Malware Detection Based On Opcode Frequency

Abhijit Yewale
Computer Science and Engineering Department
Thapar University
Patiala, India
yewaleabhijit70@gmail.com

Maninder Singh
Computer Science and Engineering Department
Thapar University
Patiala, India
msingh@thapar.edu

***Abstract*— Malware is a computer program or a piece of software that is designed to penetrate and detriment computers without owner's permission. There are different malware types such as viruses, rootkits, keyloggers, worms, trojans, spywares, ransomware, backdoors, bots, logic bomb, etc. Volume, Variant and speed of propagation of malwares are increasing every year. Antivirus companies are receiving thousands of malwares on the daily basis, so detection of malwares is complex and time consuming task. There are many malwares detection techniques like signature based detection, behavior based detection and machine learning based techniques, etc. The signatures based detection system fails for new unknown malware. In case of behavior based detection, if the antivirus program identify attempt to change or alter a file or communication over internet then it will generate alarm signal, but still there is a chance of false positive rate. Also the obfuscation and polymorphism techniques are hinderers the malware detection process. In this paper we propose new method to detect malwares based on the frequency of opcodes in the portable executable file. This research applied machine learning algorithm to find false positives, false negatives, true positives and true negatives for malwares and got 96.67 per cent success rate.**

***Index Terms*— Malware Detection, Opcode Frequency, Machine Learning, Malware Analysis, Data Mining.**

I. Introduction

Malware is a computer program which is designed to harm or disrupts the operation of a computer system. Malware detection is a system that attempts to find whether a program has malicious intent or not [5]. In the last few years, motivation behind creation of malware has changed drastically. Most malware author was financially motivated [6].There are different malware detection techniques such as signature based malware detection, specification based detection, anomaly based detection and machine leaning based detection.

In signature based malware detection, antivirus program looks for signature which is nothing but sequence of byte in a particular file to declare the file as malicious [7]. For polymorphic and unknown viruses, signature based detection system fails because polymorphic viruses are encrypted viruses and they are changing decryptor loop on each infection without changing actual code and for unknown viruses there is no signature present in antivirus database.

Anomaly based system detects any misuse of computer that falls out of the regular activity of a computer [8]. Anomaly based detection system monitor the computer program or software activity and classify it as either normal or anomalous. Behavior based detection system, identifies the action performed by the malware. The program with different syntax but have a similar behavior is collected and this single behavior signature can be used to identify different malware sample. Machine learning based detection system requires training dataset of malware attributes and according that machine learning algorithm detects the malware. There are different machine algorithms such Decision Tree, Support Vector Machine, Random Forest, Boosting, etc. [7] The obfuscation techniques such as Dead code insertion, Register Renaming, Code Transportation, Instruction Substitution hinder the malware detection mechanism. The remaining portion of a paper is organized as follows: Section 2 discusses the related work; section 3 discusses how the features are selected; Section 4 discusses about experimental setup and work flow for malware detection system; Section 5 discusses the results of machine learning algorithm and section 6 discusses the conclusion and future work.

II. Related Work

Malware detection is very important topic over the World Wide Web. Antivirus vendor has been receiving thousands of malwares on the daily basis. The research in this area is unstoppable. The survey here stated in the area of malwares detection from year 2007 to 2015. The research work done in the area of malware detection using machine learning techniques is explained in this section. The detection of malicious code can be done through statistical analysis of opcodes distribution using statistical method such as Pearson's chi square procedure, post-hoc standardized testing, Crammer's V [9]. [10] They proposed static malcode detection using decision tree classification algorithm in chronological point of view. The dataset includes malwares from year 2000 to 2007. Training and testing of classifier with malware up to respective year has done and performance has evaluated. [11] The concept of text categorization used for detection of malware.

They investigated imbalance problem about malicious and benign file. They got accuracy of 95 per

ISBN No.978-1-4673-9545-8

cent when the percentage of malicious file is below 20 per cent in training data set. [12] They used opcodes generated by disassembling the executable file, then uses n gram of the opcodes as a feature vector for classification process. Also they used concept of text categorization for detection of unknown malware. [13] The manual malware analysis is not effective and they proposed automated behavior based malware detection using machine learning techniques.

