International Journal of
Science and Engineering Investigations vol. 6, issue 62, March 2017
ISSN: 2251-8843

An Ensemble Data Mining Approach for Intrusion Detection in a Computer Network

Abisola Ayomide Adeniran[1], Solomon Olalekan Akinola[2]
[1,2]Dept. of Computer Science, University of Ibadan, Nigeria
([1]bisorlaar@yahoo.com, [2]solom202@yahoo.co.uk)

Abstract- As activities being done on the internet keep expanding every day due to the fact that we are in the era of the information age, securing sensitive and crucial data on computer networks against malicious attacks tends to be a challenging issue. Designing effective Intrusion Detection Systems (IDSs) with maximized accuracy and low rate of false alarms is an imperative need in the world of cyber attacks. This work was designed to employ an ensemble data mining technique for improving IDSs by carrying out some experiments using the KDD 99 intrusion dataset. Dataset was fragmented into five, representing the major categories of attacks: Normal, DOS (Denial of Service), Probing (Information gathering), R2L (Remote to Local) and U2R (User to Root). An ensemble classifier using the Stacking method with the Naïve Bayes and Multilayer perceptron algorithms as the base classifiers and J48 as the meta learner was developed. The base classifiers were also employed on the dataset individually, and performance comparison was done between individual classifiers and the ensemble classifier. A 10-fold cross validation for training and testing of data and Gain ratio technique for filtering of the dataset was adopted. Ensemble classifier maximized accuracy the most and helped in reduction of false positives of the U2R attack type.

Keywords- *Intrusion Detection System, Ensemble, Stacking, Network Attacks, Data Mining*

I. INTRODUCTION

Organization and companies these days have adapted internet services as their means of communication in carrying out their business, which leads to the fact that computer network is growing at a very fast rate. Thus the significance of trying to secure networks against attacks proves to be critical. As network keeps growing, so is the rate at which network attacks are growing. Firewalls, antivirus, data encryption which are different ways of securing a network are traditional techniques in computer security. Most of these techniques are vulnerable to unauthorized use. Firewalls are easily prone to errors in terms of configuration [1]. Intrusion detection System (IDSs) tends to play a vital importance in network and computer security. An IDS tends to work as a form of security mechanism by identifying intrusions, unauthorized access and malicious attacks. It is designed for the process of securing, monitoring and analyzing data and events occurring in a computer or network system in order to detect signs of security violations and problems [2].

Intrusion detection systems are of two types: Host based and network based. Host based IDSs work by examining data held on individual systems that serves as hosts [3]. A network based Intrusion detection system works by monitoring and analyzing network traffic data to protect hosts from threats and attacks when traffic is being passed through the network. In this type of intrusion detection, data exchanged between computers are being observed by the IDSs [3].

There are two approaches to intrusion detection which are the Misuse and Anomaly detections. Misuse detection works by comparing attack signatures which are kept in the database, during searching through network traffic data. If there is a pattern that matches attack signatures in the database, alarms would be flagged by the IDS system [4]. A merit of the Misuse detection is that it detects well known attacks but unable to detect unknown or new attacks. On the other hand, the Anomaly detection works based on a profile of the system's normal pattern or behaviour that has been established. If there is a deviation from the established profile, alarms would be raised [4]. One merit of the Anomaly detection is that It detects novel or unknown attacks. Its shortcoming is the high rate of false alarms due to unseen system behaviour that might be categorized as intrusion by the system even when it is not. This limitation has necessitated the application of data mining to intrusion detection field by mining network traffic data which can help in differentiating intrusions from network traffic data. Data mining can help in building effective IDSs by improving accuracy in terms of detection and lowering the rate of false alarms [5].

The primary goal of this paper is to develop a hybrid intrusion detection classification algorithm. This was done using the stacking ensemble approach by ensembling Naive Bayes and Multilayer perceptron using gain ratio for data preprocessing in terms of improving IDSS based on classification accuracy in its detection rate and reduction of false alarm rate. Decision Tree (J48) was used as the meta learner. A comparison of the performances of both individual approaches to the ensemble approach was also carried out. The rest of this paper is organized as follows: Section 2 discusses related works, Section 3 discusses about the stacking ensemble

method, Section 4 gives description of our experiments, performance criteria we used, and other techniques we adopted in our experiments. Section 5 is about results obtained and Section 6 presents the conclusion.

II. RELATED WORKS

In the work of Hui Zaho [6], bagging was used as the ensemble method with support vector machine on the KDD 99 Dataset for intrusion detection. Neighborhood roughest was used for attribute reduction and particle swarm optimization for optimizing support vector parameters. Experimental result gave a detection rate of 93.17% and a false alarm rate of 0.81%.

