# CLASSIFIER SELECTION MODELS FOR INTRUSION DETECTION SYSTEM (IDS)

Anurag Jain, BhupendraVerma and J. L. Rana

Rajiv Gandhi Technical University, Bhopal, India

## *ABSTRACT*

*Any abnormal activity can be assumed to be anomalies intrusion. In the literature several techniques and algorithms have been discussed for anomaly detection. In the most of cases true positive and false positive parameters have been used to compare their performance. However, depending upon the application a wrong true positive or wrong false positive may have severe detrimental effects. This necessitates inclusion of cost sensitive parameters in the performance. Moreover the most common testing dataset KDD-CUP-99 has huge size of data which intern require certain amount of pre-processing. Our work in this paper starts with enumerating the necessity of cost sensitive analysis with some real life examples. After discussing KDD-CUP-99 an approach is proposed for feature elimination and then features selection to reduce the number of more relevant features directly and size of KDD-CUP-99 indirectly. From the reported literature general methods for anomaly detection are selected which perform best for different types of attacks. These different classifiers are clubbed to form an ensemble. A cost opportunistic technique is suggested to allocate the relative weights to classifiers ensemble for generating the final result. The cost sensitivity of true positive and false positive results is done and a method is proposed to select the elements of cost sensitivity metrics for further improving the results to achieve the overall better performance. The impact on performance trade of due to incorporating the cost sensitivity is discussed.*

## *KEYWORDS*

*Intrusion detection system (IDS), True positive (TP), False Positive(FP), Support Vector Machine (SVM).*

## 1. INTRODUCTION

Nowadays the Internets are definitely going to be a part of our lives. It widely used and provide us with so many positive things. In general, Internet security is the risks on private property and information associated with using the internet, and the self-protection from computer crime knowledge of maximizing the user's personal security also know online security. The worldwide number of online users continues to grow; internet security is also growing or updating concern for both adults and children as time to time. Common concerns regarding security on the internet includes malicious users, websites and software and various types of obscene or offensive content. Several crimes can be committed on the Internet such as identity theft, stalking and many more. For monitoring an analysis of event occurring in the information system, any deviation from the normal uses as anomaly behavior of the system is required. For safeguarding network and connected system from intrusion activities, intrusion detection system is used as prevention or its complement. Therefore second line of defense is IDS system

DOI : 10.5121/ieij.2016.4101

In this paper, audit data logs data such as KDDCup99 have detail of the users and its behavior pattern. Using this dataset majority of information about intrusion is real time. KDDcup99 dataset contains certain redundant data which has information that may be in form of attributes. Such redundant data are not useful so we normalize it. After normalizing KDDcup99 Dataset we improve accuracy and computational time of IDS[1].

By proper selecting features subsets of classifiers that gives best classification and give multi-classifier models. This model can give improved Classification results. In section II we normalize and reduction feature for Kddcup99 dataset, In section III we discuss some of classifiers like K-Means, Bayes Net, Naïve bayes, J48, ID3, NBTree, Fuzzy Logic, Sapport Vector Macine, Decision Table, JRip, OneR, MLP, SOM, LBk and Random Forest (RF). It has been progressively shown that some classifiers that contribute better classification without any important degradation in performance of IDS. Therefore literature review is more emphasized for classification for IDS. In the section IV, various types of intrusion detection systems True Positive Rate (TPR) and False Positive Rate (FPR) in anomaly and misuse detection are discussed and also create two model for combining classifiers as per its performance and minimum time taken and give better results. The outcomes finally concluded in section.

## 2. Feature Reduction and Normalization

The best way to classify is the main objective of our work and it is tested by determining and analyzes to get high accuracy in the classification of attacks and training time in the KDD99 data set. It will also be attempted to learn a better way to classify each type of four attacks (Probe, Dos, U2R, R2L).

Several researchers have used various concepts to reduce the features. The very obvious and basic concept that can be gainfully used could be the amount of information actually contained in the different features of KDD CUP 99 data. In our work maximum information gain ratio (entropy) calculation is made the basis for minimizing the number of features [2].

We calculate the entropy set with k different values given by:

Entropy (Set) = I (Set)=$\sum_{i=1}^{k} P(valuei).\log_2 P(valuei)$

In above formula probability of getting the ith value representing by P(value i).

First we consider all the features and there after gradually reduce the number of features and compare the information gain. It is found that the change in information gain with all the features and with 18 to 20 features is almost same others are changed.