The behavior of each malware analyzed in sandbox and report is generated, then the report is preprocessed and it is use as feature vector for machine learning classifier. [14] The performance of machine learning is influenced by the feature vector and the algorithm used to generate classifier. Also they combined content based and behavior based feature. They proposed ensemble learning method called SVM-AR, it combines Support Vector Machine and association rules. [15] They proposed semi-supervised learning approach to detect malwares. This method was based on appearance of Opcode sequence frequency. [16] The distance based filtering rule used to identify the noise in training dataset then they removed the noise and checked the performance of classifier. [17] The cfg is generated from executable file then from that cfg features were extracted.

These features were given as an input to machine learning classifier to classify the executable into goodware or malware. [18] They designed modified version of perceptron algorithm to train the dataset, to get low false positive rate. [19] They identified seven features such as DebugSize, ImageVersion, IatRVA, ExportSize, ResourceSize, VirtualSize2 and NumberOfSections of Microsoft portable executable file format that can given to machine learning algorithm to classify the PE file into malware or goodware. [20] They proposed that information from kernel process control block of processes have used to detect the malwares at runtime. The selected parameters are maintained inside PCB of a kernel for each running process, which defines the semantics and behavior of an executing process.

As a result 16 out of 118 task structure parameter selected using time series analysis. These parameters are as an input to the machine learning classifier to detect malwares. [21] They used Opcode sequence frequency representation of executables to detect and classify the malwares. [22] A social graph has used to study relation between portable executable files and the relation between files used as a feature vector for machine learning classifier to detect the malwares. [23]They made comparison between several feature selection methods such as correlation based feature selection, principal component analysis, InfoGain Attribute, etc with different machine learning classifier. [24] Malware detection based on the Opcode density and system call feature has obtained dynamically by tracing the behavior of executables at runtime.

III. Feature Selection

Feature Selection means selecting subset of feature from whole. Feature vector plays major role in constructing machine learning model. Principal Component Analysis (PCA) and Feature Rank algorithm has been used to extract most important features. Goodware files are not malicious (benign) files. We have collected total 100 malwares from the malware website [1] and goodwares from the windows Operating Systems, system32 directory. Malware is unpacked and disassemble using UPX Unpacker and IDA pro respectively. We have used the Instruction Counter plug-in to get the statistics of Opcodes in text format. After Observing 100 goodware and malware samples, 20 Opcodes are frequently repeated, hence considered for further analysis. Twenty most frequent Opcodes identified such as MOV, PUSH, CALL, CMP, POP, JZ, TEST, JNZ, JMP, ADD, LEA, XOR, RETN, AND, SUB, OR, INC, MOVZX, DEC, JB. These 20 Opcodes are given input to the machine learning classifier namely BOOSTING, Support Vector Machine, Random Forest and Decision Tree to classify the portable executable file into malware or goodware.

IV. Experiment setup and working

The following tools and techniques are required for experimental setup are portable executable files, IDA pro [2], ExeInfo PE Win 32 exe identifier, PEiD, OllyDbg 32 bit PE analyser, virus total website, UPX, VMware Workstation, malwares website and rattle. We have downloaded malwares from malware website [1] and Extract assembly code from PE by disassembler. The Diag tab of ExeInfo PE [3] shows file characteristics by means of it detects compiler and packer. PEiD is used to detect packers [4]. An experiment has done in VMware. After opening the PE file in debugger OllyDbg, we have checked memory map to identify file is packed or not. Rattle is a graphical user interface based machine learning tool, in which all classifier such as SVM, RF, BOOST, Decision tree , etc have been implemented.

