Cluster Based Ensemble Classification for Intrusion Detection System

M. A. Jabbar, Rajanikanth Aluvalu, S. Sai Satyanarayana Reddy
Vardhaman College of Engineering, Hyderabad, Telangana, India
[1]jabbar.meerja@gmail.com, [2]rajanik.rkcet@gmail.com, [3]principal@vardhaman.org

ABSTRACT

Network security is a challenging task, as there is a tremendous growth of network -based services and sharing of sensitive information on the network. Intrusion throws a serious risk in the network. Even though many hardening systems are developed against intrusions, conventional approaches like firewalls, virtual private networks, and encryption techniques are not enough to provide network security. Intrusion detection is a cyber security mechanism that plays an important role in securing the network. Detection rate and false alarm rate are the challenging issues to design an Intrusion Detection System (IDS). Various data mining techniques are used to implement network intrusion detection. Classification is a supervised learning which predicts the class label in the data set. Single classifier fails to obtain high accuracy. Base classifiers are not capable of detecting the attacks accurately. Ensemble classifier is a combination of base classifiers. Ensemble classifier outperforms base classifiers. In this paper, we propose a cluster- based ensemble classifier for IDS. K means clustering is used in the experiment. Ensemble classifier is built using ADTree and KNN. The experimental results show that the proposed ensemble classifier outperforms other classifiers with 99.8% accuracy.

CCS Concepts

• Informationsystems~Clustering • Security and privacy~Intrusion detection systems

Keywords

Data mining; intrusion detection; ensemble classifier; network security; clustering.

1. INTRODUCTION

Due to the high usage of Internet and sharing of information on the web, security of the network becomes a challenging task. Intrusion is defined as an illegal attempt to access the network. Even though routers and firewalls protect the system in some way in the network, they can't detect the intrusion. Intrusion detection is to detect malicious activities in the network.

The main aim of the IDS is to monitor the network and automatically detect the anomalies. In recent years, many data mining (DM) algorithms have been developed to design IDS. Classification is a supervised learning, which predicts the class label in the data set. [1]. Base classifiers are not capable of detecting the attacks accurately. Ensemble classifier is a combination of base classifiers.

In Ensemble classifiers, the output of base classifiers is used as predictors to improve the accuracy. Ensemble classifier can produce better accuracy than base classifiers. [2]. Clustering is applied before classification due to the following reasons. If you have a dataset A, and you apply clustering to obtain clusters B and C. Then, you can search for a classification model (even different ones) for B and C, which might be more precise than the original for A. The case for doing a cluster analysis (or other dimensionality reduction methods such as PCA) would probably be to reduce the number of features in a way that the learning model is more robust.

K-means clustering is a partitioning technique in which clusters are formed with the help of centroid. K-means clustering is widely used in many applications. K-means clustering is a phase process: the first phase is to define k centroid, one for each cluster and in the second phase each point belonging to the given data set is collected and associates it to the nearest centroid.

An alternating decision tree is a new type of classification rule. It is a generalization of DT, voted decision stumps and VDT [3]. The structure of an ADTree represents decision paths.

K-Nearest neighbor (KNN) is a supervised lazy learning algorithm which can incorporate different kinds of data types. KNN is a nonparametric classification method. KNN is used in many applications like pattern recognition [4].

This paper deals with classification of IDS attacks using Ensemble classifiers using ADtree and KNN. K-means clustering is applied before applying ensemble classifier. Gure data set is chosen for simulation.

The remaining part of the paper is ordered in this way. Section 2 is dedicated to related work. Section 3 introduces methodology and Section 4 discuss about our proposed work. Experimental results are discussed in Section 5. We conclude our remarks and gives directions in Section 6.

2. RELATED WORK

Ensemble techniques are widely used for IDS. This section discusses related work on IDS using ensemble in short.

Amreen sultan et al. proposed intelligent network intrusion detection system using data mining techniques. Their approach is based on AODE algorithm [5]. NSL KDD data set is used for experiment analysis. Proposed approach classifies the four different types of attacks. Detection Rate (DR) of their proposed approach for DOS is recorded as 97.19%.

