Classification of Spyware Affected files using Data Mining Techniques

D Anil Kumar, Sisira Kumar Kapat, Susanta Kumar Das, Satya Narayan Tripathy

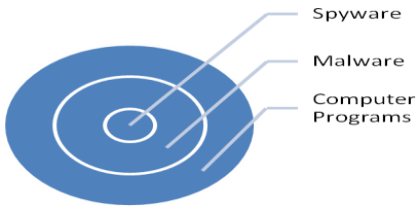
Abstract*: Spyware is a malicious computer program which collects or gathers information about a person or organization and sends them to third party without the user's knowledge and explicit consent. Spyware is a sinister malware which is mainly connected to spying activity. There are various types of spyware. So there is a need to study the Spyware. This chapter concerns with the study and classification of Spyware. The mathematical term and figures are provided as in where needed, to support the description. The experiment conducted in the chapter shows reliable accuracy and error rate in the classification. This can be used in the malware detection system.

Keywords: Spyware, Malware, Data Mining, Classification, API, Network Security.

I. INTRODUCTION

Starting from Mobile phone, all the handhold or portable devices, software is a major requirement everywhere. Software is written using some computer language. Some of the software coders take advantage of it and misuse their knowledge in creating malicious codes or mal-codes or malwares. Some software are designed for dual purpose (Chatterjee, 2018) in windows as well as android environment. This can be possible in the entire known operating system environment. The dual purpose refers that, the software acts for legitimate purpose as well as illegitimate purpose. According to McAfee, the number of new malwares increased, more than 100% and also approximately 200% of the mobile malware increased from 2016 to 2017 (McAfee, 2018).

Prior to the definition of Spyware, we have to know the meaning and definition of Malware which is a superset of Spyware; or alternatively, 'Spyware is a subset of malware, which is shown in
Figure 1 (Figure showing spyware is a subset of Malware).



Revised Manuscript Received on July 22, 2019.

D.Anil Kumar, Department of Computer Science and Engineering, GIET University, Gunupur, India
Sisir Kumar Kapat, Department of Computer Science, Berhampur University, Berhampur, India
Susanta Kumar Das, Department of Computer Science, Berhampur University, Berhampur, India
Satya Narayan Tripathy, Department of Computer Science, Berhampur University, Berhampur, India

Spyware is a type of software that can install itself or run on user computer without providing notice, consent, or control to the user. According to Microsoft windows documentation, Spyware may not display any symptom after infection; but sometimes spyware or unwanted programs can affect the smooth running of computers. The infected computer may run slow. Spyware can change the computer setting or monitor online behavior or collect information about the user; the information may be personal detail or other sensitive information about the user.

Spyware programs are usually hidden among other programs or can be unwittingly downloaded to a user's system when certain websites are visited. This is known as drive-by downloading. Once the spyware is installed it remains invisible to the user and secretly monitors the user's system and the user's activities. The spyware can transmit personal information back to the remote spy.

According to Pandey et. al. (2015), the application allows to detect whether a particular executable is spyware or not prior to their installation. Their software does not require updates from the remote server.

According to Saroiu, Gribble, and Levy (2004), "Spyware is commonly used to refer to software that, from a user's perspective, gathers information about a computer's use and relays that information back to a third party. This data collection occurs sometimes with, but often without, the knowing consent of the user."

According to McAfee (2005), the formal definition of spyware is, "It is software whose function includes the transmission of personal information to a third party without the user's knowledge and explicit consent."

All the above definitions have the similar meaning about spyware. If software is stealing the confidential information of a user and sending those data to third party for illegal purpose, then that software is considered as a spyware. Some Keyloggers are the only hardware, which is involved in spying activity. So these are also included in the category of spyware.

II. TYPES OF SPYWARE

In user prospective, spyware can be categorized as Domestic Spyware and Commercial Spyware (VanNess & Weaver, 2018).

- Domestic spyware is something which is installed by the user, employer or third parties to monitor the network activity or to collect some personal information (often confidential) of the user. In home, the guardians install this type of spyware to monitor their children or family member. The spyware collects the information of the user and send it to the admin or the controller of the spyware e.g. Dialers, Keylogger, Spybot etc.

- Commercial spyware is installed by the companies to monitor the browsing habits of the user. Accordingly they use the information for business or marketing purpose e.g. Browser Hijack, Profiling Cookies, Drone Ware etc.

