ISSN (Online) : 2348 - 2001

International Refereed Journal of Reviews and Research

Volume 4 Issue 5 September 2016

International Manuscript ID : 23482001V4I5092016-02

(Approved and Registered with Govt. of India)

An Approach for Malware Detection and Predictive Analysis Using Artificial Neural Networks

Hatinder Kaur
M.E Scholar
NITTTR Chandigarh, India
deephatin@gmail.com

Mala Kalra
Assistant Professor
NITTTR Chandigarh, India
malakalra2004@gmail.com

Abstract – **Malware Detection and Predictive Analysis is under research with escalating velocity from more than a decade still there is scope of research because of the increasing vulnerabilities on assorted media and network channels. From a long time, the monitoring of servers and forensic analysis of network infrastructure is done using packet capturing (PCAP) tools and intrusion detection systems (IDS). These activities are performed using PCAP and IDS tools available in the market which includes open source software as well as commercial products. This work underlines the assorted aspects of malware detection and avoidance approaches with the concluding remarks. In proposed work of this research, a unique and effective deep learning based approach is developed and implemented with the base of malware datasets to be fetched from real honeypots and honeynets. The proposed artificial neural net based multilayered approach is used for training and predictive analytics of malware on multiple parameters including error factor, accuracy rate and overall performance. In earlier approaches for malware detection and prior predictions, the density based clustering is used and related highest accuracy is achieved to 97%. Whereas in our proposed approach using ANN the maximum accuracy achieved is 100%.**

Keywords – Malware Detection, Network Security, Packets Forensic, Artificial Neural Networks

I. INTRODUCTION

In the era of information technology and high speed data transmission, the security and privacy is emerging as a major concern for the cyber forensic professionals. Malware Analysis is one of the utmost focused points under the sight of cyber forensic professionals and network administrators Malware is a program which is deliberately designed to be harmful. There are basically two malware analysis techniques i.e. dynamic malware analysis and static malware analysis [1]. In static malware analysis, malware analysis is done without executing the malware. In dynamic malware analysis, a malware is executed in virtual machine [2] or in an emulated environment [3].

The security against malware traffic is very important because of the escalating number of assaults on assorted network channels and damaging the digital infrastructure a lot. As per the Global Security Report released by AppRiver, "During April 1, 2016 and June 30, 2016 the security firm recorded 4.2 billion malicious emails and 3.35 billion spam emails. Meanwhile, there were 43 million unique web-borne threats daily throughout the second quarter."

The Kaspersky Security Network recently logged 10,000 malware infection attempts broadcasted worldwide. The countries which were affected most include Poland, Brazil, Colombia, Mexico, Ecuador, Greece, Portugal, Peru, Tunisia,

ISSN (Online) : 2348 - 2001

International Refereed Journal of Reviews and Research

Volume 4 Issue 5 September 2016

International Manuscript ID : 23482001V4I5092016-02

(Approved and Registered with Govt. of India)

Germany, Venezuela, and Israel as shown in Figure 1.

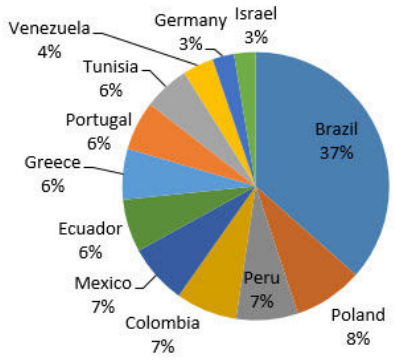

Figure 1 Malware Distribution in Year 2016[4]

Malware developers use various code obfuscation techniques like register reassignment, subroutine reordering, code transposition etc. to prevent detection of malware by traditional methods like antivirus, firewall and gateways which generally use classical approaches like signature based technique. In this domain, there are numbers of tools and technologies available by which the malware or simply malicious packets [5] can be transmitted over the network so that the data channel can be damaged using virtual attacks. Generally malware is divided into two basic categories as malicious software and non-malicious software (also called as benign), depending upon various factors like variety of attributes like replication tendency and strategy, purposes of creation, method of propagation and containment methodology. Besides various classical approaches, there are number of related terms by which the fake and malicious packets [6] can be transmitted including Beast, Trojan Horses, Scareware, Rootkits, Evasion, Backdoors, Suspicious packers, Trojan Spy, Browser Hijacker, Ransomware, Rogue Software, Botnets.

