ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.4 , No.5, Pages : 01-03 (2015)
Special Issue of ICETEM 2015 - Held on October 27, 2015 in The Solitaire Hotel, Bangalore, India
http://www.warse.org/IJATCSE/static/pdf/Issue/icetem2015sp01.pdf

Malware Image Analysis and Classification using Support Vector Machine

Aziz Makandar[1], Anita Patrot[2]
[1,2]Karnataka State Women's University
Vijayapura, Karnataka, India
azizkswu@gmail.com
patrotanita@gmail.com

***Abstract*—**The malware is one of the major concerns in computer and cyber security. The availability of various malware toolkits and internet popularity that has led to the increase in number of malware attacks day to day. Comparing with existing framework of antivirus scanners they currently used signature based a malware detection technique which is widely. In this paper, we propose an efficient framework for identification of malware variants. A new method of recognizing malicious file based on computer image processing and Support Vector Machine (SVM) is studied to improve recognition accuracy and efficiency. At first, sub band filtering was applied to the original malware image. Then a method of statistic pattern recognition Gabor filter is used to get second order gradient features were extracted, and classification method of SVM for recognition of malicious file was used. The malicious data binaries are converted to grayscale image which consists of textural patterns based on these patterns classify the variants of malware. SVM (Support Vector Machine) is used in analysis of raw malware binary. The proposed method is experiments on 3131 different variants of malware of Manhuer dataset and obtained 89.68 % accuracy.

Keywords: Gabor, Malware, Texture, SVM, Wavelet.

INTRODUCTION

Number of new variants of malware on the internet has been continuously increasing. The most of antivirus programs use signature based scanning file data to detect of malware [1][2][3]. The malware is the behavior of that particular software such as stealing the private data from various ways. More than million unique variants of malware are released per day in the internet. Analyzing more number of malware variants every day is a challenging task. Manually identifying and classifying malware samples is something which is inevitable due to growing number of malware variants every year even though the researcher need a quick and easy analysis of malware variants especially on behavioral aspects of the malware.

The machine learning techniques are used to identify different types of patterns which available in malware variants for identification and classification. The increasing use of machine learning techniques for various applications such as medical image analysis, human identification, face recognition, optical character recognition, and malware detection and classification. The machine learning techniques are k-nearest neighbor classifier, Support Vector Machine (SVM) classifier and ANN Classifier. Support Vector Machine classifier is a supervised learning algorithm which is used to analyze the data and recognize the different patterns for classification. This classifier creates a hyper plane between series of patterns of different class. The mathematical form of SVM linear classifier is shown as in (1).

$$f(x) = W^T X + b \qquad (1)$$

This is the first comprehensive introduction to Support Vector Machines (SVMs), a new generation learning system based on recent advances in statistical learning theory. SVMs deliver state-of-the-art performance in real-world applications such as text categorization, hand-written character recognition, image classification, bio sequences analysis, etc., and are now established as one of the standard tools for machine learning and data mining. Students will find the book both stimulating and accessible, while practitioners will be guided smoothly through the material required for a good grasp of the theory and its applications. The concepts are introduced gradually in accessible and self-contained stages, while the presentation is rigorous and thorough. Pointers to relevant literature and web sites containing software ensure that it forms an ideal starting point for further study.

EXISTING WORK

There are various methods to analyze malware including static and dynamic analysis. The traditional way of detecting malicious data in system is used sequence based bytes methods, instruction frequency based techniques, and API calls are used as a feature vector. As well as based on the code as well as PE faces a rigorous challenging task in [12][13]. In order to overcome defects in data mining and machine learning techniques introduce a field Antivirus [13] [14]. There are various methods are used to detect malware. The classification including graph based detection of malware [14] [15] [16], instruction sequence based classification [17, 18] API call sequence based classification [19][20]. The malware identification [21]. Recently various researchers' uses visualization technique to understand malware visually that can help antivirus software to detect malware efficiently.

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.4 , No.5, Pages : 01-03 (2015)
Special Issue of ICETEM 2015 - Held on October 27, 2015 in The Solitaire Hotel, Bangalore, India
http://www.warse.org/IJATCSE/static/pdf/Issue/icetem2015sp01.pdf

Identification of human [22] [23] [24] [25]. The visualized the malware behavior into tree maps and thread graph by using API calls [25].