ICMLC 2017, February 24-26, 2017, Singapore, Singapore
© 2017 ACM. ISBN 978-1-4503-4817-1/17/02…$15.00
DOI: http://dx.doi.org/10.1145/3055635.3056595

Nabila farnaaz et al. [6] proposed random forest based intrusion detection system. Random forest is used as a base classifier for classification of various IDS attacks. Symmetrical uncertainty is used as a feature selection measure to filter the data set. NSL KDD Data set is used for experimental analysis. Their method classifies four types of IDS attacks. After applying feature selection, DR of proposed approach is 99.68%.

A Novel Intelligent Ensemble Classifier for Network Intrusion Detection System is proposed by Jabbar et al. [7]. Authors proposed ensemble classifier with ADtree with naive Bayes. NSL KDD data set is used for experimental analysis. The interquartile range is used as a feature selection measure to improve accuracy. DR recorded by the proposed approach for U2R and R2L is 91.41%.

Identification of anomalies using outlier detection is proposed by Jabez j et al. [8]. Neighborhood Outlier Factor (NOF) technique is used to detect the outliers. KDD Dataset is used for experimental analysis.10 attributes are used in the data set. Authors compared their approach with machine learning algorithms and claimed that the proposed work is superior than other machine learning algorithms.

The ensemble of clustering algorithm for anomaly detection is proposed by Salima et al. [2]. KM -GSA, KM-PSO, FCM classifiers are used as ensemble classifier.KDD Dataset is used for simulation. Proposed method classifies four types of attacks. Average accuracy recorded by the method is 94.47%.

Morteza analoui et.al proposed two tire ensemble classifier for IDS. Authors used KDD CUP 99 Dataset. Three classifiers are used to build the ensemble classifier. [9]. Their proposed approach performs well compared with base classifiers.

Anomaly intrusion detection system using ensemble classifiers was proposed by Debojit boro et al. [10]. Decision tree, naive Bayes, and decision table are used to design ensemble classifier. TUIDS port scan data set is used for simulation. Proposed method opted a specific class by Weighted Majority Voting (WMV).

Motivated from the literature, we proposed an ensemble classifier for intrusion detection system.

The main contributions of this paper are highlighted as follows.

1) We propose a cluster based ensemble classifier for intrusion detection system
2) Our method adopts a new technique which classifies intrusion attacks.
3) We tested the proposed approach on Gure data set.

3. METHODOLOGY

This section will discuss the methodology adopted by our proposed methodology. As stated earlier, the aim of this research paper is to develop a cluster based novel ensemble classifier based on ADTree and KNN for classification of IDS to enhance accuracy. We used Gure data set, which is partitioned into training and test set.

3.1 K-Means Clustering

K-means clustering algorithm was proposed by Macqueen in 1967. This is one of the most frequently used partitioning techniques. K-means uses K-centroids to define k-clusters. In K-means we measure similarity based on the mean value of the objects of the cluster. We calculate the distance between centroid and each cluster object by using Euclidean distance as a measure. Finding the optimal centroid is the key problem in k-means. The optimal centroid is identified by implementing k-means regressively. Algorithm for k means is as follows.

Algorithm:

Step 1. Choose k data-items from Database as initial centroids.

Step 2. Repeat the process

i) Assign each item to the cluster which has the closest centroid;

ii) For each cluster calculate new mean

Step 3: Until convergence criteria is met.

3.2 Alternating Decision Tree (ADTree)

A decision tree is used in machine learning and data mining to support decision process. It is a tree-like structure, which consists of the root node, internal nodes, and terminal nodes. Alternating decision trees (ADTree) are machine learning methods used for classification, which combines DT and boosting [11]. ADTree enhances the accuracy of the DT. In addition to the classification, ADTree gives a measure of confidence called classification margin. An important feature of ADTree is the ability to merge together [12]. Unlike DT, decision node in ADTree is replaced by two types of nodes namely i) Prediction node ii) splitter node. Figure 1 shows the DT and ADTree.