An artificial neural system [7] is a framework in view of the operation of organic neural systems. A neural network is a massively parallel distributed processor made up of simple processing units, which has a natural propensity for storing experimental knowledge and making it available for use. Artificial neural networks gives an excellent performance towards the linear as well as non-linear operations, also provides parallel nature of computing and operations management. Artificial neural network can be implemented and integrated for any type of application domain and has no issues or complexities related to jitter.

II. RELATED WORK

The earlier work on malware detection primarily focus on signature based analysis and historical frequency analysis. The work in earlier approaches include clustering, rule mining and visualization while the current work is based on the predictive interpretation and prior checks on malware packets using deep learning based approach. The classical work is finding out and predicting the behaviour of malware but not with significant and higher accuracy rate. There was need to improve the classical approach using effective architecture which this work is addressing here. The classical work is around 5% less accurate than the highest level which is achieved in the proposed approach and giving utmost level accuracy and very less or negligible error rate that is improving the performance of overall proposed system.

The results and predictions in the existing work are lacking in multiple aspects including accuracy, performance, efficiency and complexity.

V. M. Afonso et al. [8] presents a unique model for logging of malware datasets so that the future predictions can be made. P. V. Shijo et al. [9] explains the uses and implementations of the positive and favorable aspect for both of static as well as dynamic methods for classification and analysis of assorted malicious technology software. E. Gandotra et al. [10] underlines the use of machine learning and advance data mining algorithms for malware predictions. A. Tamersoy et al. [11] proposed the work based on malware avoidance, prediction and evaluation using a novel

ISSN (Online) : 2348 - 2001

International Refereed Journal of Reviews and Research

Volume 4 Issue 5 September 2016

International Manuscript ID : 23482001V4I5092016-02

(Approved and Registered with Govt. of India)

model and architecture. This is very versatile that can be used for any type of malware. Using this effective approach, the multiple layers can check the data formats and attack type of malware. K. Mathur et al. [12] suggested the Bi-features approach for increasing the precision of existing static mono-features analysis and to withdraw the drawbacks of the existing techniques. This work works on assorted aspects of malicious predictions including payload, vulnerability and propagation. M. Overton et al. [13] presents the approach of using intrusion detection systems and give assurance regarding efficient and fast detection and removal of malware from a system. C. Lin et al. [14] paper proposes an approach offering a competitive malware detection procedure having generic and efficient algorithms to classify malware. In this work, the selection and the extraction of features are done to significantly reduce the dimensionality of features for training and classification. D. Stopel et al. [15] presents an approach using artificial neural networks for detecting worms on the basis of abnormal computer behaviour.

III. PROPOSED WORK

The work based in the existing malware detection includes signature analysis from historical datasets based on the heuristic functions or frequency analysis. This proposed work is based on multiple layers. The first layer of implementation is extracting the feature points from network traffic so that payload and penetration level can be extracted. A refined or cleaned dataset is used for training a neural network so that the predictions can be made.

In our proposed work, a model will be trained using two layered feed-forward back propagation neural network in MATLAB, so that the penetration level and prediction of malware can be done.

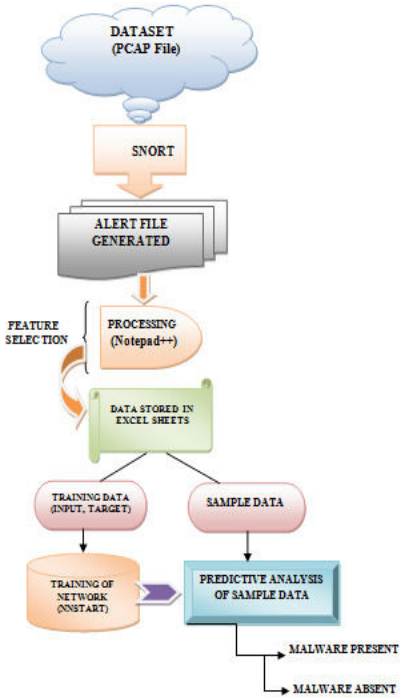

Figure 2 Flow of the Work

Steps in flow of work:

STEP I: Collection of datasets.

