Detection of Malware using the DHP Algorithm and Logistic Regression Analysis

Yeongji Ju and Juhyun Shin*
Chosun University, 309 Pilmun-daero, Dong-gu, Gwangju, 61452, Korea
juyeongji71@gmail.com, jhshinkr@chosun.ac.kr

Abstract

The proliferation of computer networks has helped to further develop the software industry. However, this has been accompanied by an increase in the numbers of several types of malware. Therefore, research efforts have been directed towards detecting malware' actions and identifying certain execution files based on their Application Programming Interface (API) data. The majority of contemporary antivirus programs employ a signature detection technique; however, the number of signatures is very limited whereas the number of malware is increasing rapidly, which leads to a very high false detection rate. To address this issue, In this paper suggests a Malware analysis and detection method using an association rule mining algorithm and logistic regression analysis. By using the Direct Hashing and Pruning (DHP) algorithm, the API of the malware and the normal codes within a Portable Executable (PE) file are compiled as a hash table. Association pattern rules are probed to group the patterns. theassociation rule patterns extracted through this research reduced false detection rates when classificationwas carried out using the logistic regression analysis, and the discrimination result was shown to begreater than 0.7

Keywords: Detection of Malware, Association rule mining, DHP Algorithm, Logistic Regression Analysis

1 Introduction

Although the proliferation of computer networks has improved the growth of the software industry, malware that cause information leakage and system errors, such as viruses, worms, Trojan horses, and ad-ware, have also increased in number[10][9]. In the case of the Trojan horse, it has the appearance of a safe, legitimate program, but contains Malware that is executed when the program runs. Of all types of malware detected in February 2016, Trojan horse viruses accounted for the highest percentage at 44%[7]. To identify such malware activities, API call information is normally used to detect whether the code may be harmful, but there are several Malware being developed and these tend to be mutated quickly by malicious programmers. Existing antivirus programs analyze and detect the characteristics of Malware collected by a signature detection technique. However, new Malware are increasing rapidly while the known signatures are not capable of identifying new malware quickly. In addition, the false detection rate is also very high[8]. Hence, this paper suggests a malware analysis and detection method using an association rule mining algorithm and logistic regression analysis. From the PE file that is used in a device's Operating System (OS), Trojan-type Malware' APIs are extracted and preconditioned. The pattern rules of the malware files and the normal files are extracted through a hashing technique that efficiently creates frequent item sets and deletes the unnecessary item sets by applying the DHP algorithm to mine the association rule patterns. It is suggested that the false detection rate will be reduced

Research Briefs on Information & Communication Technology Evolution (ReBICTE), Vol. 3, Article No. 16 (November 15, 2017)
*Corresponding author: Department of ICT Convergence, Chosun University, 309 Pilmun-daero, Dong-gu, Gwangju, 61452, Korea, Tel: +82-062-230-7162

by grouping the pattern rule, using logistic regression analysis, and classifying Malware files and normal files based on a weighting scheme. This paper is organized as follows: In Section 2, we describe a study of classifying malware and, in Section 3, we describe our proposed method for detecting malware with logistic regression analysis. In Section 4, we present our conclusions and describe future studies.

2 Related Work

The following discussion describes the association rule mining algorithm for extracting the APIs' pattern rule and detecting any existing malware. The association rule mining algorithm used to discover the pattern basically uses a hash tree to count the frequent item sets[1]. Apriori, FP-Growth, and DHP are some of the representative association rule mining algorithms[2]. The DHP algorithm is an improved Apriori algorithm that uses the bottom-up approach. As it has a small number of candidates for frequent item sets created, it can reduce any bottlenecks over the entire process by efficiently creating the item sets and reducing the size of the transaction database by using a pruning technique[12]. Comparing the performance of the FP-Growth algorithm, which is known to be the best algorithm among the association rule mining algorithms, to that of the DHP algorithm, the FP-Growth algorithm requires a large amount of memory, whereas the DHP algorithm shows superior performance in memory usage and running time regardless of the size of the database[11]. Hence, the DHP algorithm is a good choice for detecting malware programs. C. Choi, et al. used adnormal behavior pattern mining for API attack detection. [3] Research on classifying malicious code is actively being conducted[6][4][5]. If there is an API that is the same or similar in both malware and normal code, most studies using API on malware classification define the common API as a White List item and the API is excluded from the comparison to reduce classification errors. However, defining Risk List and White List items by simply counting the frequency of an API makes it difficult to discover any relationships between characteristics once the classification is complete, leading to a low detection rate[13]. Therefore, the present paper assumes that there is a pattern related to the APIs called in normal files as well as in malware files and performs groupings by using an association rule mining technique. Finally, application of logistic regression analysis to the groupings reduces the false detection rate.

