ISSN:0975-9646

Ensemble Neural Network and K-NN Classifiers for Intrusion Detection

Shalinee Chaurasia[1] , Anurag Jain[2]

[1, 2] Computer Science Dept., Radharaman Institute of Technology & Science, RGTU Bhopal (M.P.) India

***Abstract* —In this paper we present the ensemble algorithm to improve the intrusion detection precision. Ensemble classifier is a technique which uses a combination of a plurality of classifiers , to obtain a more accurate inference result as a single classifier. we propose an ensemble intrusion detection system that combines two classifiers : K- nearest neighbors and neural network for the abuse detection.We be set with bag classifier for misclassification data to enhance the detection rate.This algorithm works on the KDD -99 data set to improve . We present an approach , the user behavior that occurs advantage of the properties of bagging and get the display results on preliminary probe of our approach.In this paper our aim is to improve the efficiency of intrusion detection system. The results demonstrate that our approach achieves better performance than that of a single classifier.**

***Keywords*— Intrusion Detection System (IDS), Bagging, Neural network and k-nn classifiers.**

I. INTRODUCTON

The process can be defined by the intrusion of malicious behavior expresses the power of the system resources and targets[10].

Malicious behavior is called a system or individual action that attempts to use, or access to a computer system without authorization and the authority of those who have (for example, the insider threat) defines the access illegitimate system. Intrusion Detection System (IDS) is a system that collects and analyzes information from various areas within a computer to identify attacks made against these components. The a numer of ways in which the IDS uses a number of generic methods to monitor the exploitation of the vulnerability. IDS can be characterized according to three main conditions [15], [16].

1.The data source: In this case, we have host-based. Host based IDS monitorscomputer components (such as operating system, packet, system log, etc.). Host based IDS monitors the systems.

2.The model of intrusion detection: Here we have novelty (anomaly) detection, signature based (misuse) detection,. Anomaly based IDS monitoring depends on the behavior of system. Misuse based IDS monitoring depends on signature to data.

3.The audit collection and analysis: Here IDSs are divided into either centralized or decentralized (distributed) IDSs.

In this paper we have presented of bagged of neural network and k-nn classifiers for intrusion- detection based on machine learning. We have used firstly a neural network for classification of five class data.we are also used k-nn classifiers & than bagging of multiple classifiers. KDDcup 1999 benchmark dataset is used for testing the proposed algorithm and the results are promising and more important, especially low false alarm rate and high detection rate with take less time to create a model to achieve, that out performs the existing methods are presented [17].

II. INTRUSION DETECTION SYSTEM

Intrusion detection systems are safety management systems that are used to detect inappropriate, incorrect or anomalous activities in computers. With the rapid growth of the Internet, this harmful behavior to increase at a fast pace and can easily cause millions of dollar in damage to an organization. Therefore, the development of detection systems with the highest priority of the government, research institutions and business enterprises has set[13].

The principal aim of the intrusion detection is to detect future attacks leading to additional learning techniques. Intrusion detection is based on the principle that the characteristics of intrusion are different from normal behavior In general, IDS can be divided into two categories:anomaly detection and misuse(signature) detection. Anomaly detection attempts to determine if the difference habits normal use set can be labeled as intrusions.and misuse detection uses patterns of well-known attacks of the system to identify intrusions.

It is mainly identified following common root blocks of an intrusion detection system are: (i) Sensor probes (ii)Monitor (iii) Resolver (iv) Controller. IDSs focuses on four characteristics (i)Audit source location (ii)Detection method (iii) Behaviour on detection (iv)Usage frequency.

The rest of paper is structured as follows:The Second section explain the details of intrusion detection system.The third Section explain the detail of related work. The fourth Section explain the proposed work.The fifth section explain detail of data mining method which is including details of machine learning Neural Network and K-nn classifiers, ensemble method Bagging.The sixth Section explain the Ensemble learning. The seventh section developed the Algorithm.The eighth section showing the results. The nineth Section explain the Conclusion.

III. RELATED WORK

This section presents an overview of some of the relatedliterature published recently.

Hu Yan Lixin Li et. Ai.[1] proposed a multi classifier has been discussed. Perimeter intrusion detection system, based on the multi classifier, will get smarter. It can not only accurately discriminate nuisance events and intrusion events, suppressing nuisance alarms, but also can highly recognize intrusions, offering more valuable information to users.

