Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 7, Issue. 5, May 2018, pg.38 – 46

A STUDY ON PREDICTION OF MALICIOUS PROGRAM USING CLASSIFICATION BASED APPROCHES

N.Vaishnavi[1], K.Thiyagarajan[2]

[1]Research Scholar, Department of Computer Science, AVC College, Mayiladuthurai, India

[2]Associate Professor, Department of Computer Science, AVC College, Mayiladuthurai, India

[1] vaishumca1494@gmail.com

Abstract— Malicious programs pose a serious threat to computer security. Nowadays, malicious software attacks and threats against data and information security has become a complex process. The variety and range of those attacks and threats has resulted in providing numerous styles of relying ways in which against them however sadly current detection technologies of malware designers which use them to escape from anti-malware. However there is more needed to receive some better procedures which can guarantee the malware code recognition proficiently by testing strategy over an extensive arrangement of malicious executable. This paper explores the application of data mining methods to predict rootkits based on the attributes extracted from the information contained in the log files. In this paper, we proposed three algorithms name as Random Forest, Random Tree and Rep Tree using data mining techniques and the comparison of these algorithms.

Keywords— Data mining, Malware Code Detection, Random Forest Tree, Random Tree, Rep Tree, Classification, WEKA.

I. INTRODUCTION

Malicious code is one of the genuine dangers on the internet platform that is called malware. Malware is known as a Malicious application that has been clearly considered to harm the networks and PCs. Malware can be grouped into virus, worms, Trojans, spywares, adwares, and an assortment of different classes and subclasses that occasionally cover and obscure the limits among these classes. Data mining has been the focal point of numerous virus analysts in the current years to recognize obscure viruses. Various classifiers have been assembled and appeared to have high accuracy rates. The faust majority of these classifiers utilize highlights extricated from executable projects by applying figuring out strategies. Other than Binaries, Email corpses, Network traffic data are likewise mined for malicious activity. This research paper is presented as follows: section 2 lists on top of each other work. Section 3 presents the methodology and the aspect of detailed list algorithm. Section 4

elaborates experiments and finalizes the result produced by the algorithm. This paper, intensity on the data classification and the performance measure of the classifier algorithm based on the true positive rate, false positive rate, precision, recall, F-measure generated separately algorithm.

II. DATA MINING

Data mining is the process of discovering interesting knowledge, such as pattern, associations, changes, anomalies and significant structures from large amount of data stored in databases, data warehouses or other information repositories.

Due to the wide availability of huge amounts of data in electronic forms, and the imminent need for turning such data into useful information and knowledge for board applications including market analysis, business management, and decision support, data mining has attracted a great deal of attention in information industry in recent years[1,2]

Data mining has been popularly treated s a synonym of knowledge discovery in databases, although some researchers view data mining as an essential step of knowledge discovery.

In general, a knowledge discovery process consist of an iterative sequence of the following steps:

Data cleaning: Which handles noisy, erroneous, missing or irrelevant data.

Data integration: Where multiple, heterogeneous data sources may integrated in to one.

Data selection: Where data relevant to the analysis task are retrieved from the database.

Data transformation: Where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.

Data mining: Which is an essential process where intelligent methods are applied in order to extract data patterns.

Pattern evaluation: Which is to identify the truly interesting patterns representing knowledge base on some interestingness measures, and

Knowledge presentation: Where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

With the widely available relation database systems and data warehouses, The first four process: data cleaning, data integration, data selection, and data transformation can be performed by constructing data warehouses and performing some OLAP operations on the constructed data warehouses. The data mining, pattern evaluation and knowledge presentation processes are sometimes integrated into one (possibly iterative) process referred as data mining.

III. MALICIOUS PROGRAM

Malicious code is any code added, changed, or removed from a software system to intentionally cause harm or subvert the system's intended function [3]." It is however important to acknowledge that when fighting malicious code we care not of its intention but rather on its effect on the software. Code inserted with the most benign of intentions can have

the most malicious of effects. Take, for example, programmer negligence that manifests itself in the form of a buffer overflow. Previously, some of the more important sources of malicious code were third party renovation and remediation, off the shelve commercial/non-commercial systems that may spread pre-existing malicious code, and disgruntled employees/contractors or anyone else that might have access and the ability to insert malicious code into the system.