3 Detection of Malware

3.1 System Architecture

In this paper, we propose a method to extract association pattern rules using malware API through the DHP algorithm and to detect malware using the logistic regression analysis.

Figure 1 shows the overall system architecture that detects malware using the APIs of malware and normal files. The API of a trojan contained in the existing PE file and the API of normal code are extracted and a data preprocessing step is carried out. By using the DHP algorithm, which is an association rule mining technique and hashing technique, unnecessary item sets are deleted from the extracted API and the rules that represent the relations of the remaining item sets are probed. The rule relationships of API within the data extracted from malware and normal files are grouped and the malware is detected by using logistic regression analysis.

3.2 Clustering of API Pattern using DHP Algorithm

In this section, we describe the extraction of patterns using the DHP algorithm after carrying out a preprocessing step to clustering API patterns. The data used in this research is the API of Trojan-type

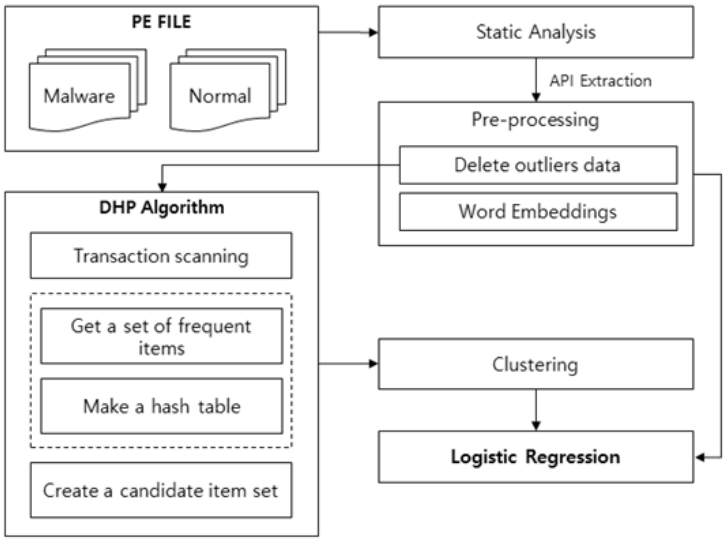

Figure 1: System Architecture for Detecting Malware

malware and normal files included in the PE file. The PE file refers to all executable files in the Windows OS and each API is extracted by static analysis. Approximately 300 APIs extracted from malware and normal files were extracted in the present study. To apply the association rule mining technique, abnormal data including special symbols were deleted and all characters were changed to lower case for each API. Moreover, to express API data as a single quantified data, the API data was converted to vector form through word embedding.

An association rule is understood to be a forecast based on the supposition that when a transaction supports X, it will also support Y with a certain probability. An association rule probe technique applies a level of confidence that expresses how often a rule can be applied and the degree of support between transactions X and Y that becomes the standard for pruning to extract a pattern. Therefore, the minimum degree of support and level of confidence is set in this research to extract a pattern. As a result, a data group excluding any unsatisfactory pattern rules is created. There is no clear method for determining the minimum degree of support. A high minimum degree of support can extract patterns with a high level of confidence but useful association patterns may be missed. Hence, a lower minimum degree of support based on the purpose of pattern extraction was chosen and applied in the present study. Tables 1 and 2 show the results of the associated rule patterns extracted by applying a minimum degree of support to APIs of malicious files and normal files, respectively.

