RESEARCHERID
THOMSON REUTERS

[Bist* *et al.*, 6(6): June, 2017]
IC™ Value: 3.00

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

ALGORITHMS FOR METAMORPHIC MALWARE DETECTION

Ankur Singh Bist[*1], Anubhav Patrick[2], Rudranshu Sharma[3], Saurabh Pargein[4] and Sanjeev Kumar Dwivedi[5]

[*1,2,3,4&5]CSE, IT & EC Department, KIET, Ghaziabad & Gehu Bhimtal

DOI: 10.5281/zenodo.809189

ABSTRACT

This survey paper describes a proper literature review of algorithms used for analysis and detection of metamorphic malware. Short descriptions of various algorithms/methods used in concerned domain are provided. Based on number of citations various research articles are taken to conduct this study. Different aspects of various algorithms are taken into consideration. Recommendations are given based on analysis done to conduct this study.

KEYWORDS: metamorphic malware, machine learning, algorithms, detection.

I. INTRODUCTION

Rapid growth of malwares is one of the major areas of concern. Lot of research articles have been written in this area. Evolution of malwares has made them very complex. At the same time researchers in same domain are finding or exploring new algorithms in order to mitigate the negative impacts of malwares. Malwares can be divided into computer viruses, worms, Trojan horses, logic bombs, botnets etc. Following figure shows the evolution of malwares from stealth to metamorphic malwares.

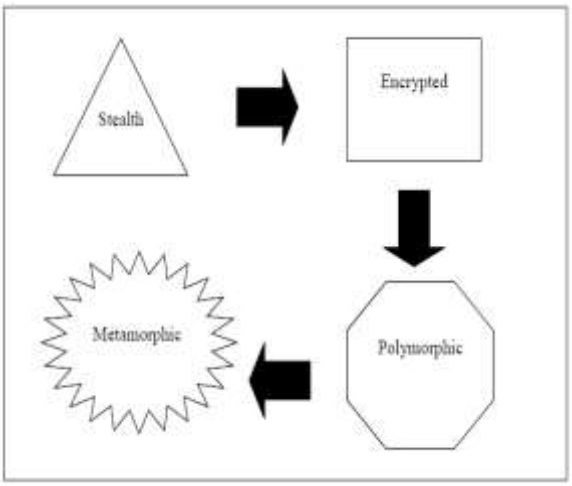

Figure 1: Evolution of malicious code

Metamorphic malwares mutate their body in order to conceal their behavior. There are lots of techniques used by metamorphic malwares like dead code insertion, variable renaming etc. Following figure explains the attack structure of malwares over time. The equation of malware designers and antimalware designers is quite complicated. The need of reverse engineering in terms of malware analysis is desired in evolved manner [1]. The techniques designed in order to detect malwares have evolved enough now it is not easy for malware designers to create mutants with past ease [12] [13].

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

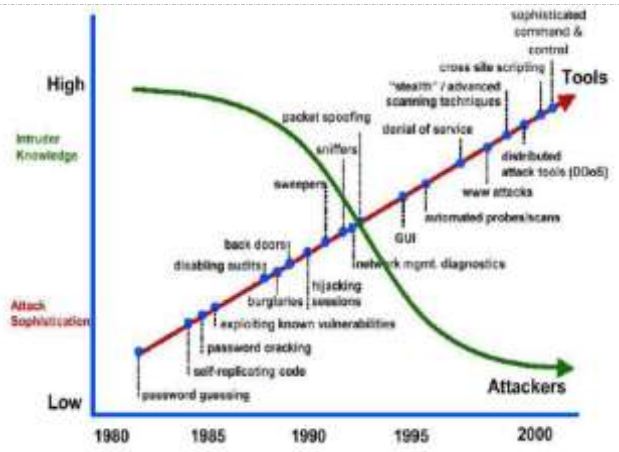

Figure 2: Attack scenario over years

II. ALGORITHMS FOR MALWARE DETECTION

a. Hidden Markov Models

Markov process is a mathematical model in which the transitions between states and internal states are known to the user. HMM comes under special cateogary of malware model. After training HMM model can be utilized for classifying metamorphic malwares.

