RESEARCH ARTICLE OPEN ACCESS

# Implementation of Spyware Detection Using Data Mining With Decision Tree Algorithm

Leena T. Patil[#1], Shamal S. Pawar[#2],Shrutika N. Lad[#3,]Nilambari Joshi[*4]

[#]Dept. of Computer Engineering, K. J. Somaiya Institute of Engineering and Information Technology, Mumbai University, Maharashtra (India)

[*]Assistant Profesoor, Dept. of Computer Engineering, K. J. Somaiya Institute of Engineering and Information Technology, Mumbai University, Maharashtra (India)

***Abstract***

Any software that monitors user behavior, or gathers information about the user without adequate notice, consent, or control from the user. " Spyware represents a serious threat to confidentiality since it may result in loss of control over private data for computer users. This type of software might collect the data and send it to a third party without informed user consent. Our approach is inspired by DM-based malicious code detectors, which are known to work well for detecting viruses and similar software We extract binary features, called n-grams, from both spyware and software and apply to train classifiers that are able to classify unknown binaries by analyzing extracted n-grams.

***Keywords****:* N-Gram, ARFF (Attribute Relation File Format) CFBE (Common Feature-based Extraction), FBFE (Frequency-based Feature Extraction), RFS (Reduced Feature Sets).

## I. INTRODUCTION

Some Programs include spyware, adware, Trojans and backdoors. They may compromise confidentiality, integrity, and availability of the system and may obtain sensitive information without informed user. Example of spyware: data collected by spyware may be used for customized ads spread through adware to an individual user. Originally, viruses represented the only major malicious threats to computer carried out in order to successfully detect and remove viruses from computer systems. However, a more recent type of malicious threat is represented by spyware. According to the Department of Computer Science and Engineering at the University of Washington, spyware is defined as "software that gathers information about use of a computer, usually without the knowledge of the owner of the computer, and relays the information across the Internet to a third party location."

Spyware differ from regular viruses that's why they are not able to detect through normal antivirus software. Traditionally two approaches have been presented for the purpose of spyware detection: Signature-based Detection and Heuristic-based Detection. These approaches perform well against known Spyware but have not been proven to be successful at detecting new spyware. Even if users have anti-virus software installed, it may not be helpful against spyware until it is designed particularly for this threat.

Our paper presents a spyware detection method inspired by data mining-based malicious code detection. In this method, binary features are extracted from executable files. A feature reduction method is then used to obtain a subset of data which is further used as a training set for automatically generating classifiers. Many data mining classification techniques are available but proposed system uses decision tree algorithm for making decision related to testing exe file. Decision tree give result based on training dataset that scanning file is spyware or not.

## II. BACKGROUND

The term spyware first appeared in a Usenet post on October 16, 1995 about a piece of hardware that could be used for espionage. In 2001, the use of data mining was investigated as an approach for detecting malware and this attempt attracted the attention of many researchers. Since then, several experiments have been performed to investigate the detection of traditional malicious software such as viruses, worms, and so forth, by using data mining technologies. Data mining is the process of analyzing electronically stored data by automatically searching for patterns. Machine Learning algorithms are commonly used to detect new patterns or relations in data, which are further used to develop a model, i.e., a classifier or a regression function. Learning algorithms have been used widely for different data mining problems to detect patterns and to find correlations between data instances and attributes.

## III. DESIGN

### 1. Goals of Application:

This application allows us to detect new or unseen spyware. It is very simple application as there is no need of internet for update the database of spyware because spyware has feature that it always have some features of previous spywares. Another feature is we can delete list of spyware so it doesn't affect other non-infected files.

