Optimized Ensemble Weak Learner Tree Based Network Efficient Intrusion Detection and Alert System using Data Mining

K. Mohanapriya, M.Savitha Devi

Abstract: ***The modern society accesses various network services through different devices. However, the services afford by the service provider faces various challenges and threats. The services are facing different network threats towards degrading the service performance or the entire network. Number of approaches discussed earlier to restrict the illegal access from malicious users which uses different properties in service level, packet level, user level features. However, they suffer to achieve higher performance in intrusion detection. To improve the performance in intrusion detection an novel tree based ensemble learner algorithm has been proposed in this paper. The method incorporates Random Forest and Random Trees, which are identified as NP complete. The method maintains the list of ensembles which are indexed under trees. At the classification, the Tabu Search algorithm has been used which measures the ensemble class weight (ECW) which has been used to perform classification. According to the result of intrusion detection, an alert has been generated to the administrator. The proposed algorithm improves the performance of intrusion detection.***

Keywords: Network Intrusion Detection Systems (NIDS), Decision trees, Random Forest, Random Trees, Ensemble Weak Learner Tree and Tabu Search (TS).

I. INTRODUCTION

Data mining has becomes a very useful technique to reduce information overload and improve decision making by extracting and refining useful knowledge through a process of searching for relationships and patterns from the extensive data collected by organizations. “The extracted information is used to predict, classify, model and summarize the data being mined. Data mining technologies, such as rule induction, neural networks, genetic algorithms, fuzzy logic and rough sets are used for classification and pattern recognition in many industries”. They have been extensively used in discriminating normal from abnormal behavior in a variety of contexts. In recent years data mining techniques have been successfully used in the context of network intrusion detection. The recent rapid development in data mining has made available a wide variety of algorithms, drawn from the fields of statistics, pattern recognition, machine learning, and database.

Firewalls are not to detect all types of malicious network traffic and computer usage . This includes network attacks such as unsafe services, privilege violations, unauthorized logins etc., IDS has the following three components: Sensors: - used to find the network traffic or system activity. Console: - to monitor every instance and control the sensors, Detection Engine: - records events in a database using sensor and uses a system of rules to generate alerts . Figure 1 presents the basic architecture of IDS.

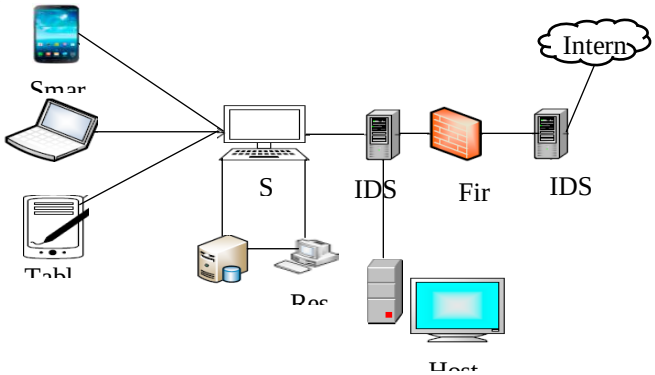

Figure 1 Intrusion Detection System Architecture

Data mining based intrusion detection techniques are generally comes under any one of the two categories: misuse detection and anomaly detection. Researchers use various classification algorithms, Cost sensitive modeling and association rules to classify every instance in a data set to detect the network intrusions first and named all other as normal activities. Anomaly detection algorithms build models of normal behavior first and automatically all others are named as intrusive and unknown attacks can be identified efficiently in this technique. Main drawback of this technique is all the dubious or unusual activities are considered as intrusive.

Supervised and unsupervised are two main anomaly detection techniques. In supervised anomaly detection, Known trained test data are used to build the normal behaviour. Unsupervised anomaly detection detects anomalous behavior without knowing knowledge about the training data and use clustering approach, outlier detection schemes, state machines, etc.

The IDS are of three types based primarily on the events monitored and the deployment. They are Network-Based, the Host Intrusion Detection System, Network Behavior Anomaly Detection. Host-based Intrusion Detection Systems (HIDS) detects traffic in network and system specific settings, while Network Behavior Anomaly Detection (NBAD) determines the existence of anomalies in the quantity and type of traffic by monitoring the traffic on network segments. The Network Intrusion Detection System (NIDS) is a familiar IDS type in which all layers of the

Revised Manuscript Received on January 2, 2020.

