Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 89 (2016) 213 – 217

Procedia Computer Science

Twelfth International Multi-Conference on Information Processing-2016 (IMCIP-2016)

Random Forest Modeling for Network Intrusion Detection System

Nabila Farnaaz* and M. A. Jabbar

MJCET Hyderabad, India

Abstract

With the growing usage of technology, intrusion detection became an emerging area of research. Intrusion Detection System (IDS) attempts to identify and notify the activities of users as normal (or) anomaly. IDS is a nonlinear and complicated problem and deals with network traffic data. Many IDS methods have been proposed and produce different levels of accuracy. This is why development of effective and robust Intrusion detection system is necessary. In this paper, we have built a model for intrusion detection system using random forest classifier. Random Forest (RF) is an ensemble classifier and performs well compared to other traditional classifiers for effective classification of attacks. To evaluate the performance of our model, we conducted experiments on NSL-KDD data set. Empirical result show that proposed model is efficient with low false alarm rate and high detection rate.

Peer-review under responsibility of organizing committee of the Organizing Committee of IMCIP-2016

Keywords: Intrusion Detection; Data Mining; Random Forest; Feature Subsrt Selection; NSL-KDD Data Set.

1. Introduction

Intrusion activities are increasing due to the increase of network usage. In recent years, attacks on computer systems are increasing and require effective and efficient intrusion detection system. Goal of IDS is to find intrusion in normal audit data. IDS was first introduced by James Anderson *et al.*[1] Intrusion detection system are classified intro two types: Host based and network based. Host based intrusion detection systems examine the data held an individual computer systems, whereas network based systems analyses the data exchanged among computer[2].

The random forest is an ensemble classifier[3]. Many current IDS are developed for classifying the attacks. However, the methods produce less accuracy in detecting intrusion. Therefore, we propose intrusion detection system using Random forest.

The major highlights of our approach are:

1) To propose a new model that apply random forest algorithm for network intrusion detection.
2) Classify various type of attacks.
3) To improve accuracy of classifier in detection different types of attacks.

Section 2 discusses the related work and Section 3 explains our proposed model. Experimental results are analyzed in Section 4. Finally, we will conclude in section 5.

*Corresponding author. Tel.: +919966350265.
E-mail address: nabila.farnaaz@gmail.com

Peer-review under responsibility of organizing committee of the Organizing Committee of IMCIP-2016
doi:10.1016/j.procs.2016.06.047

2. Related Work

In 2012, Arif Jamal Malik *et al.*, proposed network IDS using Random forest and PSO. Binary PSO is used to select appropriate features for classifying intrusions. Random forest algorithm is used as a classifier, Their method consists of two stages 1) Feature selection 2) Classification. Authors implemented proposed method in MATLAB[4].

Multistage filtering for network IDS is proposed by P. Natesan *et al.*[5] Authors used enhanced adaboost with decision tree algorithm and Naïve bayes to detect frequent attacks in networks.

A Hybrid Intelligent Approach for IDS was proposed by Mrutyunjaya Panda *et al.*[6] Authors used a combination of classifiers to improve the performance of resultant model. They used classification strategy with 10 fold cross validation. Experimental results are conducted on NSL-KDD dataset.

IDS using Random forest and SVM was proposed by Md Al Mehedi Hasan *et al.*[7] Authors developed two models for IDS using SVM and Random forest. The performance of these two approaches are compared based on their accuracy, precision and false negative rate.

Ujwala Ravale *et al.* proposed feature selection based Hybrid IDS using K-means and Radio basis function. Authors proposed hybrid technique which combines K-means and SVM. Experimental results are done using KDD Cup 99 dataset.

3. Proposed Work

In this section, we first describe IDS and RF algorithm and discuss our proposed method for IDS.

3.1 Intrusion Detection System (IDS)

IDS is defined as a malicious, externally induced operational fault[9]. IDS plays a important role in detecting various types of attacks. The main goal of IDS is to find intrusions and can be considered as classification problem[10]. IDS can be classified into various attacks such as DOS, probe, U2R, R2L[11].

3.2 Random Forest (RF)

Random fore (RF) is an ensemble classifier used to improve the accuracy. Random forest consists of many decision trees. Random forest has low classification error compared to other traditional classification algorithms. Number of trees, minimum node size and number of features used for splitting each node.

Advantages of RF are listed below.

1) Generated forests can be saved for future reference[12].
2) Random forest overcomes the problem over fitting.
3) In RF accuracy and variable importance is automatically generated[13].

When constructing individual trees in random forest, randomization is applied to select the best node to split on. This value is equal to $\sqrt{A}$, where A is no. of attributes in the data set[14]. However, RF will generate many noisy trees, which affect accuracy and wrong decision for new sample[14].

