Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 115 (2017) 226–234

Procedia Computer Science

www.elsevier.com/locate/procedia

7th International Conference on Advances in Computing & Communications, ICACC-2017, 22-24 August 2017, Cochin, India

RFAODE: A Novel Ensemble Intrusion Detection System

M A Jabbar [a]* , Rajanikanth Aluvalu[b],Sai Satyanarayana Reddy S[c]

[a,b,c] *Vardhaman College of Engineering,Hyderabad,India*

Abstract

In recent years information and communication technology (ICT) has become an important part of human life. But ICT brings a lot of cyber risks. New threats and vulnerabilities are created to attack network system. Intrusion detection system (IDS) is used to detect these attacks. Machine learning (ML) and Data Mining (DM) techniques are widely used for IDS. Current IDS algorithms result in high error rate and less accurate to classify various attacks. This paper deals with a novel ensemble classifier (RFAODE) for intrusion detection system. Proposed ensemble classifier is built using two well-known algorithms RF and AODE. Average One-Dependence Estimator (AODE) resolved the attribute dependency issue in Naïve bayes classifier. Random Forest (RF) improves accuracy and reduces the error rate. The performance of proposed ensemble classifier (RF+AODE) is analyzed on Kyoto data set. With accuracy of 90.51% and FAR of 0.14, proposed ensemble classifier outperforms AODE, Naïve bayes, and RF algorithms and efficiently classifies the network traffic as normal or malicious.

Keywords: *Intrusion Detection System;Ensemble Learning;Random Forest;AODE*

1. Introduction

ICT has become an important part of human life. Many organizations like Govt, Business, and Medical rely on ICT. With the tremendous use of ICT, information systems and network are prone to various attacks. New threats and vulnerabilities are created by individuals and organizations to attack network systems, which leads to cyber attacks.

* Corresponding author. Tel.: + 919912648686
E-mail address : jabbar.meerja@gmail.com

10.1016/j.procs.2017.09.129

To solve this problem, we require effective and efficient intrusion detection system (IDS) [1] to detect attacks and anomalies in the networks.IDS was first developed by Anderson in 1980[2].IDS are divided into two modes based on different data types and analysis methods. They are 1) Host-based 2) Network based. The former examines intrusions in individual computers and later detects intrusions in the networks. IDS can be considered as supervised learning problem [3]. An IDS is defined as the process of identifying intrusions, which violates security breach and reports the intrusions to the administrators. Basically, IDS performs the following actions [4].

1) To monitor and analyze the system and user activities

2) To audit vulnerabilities and system configuration

3) To assess the integrity of data files and critical systems

4) To analyze abnormal activities

5) To audit operating system.

Data mining techniques are widely used for IDS. A number of researches have been carried out with data mining for IDS. In data mining classification technique is widely applied to intrusion detection. Current algorithms and methods result in high false alarm rate. A high false alarm rate makes IDS ineffective.

Random forest is a tree-based ensemble classifier [5] which combines bagging and random selection of features to construct a number of decision trees. It is highly accurate with low classification error. Randomization is applied to select the best node. If A is the number of attributes in the dataset D, then randomization is equal to $\sqrt{A}$ [6].

Average one dependency estimator (AODE) is an accurate and multiclass classifier which enhances the accuracy. AODE is efficient in detecting network traffic as normal or attack. AODE is useful for large data sets.

Ensemble based learning combines multiple classifiers to classify unknown instances and to enhance the accuracy. Ensemble based learning perform well when data is huge and very scarce. These techniques are well suited for IDS [7].

In this paper, we propose novel IDS which combine RF and AODE to enhance attack detection. Section 2 discusses the related work. Our novel IDS based on RF and AODE is discussed in section 3.Results are presented in section 4 and conclusion is given in the last section.

2. Related work

Ensemble methods are well suited for IDS. This section reviews literature related to IDS which uses data mining and machine learning approaches.

A novel ensemble classifier for IDS was proposed in [8].Authors proposed ensemble IDS using AD Tree and naïve bayes. NSL-KDD data set is used for experimental analysis. Discretization and IQR techniques are used for preprocessing. Fuzzy based ensemble classifier for IDS was proposed by Kumar and selvaumar [9].Authors applied preprocessing on DARPA and CAIDA data sets which are used for experimental analysis.

