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

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

DETECTION OF METAMORPHIC VIRUSES USING BIO-INSPIRED FEATURE FUSION

Ankur Singh Bist*, Dr. Anuj Sharma
[*] SVU Gajraula
PHD Scholar SVU Gajraula

DOI: 10.5281/zenodo.438110

ABSTRACT

Increment in computer viruses is a major problem for researchers working in the field of computer virology. The existing techniques cannot provide the complete assurance towards disinfection, but still lot of advancement is going on in the direction of computer virus detection. Our proposed technique includes the test cases with 3-d biological signature flavour that are designed for the identification of metamorphic viruses. Test case contains the combination of bio inspired techniques to assure exact detection that leads to detection of metamorphic viruses with very low false positive and false negative rate. Experimental results are obtained to demonstrate that our method is suitable for identification of metamorphic viruses and achieve satisfactory performance.

KEYWORDS: Metamorphic, Computer Virology, Bio Inspired Approaches, Pairwise Alignment.

INTRODUCTION

Internet has become target of malicious codes due to its increasing use. Malicious codes are executable code and have the capability to replicate. It makes their survival strong. Viruses design and evolution attached with the area of programming. Similar to other computer programs viruses carry functions that are intelligent for providing protection in such a manner that detection remains not easy for virus scanner [1].

Viruses have to take various procedures of intellect for continued existence. That is why they may have complex encrypting and decrypting engines. These are the most frequent methods used by computer viruses in current scenario. They make use of these techniques to mask themselves from the antivirus and to adopt the certain environment for their expansion [2].

Polymorphic viruses try to hide the decrypting module. More complex methods were developed enabling the virus designers to change the code of one virus file and make multiple morphed copies while maintaining its functionalities. These are the type of viruses which have the ability to mutate itself with the code changed but without changing its functionalities. Metamorphic virus can become a serious threat considering the fact that there can be thousands of variants of one virus file with their signature being totally different.

Metamorphic viruses transform its code in a specific manner very frequently and require to be prohibited. Their analysis will lead to evolve a framework where the overall process of detection will be bounded in specific outcomes of continuing evolving results. It is essential to make a distinction between replicating programs and its similar forms. Reproducing programs will not necessarily damage your system [3] [8].

There is big fight between designers of virus and antivirus. The enhanced knowledge about the certain patterns, specifications can be designed. Various malicious codes can be evolved and incremented in well precise and efficient manner. For perfect identification of a metamorphic virus, identification routines must be written that can generate the essential instruction set of the virus code from the actual occurrence of the infection.

POPULAR ALGORITHMS

The types of sequences analysed will be prepared by finite lists that have symbols from an alphabet. In the field of bioinformatics a sequence of entities seen in the nucleotides depicted as a pattern of letters from the English

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

alphabet in which each letter corresponds for a unique entity. Similar technique can be used for the detection of metamorphic viruses. Basically a virus is made up of a sequence of op-codes which can be chosen from the disassembled version of the metamorphic virus. Opcodes is replaced with the English keywords that give rise to sequence [4] [5]. There are lot of algorithms that are being used for malware classification in past. Some of them are listed as follows:-

- n-gram
- Pairwise alignment
- Opcodes frequency analysis
- 3-D signature analysis
- Singular Value Decomposition

N-Grams

N-Grams is an algorithm for word prediction using probabilistic methods to predict next word after observing N-1 words so, computing the probability of the next word is directly related to calculating the probability of a sequence of words.

Unsmoothed N-grams

The easiest probabilistic model for word prediction can be assigning equal probability to each word. So suppose that there are N words in a language, and then the probability of any word following another word would be 1/N. However, this method neglects the truth that some words are more common than the others in languages.Ohm Sornil *et al.* explained about Malware classification using N-grams sequential pattern feature. N-grams are extracted from malicious program files, sequential n-gram patterns are determined, pattern statistics are calculated, and a classification technique is used to determine the family of malware. Classification models C4.5, multilayer perceptron, and support vector machines are used for classification.