## 2. SYSTEM FLOW

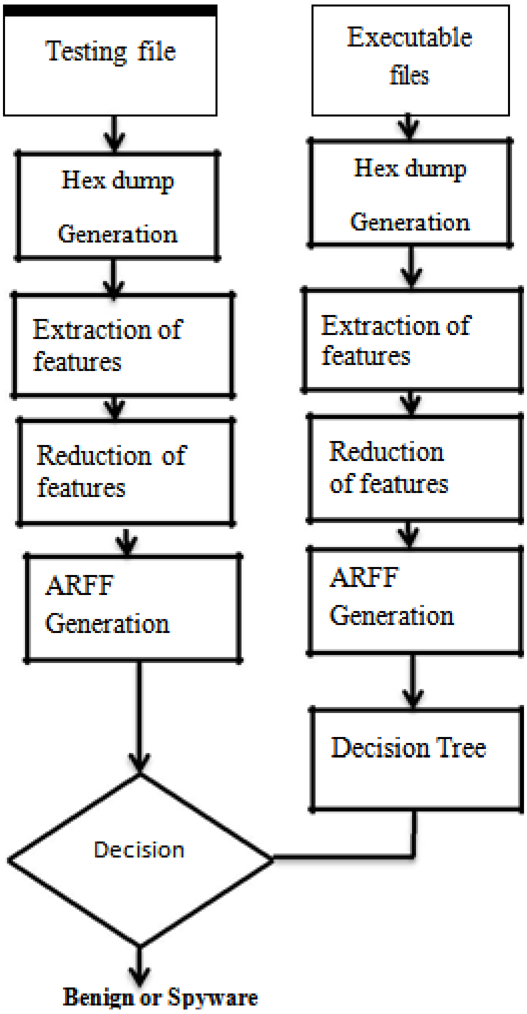

Fig 1. : SYSTEM VIEW

## IV. IMPLEMENTATION DETAILS

### *A. Data Collection*

Data set consists of set benign and malicious files. The benign files were collected from Download.com, which certifies the files to be free from spyware also System files in windows O.S and http://vl.netlux.org/v1.phplink. The spyware files were downloaded from the links provided by SpywareGuide.com, which hosts information about different types of spyware and other types of malicious software. Also malicious files were downloaded from VX heaven website and http://vx.netlux.org.

### *B. Byte Sequence Generation*

Use of xxd, which is a UNIX-based utility for generating hexadecimal dumps of the binary files. From these hexadecimal dumps we may then extract byte sequences in terms of *n*-grams of different sizes and these N-grams are saved into text file which is system generated.

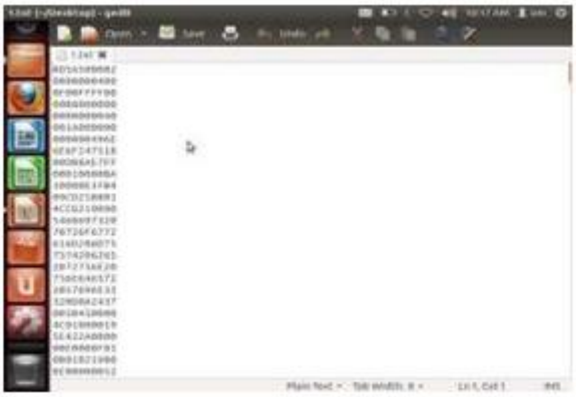

Fig. 2 shows the N-Gram Generation of file

### C. Feature Extraction

Extract the features by using two different approaches: the Common Feature-based Extraction (CFBE) and the Frequency-based Feature Extraction (FBFE). Both methods are used to obtain Reduced Feature Sets (RFSs) which are then used to generate the ARFF files.

Fig. 3 shows the sorted N-Gram of file

Feature Reduction-

Redundant features are eliminated in this step. Data Set is reduced in size at this phase. Removing redundant hex dump from hex dump file.

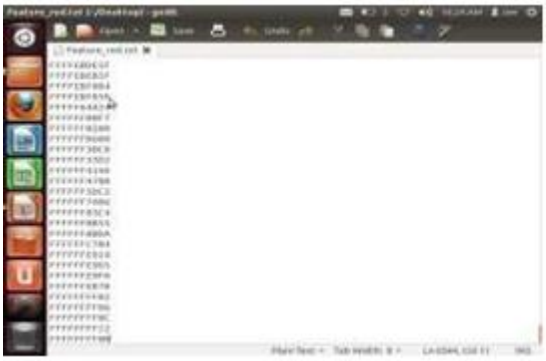

Fig. 4 shows the reduced N-Gram

**D. ARFF Generation**

Two ARFF databases based on frequency and common features were generated. All input attributes in the data set are represented by Booleans, i.e., either a certain n-gram or the n-grams within a certain frequency range are represented by either 1 or 0 (present or absent).