However, with the popularity of the Internet and the growing inter-connectivity of computers, systems are vulnerable to attacks, with or without human intervention, from just about anywhere regardless of physical distance or connectivity. Due to such connectivity, an attack can be propagated to a large number of machines in a relatively small amount of time, which may in turn propagate the attack to other machines causing a chain reaction. This ripple effect makes it equally as difficult to re-trace the attack back to its source, and there is no reason to believe it shall be easier to do so in the near future. Several categories and classifications exist to help narrow down the type of malicious code that one might be dealing with. The general categorization of malicious code has been along the lines of worms, viruses, Trojan horses, time bombs, backdoors, etc. [4].

IV. LITERATURE SURVEY

M. G. Schultz et al. [5] explained different methods of detecting a malicious executables. These malicious executables are formed at the huge rate every year and create a serious security threat. The anti-virus systems attempt to detect these malware programs with heuristics created by hand. This method is costly and sometime less-effective. In this paper, present a data-mining platform that detects new malicious files effectively and automatically. The data-mining platform automatically detects patterns in their data set and used these detected patterns to identify a set of new malicious executables. Compare these detection methods with a classical signature based method, the new method provides doubles the current detection rates for unseen malicious files.

Johannes Kinder [6] explained a model checking method for detecting malicious code. In this paper, author presents a soft method to detect malicious code sets in executables files by using model checking. While model checking was developed to check the correctness of system against specifications, author commented that it grants equally well to the identification of malicious code patterns. In the end, they introduced the specification language Computation Trees Predicate Logics which is extending the well-known logics CTL and gave description about an efficient model checking approach their practical experiments demonstrate that they are able to detect a large number of worm variants with a single specification.

Bhavani Thuraisingham [7] explained various data mining techniques for security application. These requisition include but are not limited to malicious executables detection by mining it binary executables, anomaly detecting and data stream mining process. They summarize their acquirement and present works at the University of Texas at Dalla on intrusions detection and cyber-security research.

Kirti Mathur [8] explained the techniques for detecting and analyzing Malware executables. Computer system's security is threatened by weapons named as malware to accomplish malicious intention of its writers. Various solutions are available to detect these threats like AV Scanners, Intrusion Detection System, and Firewalls etc. These solutions of malware detection traditionally use signatures of malware to detect their presence in our system. But these methods are also evaded due to some obfuscation techniques employed by malware authors. This survey paper highlights the existing detection and analysis methodologies used for these obfuscated malicious code.

Guillermo Suarez-Tangle [9] ,showed malware in current smart devices that equipped with powerful sensing, computing and networking capabilities have proliferated lately, range from famous smart android phones and tablets to Internet devices, smart TVs, and others that will soon appear. One main feature of devices is that they have ability to incorporate third-party applications from markets. This has very strong security features and secrecy problems to user and infrastructure operator, specifically via software of malicious nature that got access to the service given by the devices and gather the sensory data and personal data. Malware in latest smart devices – Smart phones and tablets– has got fame in the previous few years, in some cases supported by best techniques designed to provide better security architecture presently in use by these devices. As important advances have been made on malware detection in computers in the last decades it is still a challenging problem.

Parisa Bahraminikoo [10] implemented artificial Intelligence in anti-virus engines. Malicious software is the software which gives partial to full control of your computer to do whatever the malware creator wants. Malware can be defined as a viruses, worms, Trojans, adwares, spywares and root kits. Spyware is a class of malware which is installed on computer that is able to collect information regarding clients without having knowledge. In 1956, the purpose of establishment of Artificial Intelligence(AI) Dart muth College during a conference. Artificial Intelligence has been implemented in anti-virus engines. AI has many approaches that implemented in spyware detection systems such as Artificial Network, Heuristic Technologies and Data Mining Techniques. In this work, they focused on DM-based malicious code detectors by using Breadth-First Search approach for knowing work well for detection virus and software. BFS is the method for searching in a tree when search is very limited to essentially two operations (a) visit and inspect a node of a tree; (b) gain access to visit the nodes that are neighbour to currently visited node.

V. METHODOLOGY USED

The following steps are included in the classification process of this paper. Three different classifier is chosen Random Forest, Random Tree, Rep Tree. WEKA tool is used to analyse the predicated value by each of classifier. The Precision, Recall & F-Measure of each classifier is calculated. Finally the result is analysed and the best performance algorithm identified.