Information gain ratio is sorted in a descending order for all attribute of KDDCUP99 dataset. The average of information gain is 0.22. For most of the features we are getting under the average Information Gain Ratio (IGR). In Fig 1, shows information gain with average and red line is representing average.

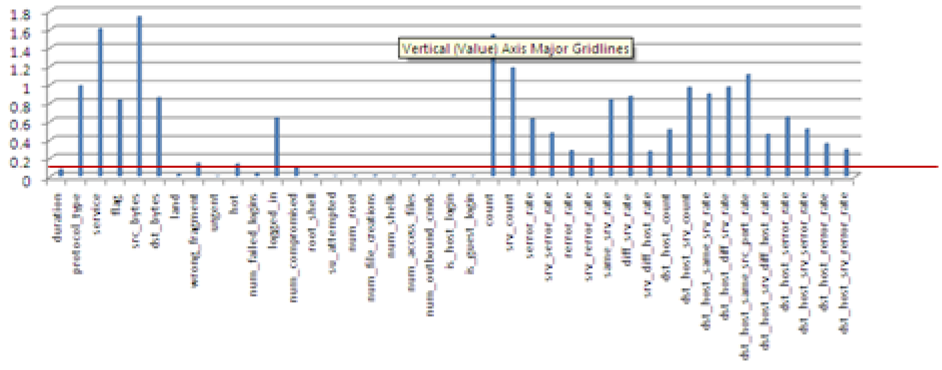

Fig1. Information Gain Ratio (IGR) under the average of the data set

## 3. Selection of Classifiers

In this paper we provide a sort introduction to some classification technique. The various techniques of intrusion detection system reported in this section.

### 3.1. K-Means

To solve the main clustering problem K-means [3] is most commonly and simplest unsupervised learning algorithms, which can also do the automatic partition of a date into k groups. The main motive of define k centroids and then relating them, is to assign each instance to its closest cluster center and to update each cluster center to be the mean of its constituent instances.

### 3.2. Bayes Net

Bayesian net also referred as belief networks belongs to the family of probabilistic graphical models. These are used to signify knowledge graphical structures of this model are used to represent knowledge about an ambiguous domain of dataset. The probabilistic dependencies among the corresponding random variables are represented by the edges of this model. Formally, nodes represent variables, and the arcs encode conditional dependencies between the variables. Bayesian networks are Directed Acyclic Graphs (DAG). The states of the random variable and a Conditional Probability Table (CPT) contain in each node.

### 3.3. Naïve Bayes

The probabilistic learning method [4] is used in the Bayesian classification. Naïve Bayes provide a simple approach which is based on probabilistic graphic models, a particular model specifies the probabilistic dependencies underlying with the help of graph structure. The naive Bayesian classifier gives us a simple approach with clear semantics, representing, using and learning probabilistic knowledge for the supervised induction tasks. This method is designed in which presentation aim is to accurately predict of class of test instances which also include class for the training instances. Such a classified and specialized form of Bayesian network as naïve because it depends on two important simplifying assumptions.

### 3.4. J48

J48 is an open source classifier of the C4.5 algorithm and its implementing in Java. C4.5 is a program that creates a decision tree based on a set of labeled input data[5]. This algorithm was developed by Ross Quinlan. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier. The J48 algorithm is designed with those features which easily address the loopholes that are present in ID3. The main disadvantage of C4.5 was that the CPU took time and a system memory was required. For the classification of problems decision tree is used. In this technique, the model is based on a tree for the classification process. Once the tree is constructing, the classification result have to applied each tuple in the database.

### 3.5. ID3

ID3 algorithm, developed by J. Ross and Quinlan back in 1979, is an example of symbolic learning and rule induction [6], machine learning technique to classify data. It is a supervised learning algorithm that employs decision tree based on mathematical calculations. it conducts top-down greedy search through given training set to test each attribute at every node, to construct a decision tree. ID3 is a very useful Decision learning algorithm.

### 3.6. NBTree

The NB Tree is a highly scalable, hybrid approach for large databases. Generally, it outperforms decision trees and naive bayes classifier alone [7]. it is suitable for cases where many attributes are relevant for classification. in such cases, database is large and interpretability of classifier is desired, and attributes are not necessarily independent (i.e. attributes are not conditionally independent). NBTree significantly improves upon the performance of its constituents by inducing highly accurate classifiers. Even though no single classifier outperforms all others in every domain, NBTree performs well in most cases as well as scales up well with respect to accuracy. As in decision trees, threshold for continuous attributes is chosen using standard entropy minimization technique.