In business prospective, spyware can also be categorized as Surveillance Spyware and Advertising Spyware.

- Surveillance Spyware is used for our daily habits or purpose. Often this type of spyware is used by the user, corporations, detectives, intelligence agencies etc. The examples include, Keylogger, Screen Recorder etc.

- Advertising spyware is used by most of the companies for advertisement of their product. These are otherwise called as adware. Adware get downloaded when there is internet connectivity on the user computer and pop-up (sometimes it is offline) the content on the user screen. Advertising spyware gets installed along with other software or via ActiveX controls on the internet. Very often adware is harmless, but at the extreme, it is sinister.

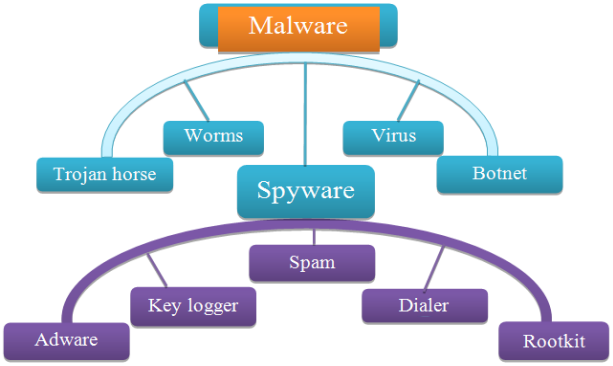


Fig. 2: Types of Spyware

III. CLASSIFICATION

Classification is a method to specify the given object into some predefined category or class based on some condition. According to the Bayes theorem, the probabilistic approach for classification may be represented as,

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

Where, 'H' represents the Hypothesis of being a class. 'X' represents the given data or condition.

Classification technique can be used to classify the data linearly or non-linearly. The linear classification and non-linear classification of data is shown in Fig. 4. Linear Classification and Fig. 3. Non-linear classification respectively.

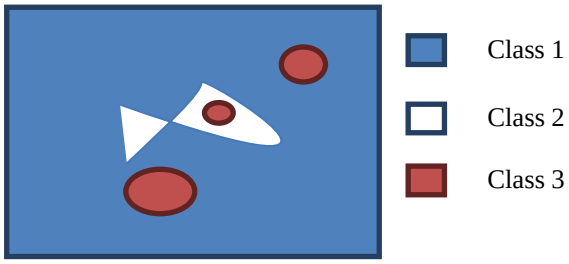


Fig. 3. Non-linear classification

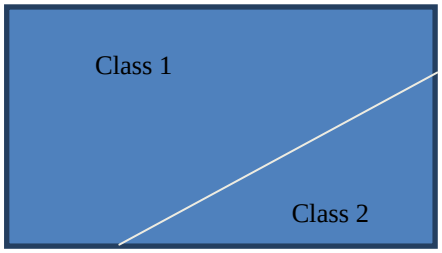


Fig. 4. Linear Classification

(binary classification), or it can be for multiple classes. Figure-3 is an example of binary classification whereas figure-4 is an example of multi-class classification.

Various algorithms are there for classification such as Naïve Bayes, Decision Tree, Decision Table etc. These classification techniques perform the simple classification based on different criteria. They perform different in different conditions. So this is hard to decide which algorithm is better for classification, unless any experiment is done with the given dataset.

IV. RELATED WORK

Signature based malware detection technique uses some specific features or unique strings extracted from binaries (Bahraminikoo, 2012) of the portable executable (PE) file to analyze and detect the malwares. Obfuscation technique can bypass the signature of a file (Dwarakanath, 2012). Malware writers use obfuscation techniques like packing, encryption or polymorphisms to avoid being detected by anti-malware engines. Hence signature based malware detection technique fails to detect some of the malicious files. Although signature based detection is faster to use, but in contrary, the frequent update of the signature database is an additional overhead and time consuming. The common obfuscation

techniques (CERT, 2014) (Balakrishnan & Schulze, 2005) are Dead-code-insertion, Code transportation, Register renaming, Instruction Substitution etc.

In previous studies Naïve Bayes, Support Vector Machine, Decision tree classifiers were used to detect new malicious executable. Park et.al. (2018) used 'Kyoto 2006+' dataset, which is collection of network packet information data to classify the various types of attacks for intrusion detection system. Hatada and Mori. (2017) used android environment to collect several PUA (potentially unwanted applications) like adware and remote monitoring tools, which are also a subset of spyware, for classification. Chatterjee et.al. (2018) used both windows as well as android spyware data for their experiment. They used machine learning approach to find the dual-purpose software which is used for both legitimate uses as well as for spying activity; and also they claim that the current anti-spyware tools are insufficient.