K. Mohanapriya, Guest Lecturer, Department of Computer Science, Government Arts College for Women, Krishnagiri, Tamilnadu, India E-mail : kmpriya4@yahoo.co.in

Dr.M.Savitha Devi, Asst. Professor, Department of Computer Science, Periyar University Constituent College of Arts & Science, Harur, Tamilnadu, India Email: madhubalasivaji@gmail.com

Open Systems Interconnection (OSI) model are analyzed for traffic in the network so that it can take decisions based on the intention of the traffic to identify suspicious activity. Most of the NIDSs are easily deployed on a network and it has the ability to examine the traffic from many systems at an instance. “Wireless Intrusion Prevention System” (WIPS) is a term commonly used by the vendors portray the Intrusion network system that detects and investigates the wireless radio spectrum for intrusions. It monitors the network traffic in a particular segment or devices of a network and takes countermeasures; it also examines the application and network layer protocol activity to identify suspicious activity .

Multiple events of interest can be identified through this. In general, deployment takes place at a boundary between networks such as near to border firewalls or routers, servers of Virtual Private Network (VPN), remote access servers and wireless networks. The other name for NIDS is passive IDS as administrator system informs these kinds of systems where an attack has taken place and sufficient measures are taken to assure the security system. The objective is to inform about an intrusion to identify IDS that can react in the post. In spite of reports of insufficient damages, IDS should react and block even doubtful traffics. Active IDS is implied through these reaction techniques.

Decision tree is a decision support model and its algorithm splits the dataset of records using any one of the approaches such as depth-first greedy approach or breadth-first approach. Decision tree structure has consists of three nodes, first one is a root node and it is in the top of the structure, next one is internal node and test condition on an attribute is represent in the node and last one is leaf node and class label is represented in it, also known as terminal node.

Tree pruning is used to minimize the over-fitting problem in decision tree. Some of the decision tree algorithms like ID3,C5.0 and CART has the quality of easy construction, better learning ability and classifying speed

The ensemble learning is a method of combining two linear regression models. AdaBoost is one of the Ensemble learning algorithm and it is strong classifier by combining two weak classifiers. Ensemble learning technique is introduced for the purpose of overcoming decision tree limitations.

A key approach to guide, alter and control other heuristics so as to obtain solutions that are better than the ones generated by local heuristics is a Meta heuristic. Their specialty is they avoid getting trapped in the local optima. They also carry out single search in the neighborhood and one befitting example for this is the Tabu Search.

II. RELATED WORKS

Kyoto 2006+ dataset which is a new labeled network dataset was put forth by Sahu&Mehtre [1]. In this set, every instant was labeled as normal (no attack), attack (known attack) and unknown attack. The Decision Tree (J48) algorithm was used to classify the network packet used for NIDS. Totally 134665 network instances were used for training and testing. In detecting the connection (no attack, known attack, unknown attack), the generated rules provided 97.2% correctness.

Desale et al., [2] presented the mechanism that improves the efficiency of the IDS using streaming data mining technique. The four selected stream data classification algorithms was used on NSL- Knowledge Discovery Database (KDD) datasets and their results were compared. To improve the efficiency of IDS comparative analysis of their results was done.

The design of distributed ID framework was first introduced by Folino et al., [3], specifically, the detector module with its basis of meta-ensemble was used in coping with the issue of intrusion detection where compared to normal connection, the number of attacks were less. This method explored role of ensemble in detecting particular attack or normal connections. To combine every specialized ensemble the Genetic Programming was adopted to generate a non-trainable function. Without any extra phase of training, non-trainable functions can be evolved which can appropriately handle concept drifts, also with regards to real-time restrictions. Preliminary experiments were conducted on well-known KDD dataset and also on a more up-to-date dataset, ISCX IDS, showed the effectiveness of the approach.

A hybrid technique for Intrusion Detection was proposed by Dubey&Dubey [4] with its basis on K-means Naive-Bayes and Back propagation neural network (KBB). The k-means which was applied initially was partition-based, unsupervised cluster analysis method. The gathered data was obtained in the form of clusters to easily process and learn in any machine learning algorithm. Fit and essential data attributes are obtained when the outcomes are processed through the bayesian classifier based on probability model which is a supervised. Back propagation neural network performed filtered data learning which was able to learn the patterns with less number of training cycles. This method used KDD cup99's dataset. The bayesian classifiers are used to detect attacks as DoS, U2R, R2L, and probe. In this method classification and performance of the classifier was focused. So, for filtering data set features, various classification algorithms are applied.

