SPYWARE DETECTION AND PREDICTION FOR USER APPLICATIONS

Mahesh V
Jain University,
Bengaluru-560069, Karnataka, India

Dr. Sumithra Devi K A
Dayananda Sagar Academy of Technology and Management
Bengaluru-560082, Karnataka, India

Abstract: A user application holds essential places in our day today life. Possibility of attacks on user application increases due to the sensitive data is on computer system. Mostly attackers uses application weakness to steels the user information. Spyware is one of the types of threat that takes user's sensitive information without user knowledge. This proposal works states the method to detect malicious code and predict vulnerabilities using data mining techniques and classifiers to predict Spyware malicious files.

Keywords: Malware, Spyware, user application, data mining, user application, classifier.

1. INTRODUCTION

An application contains a set of programs that are designed to do a specific task for the welfare of the user. The growth of the computer application and the internet tends to increase the chance of malicious code installation. Examples of malicious codes are spyware or malware, Trojan horse, virus etc. It is hard for the users to find out the installed spyware or malware in the system. Spyware is software, which is installed without the knowledge of the user. It is used for monitoring the user's action on the internet. The malicious code can be different types of data like user login, user account information or personal information. Some of the spyware result in slow internet connection and compromise the network. This type of software is available in a bundle, which is available as a free download. However, once it is installed it is difficult for the user to identify it and remove it from the system [1].

2. DATA MINING

Data mining is a process of analyzing data from different perspectives and summarizing it into useful information [2]. The main goal of a data mining process is to extract information from a raw dataset and transform it into an understandable structure for future use. Data mining tools help an organization to make a prediction from a knowledge-driven decision. These tools help to resolve business-related questions with less consumption of time than a traditional method to infer data. Raw data has hidden information so while refining it gives a valuable data.

Areas were data mining used is Retail, Finance, Banking, Manufacturing transportation and Aerospace. The Evolution of Data Mining is Data Collection, Data Access, Data Warehousing and Decision Support and Data mining [8].

2.1 CHARACTERISTICS OF DATA MINING

The characteristics of data mining are 1) A large amount of data, 2) Noisy, incomplete data, 3) Complex data structure and 4) Heterogeneous data stored in a legacy system.

2.2 MOTIVATION OF DATA MINING

The motivation of data mining are 1) A huge amount of data, 2) An important need for turning data into useful information and 3) Fast growing amount of data, collected and stored in large and numerous databases exceeded the human ability for comprehension without powerful tools.

2.3 DATA MINING CONCEPTS

Data mining model includes six main steps to extract useful information. The steps are 1.) Define data 2.) Prepare data 3.) Explore data 4.) Build models 5.) Validating models 6.) Deploying and updating models [9].

2.3.1 Define data

Defining problem is the first step to extract information, defining problems includes analyzing resource, problem scope, what to infer from raw data, what kind of data so that it helps to determine which learning method to choose i.e. classification or clustering.

2.3.2 Prepare data

Raw data contains different data are from everywhere, as the result datasets will have missing entries, inconsistent data. So cleaning datasets is required that eliminates the bad data and missing data. Thus helps to achieve accurate information.

2.3.3 Explore data

Exploring data helps the model to create correctly, as it explains the problem of the datasets. Exploring techniques includes calculating value, standard deviation and distribution of data.

2.3.4 Build models

Model or module building achieved using exploring data, information taken from exploring data helps to create modules. Module plays a major role as prediction and inferring take place here. Data sets are divided into the training data and testing data, where training data's are used to build and process the module whereas testing data will helps to predict the outcome information.

2.3.5 Validating Data

Before deploying module or models, it has to be tested to verify whether it works correctly or not. So the testing data is used to validate the module.

2.3.6 Deploying and updating data

Once module deployed in an environment, it helps to make a prediction. If possible it can be modified and updatable. Thus allows that one can update the module dynamically after reviewing and analyzing.

3. METHODS USED FOR DETECTION

Spyware detection techniques are 1.) Pattern Matching or signature-based detection technique, 2.) Heuristic Analysis based detection technique and 3.) Data Mining based detection technique [2] [8].