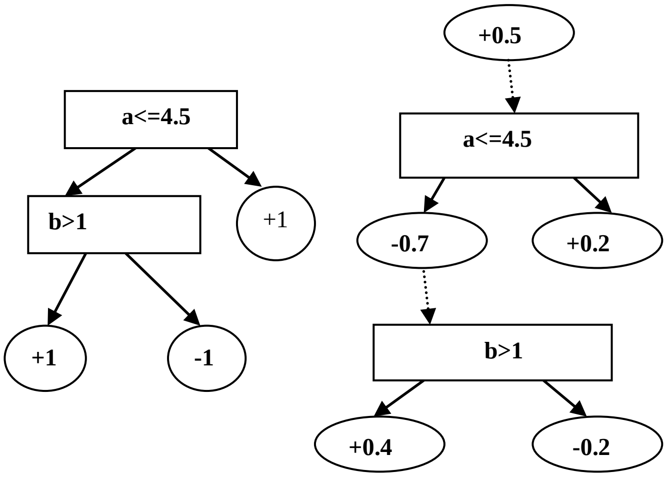

Figure 1. Decision tree and its equivalent ADTree [13]

3.3 K-Nearest Neighbor Algorithm (K-NN)

K-nearest neighbor (KNN) is one of the most widely used straight forward lazy learning classification problems. KNN classification method follows two phases

Phase 1: Find the k instances in the data set D that are closest to sample
Phase 2: These k instances then vote to determine the class of sample.

Figure 2 represents k nearest neighbor classification.

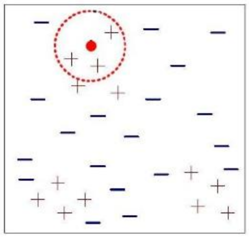

Figure 2. Identification of neighbors in KNN classifier

KNN works by estimating a fixed number of observations, which are nearer to the desired sample. For classification, knn selects most frequent neighbor, KNN suffers from the curse of dimensionality and overfitting.

3.4 Ensemble Classifier

Ensemble classifiers are introduced in the late 80s. Ensemble classifiers will obtain high accuracy if weak learners are integrated [14]. Ensemble classifiers construct the set of base classifiers and classify new sample by taking weight or vote of their prediction. Statistical, Computational, Representational are the three reasons for the success of ensemble classifiers. An advantage of combining complementary and redundant classifiers is to increase robustness and accuracy [15].

Several ensemble methods have been proposed, including mean combiner, median combiner, max combiner, majority voting, and weighted majority voting (WMV). While individual classifiers can be combined using any one of these methods, WMV is by far the most popular among them partly because of its conceptual simplicity, intuitiveness, and its effectiveness in practice [16].

3.5 Gure Dataset

Gure data set contains connections of kddcup99 but with payload to each connection. We used Gure KDD cup 6 percent data which consists of 41 attributes and one class attribute. These attributes are divided into 4 groups viz 1) intrinsic attributes 2) content attributes 3) Traffic attributes. The class attribute indicates whether the instance is normal or attack. More information on Gure data set can be found at [17].

4. PROPOSED APPROACH

As stated earlier, the aim of this paper is to propose cluster based ensemble classifier for intrusion detection system. The data set is preprocessed to remove noise and redundancy. Further, the data set is clustered using k means clustering. Two clusters are formed after applying the k means. Adtree and KNN are used to build the ensemble classifier; lastly ensemble classifier is tested using various performance measures.Steps in our proposed method are as follows.

Algorithm: Cluster based ensemble classifier for IDS

Step 1: Load the Gure data set

Step 2: Apply preprocessing to the data set

Step 3: Apply K means clustering

Step 4: Build the ensemble classifier using ADTree and KNN

Step 5: Classify the instance as normal or attack.

Loading and preprocessing will be done in Step 1 and 2. K means clustering is applied in step 3 and ensemble classifier is built using step 4. Algorithm performance is measured in step 5. Figure 3represents flow in our proposed approach. The ensemble classification is a way to build different types of base classifiers to solve the classification problem. The outputs of ADTree and KNN are used as predictors for Gure Data set and are combined to improve the accuracy of the IDS.