Sesmero et al., [5] addressed the critical issue like members of the ensemble, their learning parameters and the learning algorithm for generating the meta-classifier during the process of selecting an algorithm for classification. The suitable amalgamation of base learning algorithms and their learning parameters were selected manually generally. Automatic methods were also used in other approaches for determining the suitable stacking configurations instead of starting from these strong initial assumptions. The Stacking and its variants and several examples of application domains were presented.

III. METHODOLOGY

This section detail about the random forest, random tree and Tabu search, they are used to categories the intrusions. In this work, KDD99 dataset is used to categories the

intrusions using various classifiers and their performance is analyzed and to improve the classifiers performance, a proposed Tabu search method is introduced and the method is a combination of any of ensemble technique such as random forest or random tree with Tabu search.

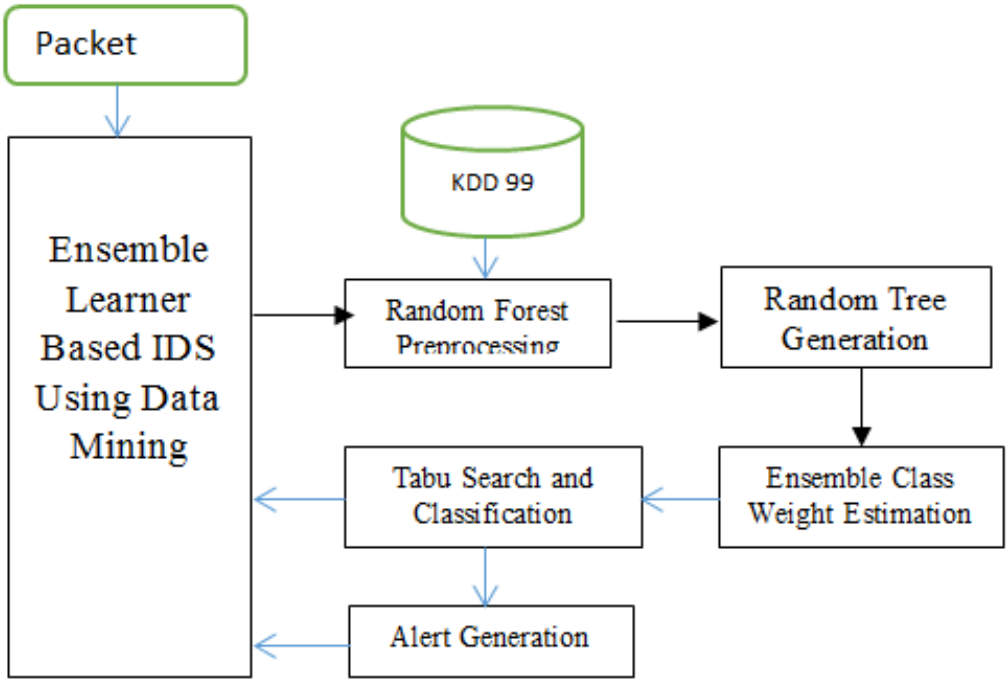

Figure 2 Block Diagram of Proposed Methodology

The Figure 2, present the block diagram of proposed Intrusion Detection algorithm. The method has different functional components and each has been discussed in detail.

3.1 Random Forest Preprocessing:

The method starts with the preprocessing of KDD 99 data set which has been introduced by the Defense academy of United States. The data set has number of attributes and all of them cannot be used in intrusion detection. The random forest algorithm is a data mining algorithm which has been used for classification. Initially, the method generates number of trees for different class of features. The input data points has been read, and each has been validated for the presence of all the features. The data points which are identified as incomplete has been eliminated from tree generation. The elimination is performed based on Gini impurity function. The points identified as complete are used to generate multiple tree of random forest. N records of training set are randomly selected from original data and Boot strap sample is done on it for constructing a tree. The tree is grown with splitting attributes to the maximum possible range to split the nodes. Randomly selected m input variables at every node is to be less than the size M. Variable m remains constant when forest growing.