Subbulakshimi *et al*. [7] carried out experiments using artificial neural network, support vector machine and Naïve Bayes on Schonlau and KDD datasets. These classifiers were applied individually, threshold values were set for different classes based on results of the single classifiers. They showed that radial basis and sigmoid kernel functions were better for anomaly detection and linear kernel function for misuse detection.

In Singh and Silakari's paper [8], hybridization of two feature selection methods was done for improving intrusion detection. The wrapper and filter method was hybridized, and K- Nearest Neighbor (KNN) classifier was used. Their model gave a good improvement in accuracy when records are huge and a slight improvement when records are small.

In the research carried out by Borji [9], he combined different classifiers in a new way and proposed a different way of combining heterogeneous classifiers. Darpa 1998 dataset was used. Artificial Neural Network, Support Vector Machine and C4.5 decision tree were applied individually and combined using three methods: the majority voting rule, Bayesian average and Belief. Experimental results showed that support vector machine performed best among other classifiers in the intrusion detection and also the Belief combination method outperformed the other two methods with a 0.87% false positive rate.

In the work of Govindarajan and Chandrasekaran [10] experiments were carried out on the NSL-KDD dataset. Radial basis function and Support Vector Machine classifiers were ensembled using Arcing, which is a generalization of the bagging and boosting ensemble approach. The authors used breadth first search for data preprocessing. Accuracy was only the performance metric used in the paper. Experimental results showed that the hybrid approach superseded the performance of the individual classifiers with an accuracy of 85.19%.

A. Ensemble Method Based on Stacking

Ensemble approach is a different way of combining a collection of classifies into a single or global classifier with the goal of increasing effectiveness better than a single classifier. Bagging, Boosting, Stacking are methods in the ensemble approach.

According to Iwan *et al.* [5], stacked generalization or Stacking for short, is a different way for combination of multiple models. Though developed years ago, it is not widely used compared to bagging and boosting. Bagging and Boosting are used for ensembling models of the same type, but Stacking is used to combine models of different types or different classifiers.

Stacking uses the meta learner to combine the output of the base classifiers in the best way. The base classifiers and the meta learner are involved in the stacking process. The predictions of the base classifiers are regarded as the meta model or level 0 model. The stacking learner model is referred to as level 1 model. Predictions from level 0 models would be fed as input to the level 1 model or meta learner which now decides how to combine the output in the best possible way for making the final prediction.

III. METHODOLOGY

This section gives insight into the intrusion dataset being used for carrying out the experiment, data preprocessing, criteria for performance evaluation and the experimental description

A. Dataset

Dataset supplied for running the simulation was the 10% of the KDD 99 dataset which consist of approximately 4,900,000 single connection vectors, each of which contains 41 features and one original feature that labels each record, which is the 42^{nd} label. According to Yimin [12], KDD 99 dataset has been the most widely used dataset for the evaluation of anomaly detection methods and can be found at http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

Dataset was fragmented into five which represents five major categories of attacks where each attack labels are grouped to their respective categories of attacks namely: Normal, DOS (Denial of Service), Probing (Information gathering), R2L (Remote to Local) and U2R (User to Root).

B. Data Preprocessing Based on Gain Ratio Technique

Data preprocessing is a crucial step in data mining before application of the desired data mining technique to be used. It is done before the dataset is being fed as input to the classifiers. Gain ratio technique was used in preprocessing of the dataset used in carrying out the experiment.

Gain ratio is a modification of the information gain to solve the issue of bias towards features with a larger set of values, exhibited by information gain. It should be large when data is evenly spread and small when all data belong to one branch attribute [11].

Gain ratio technique was applied to the dataset after been categorized into the attack categories for filtering attributes of the dataset that the base classifiers will use in making correct and accurate classification. Gain ratio filtered the attribute of the 41 features (plus the label feature) and the filtered dataset was used for classification. It ranked the attributes based on importance, though dimensionality reductions of the features were not done. Figure 1 shows the architecture of the intrusion detection model.

