# Detection of Malware by using Sequence Alignment Strategy and Data Mining Techniques

Vivek Kumar[1], Sadhna K Mishra, PhD.[2]
M Tech Scholar[1], Professor[2]
LNCT Bhopal[1, 2]

## ABSTRACT

Malware is basically malicious software or programs which are a major challenge or major threats. for the computer and different computer applications in the field of IT and cyber security. Traditional anti-viral packages and their upgrades are typically released only after the malware's key characteristics have been identified through infection. But by this time it may be too late to protect systems. Multiple sequence analysis is widely used in bioinformatics for helpful the genetic multiplicity of organisms and annotating gene functions through the identification of common genetic regions. This paper adopts a new approach to the problem of malware recognition, which is to use multiple sequence alignment techniques from bioinformatics to align variable length computer viral and worm code so that core, invariant regions of the code occupy fixed positions in the alignment patterns. Data mining (ANNs, symbolic rule extraction) can then be used to learn the critical features that help to determine into which class the aligned patterns fall. Experimental results demonstrate the feasibility of our novel approach for identifying malware code through multiple sequence alignment followed by analysis by ANNs and symbolic rule extraction methods.

**Keywords**--Malware; sequence alignment; viral signatures, HEX editor, ASCII code, ARFF, Dev C++, Java eclipse**.**

## 1. INTRODUCTION

Bioinformatics [1] is a branch of biology with computer scientists which deals with study of methods for storing, accessing or retrieval as well as analyzing the biological data, such as nucleic acid DNA, RNA and protein sequence, their structure and their function or behavior. The major research areas of Bioinformatics are sequence analysis, evolutionary biology,, analysis of gene expression, analysis of proteins , genome annotation etc. This paper is based on sequence analysis. In this analysis, sequences of DNA [2] proteins [3] and amino acid [4] has been taken. Sequence analysis is applied on these types of genetic information (DNA proteins and amino acid) in order to identify and detect the location of specific genes. For that we will try to apply better sequence alignment based algorithms in order to determine the relationship between these genes. an alignment is an adjustment of a sequence in relation to other sequences. One of the major advantages of sequence alignment based algorithms is in the conversion of variable length biological sequences can be converted into fixed length sequences through appropriate insertion and deletion techniques. .Then some Powerful data mining algorithms that assume fixed length sequences or patterns can then be applied to identify critical features that help to determine whether a sequence is malware or not. The aim of our paper is to precede our research in the determination whether the particular framed region is malware signatures and removal of malware or not by using the techniques which were used in the detection of the biological sequences.

So far we have several methods chosen for sequence analysis. Such as pair-wise alignment, multiple alignments, and standard ANNs (artificial neural networks) are used for classification and learning. Pair-wise alignment and multiple sequence alignment are used for possibly variable length sequences of DNA or proteins. A part from this Needleman-Wunsch [5] technique has proposed a technique for Global alignment which tries to align every item in every sequence. This technique is the best suited for roughly similar length. Similarly Smith-Waterman technique [6] proposed a new technique for Local alignment which is suited for dissimilar sequences. Alignment techniques can be powerful and are core to bioinformatics; they can also be complex [7] and intractable [8].

Data mining [9] [10] is the process of posing queries and extracting different patterns, often previously unknown from large quantities of data using pattern matching or other reasoning techniques. Cyber security is involved with protecting the computer and network systems against corruption due to Trojan horses, worms and viruses. By using data mining techniques variable length sequences can be converted into fixed length by inserting or deleting the gaps. However domination of fixed length alignment on the field of Data mining algorithm is new concept because not much has been reported till now. Hidden Markov Models (HMMs)[11] has suggested a probabilistic approach in this context.

### MALICIOUS SOFTWARE'S

Malicious softwares are basically programs which are written in programming languages. There are two types of malicious softwares.