The Bootstrap sample is done on the N records of the training set; these records are sampled from the original data at random but with replacement. This happens to be the training sample set for constructing the tree. The tree is grown to the largest extent possible and the process of growing the tree is by splitting the attributes to split the nodes. For an input variable of size M, a number m less than M is chosen so that m variables are selected randomly at every node. The best split on these m attributes is used to split the node. During the forest growing stage, the value of m remains constant.

Algorithm:
Input: KDD99 Data set Kds
Output: Preprocessed Kds
Start
Read Kds.
Identify list of attributes Al = $\sum_{i=1}^{size\,(Kds)}(Kds(i)(A) \ni Al \cup Al(A)$
For each attribute A
Compute values for each attribute.

$$GiniIndex(A) = Gini(Class) - \sum_{j=1}^{m} P(a_j).Gini(A = a_j)$$

End
For each data point Di
Compute impurity of data point as below:
Gini(Di) = $\sum_{i=1}^{size\,(Al)} P(Al(i))^2$
If Gini>Th then
Leave
Else
Kds =$\sum(Dk \in Kds) \cap Di$
End
Stop

The above discussed algorithm estimates the impurity value on each attribute and class. According to that the method estimates the impurity value towards all the data points and according to that the data points which are identified as incomplete has been removed from data set.

3.2 Random Tree Generation:

In this stage, the method generates random tree based on the result preprocessing. A repetitive division of the given data space is used for representing the decision tree. The Decision Tree (DT) comprises a rooted tree. A node is used for directing this, and is known as the root, the root is the main component and the leaves refer to the remainder of the nodes. Striving to optimize the cost function, the decision tree classifier determines the decision tree T, given a set of L labeled samples. Here, after optimizing the decision tree, it strives to determine an optimal class from a data set that has been given, when a query image has been provided as a test case.A 10-fold cross-validation was used since there was no training data. The levels below the roots were limited to five, parent and child nodes were set at 10 and 5. This algorithm creates a tree where the terminal nodes are events that are classified as intrusions or they can also be created in such a way that the cost of misclassification can be minimized.

3.3 Ensemble Class Weight Estimation:

In this stage, the input packet has been accepted and the features from the packet like hop count, payload, latency has been extracted. Extracted features have been measured for ECW towards different decision tree available. Based on the features, the method estimates the ECW value for various classes. Estimated ECW value has been used to perform intrusion detection.

Algorithm:

Input: Decision Tree T, Packet P

Output: ECW

Start

Read packet P.

Extract Pay load Pl = P.payload

Extract Hop count Hc = $\sum Hops \in P.Route$

Extract Latency Lat = P.Latency

Estimate ECW

=

$$\frac{pl}{\sum_{i=1}^{size\,(T)} T.pl \,/_{size\,(T)}} \times \frac{Hc}{\sum_{i=1}^{size\,(T)} T.Hc \,/_{size\,(T)}} \times \frac{Lat}{\sum_{i=1}^{size\,(T)} T.Lat \,/_{size\,(T)}}$$

Stop

The above discussed algorithm estimates the ensemble class weight towards a specific class which has been used to classify the packet.

3.4 Proposed Tabu Search (TS)

Initially, the TS was incorporated to be a technique for local search at a high level using the right approaches so that the search could be guided towards effective search space explorations, in such a way that the trapping in the local optima could be avoided To avoid the repeated movement of the solution that has been visited earlier, a Tabu list is introduced. All the moves that are successful are updated in the Tabu list. The tabu list is generated by computing the ensemble class weight ECW. Every time a new solution or move is found, it is compared to the entry in the Tabu list to check for the occurrence in the list. The new solution is discarded if it is already on the list so the repeated search for same solutions are avoided. The next movement takes places and the new solution is updated in the Tabu list if it does not match with the entry in the list stored earlier. The possibility of cycling is decreased due to the use of Tabu list hence the solutions that are visited recently are not cycled to return to the same solution forming a loop after certain iteration. All the new solutions are stored in the Tabu list and the best outcome is chosen as the best solution X_{next} . The moves that were carried out most frequently and recently are stored in the Tabu list hence the local minima problem is overcome. In Tabu search solution, each Tabu list is represented by random trees. Tabu search is used to find the optimal depth of tree and the number of trees.

