International Journal of Computer Trends and Technology
ISSN: 2231-2803 / https://doi.org/10.14445/22312803/IJCTT-V68I2P109

Volume 68 Issue 2, 59-63, February 2020

Original Article

Intrusion Detection Using Tree-Based Classifiers

Ashalata Panigrahi[1], Manas Ranjan Patra[2]

[1]*Roland Institute of Technology, Berhampur, India*
[2] *Berhampur University, Berhampur, India*

Received Date: 14 January 2020
Revised Date: 20 February 2020
Accepted Date: 24 February 2020

Abstract – *Growing cyber-crimes have become a serious concern for network users. It has become a real challenge for organizations to develop network security systems to protect data from all kinds of illegal access. Since intruders keep applying different techniques to break the security barriers, the techniques to counter such attacks are also being developed by the researchers. In this work, a model has been proposed for building an effective intrusion detection system using tree-based classification techniques, namely, BF Tree, FT, J48, NB Tree, Random Forest, and Random Tree. Further, three nature-inspired and two heuristic search-based methods have been applied for selecting important features prior to the classification process. The performance of the model has been evaluated on the NSL-KDD dataset in terms of accuracy, precision, detection rate, and false alarm rate*.

Keywords – *Best First Tree, Functional Tree, Naïve Bayes Tree, Particle swarm optimization, Heuristic search.*

I. INTRODUCTION

Nowadays, there is a tremendous growth of network-based services such as e-commerce, e-business, file sharing, social networking that improves the lives of modern society. The rapidly growing number of network attacks has become a serious threat for computer networks worldwide. Security threats come from different sources such as natural forces (such as a flood), accidents (such as a fire), failure of services (such as power) and people known as intruders. The traditional prevention techniques such as firewalls, user authentication, data encryption, antivirus, and avoiding programming errors are used as the first line of defence for computer security. Today, intrusion detection is one of the high priority and challenging tasks for network administrators and security professionals. Information security is protecting the information against unauthorized transfer or modification intentionally when it is transmitted through the network. The main function of an Intrusion Detection System is to monitor the computer system and automatically detect attacks from the network data traffic. Once the attack is detected, an alarm is raised for an administrator. Intrusion detection is classified into two types: misuse based intrusion detection and anomaly-based intrusion detection [1]. Misuse detection, where the detection process is based on known signatures or patterns, aims to distinguish legitimate instances from malicious ones. Anomaly detection is designed to detect malicious actions by identifying deviations from a normal profile behaviour. This kind of IDSs performs better in detecting novel attacks.

II. RELATED WORK

Salo et al. [2] proposed a novel hybrid model for intrusion detection based on two feature selection methods, namely, information gain and principal component analysis (PCA), which combines Support Vector Machine, instance-based k-nearest neighbours (IBK) and multilayer perceptron (MLP) classification techniques. . They reported accuracy of 98.24% on the NSL-KDD dataset. Hota and Shrivas [3] proposed a model that used different feature selection techniques to remove the irrelevant features in the dataset. The results indicate that C4.5 with Info Gain had better results and achieved the highest accuracy of 99.68% with only 17 features. Al-Yaseen et al. [4] proposed a new learning technique for developing a novel intrusion detection system using a modified k-means algorithm. The popular KDD Cup 99 dataset is used to evaluate the performance of the proposed model. The proposed model shows high efficiency in attack detection, and accuracy is 95.75%. Akyol et al. [5] have proposed an approach for IDS with the use of a discernibility function based feature selection method, and then multilayer perceptron and C4.5 algorithm were applied on KDD'CUP 99 dataset. They reported accuracy of 98.03%. Admin et al. [6] proposed an IDS based on the combination of the probability predictions of a tree of classifiers—a two-layer model. The first layer is a classification tree, and the second layer is a classifier, which combines the probability prediction of the Tree. They reported accuracy of 89.75% on the NSL-KDD dataset. Li et al. [7] proposed a model combining a Gini index and gradient boosting decision tree (GBDT) with particle swarm optimization (PSO). The optimal feature subset is selected by the Gini index, and the network attack is detected by a gradient lifting decision tree algorithm. The parameters of GBDT are optimized by the PSO algorithm. They reported accuracy of 86.10% on the NSL-KDD dataset. Shamshirband et al. [8] proposed a cooperative fuzzy Q-learning (Co-FQL) method, which was compared with the fuzzy logic controller, Q-learning and fuzzy Q-learning methods. They reported accuracy of 89.68% on the NSL-KDD dataset.