1. Dependent Malicious softwares

2. Independent Malicious softwares.

They are further classified into various forms. These softwares intends to various threats such as network attacks [12] cyber attacks and web attacks [13]. Malwares or malicious softwares may be classified into various forms. Out of them we will discuss mainly about virus & worms.

### *A. Virus:*

A computer virus [14] is usually constructed for mainly two reasons,. The first reason is to replicate itself . For example, a virus may copy itself into a useful program. A virus may invade system files and replicate itself. Secondly on the other hand a virus has a specific function (or functions) defined by the virus writer. The second objective could include displaying a message, erasing sectors of the hard disk, or expanding until it slows down other processes in the computer.

### B. *Worms:*

Worms [15] are a prime example of a malware attack. Computers are typically protected by a combination of host-based and network-based

Defenses. Self replication basics worms actively select and attack their targets through a network automatically. The capability for self replication is enabled by certain functions in the worm code (Skoudis, 2004). First, a function for target location chooses the next host for attack.

A worm that is malicious may be considered a network extension of a virus that uses a network communication mechanism for propagation.

## 2. RELATED WORK

Malware are affecting adversely in different areas of the computer fields. These areas may be related from small PCs to computer networks. Viral developers develop these viruses and worms which are injected into even mobiles systems. Keeping all these into consideration several works has been done on various areas which are described as follows:

Lin Chen & Bo Liu [16] analyze the development of virtual based model VMM-based models of malware detection. Fahad Bin Muhaya, Muhammad Khurram Khan and Yang Xiang [17] Mahinthan Chandramohan and Hee Beng Kuan Tan [19] have worked on the detection of the malware which are being transmitted in the small devices such as mobiles. He presented a survey regarding techniques for detecting mobile malware and provides some suggestions to protect the smart phones from potential security threats. He also discussed about the strengths and weaknesses of these techniques where applicable.

Te-En Wei, Ching-Hao Mao, and Albert B. Jeng [20] have proposed their work on Android Malware Detection. They proposed network spatial features of Android applications and used independent component analysis (ICA) to determine the intrinsic Android malware domain name resolution communication behavior.

Many more research has been has been worked out but most of them are based on viral signature patterns. Our research paper is based on anomaly technique which is however a probabilistic modeling approaches yet it will be very beneficial if any how we can detect the upcoming viral dataset.

## 3. PROPOSED APPROACH FOR DETECTION OF MALICIOUS SOFTWARE

This research paper is based on the detection of virus and worms by using a series of following steps:

A. At first I have downloaded viral dataset from various sites.

B. Then these dataset are converted into its equivalent hexadecimal code by using HEX editor.

C. After that these hexadecimal code are converted into ASCII code.

D. Now these ASCII codes are aligned by some sequence alignment tools. Here we have taken ARFF and enhanced data mining algorithm.

After the alignment of the particular sequences they are verify by applying various techniques so as to identify whether the particular sequence is malware or not. The basic difference between these types of techniques and the traditional anti- virus softwares (AVS) is that AVS works on the concept of the viral signature databases, which means they can detect only those viruses whose signatures are already present in the databases. These softwares are not able to detect the new incoming virus because their signature is not present in the databases. The approach of our paper is based on the anomaly approach in which we try to analysis whether the incoming patterns is consists of malware or not.

## 4. DIFFERENT SOFTWARES USED IN PROPOSED WORK FOR MALWARE DETECTION

As it is known that for detection of malware especially virus and worms include a series of steps and various types of tools, techniques and algorithms. Some of them are as follows which are used in the implementation of detection in this research paper.

### *A. HEX editor*

HEX editor is used to convert viral dataset into its equivalent HEX code. By doing this we will protect our system to get infected from the viral or worms datasets.

### *B. HEX to ASCII Converter*

ASCII stands for American Standard Code for Information Interchange. As we know that Computers can only understand numbers, therefore an ASCII code is the numerical representation of a character. We therefore we need to convert HEX into its equivalent ASCII code which can be understandable by computer.