The Figure 1 represents system block diagram for malware detection. We have Collected the malwares and goodwares, disassemble them using IDA pro. The statistics of Opcodes have been collected using Instruction Counter plug-in. After analyzing statistics of 100 goodwares and malwares, we found 20 most frequent Opcodes in each and every PE file. These top 20 Opcodes are selected as feature vector for machine learning classifier. Machine learning algorithms are classified into three categories such as supervised, unsupervised and semi-supervised learning. In supervised machine-learning technique requires training data should be properly labeled in order to build a model. Algorithms which are come into this category are decision tree, SVM, random forest, boosting, etc [26]. In unsupervised machine-learning technique does not require labeled data, in this first clusters of similar data are created using clustering algorithms. Clustering algorithm such as K-means,

hierarchical clustering, etc are come into this category. The semi-supervised machine learning is a mixture of labeled and unlabeled data to build a model. In our case data has properly labeled, we have used supervised machine learning algorithm to build a model. Total dataset have been partition in two parts 70 and 30 per cent. The 70 per cent dataset have been used to train the classifier such as RF, SVM, Boost and Decision Tree. The 30 per cent dataset have been used to evaluate the performance of all models.

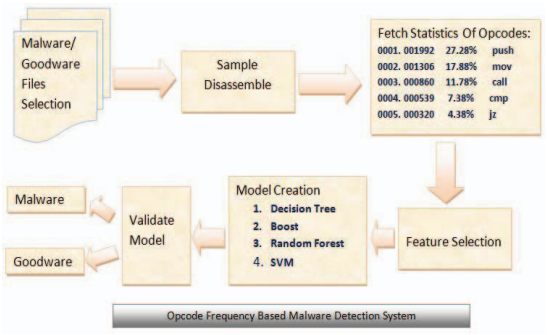

Fig. 1. System Block Diagram for Malware Detection

V. RESULTS AND DISCUSSIONS

This research aims to detect unknown malware using machine learning techniques. We analyzed 100 goodwares and malwares files each containing opcodes approximately 30 to 40 thousands. All malware were verified from virus total website. The virus total website contains database of approximately 50 antiviruses. After testing malwares on virus total, mostly antivirus identified malwares as Trojan hence we have not done further categorization of malwares. Our focus in this research is opcodes and their frequency. We figured out 20 most frequent opcodes and theses opcodes were extracted from assembly file of each malware and goodware file. These 20 opcodes were used as feature vector for machine learning classifier. The machine learning classifier which used for this research are support vector machine, random forest, decision tree and boosting. Next, we divided the dataset in two parts are training and testing dataset. Training and testing dataset percentage are 70 and 30 per cent respectively. In specifically we measured True Positives (TP), True Negatives (TN), False Positives (FP), False Negatives (FN), True Positive Ratio (TPR), False Positive Ratio (FPR) and Accuracy (AC). True Positive Ratio is nothing but number of malwares correctly classified divided by total number of malwares. TPR=TP/(TP+FN) where TP is the number of malwares correctly classified and FN is the number of malwares misclassified. False Positive Ratio is nothing but number of goodware files misclassified as malwares divided by total number of goodware files. FPR=FP/(FP+TN) where FP is the number of goodware files misclassified as malware and TN is nothing but number of correctly classified goodware files. Accuracy is nothing but total number of correctly classified observation is divided by total number of observation in whole dataset. Accuracy (%) = (TP+TN)/(TP+FN+FP+TN) (Santos, Igor, et al., 2011). The Table I represent the how TP, TN, FN, FP, TPR, FPR and AC vary with respect to classifiers. The Table I represent the performance evaluation of models based on testing dataset.

PERFORMANCE EVALUATION OF MODELS

Model	TP (%)	TN (%)	FN (%)	FP (%)	TPR (%)	FPR (%)	AC (%)
DT	30	50	7	13	82	21	80
BT	30	57	7	7	82	11	87
RF	33	63	3	0	91	0	97
SVM	33	60	3	3	91	5	93

Where DT- Decision Tree, BT- Boost, RF- Random Forest, SVM- Support Vector Machine