3.1 PATTERN MATCHING OR SIGNATURE BASED DETECTION TECHNIQUE

In pattern matching or Signature-based detection, it has an algorithm or hash that finds malicious code in the system. This can be developed by a group of experts who ha familiarity and experience in related field. It contains virus pattern or virus signature, so whenever a system experience a spyware this allows finding a matching in the signature pattern database. Thus, it becomes simple to find malicious codes by comparing the file code to the signature pattern database.

All the anti-virus or anti-spyware software uses the signature method to find malicious. But some false error will occur in signature pattern method. There is possible in the occurrence of an error due to the manual coding. In order to secure, it has to be updated regularly to avoid growing false negative rate.

3.2 HEURISTIC ANALYSIS BASED DETECTION TECHNIQUE

The heuristic detection method is similar to signature-pattern method but here in the heuristic method it also searches for some specific command in a program. It helps the heuristic engine to detect a virus or malicious code. There are many methods used in the heuristic analysis to detect maliciously but some of the detection techniques used in the heuristic analysis are File analysis, behaviour-based, and rule and weight based.

3.2.1File analysis based heuristic-based detection methods

Files are analyzed in-depth to find malicious code and determine the purpose of why this created.

Weight and rule-based detection

The weight-based detection method is an old technique to find malicious code in the file where it detects using weights according to the possibility of the danger it occurs.

A rule-based heuristic detection method used now a day as it detects using certain rules that help to detect malicious code by comparing it with the heuristic engine.

3.2.2 Behavioural-based detection techniques

Signature-based techniques used to find or detect malicious code. It only detections of malicious code in need then signature–pattern method used. However, to detect the behaviour of the malicious and how it will attack the behaviour-based detection is used.

Most of the anti-spyware and anti-virus use behaviour-based detection methods as it is compared with signature-based to detect spyware class, this is due to independence on binary representation.

However, it is used well but it also has some negative reports as it is very complex to build and same as signature-based where it needs a frequent update to keep the signature database to avoid false positive rate.

3.3 DATA MINING BASED DETECTION TECHNIQUES

In this detection technique, it uses data mining techniques to detect spyware. Data mining based detection techniques are used as classifiers to detect the application whether spyware is present or not. A module is created where a classifier is built by a training set of data's to detect spyware. This method of detection is fast to find malicious code.

This technique does not depend on signature method so false positive occurrence is impossible and provide accurate detection than any other method.

3.4 MALWARE TYPES

Malware's are the set of programs that contain some malicious code in order to gather information from the user without their knowledge.

Few malicious threats are given below [3][4][5].

3.4.1 Adware

It is an automatic delivery system that delivery's advertisement as a pop-up ad on websites and it will appear pop-up ads when the user clicks for some links to download songs, movies etc.

3.4.2 Bot

This is created to perform specific operations automatically on websites that scrape the server data. To secure mostly websites uses CAPTCHA and block the automatic operations.

3.4.3 Bug

Bugs are the set of procedure that affects the source code or compiler. A slight bug affects the behaviour of the program and major bugs results in steel security related issues like authentication, access privileges and steal user information.

3.4.4 Ransomware

It is a type of malware, which holds the computer system until the user pays some amount. This ransom blocks user system by encrypting its hardware, locking the system and pop-up a message that has a message to pay the amount.

3.4.5 Rootkit

This is one type of software that is designed to access the system remotely without noticed by user and security software. If a rootkit is installed then that can be accessed remotely access system by changing the configuration of a system, execute files to access information.

3.4.6 Spyware

Spyware is a malicious codes that eavesdropping user information without their knowledge. This will monitor, gather information, and extract sensitive information.

3.4.7 Trojan horse

It is normally called as "Trojan" that acts as a normal file to steal the user information, modify the files, monitor the user actives.

3.4.8 Spam

A Spam is a message that sends to user application by either emails or a multimedia message. This kind of massages are normally seen in Gmail scam, it contains a link that directs to some website pages where it steals user information like name, password, card numbers, and some sensitive information's.

3.4.9 Phishing

It is a website that is created to steal the user information. Once user visit to this website it asks to enter user information is like name, password, mail ID. As all the sensitive information, contact details, important documents, bank account and card

information are usually stored in emails. When an attacker gets user ID, password, and it is easy to get their information's. This kind of links can be seen in regular web pages as an ad, which directs to a phishing website to gather information.