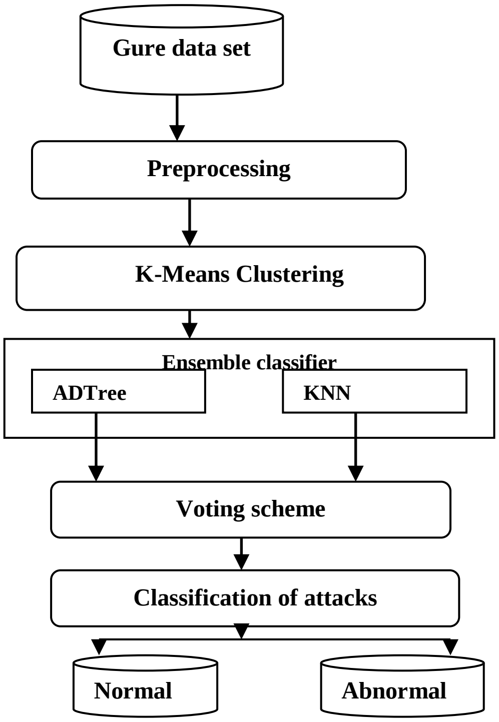

Figure 3. Flow in our proposed approach

The ensemble classifier is a linear combination of ADTree and KNN. Base classifiers ADTree and KNN are constructed from training dataset and classification is performed by taking a vote on the predictions made by each classifier.

5. EXPERIMENTAL RESULTS

Experiments for proposed approach were conducted using Gure data set. Gure data set is partitioned into training and testing. Training data set is used to build ensemble classifier, and test data set is used to test the instances. 10-fold cross validation is applied. The proposed approach is evaluated using the following metrics.

1) Detection Rate (DR) 2) False alarm rate 3) Accuracy 4) Rand index 5) Hubert's index. All the measures are based on confusion matrix. The confusion matrix is shown in Table 1.

Table 1. Confusion matrix

	Classified as Normal	Classified as Attack
Normal	TP	FP
Attack	FN	TN

where

TN –Instances correctly predicted as non-attacks.

FN – Instances wrongly predicted as non-attacks.

FP –Instances wrongly predicted as attacks.

TP –Instances correctly predicted as attacks.

Various metrics are defined as

1) Accuracy = $\frac{\text{Number of samples correctly classified in test data}}{\text{Total number of samples in test data}}$

2) Detection Rate (DR) = $\frac{TP}{(TP+FN)}$

3) False Alarm Rate (FAR) = $\underline{\quad FP \quad}$

(FP+TN)

4) Rand Index (RI) = TP+TN/(TP+TN+FP+FN)

RI is defined as the ratio of agreeing pairs and all possible pairs [18].

5) Hubert's Index (HI):= (TP+TN)-(FP+FN)/(TP+TN+FP+FN).

HI is defined as the difference of the agreeing pairs and disagreeing pairs [19]

Results of proposed ensemble classifier are listed in Table 2.

Table 2. Results of proposed ensemble classifier (ADtree+KNN)

Sl.no	Metric	Value
1	Detection Rate	99.8
2	False alarm rate	0.0003
3	Huberts index	99.8
4	Rand index	99.9
5	Accuracy	99.93

Table 2 shows that proposed approach records DR of 99.8% and false alarm rate of 0.0003. An intrusion detection system should have high detection rate and low false alarm rate. From the Table 2, it is evident that the proposed system has high DT of 99.8% and a very low false alarm rate of 0.0003. Accuracy and Rand index are recorded as 99.9%, which is high when compared with other existing techniques.

From the Huberts index, we will learn that difference between agreeing pairs and disagreeing pairs is high, it show that the proposed system effectively identifies the normal and malign attacks. Comparisons of proposed approach with other approaches are listed in Table 3.

Table 3. Detection rate and false alarm rate comparison

Sl no	Author name	Detection rate	False alarm rate
1	Warusia Yassin[20]	98.8	2.2
2	S S Sivatha[21]	98.38	1.62
3	Mohammad Sazzadul Hoque[22]	95	5
4	Proposed approach	99.8	0.0003

From the Table 3, it has been observed that our approach outperforms other existing approaches in terms of detection rate (DR).

False alarm rate for Warusia Yassin approach is 2.2, where as for S S Sivatha it is recorded as 1.62. Mohammad Sazzadul Hoque approach has a very high false alarm rate of 5%. Our approach records false alarm rate of 0.0003, which is very less. Good IDS system should have low false alarm rate in order to effectively identify the intrusion attacks.