After analyzing the performance evaluation of all four models, we come to conclusion that Random Forest classifier provides highest accuracy about 97 per cent as compare to all the models. After this we checked how FPR, TPR and Accuracy parameter affects performance of random forest by drawing graphs as shown in Figure 2 and Figure 3. The Figure 2 represents graph between Training Dataset versus TPR and Accuracy. If we observe in the Figure 3 as Training Dataset percentage increases the True Positive Rate also increases. In case of Accuracy, as training data percentage increases the accuracy also increases. At 60 per cent training data we got highest accuracy of 98 per cent. The Figure 3 represents the graph between training dataset and false positive ratio. If we observed the graph that as Training dataset increase the False Positive Ratio become approaching to zero. In particular, at 60 per cent training dataset the FPR is 0. From both the graph we come to conclusion that if you train the classifier with more dataset then it will provide high accuracy and low false positive rate.

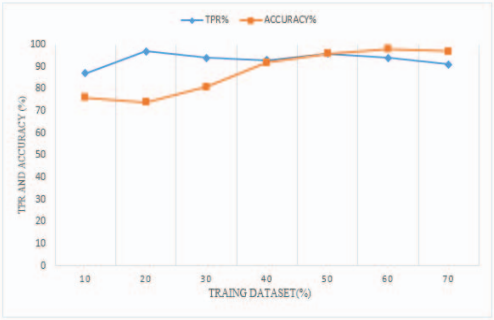

Fig. 2. Graph for Training Dataset Vs. TPR and Accuracy

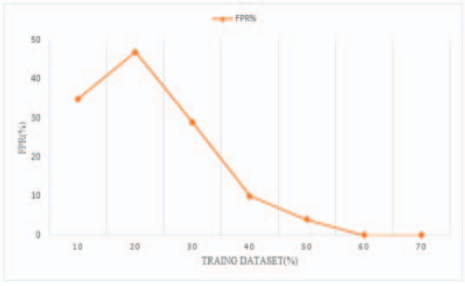

Fig. 3. Graph for Training Dataset vs. FPR

VI. Conclusion and Future Work

Malware detection is growing research area because of concern over increase of malwares on the daily basis. Signature based antivirus system is not useful for unknown malware detection and they are facing difficulties because of polymorphic viruses and zero day attack. So signature based methods must be along with other approach that can able to detect unknown malwares. The machine learning is a suitable approach to complement classical signature based malware detection system. In this paper we collected dataset from 100 goodware and malware files for machine learning classifier. We identified that, Opcode frequency can be used to detect the unknown malwares. We found 20 most frequent opcodes can be used as feature vector for machine learning classifier. The dataset for goodwares and malwares are containing 20 most frequent Opcode with their frequency. By using our dataset we have constructed four models are SVM, RF, BOOST and Decision Tree. Out of four models Random Forest has provided 97 per cent accuracy and zero per cent false positive ratio. In this research we have classified portable executable file into two categories are either goodware or malware. In future, we plan to apply our dataset to more machine learning classifier, in this paper we have classified PE file in only two categories either goodware or malware but we are planning to classify malware into further categories like Trojan, Spyware, Backdoor, worm, etc.

References

[1] Sandbox, Cuckoo. "Malwr-Malware Analysis".Guilfanov, Ilfak.

[2] "The IDA Pro disassembler and debugger version 5.0, March 2006".

[3] ExEinfo PE. http://www.exeinfo.go.pl/.

[4] Jibz, Qwerton, XineohP Snaker, and PEiD BOB. "Peid." Available in: http://www. peid. info/. Accessed in 21 (2011).

[5] Christodorescu, Mihai, Somesh Jha, Sanjit A. Seshia, Dawn Song, and Randal E. Bryant. "Semantics-aware malware detection." In Security and Privacy, 2005 IEEE Symposium on, pp. 32-46. IEEE, 2005.

[6] Ollmann, Gunter. "The evolution of commercial malware development kits and colour-by-numbers custom malware." Computer Fraud & Security 2008.9 (2008): 4-7.

[7] Vinod, P., R. Jaipur, V. Laxmi, and M. Gaur. "Survey on malware detection methods." In Proceedings of the 3rd Hackers' Workshop on Computer and Internet Security (IITKHACK'09), pp. 74-79. 2009.

