An Improved Intrusion System Design using Hybrid Classification Technique

Rita Mandal*
Department of Information Technology
Mahakal Institute of Technology, Ujjain
Madhya Pradesh, INDIA.

Shweta Yadav
Department of Information Technology
Mahakal Institute of Technology, Ujjain
Madhya Pradesh, INDIA.

ABSTRACT
The data mining is an essential tool for current age technology. It is very useful for various applications such as business intelligence, computational cloud and other research and science based projects. In this presented paper a new kind of intrusion system design is proposed and their implementation is presented using MATLAB. The proposed attack classification technique is based on the C4.5 decision tree and the supervised back propagation neural network. During the attack classification the proposed system first pre-processes the input dataset. Therefore, in order to reduce the dimensions of the dataset, the KPCA algorithm is applied and then the C4.5 data model is implemented for extracting the classification rules form the C4.5 classifier. These rules are in the form of IF THEN ELSE format. Now the defined rules are processed by the using the back propagation neural network. The proposed model is tested on the different size of KDD cup dataset and the performance is provided. According to the obtained results the proposed data model provides the improved classification rates.

General Terms
Security, Data mining, IDS design, Implementation, performance evaluation

Keywords
Decision Trees, C4.5, back propagation neural network, IDS, results analysis

1. INTRODUCTION
Nowadays the network communication is growing continuously and adopted rapidly. The network technology having a large number of applications, this is now used for banking applications, shopping and others. Thus a significant amount of sensitive and private data is traversing through these networks. Due to data traversing through untrusted network, the loss of security and data is an essential concern in the network technology [1]. The presented study provides a detailed investigation of the security aspects and their flaws. Therefore, in this study intrusion detection systems are learned and a new concept of intrusion system design is presented using hybrid classification technique [2].

The proposed hybrid classification technique involves the implementation of cluster analysis, Bayesian classification and the back propagation algorithm for classifying the KDD cup dataset [3]. KDD cup data set includes the 41 attributes and a class label, thus total 42 attributes are available for classification. In this data set the classes are divided into two major classes: normal and anomaly. The data set basically contains the basic network packets and their values for detection.

In this proposed work a data mining based IDS system [4] is prepared and implemented, the proposed data model is able to accept the KDD cup dataset as input training set and produces the outcomes of classifiers. The proposed implementation of IDS uses the concept of supervised learning for improving the classification ability of detection.

2. PROPOSED WORK
In recent development there is various kinds of IDS design concepts are appeared but in these systems some issues are observed. Some of them are considered for improvements in traditional systems are listed below.

1. The implementation of IDS systems leads to store a significant amount of data for processing. Thus pattern detection and recognition from such huge set of data is a time consuming as well as high memory resource consuming process.
2. Learning with large data and noisy data is affecting the performance of classifiers in terms of accuracy. Therefore, using dimensionality reduction and noise reduction methods required to keep preserve the performance during the data analysis and learning.
3. Due to weak learning situations there are less false alarm rate are obtained, therefore learning ability is required to improve by which thedetection rate can be improvable.

In order to overcome the addressed issues and challenges the following suggestions are made.

1. Pre-process data by which size of data attributes can be reduced and only selected features are extracted for classifiers training or learning.
2. Hybridization of algorithms inherits properties of parent algorithms by which performance of classification technique are improved by combining more then on classification approach
3. Search for strong and accurate classifier that provides the significant improvement on classification rate and the attack detection rate of IDS

The proposed system is constructed by implementing the number of different components into a single system. The individual components of the proposed model are discussed as:

1. **Input dataset:** The input dataset for training of the proposed classification scheme is KDD CUP dataset which includes 41 attributes and the huge number of instances which is approximately 1.5 lakes thus for experimentation purpose a number of datasets are constructed from the original KDD Cup data for experimentation. The preferred size of input data set is 1250, 2500, 4725, 5000, 10000.
2. **KPCA:** KPCA (Kernel Principal Component Analysis) [5] is kernel based principal component analysis technique which is used to reduce the dimensions of the training data. Thus the input dataset is reduced in dimensions by which the performance of the learning can be optimized in terms of the memory consumption.
3. **BPN:** It is an opaque data model for supervised learning. This algorithm is used to recognize the input data patterns and constructs the weights for estimating the pattern outcomes as given in test sets. Therefore first the BPN algorithm takes training from the KPCA selected attributes.
4. **C4.5 Decision tree:** Decision trees are having an essential property by which the data pattern can be distinguished by evaluating a limited set of attributes. Therefore the BPN generated patterns are mounted using the decision tree algorithm. This algorithm generates the tree data structure to support the classification.
5. **Rule generation:** The tree data structure is converted into a set of rules for generating the classification rules. These rules are used to evaluate the limited amount of attributes to find the target data pattern.