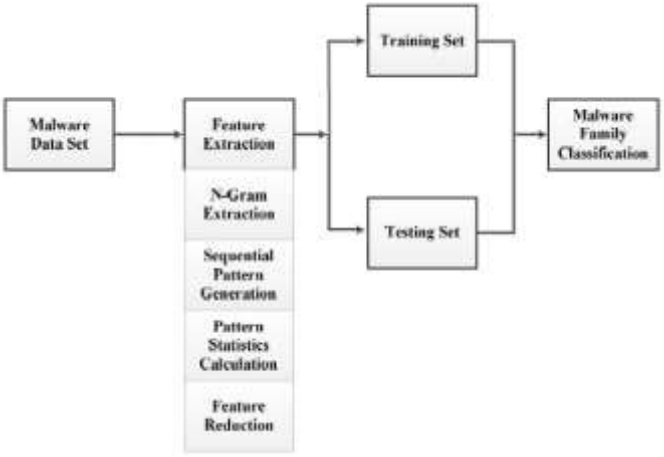

Figure1: Malware classification method

Markov Assumption

Some words are more likely to follow a word in certain contexts. It would be correct to know all the words up to the word that we are trying to forecast, but it would be inefficient to know complete history, because infinitely many sequences of sentences can be found and the history we know would have never occurred before. Therefore, approximation of the history is done by only a few words. Bigram, also called Markov assumption, assumes that we can foresee the probability of the future word by only looking at the last word encountered. Generalization of bigram to trigram can be made by looking last two words in the past, and to N-gram by looking N-1 words in the past [21-22].

The simplest way to estimate the probabilistic is to use Maximum Likelihood Estimation (MLE), based on taking counts from the corpus and normalizing them to lie in the interval [0, 1]. For example to calculate the bigram probability of word y following x is to count the bigrams c(x y) from the corpus and normalize it with the number of bigrams that starts with x.

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

$$P(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n)}{C(w_{n-1})}$$

In the denominator, $C(w_{n-1})$ represents the count of bigrams starting with w_{n-1} because the bigrams starting with w_{n-1} is equal to the number that w_{n-1} occurs in dataset. The general equation for estimating probability for a MLE N-gram is:-

$$P(w_n|w_{n-N+1}^{n-1}) = \frac{C(w_{n-N+1}^{n-1}w_n)}{C(w_{n-N+1}^{n-1})}$$

Hidden markov model are widely used for protein sequence analysis, speech recognition, software piracy detection and malware detection.

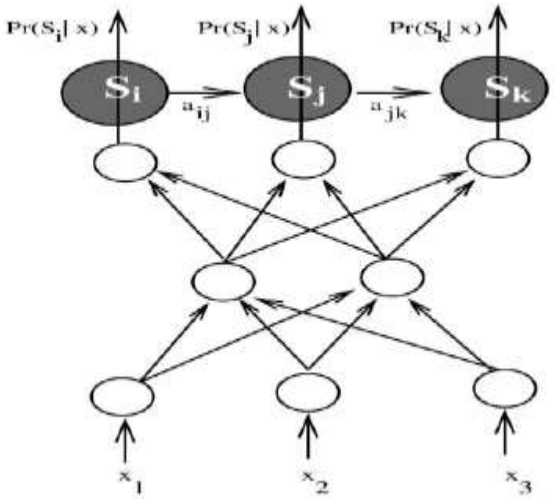

Figure 2: Basic Hybrid Architecture where a Two Layer Feed forward ANN estimates the posterior probabilities of states Si, Sj, Sk, of a left to right HMM given an hypothetic acoustic observation

Pairwise Alignment

A sequence alignment is widely used for arranging various sequences like proteins, RNA or DNA to calculate similarity index that may be due to following process between the sequences.

- Functional
- Structural
- Evolutionary

Aligned sequences of proteins, amino acid or nucleotide residues are typically represented as rows within a matrix. Gaps are inserted between the residues so that identical or similar characters are aligned in successive columns. The Next Generation Virus Creation Kit is a virus generator kit which can automatically generate variations of a virus. The NGVCK generator involves assembly morphing engine of source which includes-

- Random function ordering
- Junk code insertion
- User-defined encryption methods

This means that all variants of the viruses generated by NGVCK will have the similar functionality, but they appearance will be different from one another [9] [10]. A similarity score was introduced which shows NGVCK viruses have significantly more variation than many other commonly known virus generators. The NGVCK virus generator can be considered a metamorphic virus, and because of its significant variation it will be considered the complex case which should sufficiently test the theory that a meaningful alignment can be created for a highly

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

metamorphic virus [11-20]. Tests on the viruses generated by NGVCK will provide valuable information since the virus demonstrates all 4 of the mutational processes [12]-

- Substitutions – a subsequence in the original is substituted for a new subsequence
- Insertions – the subsequence was inserted into the original
- Deletions – the subsequence was removed from the original
- Permutation – a random re-ordering of the original sequence

Description Of Pairwise Alignment Algorithm

Definitions

s1 = first sequence
s2= second sequence
|L|=length of sequence a
a_i *=indicates the* i^{th} *symbol of sequence a*
$a_{i\ldots j}$ *= subsequence of a with indices i to j*
where a= $a_{i\ldots|L|}$
scor(a, b) = score assigned to substituting symbols a with b
gap(n) = cost of adding one gap to sequence with n-1 gaps
*F and G = matrix of size |s1|+1 * |s2|+1 (indices will be 0 based)*
F(i, j) =optimal score for aligning $s1_{1\ldots i}$ *with* $s2_{1\ldots j}$
G (i, j) = number of subsequent gaps used to generate F(i, j).

