PROPOSING AN HMM-BASED APPROACH TO DETECT METAMORPHIC MALWARE

Abdumuminov Abdurafiq Abdurashidovich
Republican center for management of telecommunications networks of Uzbekistan. SUE.

Ibragimov Jalaliddin Obidjon o'g'li
Teacher of the Department, "Systematic and Practical Programming", Tashkent University of Information Technologies named after Muhammad Al-Khwarizmi, UZBEKISTAN

Shoraimov Khusanboy Uktamboyevich
Teacher of the Department, "Systematic and Practical Programming", Tashkent University of Information Technologies named after Muhammad Al-Khwarizmi, UZBEKISTAN

ABSTRACT

Previous research has shown that hidden Markov model (HMM) is a compelling option for malware identification. However, some advanced metamorphic malware have proven to be more challenging to detect with these techniques. In this paper, we separated the importance of the some part of the malware files to train the HMMs aiming at extracting the significant sequences of malware opcodes. These parts have been deemed important according to their dissimilarity to the benign files, as all parts of a malware file are not representative of the malicious nature. Extracting these parts has been performed using the methods similar to sound processing. The results demonstrate that the proposed method has the higher accuracy to the metamorphic malware detection and also has the higher speed at classification, compared to the previous methods.

Keywords—malware detection; metamorphic malware; hidden Markov model

I. INTRODUCTION

Today, malware are considered the serious threat for the personal security of data and computer's systems by creating a large-scale failure. With regard to the growth of the malware and also the various and complex techniques are being used by malware designers, such as obfuscation in which the malware are different in structure, however they have the same functionality, recognizing the malware is significant prior to take a wide action. The malware is the shortened form of malicious software and includes Viruses, Worms, Trojans, Adwares, Spywares, etc. There are two major trends for malware detection: traditional signature-based methods and behavior-based methods.

The malware's signature which is utilized by the signature- based method to detect the malware is something like fingerprint. In other words, it is a unique characteristic and is the sequence of bytes comprising the malware in technique which should be found in the same malware to maintain its monopoly. The malware's authors sought for the ways to get around this method. Having considered the intention, polymorphic malware were created. They didn't have the stable codes, as they encrypt their code by one algorithm per pollution; they decrypt the same code during the runtime. The next generation of malware will not even have the fixed encoding and decoding engine and they will generally transform their code per spread, so that it is very difficult to recognize them through the traditional signature-based methods and calls for creating the complex signatures; sometimes the recognition is impossible. These types of malware are called Metamorphic.

The behavior-based methods are also able to recognize the malware use the obfuscation techniques, analyzing the programs' behavior. The most methods proposed to achieve the objective uses the modeling and data mining to recognize the malware; that is a set of features are being extracted from the malware files and make effort to learn the malware's behavior using the machine learning algorithms, then a model of destructive behaviors is created by which malware are being separated from benign programs. One of the effective ways in this field is to use the HMMs to the metamorphic malware detection. The research indicates that the method has been led to the suitable results.

This paper is organized as follows. In Section 2, we give an overview of the most important previous works based on HMMs. Section 3 discusses our proposed method in more detail. In Sect. 4, we present the experimental results for our proposed method based on HMMs. Finally, Section 5 gives our conclusions.

II. RELATED WORK

In this section, it has been reviewed the works to detect metamorphic malware based on HMMs. The HMMs was introduced in the late 1960s and now it is expanding the range of its applications rapidly. The HMMs can have various applications in the field of modeling and learning and they are most well-known because of pattern recognition such as recognizing the voice and handwriting, recognizing the points and movement, labeling the speech and bioinformatics.

Several HMMs are being trained based on phonemes in the application of speech as a solution and the input signal similarity is computed by the trained phonemes. In the solution, the input signal is divided into fixed-size overlapped frames and each frame is analyzed through the HMMs and is classified as a phoneme [9]. The solution can be applied by some transformation to recognize the metamorphic destructive software. It is challenging to classify the malware automatically. Annachhatre et al used the HMMs and cluster analysis to solve the challenge. They evaluated the HMMs according to scoring technique and they clustered by K-means algorithm and achieved the acceptable results.