3.3 Feature selection

Feature selection (FSS) is a pre processing step commonly used in data mining. It is effective in dimensionality reduction and removes irrelevant features thus increases accuracy. It refers to the problem of identifying those features that are useful in predicting class. Feature selection methods can be classified into three categories 1) filter method, 2) wrapper method and 3) embedded method[15].

3.4 Proposed approach

Research framework for our proposed method is shown in Fig. 1.

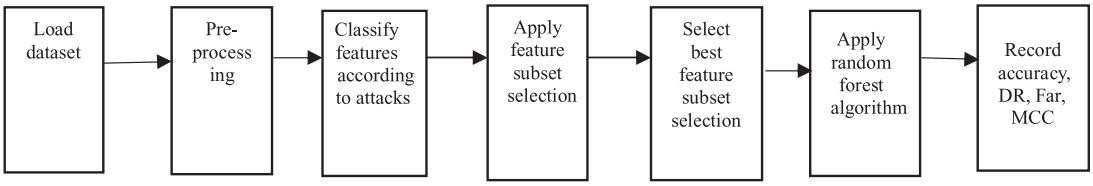

Fig. 1. Sequence of Steps for Proposed Approach.

Our proposed algorithm is described below

Algorithm: Random forest modeling for network IDS
Input: NSL-KDD dataset
Output: Classification of different type of attacks
Step 1: Load the dataset
Step 2: Apply pre-processing technique Discretization
Step 3: Cluster the dataset into four datasets.
Step 4: Partition the data set into training and test
Step 5: Select the best set features using feature subset selection measure Symmetrical uncertainty (SU)
Symmetrical uncertainty compensates information gain

$$SU(X,Y) = 2[IG(X/Y)/H(X)H(Y)]$$

Step 6: Data set is given to Random forest for training
Step 7: The test data set is then fed to random forest for classification
Step 8: Calculate accuracy, Detection rate, False alarm rate, Mathew correlation coefficient

For our experimental analysis, we downloaded the NSL-KDD dataset in ARFF format. We adopted the following preprocessing techniques to run the experiment.

1) Replace missing values:
   In weka, we used replace missing values filter to replace all missing feature values in NSL-KDD dataset. This filter replaces all missing values with the mean and mode from the training data.
2) Discretization:
   Numeric attributes were discretized by discretization filter using unsupervised 10 bin discretization.

4. Experimental Results

All experiments were carried using weka tool. We used NSL-KDD dataset for our analysis. NSL-KDD dataset consists of 42 attributes; last attribute consists of class label. We tested for various number of Random forest trees. Following performance measures are used to evaluate the classifier. 10 cross validation is adopted for classification.

1) Accuracy – Defined as the ratio of correctly classified samples to total number of samples

$$\text{Accuracy} = \frac{\text{Samples correctly classified in test data}}{\text{Number of samples in test data}}$$

2) Detection rate – It is the ratio between total numbers of attacks detected by the system to the total number of attacks present in the dataset

$$\text{DR} = \frac{TP}{TP + TN}$$

3) False alarm rate – false alarm rate is defined as

$$\text{FAR} = \frac{FP}{TN + FP}$$

4) Mathews correlation coefficient (MCC) – This is defined as ratio between the observed and predicted binary classifications.

$$\text{MCC} = \frac{TN \times TP - FN \times FP}{\sqrt{(FP + TP)(FN + TP)(TN + FP)(TN + FN)}}$$

The above performance measures are derived from confusion matrix which is shown below.

	Normal	Attack
Normal	TP	FP
Attack	FN	TN

where

TN → True negative

FN → False negative.

FP → False positive

TP → True Positive.

Performance of our proposed approach is shown in Table 1.

It is evident from Tables 1, 2 and 3 that our proposed model yielded high DR and low FAR to classify the attacks. For DOS attack our proposed model achieved an accuracy of 99.67%, which is 7% more than J48 algorithm. FAR recorded for J48 is more than our proposed model. For a good classifier to detect attacks it should have high DR and low FAR. After applying SU feature selection measure accuracy and DR has been increased and FAR is reduced.

Table 1. Performance Measure for Random Forest (No. of trees = 100).

SNO	Attack Type	Accuracy	DR	Far	MCC
1	DoS	99.67	99.84	0.00527	0.99
2	Probe	99.67	99.82	0.00502	0.99
3	R2L	99.67	99.82	0.00505	0.99
4	U2R	99.67	99.84	0.00552	0.99

Table 2. Performance Measure for J48 Tree.