Recursive definition of F and G for i, j≥0

G(i,0)=F(i,0)=0
G(0,j)=j
F(0,j)= $\sum_{n=1}^{j} gap(n)$ *(the cost of aligning j gaps)*
F(i, j)=max((F(i-1, j-1)+scor(s1i, s2i),F(i-1, j)-gap(G(i-1,j))),F(i, j-1)-gap(G(i,j-1))

if case1: G(i, j)=0
if case2: G(i, j)=G(i-1,j)+1
if case3: G(i, j)=G(i, j-1)+1

Pseudo code

Initialize the first row in F and G: G (0, j)= j and F(0,j)= $\sum_{n=1}^{j} gap(n)$
For each row i , 1….|s1|
Initialize F(i, 0)=0 and G(i,0)=0
For each column j, 1….|s2|
(i-1, j-1), (i-1,j) and (i, j-1) for F and G are all known.
Calculate F(i, j) and G(i, j) using the recursive definition

3-D Signatures

The analysis of metamorphic virus dataset in the form of 3-D signatures proved itself revolutionary in the field of computer virology as experimented by us. This scheme is not only new in the sense of new viral detection techniques but it reveals the hidden relationship between the Computer viruses and biological diseases. The reduction of problem from computer domain to biological domain and to revert back in same domain with some fruitful results brings the sense of special kind of bridge between patterns. 3-d signatures can be designed from opcode sequences. For this opcode to protein mapping is required in conceptual domain.

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

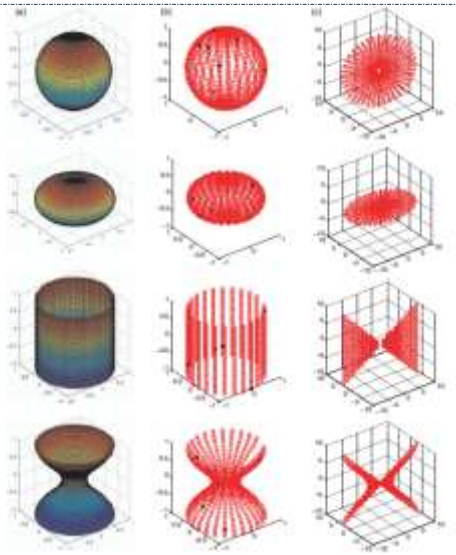

Figure3: Random collection of points in 3-D space with known geometrical origins.

Opcode Frequency Analysis

The analysis of opcodes has been used to study the malicious codes. The distribution of opcodes depicts the classifying feature that can be used for the purpose of detection. Following figures shows the average opcodes distribution in executables and opcodes frequency analysis in different malware family and normal files.

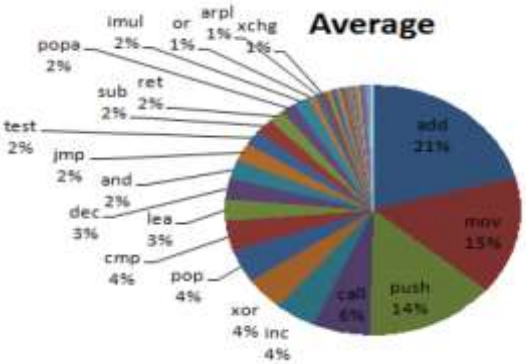

Figure4: Average opcodes distribution in executables

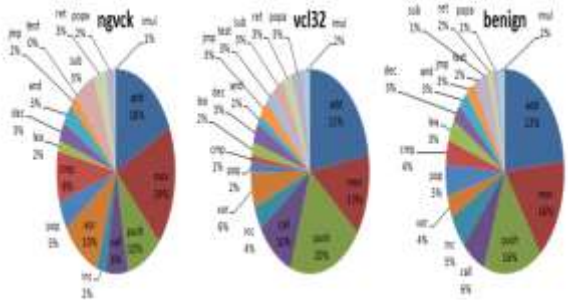

Figure5: Opcode frequency analysis in different malware family and normal files

Singular Value Decomposition

Singular value decomposition can be defined in terms of matrix factorization. The factorization form of matrix of order m*n is:-