Borji Ali proposed a network IDS which combines heterogeneous classifiers [10].Authors combined ANN, K-NN, DT and SVM classifier and tested on DARPA98 datasets. DR and FAR metrics are used to evaluate the performance of their approach.

Perdisci et.al [11] proposed an ensemble based on SVM for IDS. To reduce the dimensionality of data sets, authors adopted clustering algorithm. DARPA and GATECH data sets are used for experimental analysis.

Cluster-based ensemble classifier for IDS was proposed by Jabbar et.al[12]. ADTree and K-NN classifiers are combined to detect intrusions. K-Means clustering algorithm is applied on the Gure data set which is used for simulation. DR, FAR, Hubers Index, Rand index, and accuracy metrics are used to evaluate the method. Summary of related work is shown in Table 1.

Table 1: Summary of earlier work on ensemble IDS.

S.No	Author	Method	Data set used	Metrics used
1	Jabbar	ADTree and Naïve bayes	NSL-KDD	DR,FAR, Accuracy
2	Kumar and selvakumar	Fuzzy based	DARPA and CAIDA	Accuracy
3	Borji Ali	ANN,K-NN,DT and SVM	DARPA98	DR and FAR
4	Perdisci	SVM based	DARPA and GATECH	Accuracy
5	Jabbar	ADTree and K-NN	GURE Data set	DR,FAR,Hubers Index ,Rand index and accuracy

Motivated from above literature, in this research paper we propose a novel ensemble classifier based on RF and AODE for intrusion detection.

3. Methodology

As stated earlier, goal of this research paper is to develop a novel ensemble classifier for IDS. This section will discuss AODE, RF, and our proposed methodology in brief. We used Kyoto data set which is classified into training and testing.

3.1. Random Forest(RF)

In recent times Random forest, (now onwards we denote as RF) have been used widely for IDS problem. RF combines bagging and random selection of features. RF consists of many classification trees. RF improves the accuracy and reduces error rate for large data sets. RF generates out of bag error during training phase. Three Turing parameters are used in RF: No of trees, minimum node size, numbers of descriptors are used for splitting each node.

Steps in Random forest are as follows

```
-------------------------------------------------------------------------------------------
Steps in Random Forest
-------------------------------------------------------------------------------------------
Step 1: From the training set, select a bootstrap sample
Step 2: On this bootstrap sample, grow an unpruned tree
Step 3: Randomly select predictors at each node and determine the best split.
Step 4: Save tree as it is and don’t apply cost complexity.
-------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------
```

3.2. Average One Dependency Estimator(AODE)

Average one dependency estimator resolved the attribute independence issue in naive bayes. As naïve bayes classifier does not consider attribute interdependency, this may affect the accuracy of the IDS. High computational overheads of naïve bayes and augmented naïve bayes are overcome by AODE. AODE is capable of accurately predicting whether network traffic is normal or anomalous [13].Structure of AODE is shown in figure 1.

