Permission Based in Android Malware Classification

Zolidah KASIRAN*, Norkhushaini AWANG and Fatin Nurhanani RUSLI
Faculty of Computer & Mathematical Sciences Universiti Teknologi MARA 40450 Shah Alam
*Corresponding author

Keywords: Malware detection, Android malware classification.

Abstract. Android malware is growing in such an exponential pace which lead out for automated tools that can aid the malware analyst in analysing the behaviours of new malicious applications. This project had proposed clustering in intrusion detection method using hybrid learning approaches combining K-Means clustering and Naïve Bayes classification had been proposed. The result had shown the improved false rate alarm in malware detection.

Introduction

Android malware is growing in such an exponential pace which lead out for automated tools that can aid the malware analyst in analysing the behaviours of new malicious applications. Malware threats against the smartphone users are increasing especially on Android users. Due to the lack of awareness on the users, the Android malware was spreading during permission stage [1].

According to [2], DroidKungfu was one of the most harmful Android malware and it has several names. Though there are many clustering based method used in detecting malware, the issue of false alarm rate and accuracy is still a topic of discussion. Besides, [3] also concluded there are not even a single clustering method yield low false alarm rates with high detection rates. Few researcher found that the accuracy of clustering could be improved by using K-Means [4]. Though another researcher found combination of techniques could improved the precision and reduces false alarm rates compared to single classification [5,6, 7].

Therefore, clustering intrusion detection method using hybrid learning approaches combining K-Means clustering and Naïve Bayes classification had been proposed since K-Means is lightweight, easy to implement and fast-iterative algorithms compared to other clustering methods [8].

Related Works

Clustering

Clustering is one of the techniques that help to categorizing Android malware. It is one of a common data mining and statistical data analysis techniques that have been widely used by many researchers. Researcher name [8] had used this techniques to detect malware in androids platform and reported that the techniques is possible in classifying malware. Another researcher [9] used clustering techniques and compared with Mini Batch K-means algorithm in analyzing network traffic.

Classification

Classification is a data mining function that assigned items in a collection to target categories, families and classes. Few classification techniques such as; Naïve Bayes, and Supported Vector Machine(SVM), are commonly employed in classification research.

Naïve Bayes. Naïve Bayer’s Classification had been studied widely and remained a common method with appropriate pre-processing. Naïve Bayes uses Bayesian Theorem that are used in many field of classification study, [10] and [11]. Naïve Bayes is a simple technique to produce classifiers, assign class labels to problem instances and it represented as vectors of feature values.
**Research Method**

This project had adopted three main phases as shown in Figure 1 which was Feature Extraction, Clustering & Classification. The first phase was feature extraction of the permission from the manifest file MalGenome datasets. The result from the data extraction was passed to the clustering phase adopting the K-Means clustering technique. The last phase was classification using Naïve Bayes.

![Diagram](image)

Figure 1. Reseach Flows Phase.

In the feature extraction phase, the static analysis was done manually by converting the apk file into a normal folder that Windows can work on Malware dataset. Then, the apk files can be opened and viewed. There are a few files contains in each apk file which are “Android Manifest.xml”, “classes.dex” and yml files. From this extracting files, then the static analysis was started by observed all xml files from all differents 427 apk files choosen. Subsequently, from xml files a collected of data with permission list on the apk files then accumulated in comma demilated (csv) files. The samples of Android malware dataset were in apk which only readable in Android environment. Each sample provided was in sha256 unique names. In the manifest file, the application was asking permission on “READ_CONTACTS”, “READ_PHONE_STATE”, “RECEIVE_SMS” and “SEND_SMS”. The permissions were requested to read the contact information, write, receive and send text messages which looks kind of suspicious.

Afterwards, the csv file was saved into a.rff file format through .arff viewer in Weka. Then, all experimentation was done and K-Means clustering are used to cluster the data into clustered sets. In this hybrid learning approach, similar data was grouped based on their behaviors by applying K-Means clustering as a first classification step. Then, Naïve Bayes was used to classified the resulting clusters as a second classification step.

**Results and Analysis**

The fist clustering method was K-Means clustering methodsthat yield into clustered sets. The n the clustered set were classified the using Naïve Bayes classification. K-Means clustering result with value of cluster, k=2 which cluster 0 (C0) as malware, cluster 1 (C1) as non-malware apps. After the cluster sets obtained, the data then labelled into two definite group which are malware (A) and non-malware (B). Another parameter that is commonly used is Receiver Operating Characteristics (ROC) Curve. ROC Curve is a commonly used graph that summarizes the performance of a classifier over all possible thresholds. It is generated by plotting the True Positive Rate (y-axis) against the False Positive Rate (x-axis) as the threshold adapted for assigning observations to a given class.
Table 1. Results of Hybrid Learning Approaches for Training Dataset.