### 3.7. Fuzzy Logic

Fuzzy logic, proposed by LoftiZadeh in the 1960s, although a relatively newer theory, has proven its worth in a number of industrial applications [8]. Fuzzy approach lends ability to apply logic on soft values (fuzzy sets or “degrees of truth”) rather than hard values (crisp or true/false), to make the underlying reasoning framework ( expert systems, decision trees, etc.) more generalized, enabling it to be suitable for a wide range of problems.

### 3.8. Support Vector Machine

Support Vector Machines (SVMs), a new generation of learning algorithms, set of related supervised learning algorithms, was developed by Vladimir Vapnik in the mid 90’s [9]. SVMs are used for classification and regression. SVM are at the forefront of the Machine Learning field, owing to its elegance and rigorous mathematical foundations from optimization and statistical learning theory. It is formally defined by a separating hyperplane, making it a discriminative classifier. Thus, For a given set of labeled training data, SVM gives an optimal hyperplane to categorize new examples.

### 3.9. Decision Table

Comprising of two attributes at each level of the hierarchy [10], decision table is a hierarchical breakdown of the data. Important columns (Attributes) for classifying data, are identified, and the resulting model is graphically displayed as series of cake charts, with the help of accompanying visualize. Several levels, representing decreasingly important attributes, can be contained in the visualization. This is done with the help of cakes, wherein, each cake can be subdivided into smaller cakes to represent next most important attributes Pair.

### 3.10. JRip

JRip is a rule induction algorithm, proposed by Cohen W.W. in 1995. It was introduced as a successor of IREP algorithm. It implements a propositional rule learner. The initial set of rules for the RIPPER (Repeated Incremental Pruning to Produce Error Reduction) class is generated using incremental reduced-error pruning [11]. RIPPER employs separate-and-conquer strategy to learn such rules in greedy manner. Based on corresponding class frequencies, the training data is sorted in ascending order by class labels. Starting with the smallest, rules are learnt for m-1 classes. The instances covered by the rule thus created are then removed from the training data set. this is repeated until all the instances from the target class are removed. This is repeated for the remaining classes till all the rules are learnt. For the last class, i.e. most frequent class, a default rule with empty antecedent is added.

### 3.11. OneR

One Rule (OneR, Witten I H, 2005) algorithm is an algorithm based on Rule based model, wherein, a one-level decision tree is generated in the form of a set of rules that tests a particular attribute[12]. It finds one attribute to base prediction upon, that makes fewest prediction errors. OneR is a simple yet effective method that generates efficient rules for characterizing structures in data. A single predictor value is used to induce classification rules. For each predictor in the data, a single rule is generated. The rule with smallest error is then selected. The rule is generated as follows:

- Create a frequency table for each predictor
- Determine the most frequently occurring class
- For each predictor, compute total error for the rules
- Select the predictor with smallest total error

### 3.12. MLP

Multi Layer Perceptron (MLP), a feed-forward ANN (Artificial Neural Network) model, is a widely used neural network classification algorithm. MLP maps set of input data to suitable output set[15]. MLP is a directed graph comprising of multiple layers, where, each layer is fully connected to the next one. MLP employs back propagation, a supervised learning technique for training the network. It can classify data that are not linearly separable.

### 3.13. Self Organizing Map (SOM)

Self Organizing Map (SOM), an unsupervised learning technique, is a type of Artificial Neural Network (ANN). Instead of error-correction learning approach, it employs competitive learning approach[14]. In competitive learning, output neurons compete with each other to get activated as only one output neuron is activated at a time. Negative feedback paths or (lateral) inhibition connections between neurons are used to ensure that there is only one winning neuron. SOM maps input of arbitrary dimensions onto output of regular, low-dimensional array. Clustering of given input data can be visualized with the help of SOM. In SOM, the neurons are represented by K-Dimensional vectors, where K is the number of parameters used to characterize the input space.

### 3.14. IBK (K - Nearest Neighbour)

IBK also known as k-nearest-neighbour Algorithm [13], is a supervised learning algorithm. In its training phase, a supervised learning algorithm generates classifying function from the training set. In IBK, in the training phase, feature vector of the training set is stored. Then, in classifying phase, for given input data (unclassified), represented as a vector, the class label most frequent among its K nearest neighbors (initial neighbourhood information from training phase) is assigned to it. IBK gives strongly consistent results. However, one of the drawbacks of K-NN is that equal weightage is given to each of the attributes, whereas in some cases it might be desirable to afford more weightage to some parameters over others.