The proposed system can be simulated using the below given figure 1 simulation architecture.

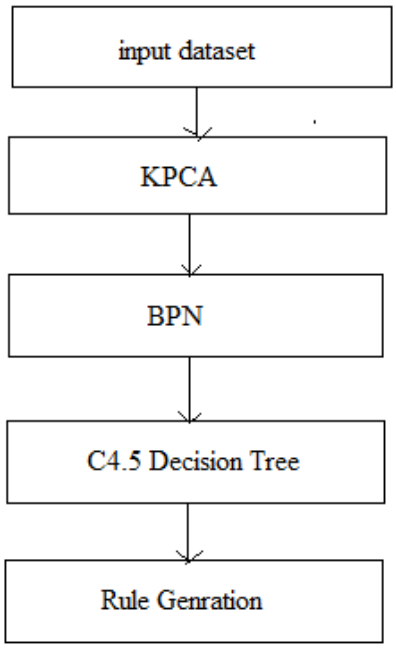

Figure 1 The proposed system

3. IMPLEMENTED ALGORITHMS

This section provides the study of the C4.5 algorithm and the BPN algorithm which is used in proposed system design.

3.1 C4.5 algorithm

C4.5 (developed by Quinlan, 1993) is an algorithm that learns the decision-tree classifiers. It has been observed that C4.5 performs short in the domain where there is pre-entrance of continuous attributes compared with the learning tasks with mostly separate attributes. For instance, a system which looks for well-defined decision tree with 2 levels and then put comments [6]:

"The accuracy of trees made with T2 is equalized or even exceed trees of C4.5 upon 8 out of all the datasets, with the entire except one that have incessant attributes only."

INPUT: An exploratory data set of data (D) portrayed with the means of discrete variables.

OUTPUT: A decision tree say T which is constructed by means of passing investigational data sets.

1. A node (X) is created;
2. Check if the instance falls in the same class.
3. Make node (X) as the leaf node and assign a label CLASS C;
4. Check IF the attribute list is empty, THEN
5. Make node(X) a leaf node and assign a label of most customary CLASS;
6. Now choose an attribute which has highest information gain from the provided attribute List, and then marked as the test_attribute;
7. Confirming X in the role of the test_attribute;
8. In order to have a recognized value for every test_attribute for dividing the samples;
9. Generating a fresh twig of tree that is suitable for test_attribute = atti from node X;
10. Take an assumption that Bi is a group of test_attribute=atti in the samples;
11. Check If Bi is NULL, THEN
12. Next, add a new leaf node, with label of the most general class;
13. ELSE a leaf node is going to be added and returned by the Generate_decision_tree.

This section provides the implementation of the classical C4.5 algorithm the next section provides BPN algorithm.

3.2 Back Propagation Neural Network

The implementation of neural network is defined in two phases' first training and second prediction: training method utilizes data and designs the data model. By this data model next phase prediction of values is performed [7].

Training:

1. Prepare two arrays, one is input and hidden unit and the second is output unit.

2. Here first is a two dimensional array W_{ij} is used and output is a one dimensional array Y_i.

3. Original weights are random values put inside the arrays after that the output is given as.

$$x_j = \sum_{i=0} y_i W_{ij}$$

Where, y_i is the activity level of the j^{th} unit in the previous layer and W_{ij} is the weight of the connection between the i^{th} and the j^{th} unit.

4. Next, action level of y_i is estimated by sigmoidal function of the total weighted input.

$$y_i = \left[\frac{e^x - e^{-x}}{e^x + e^{-x}}\right]$$

When event of the all output units have been determined, the network calculates the error (E) given in equation.

$$E = \frac{1}{2}\sum_i (y_i - d_i)^2$$

Where, y_i is the event level of the j^{th} unit in the top layer and d_i is the preferred output of the j_i unit.

4. RESULTS ANALYSIS

This section provides the complexity of classification according to the input training set. The evaluation of the performance is given in terms of memory and training time of implemented classifiers.