<table>
<thead>
<tr>
<th>Hybrid Method</th>
<th>True Positive Rate (TPR)</th>
<th>False Positive Rate (FPR)</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
<th>Accuracy</th>
<th>ROC Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>KM+NB</td>
<td>0.989</td>
<td>0.001</td>
<td>0.991</td>
<td>0.989</td>
<td>0.989</td>
<td>98.8</td>
<td>0.889</td>
</tr>
<tr>
<td>KM+1R</td>
<td>0.996</td>
<td>0.000</td>
<td>0.997</td>
<td>0.996</td>
<td>0.996</td>
<td>99.6</td>
<td>0.998</td>
</tr>
<tr>
<td>KM+J48</td>
<td>0.993</td>
<td>0.050</td>
<td>0.993</td>
<td>0.993</td>
<td>0.993</td>
<td>99.3</td>
<td>0.956</td>
</tr>
</tbody>
</table>

Table 1 and Table 2 shows the results from the three experiments where the highest accuracy was from the results of experiment 1. Experiment 1 was using K-Means clustering and Naive Bayes (KM+NB) classification yield 99.4% of accuracy compared with two other hybrid methods.

Table 2. Results of Hybrid Learning Approaches for Testing Dataset.

<table>
<thead>
<tr>
<th>Hybrid Method</th>
<th>True Positive Rate (TPR)</th>
<th>False Positive Rate (FPR)</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
<th>Accuracy</th>
<th>ROC Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>KM+NB</td>
<td>0.994</td>
<td>0.00</td>
<td>0.995</td>
<td>0.994</td>
<td>0.995</td>
<td>99.4</td>
<td>1.0</td>
</tr>
<tr>
<td>KM+1R</td>
<td>0.994</td>
<td>0.00</td>
<td>0.995</td>
<td>0.994</td>
<td>0.995</td>
<td>99.4</td>
<td>0.997</td>
</tr>
<tr>
<td>KM+J48</td>
<td>0.983</td>
<td>0.138</td>
<td>0.985</td>
<td>0.983</td>
<td>0.984</td>
<td>98.3</td>
<td>0.923</td>
</tr>
</tbody>
</table>

The formula for False Positive Rates (FPR) was similar with False Alarm Rates. This project measures the performance of each method using accuracy, Precision, TPR and FPR only. Third objective was to evaluate the hybrid learning approaches.

Table 3 shows shows the differences on accuracy between three experiments held. The first experiment (KM+NB) shows the highest accuracy compared to the other. According to researcher [6], the combination of techniques may increase the value in term of accuracy and precision. The Decision tree (J48) classification resulted better compared to Rule-based (1R) classification [12].

Table 3. Results of Hybrid Learning Approaches.

<table>
<thead>
<tr>
<th>Measurement</th>
<th>KM+NB</th>
<th>KM+1R</th>
<th>KM+J48</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy (%)</td>
<td>99.4</td>
<td>99.4</td>
<td>98.3</td>
</tr>
<tr>
<td>Precision</td>
<td>0.995</td>
<td>0.995</td>
<td>0.985</td>
</tr>
<tr>
<td>True Positive Rates (TPR) (%)</td>
<td>0.994</td>
<td>0.994</td>
<td>0.983</td>
</tr>
<tr>
<td>False Positive Rates (FPR) (%)</td>
<td>0.00</td>
<td>0.00</td>
<td>0.138</td>
</tr>
</tbody>
</table>

Conclusion and Recommendation

After the experiment done in this research project, it can be concluded that using hybrid learning method yield better result compared to single method as mention in previous chapter. This research project did not covered any single classifier result. However, as we can see in the results from K-Means clustering the results from the clustered only had predicted so many malware apps as a normal apps. Therefore, a classification method had been proposed to overcome the flawlss of the single method.

In the future, a suggestion on a combination of static and dynamic analysis in Android malware detection should be considered. Moreover, use many datasets for example MalGenome, Drebin and
Andrubis may expand the results of both methods. Besides, the methods in clustering can be other than k-mean and hierarchical. Future works, another complex clustering method for example Expectation Maximization (EM) clustering and X-Means can be choose to replace K-Means clustering method for better results.

Acknowledgement

This research was financially supported by the Ministry of Higher Education through Fundamental Research Grant Scheme (500-RMI/FRGS 5/3 0006/2016).

References