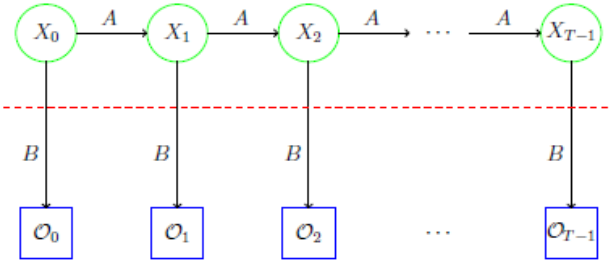

Figure 3: Simple view of HMM

Annachhatre solved the problem of malware detection with the help of hidden markov models. Malware samples and compiler dataset is used to train hidden markov model. More than 9000 samples are taken for analysis and designed clusters from scores. Results reflect the importance of HMM for malware classification [2].

Sridhara explained about the classification of malwares based on MWOR kit. The results are taken on various parameters in order to identify the utility of algorithms for practical use. Results show the importance of hidden markov models for malware detection [3].

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

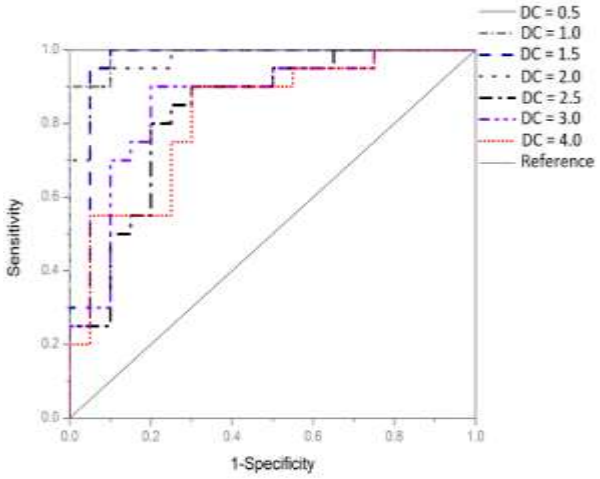

Figure 4: Padding ratios in ROC curve

b. K- Means and Expectation Maximization Clustering Techniques

K- Means clustering is a technique for designing k clusters from data samples. There are various steps involved for cluster formation.

1. Mention the number of clusters.
2. Select initial centroids.
3. Relate each data point to the closest centroid.
4. Re-calculate the cluster centroids based on current clustering of data samples.
5. If there exist major change in centroids, re –arrange the data points to the nearest centroids.
6. The method is repeated until there is no major change in the centroid positions.

Expectation minimization clustering is a popular unsupervised method in the field of machine learning. This method utilizes Gaussian mixture models to calculate the maximum likelihood estimates of the parameters in the data. Unlike *k*-means clustering, this calculates distance index to classify the data sets into different clusters, EM clustering uses existing probability distributions of the data. That is, instead of conveying a data point completely to a single cluster, EM clustering estimates the probability with which a data point maps to each cluster. The data points are then allocated to the cluster with maximum probability. Thus, EM clustering technique can be visualized as soft clustering technique. EM clustering technique has two steps, namely E-step and M-step. The clustering technique iterates between these two steps to calculate the maximum likelihood of the parameters of the data. The process moves until the parameters converge or the maximum number of iterations is attained.
Narra proposed a malware clustering techniques using k-means and Expectation Maximization clustering. Support vector machine classifier is also used for malware classification and compared with clustering methods. In the initial phase HMM is trained with GCC, TurboC, MinGW, TASM, Clang, MWOR and NGVCK. K-Means is used to cluster HMM scores with reasonable accuracy. The main problem was that k-means does not take into account data distribution. To mitigate this technique EM clustering is used [4].

c. Artificial Neural Network

ANNs are based on the artificial neurons capable of performing classification task based on internal computations. ANN takes data from input layer and transmit data to next layer final layers generates the classification output. The layers between first layer and final layers are termed as hidden layers. With new advancement in this field like feed forward, recurrent and convolution neural network, NNs are now in wide used in industry as well as in research.

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

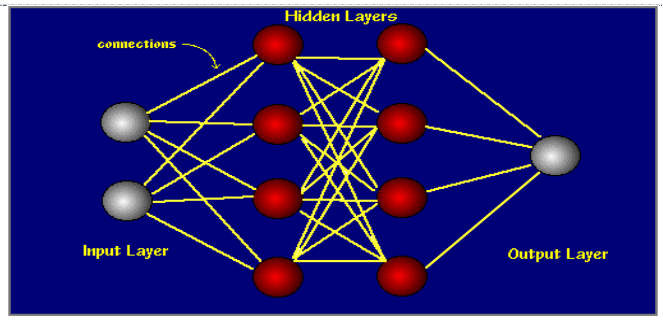

Figure 5: Simple representation of neural network

Golovko proposed the intelligent adaptive and self learning technique based on artificial immune system and artificial neural network. Experimental results show that proposed method worked well for malware detection [5].