The authors of metamorphic viruses endeavour to frustrate the methods based on the HMMs, using the dead benign codes. Vinod et al presented a new approach to recognize the unseen malware and benign samples, using the discriminating linear analyzing to rank and produce the most prominent features of opcode which can develop the detection rate compared to the general scanners.

Wong et al utilized the threshold approach to manifest the malware, using the HMMs successfully. They demonstrated that the approach presents a practical solution for some metamorphic viruses which can't be recognized by the signature-based method. In this method, the opcode sequence has been considered unique in destructive software and has been learned by the HMMs. Then it is examined and determined the probability of observing the opcodes sequences through its similarity with the sequences learned by this model for every new file. It is recognized as a virus, if the probability is less than the determined threshold, otherwise it is classified as a benign file. The most important benefit of this method is that the analysis is performed at a high speed, as it uses just one HMMs. However some of the metamorphic malware such as MWOR are not able to escape from the approach.

Kalbhor et al used the dual HMMs as a tool to recognize the metamorphic viruses and they could recognize them with a high precision. However, the approach is very overwhelming and the identification time in the approach is about twice the threshold approach, as it creates a separated model for each virus family and every family of benign files rather than using the HMMs. Then it computes the probability of observing the each file opcodes sequences for each model and file belongs to the family that the model corresponding to it has the highest probability.

III. PROPOSED METHOD

In the recent years, different ways have been developed used an executable files to recognize the malware. Although each part of an executable file contains the important information on the file, all the information doesn't facilitate the detection of destructive behavior. The opcodes sequence extracted from an executable file explains the same file behavior and can be shown as a set of simpler tasks through a few opcode. The malware files include the various opcode commands, some of which facilitate the detection and the rest help to interference. As a result, if the commands could be separated from each other and the HMM is trained based on the important commands of opcodes, it will lead to better achievements. However, the matter is that we are not aware of the important sequence of opcodes in the malware.

The proposed method uses the less important sequence of opcodes in malware to extract the important sequences. It can be done based on the fact that the less important sequence of opcodes is the one which has more similarity with the benign files. Then we could achieve the important sequences by recognizing and separating the less important sequences. To do so, we divide each malware file into several parts or frames and evaluate the similarity of each frame to the benign files and we remove each frame with the more similarity. Thus, it is remained the only parts of malware including the important sequences. Fig. 1 shows the opcodes sequence extracted from an executable file.