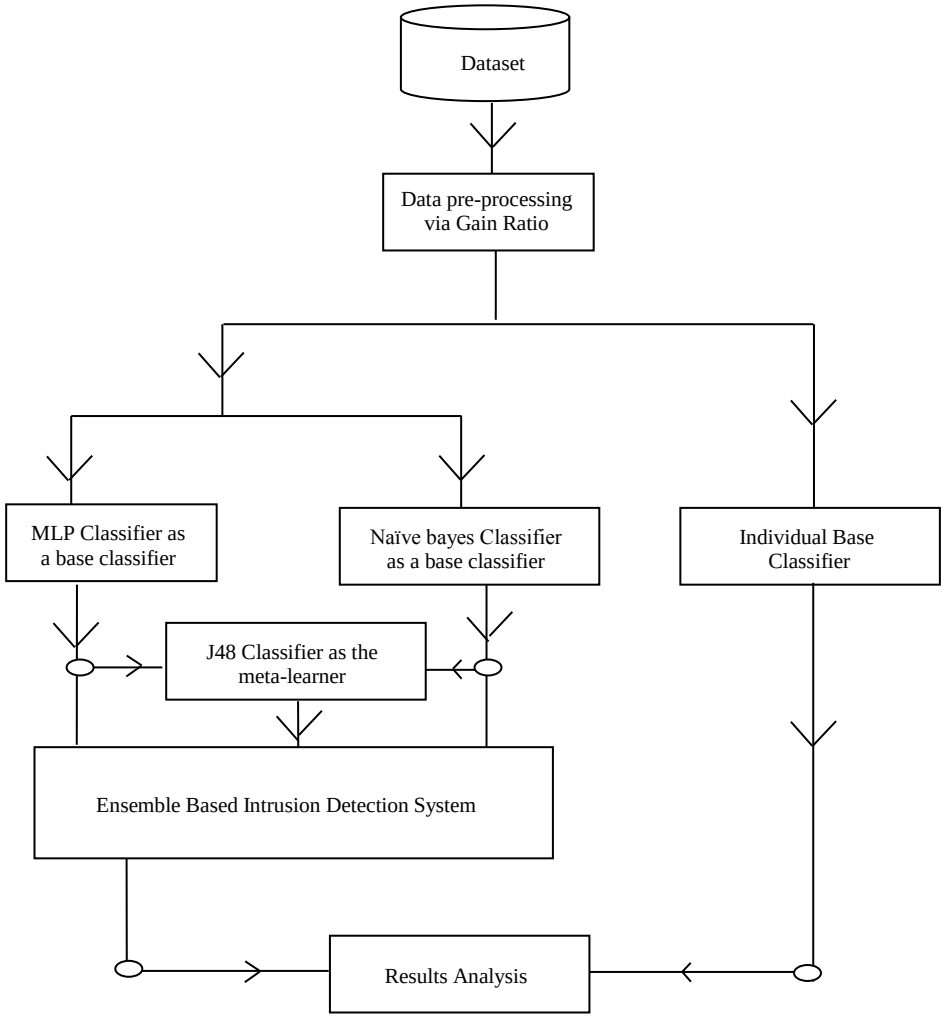

Figure 1. The Architecture of the intrusion detection model

From Figure 1, the dataset was passed through the pre-processing stage, which involved fragmenting the KDD 99 dataset into various categories of attacks, performing data filtering using gain ratio technique which is an advancement of information gain technique. After pre-processing, the dataset was classified using two classification algorithms using Naive Bayes and Multilayer perceptron classifier individually and ensembled using the Stacking approach. The two single classifiers, that is, the Multilayer perceptron and the Naive Bayes served as the base classifiers for the stacking method, their results fed as input to the J48 algorithm which stands as the meta learner. A 10-fold cross validation for training and testing the dataset was then applied. All experiments were carried out using Weka 3.6.9 on a 64-bit Windows 7 Professional Operating System with 4GB of RAM and an Intel core 'i5 CPU @ 2.60GHz.

C. Dataset

According to Yimin (2004) Kdd'99 has been the most widely used data set for the evaluation of anomaly detection methods. KDD99 data set is built based on the data captured in DARPA'98 IDS evaluation program (KDD, 1999). DARPA'98 is about 4 gigabytes of compressed raw (binary) TCP dump data of 7 weeks of network traffic. The two weeks of test data have around 2 million connection records.

The said dataset was fragmented into five which represents the categories of attacks namely: NORMAL, DOS (Denial of Service), PROBING (Information Gathering), R2L (Remote to Local) and U2R (User to Root).

D. Data Pre-Processing Based on Gain Ratio Technique

As it is known that data pre-processing is an important prerequisite before data evaluation and analysis is carried out, hence the use of gain ratio technique was applied as part of the pre-processing step in filtering features of the dataset which helped in filtering of the number of features that the base classifiers used in making correct and accurate classification. The feature selection technique filtered the attribute of the 41 features (plus the label feature) and the filtered dataset was used in classification.

