference.

The random vectors specified the k-values of random field collection in the form of tree representation. An ensemble of decision trees is generated in RF. Following are the steps followed by Breiman to generate every single tree in RF: The Bootstrap sample is done on the N records of the training set; these records are sampled from the original data at random but with replacement. This obtains to practice the sample set for constructing the tree. The extension of tree structure be elaborated and the process of growing the tree is by splitting the attributes to split the nodes. For an input variable of size M is the data, a number m-tree node less than M is chosen so that m-tree variables are selected randomly at every node. The attributes of M is split into node. During the forest growing stage, the value of m remains constant. Pruning is not applied here.

The random forest resembles the tree in various forms is determined by the parameter N-tree. Multiple trees are grown in the similar way as an individual tree grows. The (m) represent the variable point used at the node can also be called as m-tree or k. The number of the instances in at leaf node i.e. parameter node size decides the depth of the tree. The new instance can be classified after training i.e. the building of the tree by running a search process across the entire forest build by several trees. The vote is the classification value of the new sample done by a tree and is recorded at each node. The maximum counts of votes for a class obtained after combining all the votes from all the tree is the classification of the new instance.

A.3 Random Tree

A repetitive division of the given data space is used for representing the decision tree. The Decision Tree (DT) comprises a rooted tree. A node is used for directing this, and is known as the root, the root is the main component and the leaves refer to the remainder of the nodes. A decision tree is constructed for the decision tree algorithm, automatically for the given dataset in such a way that the error is least. Striving to optimize the cost function, the decision tree classifier determines the decision tree (T), given a set of L labelled samples. Here, after optimizing the decision tree, it strives to determine an optimal class from a data set that has been given, when a query image has been provided as a test case.

Also known as the non-parametric method, the Decision Tree model is represented as a hierarchical tree structure. It not only has the ability to identify the non-additive and non-linear relationships between the input factors, but can also target predictive factors. Representing the protocols, decision tree delineates structural data patters. A set of rules contain the decision tree to be used by identifying the class on unseen records, for a given dataset of factors along with its classes. The model/structure along with the likely outcome off the decision tree can be elaborated as a tree structure. The tree comprises the following: the higher node representation corresponding to observed extracted value, also known as the root node, a decision st-node value a, intermediate value of node comprising more than groups as well as a set of pointed tree nodes and the bud nodes which are used for denoting the decision or the classification. A binary tree structure of every node generates a binary decision which can split one or the terminal node is encountered.

IV. RESULTS AND DISCUSSION

The resultant tested with positive negative cases of search content for which the anomalies to look at the part of irregular activities of content to detect. The proposed optimized ensemble weak learner tree attained the classification result be developed. The experiments were carried out using the subset of KDD99 dataset consisting of 4855 normal and 6417 abnormal data stream. Random Forest and Random Tree is used to find the relational supportive depth measure of tree. Table 1 shows the Classification

Table 1 Experimental evaluated Results

Classification Function	Random Forest	Random Tree
True Positive	5151	5222
True Negative	5364	5416
False Positive	381	319
False Negative	376	315
Total Activities	11272	11272
Method Parameters	Precision	Recall
Random Forest	0.9311	0.932
Random Tree	0.9424	0.9431

Precision value = $\frac{\text{TP}}{(\text{TP+FP})}$............. (4.1)

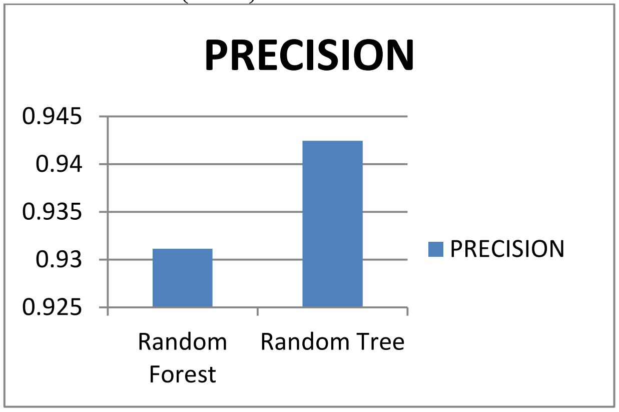

Figure 2 Impact of Precision analysis

From table 1 and figure 2retains the precision intent system Random Tree performs better rate of Random Forest by 1.012%, Recall value = $\frac{\text{TP}}{(\text{TP+FN})}$........... (4.2)

