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. 8, Issue. 3, March 2019, pg.27 – 30

Automatic Malware Detection using Data Mining Techniques Based on Power Spectral Density (PSD)

Asst Prof. Dr. Sefer Kurnaz[1], Yaseen Ahmed[2]

[1,2]Computer Engineering & Altinbaş University, Turkey
[1] sefer.kurnaz@altinbas.edu.tr; [2] ensoftware_alsomaidae@yahoo.com

Abstract— A malware is a software that furtively achieves its process below the appearance of an honest software. Classic approaches apply signatures to separate these software's denote tiny risk to new and hidden instances whose signatures are not offered. The emphasis of malware investigation is instable from applying signature designs to categorising the hateful conduct showed by these malwares. Numerous data mining methods proposed to notice malware mechanically in effectual face. In this thesis, PSD applied to extract the landscapes of malware database and the yield of PSD confidential applying several of data mining methods: Support Vector Machine (SVM), Radial Basis Network (RBF) and multi-layer perceptron (MLP). These techniques presented remarkable results when compared with common researches in this field.

Keywords— Malware, Data mining, Power spectral density, computer security.

I. Introduction

Over the previous couple of years, cyber danger scenery has altered melodramatically and removed from economically interested attacks to targeted attacks, especially Advanced Persistent Threats. Starting from the year 2010 with Operation Aurora, we have witnessed increasing number of such targeted attacks including Stuxnet, Duqu, Flame, Red October, Snake, etc. [1]. This new class of attacks become top priority cyber risks for governments and commercial entities because of its sophistication in terms of tools and techniques employed and well-funded and skilled threat actors. In targeted attacks, efficient workers goal exact objects obstinately with tall motivation, evades security fortifications in home, employs progressive gears and strategies, maintains long-time attendance in board setting and operates sluggish and furtive to evade discovery [2].

Malware theatres an energetic part in achievement of a beleaguered bout and is working nearly in each stage of a bout lifespan pending the operator's goal is realized. It carries out extensive variety of errands counting cooperating model, mounting freedoms, upholding attendance, exfiltrating data, interactive with the workers over knowledge and switch waiters, loud out instructions, etc. Smooth however these errands are not odd to beleaguered malware only and also traditional malware could carry out most of these tasks throughout its execution, beleaguered malware is still predictable to entertainment different than the classical malware because of its stealthy nature.

Informal but too actual effectual way to disclose malwares performance when it contaminates an organization is running it in a controlled atmosphere and detention all the vicissitudes on the system and net throughout the examination procedure. This analysis method is called dynamic analysis and quick insight into malware can be gained by running it in a dynamic analysis sandbox for a very short time (usually 3 minutes). However, it does not work well for all kinds of malware, because there are some malwares that could detect it is running in a sandbox and stop or delay its execution, or could not complete all its tasks within the analysis period, or try to deceive analyst by doing nothing malicious or suspicious. Even with its all limitations dynamic analysis is extensively used in malware analysis field because huge number of new malwares are discovered every day and they are needed to be analyzed in a wild and automatic method. In our thesis work, we used a modified version of a popular open-source automatic malware examination organization, Cuckoo Sandbox, for capturing malware behavior [3].

II. Power spectral density (PSD)

In this chunk fleetingly we will clarify how the forte of a signal is separated in the frequency domain, relative to the fortes of any signals in the atmosphere, is principal to the project of all Linear time-invariant filter deliberate to suppress or assortment the signal [4]. This decent once signal is deterministic, and it change out to be only as precise in the vigorous of accidental signals. For instance, if a musical waveform is audio signals with collective disorder signals, it’s necessarily make a little permit Linear time-invariant filter for mining the audio and curb the complaint signals [5]. Power spectral density purpose current the power of the energy in the signal as a function of frequency. The unit of Power spectral density function is energy (difference) for each frequency(width) and can gain energy in a certain frequency domain by merge Power spectral density in that frequency domain. Figure 3.3 is example about power spectral for a signal.

III.DATA MINING

Data mining, also termed knowledge discovery in databases, in computer science, the procedure of learning stimulating and valuable forms and relations in huge dimensions of data. The arena associations tools from statistics and machine learning with database supervision to investigate huge numerical groups, recognized as data sets. Data mining is extensively applied in commercial, science study, and government safety (detection of network attacks and terrorists).