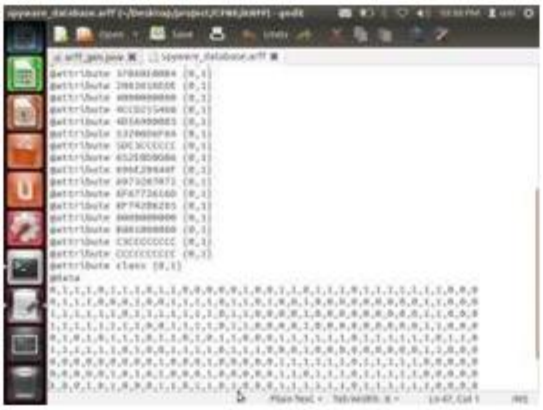

Fig. 5 shows the ARFF training file

**E. ARFF Generation for scanning file**

ARFF file of scanning executable file is generated. Based on the ARFF training file value of last? Position is placed. If value return by decision tree is 1 then file is spyware and if 0 then it is benign.

Fig. 6 shows the ARFF Scanning files and output is generated with binary values 0 and 1.

Fig. 6 shows the ARFF Scanning file

F. Decision Tree

Decision tree algorithm is a data mining induction techniques that recursively partitions a data set of records. Decision trees can handle high dimensional data, the learning and classification steps are simple and fast [8]. Decision tree is a flow-chart-like tree structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes [8].

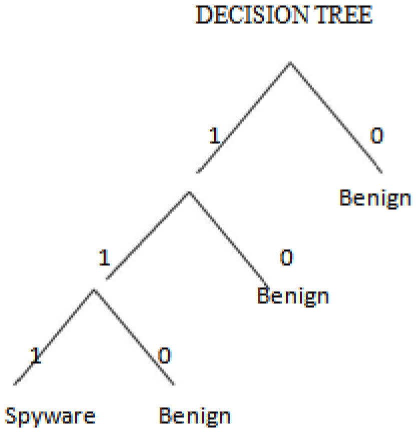

Fig. 7 Decision Tree based on feature

The results are generated depending on the decision tree and calculate no. of spyware files and give result.

Fig-7 shows how the decision tree is generated based on the feature. If pattern of new arrived file is matched with the pattern which is generated by decision tree i.e. if new arrived file contain 111 patterns which is exactly match with decision tree pattern then it is spyware file as well as based on these patterns it is decided that whether it is benign or spyware file.

Fig 8 : User Interface

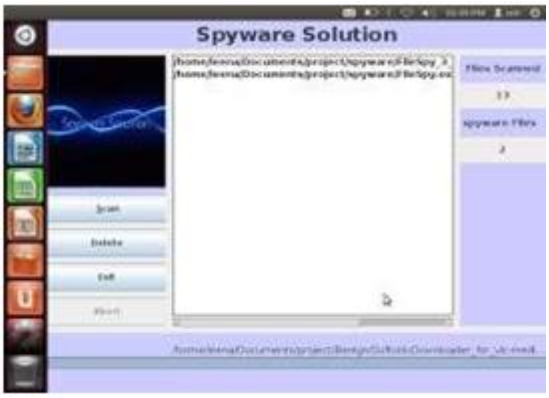

Fig 9: Spyware Detected in Application

Fig. 8 Shows how Application Interface look like. Here scan button option is provided and by clicking on it you can select any drive or folder for scan and when you open it scan will automatically start. If you want to stop scan click on Abort button and for deleting scanned spyware click on spyware and click delete button.

Fig. 9 Shows spyware files which are detected by this application.

## V. CONCLUSION FUTURE SCOPE

Detection rate for data mining methods will be more than signature-based method. Reduce the false negative alert Detecting known as well as new unseen spyware.

1. Make a project platform independent

Develop module for equivalent work of XXD tool which make a project platform independent.

2. The Application of Data Mining Technology in Network Security Management

In compression rate, false alarm rate, construction rate and scene detection rate, and significantly improves the accuracy, intelligence and adaptability of the network security management.

## REFERENCES

[1] http://www.security-informatics.com/.
[2] Download, http://download.com.
[3] Spyware Guide, http://Spywareguide.com.
[4] Linux/Unix Command: xxd, http://linux.about.com/library/cmd/blcmdl1_xxd.htm.
[5] http://vl.netlux.org/v1.php
[6] M. Boldt and B. Carlsson, ““Privacy-invasive software and preventive mechanisms,”” 2$^{nd}$ International Conference on Systems and Networks Communications, (ICSNC 2006), Oct. 28- Nov.2, IEEE Computer Society.
[7] www.researchgate.net/.../221548426_**Detection**_of_**Spyware**_by_**Mining**
[8] J. Han and M. Kamber, *Data Mining: Concepts andTechniques*.: The Morgan Kaufmann, 2006