According to a report of Symantec (ISTR, march-2018), among all the malwares only 6% of the malwares are involved in disruptive activities like disk wipe-off, 27% of the total malware are involved in zero-day vulnerability. The total new malware increased by 54%, botnet traffic increased by 62.3%, Phishing URL traffic in web traffic increased by 182.6%, overall spam rate increased by 1.2% from 2016 to 2017. The total mobile spyware increased by 20%. In mobile environment, 63% of the apps leak phone number, 37% of the apps leak location information, 35% of the apps leak installed app information.

Spyware exist in windows as well as android environment. This chapter uses API calls collected from different executable from windows environment to classify them. API function calls are the set of instructions which defines the behavior of an executable. So by classifying the APIs, the malicious as well as benign programs can be differentiable.

III. EXPERIMENTAL SETUP

WEKA is a reliable open source data mining tool which can be used for the experiment. Some other tools can be used as well. A debugger or disassembler is needed to disassemble the application so that the API calls can be recorded. X32/X64 dbg is an open source debugger which can be used to collect the API calls. X32 dbg is used for 32 bit applications whereas X64 dbg is used for 64 bit applications. The API calls should be collected into a database/datasheet. The database or datasheet must be in such a format that, it must be able to process by the data mining tool. WEKA is able to process arff, C4.5, CSV, JSON and libsvm files.

The database should contain the spyware as well as benign data. To get better accuracy, the dataset must contain a large amount of data. A number of online sources provide different dataset for reliable experimental purpose. Some examples include,

- http://ocslab.hksecurity.net/apimds-dataset,
- http://www.csmining.org/tl_files/Project_Datasets/task3/task3/ . etc.

'Virustotal' is a platform to find whether an application is malicious or not. A number of antimalware engines are exist in virustotal in a single platform. For the experiment, one can use it with the assumption that the anti-malware engines are updated one.

After completion of the classification, confusion matrix is generated. The terms used in the confusion matrix are, TP (True positive rate), FP (False positive rate), TN (True Negative rate) and FN (False negative rate). These terms are well understood from **Error! Reference source not found.**.

Table – I: Confusion Matrix

		Predicted Class	
		Malware	**Benign**
Actual Class	**Malware**	TP	FP
	Benign	FN	TN

From Table-1, the above terminologies can be well described as,

- TP: Number of correctly identified malware programs
- FP: Number of wrongly identified malware programs
- TN: Number of correctly identified benign programs
- FN: Number of wrongly identified benign programs

Now, the user can be able to find the Accuracy, Error rate, Precision etc.

Accuracy is the ratio of the number of correctly classified instances (N_c) to the total instances (N_t), and is represented as,

$$Accuracy = \frac{N_c}{N_t} = \frac{TP + TN}{Total} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

Error rate is the ratio of the number of incorrectly classified instance to the total instances in the dataset, and is represented as

$$\text{Misclassification rate or Error rate} = \frac{N_t - N_c}{N_t} = \frac{FP + FN}{Total} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4)$$

VI. RESULT ANALYSIS

In this chapter, WEKA data mining tool is used for the experiment. The dataset is collected by processing various executable files in X32/X64 dbg. The API data collected from the X32/X64 dbg is stored in a database for the experiment. The experiment described in this chapter uses the datasheet in '.csv' format data. This file format is simple, reliable and

easy to handle. The sequence of API, of an application is stored as a single record in the datasheet. Hundreds of attributes (API calls in average) are present in the datasheet. In this experiment, more than 8000 data of malwares and hundreds of benign data are used. The dataset is formulated using the above mentioned process as well as some online sources; and it contains the data of several spyware like adware, botnet etc. So basically this experiment is concerned with multi-class classification. The example of the dataset is provided by Fig. 5. Example of API dataset.

