International Journal of Hybrid Information Technology
Vol.9, No.3 (2016), pp. 171-178
http://dx.doi.org/10.14257/ijhit.2016.9.3.16

Data Mining Methods for New Feature of Malicious Program

Haixu Xi and Hongjin Zhu

*Jiangsu University of Technology, No.1801, Zhongwu Avenue, Changzhou City,
Jiangsu Province, China
jsut_xhx@126.com, zhuhongjin0427@hotmail.com*

Abstract

Rapid Propagation of malicious program has caused great harm to the security of user information, the traditional way of killing methods, which is lagging behind and non-intelligent, has been unable to meet the demand of current detection. Studying the new malicious detection method on Windows Platform, screening out intelligent detection rules model feature of malicious executable and extracting the new malicious program detection methods based on data mining. Introducing the sample data processing and feature selection process, analyzing and simulating the new classification method, the result shows that the malicious program model can effectively improve the detection accuracy and reduce the rate of false negatives and false positives.

Keywords*: Coal mining data, Data mining, Class label prediction, Naïve bays classifier, Artificial neural network, Decision tree model*

1. Introduction

Malicious programs are malicious attack executable programs, such as damage the system or illegal access to sensitive user information. In the malicious programs detection, using the data mining technology is in order to be able to set up a method of automatic detection malicious executable file. Detection model based on data mining of massive data, and use these patterns to detect similar data detection. In this paper, the design of test system framework is the use of classifier to detect new characteristics of malicious programs; classifier is a kind of training set of rules by data mining algorithm.

The current has eight to ten new malicious programs. The new Trojan malicious programs has reached tens of millions in the year of 2012, these malicious programs has damaged bigger and bigger. It has brought huge losses to the user every day .And many types of attack are using malicious programs, DARPA test evaluation of attack, the Windows platform of malicious attack is based on malicious programs. At present, Microsoft released based on the kernel network vulnerabilities; malicious programs can use the vulnerability to open the back door into the Microsoft to steal all kinds of information within the network. The number of malicious programs is growing exponentially that caused great harm to the users of information security.

The traditional *get samples* - *analysis features* - *update the deployment* methods have been unable to meet the needs of the current killing virus, In order to solve the various problems of anti-virus software, this paper designs a kind of data mining technology on the basis of the new features of malicious programs intelligent detection rules, the method is using the Windows platform executable file format as the main characteristics, extraction of executable file, then analyzes and gets the new characteristics of malicious programs, using data mining technology to extract the detection rules, finds out the hidden malicious program rules, improves the accuracy.

ISSN: 1738-9968 IJHIT

2. The Key Technology

2.1 Malicious Programs

Malicious programs are divided into six types: viruses, worms, Trojans, botnets, spyware and malware, these six kind of malicious programs bring huge loss to the user. In order to counter anti-virus program, these malicious programs with anti-debugging techniques, anti-virtual machine technology, shell technology and antagonism safety software technology.

2.2 Data Mining

In order to deal with the anti-detection technology of malicious programs, to improve the accuracy of malware detection, at present, the data mining technology with better comprehensive properties are as follows: the neural network, Bayes and support vector machines, decision tree and association rules and so on. But these technologies still have many defects, such as ignoring screening malicious programs feature method, unable to detection with the client and the low efficiency of detection. Based on this, this paper designs a method of screening malicious programs' new characteristics, and based on the extract intelligent planning of the malicious programs, by the rules can detect much kind of malicious programs, and the versatility. The method first extracts the sample characteristics, then the sample feature data preprocessing, by processing the results to screening out new features, finally extracts intelligent planning.

3. Sample Data Processing

3.1 Sample Data Screening

In this paper, the author studies on the source data using the VX samples of their data, downloaded from the website of PE format 25585 samples of malicious programs, include normal application number is 4730, Figure 1 depicts the distribution of all kinds of programs in the sample.