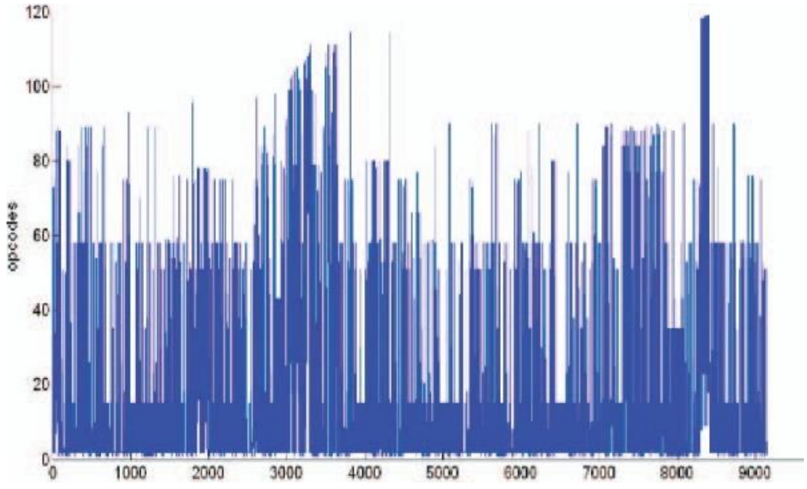

Fig. 1.sequence of opcodes

The presented method requires explaining three fixed coefficients: the size of frame, importance threshold and classification threshold. The frame size is the opcodes sequence length extracted from the destructive software. Thus, if the frame size is considered three, each extracted frame includes just the three opcodes. In the method, we use the sliding window technique to extract the frames . The importance threshold is the cutting point by which the unique and important part of malware is removed from the less unique part. If the importance threshold is zero which is very large, malware code is extracted completely and if it is selected very small, just the small part of unique sequences of opcode is extracted. The classification threshold is the amounts by which the files exist in the test set are classified as a malware or benign file. It has been summarized the training and classification process of the proposed method as follows:

1. We train one HMM based on the benign files.
2. We determine a frame size and extract the frames from the malware files existing in the training set through the sliding window method. Then we determine the similarity of each malware frame with the trained model.
3. We remove the frames with more similarity to the benign files using the explained importance threshold. So the important sequence of opcodes is separated from the less important ones. Then we used the remained important sequence of opcodes to train the new HMM.
4. We classify the files of test set using the classification threshold and the new HMM which is only trained on special parts of malware and we separate the members of virus family from the nonmember.

IV. EXPERIMENTAL RESULTS

The proposed method has been examined and tested on the well-known malware MWOR and malware constructed by kits NGVCK, G2, MPCGEN which are downloadable the executable random benign files which have been selected from the Cygwin. In total, it has been used benign and malware file in the training and test sets out of 740 samples.

Effectiveness of the proposed approach is determined by detection rate, false-positive rate and overall accuracy. Detection rate equals to divide numbers of viruses known through the model by the total number of viruses in the test set. The false-positive rate corresponds to the model features and it is obtained through dividing the number of false positive by the total number of benign programs in the test set. The general precision is defined as the divide number of correct predictions by the total numbers of benign and non-benign programs. The three cases are computed based on true positive (TP), true negative (TN), false positive (FP) and false negatives (FN) as follows:

$$Detection\ Rate = \frac{TP}{TP + FN} \qquad (1)$$

$$False\ positive\ rate = \frac{FP}{FP + TN} \qquad (2)$$

$$Overall\ accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (3)$$

The proposed method is able to recognize the destructive files by the 92 precision and yet it remains the false-positive rate at minimum.

V. CONCLUSION

As it was stated, some of the metamorphic malware which are able to get around the threshold approach are recognized by dual approach with high precision. However, the approach is very overwhelming. In addition to taking the advantage of both approaches benefits, we could present a new method by detachment of malware files parts importance based on dissimilarity to the benign files and extracting the important sequences of opcodes from the malware files to overcome the challenge. The results demonstrated that the HMM trained based on the important sequences of opcode has been able to act with the higher speed and the most cases more accurate than the dual approach. We hope the proposed method allows us to improve the malware recognition tool based on training the HMM better.

REFERENCES

1. A. Kalbhor, T. H. Austin, E. Filiol and M. Stamp, "Dueling hidden Markov models for virus analysis," Journal in Computer Virology Hack Tech:Springer, 2018.
2. C. Annachhatre, T. H. Austin and M. Stamp, "Hidden Markov models for malware classification," Journal in Computer Virology Hack Tech:Springer, 2019.
3. Cygwin, Available: http://cygwin.com
4. D. Baysa, "Structural Entropy and Metamorphic Malware," M.S. dissertation, Dept. Comp. Sc., Univ. San Jose State, 2019.
5. J. Aycock, "Computer Viruses and Malware," Advances In Information Security:Springer, 2019.
6. J. Kuriakose and P. Vinod, "Ranked Linear Discriminant Analysis Features for Metamorphic Malware Detection," IEEE International Advanced Computing Conference, pp. 112-117, 2020.
7. K. Mathur and S. Hiranwal, "A Survey on Techniques in Detection and Analyzing Malware Executables," International Journal of Advanced Research in Computer Science and Software Engineering, pp. 422-428, 2018.