RANDOM FORESTS CLASSIFIER

Random Forests [14] are broadly believed to be the finest "off-the-shelf" classifiers for high-dimensional data. Random forests are a mixture of tree predictors such that each tree depends on the values of a random vector sampled autonomously and with the same distribution for all trees in the forest. The generalization error for forests converges to a limit as the number of trees in the forest becomes large. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the association between them. A different subset of the training data are selected, with replacement, to train each tree. Remaining training data are used to estimate error and variable importance. Class assignment is made by the number of votes from all of the trees and for regression the average of the results is used. It is similar to bagged decision trees with hardly some key differences as given below:

1.For each split point, the search is not over all p variables but just over mtry variables (where e.g. mtry = [p/3])

2. No pruning necessary. Trees can be grown until each node contains just very few observations (1 or 5).

Advantages of Random Forest over bagged decision trees are listed below:

1. better prediction.
2. almost no parameter tuning necessary with Random Forest.

RANDOM TREE CLASSIFIER

A Random tree [15] is a collection of tree predictors that is called forest. It can deal with both classification and regression problems. The classification works as follows: the random trees classifier takes the input feature vector, classifier it with every tree in the forest, and outputs the class label that received the majority of "votes" . In case of a regression , the classifier response is the average of the responses over all trees in the forest. All the trees are trained with the same parameter but on different training sets.

REP TREE CLASSIFIER

Reduces Error Pruning (REP) Tree Classifier is a fast decision tree learning algorithm and is based on the principle of computing the information gain with entropy and minimizing the error arising from variance [11]. This algorithm is first recommended in [12]. REP Tree applies regression tree logic and generates multiple trees in altered iterations. Afterwards it picks best one from all spawned trees. This algorithm constructs the regression/decision tree using variance and information gain. Also, this algorithm prunes the tree using reduced-error pruning with back fitting method. At the beginning of the model preparation, it sorts the values of numeric attributes once. As in C4.5 Algorithm, this algorithm also deals the missing values by splitting the corresponding instances into pieces. [13].

VI. STAISTICAL MEASURE

The accuracy of the classifier is given by TP rate, FP rate, Re-Call, F-Measure, Precision using WEKA tool. WEKA was developed at the university of Waikato in New Zealand; The name stands for Waikato Environment for Knowledge Analysis. The system is written in java and distributed under the terms of the GNU General public license.

It provides extensive support for whole process of experimental data mining, including preparing the input data, evaluating learning schemes statistically, and visualizing the input data and the result of learning as well as a variety of learning algorithms, it includes a wide range of preprocessing tools.

Before implementation of database some important terminology are-

N- Total number of classified instances.
True Positive (TP) - correctly predicated of positive classes.
True Negative (TN) - correctly predicated of negative classes.
False Positive (FP) - wrongly predicated as positive classes.
False Negative (FN) - total wrongly predicated as negative classes.
False Positive Rate (FPR) – negatives is correctly classified positives.

1) **Accuracy (A):** It shows the proportion of the total number of instance predictions which are correctly predicted

$$A = \frac{TP+TN}{N}$$

2) **Precision (p):** It is a determined of exactness. It is the ration of the predicated positive cases that were correct to the total number of predicated positive cases.

$$P = \frac{TP}{TP+FP}$$

3) **Recall (R):** Recall is determine of completeness. It is the proportion of positive cases that were correctly recognized to the total number of positive cases. It is also known as sensitivity or true positive rate(TPR)

$$R = \frac{TP}{TP+FN}$$

4) **F-Measure:** The harmonic mean of precision and recall. It is an important measure as it gives equal importance to precision and recall.

$$\text{F-Measure} = \frac{2 \times recall \times precision}{precision + recall}$$

VII. EXPERIMENTAL RESULTS

The Malicious data [16] is used to evaluate the performance of Random Forest, Random Tree and REP Tree for malicious risk predictions used weka tool. The data set for experiment has been collected UCI repository. This data set contains 42 Attributes. The data set comprises 34041 instances of malicious data with class details.