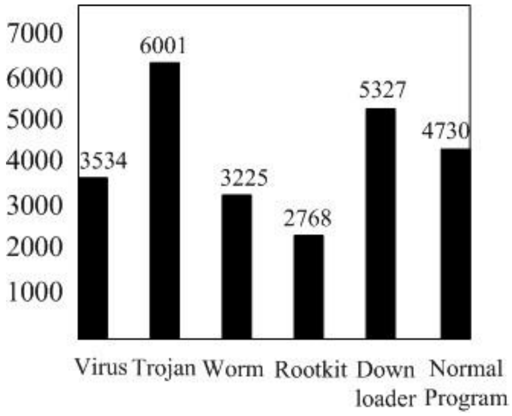


Figure 1. Table of Sample Data Distribution

3.2 Sample Data Standardization

In order to obtain the characteristics of PE, first of all, we want to extract characteristics of sample data. The Extraction process involves five steps: modifying the PE file, PE header information acquisition, PE section analysis, Dubious and malicious values reported and detecting packed files. This article involves the data extraction method is to use PE file provide python library for implementation. After extracting the sample data, some of the properties exist in the form of text, that can't be identify by data mining tools. So the data must be standardized by the way of Data Dictionary Mapping.

For example, in describing the entry point feature there usually have three ways which valid, effective, fuzzy. So replace them with 0, 1, 2; In the case of whether abnormal, to replace it to 0, 1.

3.3 Sample Data Analysis and Processing

3.3.1 Missing Value Processing

Caused by the damage is due to the file integrity missing value belongs to the abnormal value On the Windows platform. In order to remove these abnormal lack of value, when dealing with missing values of the non-normal data we use the delete method. Handling missing values don't affect the execution of the program other than the non-normal missing values. We use statistical filling method to handle the normal value data missing.

Missing value processing algorithm is as follows:

```
Foreach y Y
    Foreach ai y
        If ai L
            ai = average (a);
        If ai U
            Delete (ai);
```

Among them, Y is a sample set, y is the sample instance, a is a property of y, L is the lack of a reasonable values set, U is the unreasonable lack of value of the property set, average (a) is the average value of a property of the column.

According to the algorithm, it can find 10372 missing samples. There are 3017 reasonable missing values and 7310 unreasonable missing values.

3.3.2 Detecting and Removing Outliers

The sample data has isolated points inevitably, it will cause exists in the sample data. In order to eliminate the noise, we can use the method of based on distance and outlier detection to removing the isolated points. After the way of standardized the raw sample data sets, calculating the n of the distance between the two that named dij and thus form a distance matrix called R, such as formula 1.

According to the distance matrix R, $p_i = \sum_{j=1}^{n} d_{ij}$ when the greater the value of Pi that illustrated object I farther away from the other object distance. If the Pi maximum number of deleted, it can know the outlier sample data.

$$R = \left\{ \begin{matrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{matrix} \right\} \qquad (1)$$

According to this method it can be find out the number of 6224 isolated data, it can guarantee the accuracy of the classification rules by knowing these isolation data items.

Bases on the above two methods, the results of processing the sample data set as shown in Figure 2.

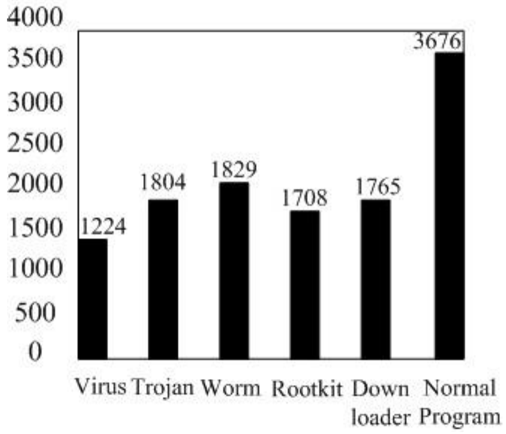


Figure 2. The Sample Data After Processing

4. Screening of New Features

4.1 Delete the Data of the Same Attributes and the Data of Linear Correlation between Attributes

1) Some data of the sample data set may have the same attributes after processing data. These data have practical significance for classification. For example, each PE file will be marked "MZ", this part of the data will enhance the characteristics of complexity. So we direct delete this portion of the data which can reduce the dimension of feature vector.

2) Linear correlation data divided into two categories, function and statistical relations. Function relation representation deterministic relations on the number of variables, the statistical relationship said statistical regularity between the variables with a certain amount of changes. There will be some of the data after the sample data set processing. If the PE file head tag a file for the DLL, this file must have derived function. So there exists linear correlation in these two properties.