3.4.10 Spoofing

Spoofing will act as a trusted party; here an attacker will act as a trusted party. They will send a spam message as from a trusted party.

4. PROPOSED SYSTEM

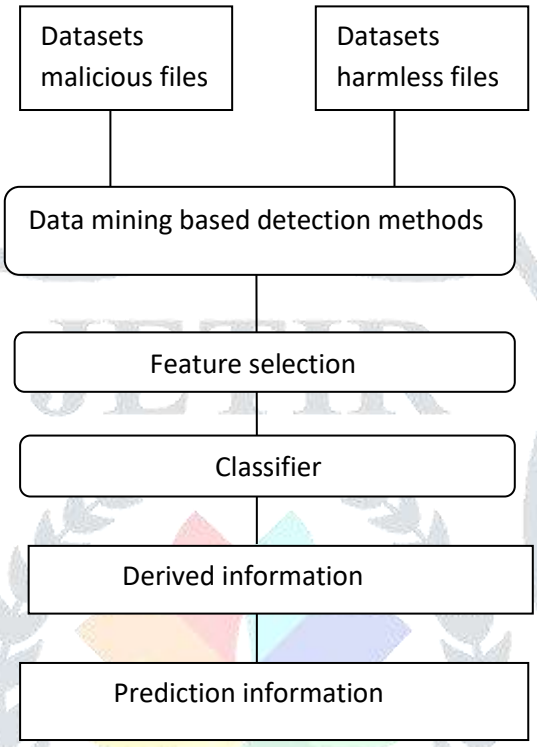

Fig.1 Workflow

The selected datasets contain malicious files and innocuous files that passed to data mining detection to detect the malicious files. Then the malicious files are passed to feature selection to extract malicious feature that is trained to detect and classifier to predict malicious file. Finally, the derived information contains the predicted result.

5. CLASSIFICATION

Classification is a machine learning technique used to predict group membership for data instances. Classification may refer to categorization. The classification technique is a systematic approach to build classification models from an input data set [6].

5.1 Naive Bayes

In supervised learning, Naïve Bayes classifier is a family of probabilistic classifiers based on Bayes theorem where there is an independence assumption between the features [10]. It has been proposed since the 1950s and it remains a popular method for text categorization where the problem in categorization between the documents as belongs to one category to another.

The main application of naïve Bayes classifier is in automatic Medical diagnosis. These are highly scalable, requiring a number of parameters linear in the number of variables in a learning problem.

The advantages of Naive Bayes are super simple as we are just doing a bunch of counts. If the NB conditional independence assumption actually holds, a Naive Bayes classifier will converge quicker than discriminative models like logistic regression, so one need less training data. In addition, even if the NB assumption doesn't hold, an NB classifier still often performs surprisingly well in practice. A good bet is that if we want to do some kind of semi-supervised learning, or want something embarrassingly simple that performs pretty well.

6. RESULT AND EVALUATION

Data mining plays an essential role in detection techniques, Naive Bayes classifier helps in analyzing malware detection that helps to gain data detection from a sample sets. This classifier explains how human prediction exactly used. For prediction, it chooses a set of words in specific order i.e. malicious codes can be arranged in order so that while detecting malware the module could predict easily. Probability prediction is good in Naive Bayes classifier as it research patterns using probability in statistical distribution. Example consider "win prize" as a malicious code, when visiting websites or downloading songs or movie from mobiles if fortunately or unfortunately an user has pressed link that contain malicious code the module which receive this code will detect from training datasets. Each predefined malicious code is assigned with keys and a these keys are designed in module that analysis with training sets and if detects found then it says that malicious code attacks the user system.

CONDITIONAL PROBABILITY

This helps classifier in detection of accurate answer. Consider an example on malware prediction as class and benign so P (CLASS| BINARY) where CLASS is given as binary that detects malware.

P (CLASS|BINARY) is a direct proportional to P (BINARY|CLASS) * P (CLASS)

P (BENIGN|BINARY) = P (BINARY|BENIGN) * P (BENIGN)

P (MALWARE|BINARY) = P (BINARY|MALWARE) * P (MALWARE)

All these conditions are proportional, so detecting malware by comparing values and assigning higher values to finalizing detection. This is regressing and ranking method to detecting malware.

