**Significance of Feature Extraction in Classification of
Rootkit sub-families of Malware**

Prasenjit Das1, Chetan Sharma2

Chitkara University

Himachal Pradesh (India)

**Abstract**: Modern malware sharing common code within sub-families , in order to nullify the anti-malware has redundant features in them. These features are added in the code to obfuscate the anti-malware. Feature selection techniques remove these redundant insignificant features from the binary exe. Thus leading to better classification results. We have shown experimentally that the classification of two classes of rootkit family of malware generate better accuracy when feature selection techniques have been applied . The accuracy of 84.17% as against 66.67% when feature selection is not applied, shows the significance of feature selection in malware classification.

**Keywords:** Feature selection, feature extraction, classification, malware.

1 Introduction

Data or information that is stored on devices connected over a network or in a standalone system is always a potential target of being attacked by malicious software which try to steal the data or corrupt the same. In general, most of the malwares (malicious software’s) are commonly referred to as a virus. Considering the detrimental effects of these malicious software’s, organization largely rely on intermediate intrusion detection systems to detect and block malwares from entering and causing damage. It should be noted that not all ‘virus attacks’ are caused by viruses, requiring intrusion detection systems to be constantly up-to-date with known malware families and their corresponding morphological changes as these malwares constantly evolve, rendering detection ineffective. Malwares are broadly categorized based on the criteria such as: the manner by which these malware enters into the system, the manner by which they execute, and by the associated damage that they can cause.

Reported by Microsoft TM, there exists close to two hundred families of malwares. Some of the major malware families are *Computer Virus, Trojan horse, Worms, Spyware, rootkits.* Furthermore, as per this study conducted in 2006 by Microsoft, approximately 75% of the malware detected belong to 25 prominent families of malwares and their variants [1]. This raises the question that if the variants of malwares are indeed related to common families, then they must share common behavior. This then logically implies that the code of the malwares must share something in common. A rootkit is a malware which is designed to hide the fact that a system has been infected by the malware by changing its important executable files required by the operating system to deliver the desired outcome. Rootkit is different from other malwares as this malware attacks the operating system files.

The motivation of this work is to explore the significance of feature extraction in rootkit malware sub-families, and to analyze the effect on classification of rootkit families accurately. We hypothesize that there are redundant features in root kit families of malware and if the same are not discarded while classifying the class of malware, the accuracy is reduced significantly. Also the existence of redundant features does hinder the classification amongst the families of malwares. In the present work we have tried to establish that the classification percentage improves when we are able to remove the redundant features by applying feature extraction and feature selection methods. Feature extraction removes the redundant features in the data set which is an hindrance in the classification process. The percentage of accuracy improves significantly after we apply the feature selection and extraction techniques as the redundant features are removed from the data set. The following manuscript is structured as follows: Section 2 provides an overview of recent advances in malware detection techniques. Section 3 explains the feature extraction and feature selection along with the classification carried out. Our results and discussions are explained in section 4, followed by conclusion in section 5.