The algorithm is as follows, p and q is constant

```
//Delete the attributes of the same column
For (a=0; a<i; a++)
        For (b=0; b<j; b++)
                If (mab = ma0)
                        Delete_col(a);
//Delete the linear correlation column
For (a=0; a<i; a++)
For (b=0; b<i; b++)
For(c=0; c<j; c++)
        If (mbc = mcc*p+q)
                Delete_col (b);
```

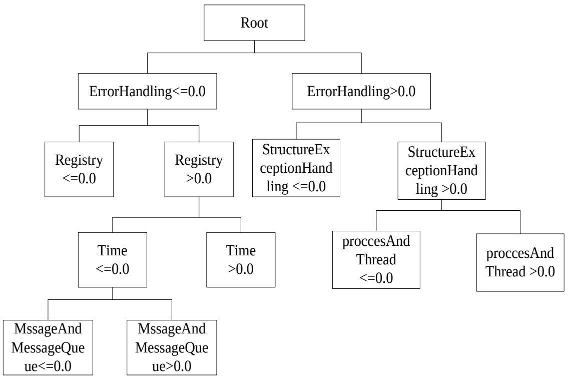


Figure 3. Intelligent Rule Tree by the Knime Generate

4.2 Redundant Data Processing

Redundant features to eliminate uses Principal Component Analysis (PCA), the algorithm selects the fewer number of important variables from multiple variables by linear transformation. Its optimality is that extracted the N main features from the N training focus, so as to dimension reduction. Suppose the N d-dimensional of original samples x1, x2... xn compose a matrix X (d * n). Project X onto vector Y of a low dimensional space and calculate the sample mean μ by formula (2), then by formula (3) to get the covariance matrix ST, finally calculate eigenvalue ei of ST by formula (4).

$$Y = W^T X \qquad (2)$$

$$S_T = \sum_{I=1}^{n} (x_i - \mu)(x_i - \mu)^T \qquad (3)$$

$$\lambda_i e_i = S_T e_i, i \in [1, N) \qquad (4)$$

After applying comprehensive evaluation for features of malicious program of the sample set through the above method and SPSS statistical software, we get the characteristic value of the sample set. As shown in Table 1.

Table 1. Analysis Results of Eigenvalue

Number	Principal components	Eigenv alue	Contribution rate	Cumulative contribution rate
1	Remote thread	8.36	0.06022	0.06022
2	Write service	7.61	0.05481	0.11503
3	Set the hook	7.552	0.05440	0.16943
4	File download	7.47	0.05381	0.22323
5	Hidden files	6.246	0.04499	0.26822
6	Release files	5.783	0.04165	0.30988
7	Modify the registry	5.78	0.04163	0.35151
…	…	…	…	…
41	Exception table size		0.00802	0.91752
42	Counter	1.112	0.00801	0.92554
…	…	…	…	…
59	Timestamp	0.224	0.00161	1.00000

From the table above, retention the coefficient of principal component greater than ninety-two percent and the eigenvalues greater than one, and delete the other eighteen feature items that do not conform to the requirements. Got forty-one attributes as the new feature of malicious programs after screening the characteristic values. And the feature vector of malicious programs includes the following attributes: the serial number, the file name, the file categories, the file header information, API function sequence and API function name. Including file categories expressed in 0, 1, and 0 indicates normal procedures and 1 indicates malicious programs.

5. Feature Evaluation

In order to assess forty-one new features after screening and verify its validity, we will make a simulation experiment on the new features of this article and then get the results of the experiment, including the precision, the false alarm rate and the missed alarm rate.

The dataset this article uses downloaded from VX Heavens. They are eleven thousand nine hundred and fifty-five PE file information after data processing and choose Knime as the data mining platform. The classification workflow of Knime is first the original data after filtering columns and distinguishing colors get the training and test sets, then the training model trains the training and test sets and scores the prediction, last get the assessment model. There are eleven thousand nine hundred and fifty-five data of the original dataset in this article, including three thousand six hundred and seventy-six normal procedures and eight thousand two hundred and seventy-nine malicious programs. As shown in Table 2 of the test results of several kinds of classification algorithm.

Table 2. Assessment Results of Eigenvalue

Classi fier	Sample number	Test result	Detection rate	False alarm rate	Missed alarm rate
Bayes	3676/8279	45/6	96.8%	0.4	3.2%
MLP	3676/8279	28/9	98.8%	0.5	1.2%
SVM	3676/8279	34/6	97.4%	0.4	2.6%
C4.5	3676/8279	18/4	99.7%	0.1	0.3%

PS: The test result is the false alarm rate or missed alarm rate

6. Extract Rules

The purpose of extracting rules is to separate normal procedures and malicious programs. And it needs to have a higher detection rate and a lower false alarm rate. According to the test results in Table 2, we found that using the C4.5 algorithm of the decision tree can get the best test result. So this article chose the C4.5 algorithm as an algorithm to build a classifier. It generated the results of the decision tree on the Knime platform as shown in Figure 3. Since the result data is bigger, the figure is only for part of the decision result.