Datasets can be captured from several open data portals. Packet capturing is one of classical and most frequently used task performed by the network administrators. To analyze the penetration level of packets in network channels, there is need to have log or record of raw packets. For this purpose, PCAP file is used which keep the raw network traffic in binary format. PCAP files are cross platform by which multiple operating systems and software tools can read and analyze the internal patterns of PCAP.

STEP II: Generation of alert file using Snort.

PCAP files can be analyzed using Snort to generate an alert file. Snort IDS generates the full alert file which is classically not available in other tools. The alert file contains the overall behavior of packets in PCAP which is the base and key focus for training and predictions using ANN in our proposed work.

ISSN (Online) : 2348 - 2001

International Refereed Journal of Reviews and Research

Volume 4 Issue 5 September 2016

International Manuscript ID : 23482001V4I5092016-02

(Approved and Registered with Govt. of India)

STEP III: Cleaning of the data.

The alert file generated through Snort comprises of all the parameters in the PCAP file. Features of the fetched dataset in the alert file will be investigated and only the required features are selected from the text file (alert file) using notepad++ and saved on to an excel sheet. Feature selection is important for categorization or classification of data as it enhances the accuracy of the results.

STEP IV: Categorization of data.

The data in the excel sheets is then categorized as training dataset and sample datasets. The training dataset comprises of input data and the target data for training the network, whereas sample datasets consists of different samples used for prediction analysis.

STEP V: Training of the neural network.

An artificial neural network is trained using ANN toolbox in MATLAB. The training of ANN model comprises of all the vulnerabilities which will generate the prediction on other datasets (sample).

STEP VI: Prediction of sample data.

Prediction analysis is performed on trained network using sample datasets.

V. EXPERIMENTAL SETUP AND RESULTS

a) Experimental setup

The proposed work is implemented using nprtool in MATLAB. The training is done through a two layered feed forward back propagation neural network.

The neural network to be trained takes PCAP files captured from various sources as an input. The two parameters used for training the neural network are input dataset and the target dataset (output).The training of the neural network comprises of three stages. Training the network is the initial stage in which the inputs are fed to the network, based on which the neurons get trained by identifying the pattern of the input. The second stage is validation which is used to measure network generalization. This stage also determines when to stop training the network in order to get maximum efficiency. Testing is the last stage, in this stage the neurons which are trained during the first stage are tested in order to measure the performance of the trained network.

The total data used to train the network is also divided into three sub-parts, such that 70% of the input data is used for training, 15% is used for validation and the rest 15% is used for testing. The ratio of input data used for training, validation and testing can be changed.

b) Results

Training the neural network comprises of various parameters such as epochs, hidden layers, number of neurons and validation checks.

The network to be trained is a two layered neural network, the maximum number of epochs in training the network are 1000 whereas the total validation checks are 6.The training of network automatically stops when generalization stops improving, as indicated by an increase in the mean square error of the validation samples. Whereas a network can be retrained again and again until the desired results are achieved.

During the training of the network, the number of layers are kept static but the number of neurons (N) in the hidden layer are varied in an increasing order i.e, training starts at N=10 to N=300 as shown in Table 1.

Table 1. Training the network

Neurons	Epoch	Training accuracy	Validation accuracy	Test accuracy
10	36	99.7%	99.7%	99.7%
30	28	99.9%	99.7%	99.8%
50	43	99.9%	99.9%	99.8%
70	50	99.9%	99.9%	99.8%
90	62	99.9%	99.9%	99.9%
100	38	99.9%	99.9%	99.9%
130	38	99.9%	99.9%	99.9%
150	34	99.9%	100%	99.9%

ISSN (Online) : 2348 - 2001

International Refereed Journal of Reviews and Research

Volume 4 Issue 5 September 2016

International Manuscript ID : 23482001V4I5092016-02

(Approved and Registered with Govt. of India)

180	45	100%	99.9%	100%
200	41	100%	99.9%	99.9%
250	**39**	**100%**	**100%**	**100%**
300	32	99.9%	99.8%	99.9%

As the network was trained and retrained several times using different number of neurons, it was found that the accuracy in terms of training, validation and testing was highest (100%) at N=250 and also the value of mean square error was minimum. Figure 3 shows the performance graph at N=250 and Figure 4 shows the confusion matrix of trained network at N=250.