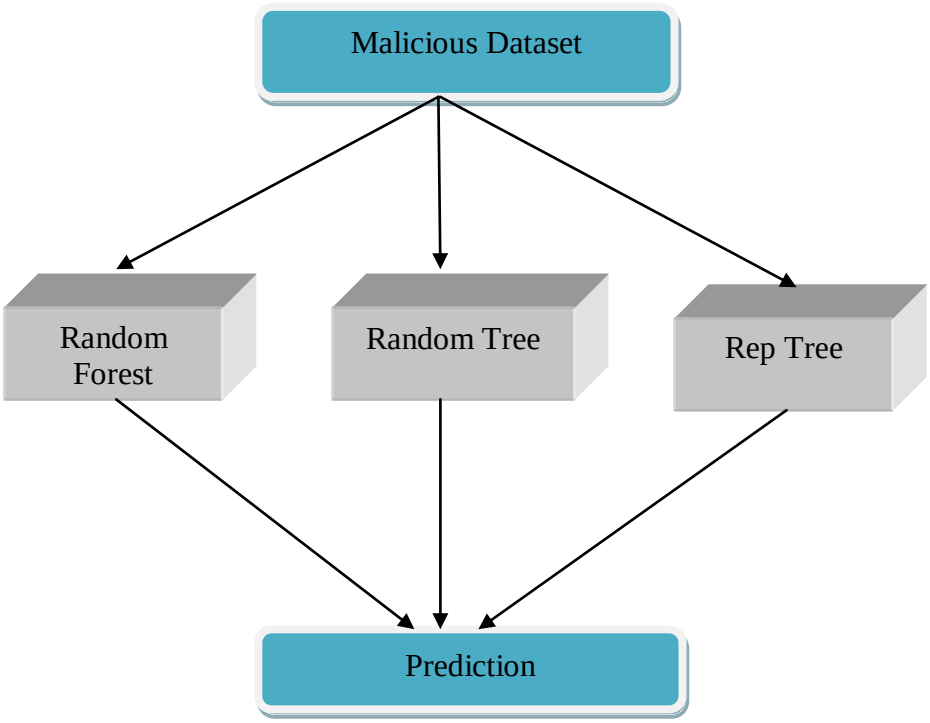

Figure. 1. Working Architecture for Proposed work

The result have been tabulated

Method	Correctly Classified Instances	Incorrectly Classified Instances	Time in sec taken with 20 attributes
Random Forest	34022	19	2.42 sec
Random Tree	34016	25	0.3 sec
Rep Tree	33726	315	0.47 sec

Table 1. Performance of accuracy measure for different algorithm.

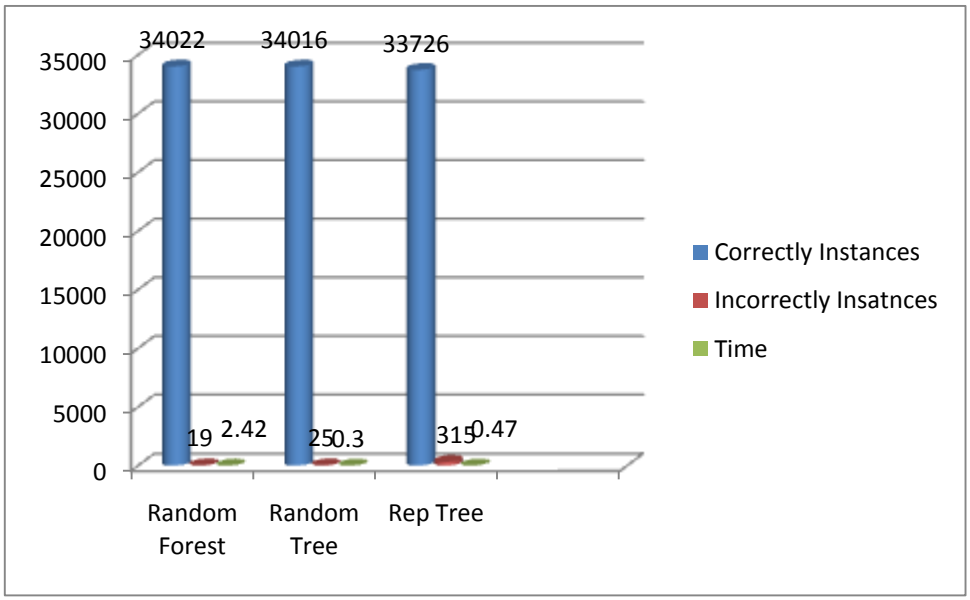

Figure.2. Performance diagram of different algorithm

The comparison of performance between Random Forest, Random Tree and Rep tree classifiers is depicted in table I and figure 2. The ranking is prepared based on time. The Random tree classifier outperforms the other two classifiers. It performs better time 0.3 sec comparing with other two algorithms.

