International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 3, Issue 3, March 2016

Malware Categorization Based On Improved Clustering Technique Using Feature Selection in Clustering Process

[1] Ms. Shubhangi Y. Chaware [2] Mr. Sagar G. Chunne
[1][2] M.Tech, Department of Computer Science & Engineering RGPV Bhopal
[1] shubhangi.chaware@yahoo.co.in [2] sagar_chunne@yahoo.com[2]

Abstract: **Malware is basically malicious software or programs which are a major challenge or major threats for the computer and different computer applications in the field of IT and cyber security. Traditional anti-viral packages and their upgrades are typically released only after the malware's key characteristics have been identified through infection. The most common detection method is the signature based detection that makes the core of every commercial anti-virus program. To avoid detection by the traditional signature based algorithms, a number of stealth techniques have been developed by the malware writers. The inability of traditional signature based detection approaches to catch these new breed of malwares has shifted the focus of malware research to find more generalized and scalable features that can identify malicious behavior as a process instead of a single static signature.**

The goal of proposed work is to create a hybrid model for feature selection and Malware categorization. Feature selection is important issue in Malware categorization. The selection of feature in attack attribute sand normal traffic attribute is challenging task. For the test of our hybrid method, we used DARPA KDDCUP99 dataset. This data set basically set of network Malware and host Malware data. This data provided by UCI machine learning website. Our proposed method compare with exiting ISMCS, HC and KM technique and getting better result such as F-measure, precision and recall value.

***Keywords*— Subspace clustering, F-measure, Recall, Precision**

I. INTRODUCTION

Automated malware categorization methods and an industry-wide categorization convention have been the computer security topics that are of great interest. However, malware categorization cannot be reliable unless the virus analysts can classify a new sample to a certain family in a reasonable amount of time. Malware is defined as computer software that has been explicitly designed to harm computers or networks. In the past, malware creators were motivated mainly by fame or glory. Most current malware, however, is economically motivated. Commercial anti-malware solutions rely on a signature database for detection. An example of signature is a sequence of bytes that is always present within a malicious executable and within the files already infected by that malware. In order to determine a file signature for a new malicious executable and to advise a suitable solution for it, specialists must wait until the new malicious executable has damaged several computers or networks. In this way, suspect files can be analyzed by comparing bytes with the list of signatures. If a match is found, the file under test will be identified as a malicious executable. This approach Has proved to be effective when the threats are known beforehand.

Malware writers use code obfuscation techniques [5] to hide the actual behavior of their malicious creations. Examples of these obfuscation algorithms include garbage insertion, which consists on adding sequences which do not modify the behavior of the program (e.g., nop instructions1); code reordering, which changes the order of program instructions and variable renaming; which replaces a variable identifier with another one [9]. Data-mining-based approaches rely on datasets that include several characteristic features for of both malicious samples and benign software to build classification tools that detect malware in the wild. Machine-learning algorithms can be classified into three different types: supervised learning, unsupervised learning and semi-supervised learning algorithms. First supervised machine-learning algorithms, or classifying algorithms, require the training dataset to be properly labeled (in our case, knowing whether an instance is malware) [27]. Second, unsupervised machine-learning algorithms, or clustering algorithms, try to assess how data are organized into different groups called clusters. In this type of machine-

learning, data do not need to be labelled. Finally, semi-supervised machine-learning algorithms use a mixture of both labelled and unlabeled data in order to build models, thus improving the accuracy of unsupervised methods.

II. MALWARE DETECTION TECHNIQUES

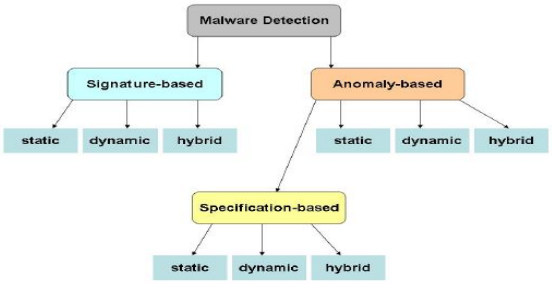


Figure 1.1: A classification of malware detection techniques.