The term "data mining" is in detail a contradiction, because the objective is the mining of patterns and information from huge quantities of facts, not the selection of feature itself. [6] It also is a slogan [7] and is regularly used to any procedure of important feature or data processing, selection, warehousing, examination, and indicators) as well as any submission of computer decision provision method, counting machine learning and commercial intelligence.

In this thesis, several data mining techniques for data classification such as RBF, SVM, NN.

IV.Proposed Method

The proposed method involves from 2 portions: feature mining and classifier. Formerly, the feature mining built by applying the PSD which attempt to get the greatest and subtle topographies from input features by measuring power spectral density. see equation 1.

$$\boldsymbol{P} = \int_{-\infty}^{\infty} x(f)^2 df \qquad (1)$$

Where ∞ and - ∞ are the areas of purpose, is the frequency standards. The PSD applied to measure PSD for respectively dated strongminded by operator rendering to the measurement and conduct of the input feature, Also, the production of the wired to the classifier which applied SVM, RBF and MLP. Formerly, numerous tests will consume complete to control planned technique (PSD based MLP and RBF and SVM), which consequences are related with greatest presented approaches in this arena that’s proposed in section 2.

The flowchart shown in Figure 1.

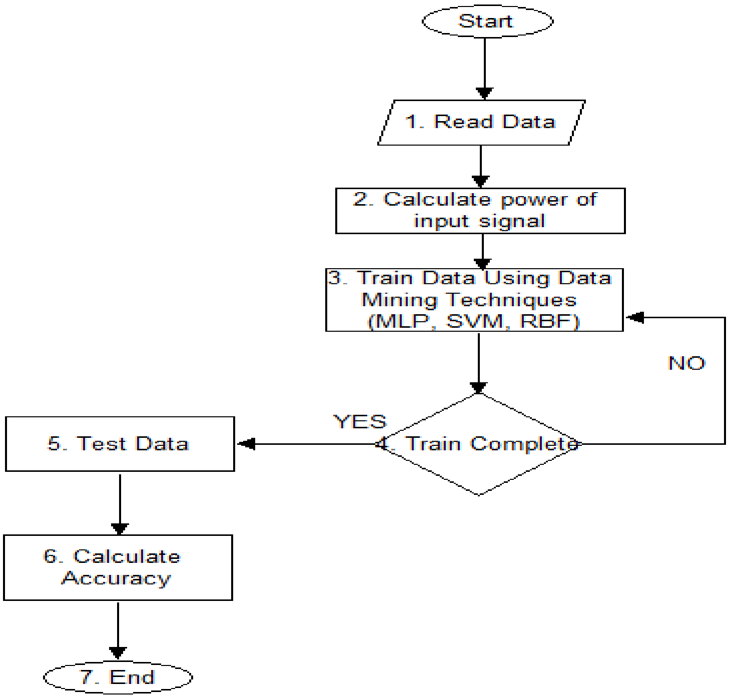

Start
1. Read Data
2. Calculate power of input signal
3. Train Data Using Data Mining Techniques (MLP, SVM, RBF)
NO
YES
4. Train Complete
5. Test Data
6. Calculate Accuracy
7. End

Figure 1: Proposed Method

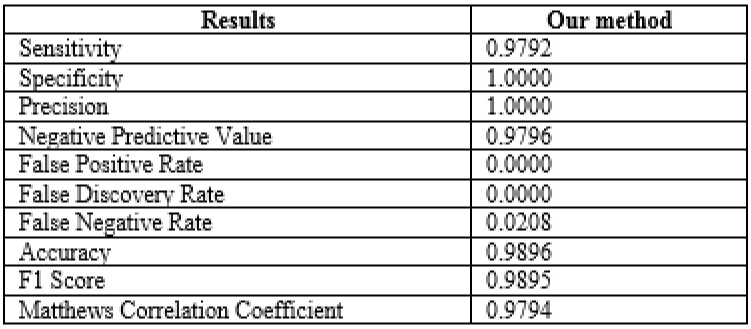
The details step of Figure 1 presented in detail in the flowing:

1. Read the data by using MATLAB program.
2. Calculate the input data power by using Eq (1).
3. Train classifier to classify the features that extracted by Eq(1).
4. If train complete continue with testing section or return to the training.
5. Then, test the trained model malware dataset.
6. Calculate the accuracy of the results.
7. Complete.

