Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 125 (2018) 346–356

Procedia Computer Science

www.elsevier.com/locate/procedia

6th International Conference on Smart Computing and Communications, ICSCC 2017, 7-8 December 2017, Kurukshetra, India

Performance Evaluation of Filter-based Feature Selection Techniques in Classifying Portable Executable Files

Shiva Darshan S.L.* and Jaidhar C.D.

Department of Information Technology, National Institute of Technology Karnataka, Surathkal, Mangalore, India

Abstract

The dimensionality of the feature space exhibits a significant effect on the processing time and predictive performance of the Malware Detection Systems (MDS). Therefore, the selection of relevant features is crucial for the classification process. Feature Selection Technique (FST) is a prominent solution that effectively reduces the dimensionality of the feature space by identifying and neglecting noisy or irrelevant features from the original feature space. The significant features recommended by FST uplift the malware detection rate. This paper provides the performance analysis of four chosen filter-based FSTs and their impact on the classifier decision. FSTs such as Distinguishing Feature Selector (DFS), Mutual Information (MI), Categorical Proportional Difference (CPD), and Darmstadt Indexing Approach (DIA) have been used in this work and their efficiency has been evaluated using different datasets, various feature-length, classifiers, and success measures. The experimental results explicitly indicate that DFS and MI offer a competitive performance in terms of better detection accuracy and that the efficiency of the classifiers does not decline on both the balanced and unbalanced datasets.

Peer-review under responsibility of the scientific committee of the 6th International Conference on Smart Computing and Communications

Keywords: Feature Selection Technique; Malware; Malware Detection System; Machine Learning; Portable Executable Files;

1. Introduction

Malware is a computer program designed to harm the host system without the user consent. It can morph itself to gain control of the host system, whereby it can access the system level operations in multiple dimensions. It can easily evade the existing detection techniques using various modern obfuscation characteristics. It has grown drastically and has emerged as an insurmountable issue for many anti-malware defensive solutions. Therefore, there is an immediate need of an intricate MDS [9] to resist the attacks caused by such malware.

* Corresponding author.
E-mail address: it15f02.shivadarshan@nitk.edu.in

1877-0509
Peer-review under responsibility of the scientific committee of the 6th International Conference on Smart Computing and Communications
10.1016/j.procs.2017.12.046

Generally, traditional malware defensive solutions rely on signature-based detection technique and thus, are vulnerable to unknown malware, if the malware database has not been updated. It extracts static features from the executable file, including binary sequences, function calls, and any other information to determine whether the executable file is a malware. These techniques are said to be more resilient to the malevolent activity of malware, but are easily disrupted by the obfuscation characteristics [14].

The behavioural-based detection technique detects the malware by monitoring the behaviour of the executable file during its runtime [6]. It isolates the malware in an environment called the sandbox [18] and records behaviours such as API calls, system calls, or any other function-based calls triggered upon the operating systems. Thus, it provides a new perspective to analyze the unknown malware. However, it fails to balance between False Positive Rate (FPR) and malware detection rate.

The heuristic-based detection techniques employ the machine learning method to learn the behaviour of an executable file. To detect malware, it deliberately analyzes features such as system calls, API calls, Opcodes, and structural information like header information, etc. [5]. However, in a real scenario, it becomes tedious to examine all the recorded features to acquire the most predominant features for the purpose of the classification operation. Under such circumstances, FST plays a vital role in minimizing the dimensionality of the original feature space and boosts the predictive performance of the classifiers [13]. Information Gain [19], MI [12], Fisher Score [20], Chi-square [7], etc. are examples of FSTs.

In this paper, an MDS has been designed that detects malware based on the extracted information related to Portable Executable Optional Header Fields (PEOHF). Moreover, our prime focus is on the performance analysis of FSTs that are capable of selecting the most relevant features, which are crucial in discriminating between benign and malware PE files. We have employed Single-Stage-Feature-Selector that acquires significant features by adapting the filter-based FST. Further, we compute the score for those extracted PEOHF (features) and then, choose the topmost features as predominant features based on the highest score. From the experimental results, we observe that the features suggested by the DFS and MI were successful in attaining malware detection accuracy of 98.677% for the Balanced Dataset (BD) and 99.308% for the Unbalanced Dataset (UBD) under the 10-fold cross-validation test. Finally, the accuracy shown by the classifiers for both the BD and UBD was considered and the difference was computed to convey that the efficiency of the classifiers does not change much on the BD and the UBD.

The major contributions of this paper are summarized as follows:

- In this work, MDS is designed, implemented, and evaluated using real-world malware samples. The MDS is proficient in precisely distinguishing between malware and benign PE files based on the features recommended by the Single-Stage-Feature-Selector.
- We have employed different FSTs such as DFS, MI, CPD, and DIA to select a compact set of the most significant features to boost the efficiency of the classifier for better accuracy. Four filter-based FSTs were adopted to measure the comparison with the intention of identifying the better one. To evaluate the performance of the different FSTs, two sets of experiments were conducted on the BD and the UBD.
- The experimental results demonstrated that the features recommended by the DFS and MI were successful in achieving malware detection rate of 98.677% for the BD and 99.308% for the UBD under the 10-fold cross-validation test.
- Lastly, evaluation on the BD and the UBD was made to measure the accuracy variation between them. The results clearly indicated that the accuracy difference range of <1% was not of much affect on the efficiency of the classifiers.

The rest of this paper is organized as follows. In Section 2, we study the background of the Portable Executable files. In Section 3, we review earlier research work on filter-based FSTs used in classification. Section 4 elucidates the methodology to effectuate the performance analysis of the filter-based FSTs. Section 5 provides a brief description of the filter-based FSTs used in our experimental work. The obtained empirical results are presented in Section 6. Finally, the conclusion is summarized in Section 7.