[8] Teng, H.S., Chen, K. and Lu, S.C., 1990, May. Adaptive real-time anomaly detection using inductively generated sequential patterns. In Research in Security and Privacy, 1990. Proceedings., 1990 IEEE Computer Society Symposium on (pp. 278-284). IEEE.

[9] bBilar, D. (2007). Opcodes as predictor for malware. International Journal of Electronic Security and Digital Forensics, 1(2), 156-168.

[10] Moskovitch R, Feher C, Elovici Y. Unknown malcode detection—a chronological evaluation. InIntelligence and Security Informatics, 2008. ISI 2008. IEEE International Conference on 2008 Jun 17 (pp. 267-268). IEEE.

[11] Moskovitch, Robert, et al. "Unknown malcode detection via text categorization and the imbalance problem." Intelligence and Security Informatics, 2008. ISI 2008. IEEE International Conference on. IEEE, 2008.

[12] Moskovitch, Robert, et al. "Unknown malcode detection using opcode representation." Intelligence and Security Informatics. Springer Berlin Heidelberg, 2008. 204-215.

[13] Firdausi, Ivan, et al. "Analysis of machine learning techniques used in behavior-based malware detection." Advances in Computing, Control and Telecommunication Technologies (ACT), 2010 Second International Conference on. IEEE, 2010.

[14] Lu, Yi-Bin, Shu-Chang Din, Chao-Fu Zheng, and Bai-Jian Gao. "Using multi-feature and classifier ensembles to improve malware detection." Journal of CCIT 39, no. 2 (2010): 57-72.

[15] Santos, Igor, et al. "Opcode-sequence-based semi-supervised unknown malware detection." Computational Intelligence in Security for Information Systems. Springer Berlin Heidelberg, 2011. 50-57.

[16] Gavriluţ, D. and Ciortuz, L., 2011, September. Dealing with Class Noise in Large Training Datasets for Malware Detection. In Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), 2011 13th International Symposium on (pp. 401-407). IEEE.

[17] Zhao Z. A virus detection scheme based on features of Control Flow Graph. InArtificial Intelligence, Management Science and Electronic Commerce (AIMSEC), 2011 2nd International Conference on 2011 Aug 8 (pp. 943-947). IEEE.

[18] Gavrilut, D., Benchea, R., & Vatamanu, C. (2012, September). Optimized zero false positives perceptron training for malware detection. In Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), 2012 14th International Symposium on (pp. 247-253). IEEE.

[19] Raman, K., 2012. Selecting features to classify malware. InfoSec Southwest, 2012.

[20] Shahzad, F., Shahzad, M. and Farooq, M., 2013. In-execution dynamic malware analysis and detection by mining information in process control blocks of Linux OS. Information Sciences, 231, pp.45-63.

[21] Santos, I., Brezo, F., Ugarte-Pedrero, X., & Bringas, P. G. (2013). Opcode sequences as representation of executables for data-mining-based unknown malware detection. Information Sciences, 231, 64-82.

[22] Jiang, Q., Liu, N. and Zhang, W., 2013, December. A Feature Representation Method of Social Graph for Malware Detection. In Intelligent Systems (GCIS), 2013 Fourth Global Congress on (pp. 139-143). IEEE.

[23] Khammas, Ban Mohammed, et al. "FEATURE SELECTION AND MACHINE LEARNING CLASSIFICATION FOR MALWARE DETECTION." Jurnal Teknologi 77.1 (2015).

[24] Ranveer, S., & Hiray, S. SVM Based Effective Malware Detection System. In: 2015 International Journal of Computer Science and Information Technologies, Vol. 6 (4) , 2015, 3361-3365.

[25] Williams, G. (2011). Data mining with Rattle and R: The art of excavating data for knowledge discovery. Springer Science & Business Media.

[26] S. Kotsiantis, Supervised machine learning: a review of classification techniques, in: Proceeding of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies, 2007, pp. 3–24.