6. CONCLUSION

In this paper, we proposed a cluster-based ensemble classifier for intrusion detection system. Use of conventional approaches is not enough to provide network security. Dynamic approaches like IDS are widely used to enhance security. Data mining and machine learning techniques are widely used for IDS. Proposed ensemble approach is build using ADTree and KNN classifiers. ADtree is a combination of DT and boosting, where as KNN is a lazy learning approach. Combining these two approaches proved that proposed approach is successful for IDS. Proposed approach recorded high detection rate with low false alarm rate. In future, our work will concentrate combining IDS with evolutionary techniques.

7. REFERENCES

[1] M. A. Jabbar, deekshatulu B L, Priti Chandra," Prediction of risk score for heart disease using associative classification and hybrid feature subset selection," IEEE, ISDA2012, pp. 628-634,(2012)

[2] Salima Benqdara et al," Ensemble of Clustering Algorithms for Anomaly Intrusion Detection System," JATIT, Vol 70, No 3, (2014)

[3] MA Jabbar, BL Deekshatulu, P Chndra," Alternating decision trees for early diagnosis of heart disease," In. Conf IEEE I4C, 2014, pp. 322-328 (2014)

[4] BL Deekshatulu, P Chandra, "Classification of heart disease using k-nearest neighbor and genetic algorithm," Elsevier Procedia Technology, Vol 10, pp. 85-94 (2013)

[5] Amreen sultana, M A Jabbar, "Intelligent network intrusion detection system using data mining techniques," IEEE, pp. 327-331 (2016)

[6] N Farnaaz, "Random Forest Modeling for Network Intrusion Detection System," Procedia Computer Science, Vol 89, pp 213-217 (2016)

[7] MA Jabbar, "A Novel Intelligent Ensemble Classifier for Network Intrusion Detection System," To appear in Scopar 2016 India, Springer, (2016)

[8] Jabez J, "Intrusion Detection System (Ids): Anomaly Detection Using Outlier Detection Approach," Procedia Computer Science, Vol 48, ICCC2015, pp. 338-346 (2015)

[9] Morteza analoui, "Hierarchical Two-Tier Ensemble Learning: A New Paradigm for Network Intrusion Detection," PIKM2007, ACM, pp. 33-39(2007)

[10] Debojit Boro et al, "Anomaly Based Intrusion Detection Using Meta Ensemble classifier," SIN2012, ACM, pp. 143-147 (2012)

[11] Freund, Y mason L, "The alternating Decision tree learning algorithm," In proc of 6th ICM Bled, Slovenia, pp. 124-133 (1999)

[12] Geoffrey holmes, "Multiclass alternation decision trees," University of Waikato, pp. 1-12, last accessed 27-09-2016

[13] Freund, Y mason L, "The alternating Decision tree learning algorithm," In proc of 6, ICM Bled, Slovenia, pp. 124-133 (1999)

[14] R. E. Schapire, "The strength of weak learnability," Machine learning, vol. 5, no. 2, pp. 197–227, Jul. 1990

[15] Anazida Zainal et al, "Ensemble Classifiers for Network Intrusion detection system," Journal of Information Assurance and Security, 4, pp. 217-225 (2009)

[16] Abdulla Amin Aburomman et al, "A novel SVM-KNN-PSO ensemble method for intrusion detection system," Applied Soft Computing Journal, pp. 1-13 (2015)

[17] GureKDDCup Dataset: http://www.sc.ehu.es/acwaldap/.

[18] Rand, W. M, "Objective criteria for the evaluation of clustering methods", Journal of American statistical association, " Vol. 66, pp. 846-850 (1971)

[19] L. Hubert, P Arabie, "Comparing partitions, Journal of classification, Vol. 2, pp. 193-218 (1985)

[20] Warusia Yasin et al, “Anomaly-Based Intrusion Detection through K means Clustering and Naives Bayes Classification,” in ICOCI2013, pp. 298-303 (2013)

[21] Siva S. Sivatha Sindhu et al, “Decision tree based light weight intrusion detection using a wrapper approach,”,Expert System with Applications, Vol. 39, pp. 129-141 (2012)