### *C. ARFF FILE*

ARFF file is also known as attribute-relation file format which stores ASCII data in such a way that the data are stored in the format of rows and columns. By default each row contains 32 characters of ASCII data. This length of character may be changed according to the attribute that can be taken. An Arff file is divided up into two parts first header part and second is data part. The data part consists of data that may be individually identified by the comma as the separation.

### *D. Java eclipse programming language*

Java eclipse programming language is used in this research as ANN simulator which is similar to Java NNS or other GUI based softwares. We can train and test the dataset by using several data mining algorithms. These algorithms are used for the classification of the virus and worms and try to find out that up to what extent this tool is able to classify the the input data correctly.

## 5. IMPLEMENTATION AND RESULT

After the final alignment we have used some improved data mining algorithms in order to detect whether the incoming pattern consists malware or not. First one is to trained the dataset and by using data mining algorithms and then to test that dataset by using enhanced version of the respective algorithms. the feature of these types of techniques is cross validation. in the cross validation we usually determine that what training we have given on the dataset in order to distinguish between computer virus & worms ,what is the percentage of the training the dataset & how efficiently the respective algorithms could able test the test dataset. This will be calculated on the percentage basis.

Later on the main feature of the cross validation is to construct the huge amount of information as a database. This or that we have taken a group of virus and worms and they are being converted into its hex code. These hex codes are transforming into its ASCII code as arff file which is the input for Weka tools.

We have taken near about 323 combined instances of virus and worms which have been stored in the relation named "vv" and apply different data mining algorithms and got following results:

I have taken enhanced version of IBK as a data mining algorithm and tried to find out that IBK algorithms gives approximately 96% result. That means IBK is able to classify 96% correctly or out of 100 different pattern enhanced IBK algorithm is able differentiate 96 patterns between patterns of computer virus & computer worms correctly. Also this algorithm is able to test 85% correctly.

The results are as follows:

Scheme: IBK

Test mode: evaluate on training data

Time taken to test model on training data: 0.2 seconds

Summary

| | | |
|---|---|---|
| Correctly Classified Instance | 310 | 95.9752 % |
| Incorrectly Classified Instances | 13 | 4.0248 % |
| Kappa statistic | | 0.9004 |
| Mean absolute error | | 0.0484 |
| Root mean squared error | | 0.151 |
| Relative absolute error | | 11.5049 % |
| Root relative squared error | | 32.9516 % |
| Coverage of cases (0.95 level) | | 100% |
| Mean rel. region size (0.95 level) | | 54.644 % |
| Total Number of Instances | | 323 |

| TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.134 | 0.946 | 1 | 0.972 | 0.905 | 0.995 | 0.996 | VIRUS |
| 0.866 | 0 | 1 | 0.866 | 0.928 | 0.905 | 0.995 | 0.98 | WORMS |
| 0.96 | 0.094 | 0.962 | 0.960 | 0.959 | 0.905 | 0.995 | 0.991 | Average |

Time taken to test model on training data: 1.19 seconds

Summary

| | | |
|---|---|---|
| Correctly Classified Instances | 280 | 86.6873 % |
| Incorrectly Classified Instances | 43 | 13.3127 % |
| Kappa statistic | | 0.6514 |
| Mean absolute error | | 0.133 |
| Root mean squared error | | 0.3645 |
| Relative absolute error | | 31.6064 % |
| Root relative squared error | | 79.5073 % |
| Coverage of cases (0.95 level) | | 86.6873 % |
| Mean rel. region size (0.95 level) | | 50 % |
| Total Number of Instances | | 323 |

Detailed Accuracy by Class

| TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|---|---|---|---|---|---|---|---|---|
| 0.973 | 0.381 | 0.856 | 0.973 | 0.911 | 0.673 | 0.883 | 0.915 | VIRUS |
| 0.619 | 0.027 | 0.909 | 0.619 | 0.736 | 0.673 | 0.940 | 0.856 | WORMS |
| 0.867 | 0.275 | 0.872 | 0.867 | 0.858 | 0.673 | 0.9 | 0.897 | Average |

## 6. CONCLUSION AND FUTURE WORK

We can conclude our research work that although many good & efficient anti- virus softwares are available in the market but they are working on signature based. That means they are able to detect only those malwares whose characteristics is known to the AVS. The motive of my research is basically concentrated on the two facts.