### 3.15. Random Forest (RF)

Random Forest (RF) employs ensemble approach to combine bagging and random selection of features, to generate a forest (multitude) of decision trees with controlled variance to correct the problem of over-fitting as in case of decision trees[13]. A tree is grown by splitting a node to search a random subset of available decisions. A Random Forest (RF) of such trees is grown by projecting the training data into a randomly selected subspace before fitting each tree. The object to be classified is pushed down each of the trees. To get accurate results, the RF must have large number of trees, which makes it slower. Thus, even though RF are accurate and can be trained fairly quickly, they take longer to make predictions.

## 4. Results and Discussion

We have analyzed the performance comparison with all fifteen classifier algorithm. Further, we generalized the empirical results with two models for algorithm selection. Here we have observed that for a given attack category certain subset of classifier algorithms offer enhance performance over single classifiers. We identified the best result in form of True Positive (TP) shown in Fig 2, False Positive (FP) in Fig 3. In fig 4, Total correctly classified (CC) and total Time Taken (TT) algorithms for each attack in Simulation results are shown.

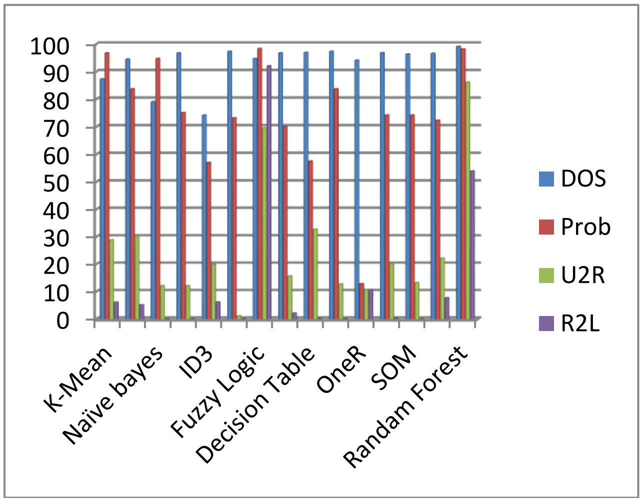

Fig 2: True positive result in different classifier

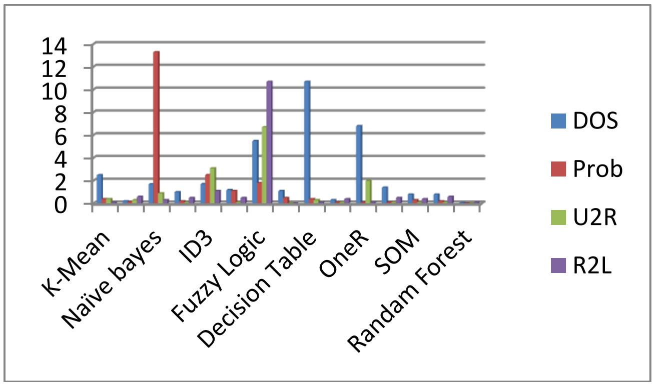

Fig 3: False positive result in different classifier

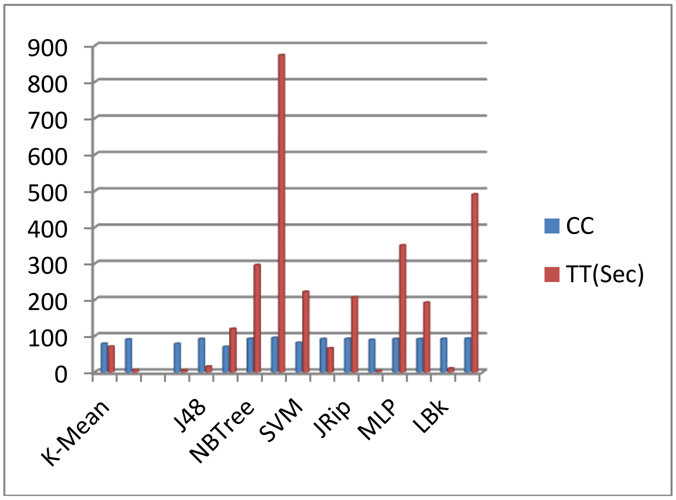

Fig4: Total correctly classified instances and total time taken

## 5. MODEL EVALUATION AND DISCUSSION

The Best Results are shown after simulation Fuzzy Logic and Random Forest are best but total time taken is very large. We create two model first best performances and other model is use as per min time taken and gives better results.