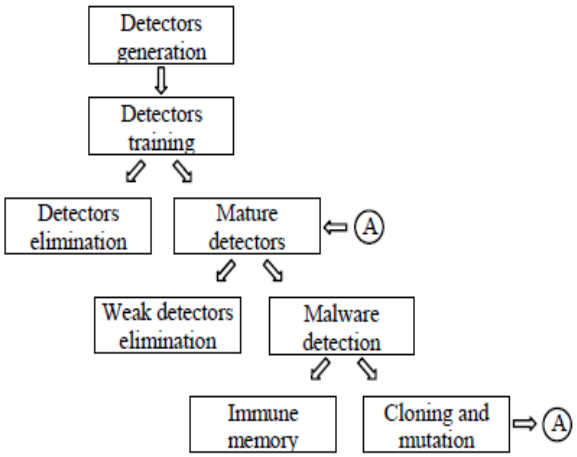

Figure 5: Artificial immune system

d. Association Rules and Fuzzy Association Rules

In order to find out unknown association rules from data sets association rule mining concept is used. If and else architecture is used to trace out various situations. Literature shows that association rule mining could be helpful to classify malicious data sets with higher accuracy. This technique is used in fusion with other conventional techniques like support vector machines, hidden markov models etc. for malware detection.

e. Bayesian Network

A Bayesian network is based on probabilistic theory that represents the relation between various entities. In Bayesian graphical model node represents the random variables and edges represent the relationship between them.

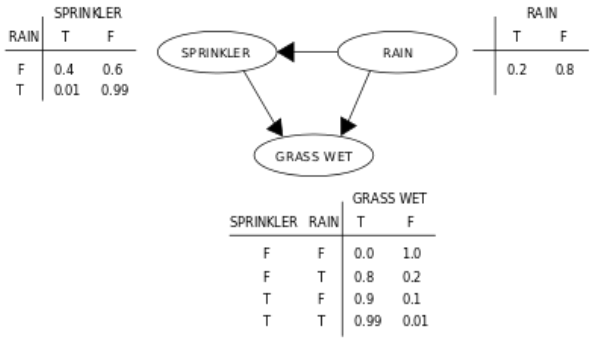

Figure 6: Bayesian model

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

NedaShabani proposed Bayesian network to identify metamorphic malware. Experiment showed that Bayesian network can detect metamorphic malware with better accuracy. Following figure reflects the results of experimental analysis [6].

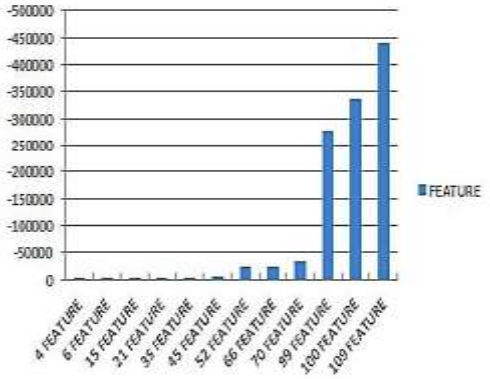

Figure 7: Different Bayesian network accuracy states

f. Support Vector Machines

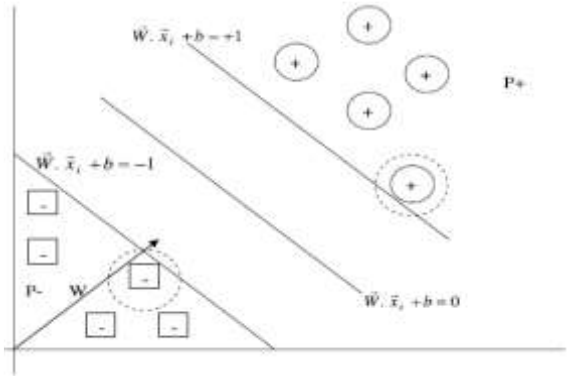

Figure 8: Wrapper heuristics using SVM

Tanuvir used support vector machines for classification of metamorphic malware. Hidden markov model, opcode similarity and simple substitution started miss-classification with morphed malware samples. SVM accuracy falls with morphing but gives better accuracy as compared to other classification techniques mentioned in paper [7].