Random Forest and Tabu Search Algorithm:

$\{User\ Setting\}$

$Input\ N,\ M,\ S,\ NG$

$\{Process\}$

$\vec{RF} = Call$ Random Forest (N, M)

for i=1 $\rightarrow$ S do

for k=1 $\rightarrow$ n do

x=Random(1 $\rightarrow$ N)

Add tree RF_x to forest i in the Tabu Search's Population $\vec{P}_i$

end for

end for

Evaluate each forest in the initial population $\vec{P}$

for j=1 $\rightarrow$ NG do

{Generate a new population by applying Tabu Search: Neighbor generating, random solution, tabu list creation}

$\vec{PNew} = Tabu$ operation ($\vec{P}$)

Evaluate each forest in $\vec{P}$

$\vec{bestforest} \leftarrow copy$ of best $\vec{P}$

$\vec{P} = \vec{PNew}$

end for

{Output}

A vector of Trees $\vec{bestforest}$

Where N is the size of the training set, NG is the size of neighbors, M is the total number of attributes in training set, S is the solution space.

The class labels are assigned with random values initially and an arbitrary parameter set is taken. The class labels are then estimated by the proposed Tabu search algorithm, they are then evaluated and the parameters are updated until the termination criterion is achieved that is a maximum number of iterations. In the hybrid algorithm, two approaches have been validated i.e using random trees in the Tabu list and random trees in Tabu list.

The initial random class labels are subjected to random forest and the parameters of the forest are updated in the Tabu list, the forest is evaluated for all events in the dataset to find the best solution, this best class label is updated in the Tabu list and the Tabu search process takes place and the new solution is compared with the list created to ultimately find the best forest as the solution to the classification process. Similarly, the initial random class labels are subjected to the random tree and get updated in the Tabu list, the DT is evaluated for all events in the dataset to find the best class label. The Tabu search process takes place and the new solution is compared with the list created to ultimately find the best Random trees as the solution to the classification process.

IV. RESULTS AND DISCUSSION

Analysis is carried out with 9711 normal and 12833 abnormal data stream and it was taken from KDD99 dataset. Tabu search is used to find the optimal depth of tree and the number of trees. Table 1 and Figure 2 to 5 shows the Classification Accuracy, Precision, Recall and F Measure respectively.

Table 1 Results

	Classification accuracy	Precision	Recall	F Measure
Random Forest	0.9328	0.9311	0.932	0.9315
Random Tree	0.9438	0.9424	0.9431	0.9427
Proposed RF-TS	0.9555	0.9541	0.9554	0.9547
Proposed RT-TS	0.9626	0.9617	0.9621	0.9619

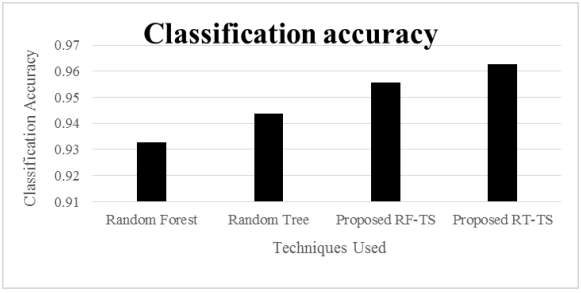

Figure 2 Classification Accuracy

It is observed From table 1 and figure 2 that the classification accuracy of Proposed RT-TS performs better than Random Forest by 3.14%, better than Random Tree by 1.97% and better than Proposed RF-TS by 0.74%.

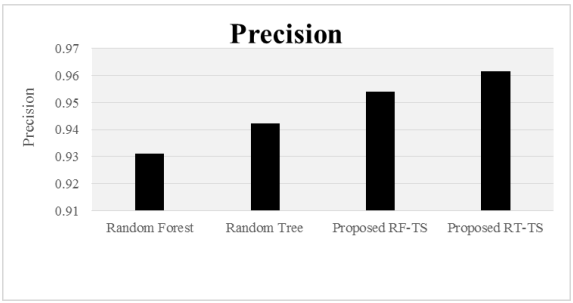

Figure 3 Precision

From table 1 and figure 3, it is observed that the precision of Proposed RT-TS performs better than Random Forest by 3.23%, better than Random Tree by 2.03% and better than Proposed RF-TS by 0.79%.