Herve Debar et. At. [2] proposed a neural networks are of using an intrusion detection system.The user model that we have developed here, is the complement of a statistical model, because neural networks can not treat adequately all available data. The close coupling between the neural network and the expert system is necessary to analyze the output of the net and propose explanations and a clear diagnosis about the security administrator.

Manasi Gyanchandani et. At.[3] proposed a Error due to variance has been reduced using classifier combinations thus improving the performance of the system classification using the NSL-KDD dataset.Bagging provides better results. NSL-KDD dataset can be used for resultants evaluation for five classes.

Sufyan T. Faraj Al-Janabi et. Al. [4] proposed the detection of unknown attack anomalies by the developed IDS. Indeed, the modular architecture of the system enables it to be easily extended, configured, and/or modified. This is can be done by adding new features or by replacing features when they need to be updated. However, the training of the ANN requires a very large amount of data and considerable time to ensure that the results are accurate. Another issue is that there is some kind of compromise between increasing the classification levels and the percentage of detection.

Benjamin Thirey et. Al. [5] proposed an ensemble of classifiers trained to detect membership in a given class can achieve high rates of classification. We have shown that we can achieve greater classification rates by combining a series of classifiers optimized to detect class membership, than by using single instances of classifiers. Our model is best adapted towards classification problems involving three or more classes since a two class model can be readily handled by a single classifier instance.

Te- shun chou et Al.[6] Proposed classification of the technical set (ensemble) is applied to the intrusion detection task. We develop a three-layer hierarchy structure That includes three groups of classifiers and Each Consists of three classifiers based feature selection. Each base in Selecting feature classification, we apply different machine learning algorithm and feature subset to solve uncertainty problem and maximize the diversity.

P Amudha et al.[7] Proposed Classification and measurement in relation to the attack and ensemble of classifiers in order to analysis the efficiency of a series of experiments on KDD Cup'99 dataset. The experimental results, NBTree have small training data and a better detection rate and false alarm rate for R2L dataset and U2R datasets that allows for better accuracy. They also build the model to the time taken by NBTree is further observed that compared to other classifiers. They conclude the random forest for DOS and probe dataset and NBTree for U2R and R2L dataset gives better performance.

IV. Proposed Work

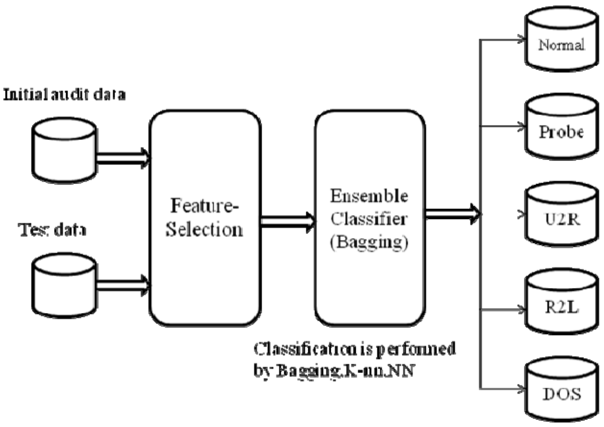
Initial audit data
Test data
Feature-Selection
Ensemble Classifier (Bagging)
Normal
Probe
U2R
R2L
DOS
Classification is performed by Bagging,K-nn,NN

Fig. 1. Experimental Flow

There is a whole data set,each record include a set of features, one of the features of the class. This data set is divided into initial audit data(trainning) and test sets, with training to be used for the construction of the model and the test used to validate it. We train different layers our model to select different characteristics.The feature selection phase aims to reduce more n more data and provide a better accuracy. We had trained each of the classifiers, using the same training data. Classification is performed with the help of ensemble classifiers. Here we have ensemble neural networks and using KNN classifier ensemble. Training Data was presented for the machine classifiers. This training database has five classes.it is normal, probe, Remote to local(R2L), user to root(U2R), and denial of service(DOS)[11].