III. METHODOLOGY

A. Best First Tree (BF Tree)

Best First trees[9] are constructed in a divide-and-conquer approach. First, select the best feature and place it at the root node. The best node is the node whose split leads to a maximum reduction of impurity among all nodes available for splitting. Second, to determine the node to be expanded next, and finally, decide on the stopping criteria for the Tree to grow. Best first tree learning selects the "best" node to split at each step.

B. Functional Tree (FT Tree)

The FT algorithm [10] uses a standard top-down recursive partitioning strategy to construct a decision tree. Splitting at each node is univariate but considers both the original attributes in the data and new attributes constructed using an attribute constructor function: multiple linear regression in the regression setting and linear discriminants or multiple logistic regression in the classification setting.

C. J48

J48 [11] builds decision trees from a set of labelled training data using the concept of information entropy. It uses the fact that each attribute of the data can be used to make a decision by splitting the data into smaller subsets. J48 examines the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. The attribute with the highest normalized information gain is chosen to make the decision. The splitting procedure stops if all instances in a subset belong to the same class.

D. Naïve Bayes Tree (NB Tree)

The NB Tree [12] is a hybrid learning approach of decision tree and Naïve Bayes Classifier. NB Tree splits the dataset by applying an entropy-based algorithm and using standard NBC at the leaf node to handle attributes. The advantages of both decision tree and NBC can be utilized simultaneously.

E. Random Forest

The Random Forests [13] is an ensemble of un-pruned 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. Each Tree gives a vote that indicates the Tree's decision about the class of the object. The forest chooses the class with maximum votes for any object.

F. Random Tree

Random Tree [14] is an ensemble learning algorithm that generates many individual learners. It employs a bagging idea to produce a random set of data for constructing a decision tree. In a standard tree, each node is split using the best split among all variables.

IV. THE PROPOSED MODEL

The objective of the proposed model is to build an efficient intrusion detection model that can achieve high accuracy, high detection rate and low false alarm rate. The model consists of two phases, as depicted in figure 1. In phase 1, important features are selected using three nature-inspired search-based techniques such as Particle Swarm Optimization search, Genetic Search, and Ant search and two heuristic search-based techniques such as Best first search and Greedy stepwise search. In phase 2, the reduced dataset is classified using six Tree-based classifiers. Further, a 10-fold cross-validation technique is used for training and testing of the model, and the performance of the model is evaluated using certain standard criteria.

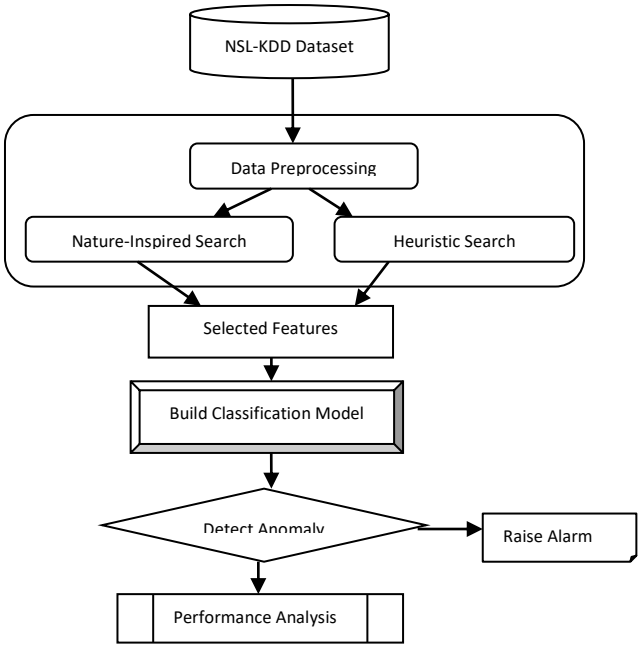

Fig. 1 Proposed Model

V. EXPERIMENTAL SETUP

A. NSL-KDD Dataset

The NSL-KDD intrusion dataset [15], which is a reduced version of the original KDD'99 dataset, has been used for experimentation. The dataset consists of 41 features and a total of 125973 records, of which 67343 are normal, and 58630 are attacks. The dataset contains 24 different attack types, which can be classified into four categories viz. Denial of Service (DoS), Remote-to-Local (R2L), Probe and User-to-Root (U2R). Each attack category consists of different attack types. A DoS attack is a type of attack that tries to shut down traffic flow to and from the target system, for example, ping-of-death, smurf, etc. A remote to local attack is an attack in which the intruder tries to exploit the system vulnerabilities in order to control the remote machine through the network as a local user, for example, guessing passwords etc. Probe or surveillance is an attack that tries to get information from a network, for example, a port scan. A user to root attack is an attack that starts off on the system with a normal user account and tries to gain access to the system or network as

a super-user (root). The attacker attempts to exploit the vulnerabilities in a system to gain root privileges/access, for example, phf etc.

