DATA MINING: A 'RIM' ALGORITHM FOR SPYWARE DETECTION WITH PRUNING

Mrs. Indulakshmi[1]
(M.Tech) Scholar,
indu.rajnish@gmail.com
Department of Computer Science Engineering,
Bharath University-Chennai, Tamilnadu

Dr.Nalini[2]
Professor
Department of Computer Science Engineering,
Bharath University-Chennai, Tamilnadu

ABSTRACT:

The aim of this paper is to employ the principles of data mining and classify a new algorithm. In this method, we have proposed a new anti-spyware system (Spyware Detection), which is capable for classifying spyware files from legitimate files. Also we have evaluated the performance of our anti-spyware system with the existing anti- spywares in terms of overall accuracy (ACC) and false positive rate (FPR)... This paper describes the use of classification trees and shows the methods of pruning them using the new Algorithm.The RIM algorithm which can be classified and compared with other algorithms like Naïve Bayes, HNB (Hidden Naïve Bayes) , Random Tree Algorithm The experimental results suggest that our method is better than the existing methods and shows the best result.

KEYWORDS: CSV –Comma separated Values; ARFF- Attribute RelationFile Format ; HNB-Hidden Naïve Bayes; RFS- Reduced Feature Set

1.INTRODUCTION:

The journey of the project explains how to implement a new algorithm and use it with the existing Algorithm for a comparative study.In recent years the malware practice The computer experts have thought over the classification of malware by using pattern matching. Let us see detailed about the each step of the project.

In our project pruning method is used. Pruning is the method that reduces the size of the tree which is not important and avoid intricacy of the dataset. Complication of the tree is measured by pruning. There are 2 types of pruning methods.

1. Forward Pruning
2. Backward Pruning

Where Forward pruning does not generate full tree. In our Algorithm RIM method which gives us reduced pruning method which also generates decision tree to the dataset given .Compared to other algorithm with Naïve Bayes , HNB and Random Tree methods.

Weka tool is widely used to classify all these algorithm by extracting few features from it and implement it using java swing. The front end is designed using Java Swing and run in command prompt using some software like Eclipse,Ant,Java.

Most spyware detector which uses n-gram sequence of any text of variable length to hexadecimal dump of the file . Naive Bayes (NB) uses string data and n-grams of byte sequence. Naive Bayes using strings performed best with 97.11% accuracy. They are also using signature-based algorithm to match

their results from data mining algorithms over new Malware.

2.PROPOSED WORK

Proposed work will explain how to collect spy data from different system and make use of spyware Detector which helps to check for Original files and files which is infected. Here is the detailed explanation of how to build this project.

2.1 DATASET CREATION:

Database is created according to the files that is present in the system where some file are malicious .

These files are given as input to java file and converted as Hexadecimal value using a program.

strToHex.convertStringToHex(str).

These Hexadecimal is further converted to arff format by using another program.

if(args.length==4){

System.out.println("\nUsage: <input><output>\n");

System.exit(1); }

ArffSaver saver =newArffSaver(); }

The sample arff format is given here

@relation Spyware

@attribute Feature {4D5A,0000,2A00,1B00,2000,B804,B805,A808,FFFF}

@attribute Filename {GHOST,PlayFLV,CamStudio20,dap,dopdf,googledesktop,googletoolbar,googlevoiceandvideo,installflashplayer,pdf2word,pwviewer,qdweb33,shutdowntimer,skypesetup,smartdraw,winamp,wrar,youtubedownloader}

@data

4D5A GHOST

1B00PlayFLV

0000 PlayFLV

4D5A CamStudio20

B805 shutdowntimer

FFFF shutdowntimer

2A00 pwviewer ….

3. ALGORITHM DESCRIPTION

The study will show us how to differentiatethe newalgorithm (RIM) with other Algorithm.

3.1 NAÏVE BAYES:

A Naive Bayes classifier is classified based on probability method. It checks for number of possible outcome with given number of data.

The Bayes Theorem:

$$P(h/D) = \frac{P(D/h)\,P(h)}{P(D)}$$

P(h) : probability of hypothesis h

P(D) : probability of training data D

P(h/D) : Probability of h given D

P(D/h) : Probability of D given h

3.2 HIDDEN NAÏVE BAYES(HNB):

HNB is created with aunknown parent for each attribute, which joins theinfluencesfromallother features.Thismodel is called hidden naïve Bayes (HNB).The attribute dependencies are represented by hidden parents of attributes

The distribution represented by an HNB is defined as follows.

$P(A1,\cdots,An,C) = P(C)\prod_{i=1}^{k} P(Ai|Ahpi, C)$

The classifier corresponding to an HNB

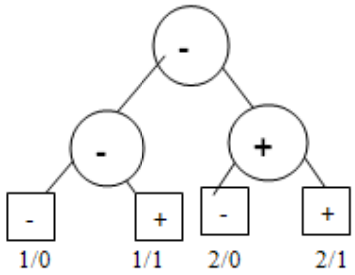
$$c(E) = \underset{c \in C}{\operatorname{argmax}} P(C)\prod_{i=1}^{k} P(Ai|Ahpi, C)$$

3.3 RANDOM TREE:

Random Tree is an algorithm for constructing a tree that considers K random features at each node. A random number of attributes are chosen for each tree. These attributes form the nodes and leafs and itPerforms no pruning

3.4 RIM ALGORITHM:

RIM Algorithm which tells you how to classify the trees with the given file format. RIM starts at the leaves and each node is replaced with most classes. RIM has the advantage of simplicity and speed. It uses a pruning method to evaluate the sub trees of a parent tree.