Visualization Technique	Type of Analysis	Types of Feature	Limitations
Malware Tree Map	Dynamic	API Calls	Low granularity
Malware Thread Graph	Dynamic	API Calls	Low granularity to 550 Operations
Malware Image	Static	Raw malware binary	Does not represent actual malware behavior
VERA	Dynamic	Memory address, memory state, code entropy	Not meant for representing malware behavior

Table.1 VISUALIZATION TECHNIQUES OF MALWARE

METHODOLOGY

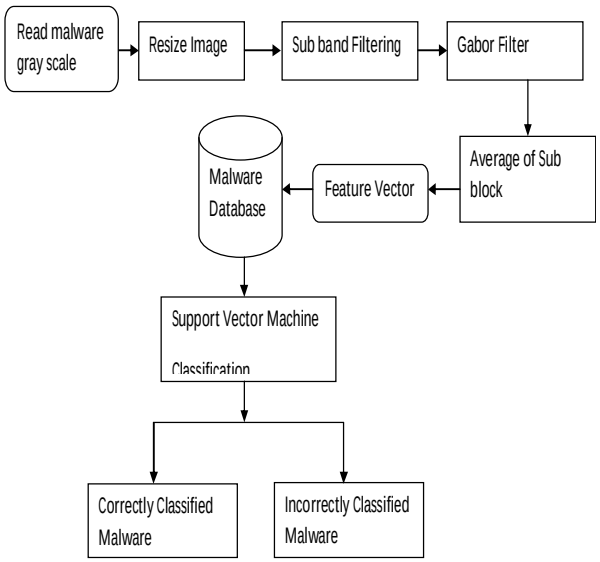


Fig. 1Example of a figure caption

The analysis of malware using various image processing techniques is used in this proposed work. The classification of malware samples using machine learning technique support vector machine. This is most used for classification of different variants of malware. The proposed work is fallows as shown in Fig.1. The malware binaries are converted into gray scale image on the basis of raw binary data which is extracted from executable file. The malware gray scale image is resized then applied sub band filter to get various bands and by using Gabor wavelet we extract second order gradient features. Then average of sub block consider as feature vector based on this vector SVM is classification is done on 3131 samples, in that 2808 samples are classified true positive and 323 are false positive. The feature vector is generated by using fallowing equation (2).

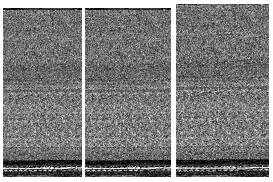
$$iL(x) = \{i1(x), i2(x), \ldots\ldots, ij(x), \ldots, iN(x)x_i)\} \quad (2)$$

RESULTS AND DISCUSSION

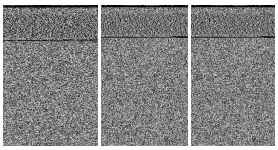
After the experimental results we obtained the accuracy of 89.68% on Manhuer dataset of 3131 samples. Table II describe the experimental result on Manhuer dataset of 3131 samples which consists of only malicious executable files, which is converted into gray scale image all samples look like different patterns in texture. Based on the global features of malware is used in this work.The variants of malware family is shown in Fig.2

TABLE.2 EXPERIMENTAL RESULTS

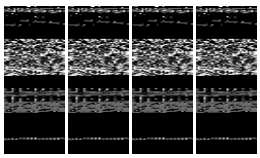
Dataset	Experimental Results			
	Samples	*TPR*	*FPR*	*Accuracy*
Manhuer Dataset	3131	2808	323	89.68%



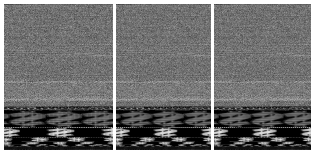
(a) ADULTBROWSER



(b) ALLAPALE



(C) DORFDO



(d) CASINO

Fig. 2 Variants of malware family.

CONCLUSION AND FUTURE WORK

The proposed work the texture features are extracted by applying Gabor wavelet which gives gradient features of texture of different parts of the malware image. Feature vector is formed with 512 dimensional vectors from 3131 malware dataset which contains the different malware variants in 24 malware families. After construction of feature vector, the classification is done on malware samples based on machine learning technique SVM. The experimental result shows classification of accuracy is 89.68%. The future work we planned to concentrate on packed and unpacked malware detected executable file for further research study on static analysis of malware.

Acknowledgment

This research work is supported by UGC under Rajiv Gandhi National Fellowship (RGNF) UGC Letter No: F1-17.1/2014-15/RGNF-2014-15-SC-KAR-69608, February, 2015.

References

[1] Malware- Wikipedia, the free encyclopedia https://en.wikipedia.org/wiki/Malware.

[2] Tantan Xu, (2014). A file fragment classification method based on grayscale image. Journal of computers, vol. 9, No. 8

[3] M. Labs. McAfee threats report: Second quarter (2013). Technical report, McAfee.