E. Classification by Individual Base Classifier

The new dataset served as input to each base classifier i.e. Multilayer perceptron and Naïve Bayes machine learning algorithm and their individual performance recorded respectively.

F. Ensemble Based Intrusion Detection System

The feature extracted from the original KDD cup dataset based on gain ratio technique was fed as input to the base classifiers i.e. MLP and Naïve Bayes algorithms, and the training and testing of the stacking ensemble method was done using 10-fold cross validation technique.

The results of the base classifiers which is the level-0 model were supplied as input to the meta-learner (J48 algorithm) i.e. the level-1 model so that the meta-model can combine the inputs and make the final prediction.

G. Result Analysis

Conclusively, the results derived from both separate operation were subjected to analysis and comparison. Basically, the false positive and the accuracy of each operation were determined as they are the variables used to measure the performance of the algorithms and the ensemble technique.

H. Criteria for Performance Evaluation

The main criteria for performance evaluation used in this experiment are the accuracy and false positive rate; the accuracy gives the percentage of correctly classified instances and the false positive rate, also known as the false alarm rate, is the number of undetected attacks which are indeed normal attacks.

I. N-Fold Cross Validation

It is a mechanism in machine learning for training and testing an algorithm by dividing the dataset into random N partitions. The value for N for cross-validation mostly is always set to 10 which were adopted in this experiment. That is, first 9 partitioned dataset trains the algorithm, while the last group is used for testing which would be repeated 10 times.

IV. RESULTS

Table 1 and Figure 2 show the accuracies obtained with the classifiers on the different intrusion attacks.

TABLE I. THE ACCURACY OF EACH ALGORITHM

CLASSIFIER	ATTACK TYPES				
	DOS	NORMAL	PROBING	R2L	U2R
MLP	99.9930	100	99.1234	97.9574	78.8462
NAÏVE BAYES	99.8733	100	95.0816	75.222	69.2308
ENSEMBLE	99.9946	100	99.3182	98.0462	73.0769

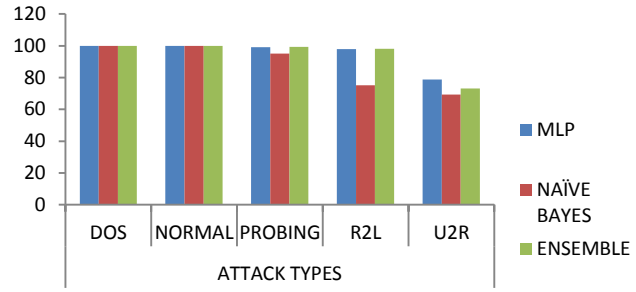

Figure 2. The classification accuracy

Table 2 and Figure 3 show the False Positive rates of each algorithm on the different intrusion attacks.

TABLE II. THE FALSE POSITIVE RATES

CLASSIFIER	ATTACK TYPES				
	DOS	NORMAL	PROBING	R2L	U2R
MLP	0	0	0.002	0.077	0.201
NAÏVE BAYES	0	0	0.020	0.045	0.174
ENSEMBLE	0	0	0.002	0.09	0.163

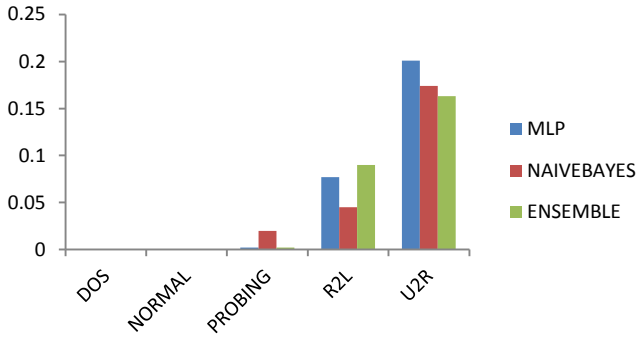

Figure 3. The false positive rate for each classifier

Evaluating the accuracy of each Classifiers and comparing them show that there is a positive improvement in the ensemble approach in terms of classification accuracy compared to other algorithms. The ensemble classifier significantly improved the detection accuracy, the effectiveness of the ensemble approach gives a more reliable result in terms of accuracy which is a very key factor in IDS. An IDS system should be able to have a high detection rate and detection should be accurate.

Where the Naïve Bayes detection of the U2R attack was low, the ensemble approach did improve it. Naive Bayes performed poorly the most in terms of detection accuracy compared to the other two classifiers, Multilayer Perceptron performed excellently well individually in comparison with the Naïve Bayes classifier, but the ensemble did outperform it in some other category of attacks. The ensemble classifier maximized accuracy detection.