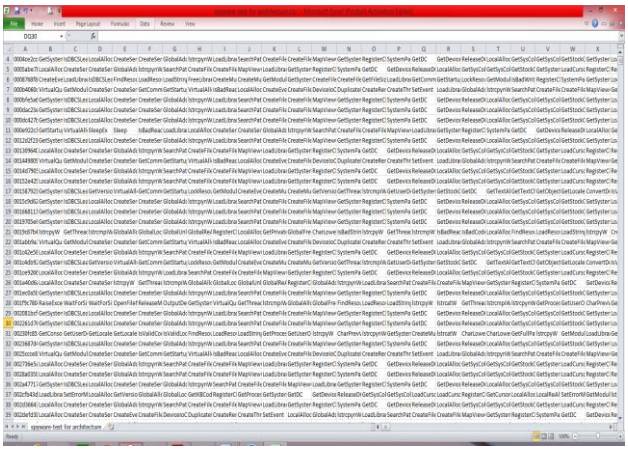


Fig. 5. Example of API dataset

With the assumption that the anti-malware engines are updated one, this experiment uses virustotal to find the malware family (class value) in this experiment. CAT-QuickHeal antimalware engine is used for the class attribute in this chapter.

J48 Decision Tree classification technique of WEKA is considered for the classification purpose. The output of the J48 decision tree is represented by the Fig. 6. Decision Tree generated from WEKA.

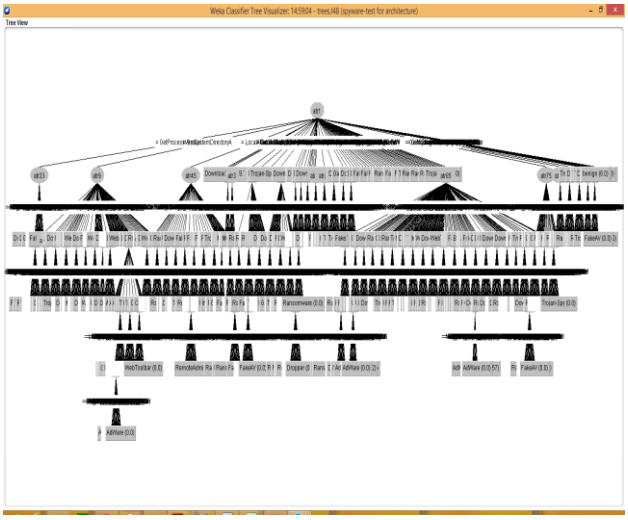


Fig. 6. Decision Tree generated from WEKA

The accuracy of this experiment is 86.93%, whereas the error rate is 13.06% which is shown in Fig. 7. Accuracy and Error value of the experiment.

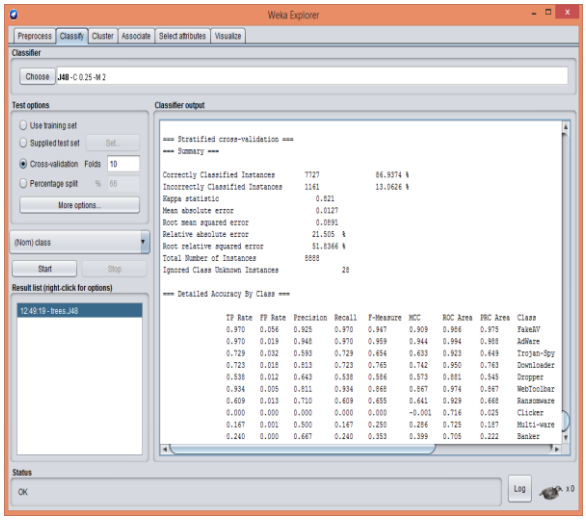


Fig. 7. Accuracy and Error value of the experiment

The summary of the experiment is represented by the Table-2: Summary of the experiment.

Table-2: Summary of the experiment

Accuracy	86.93%
Error rate	13.06%
True positive rate	0.869
False positive rate	0.033

From the decision tree it is clear that, the spyware data is categorized into some families according to some rules. When a new executable is entered into the system, these rules can be applied so that the new executable can be classified into some predefined class.

VII. SUMMARY AND CONCLUSION

The number of spyware is increasing in both the windows as well as mobile environment. Hence this type of classification of spyware is necessary to classify the incoming spywares. From the experiment approximately 87% of accuracy and approximately 3% of false positive rate achieved.