V. Datamining Machine Learning Techniques

Data mining is also called Knowledge Discovery andData Mining is the process of automatically searching large amount data to models using association rules [14] .In data mining based approaches to modeling intrusion detection . When data mining is introduced in intrusion detection ,data mining focused on two main problems : i) to determine the adaptive beam feature , ii) to improve the detection rate . The intrusion detection system based on data mining is presented in Figure 2 .The process starts with an initial set of test data . Data and Method Selection page preprocessed correlation is used to obtain the optimal set of features for classification [8] .

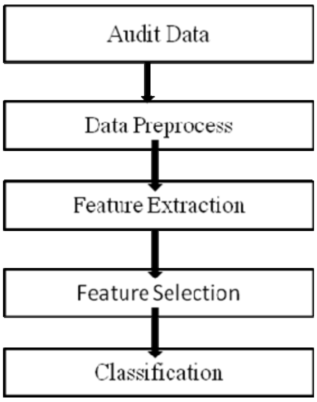

Fig. 2. Intrusion detection system based on datamining

A. Neural Network

A neural network [2] is a collection of simple units called neurons. A neuron is a linear machine that, a weighted sum of several inputs for a set of weights, and then computes the heaviside function or a sigmoid function value obtainan output, it is said activation of neurons. The choice of the transfer function determines whether the neuron is rated binary or continuously. In the form of a neural network, these neurons are connected to each other by a predetermined Topologie.Dieser topology defines an input layer, in which the activations of the neurons on the input values and one output, where the reading of the activations of the neurons are the response of the net. Neural networks can therefore as a non-linear transfer function of a space vector in another. The weights of the neurons are the parameters of this function folding[2].

Learning is the characteristic of neural networks. It will allow us to learn the law of the behaviour without explicitly expressing it using the sample we get from the audit trail. The model will therefore adapt itself to the user with very few changes, thus providing a good genericity. Also, the behavior changes slowly as the user slips into his habits. It changes quickly when the user gets a new job. The learning algorithm allows the network to follow the behavior patterns closely and adapt itself to the constantly occurring changes[2].

Neural networks have been used both in novelty intrusion detection as well as in Signature based intrusion detection. For novelty intrusion detection, neural networks were modeled to learn the typical characteristics of system users and identify statistically significant variations from the user's established behavior[12].

A neural network for signature-based detection is implemented in two ways. The first approach incorporates the neural network component into an existing or modified expert system. This method uses the neural network to filter the incoming data for suspicious events and forward them to the expert system. This improves the effectiveness of the detection system The second approach uses the neural network as an independent signature-based detection system. In this method, the neural network would be to receive data from the electricity network and to analyze abuse penetration. There are several advantages to this approach. It has the ability to learn the properties of misuse attacks and identify which, in contrast to all instances that have previously been observed by the network. It has to recognize a high accuracy to known suspicious events. Neural network works well on noisy data [12].

K-NN

The k - nearest neighbor or K-NN algorithm is known for the data mining community , and is one of the top algorithms in [13] . The algorithm achieves ranking among m different categories.Each case should be classified is an element that contains a collection of r distinct features in the set A = { a1 , a2 , ..., ar} where aj corresponds to the j-th feature. Therefore ,an example is a vector p = <p1, p2, …, pr> property values . For some predetermined value of k, the k nearest neighbors is determined using a distance measure which is calculated using the difference in the distances between each of the features of the instance and its neighbors . Euclidean distance is by far the most popular metric for calculating membership proximity.An example in a particular category can be calculated either as a possibility or a simple majority of the class with the most representation in the nearest neighbors k. At the simplest level , this is a problem of binary classification where the data is classified as a specific course or not[5].

VI. ENSEMBLE LEARNING

"In fact there are several types of intrusions and several detectors are required to identify them" . Improve some features such as precision of a data mining system is to use a various techniques and combine the results together. The combined use of several data mining methods is known as a complex approach, and the procedure of learning the correlation between ensemble of these techniques is known by such names as multistrategy learning [17]. There are many ensemble learning technique that is boosting, bagging, stacking. We are using BAGGING learning technique together.

A. Bagging

Bagging means bootstrap aggregation is single easiest but extremely successful set of scheme for improving the problems of classification. For example the weak classifiers ,as like decision tree algorithms can be changeable , specially when the designation of a point changes little and training can conduct to many different tree . This method is commonly used to decision tree algorithms , but also can be applied with other classification algorithms as like Naive Bayes, nearest neighbor , induction rule , etc. The ensemble technique(bagging) is too usable for wide and high dimensional data , as like intrusion datasets , where searching a fair model or classifier that can operate in one step , is laborious because of the complexity and level of the problem.