A. We able to test the new upcoming pattern on the basis of our trained dataset in the databases.

B. Upto large extent we have tried to demolished the confusion between computer virus & computer worms which was not able to distinguished between them by using some of the traditional AVS.

In our research work we have succeeded trained our dataset upto 96%. Also we are able to test the new upcoming dataset upto 85%. Furthermore, as we know that anomaly technique is a probabilistic approach therefore still a lot of work has to be done in this context. In future keeping all these things in our mind we will try to take a large amount of new dataset as a database in which it may consists of malicious & non malicious softwares and try to apply different data miniing enhanced algorithms in order to classify between malicious & non malicious softwares. If the pattern falls in malware group, they must be classified as computer virus & computer worms. Also it may be applicable in the classification of spyware, Trojans, Backdoors etc.

## 7. REFERENCES

[1] Vivek kumar, Dr. Sadhna K Mishra, Prof. Vineet Ricchariya “Detection of malicious software by Using Data Mining Tools and Other Techniques- a Survey”, IJCSMR volume (1) issue 4, 2012, pp-746-750.

[2] Belleville, Callicut et, al. “Active CMOS biochips: an electro-addressed DNA probe” IEEE conference 1998 pp-272-273.

[3] Cuff, J, Barton, G., (1999) “Evaluation and Improvement of Multiple Sequence Methods for Protein Secondary Structure Prediction”, *Proteins: Struct. Funct. Genet.* **34**, 508-519.

[4] Fooks, H. M., Martin, A., woolfson, D., Sessions, R., Hutchinson, E. (2006) Amino Acid Pairing Preferences in Parallel $\beta$-Sheets in Proteins, *J. Mol. Biol.*, **356**, 32-44.

[5] S.B.Needleman and C.D. Wunsch, “A general method applicable to the search for similarities in the amino acid sequence of two proteins", Journal of Molecular Biology 48 (3), 1970, pp. 443–53.

[6] T.F. Smith and M.S. Waterman, "Identification of Common Molecular Subsequences", Journal of Molecular Biology 147, 1981, pp. 195–197.

[7] L.Wang and T. Jiang T. “On the complexity of multiple sequence alignment”, J Comput Biol, 1 (4), 1994, pp 337–48.

[8] I. Elias, "Settling the intractability of multiple alignment", J Comput Biol 13 (7), 1996, pp. 1323–1339.

[9] Data Mining for Malicious Code Detection and Security Applications

[10] Data Mining: “Concepts and Techniques Jiawei Han and Micheline Kamber”, Morgan Kaufmann, 2001.

[11] S. McGhee, “Pairwise Alignment of Metamorphic Computer Viruses”. Masters Project Paper 37 2007. Faculty of the Department of Computer Science San Jose State University. http://scholarworks.sjsu.edu/etd_projects/37

[12] Symantec Internet Security Threat Report: Trends for 2010. http://www.symantec.com/business/threatreport/index.jsp

[14] Marshall D. Abrams and Harold J. Podell “Malicious Software.pdf” pp 116-120.

[15] Thomas M. Chen and Gregg W. Tally” Malicious Software. pdf”.

[16] Lin Chen and Bo Liu “A layered malware detection model using VMM” IEEE Journal 2011,pp-232-236.

[17] Fahad Bin Muhaya, Muhammad Khurram Khan and Yang Xiang” Polymorphic Malware Detection Using Hierarchical Hidden Markov Model” IEEE, 2011,pp-151-155.