pattern number	Association Rule pattern of Malware File API	Lift
1	[LoadLibraryA, GetProcAddress]	1.0
2	[Getcurrentprocess, getlasterror]	1.44
3	[Exitprocess, writefile]	1.34
⋮	⋮	⋮
22	[Getprocaddress, writefile, exitprocess]	1.36

Table 1: API pattern rule of Malware file with 60% minimum support

For the case of normal code, 70% of the degree of support where the most useful pattern was extracted

pattern number	Association Rule pattern of normal File API	Lift
1	[Sleep, GetCurrentThreadId]	1.20
2	[GetLasstError, Sleep]	1.15
3	[GetTickCount, GetLastError]	1.15
⋮	⋮	⋮
22	[GetCurrentThreadId, GetLastError, GetCurrentProcessId]	1.25

Table 2: API pattern rule of Normal file with 70% minimum support

was applied and 35 association rule patterns were extracted. As the malware has many mutated API, 60% of the degree of support was applied (slightly less than that for normal code) and 22 patterns were extracted.
Scaled Lift refers to the increased rate of the probability of an associated result over the probability of a pattern occurring alone. If the scaled lift is higher than 1, it means it is an inevitable relationship and not a random relationship. As the lifts in Tables 1 and 2 are all greater than 1, it is clear that a meaningful association rule pattern has been created. From the extracted associated rule patterns, a group is selected and plays a role in reducing the fault rate in malware detection, as discussed below.

3.3 Detection of Malware using Logistic Regression Analysis

This discussion proposes a method to analyze and identify malware by using the API extracted from malware and normal files and the grouped data obtained by the DHP algorithm using logistic regression analysis. The extracted associated rule patterns play a role in reducing false detection rates by assigning weights to them when being classified through logistic regression analysis. Logistic regression analysis forecasts the probability of occurrence by using a linear combination and is similar to linear regression analysis but is also used as a classification technique because its dependent variable is category type data. The logistic regression model assumes the parameter (a Bernoulli random variable) to be dependent on x, which is an independent variable. The sigmoid function is used to describe the parameter, because the parameter can only have a real value between 0 and 1, while x is represented as an integer. A sigmoid function refers to a function that has a limited value and positive slope of a finite interval for all real values (between 0 and 1 in this case). equation 1 indicates the sigmoid function used for the regression analysis.

$$logistic(z) = \frac{1}{1+e^{-z}} \tag{1}$$

The rate of probability of 1 over the probability of 0, $1-p$, is called the odds ratio, and Formula 1 expresses the reversed function of the logit function, which is the odds ratio after it has been log transformed to represent a sigmoid function. As $Logistic(z)$ is always greater than 0.5 when z is greater than 0 and less than 0.5 when z is less than 0, such characteristics are used in classifying malwares, where a value larger than 0.5 indicates malware is present.
The effectiveness of the malware detection performance was studied by evaluating the precision and recall factor using the mixed matrix method. equation 2 indicates the precision and recall.

$$precision = \frac{TP}{TP+FP} \qquad Recall = \frac{TP}{TP+FN} \tag{2}$$

Precision refers to how far the real value deviates from the forecast value when the latter is equal to 1. The recall factor refers to the deviation between the forecast value and the real value when the latter is equal to 1. If both the precision and the recall factor approach a value of 1, the classification

being evaluated tends to be valid. equation 2 was used to determine whether the forecast result using the association rule probe technique and logistic regression analysis proposed in the current paper is accurate. Table 3 shows the results of the classification method based on precision and recall factor for malware and normal code.

classification	Precision	Recall
Malware File	0.73	0.76
Normal File	0.75	0.72

Table 3: Result of discrimination using association rule patterns and logistic regression analysis

In total, 100 malware and normal files were used as a test set for the experiment. The APIs extracted from the normal files were higher in number than malware files in a single program and had slightly higher precision than malware files using the same base of informative pattern rules. Both malware and normal file classifications showed higher than 0.7 classification results, indicating a valid procedure was applied.