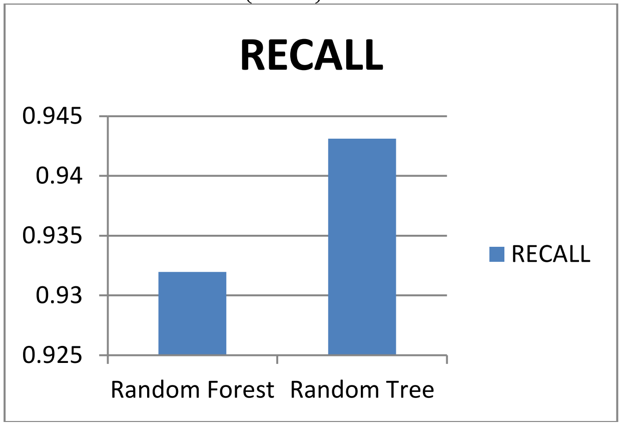

Figure 3 Impact of Recall analysis

From table 1 and figure 3retains the recall intent system Random Tree performs better rate of Random Forest by 1.012%,

V. Conclusion

Recent years, increase in the usage of computers and mobile over Internet for social networking, healthcare, e-commerce, bank transactions, and many other services. The networks allow users to create a profile including their personal information, to add other users as friends and to exchange messages. It opens the door for unlawful activities. Existing Decision tree algorithms which are applied to online networks are limited because of issues such as complexity, low accuracy; privacy and it mostly follow the signature based analysis. Using Ensemble methods for network security makes us secure communication in public network. In Ensemble methods, Result shows Random Tree is better than Random Tree

References

[1] Raghunath, B. R., &Mahadeo, S. N. (**2008**, July). Network intrusion detection system (NIDS).

[2] Chandola.V, Banerjee.A&Kumar.V (**2009**) "Anomaly detection: A survey", ACM Computing Surveys (CSUR

[3] G.K. Gupta "Introduction to data mining with case studies.

[4] Tan .P. N, Steinbach .M & Kumar .V (**2005**) "Introduction to data mining" Pearson Addison Wesley

[5] Rokach, L. (**2010**). Ensemble-based classifiers. Artificial Intelligence Review, **33(1), 1-39**.

[6] Jones JM, Fielding A, Sullivan M (**2006**) Analyzing extinction risk in parrots using decision trees. BiodiversConserv **15(6):1993–2007**.

Authors Profile

TITLE	:	NIDS USING RANDOM FOREST AND RANDOM TREE
NAME	:	MOHANAPRIYA .K
DESIGNATION	:	GUEST LECTURER DEPARTMENT OF COMPUTER SCIENCE GOVERNMENT ARTS COLLEGE FOR WOMEN, KRISHNAGIRI – 635002.
ADDRESS	:	5/876 ANNAI TERESA STREET, SENTHIL NAGAR, DHARMAPURI – 636 705.
E-MAIL ID	:	kmpriya4@yahoo.co.in
CONTACT NO.	:	9443496196
GUIDE NAME	:	Dr.M.SAVITHA DEVI, M.Sc., M.Phil., M.C.A., B.Ed., P.hd.,
DESIGNATION	:	ASST. PROFESSOR, DEPARTMENT OF COMPUTER SCIENCE, PERIYAR UNIVERSITY CONSTITUENT COLLEGE OF ARTS & SCIENCE,HARUR.