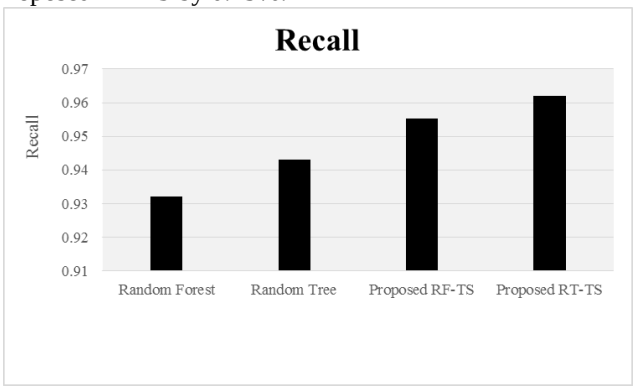

Figure 4 Recall

It is noted From table 1 and figure 4 that the recall of Proposed RT-TS performs better than Random Forest by 3.18%, better than Random Tree by 1.99% and better than Proposed RF-TS by 0.69%.

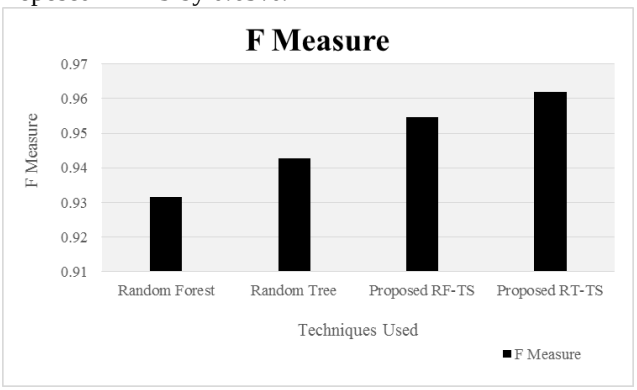

Figure 5 F Measure

From table 1 and figure 5,states that the F Measure of Proposed RT-TS performs better than Random Forest by 3.21%, better than Random Tree by 2.02% and better than Proposed RF-TS by 0.75%.

V. CONCLUSION

The key ideas is to use data mining techniques and discover consistent and useful patterns of system features that describe network behaviour, and use the set of relevant system features to recognize anomalies and known intrusions. Being a Meta heuristic technique, the Tabu

Search locally guides the search procedure for exploring a solution space and avoiding the local optima. Thus, the search is more flexible because of the list is identified by computing the ECW value. Once the method identifies the intrusion, an alert mail has been sent to the administrator to take necessary action. The output of the work revealed that the classification accuracy of Proposed RT-TS performs better than Random Forest by 3.14%, better than Random Tree by 1.97% and better than Proposed RF-TS by 0.74%.

REFERENCES

1. Sahu, S., &Mehtre, B. M. (2015, August).Network intrusion detection system using J48 Decision Tree.In Advances in Computing, Communications and Informatics (ICACCI), 2015 International Conference on (pp. 2023-2026).IEEE.
2. Desale, K. S., Kumathekar, C. N., &Chavan, A. P. (2015, February). Efficient intrusion detection system using stream data mining classification technique. In Computing Communication Control and Automation (ICCUBEA), 2015 International Conference on (pp. 469-473). IEEE.
3. Folino, G., Pisani, F. S., &Sabatino, P. (2016, March). A distributed intrusion detection framework based on evolved specialized ensembles of classifiers. In European Conference on the Applications of Evolutionary Computation (pp. 315-331).Springer, Cham.
4. Dubey, S., &Dubey, J. (2015, September). KBB: A hybrid method for intrusion detection. In Computer, Communication and Control (IC4), 2015 International Conference on (pp. 1-6). IEEE.
5. Sesmero, M. P., Ledezma, A. I., &Sanchis, A. (2015). Generating ensembles of heterogeneous classifiers using stacked generalization. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 5(1), 21-34.
6. Han, J., Pei, J., &Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.
7. Dr.M. Savithadevi, Mohanapriya.K(2018,August). Survey of classification on network intrusion detection using data mining techniques, IARA India Volume VI.16-21

AUTHORS PROFILE

K. Mohanapriya, Guest Lecturer, Department of Computer Science, Government Arts College for Women, Krishnagiri, Tamilnadu, India E-mail : kmpriya4@yahoo.co.in

Dr. M.Savitha Devi, Asst. Professor, Department of Computer Science, Periyar University Constituent College of Arts & Science, Harur, Tamilnadu, India Email: madhubalasivaji@gmail.com