$$M = USV^t$$

U= Left singular vectors of M

V= Right singular vectors of M that contains Eigen vector of matrix $M = M^tM$

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

S= Diagonal Matrix
M= Square Covariance Matrix

LAYOUT OF PROPOSED WORK

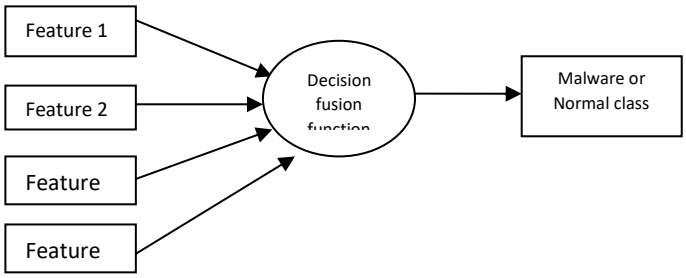

Figure6: Classification using feature Fusion

Proposed approach based on similarity analysis with static scanning of code. This improvised approach can be performed in the particular direction of metamorphic viruses' detection. For identification of metamorphic virus in effective way detailed analysis has been done. Our contribution involves the addition to similarity detection methods such that rate of false positive and false negative could be reduced. To do this five cases are considered. Up to what extent our theoretical approach works is checked through experimental analysis. It involves the threshold criteria in restricted domain to calculate the required results in appropriate manner. Five cases taken for analysis are as follows-

- Fusion of Pairwise alignment and 3-D biological signature
- Fusion of HMM with 3-D biological signature
- Fusion of opcode frequency with 3-D biological signature
- Fusion of SVD with 3-D biological signature
- Fusion of n-gram with 3-D biological signature

Experimental result shows that these hybrid techniques will give rise to better detection accuracy. These approaches are examined using the metamorphic viruses generated through NGVCK kit and then IDA-Pro is used for disassembly. The analysis is done by algorithms and the appropriate results obtained by sequence analysis of opcodes.

CONCLUSIONS

The approach presented in this paper is based on feature fusion analysis of computer viruses. Our method covers static analysis approach which is based on the fundamental theme of our analysis that is the pattern observation approach of opcodes. The advancement in this phase is left open for future work which includes the addition of functionalities regarding detection of semantic based pattern.

It is sure that the certain updation will lead to revolutionary results that will make further improvement in metamorphic viruses' detection. To make the analysis of metamorphic viruses NGVCK kit is used because it generate viruses have distinct pattern thus have higher tendency to hide them. It is typical to detect them as compare to viruses generated from other kits like G2.

REFERENCES

[1] J. Aycock, Computer Viruses and Malware, Vol 22, New York, NY, USA: Springer, pp. 5-32. 2006.
[2] H. Bidgoli, Handbook of information security, Wiley. 2006.
[3] F. Cohen, Computer Viruses. PhD thesis, University of Southern California. 1986.
[4] Donabelle, Richard M. Low, and Mark Stamp. "Structural entropy and metamorphic malware." *Journal of Computer Virology and Hacking Techniques*: 1-14. 2012.
[5] R. Durbin, Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. 1998.
[6] HiddenMarkovModels: http://en.wikipedia.org/wiki/Hidden_Markov_model, Last visited 20 Jan 2013.
[7] IDA Pro, http://www.hex-rays.com/idapro/., Last visited 20 Jan 2013.
[8] M. Jordan, Dealing with Metamorphism, Virus Bulletin, 2(1) pp. 4-6. 2002.

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

[9] A. Lakhotia, and M. Moinuddin, Imposing order on program statements to assist anti-virus scanners. In WCRE '04: Proceedings of the 11th Working Conference on Reverse Engineering (WCRE'04), Washington, DC, USA. IEEE Computer Society. 2004.
[10] P. Mishra, Taxonomy of software uniqueness transformations. Masters Thesis, of Computer Science, San Jose State University. 2003.
[11] Bist, Ankur Singh, and Sunita Jalal. "Identification of metamorphic viruses."*Advance Computing Conference (IACC), 2014 IEEE International*. IEEE, 2014.
[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.
[14] Bist, Ankur Singh. "Fuzzy Logic for Computer Virus Detection." *IJESRT, ISSN*: 2277-9655.
[15] Bist, Ankur Singh. "Hybrid model for Computer Viruses: an Approach towards Ideal Behavior." *International Journal of Computer Applications* 45 (2012).
[16] Sharma, Rudranshu, et al. "Genetic Algorithm based Weighted Extreme Learning Machine for binary Imbalance Learning."
[17] Sharma, Rudranshu, and Ankur Singh Bist. "INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY MACHINE LEARNING: A SURVEY.
[18] S. McGhee, Pairwise alignment of metamorphic viruses, Department of Computer Science, San Jose State University: pp. 12- 51. 2007.
[19] J. Maries, Borello and L. Me, Code Obfuscation Techniques for Metamorphic Viruses. 2008.
[20] VX Heavens. http://vx.netlux.org/ , Last visited 20 Jan 2013.
[21] W. Wong, and M. Stamp, Hunting for metamorphic engines. Journal in Computer Virology, 2(3):211-229. 2006.
[22] Jurafsky, D. and Martin, J. (2006). An introduction to speech recognition, computational linguistics and natural language processing, Draft.