Method	RMSE	MAE	KAPPA STATISICS
Random Forest	0.0058	0.0015	0.9973
Random Tree	0.0079	0.0001	0.9539
Rep Tree	0.0283	0.0016	0.9965

Table 2. Comparison of error rate measure for different algorithm in RMSE, MAE & KAPPA STATISIC

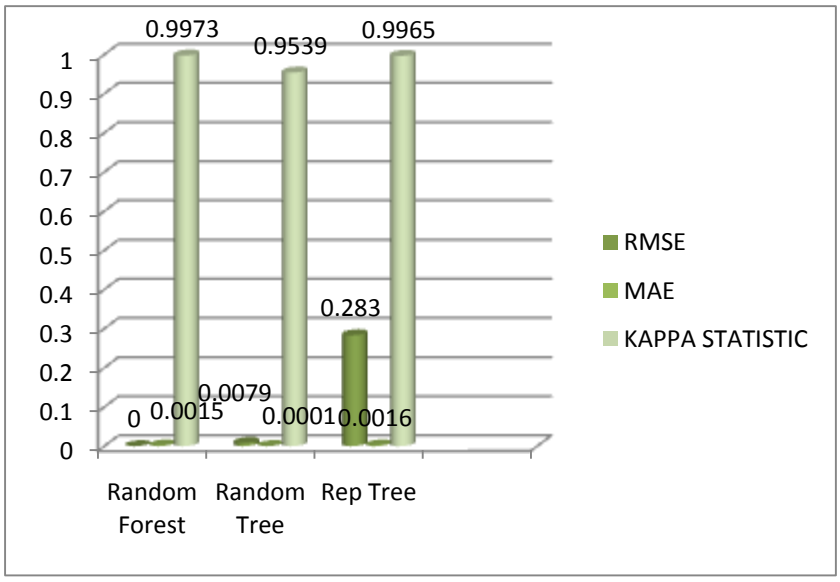

Figure.3. Diagram of comparison RMSE, MAE & KAPPA STATSTIC

From the experiment we have got thought of that Random tree takes the minimal time with middle value of RMSE wherever as Rep tree takes most RMSE value. Table II verified the results of Kappa statistic evaluating with two algorithms in Random tree can perform of lowest accuracy in kappa statistic. Experiments square measure performed on the malicious dataset by using Random Forest, Random Tree and Rep tree using weka tool.

Method	Precision	Re-call	TP Rate	FP Rate	F-Measure
Random Forest	0.996	0.996	0.996	0.003	0.999
Random Tree	0.999	0.999	0.999	0.002	0.999
Rep Tree	0.982	0.991	0.991	0.069	0.987

Table 3.Comparison of performance measure for the classification algorithm

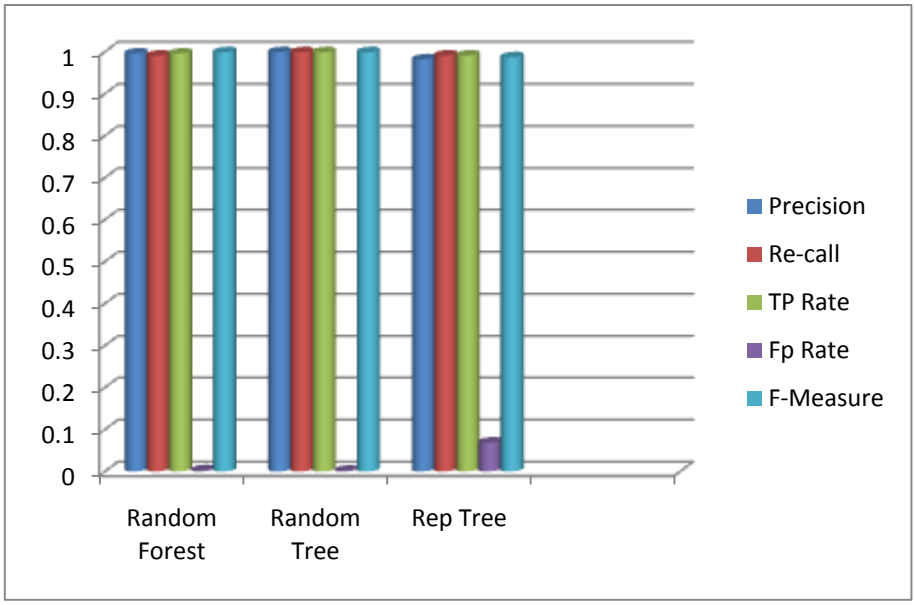

Figure. 4. Diagram of performance measure for the classification algorithm

The results of following analysis on the dataset is clearly given by the table II and III. Table II have given the RMSE, MAE & KAPPA STASTIC in dataset using totally different classifier. Table III listed the Precision, Re-call, True Positive Rate, False Positive rate, F-measure to analyse the classifier.