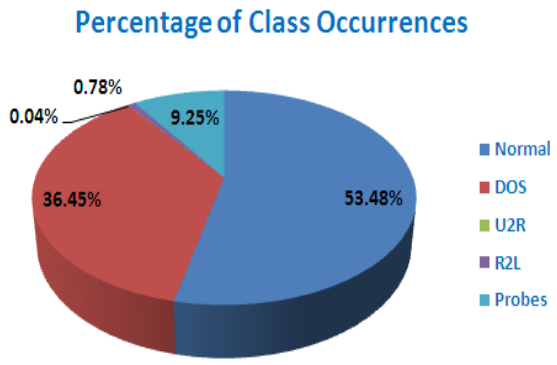

Fig. 2 Distribution of Records

B. Feature Selection

Feature selection is an effective and essential step in successful high dimensionality data mining applications. [16]. The feature selection process reduces the dimensionality of feature space, removes irrelevant and noisy data and select the most important features. In this work, three nature-inspired search-based feature selection methods, namely, PSO search, Genetic Search, and Ant search and two heuristic search-based methods, namely, Best first and Greedy stepwise methods, have been applied for the selection of important features.

C. Confusion Matrix

To evaluate the results of classifier techniques confusion matrix is used. The confusion matrix is a table with two rows and two columns that reports the number of True Positive, True Negative, False Positive, False Negative. The matrix maintains the information about actual and predicted classes. An IDS is evaluated by its ability to make an accurate prediction of attacks. Intrusion detection systems mainly discriminate between two classes, attack class (abnormal data) and normal class (normal data). The performance of the model is measured by computing the accuracy, precision, recall/detection rate and false alarm rate.

The accuracy, detection rate, precision, false alarm rate, and F-measure are calculated as follows:

Accuracy = (TP + TN) / (TN + TP + FN + FP)

Detection Rate or Recall = TP / (TP + FN)

Precision = TP / (TP + FP)

False Alarm rate = FP / (TN + FP)

Where

TP represents True Positive when an attack is detected successfully, and an alarm is raised.

FP represents False Positive when a normal connection is wrongly detected as an attack, and a false alarm is raised.

TN represents True Negative when a normal connection does not raise any alarm.

FN represents False Negative when an attack is not detected, and an alarm is not raised

VI. RESULTS AND DISCUSSION

Six different Tree-based classifiers, namely, Best First Tree Algorithm (BF Tree), Functional Trees (FT), J48, Naive Bayes Trees (NB Tree), Random Forests, Random Tree with two categories of feature selection methods, namely, natured-inspired and heuristic search-based methods were applied on the NSL-KDD dataset, and their performance was measured in terms of accuracy, precision, detection rate and false alarm rate. A comparative analysis of different combinations of classifiers and feature selection methods are depicted in table 1 and 2. The result shows that the Random Forest technique with the Genetic search feature selection method gives the highest accuracy of 99.8738 %, the highest detection rate of 99.8192%, and a low false alarm rate of 0.0787%. Detection Rate and False Alarm Rate of different combinations of classifiers with nature-inspired search-based feature selection methods and heuristic search-based feature selection methods are presented in figures 3, 4, 5, and 6, respectively.

Table 1. Comparison of Tree-based Classifiers with Nature-Inspired Search based Feature Selection Method

Feature Selection Method	Test Mode	Classifier Techniques	Evaluation Criteria			
			Accuracy in %	Precision in %	Recall or Detection Rate in %	False Alarm Rate in %
PSO Search	10-Fold Cross Validation	BF Tree	99.5944	99.7432	99.3843	0.2227
		FT	99.4586	99.6096	99.2256	0.3386
		J48	99.6047	99.8081	99.3416	0.1663
		NB Tree	99.6325	99.8252	99.3843	0.1515
		Random Forest	**99.6547**	**99.8424**	**99.415**	**0.1366**
		Random Tree	99.592	99.7245	99.3979	0.2391
Genetic	10-Fold Cross	BF Tree	99.7992	99.8446	99.7237	0.1351