It was also observed that higher the number of neurons in neural network better the accuracy, whereas more neurons need greater training time.

The higher number of hidden layers leads to hard training and also takes much time, so less hidden layers (1 0r 2) performs well.

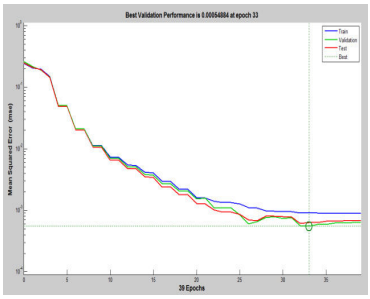

Figure 3 Performance graph of the Optimal and Higher Efficiency ANN Model

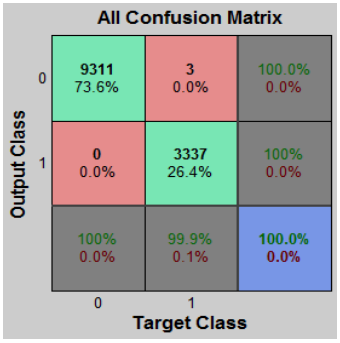

Figure 4 Cumulative Confusion Matrix

The values in the above confusion matrix represents very less or null equivalent values in the false positive and false negative blocks which shows that the dataset is accurate for predictions and there is no ambiguity.

The prediction analysis was performed on the trained network with three different samples of datasets using sim function so that the network can be tested and predicted values can be fetched from the sample datasets.

PREDICTION USING SAMPLE:

```
% prediction of sample dataset of value 1064
myprediction=sim(netP,sampledifferent);
malware_present=0
malware_absent=0
for i=1:1064
if(myprediction(i)>=0.5)
   malware_present=malware_present+1;
disp('Malware Present')
else
   malware_absent=malware_absent+1;
disp('Malware Absent')
end
end
disp('Prediction of.....')
malware_present
disp('Prediction of.....')
malware_absent
Data =[malware_present  malware_absent]
mylog=[malware_present malware_absent]
figure
subplot(1, 3, 1)
pie(mylog)
subplot(1, 3, 2)
plot(mylog)
subplot(1, 3, 3)
bar(mylog)
```

ISSN (Online) : 2348 - 2001

International Refereed Journal of Reviews and Research

Volume 4 Issue 5 September 2016

International Manuscript ID : 23482001V4I5092016-02

(Approved and Registered with Govt. of India)

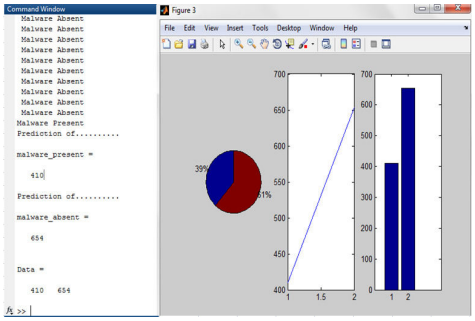

Figure 5 Result of Prediction using sample.

While comparing the results of the proposed approach with the previous approach (Density Based Clustering) it is evident that the results in increasing number of records in the dataset has the highest accuracy of 97% [16].Whereas in the proposed approach as the number of neurons are increased the accuracy also increases to 100% as shown in Table 2.

Table 2 Comparison of Accuracy (%) between Density Based and Proposed Approach

Scenario	Density Based Clustering Approach [16]	Proposed Approach
1	94	99.7
2	95	99.7
3	95.8	99.9
4	96.8	99.9
5	95.8	99.9
6	95.7	99.9
7	95.7	99.9
8	97	100

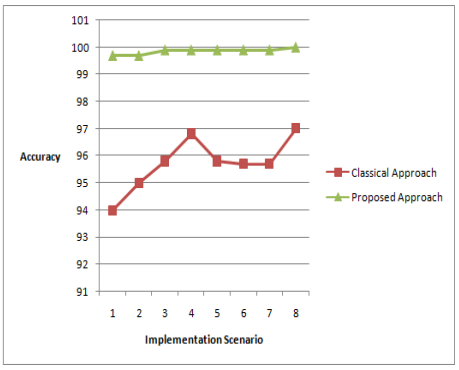

Figure 6 Comparison on Accuracy (%) Parameter between and Proposed Approach

In the graphical (Figure 6) comparison, it is also evident that with the increasing order of datasets and neurons, the performance of proposed approach is effective and moving towards 100% as compared to maximum 97% accuracy level in the earlier approach [16]. The proposed approach is better in terms of faster execution and minimum error rate which is generally required in the fault tolerant and security specific domains.