VIII. CONCLUSION AND FUTURE WORK

The dataset this article uses download from UCI repository. They are 34016 instances and 42 attributes. This work investigate the efficiency of three different classifiers namely Random Forest, Random Tree and Rep Tree classifiers. Testing is accomplished using the WEKA tool. The motivation behind removing rules is to isolate ordinary methods and malicious programs. According to the test results, we found that using the Random Tree classifier can get less time (0.3sec) but Random Forest gives maximum accuracy in classification, however it takes more time to process the data (2.42sec). At last, it is observed that Random Tree Classifier performs best followed by Random forest classifier and the Rep Tree classier. In future classifier algorithm can be combined to generate the results of the decision tree on the Knime platforms.

REFERENCES

[1]. U.M. Fayyad, G. Piatetsky – Shapiro, P. Smyth and R. Uthurusamy(eds.), "Adavnces in Knowledge Discovery and Data Mining", AAAI/MIT press,1996.

[2]. G. Piatetsky – Shapiro and W.J. Frawley(eds.),"Knowledge Discovery in DataBases". AAAI/MIT press, 1991.

[3]. G. McGraw and G. Morrisett, "Attacking malicious code: a report to the Infosec Research Council," *IEEE Software***,** Volume 17, Issue 5, pp. 33 – 41, September 2000.

[4]. D. M. Kienzle and M. C. Elder, "Recent worms: a survey and trends," in Proceedings of the 2003 ACM workshop on Rapid malcode, pp. 1 - 10, Washington, DC, USA, October 2003.

[5]. M. G. Schultz, E. Eskin, E. Zadok and S. J. Stolfo, "Data Mining Methods for Detection of New Malicious Executables", Proceedings of the 2001 IEEE Symposium on Security and Privacy, IEEE Computer Society.

[6]. Johannes kinder, "Detecting Malicious Code by Model Checking", pure.rhul.ac.uk/portal/files/17566588/mcodedimva05.pdf.

[7]. Bhavani Thuraisingham, "Data Mining for Security Applications", IEEE/IFIP International Conference on Embedded and Ubiquitous Computing,2008 .

[8]. Kirti Mathur, " A Survey on Techniques in Detection and Analyzing Malware Executables", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 4, April 2013.

[9]. Guillermo Suarez-Tangle, "Evolution, Detection and Analysis of Malware for Smart Devices" IEEE communications surveys & tutorials, accepted for publication, pp.1-27, 2013.

[10]. Parisa Bahraminikoo "Utilization Data Mining to Detect Spyware", IOSR Journal of Computer Engineering (IOSRJCE),Volume 4, Issue 3, pp.01-04,2012.

[11]. IH. Witten, and E. Frank, "Data mining: practical machine learning tools and techniques" 2nd ed. the United States of America, Morgan Kaufmann series in data management systems,2005.

[12]. J. Quinlan, "Simplifying decision trees", International Journal of Man Machine Studies, 27(3), 221–234, 1987.

[13]. S.K. Jayanthi and S.Sasikala, "REPTree Classifier for indentifying Link Spam in Web Search Engines", IJSC, Volume 3, Issue 2, (Jan 2013), 498 – 505,2013.

[14]. Leo Breiman, "Random Forests. Machine Learning",45(1): 5-32,2001.

[15]. Wikipedia contributors, "Random _tree," Wikipedia, The Free Encyclopedia. Wikimedia Foundation, 13-jul-2014.

[16]. UCI Machine Learning Data Repository – http://archive.ics.uci.edu/ml/datasets.html.

[17]. C. Lakshmi pevasena,"Comperative Analysis of Random Forest, REP Tree and J48 classifiers for credit risk prediction",International Journal of Computer Applications (0975 – 8887) International Conference on Communication, Computing and Information Technology (ICCCMIT-2014).

[18]. Haixu Xi and Hongjin Zhu, "Data Mining Methods For New Feature of Malicious Program", interntional journal of Hybrid information technology, Vol.9, No.3(2016)pp, 171-178, http://dx.doi.org/10.14257/ijhit.2016.9.3.16.