V. Conclusions

In this study, several experiments are done by using RBF, SVM, NN. The RBF presented best results which is 99.00, SVM presented 96.9 and NN 94.9. The results of our method shown in Table 1.

TABLE I
CLASSIFICATION RESULTS



Results	Our method
Sensitivity	0.9792
Specificity	1.0000
Precision	1.0000
Negative Predictive Value	0.9796
False Positive Rate	0.0000
False Discovery Rate	0.0000
False Negative Rate	0.0208
Accuracy	0.9896
F1 Score	0.9895
Matthews Correlation Coefficient	0.9794

Then, the obtained results compared with well known researches in this field and prove that our method presented best results then other previous researches.

Methods		ACC
[8]	SVM	96.30
[9]	Naïve-Bayes	93
[9]	J48	98
Our Method	RBF SVM NN	99 96.9 94.9

VI. CONCLUSIONS

In this study, we offered a data mining technique to distinguish malware. The power of signal applied to mine features and improve the performance of the recognition correctness. The proposed method improves the recognition accuracy and processing time of malware recognition.

In future work, various data mining techniques can be used by using PSD to increase the classification accuracy. The proposed method can be applied to other fields such as medical data, recognition and computer vision.

REFERENCES

[1] S. Stolfo, W. Fan, W. Lee, A. Prodromidis, and P. Chan, “Cost-based modeling for fraud and intrusion detection: results from the jam project, in: DARPA Information Survivability Conference and Exposition,” DISCEX’00, Proc., vol. 2, no. IEEE, 2000, pp. 130–144, 2000.

[2] M. Alkasassbeh, A. B. A. Hassanat, and G. Al-naymat, “Detecting Distributed Denial of Service Attacks Using Data Mining Techniques,” vol. 7, no. 1, pp. 436–445, 2016.

[3] A. Abraham, C. Grosan, and C. Martin-Vide, “Evolutionary design of intrusion detection programs,” Int. J. Netw. Secur., vol. 4, no. 3, pp. 328–339, 2007.

[4] A. M. Karim, Ö. Karal, and F. V Çelebi, “A New Automatic Epilepsy Serious Detection Method by Using Deep Learning Based on Discrete Wavelet Transform,” no. 4, pp. 15–18, 2018.

[5] Karim, A. M., Güzel, M. S., Tolun, M. R., Kaya, H., & Çelebi, F. V. (2019). A new framework using deep auto-encoder and energy spectral density for medical waveform data classification and processing. Biocybernetics and Biomedical Engineering, 39(1), 148-159. doi:10.1016/j.bbe.2018.11.004

[6] C. Khammassi and S. Krichen, “A GA-LR wrapper approach for feature selection in network intrusion detection,” Comput. Secur., vol. 70, pp. 255–277, 2017.

[7] Ahmad M. Karim, Mehmet S. Güzel, Mehmet R. Tolun, Hilal Kaya, and Fatih V. Çelebi, “A New Generalized Deep Learning Framework Combining Sparse Autoencoder and Taguchi Method for Novel Data Classification and Processing,” Mathematical Problems in Engineering, vol. 2018, Article ID 3145947, 13 pages, 2018. https://doi.org/10.1155/2018/3145947.

[8] Usukhbayar Baldangombo, Nyamjav Jambaljav, and Shi-Jinn Horng “A STATIC MALWARE DETECTION SYSTEM USINGDATA MINING METHODS”, International Journal of Artificial Intelligence & Applications (IJAIA), Vol. 4, No. 4, July 2013, 10.5121/ijaia.2013.4411.

[9] Mohammad M. Masud, Latifur Khan, and Bhavani Thuraisingham., A Hybrid Model to Detect Malicious Executables, IEEE Communications Society subject matter experts for publication in the ICC 2007 proceedings.