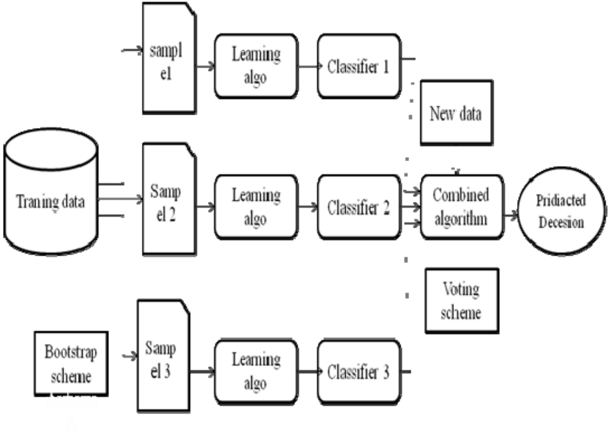

Figure 3. Bagging (bootstrap aggregation).

VII . ALGORITHM

Bagging Procedure

Classifier Generation

Step 1. Make t datasets from a database application of sampling replacement arrangement.

Step 2 Set a learning algorithm to individual pattern Bagging Procedure.

Step 3. Set a learning algorithm to individual pattern training dataset.

Classification

Step 3. For an object with unknown decision to do with each of the t classifiers predictions.

Step 4. Choose the most repeatedly predicted decision.

Algorithm-

Stp.-1. In first step We trained K-nn classifier and apply testing on 5-class data which is classified or misclassified data.

Stp.- 2. K-nn works For each training example <x,f(x)>, add the example to the list of training_examples.

Given a query instance xq ¨ Given a query instance x to be classified, q to be classified, Let x1,x2….xk denote the k instances from training_examples that are nearest to xq.Return the class that represents the maximum of the k.

```
Stp.-3 If knn_weight(class_knn)>nn_weight(class_nn)
            data_class = class_knn;
      else
            data_class = class_pnn;
```

Stp.- 4. Now in this step we trained Neural network classifier and apply testing on 5- class data, it is classifies data which are misclassified by k-nn.

Step 5. Than now we used Ensemble technique Bagging for misclassifiers.

- *Training*

  i)In each iteration t, t=1,…T

  Randomly sample with replacement N samples from the training set

  ii)Train a chosen “base model” (e.g. neural network, decision tree) on the samples.

- *Test*

  i)For each test example Start all trained base models

  ii)Predict by combining results of all T trained models:

  Regression: averaging

  Classification: a majority vot

VIII . EXPERIMENT AND RESULT

The data set used for our research is KDD Cup 99 dataset.KDD 99 dataset has been the most widely used for the evaluation of signature based intrusion detection and novelty intrusion detection methods. We have used KDDCup’99 intrusion detection dataset, which contains 25621 records with .8 training ratio..KDD contains a totally 39 attack types and are fall into four categoies:

1. Denial of Service(DoS)
2. Remote to Local (R2l)
3. User to Root(U2R)
4. Probe

The proposed IDS has beem implemented using MATLAB tool and These algorithms were executed on a PC on a Pentium core processors under window 7 ,2.20 GHz and 4 Gb RAM. We have first used a k-nn classifier for classification of five-class data which is (normal, dos, u2r, r2l, and probe). This did classified or misclassified. We also used experiments with Neural network, which created multiple classifiers and then classified data using a voting technique. K-nn classified data which were misclassified by NN & then apply bagging on multiple classifiers. This was focused on misclassified classifiers. Bagging creates a number of rounds and continues parellell the process until all data sets are correctly classified.

Evaluation methods

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Detection Rate(TPR)} = \frac{TP}{TP + FP}$$

$$\text{False Alarm Rate (FPR)} = \frac{FP}{FP + TN}$$

Where TP= true positive,TN= true negative,FP=false positive,FN= false negative.