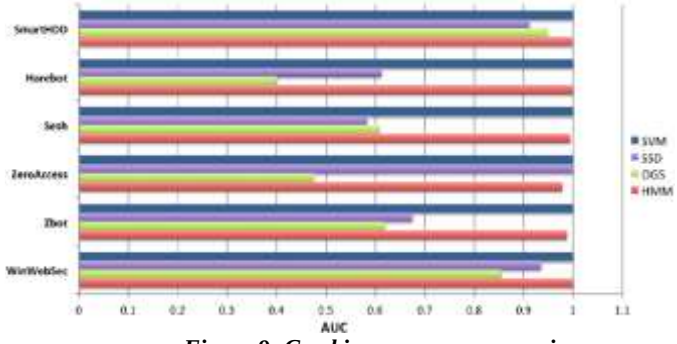

Figure 9: Combine accuracy comparison

g. Decision Trees

A decision tree is designed which is used for classification task. ID3 and C4.5 are very popular algorithms in decision tree category. Intuitive knowledge expression is an advantage of decision tree.

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

Figure 10: Decision tree

Vinod proposed robust feature selection technique for metamorphic malware detection. J48, Bagging and Random Forest implemented in WEKA are used for classification. Proposed technique proved itself efficient for metamorphic malware detection. 250 benign and 360 malware samples are taken for classification [8].

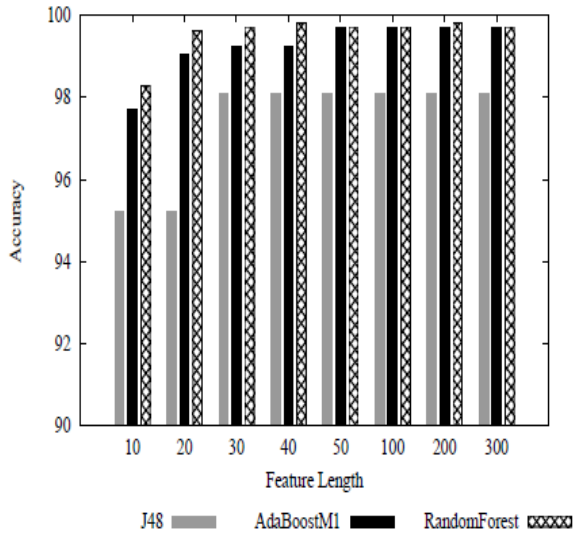

Figure 11: Weight of evidence of text evaluation metric

h. Prediction by Partial Matching (PPM)

Compression is a technique used for minimizing the size of file. Unique symbol assignment technique is used to reduce the size of file in significant manner. Prediction by partial matching designs compression models. Based on compression models these are used for classification. Lee used Adaptive data compression method to analyze malwares. Classification is done with the help of segment sequences. Results reflect efficiency of proposed technique for metamorphic malware detection. Following figure give one important aspect of obtained results [9].

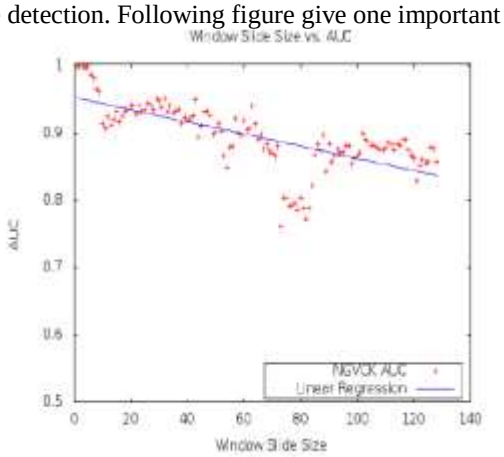

Figure 12: Window slide size vs AUC

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

i. Edit distance and Pairwise alignment

Patel used edit distance and pairwise alignment technique for the detection of metamorphic malware and found good results on morphed malware samples [10].

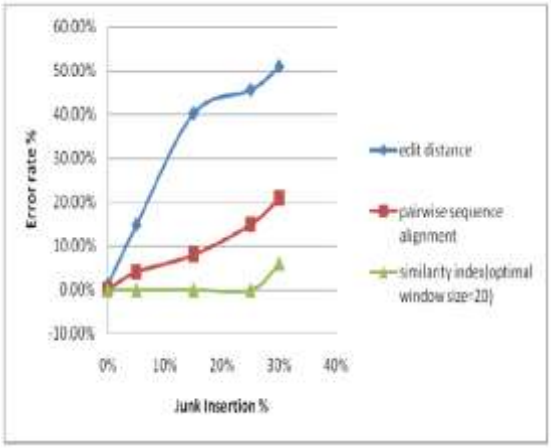

Figure 13: Error rates

j. Vigenère Score

Deshmukh proposed a technique based on index of coincidence which is used to determine the length of keyword in a Vigenère ciphertext. Proposed method generated interesting results. Following figure presents one of the important aspects generated from experiment [11].