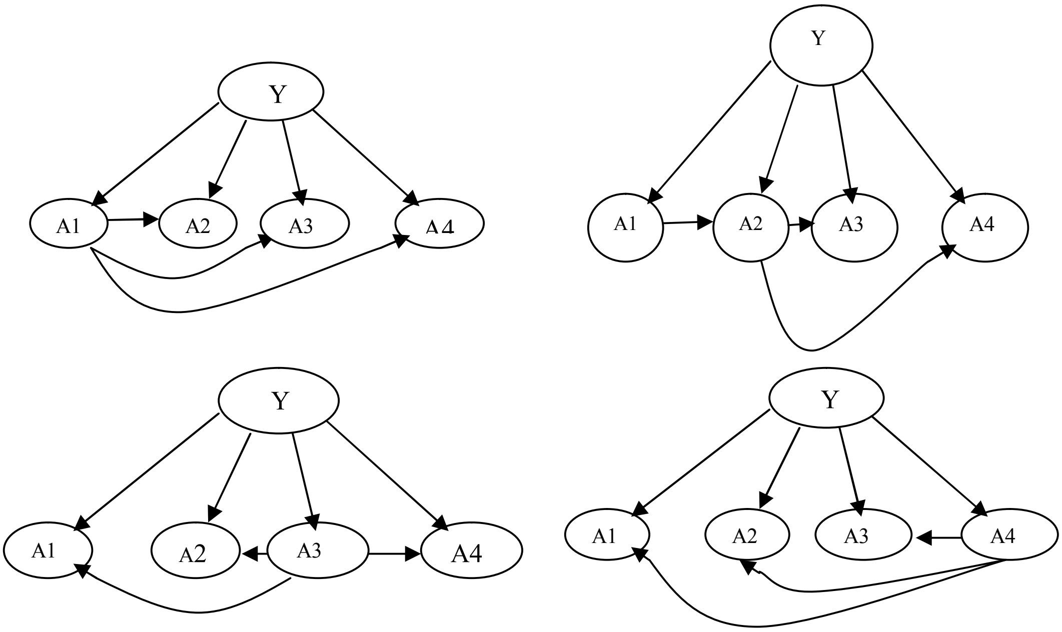

Fig. 1 structure of AODE [13]

AODE achieves high accuracy by averaging the aggregation of many special tree augmented naïve bayes. AODE has a low variance and supports incremental learning and can effectively handle missing values. It has classification time of O (cn^2) and has a computational complexity of O (tn^2).

3.3. Ensemble Learning

Ensemble learning is a new trend in AI and data mining, in which several weak learning algorithms are combined [14]. Idea behind ensemble classification is to exploit the strength of weak learning algorithms to obtain a robust/efficient classifier. A single IDS developed with weak learning algorithm can cover and identify limited input data and no. of attacks [15].

Ensemble classifiers are constructed by a set of weak classifiers and decision function which combines the classification results. Majority voting is simple and efficient decision function used in many ensemble techniques.

3.4. Proposed Approach(RFAODE)

In this subsection, we will discuss our proposed RFAODE: A novel ensemble IDS using RF and AODE.
The proposed approach operates in three stages.
1) Data preprocessing (For Kyoto Data set)
2) Training and evaluating ensemble classifier
3) Testing
Prior to applying ensemble classifier, it is essential to convert the features to a format which are perceivable by the ensemble classifier. In our proposed approach, features are converted from numeric to binary. Our proposed ensemble IDS will operate on this data set for classification of network traffic data. Inter quartile range (IQR) is used to remove noise and outliers in the data set. Variability is summarized by IQR.

Steps in our proposed approach are as follows

RFAODE: A Novel Ensemble Intrusion Detection System

```
Step 1: Kyoto data set is loaded
Step 2: Apply for preprocessing
  a) Convert the features from numeric to binary.
Step 3: Build Ensemble classifier (RF +AODE)
Step 4: Consider the Average of Probabilities
Step 5: Classification of network attacks as normal or malicious.
```

Figure 2 represents flow in our proposed approach. In the third stage RF and AODE classifiers are integrated to enhance the classification accuracy.

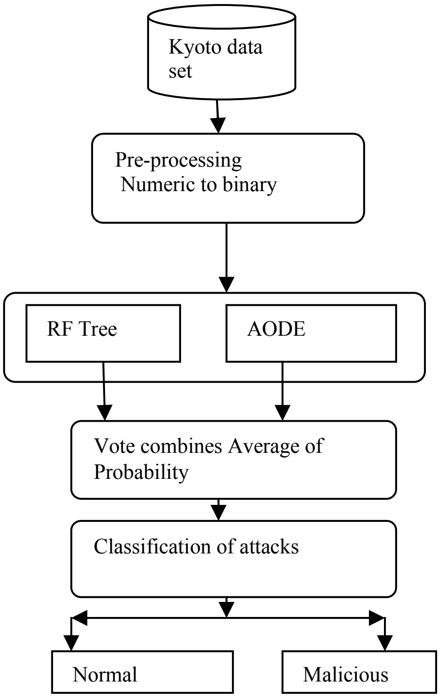

Figure 2: Proposed model (RFAODE)

After pre-processing, the data set is used for training using ensemble classifier built using AODE and RF. Average of probability combination rule is applied to test the class of a sample. During the testing phase, instances of the Kyoto data set are introduced to proposed ensemble (RFAODE) classifier by hiding their class to which they belong. This ensemble classifier predicts the networks traffic data as normal or malicious.

4. Experimental Analysis

This section analyzes the result obtained from the experiments performed to test the performance of our proposed ensemble classifier. For the experimental analysis, we used Kyoto data set which is refined version for traffic data. Description about Kyoto data set is shown in table 2. Kyoto data set (2006) consists of 24 features, but we used only 15 features and excluded the features which are related to security analysis. Kyoto data set is presented by Song et.al [16].