From the table 1 it is clear that the following methods give the best results for each of the above 4 attack categories.

| | Best true positive result | Worst false positive result |
|---|---|---|
| DoS: | Random Forest Classifier (TP 99.2) (TT 491) | Random Forest Classifier(FP .05) (TT 491) |
| Probe | Fuzzy logic (TP 98.4)(TT 873.9) | Random Forest Classifier (FP .01) (TT 491) |
| R2L | Fuzzy logic (TP 92.1) (TT 873.9) | Support Vector Machine (SVM) (FP 0)(TT 222.28) |
| U2R | Random Forest Classifier (TP 86.2) (TT 491) | Support Vector Machine (SVM) (FP 0.02) (TT222.28) |

Table1 Best TP and Worst FP in all class

We then propose a model for classifier selection as in Fig.5.

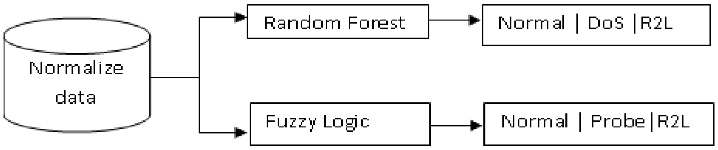

Fig 5 Model 1 as per good performance

The fig 5 depicts that IDS system with data mining capabilities are flexible in choosing the classifying method that best to deal with attack. Moreover, it is equally important to judge whether the selected algorithm can be implemented in real time IDS system. We have also suggested another model for real time algorithm selection showing in fig 6. This model has significant meaning with low TT for each attack.

| | Best true positive result | Worst false positive result |
|---|---|---|
| DoS: | J48 (TP 96.8) (TT 15.85 Sec) | Bayes Net(FP .2) (TT 6.28 Sec) |
| Probe | K-Means (TP 96.8)(TT 70.7 Sec) | J48 (FP .2) (TT 15.85 Sec) |
| R2L | One-R (TP 10.7) (TT 3.75 Sec) | One-R (FP 0.1) (TT 3.75 Sec) |
| U2R | Decision Table (TP 32.8) (TT 66.24 Sec) | Decision Table (FP 0.3) (TT 66.24 Sec) |

Table 2 Best TP and Worst FP in Minimum TT

In table 2 we find best classifiers that give best result when considering minimum time taken.

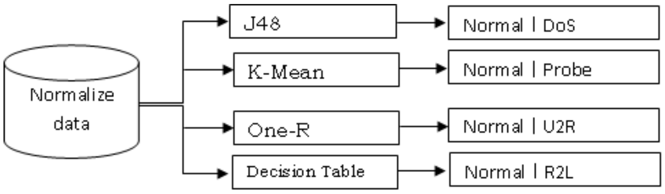

Fig 6 Model 2 as per min Time taken

Table 3 Performance comparison between the two models and Max Positive results Models with KDD Cup Winner.

| | | Dos | Prob | U2R | R2L |
|---|---|---|---|---|---|
| Max Positive Results | TP | 99.2 | 98.4 | 92.1 | 86.2 |
| | FP | 0.05 | 0.01 | 0 | 0.02 |
| Model 1 | TP | 99.2 | 98.4 | 92.1 | 86.2 |
| | FP | 0.05 | 1.8 | 10.7 | 0.17 |
| Model 2 | TP | 96.8 | 96.8 | 30.30 | 10.70 |
| | FP | 1.00 | 0.13 | 0.30 | 0.10 |

Table 3 shows the performance comparison of the two proposed multi-classifier model

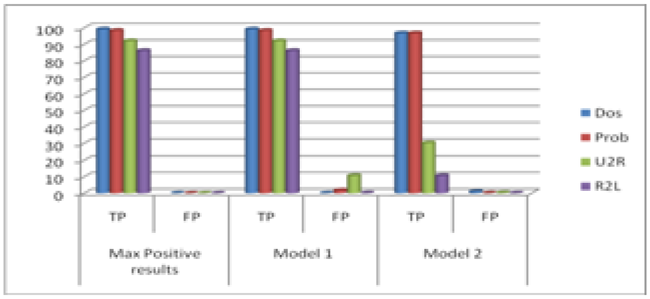

Fig 7: shows the performance comparison of the two proposed multi-classifier

The results suggest that the two proposed models showed in fig 7. In model 1 minor improvement in best TP for other single classifiers for DoS and Probe and significant improvement for U2R and R2L attack categories. Also, FP was reasonably small for all attack categories.