V. CONCLUSION

Despite of number of algorithms and predictive model developed for analysis and pre-checking of probability, there is huge scope of further research.

As this work is relying on the training of malware fingerprints captured from network traffic, the upcoming future work can be done on the predictive analysis using soft computing approaches or nature inspired algorithms which are generally meant and developed for optimization in assorted random iterations.

There are number of optimization approaches using which the efficiency, accuracy and performance factors can be improved. The integration of soft computing approaches are prevalent in the research community which provides fuzzy based execution and global optimization from existing results.

There exist approaches like metaheuristics and hyper-heuristic that can be integrated for deep learning of malware and predictive analysis.

REFERENCES

[1]S. Gadhiya, K. Bhavsar,"Techniques for Malware Analysis", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3, Issue 4, pp. 972-975, 2013.

[2]R. P. Goldberg,"Survey of virtual machine research", IEEE Computer magazine, pp. 34-45, 1974.

[3]F. Bellard,"QEMU, a Fast and Portable Dynamic Translator", In FREENIX Track of the USENIX Annual Technical Conference, 2005.

[4]http://www.statista.com/statistics/266164/crime ware-infections-of-desktop-pcs-by-type-of-malware/

[5] H. Mahmassani, J. C. Williams, and R. Herman, "Investigation of network-level traffic flow relationships",1984.

[6]D. Emm, M. Garnaeva, R. Unuchek, D. Makrushin, and A. Ivanov, "IT THREAT EVOLUTION IN Q3", 2015.

[7]T. Hill and M.Leorey. "Artificial neural network models for forecasting and decision making." International Journal of Forecasting, Vol. 10, Issue, pp. 5-15, 1994

[8] V. M. Afonso, D. S. F. Filho, A. R. A. Gregio, P. L. de Geus, M. Jino, "A hybrid framework to analyze web and operating system malware", IEEE International Conference on Communications (ICC), Ottawa, pp. 966-970, June 2012.

[9] P.V. Shijo, A. Salim, "Integrated Static and Dynamic Analysis for Malware Detection", International Conference on Information and Communication Technologies, Kochi, pp. 804-811, December 2015.

[10] E. Gandotra, D. Bansal, S. Sofat, "Malware analysis and classification: A survey", Journal of Information Security, Vol. 5, Issue 2, pp. 56-64, 2014.

[11] A. Tamersoy, K. Roundy and D. H. Chau, "Guilt by association: large scale malware detection by mining file-relation graphs", 20th International ACM Conference on Knowledge Discovery and Data Mining, pp. 1524-1533, August 2014,

[12]K. Mathur, S. Hiranwal, "A Survey on Techniques in Detection and Analyzing Malware Executables", International Journal of Advanced Research in Computer Science and software engineering, Vol. 3, Issue 4, April 2013.

[13]M. Overton, "Anti-Malware Tools: Intrusion Detection Systems", EICAR Conference, Malta, pp. 1-22, May 2005.

[14]C. Lin, N. J. Wang, H. Xiao and C. Eckert, "Feature Selection and Extraction for Malware Classification", Journal of information science and engineering, , Vol. 31, Issue 3, pp. 965-992 , May 2015.

[15]D. Stopel, R. Moskovitch, Z. Boger, Y. Shahar, Y. Elovici, "Using artificial neural networks to detect unknown computer worms", International Journal of Neural Computing and Applications, Vol. 18, Issue 7, pp. 663-674, October 2009.

[16] S. C. Ming, "Network Malware Detection and Accuracy Predictions using Dynamic Density Based Clustering", International Journal and Bulletin of Multidisciplinary Research (IJNMR), Vol. 4, Issue 3 September – December 2015.