1. System Model

This system is to accurately detect new malware (unknown malware) binaries using a number of data mining techniques. The architecture of malware detection system consists of three main modules: (1) PE-Miner, (2) feature selection and data transformation, and (3) learning algorithms.

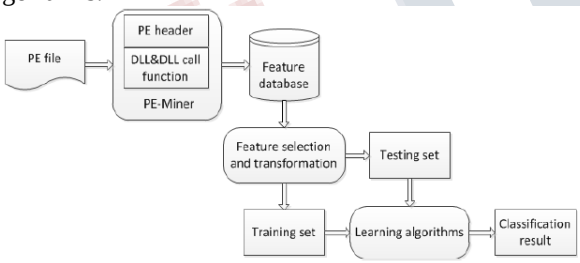


Figure 2.1: Architecture of the Malware Detection System.

2.1 Malware Categories And Behaviour

Most malware families have similar behaviour and properties, which the majority of scanners use as signatures to detect malware variants. For instance, one of the properties of a worm is self-replication – a worm tries to spread by simply copying itself to a host machine through the communication channels of other infected machines. On the other hand, a virus will attempt to spread by a carrier such as an infected file or a media drive. In the following we will examine some common environments and the behaviour of malware.

2.2 Malware Environments

In order for malware to perform its malicious functionality and to infect other victims, some components or resources should exist. Malware writers usually develop their code for a particular operating system. For instance, Win32 viruses are effective against Microsoft Windows and may not work on other operating systems. Moreover, a malware may require that some particular applications are running on the victim system in order to be effective. For example, some virus attacks are only effective if a scripting language such as Microsoft VB script or JavaScript (.vbs, .js, etc.) files can execute on the local machine.

2.3 Means Of Infection

Malware uses common methods of transmission between computer systems. One of the traditional methods, and the easiest, of transmit-ting malicious programs is via external media such as USB devices and memory disks; however, the rate of spreading malware using this method is considered low compared to other methods such as through networked systems. Malware writers find networked computer systems an excellent environment to replicate and spread their viruses and worms; therefore, inadequate security on a network means that a large number of systems are vulnerable to malicious attacks. Another means of malware infection between computer systems is electronic mail (e-mail). Malicious code can spread easily as a file attachment sent with an e-mail message to as many as possible e-mail users. This type of spreading mechanism requires only a little effort from malware writers to make successful attacks. E-mail-based malware falls into two categories: mailer and mass mailer malware. The first category uses mail software such as Microsoft Outlook; the list of e-mail addresses on the host machine is used by the virus to e-mail itself to other users. The second category uses its own SMTP engine to send malicious code to many e-mail addresses.

III. PREVIOUS WORK

Several features have been proposed for representation of malware. N-Grams [3-5], the length of the functions and the frequency of the function length [6] been proposed as an effective methods to represent malware. But these approaches also have some disadvantages: N-Grams extracts a lot of interfering data from the original file, and transforms the file into a very high dimension feature space; the length of a function or the frequency of the function length is not an interpretable feature, which is an obstacle for signature generation in malware categorization.

Comparing to the ordinary clustering algorithms, subspace clustering algorithm can automatically choose the most important dimensions for every cluster and find the hidden cluster in any subspace. So, the main phases of subspace clustering algorithm are searching the potential subspace and the corresponding cluster. According to the

ways how the subspaces are identified, there are two categories of subspace clustering algorithm: hard subspace clustering algorithm and soft subspace clustering algorithm. The first one uses a subset of the entire dimension as subspace, the second one assigns a weight for every dimension.

On experimenting with different dataset, the number of normal/abnormal packets is being monitor. We have examined five different dataset in our experiment, with each having corresponding number of rejected or normal packets. In our conducted test the packets could either fall under normal packet type or in the category of attack (DOS, R2L.U2R.PROB).