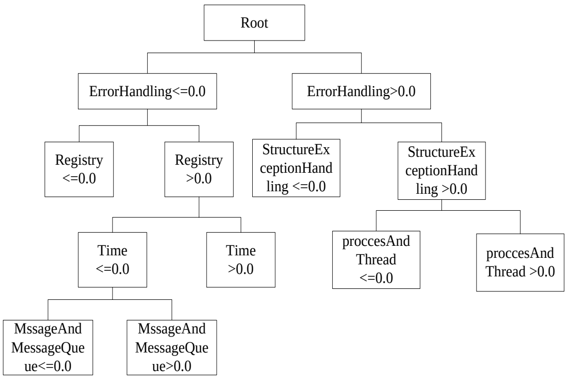


Figure 4. The Intelligent Rule Tree Knime Generated

7. Epilogue

In this article, the PE file format as the main feature of the source. Through screen the new features got the new feature of malicious programs as well as reusable intelligent test rules, and proved them through classification methods such as Bayes, MLP, SVM, and C4.5. The results of the experiment show that this new feature has the lower missed alarm rate and false alarm rate and the higher detection rate. It is very important to improve the efficiency and precision of the test.

Acknowledgements

This work was supported by Natural Science Fund of Jiangsu Province (BK20130235), Natural Science Fund of Changzhou (CJ20140049), China National Natural Science Fund of China (61302124) and Scientific Research Foundation of Jiangsu University of Technology (KYY14020).

References

[1] L. Hui, “The Network Computing Model Based on The Virtual Machine”, Science, Technology and Engineering, vol. 16, no. 5, **(2005)**, pp. 1209-1211.

[2] B. Thuraisingham, “Data Mining for Malicious Code Detection and Security Applications”, Intelligence and Security Informatics Conference, European 2011 IEEE Conference Publications, **(2011)**.

[3] Y. Fu-Qiang and W. Hao, “Application of Genetic Programming on Data Mining”, Science Technology and Engineering, vol. 14, no. 8, **(2008)**, pp. 3966-3969.

[4] L. Wen-Hua, “Malicious Detection Method Based on Reverse Technology and Research”, the Police Technology, no. 6, **(2012)**, pp. 26-28.

[5] L. Wen-Hua, “Malware Analysis Method Based on Reverse Technology”, Computer Application, no. 11, **(2011)**, pp.63-64.

[6] M. M. Masud, L. Khan and B. Thuraisingham, “A Hybrid Model to Detect Malicious Executables”, Communications, ICC’07, IEEE Conference Publications, **(2007)**.

[7] L. Wen-Hua, “Extraction of Polymorphic Malware Signatures using Abstract Interpretation Theory”, Information Network Security, no. 1, **(2013)**, pp. 18-21.

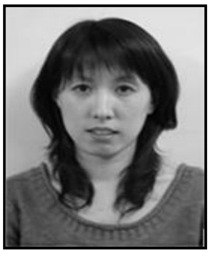
[8] L. Peng and W. Ru-Chuan, “Malicious Code Dynamic Analysis Based on Self Similar Characteristics”, Journal of Nanjing University of Posts and Telecommunications (NATURAL SCIENCE EDITION), vol. 6, no. 1, **(2012)**, pp. 24-26.

[9] Z. Yi-Chi and P. Jian-Min, “Program Malicious Behavior Recognizing Method Based on Model Checking”, Computer Engineering, vol. 9, no. 20, **(2012)**, pp. 35-38.

[10] F. Shahzad and M. Farooq, “ELF-Miner: Using Structural Knowledge and Data Mining Methods to Detect New (Linux) Malicious Executables”, Knowl.Inf.Syst, no. 30, **(2012)**, pp. 189-192.

Authors

Haixu Xi, He received the master’s degree in Educational Technology from Nanjing normal University in 2006.Currently, he is an lecturer at the School of Computer School of Computer Engineering in Jiangsu University of Technology. His interests are in digital teaching resource development and multimedia technology.



Hongjin Zhu, She received the M.Sc. and Ph.D. from the Yamgata University of Japan in 2007 and 2010, respectively. She was employed as a special researcher in the Department of Engineering, Yamagata University of Japan in 2010. She is currently anassociate professor in Jiangsu University of Technology, Changzhou, China. Her research interests include image processing, computer vision, pattern recognition and evolutionary computation.