4.1 Memory consumption

The amount of main memory consumed during algorithm processing is known as memory consumption of space complexity. The proposed classification performance is given with increasing size of dataset instances and their respective memory values.

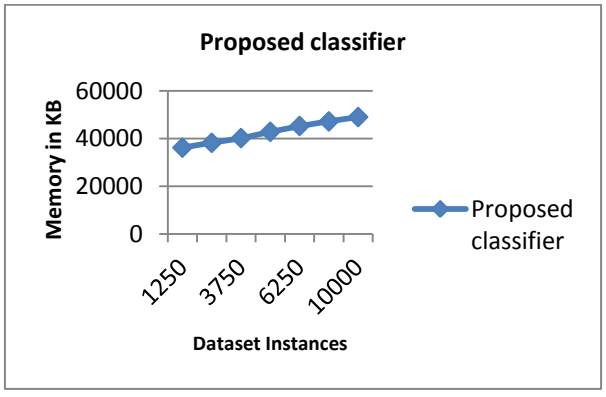

Figure 2 Memory Consumption

The evaluated memory consumption is provided using figure 2, in this diagram X axis provides the information about the dataset size in terms of data instances and respective memory consumption in terms of KB is given using Y axis. According to the evaluated results the obtained memory consumption is increases as the size of data set increase.

4.2 Space complexity

The amount of time required to perform classification task is known as the space complexity of the system. In this system three different classifiers are implemented one after another therefore the complexity is computed in terms of seconds with increasing size of data instances.

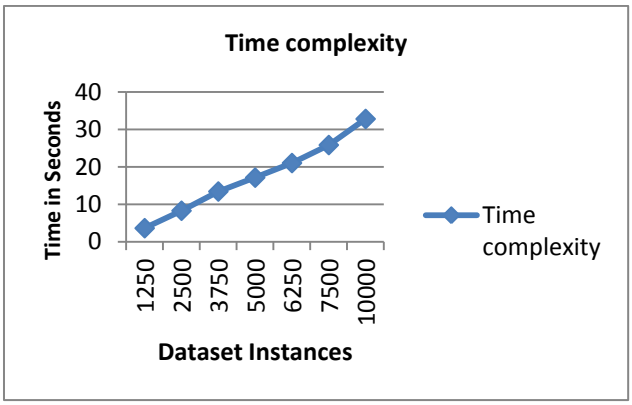

Figure 3 Time Complexity

The time consumption of the proposed classification model is given using figure 3 in this diagram the amount of time is given in Y axis and the X axis demonstrate the size of dataset. According to the evaluated performance the classifier needs more time as the size of data is increases. Thus that can be conclude time complexity of the proposed classifier system is optimal and consumes an adoptable amount of time.

4.3 Accuracy

The amount of data correctly classified over the given input patterns is known as accuracy of the system. That can also derive using the formula given:

$$Accuracy = \frac{total\ correctly\ classified\ samples}{total\ input\ samples} X100$$

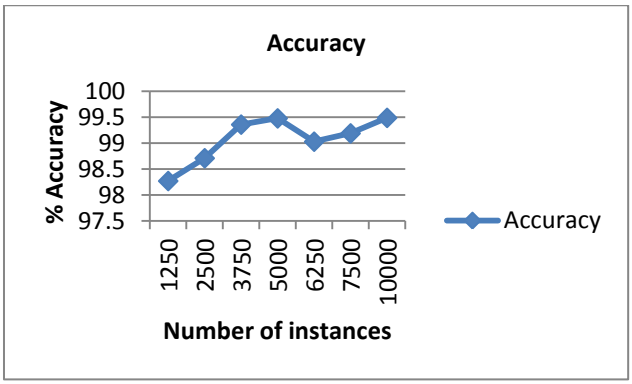

Figure 4 Accuracy of System

The accuracy of the system is given using figure 4. In this diagram the percentage accuracy of the system is given in Y axis and the number of dataset instances are given using X axis. According to the evaluated results the performance of classification is improved if the class distribution and all the attributes are contains the significant information otherwise the performance of classification is decreases as given for 6250 instances. Thus the accuracy of the system is depends upon the data provided as input for learning.