We have supervised on data set with each 7000 instances of data under .the result of predicted normal and abnormal data is form of confusion matrix.

TP: True Positive TN: True Negative
FP: False Positive FN: False Negative

IV. PERFORMANCE PARAMETERS

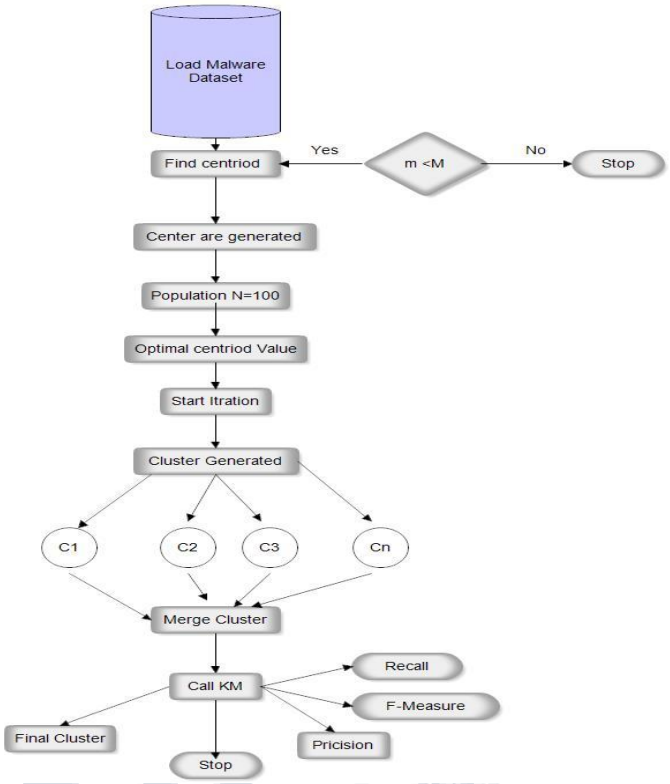
Earlier application of isolated feature reduction on dataset has much greater Accuracy, than later by integrating both feature reduction and Improved ID3 Methods. Also there is a considerable enhancement in the true positive and true negative detection ratio and minimizes in false positive and false negative ratio .Thus this gives the direct improvised accuracy in the result. Basis the result of confusion matrix (true positive, true negative, false positive, false negative).We are showing the consequence for the following parameters i.e. - Accuracy, Precision, Recall for data sets.

Precision- Precision measures the proportion of predicted positives/negatives which are actually positive/negative.

Recall -It is the proportion of actual positives/negatives which are predicted positive/negative.

$$\text{Precision} = \frac{TP}{TP+FP}; \text{Recall} = \frac{TP}{TP+FN}$$

1. *Proposed Model*



4.1 Processing Step Of Algorithm

Step1: In first step the data are load the data and randomly assign the center point of cluster. The randomly generated centers are passes through the genetic algorithm, the genetic algorithm assigned the population set N=100;

Step 2: After the process of population set the selection of center point find. The center point finds the position location of Centered value.

Step 3: The variable of auto level chose by the fitness constraints function, the fitness constraints decide the selection of parameter value for centroid ratio.

Step4: The selection of centroid process done and after that the processing of data are from and iteration process are done.

Step5: The process of iteration generate the number of maximum cluster and merging process are done.

Step 6 finally calls KM.

2. *Experimental Result*

Input value	Method Name	F-Measure	Precision	Recall
K=1	ISMCS	89.7999	86.2737	83.8146
	HC	95.3094	88.7004	84.8546
	KM	96.3094	89.8004	88.0657
	PROPOSED	97.3094	89.8004	88.0657

Input value	Method Name	F-Measure	Precision	Recall
K=2	ISMCS	90.469	86.9424	84.484
	HC	95.9788	89.3697	85.5239
	KM	96.9788	90.4697	88.735
	PROPOSED	97.9788	90.4697	88.735

Figure 5.1 Comparative result for input value is 1 and 2 using different number of malware detection methods.