SNO	Attack Type	Accuracy	DR	Far	MCC
1	DoS	99.25	99.3	0.00829	0.985
2	Probe	99.29	99.4	0.0092	0.985
3	R2L	99.24	99.3	0.008	0.99
4	U2R	99.28	99.3	0.0072	0.99

Table 3. After Applying FSS-Symmetric Uncertainty.

SNO	Attack Type	Accuracy	DR	Far	MCC
1	DoS	99.68	99.82	0.00477	0.99
2	Probe	99.63	99.73	0.00477	0.99
3	R2L	99.69	99.86	0.00502	0.99
4	U2R	99.68	99.82	0.00477	0.99

For a probe attack after applying feature selection DR is recorded as 99.73%. For R2L and U2R MCC has been recorded as 0.99 which shows our approach is good in classifying the attacks in IDS. Average accuracy obtained by our proposed approach without feature selection is 99.67%, where as for j48 it is recorded as 99.26% only. Mathews correlation coefficient recorded by our model is high compared with j48 classifier. The experimental result shows that our approach can achieve good accuracy, high DR with low FAR.

5. Conclusions

This paper deals the Random Forest (RF) algorithm to detect four types of attack like DOS, probe, U2R and R2L.We adopted 10 cross validation applied for classification. Feature selection is applied on the data set to reduce dimensionality and to remove redundant and irrelevant features. We applied symmetrical uncertainty of attributes which overcomes the problems of information gain. The proposed approach is evaluated using NSL KDD data set. We compared our random forest modelling with j48 classier in terms of accuracy, DR, FAR and MCC. Our experimental result prove that accuracy, DR and MCC for four types of attacks are increased by our proposed method. For future work, we will apply evolutionary computation as a feature selection measure to further improve accuracy of the classifier.

References

[1] J. P. Anderson, Computer Security Threat Monitoring and Surveillance, Technical Report, *James Anderson Report, Pennsylvania*, (1980).
[2] G. V. Nadiammai, S. Krishnaveni and M. Hemalatha, A Comprehensive Analysis and Study in IDS Using Data Mining Techniques, *IJCA*, vol. 35, pp. 51–56, November–December (2011).
[3] L. Breiman, Random Forests, Machine Learning, vol. 45, no. 1, pp. 5–32, (2001).
[4] Arif Jamal Malik, Waseem Shahzad and Farrukh Aslam Khan, Network Intrusion Detection Using Hybrid Binary PSO and Random Forests Algorithm, *Security and Communication Networks*, (2012).
[5] P. Natesan and P. Balasubramanie, Multi Stage Filter Using Enhanced Adaboost for Network IDS, *International Journal of Network Security and its Applications*, vol. 4, no. 3, (2012).
[6] Mrutyunjaya Panda, Ajith Abraham and Manas Ranjan Patra, A Hybrid Intelligent Approach for Network Intrusion Detection, *UCCTSD*, pp. 1–9, (2012).
[7] Md. Al Mehedi Hasan, Mohammed Nasser, Biprodip and Shamim Ahmad, Support Vector Machine and Random Forest Modeling for IDS, *JILSA*, pp. 45–52, (2014).
[8] Ujwala Ravale, Nilesh Marathe and Puja Padiya, Feature Selection Based Hybrid Anomaly Intrusion Detection System Using K Means and RBF Kernel Function, *ICACTA*, pp. 428–435, (2015).
[9] D. Powell and R. Stroud, Conceptual Model and Architecture, IBM Zurich Laboratory Research Report RZ 3377, November (2001).
[10] Aleksandar Lazarevic, Vipin Kumar and Jaideep Srivastava, Intrusion Detection: An Survey, p. 31.
[11] Araujo, Oliviera, Shinoda and Bhargava, Identifying Important Characteristics in the KDD99 Intrusion Detection Dataset by Feature Selection Using a Hybrid Approach, *International Conference on Telecommunications*, (2010).
[12] M. A. Jabbar and B. L. Deekshatulu, Priti Chandra, Alternating Decision Tree for Early Diagnosis of Heart Disease, *IEEE*, pp. 322–328, (2014).
[13] Jehad Ali, *et al.*, Random Forest and Decision Trees, *IJCSI*, vol. 9, no. 3, pp. 272–278, (2012).
[14] Kahled Fawagreh, Mohamded, Medhat Gaber and Eyad Elya, RF: From Early Developments to Recent Advancements, *System Science And Control Engineering*, 2:1, pp. 602–609, (2014).
[15] Yvan Saeys, Inaki Inza and Pedro Larranaga, A Review of Feature Selection Techniques in Bioinformatics, vol. 23, no. 19, pp. 2507–2517, (2007).