Table-I and II shows the performance of two classifiers and ensemble approach (bagging) based on correctly classified instances. Table I showing the five class indivisual accuracy result. In table I we can show that bagging gives better accuracy other than single classifier. From table II, bagging gives the better result. Now we compare the result of the K-NN, NN & Bagging.K-NN, NN, Bagging performance of TPR ,FPR, precision and recall for each specific class.

Table-I Accuracy of different attacks through K-nn ,NN and Bagging

Method	Dos	Normal	Probe	R2l	U2r
k-nn	0.904	0.94	0.936	0.938	0.964
NN	0.838	0.996	1	0.85	0.984
BAGGING	0.928	0.992	0.996	0.942	0.984

Table-II Shows the result for K-nn,NN & bagging.

parameter	K-NN	NN	PROPOSED BAGGING	BAGGING[9]
TPR	0.878	0.863	0.8852	0.712
FPR	0.04	0.032	0.021	0.027
PRECISION	0.843	0.856	0.8888	0.712
RECALL	0.8995	0.9085	0.9152	0.972

The graph in figure 4 shows the performances of K-NN, NN and Bagging in terms of accuracy. Now we compare the results of the K-NN, NN and Bagging algorithms. Especially in case of Bagging accuracy is 96.84 %. It provides the better accuracy for good intrusion detection.

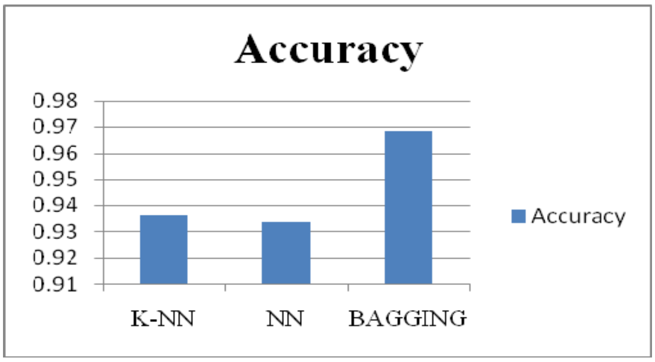

Fig. 2 Comparison accuracy for K-NN, NN & BAGGING

The graph in Figure 5 shows that the performances FPR of bagging classifiers is 0.021 near about zero which is best for desirable intrusion detection. For good IDS FPR should be Low.

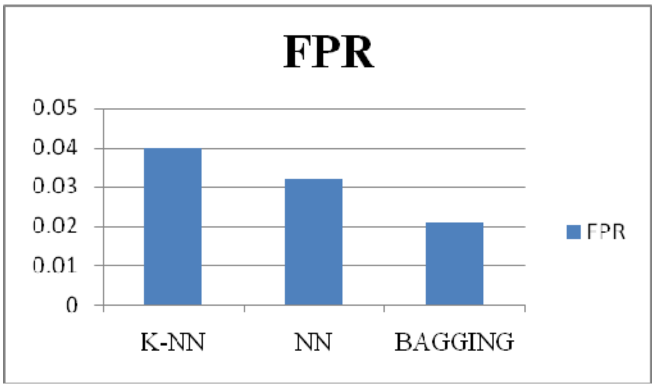

Fig. 5 FPR comparison of K-NN, NN & BAGGING

The graph in figure 6 we compare the TPR of the K-NN, NN & bagging. For a good IDS TP rate should be high. Shows that the TP rate of bagging classifiers is a high comparision to K-nn and NN.

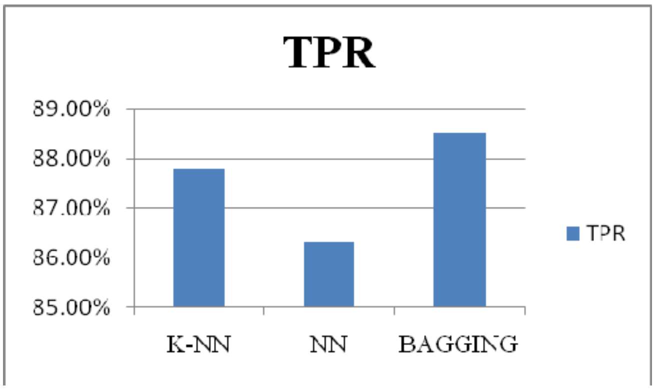

Fig. 6 TPR comparison of K-NN,NN & BAGGING

From above figure 4, 5 and 6 clearly show that the bagging classifier algorithm accuracy, TPR, FPR is better than single classifier.