To verify malicious detection, a dataset is taken and compared to give higher keys values that could give best result in assumption of malicious code of 1500 data's where top count is detected and marked high value, which is malware.

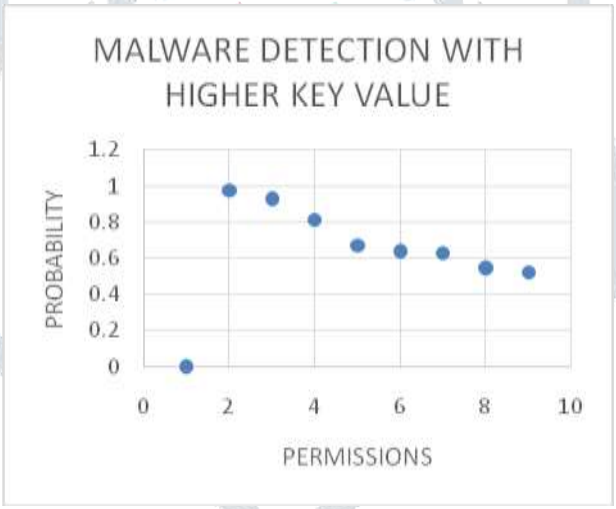

Figure 2. Malware detection with higher key value

Malware detection result with key values, probability of the malicious code from sample malware sets. It shows the malware detection using key values as prioritizing in Naïve Bayes.

7. CONCLUSION AND FUTURE WORK

In this proposed work, a malware detection module based data mining techniques are used to detect malicious software. Malware detection with data mining module is assumed that it can give an accurate result compared to other traditional methods. For home users and small companies uses anti-virus or anti-cyber methods to detect and predict malicious. However, the large organization uses data mining along with machine learning techniques on gateway level to predict. User application needs a special method to secure their personal details. So it is assumed that proposed detection method by sample sets using key values gives an accurate result. In order to have high accuracy proposed data mining system is combined with other detection techniques for user application. Future work of this will be detection methods with machine learning and combined with detection techniques.

REFERENCES

[1] Karishma Pandey, Madhura Naik, Junaid Qamar ,Mahendra Patil (2015), Spyware Detection using Data Mining, International Journal of Engineering and Techniques.

[2] Ms. Milan Jain, Ms. Punam Bajaj (2014), Malicious Code Detection through Data Mining Techniques, International Journal of Computer Science & Engineering Technology (IJCSET)

[3] Saba Arshad, Abid Khan, Munam Ali Shah, Mansoor Ahmed (2016), Android Malware Detection & Protection: A Survey, (IJACSA) International Journal of Advanced Computer Science and Applications.

[4] Z. Bakdash, Steve Hutchinson, Erin G. Zaroukian, Laura R. Marusich, Saravanan Thirumuruganathan , Charmaine Sample, Blaine Hoffman , and Gautam Das, Malware in the future? forecasting of analyst detection of cyber events Jonathan, University of Texas Dallas Dallas, TX, USA

[5] Niklas Lavesson, Martin Boldt, Paul Davidsson, Andreas Jacobsson (2009), Learning to detect spyware using end user license agreements, Springer.

[6] Androutsopoulos I, Paliouras G, Karkaletsis V, Sakkis G, Spyropoulos CD, Stamatopoulos P (2000), Learning to filter spam E-mail: a comparison of a naive bayesian and a memory-based approach.

[7] Kirti Mathur (2013), A Survey on Techniques in Detection and Analyzing Malware Executables, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 4.

[8] M. G. Schultz, E. Eskin, E. Zadok and S. J. Stolfo (2001), Data Mining Methods for Detection of New Malicious Executables, Proceedings of the 2001 IEEE Symposium on Security and Privacy, IEEE Computer Society.

[9] Parisa Bahraminikoo (2012),Utilization Data Mining to Detect Spyware, IOSR Journal of Computer Engineering (IOSRJCE), Volume 4, Issue 3.

[10] Robert Moskovitch (2012), Detecting unknown malicious code by applying classification techniques on OpCode patterns, Springer-Verlag http://link.springer.com/article/10.1186%2F2190-8532-1-1.

[11]Milan Jain1, Punam Bajaj (2014), Techniques in Detection and Analyzing Malware Executables: A Review, International Journal of Computer Science and Mobile Computing.