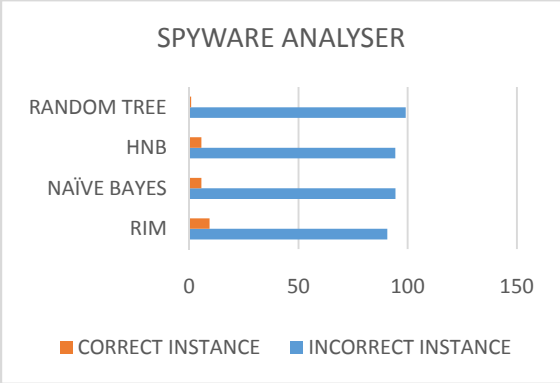


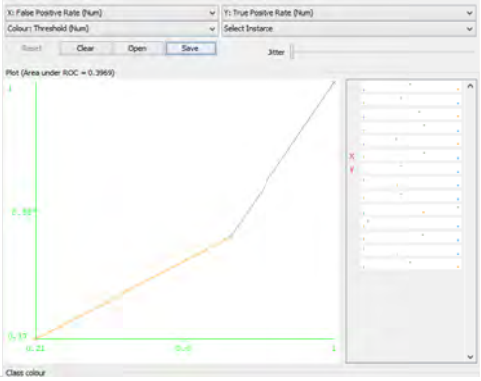
we see that the expected number of leaves in T is

$$EA + 1 = 2(t- 1)p(1- p)+1$$

In particular, this is linear in the size of the sample S; |S| = t.

3.5 COMPARATIVE STUDY:





Comparative study between 4 algorithms **Threshold Point**

Output:

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	10	9.3458 %
Incorrectly Classified Instances	97	90.6542 %
Kappa statistic	-0.0136	
Mean absolute error	0.1027	
Root mean squared error	0.2269	
Relative absolute error	99.6973 %	
Root relative squared error	100.017 %	
Total Number of Instances	**107**	

4.EVALUATION MEASURES:

True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN). TP represents the correctly identified benign programs while FP represents the incorrectly classified spyware programs. TN represents the correctly identified spyware programs and FN represents the wrongly identified benign programs.

Accuracy: ACC = (TP+TN)/(P+N)
Positive predictive value: PPV = TP/ (TP+FP)
Negative predictive value: NPV = TN/ (TN+FN)
False discovery rate: (FDR) = FP/ (FP+TP)
Specificity: (SPC) = TN/ (FP + TN)

5.CONCLUSION AND FUTURE WORK:

The techniques and algorithms that were executed under different circumstances where RIM algorithm gave a good result because of Pruning Method. The future work can be modified into MATLAB where it gives us the exact prediction of how each and every model works. The experimental results indicate that the approach is successful, achieving 92% of accurate result using RIM algorithm. In future, we aim to develop a hybrid Spyware identification method using MATLAB.

6.REFERENCES:

[1] S M. Boldt and B. Carlsson ,, “Privacy-invasive software and preventive mechanisms,” IEEE Computer Society, 2nd International Conference on Systems and Networks Communications, (ICSNC 2006), Oct. 28- Nov.2

[2] M. Wu, Y. Huang, S. Kuo, “Examining Web-based Spyware invasion with statefulbehavior monitoring,” 13th Paci_c Rim International Symposium on Dependable Computing (PRDC '07) , 17-19 Dec. Piscataway, NJ, USA: IEEE, pp. 275-81.

[3] R. Sandhu, “Lattice-based access control models,” Computer , vol. 26, Nov. 1993,pp. 9-19.

[4] M. Schultz, E. Eskin, F. Zadok, and S. Stolfo, “Data Mining methods for detection of new malicious executables,” Proceedings of IEEE Symposium on Security and Privacy , 14-16 May 2001, Los Alamitos,CA, USA: IEEE Computer Society, pp.38-49.

[5] I.H. Witten, E. Frank, ““Data Mining: Practical Machine Learning Tools and Techniques,” 2nd ed. Morgan, Kaufmann , 2005.

[6] O. Henchiri, N. Japkowicz, “A Feature Selection and Evaluation Scheme for Computer Virus Detection,” International Conference on Data Mining , 8-22 Dec., Piscataway, NJ, USA: IEEE, pp. 918-922, (ICDM2006).

[7] R. Moskovitch, C. Feher, N. Tzachar, E. Berger, M. Gitelman, S. Dolev, and Y.Elovici, “Unknown malcode detection using OPCODE representation,” (EuroISI 2008) , 3-5 Dec., Berlin, Germany: Springer- Verlag, pp. 204-215.

[8] Y. Elovici, A. Shabtai, R. Moskovitch, G. Tahan, and C. Glezer, “Applying Machine Learning Techniques for Detection of Malicious Code in Network Traffic,” Proceedings of the 30th annual German conference on Advances in Arti_cial Intelligence , 2007, 10-13 Sept., Berlin, Germany: Springer-Verlag, pp. 44-50.

[9] C. D. Bozagac, “Application of Data Mining based Malicious Code Detection Techniques for Detecting new Spyware,” White paper ,Bilkent University 2005.

[10] N. Lavesson, M. Boldt, P. Davidsson and A. Jacobsson, “Learning to Detect Spyware using End User License Agreements,” Knowledge and Information Systems , in.press.

[11] D. Reddy, S. Dash, and A. Pujari, “New Malicious Code Detection Using Variable Length n-grams,” Information Systems Security , 2006, pp. 276-288.

[12] M. Siddiqui, M. Wang, and J. Lee, “Detecting Trojans Using Data Mining Techniques,” Wireless Networks, Information Processing and Systems , 2009, pp. 400-411.