Search	Validation	FT	99.7356	99.7559	99.6759	0.2123
		J48	99.8325	99.8668	99.7731	0.1158
		NB Tree	99.8492	99.8958	99.78	0.0906
		Random Forest	**99.8738**	**99.9095**	**99.8192**	**0.0787**
		Random Tree	99.7832	99.7816	99.7527	0.1901
Ant Search	10-Fold Cross Validation	BF Tree	99.5443	99.5934	99.4269	0.3534
		FT	99.3713	99.4393	99.2086	0.487
		J48	99.5539	99.5765	99.4644	0.3683
		NB Tree	**99.5682**	99.6326	99.4388	0.3193
		Random Forest	99.5642	**99.6342**	99.4286	**0.3178**
		Random Tree	99.4705	99.4691	99.3928	0.4618

Table 2. Comparison of Tree-based Classifiers with Heuristic Search based Feature Selection Method

Feature Selection Method	Test Mode	Classifier Techniques	Evaluation Criteria			
			Accuracy in %	Precision in %	Recall or Detection Rate in %	False Alarm Rate in %
Best First Search	10-Fold Cross Validation	BF Tree	99.8071	99.8651	99.7203	0.1173
		FT	99.7174	99.7694	99.6231	0.2005
		J48	99.7801	99.8445	99.6827	0.1351
		NB Tree	**99.8396**	**99.8975**	99.7578	**0.0891**
		Random Forest	99.838	99.8583	**99.7936**	0.1232
		Random Tree	99.7976	99.7851	99.78	0.1871
Greedy Stepwise	10-Fold Cross Validation	BF Tree	99.8063	99.8599	99.7237	0.1218
		FT	99.7229	99.7694	99.635	0.2005
		J48	99.7793	99.836	99.6896	0.1425
		NB Tree	99.8404	**99.8924**	99.7646	**0.0935**
		Random Forest	**99.8468**	99.8856	**99.7851**	0.0995
		Random Tree	99.796	99.78	99.7817	0.1915