## 6. CONCLUSIONS

In real system, When the models are practically deployed there might have certain potential problem though there is superiority in numeric comparison between the proposed models. We have to hardcode the algorithms for deploying of a system with multiple algorithms which is inflexible. The resource requirements are another problem when the models are implemented and finally, a comparison between the proposed models and a multiple classifiers selection (MCS) system can be made. Above mentioned issues may be well solved if we will develop one another model which adoptive and scalable. The approach will be adaptive because depending up on the system load and use scenarios, less or more number of detectors could be applied thus adapting according to system load and level and type of intrusions. The detection model architecture is such that any number of patterns/detectors can be easily deployed without much computational overhead, the approach is scalable.

## REFERENCES

[1] MahbodTavallaee, EbrahimBagheri, Wei Lu, and Ali A. Ghorbani" A Detailed Analysis of the KDD CUP 99 Data Set" Proceedings of the 2009 IEEE Symposium on Computational Intelligence in Security and Defense Applications (CISDA 2009).

[2] Chebrolu, Srilatha, Ajith Abraham, and Johnson P.Thomas."Feature deduction and ensemble design of intrusion detection systems." Computers & Security24, no. 4 (2005): 295-307.

[3] FarhadSoleimanianGharehchopogh, Neda Jabbari, ZeinabGhaffari Azar "Evaluation of Fuzzy K-Means And K-Means Clustering Algorithms In Intrusion Detection Systems" INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 1, ISSUE 11, DECEMBER 2012 pp 66-72.

[4] John, G.H., Langley, P.:Estimating Continuous Distributions in Bayesian Classifiers. In: Proc. of the 11th Conf. on Uncertainty in Artificial Intelligence (1995).

[5] M.Revathi and T.Ramesh" NETWORK INTRUSION DETECTION SYSTEM USING REDUCED DIMENSIONALITY" Indian Journal of Computer Science and Engineering (IJCSE) Feb2011 PP 61-67.

[6] Mary Slocum "Decision making using ID3" RivierAcadmic Journal, Vol 8, No 2, 2012.

[7] Dewan Md. Farid, Jerome Darmont and Mohammad Zahidur Rahman" Attribute Weighting with Adaptive NBTree for Reducing False Positives in Intrusion Detection" International Journal of Computer Science and Information Security, Vol. 8, No. 1, 2010 PP 19-26.

[8] Alma Husagic-Selman” Intrusion Detection System using Fuzzy Logic” SOUTHEAST EUROPE JOURNAL OF SOFT COMPUTING Vol 2 No 1 March 2013 PP 14-20.

[9] Daniele Loiacono, Andrea Marelli, Pier Luca Lanzi” Support Vector Regression for Classifier Prediction” ACM GECCO’07, July 2007 pp 1806-1813.

[10] VANTHIENEN, J., G.WETS & G. CHEN (1996) “Incorporating fuzziness in the classical decision table formalism”. International Journal of Intelligent Systems. Vol. 11 (11), pp. 879-891.

[11] W.NorHaizan W. Mohamed, MohdNajibMohdSalleh, Abdul Halim Omar” A Comparative Study of Reduced Error Pruning Method in Decision Tree Algorithms” IEEE International Conference on Control System, Computing and Engineering, 23 - 25 Nov. 2012, Penang, Malaysia.

[12] MsS.Vijayarani ,MsM.Muthulakshmi “Comparative Analysis of Bayes and Lazy Classification Algorithms” International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 8, August 2013 pp 3118-3124.

[13] PhyuThiHtun, KyawThetKhaing “Anomaly Intrusion Detection System using Random Forests and k-Nearest Neighbor” International Journal of P2P Network Trends and Technology Vol. 3, Issue 1, August 2012 pp 67-71.

[14] Mia Louise Westerlund “Classification with Kohonen Self-Organizing Maps” Soft Computing, Haskoli Islands, April 24, 2005

[15] GurselSerpen and Zhenning Gao “Complexity Analysis of Multilayer Perceptron Neural Network Embedded into a Wireless Sensor Network” Conference Organized by Missouri University of Science and Technology 2014- Philadelphia, PA Procedia Computer Science 36 ( 2014 ) pp 192 – 197.