Analyzing the single classifiers, Multilayer Perceptron out performed Naïve Bayes in terms of classification accuracy for all categories of attacks but in the Ensemble approach there is an improvement in detection. Ensemble approach did not perform well for U2R compared to other attacks due to the poor performance of Naïve Bayes on the said category which might be as a result of the size of the U2R attack category. The Multilayer Perceptron as a single classifier still outperformed the ensemble approach in detection of U2R attacks. Ensemble detection accuracy gives almost a 100% in all attacks except for the U2R type attack.

V. CONCLUSION

The field of intrusion detection has found a lot of researchers working tremendously on improving its standard and increases the accuracy of the detection system. For the improvement of the performance of an algorithm, another algorithm can be used in conjunction to remove its deficiency while performing its functions. Thus, this study was based on performing feature selection via gain ratio technique for filtering the given dataset (KDD Cup'99 dataset) and classifying the dataset using Multilayer Perceptron, Naïve Bayes, and the Ensemble method via Stacking.

The empirical result showed that the ensemble method accompanied with gain ratio technique for filtering performed better than each individual base classifier in the classification of various attack. It helped in increasing the accuracy in terms of detection and classifying more than the other base classifiers and in terms of lowering false alarm rate. It also lowered the false alarm rate reducing it more in the U2R category of attack which had the highest false positive rate with the Naïve Bayes and Multilayer Perceptron. The overall performance of the ensemble algorithm based on stacking in terms of detection accuracy did improve accuracy and was the most effective in the reduction of false positive attack for the U2R category of attack. Conclusively the Ensemble approach improved Intrusion Detection not only in maximizing the accuracy to almost a 100% but also minimizing the rate of false alarms for the U2R attack category which was high with the multilayer perceptron and Naïve Bayes classifier.

This work can be extended by combining other feature selection techniques for the purpose of reducing the dimensionality of the dataset (which will be relative to the dataset used) before carrying out the classification process.

REFERENCES

[1] Summers R. C. (1997). Secure computing: threats and safeguards. New York: McGraw Hill:

[2] Mounji. A. (1997). Languages and Tools for Rule Based Distributed Intrusion Detection .PhD Thesis ,Faculties Universaitres Notre-Dame dela Paix Namur.

[3] Nadiammai G. V., Krishaveni S., .Hemalatha M. (2011) – "A comprehensive Analysis and study in intrusion detection system using data mining Techniques". *IJCA*, Volume 35 –No.8.

[4] Youssef A. and Emam. A. (2011). Network intrusion detection using data mining and network behaviour, *International Journal of Computer Science & Information Technology (IJCSIT)*, Vol. 3, No 6.

[5] Iwan, Syarif Ed, Zaluska, Adam Prugel-Bennett, Gary Wills (2012): Application of Bagging, Boosting and Stacking to Intrusion Detection, School of Electronics and Computer Science, University of Southampton, UK.

[6] Zhao H. (2014). Intrusion Detection Ensemble Algorithm based on Bagging and Neighbourhood Rough Set, *International Journal of Security and Its Applications* Vol.7, No.5 pp.193-204.

[7] Subbulakshmi T, Ramamoorthi A, and Shalinie S. M. (2009). Ensemble design for intrusion detection systems, *International Journal of Computer science & Information Technology (IJCSIT)*, Vol 1, No 1.

[8] Singh S and Sanjay (2009). An ensemble approach for feature selection of Cyber Attack Dataset, *International Journal of Computer Science and Information Security IJCSIS*, Vol. 6, No. 2.

[9] Borji. A. (2007). Combining Heterogeneous Classifiers for Network Intrusion Detection Cervesato (Ed.): ASIAN 2007, *LNCS* 4846, pp. 254 – 260.© Springer-Verlag Berlin Heidelberg.

[10] Govindarajan M. and Chandrasekaran R. M. (2012). Intrusion Detection using an Ensemble of Classification Methods, *Proceedings of the World Congress on Engineering and Computer Science* Vol. I WCECS , October 24-26, San Francisco, USA.

[11] Ibrahim H, Badr. M , Shaheen. A. (2012). Adaptive Layered Approach using Machine Learning Techniques with Gain Ratio for Intrusion Detection Systems, *International Journal of Computer Applications* (0975 – 8887), Volume 56– No.7.

[12] Yimin Wu, (2004). High –dimensional Pattern Analysis in Multimedia Information Retrieval and Bioinformatics, Doctoral Thesis, State University of New York, January 2004.