Table 2: Kyoto data set description

Number of instances :85506 Number of features :15	**Features :** 1. Duration 2. Service 3. Source_bytes 4. Destination_bytes 5. Count 6. Same_srv_rate 7. Serror_rate 8 Srv_serror_rate 9.Dst_host_count 10. Dst_host_srv_count 11. st_host_same_src_port_rate 12. Dst_host_serror_rate 13. Dst_host_srv_serror_rate 14.Flag 15. Label

In our analysis we used RF and AODE algorithm to build ensemble classifier. We build the ensemble classifier using WEKA machine learning tool [17].10 fold cross validation is used for testing the classifier.10 Trees are used to build the RF. Following evaluation indices are defined from the confusion matrix. The confusion matrix shows the distribution of instances that are either attack or normal.

		Predicted class	
		Attack	Normal
Actual class	Attack	TP	FN
	Normal	FP	TN

Where
TN – Number of instances correctly predicted as non-attacks.

FN – Number of instances wrongly predicted as non-attacks.

FP – Number of instances wrongly predicted as attacks.

TP – Number of instances correctly predicted as attacks.

FN and FP cause a problem for IDS. FN represents an attack which is classified as normal by IDS.

Evaluation indices are defined as

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

2) False alarm rate =FP/ (FP+TN)

3) 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 DR= TP/TP+FN

4) Hubert Index (HI) = (TP+TN)-(FP+FN)/ (TP+TN+FP+FN)

Hubert Index is defined as the difference of agreeing pairs to disagreeing pairs. Performance of our proposed method is shown in table 3.

Table 3: Accuracy of various approaches on Kyoto data set

Sl.no	Approach	Accuracy	DR	FAR	HI
1	RFAODE	90.51	92.38	0.14	80

Values for DR, FAR, FAR, HI for various approaches are recorded in table 4.

Table 4: Evaluation Indices for four 4 approaches for Kyoto data set.

Sl no	Approach	Accuracy	DR	FAR	HI
1	AODE	89.68	68.8	0.15	79
2	RF	89.34	68.7	0.17	78
3	RF-AODE	90.51	92.38	0.14	80
4	Naive bayes	82.5	91	0.15	79

We compared our approach with other classification algorithms. Comparisons of accuracy, DR, HI for various approaches are recorded in table 4, table 5 and figure 3.

Table 5: Accuracy comparison of various approaches for Kyoto data set

Sl no	Approach	Accuracy
1	Naïve bayes	82.5
2	AODE	89.68
3	RF	89.34
4	Proposed approach (RFAODE)	90.51

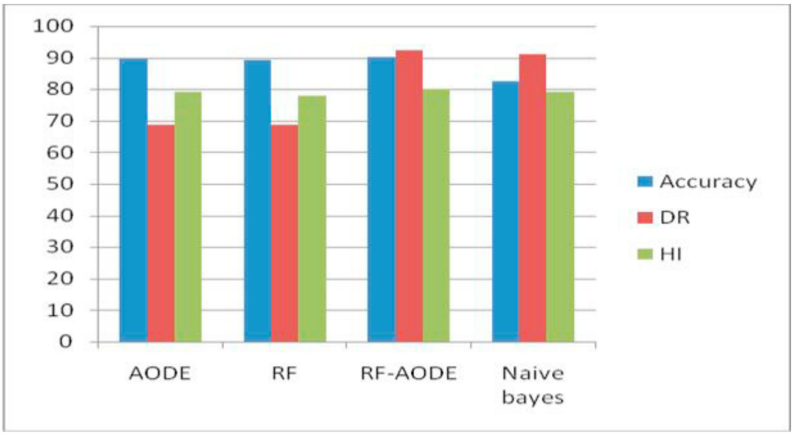

Fig 3: comparison of various metrics (accuracy, DR, HI)

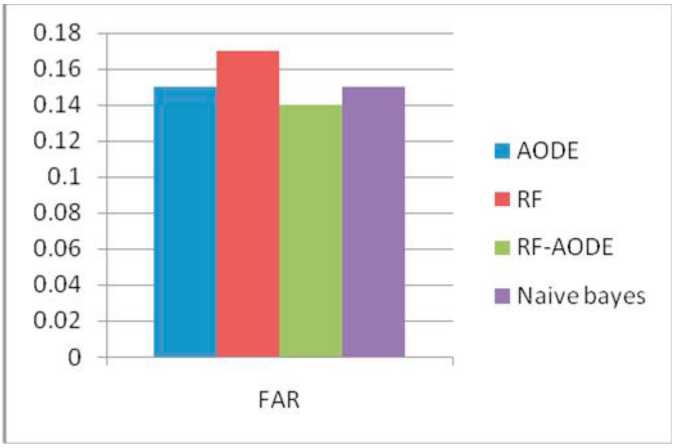

Fig 4: Comparison of FAR metric for various approaches.