REFERENCES

1. Arini Balakrishnan, Chloe Schulze. (2005) Code Obfuscation Literature Survey. Retrieved from, http://pages.cs.wisc.edu/~arinib/writeup.pdf
2. Brian VanNess, Joanne C. Weaver. (Accessed on 30th July 2018) What Types of Spyware are Out There?. Retrieved from http://www.toptenreviews.com/software/articles/types-of-spyware/
3. CERT-UK (2014) Code obfuscation. retrieved from www.cert.gov.uk ISTR (March 2018). Internet Security Threat Report, volume 23. Retrieved from https://www.symantec.com/content/dam/symantec/docs/reports/istr-23-2018-en.pdf
4. Karishma Pandey, Madhura Naik, Junaid Qamar , Mahendra Patil. (March 2015) Spyware Detection Using Data Mining. International Journal for Research in Applied Science & Engineering Technology (IJRASET), 3(III), pp 488-492
5. Kinam Park, Youngrok Song, Yun-Gyung Cheong (2018) Classification of Attack Types for Intrusion Detection Systems using a Machine Learning Algorithm. IEEE Fourth International Conference on Big Data Computing Service and Applications. DOI 10.1109/BigDataService.2018.00050, pp 282-286
6. McAfee® Proven SecurityTM. (September 2005) white paper on Potentially Unwanted Programs Spyware and Adware. Retrieved from www.mcafee.com
7. McAfee (march-2018) McAfee Labs Threats Report, Retrieved from https://www.mcafee.com/enterprise/en-us/assets/reports/rp-quarterly-threats-mar-2018.pdf

A. Mr. B. Dwarakanath, Mr. A. Suthakar. (2012) Prediction and Detection of Malware Using Association Rules. International Journal of Power Control Signal and Computation (IJPCSC), 3(1), pp 45-50

8. Parisa Bahraminikoo, Mehdi Samiei yeganeh, G.Praveen Babu. (2012) Utilization Data Mining to Detect Spyware. IOSR Journal of Computer Engineering (IOSRJCE), 4(3), PP 01-04
9. Stefan Saroiu, Steven D. Gribble, Henry M. Levy. (2004) Measurement and Analysis of Spyware in a University Environment. Proceedings of the first Symposium on Networked Systems Design and Implementation (NSDI), San Francisco, CA.
10. Mitsuhiro Hatada, Tatsuya Mori. (2017) Detecting and Classifying Android PUAs by similarity of DNS queries. IEEE 41st Annual Computer Software and Applications Conference, DOI 10.1109/COMPSAC.2017.103, pp-590-595
11. Rahul Chatterjee, Periwinkle Doerfler, Hadas Orgad, Sam Havron, Jackeline Palmer, Diana Freed, Karen Levy, Nicola Dell, Damon McCoy, Thomas Ristenpart. (2018) The Spyware Used in Intimate Partner Violence. IEEE Symposium on Security and Privacy, DOI 10.1109/SP.2018.00061, pp 441-458

AUTHORS PROFILE

Mr. D Anil Kumar is working as an Associate Professor in GIET University. He is pursuing his Ph.D from Berhampur University. He has 5 publications in national and international journals. He is the member of IE, CSI, and ISTE. His Research area is data mining.

Mr. Sisir Kumar Kapat is a research scholar from Berhampur University. He has 6 publications in national and international journals. He is a member of ISTE and CSI. His research area is data mining.

Dr. Susanta Kumar Das joined the Dept. of Computer Science in 1993. He has teaching experience of 23 years in the department. He has attended no. of national & international conferences. To his credit, he has served as H.O.D for 2 years in the department. At present he is the coordinator of M.Tech(S.F) course & as coordinator of spoken tutorial project conducted by IIT Bombay & funded by MHRD, Govt of India. Fourteen no. of scholars are awarded Ph.D under his guidance. One D.Sc degree is awarded in Computer Science under his guidance. He has been felicitated award of honour by Dept. of Mathematics, Maharshi Dayanand University Rohtak, Haryana in the international conference on History & Development of Mathematical Science & Symposium on Nonlinear Analysis. His research are in Software Engineering & Network Security.

Dr. Satya Narayan Tripathy received his M.C.A. and Ph.D. degrees in Computer Science from Berhampur University, Berhampur, Orissa, India in the years 1998 and 2010, respectively. He has been teaching in the Department of Computer Science, Berhampur University since 2011. Currently, he as a Lecturer in the Department of Computer Science, Berhampur University. in addition to his normal duty he is assigned as the Web Administrator of Berhampur University web portal. Dr. Tripathy serves on the advisory boards of several organizations and conferences. He is an Life Member of Computer Society of India (LMCSI), Life Member of Orissa Information Technology Society (LMOITS) and Member of several professional bodies.