IX . CONCLUSION

This paper proposed Ensemble classifier technique for intrusion detection. We have demonstrated ensemble of different classifiers for increase the accuracy. In the algorithm K-NN and NN classifier are used and evaluated their performance on the KDD Cup99 dataset. We have first used a K-NN for classification of five-class data and We also used neural network (NN) & then bagging uses of multiple classifiers.Bagging provides better results .kdd cup 99 dataset can be used for performance evaluation for 5 classes (normal, dos ,probe, u2r and r2l).bagging classifier provides better result and provide better accuracy for intrusion detection system.the results shows 96.84 % accuracy by bagging and 0.021 false positive rate.

In the future we will continue the research of improving the performance using several combination of classifiers by ensemble technique.

REFERENCES

[1] Hu yan, Lixin Li, “ ANN-based Multi Classifier for Identification of Perimeter Events,” Fourth International Symposium on Computational Intelligence and Design,pp.158-161,2011.

[2] Herve Debar, “ A Neural Network Component for an Intrusion Detection System,” pp. 240-250,IEEE-1992.

[3] Manasi Gyanchandani, R. N. Yadav, and J. L. Rana, “Intrusion Detection using C4.5: Performance Enhancement by Classifier Combination,” Vol. 01, No. 03, ACEEE Dec 2010.

[4] Sufyan T. Faraj Al-Janabi and Hadeel Amjed Saeed, “ A Neural Network Based Anomaly Intrusion Detection System,” pp. 221-226,IEEE-2011.

[5] Benjamin Thirey and Christopher Eastburg, “Increasing Accuracy Through Class Detection: Ensemble Creation Using Optimized Binary Knn Classifiers,” IJCSEA, Vol.1, No.2, April 2011.

[6] Te-Shun Chou, Jeffrey Fan, Sharon Fan, and Kia Makki, “Ensemble of Machine Learning Algorithms for Intrusion Detection,” IEEE 2009 International Conference on Systems, Man, and Cybernetics San Antonio, TX, USA - October 2009

[7] H. Günes Kayacık, “ Selecting Features for Intrusion Detection: A Feature Relevance Analysis on KDD 99 Intrusion Detection Datasets”.

[8] P Amudha, H Abdul Rauf, “*Performance Analysis of Data Mining Approaches in Intrusion Detection*,” IEEE-2011.

[9] Manasi Gyanchandani, R. N. Yadav and J. L. Rana, “Intrusion Detection using C4.5: Performance Enhancement by Classifier Combination,” ACEEE Int. J. on Signal & Image Processing, Vol. 01, No. 03, Dec 2010.

[10] Kruegel Christopher, Darren Mutz William, “Bayesian Event Classification for Intrusion Detection,” 2003.

[11] Shalinee Chaurasia and Prof. Anurag Jain, “*Review: Ensemble Neural Network and KNN Classifiers for Intrusion Detection*,” IJSER(ISSN 2229-5518) ,Volume 4,pp. 213-217, Issue 12, December-2013.

[12] Sandhya Peddabachigari, Ajith Abraham, Johnson Thomas, “Intrusion Detection Systems Using Decision Trees and Support Vector Machines,” Department of Computer Science, Oklahoma State University, USA.

[13] Xindong, W., Kumar and Hand, D. J. and Steinberg, D. (2008) Top 10 Algorithms in Data Mining. Knowledge and Information Systems, 14 1: 1-37.

[14] Theodoros Lappas and Konstantinos Pelechrinis,“Data Mining Techniques for (Network) Intrusion Detection Systems,” Department of Computer Science and Engineering ,UC Riverside, Riverside CA 92521.

[15] J. Daejoon , H. Taeho, and H. Ingoo “The neural network models for IDS based on the asymmetric costs of false negative errors,” Pergamon, Journal of Expert Systems with Applications, No. 25, pp. 69–75, 2003.

[16] R. Bace and P. Mell, "NIST Special Publication on Intrusion Detection Systems," 2002.

[17] Mrutyunjaya Panda1 and Manas Ranjan Patra,” *Ensemble Voting System for Anomaly Based Network Intrusion Detection*,” Vol 2, No. 5, ACEEE-November 2009.