False Error rate (FAR) values for AODE, RF, Naive bayes and Ensemble classifier RF-AODE are shown in figure 4. FAR of our approach is recorded as 0.14, whereas for RF and AODE it is recorded as 0.17 and 0.15 respectively. FAR values for the proposed ensemble approach is very less when compared with other algorithms. HI values for naive bayes is 79. For Kyoto data set our proposed ensemble classifier out performs naïve bayes, AODE, RF with good classification accuracy for attack detection for network traffic data. Naïve bayes recorded the accuracy as 82.5% only whereas AODE recorded as 89.68% and our proposed approach recorded accuracy as 90.51%. It is evident from the results obtained from the experimental analysis that our proposed approach performs well for Intrusion detection system and overcomes the shortfalls of Naïve bayes and Random forest algorithms. Proposed approach recorded high detection rate with low false alarm rate. Combining RF and AODE approaches proved that proposed approach is successful for IDS. Proposed approach can be used to detect network intrusions and classify traffic data as normal or malicious.

5. Conclusion

In this research paper, we proposed a novel ensemble classifier (RFAODE) for intrusion detection system. The proposed approach efficiently classifies network traffic as normal or malicious. The results indicate that proposed classifier is accurate than naïve bayes, J48, PART classifiers. We considered Kyoto data set for experimental analysis. As Base classifiers are not capable of detecting the attacks accurately, proposed Ensemble classifier outperforms base classifiers RF and AODE. The results presented in this paper show that integration of RF, AODE and pre-processing technique will yield the good result for IDS. For future work, we will apply evolutionary approaches for IDS to classify network traffic data on various data sets.

References

[1] Long –Sheng et.al"Feature extraction based approaches for improving the performance of intrusion detection systems",IMECS2015,volume I,pp1-6(2015)

[2] J.P Anderson, "Computer Security Threat Monitoring and Surveillance", Technical report, James Anderson Report, Pennsylvania(1980)

[3] Amreen Sultana et.al,"Intelligent network intrusion detection system using data mining techniques", pp327-331,IEEE(2016)

[4] Shveta et.al,"Applying genetic algorithm in intrusion detection system: A comprehensive review",ACEEE,pp102-112(2014)

[5] L.Breiman,"Random Forest", A machine learning,vol 5,no 1,pp5-32(2011)

[6] Kahled Fawagreh,"Random forest from early development to recent advancements", System science and control engineering",Vol2(1),pp 602-609(2014)

[7] Gianluigofolino, "Ensemble based collaborative and distributed intrusion detection system",JNCA,Vol66,pp1-16(2016)

[8] M.A.Jabbar et.al,"A novel Intelligent Ensemble Classifier for Network Intrusion Detection System",SOCPAR Springer (2016)

[9] Kumar P Arunraj Selvakumar, "Detection of DDOS attacks using an ensemble of adaptive and hybrid neuro fuzzy systems", Computer communication36(3),pp303-309(2013)

[10] Borji ,Ali," Combining heterogeneous classifiers for network intrusion detection",LNCS,Vol4846,pp 254-260(2007)

[11] Perdisci Roberto et.al, "A multiple classifier system for accurate payload based anomaly detection", Com.Network (6),864-81(2009)

[12] M.A.Jabbar et.al,"Cluster based Ensemble classification for Intrusion Detection System", ACM, *ICMLC 2017,* February 24-26, 2017, Singapore, Singapore, pp253-257

[13] Jia Wu et.al,"Attribute weighting: How and when does it work for Bayesian network classification ",IJCNN2014, pp 4076-4083(2014)

[14] Alexander Balon-Perin",Ensemble based methods for IDS",NTNU,Trandheim(2012)

[15] Christopher et.al ",IDS and Correlation", Challenges and solutions",Vol14,AISC,Springer(2005)

[16] http://www.takakura.com/Kyoto_data

[17] http://www.cs.waikato.ac.nz/ml/weka/