Input value	Method Name	F-Measure	Precision	Recall
K=3	ISMCS	90.7122	87.1854	84.7269
	HC	96.2217	89.6121	85.7669
	KM	97.2217	90.7127	88.978
	PROPOSED	98.2217	90.7127	88.978

Input Value	Method Name	F-Measure	Precision	Recall
K=4	ISMCS	90.8344	87.3096	84.8492
	HC	96.3439	89.7349	85.8891
	KM	97.3439	90.8349	89.1002
	PROPOSED	98.3439	90.8349	89.1002

Figure 5.2 Comparative result for input value is 3 and 4 using different number of malware detection methods.

V. CONCLUSION

In this paper we perform experimental process of proposed Malware detection and categorization algorithm. The proposed method implements in mat lab 7.14.0 and tested with very reputed data set from KDD Cup data set . In the research work, I have measured Precision and recall for the ISMCS, HC, KM and proposed method. To evaluate these performance parameters I have used KDDCUP99 datasets from UCI machine learning repository [43].

REFERENCES

[1] Kai Huang , Yanfang Ye , Qinshan Jiang "ISMCS: An Intelligent Instruction Sequence based Malware Categorization Sy stem" the National Science Foundation of China

[2]Tawfeeq S. Barhoom, Hanaa A. Qeshta "Adaptive Worm Detection Model
Based on Multi classifiers "Palestinian International Conference on Information and Communication Technology 2013, Pp57-65.

[3]StanislavPonomarev, Jan Durand, Nathan Wallace, Travis Atkison" Evaluation of Random Projection for Malware Classification" 2013, Pp 68-73

[4]Aiman A. Abu Samra, KangbinYim , Osama A. Ghanem, " Analysis of Clustering Technique in Android Malware Detection" Seventh International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing 2013, Pp 729-733

[5] Jonghoon Kwon, Heejo Lee ," BinGraph: Discovering Mutant Malware using
Hierarchical Semantic Signatures" *7th International Conference on Malicious and Unwanted Software,* 2012, Pp 104-111.

[6] P.R.LakshmiEswari , N.Sarat Chandra Babu "A Practical Business Security Framework to Combat Malware Threat "World Congress on Internet Security, 2012, Pp 77-80.

[7] Ahmed F.Shosha, Chen-Ching Liu, PavelGladyshev, Marcus Matten "Evasion-Resistant Malware Signature Based on Profiling Kernel Data Structure Objects" 2012, 7th International Conference on Risks and Security of Internet and Systems (CRiSIS)

[8] Vinod P., V.Laxmi , M.S.Gaur , GrijeshChauhan "MOMENTUM :MetamOrphicMalware Exploration Techniques Using MSAsignatures"International Conference on Innovations in Information Technology (IIT), 2012, Pp 232-237.

[9] HiraAgrawal, Lisa Bahler, Josephine Micallef, Shane Snyder, and AlexandrVirodov
"Detection of Global, Metamorphic Malware Variants Using Control and Data Flow Analysis" 2013 ,Pp 1-6.

[10] Yanfang Ye, Tao Li, Qingshan Jiang, and Youyu Wang," CIMDS: Adapting Postprocessing Techniques of Associative Classification for Malware Detection" IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS, VOL. 40, NO. 3 ,2010,Pp 298-307

AUTHORS PROFILE

Ms.Shubhangi Y. Chaware has received his Bachelor of Engineering degree in Information Technology from Radhikatai Pandav college of Engineering College (RTMNU) Nagpur in the year 2008. At present pursuing M.Tech. with the specialization Software Engineering in Patel Institute of Technology ,Bhopal. Area of Computer security, Network Security.

Mr. Sagar G. Chunne has received his Bachelor of Engineering degree in Computer Science & Engineering from BAMU Aurangabad in the year 2008. At present pursuing M.Tech. with the specialization Information Technology from NRI institute of Technology science and research ,Bhopal. Area of Computer security, Computer Network Security.