[4] Symantec Global Internet Security Threat Report (2010).

[5] Kyoung Soo Han, Jae Hyun Lim, Boojoong Kang, and Eul Gyu Im, (2015). Malware Analysis Using Entropy Graphs. Springer-Verlag Berlin Heidelberg, International Journal of Information Security. 14:1-14, DOI: 10.1007/s10207-014-0242-0.

[6] Said Zainudeen Mohd Shaid, Mohd Aizaini Maarof (2014). Malware Behavior Image for Malware Variant Identification. IEEE, International Symposium on Biometric and Security Technologies (ISBAST).

[7] Infographic: The State of Malware.(2013)

[8] Natraj. L, Yegneswaran.V, Porras.P and Zhang. J. (2011). A Comparative Assessment of Malware Classification Using Binary Texture Analysis. Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence, pp.21-30.

[9] Nataraj, L. (2013). SigMal: A Static Signal Processing Based Malware Triage.

[10] Kong, D. and Yan, G. Discriminant. (2013). Malware Distance Learning on Structural Information for Automated Malware Classification. Proceedings of the ACM SIGMETRICS/International Conference on Measurement and Modeling of Computer Systems, pp. 347-348.

[11] G.Conti, S. Bratus, A. Shubina, A.Lichtenberg, R. Ragsdale, R.Perez-Alemany, B.Sangster, and A. M. Supan. (2010). A Visual Study of Primitive Binary Fragment Types. In Black Hat USA.

[12] Shui Yu, Guofei Gu, Barnawi, Song Guo, Stojmenovic. (2013). Malware Propagation in Large Scale Network. Knowledge and Data Engineering, IEEE Transaction, 27(1):170-179.

[13] Acar Tamersoy, Kevin Roundy, Duen Horng Chau, Guilt by Association. (2014). Large Scale Malware Detection by Mining File-relation Graphs. In Proceedings of KDD'14, August 24-27, New York, NY, USA, Pages: 1524-1533.

[14] Cesare, S.,Xiang, Y. (2010). A fast flow graph based classification based classification system for packed and polymorphic malware on the end host. Advanced Information Networking and Applications (AINA), 24th IEEE International Conference, pp.721-728. IEEE.

[15] Shang,S. Zheng,N. Xu, J ,Xu M. Zhang, H. (2010). Detecting malware variants via function-call graph similarity Malicious and Unwanted Software (MALWARE), 5th International Conference,pp.113-120.IEEE

[16] Santos,I. ,Brezo ,F. Nieyes, J. ,Penya, Y. K, Sanz, B.Laorden, C., Bringas, P.G. (2010). Opcode-sequence-based malware detection. Engineering Secure Software and Systems, pp.35-43.Springer Berlin.

[17] Egele, M., Kruegel, C., Kirda, E., Yin, H., Song, D.(2007). Dynamic spyware analysis. Usenix Annual Technical Conference.

[18] Miao, Q.-G., Wang, Y., Cao, Y., Zhang, X.-G., Liu, Z.-L. (2010). API Capture a tool for monitoring the behavior of malware," Advanced Computer Theory and Engineering (ICACTE), 3rd International conference, pp.V4-390-V394-394. IEEE.

[19] Aziz Makandar and Anita Patrot. (2015).Overview of Malware Analysis and Detection. International Journal of Computer Applications (0975-8887) National Conference on Knowledge, Innovation in Technology and Engineering (NCKITE).

[20] A. Torralba, K. Murphy, W. Freeman, and M. Rubin. (2003). Context-based vision systems for place and object recognition. In Proceedings of ICCV.

[21] GIST Code. http://people.csail.mit.edu/torralba/code/spatialenvelope.

[22] Z. Wen, Y.Hu and W.Zhu. (2013). Research on Feature Extraction of Halftone Image. Journal of Software, vol. 10, pp.2575-2580.

[23] Y. Lan, Y.Zhang and H.Ren.(2013). A Combinational K-View Based Algorithm for Texture Classification. Jornal of Software, vol. 8, pp.218-227.

[24] Jang, J., Brumley, D., Venkataraman, S. (2011). Bitshred: feature hashing malware for scalable triage and semantic analysis," In Proceedings of the 18th ACM Conference on Computer and Communications Security, pp.309-320. ACM.

[25] Trinius. P., Holz. T., Gobel. J., Freiling. F.C. (2009).Visual analysis of malware behavior using tree maps and thread graphs. Visualization for Cyber Security. VizSec. 6th International Workshop, pp.33-38, IEEE.