4.4 Error Rate

The error rate provides the information about the misclassified data over the given samples to classify, in this scheme the N-cross validation processes used for calculating the accuracy and error rate. The obtained error rate can be calculated using the formula:

$$error\ rate = \frac{total\ incorrectly\ classified\ samples}{total\ samples\ given\ for\ classification} X100$$

Or

$$error\ rate = 100 - accuracy$$

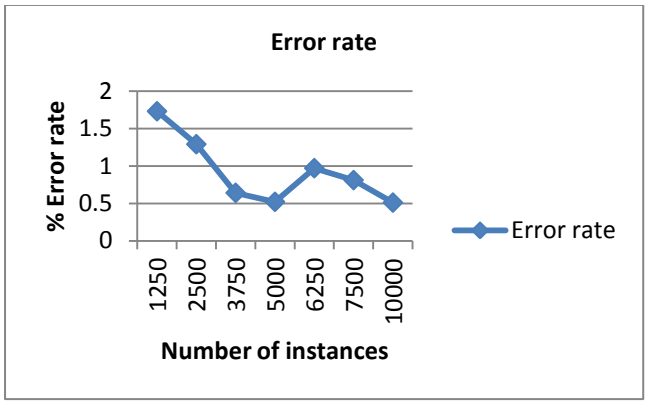

Figure 5 error rate

The obtained error rate from the classification system is given using figure 5 in this diagram the Y axis demonstrate the percentage error rate of the system and the X axis shows the number of instances in dataset.

5. CONCLUSION

These days the communication system is affecting applications and use of applications, in this context security in this domain is a primary aspect in communication network. The proposed study is an investigation of IDS (intrusion detection system) and their design concept. For that purpose an intrusion detection system is developed using the analysis of KDD CUP 99's dataset. In this intrusion detection system the main focus is given over classification and performance improvement of classifiers. Therefore, different algorithms are applied for filtering the data set features.

Thus the proposed KDD CUP dataset classification technique is used to demonstrate the effective classification accuracy. The proposed model incorporates the KPCA (kernel principal component) analysis technique, BPN (back propagation neural network) and the decision tree rule generation technique. the proposed data model first reduce the data set dimensions for optimizing the performance of the learning systems. Then after the data model is prepared using back propagation neural network and decision tree data model. This model generates the classification rules for classifying the KDD CUP dataset using less number of attributes evaluations. Thus the presented data models are an efficient classification technique.

The implementation of the proposed concept is provided using MATLAB simulation environment and the performance of the system is evaluated in different performance parameters. The proposed model provides the optimum classification rule generation technique with less resource consumption. Thus the given model is adoptable by using the accuracy parameters and the resource consumption.

The proposed KDD CUP 99's classifier is prepared in this study and performance of the proposed system is evaluated, according to the obtained performance the system is adoptable and efficient. In near future the performance of the classification is improved more for reducing the steps of algorithm processing and the time consumption.

6. REFERENCES

[1] Mala Bharti Lodhi, Prof. Vineet Richhariya, Prof. Mahesh Parmar, "An Implementation of IDS in a Hybird Approach And KDDCUP Dataset", International Journal ofResearch –Granthalayah, Vol.2(Iss.3):December,2014

[2] Mala Bharti Lodhi, Vineet Richhariya and Mahesh Parmar, "A survey on Data Mining based Intrusion Detection Systems", International Journal of Computer Networks and Communications Security,VOL. 2, NO. 12, DECEMBER 2014, 485–490

[3] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set", proceeding of the 2009 IEEE symposium on computational intelligence

[4] Mradul Dhakar and Akhilesh Tiwari, "A Novel Data Mining based Hybrid Intrusion Detection Framework", A Novel Data Mining based Hybrid Intrusion Detection Framework, Vol. 9, No. 1, 2014, pp. 037-048

[5] Long-Sheng Chen, Jhih-Siang Syu, "Feature Extraction based Approaches forImproving the Performance of Intrusion Detection Systems", Proceedings of the International Multi Conference of Engineers and Computer Scientists 2015 Vol I,IMECS 2015, March 18 - 20, 2015, Hong Kong

[6] Kundan Kumar Mishra, Rahul Kaul, "Audit Trail Based on Process Mining and Log", International Journal of Recent Development in Engineering and Technology, Volume 1, Issue 1, Oct 2013

[7] AnkitaAgrawal, "Host Load Prediction in Computational Grid Environment", International Journal of Computer Applications (0975 – 8887) Volume 77– No.10, September 2013