4 Conclusion

In this research, we detect malware using the DHP algorithm and logistic regression analysis based on the APIs extracted from malware and normal files. Malware and normal code APIs of a Trojan contained in the PE file were extracted and, after the data are preprocessed, association rule mining was applied. The association rule patterns of the malware and normal code APIs were extracted from the processed data with the DHP algorithm, and a weight was given to relevant association rule patterns; malware and normal files were discriminated through classification by logistic regression analysis. Accordingly, the association rule patterns extracted through this research reduced false detection rates when classification was carried out using the logistic regression analysis, and the discrimination result was shown to be greater than 0.7. In future research, a study on detecting malware by extracting characteristics from a malware behavior base and constructing a hybrid classification model will be conducted.

5 Acknowledgments

This research was supported by the Human Resource Training Program for Regional Innovation and Creativity through the Ministry of Education and National Research Foundation of Korea (2015H1C1A1 035823) and supported by the MSIP(Ministry of Science, ICT & Future Planning), Korea, under the ”Employment Contract based Master’s Degree Program for Information Security” supervised by the KISA(Korea Internet Security Agency)(H2101-16-1001)

References

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proc. of the 20th International Conference on Very Large Data-bases (VLDB’94), Santiago de Chile, Chile*, pages 478–499. Morgan Kaufmann Publishers Inc., September 2016.

[2] R. A. Bell. *Machine Learning: Hands-On for Developers and Technical Professionals*. John Wiley & Sons, 2014.

[3] C. Choi, J. Choi, and P. Kim. Abnormal behavior pattern mining for apt attack detection. *Computer Systems Science & Engineering*, 32(2), March 2017.

[4] J. Choi, C. Choi, B. Ko, and P. Kim. A method of ddos attack detection using http packet pattern and rule engine in cloud computing environment. *Soft Computing*, 18(9):1697–1703, September 2014.

[5] J. Choi, C. Choi, K. Yim, J. Kim, and P. Kim. Intelligent reconfigurable method of cloud computing resources for multimedia data delivery. *INFORMATICA*, 24(3):381–394, September 2013.

[6] C. Esposito and C. Choi. Signaling game based strategy for secure positioning in wireless sensor networks. *Pervasive and Mobile Computing*, 40:611–627, September 2017.

[7] INCA Internet. Detection statistics of malware in imca, February 2016. `http://erteam.nprotect.com/` [Online; Accessed on October 3, 2017].

[8] H.-J. Ji, J.-Y. Choi, S.-K. Kim, and B.-J. Min. Signature effectiveness description scheme. In *Proc. of General Spring Conference of Korea Multimedia Society*, volume 13, pages 41–43, 2016.

[9] S. Jo. Evolution of malicious code for corresponding technology and standardization trend. *Journal of TTA*, 118:47–57, 2008.

[10] H. Kim, J. Park, and Y. Won. A study on the malware realtime analysis systems using the finite automata. *Journal of the Korea Society of Computer and Information*, 18(5):69–76, 2013.

[11] H. Lee and J. Kim. Performance evaluation of the fp-tree and the dhp algorithms for association rule mining. *Journal of KIISE:Database*, 35(3):199–207, 2008.

[12] J. Park, M. Chen, and P. Yu. An effective hash-based algorithm for mining association rules. In *Proc. of the 1995 ACM SIGMOD International Conference on Management of Data (SIGMOD'95), San Jose, California, USA*, pages 175–186. ACM, May 1995.

[13] J. Park, S. Moon, G. son, I. Kim, K. Han, E. Im, and I. Kim. An automatic malware classification system using string list and api. *Journal of Security Engineering*, 8(5):611–626, 2011.

Author Biography

Yeongji Ju received the B.S. degrees in Control and Measuring Robot Engineering from Chosun University in 2016, Currently she is taking a master's course at Graduate School of Software Convergence, Chosun University. Her research interests include Data Mining, Big-data processing and Security.

Juhyun Shin received the Ph.D. degree in Computer Engineering from Chosun University, Gwangju Korea in 2007. From 1986 through 2011 she worked at TRUTEC CO. LTD. as Technical Director. Currently She is a professor in department of ICT convergence at Chosun University, Gwangju, Korea. Her research interests are in the areas of Multimedia database, Sentiment analysis, and Big-data processing.