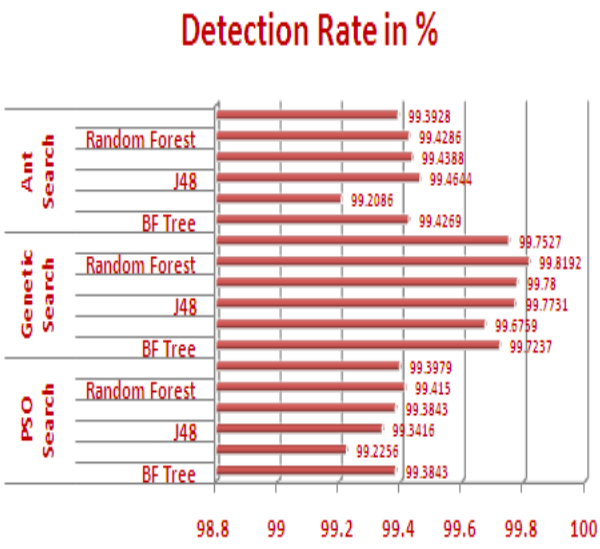

Fig. 3 Comparison of Detection Rate

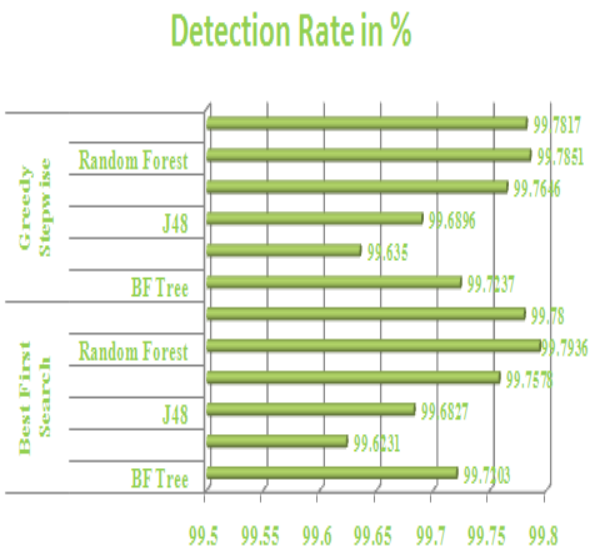

Fig. 4 Comparison of Detection Rate

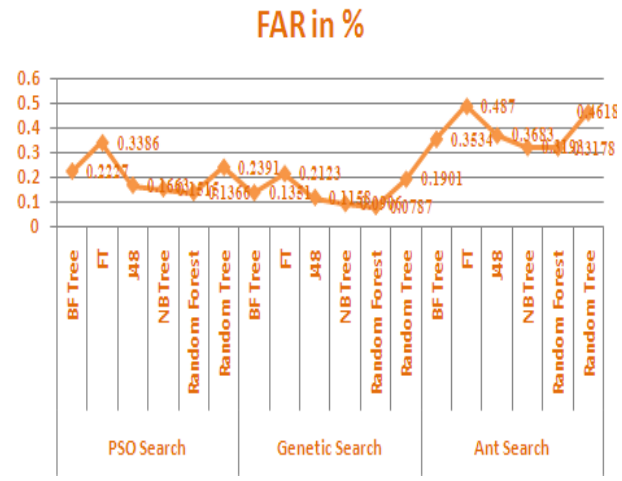

Fig. 5 Comparison of FAR

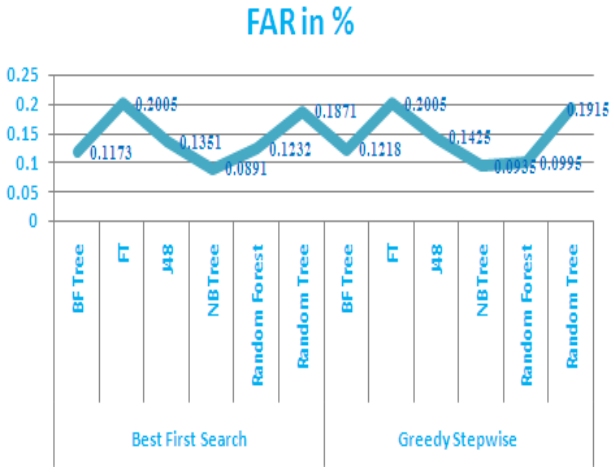

Fig. 6 Comparison of FAR

REFERENCES

[1] O. Joldzic, Z. Djuric, and P. Vuletic, A transparent and scalable anomaly-based dos detection method, Computer Networks. 104 (2016) 27–42.

[2] F. Salo, A. B. Nassif, A. Essex, Dimensionality reduction with IG-PCA and ensemble classifier for network intrusion detection, Computer Network. (2018) 164–175.

[3] H. Hota and A. K. Shrivas, Decision tree techniques applied on NSL-KDD data and its comparison with various feature selection techniques, Advanced Computing, Networking and Informatics,Springer. 1 (2014) 205–211.

[4] W. L. Al-Yaseen, Z. A. Othman, and M. Z. A. Nazri, Multilevel hybrid support vector machine and extreme learning machine based on modified K-means for intrusion detection system, Expert Systems with Applications. 67 (2017) 296–303.

[5] A. Akyol, M. Hacibeyoglu, and B. Karlik, Design of multilevel hybrid classifier with variant feature sets for the intrusion detection system. IEICE Trans. Inf. Syst. (2016) 1810–1821.

[6] Ahmim M, Derdour, and M. A. Ferrag, An intrusion detection system based on combining probability predictions of a tree of classifiers, International Journal of Communication System. 31 (2018) 1–14.

[7] L. J. Li, Y. Yu, S. S. Bai, J. J. Cheng, and X.Y. Chen, Towards effective network intrusion detection: A hybrid model integrating Gini index and GBDT with PSO, Journal of Sensors. 6 (2018) 1–9.

[8] S. Shamshirband, B. Daghighi, N.B. Anuar, M.L.M. Kiah, A. Patel, A. Abraham, Co-FQL: Anomaly detection using cooperative fuzzy Q-learning in-network, Journal of Intelligent and Fuzzy System. 28 (2015) 1345–1357.

[9] S. Haijian, Best-First Decision Tree Learning. Masters Degree Theses. University of Waikato Masters Theses. (2007).

[10] J. Gama, Machine Learning, Kluwer Academic Publishers. (2004) 219-250.

[11] R. Quinlan, C 4.5: Programs for Machine Learning, Morgan Kaufmann Publishers. San Mateo, CA. (1993).

[12] R. Kohavi, Scaling up the accuracy of Naive Bayes classifiers: a Decision tree hybrid, In Proc. Of the 2nd international conference on knowledge discovery and data mining. (1996) 202-207.

[13] L.Breiman, Random Forests, Machine Learning. 45 (2001) 5–32.

[14] H. W. Ian, E. Frank, and M.A. Hall, Data Mining Practical Machine Learning Tools and Techniques, Third Edition. Morgan Kaufmann Publishers. (2012).

[15] M.Tavallaee, E. Bagheri, W. Lu, and A. Ghorbani, A detailed analysis of the KDD CUP 99 data set, Proceedings of the 2009 IEEE Symposium on Computational Intelligence in Security and Defence Applications. (2009) 1-6.