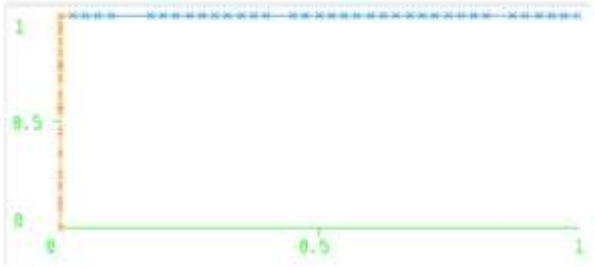

Figure 14: NGVCK ROC

III. CLASSIFICATION PARAMETERS

Confusion Matrix- In predictive analytics, it is a table with two rows and two columns that signifies the number of false negatives, false positives, true positives, and true negatives.

True Positive (TP)- It defines how correctly an algorithm is in identifying a malware as virus.
False Positive (FP)- It is the number of mistakenly classified instances as positive.
True Negative (TN)- It is the number of correctly classified instances as negative.
False Negative (FN)- It is the number of mistakenly classified instances as negative.
Accuracy=(TP+TN)/(TP+TN+FP+FN)
Precision=TP/(TP+FP)
Recall=TP/(TP+FN)
F-measure=(2*precision*recall)/(precision+recall)

IV. CONCLUSION

There are lots of research articles that explained different solutions regarding metamorphic malware detection. This article includes some of important results observed by researchers in past years. This study will be helpful for those working in the field of metamorphic malware detection.

V. ACKNOWLEDGMENT

I would like to thank all who directly or indirectly supported in this work

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

VI. REFERENCES

[1] Thomas, C., & Balakrishnan, N. (2009). Performance enhancement of intrusion detection systems using advances in sensor fusion. *Supercomputer Education and Research Centre Indian Institute of Science, Doctoral Thesis, 304pp. Available at: http://www. serc. iisc. ernet. in/graduation-theses/CizaThomas-PhD-Thesis. pdf*.
[2] Annachhatre, C., Austin, T. H., & Stamp, M. (2015). Hidden Markov models for malware classification. *Journal of Computer Virology and Hacking Techniques*, *11*(2), 59-73.
[3] Sridhara, Sudarshan Madenur, and Mark Stamp. "Metamorphic worm that carries its own morphing engine." *Journal of Computer Virology and Hacking Techniques* 9.2 (2013): 49-58.
[4] Narra, U., Di Troia, F., Corrado, V. A., Austin, T. H., & Stamp, M. (2016). Clustering versus SVM for malware detection. *Journal of Computer Virology and Hacking Techniques*, *12*(4), 213-224
[5] Golovko, Vladimir, et al. "Neural network and artificial immune systems for malware and network intrusion detection." *Advances in Machine Learning II*. Springer Berlin Heidelberg, 2010. 485-513.
[6] Shabani, Neda, and Majid Vafaei Jahan. "Metamorphic virus detection based on Bayesian network." *Technology, Communication and Knowledge (ICTCK), 2014 International Congress on*. IEEE, 2014.
[7] Singh, T., Di Troia, F., Corrado, V. A., Austin, T. H., & Stamp, M. (2016). Support vector machines and malware detection. Journal of Computer Virology and Hacking Techniques, 12(4), 203-212.
[8] Kuriakose, J., & Vinod, P. (2015). Unknown metamorphic malware detection: Modelling with fewer relevant features and robust feature selection techniques. *IAENG International Journal of Computer Science*, *42*(2), 139-151.
[9] Lee, Jared, Thomas H. Austin, and Mark Stamp. "Compression-based analysis of metamorphic malware." *International Journal of Security and Networks* 10.2 (2015): 124-136.
[10] Patel, M. (2011). *Similarity tests for metamorphic virus detection* (Doctoral dissertation, San Jose State University).
[11] Deshmukh, Suchita. "Vigenère Score for Malware Detection." (2016).
[12] Bist, Ankur Singh. "Detection of metamorphic viruses: A survey." *Advances in Computing, Communications and Informatics (ICACCI, 2014 International Conference on*. IEEE, 2014.
[13] Bist, Ankur Singh. "Classification and identification of Malicious codes." *IJCSE* (2012)..

CITE AN ARTICLE

Bist, A. S., Patrick, A., Sharma, R., Pargein, S., & Dwivedi, S. K. (2017). ALGORITHMS FOR METAMORPHIC MALWARE DETECTION. *INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY*, *6*(6), 325-332. doi:10.5281/zenodo.809189