Analysis of Data Mining Tools for Android Malware Detection

Robiah Yusof[1], Nurul Syahirrah Adnan [2], Nurlaily Abd. Jalil [3] and Raihana Syahirah Abdullah[4]
[1,2,3,4]Centre for Advanced Computing Technology,
Faculty of Information and Communication Technology
Universiti Teknikal Malaysia Melaka, Malaysia.
Email[1]: robiah@utem.edu.my

Abstract - There are various data mining tools available to analyze data related android malware detection. However, the problem arises in deciding the most appropriate machine learning techniques or algorithm on particular tools to be implemented on particular data. This research is focusing only on classification techniques. Hence, the objective of this research is to identify the best machine learning technique or algorithm on selected tool for android malware detection. Five techniques: *Random Forest, Naive Bayes, Support Vector Machine, Forest, K-Nearest Neighbour* and *Adaboost* are selected and applied in selected tools namely *Weka* and *Orange*. The result shows that *Adaboost* technique in *Weka* tool and *Random Forest* technique in *Orange* tool has obtained accuracy above 80% compare to other techniques. This result provides an option for the researcher on applying technique or algorithm on selected tool when analyzing android malware data.

Index Terms— android malware, machine learning tools, data mining, weka, orange

I. INTRODUCTION

Nowadays, the usage of mobile device has increasing rapidly and android has becoming the fastest growing mobile operating system due to it open source. This open nature has attracted the attention of developer and consumers to use this platform. In order to meet the latest mobile technology needs, the android itself can easily modify and enhance features for software developers. In the other hand, android malware known as malicious software can cause harmful to mobile device and become a challenge in the field of information security. There are several types of android malware such as rootkit, adware, spyware, worms, botnet and Trojan [18]. They were built to distract the application and user privacy information in mobile phone. The main target for malware attack on mobile device is on android platform.

This research is focuses on analyzing the android malware dataset obtained from the previous research [1]. Many researcher has lack of knowledge in selecting the best machine learning techniques in tools to analyze this type of dataset. In order to identify the best machine learning techniques and tool, five machine learning techniques are selected and then applied in the selected tools. The machine learning tools used are *Weka* and *Orange*. After successfully applied the machine learning techniques on the dataset, the results will be further analyzing to determine the best technique and tool.

The rest of the paper is structured as follows. Section 2 discusses the related work on the machine learning tools, classification techniques, dataset and parameter. Section 3 presents the methodology used in this research. Section 4 discusses the result of the analysis. Finally, Section 5 concludes and summarizes future directions of this work.

II. RELATED WORK

A. Machine Learning Tools

In this research, two machine learning tools were used: *Weka* and *Orange* tool. *Weka* (Weka Waikato Environment for Knowledge Analysis) is developed by University of Waikato, New Zealand. It is a famous machine learning software written in Java and it is an open source platform. *Weka* provides algorithms for analyzing predictive data and modelling. It also provides a collection of visualization tools and graphical user interfaces where it is easier for users to use [2]. Data mining task such as data pre-processing, clustering, classification, regression, visualization, and feature selection is a collection of machine learning algorithms in *Weka*. In addition, the algorithm can either be applied directly to the dataset or called from your own Java code [3].

Orange is another data mining tool used in this research and it is written in *C++* and *Phyton*. It is developed in 2009 and it has Pre-processing data, feature scoring and filtering, modelling and model evaluation and exploration techniques [4]. In *Orange* tool, Visual programming and explorative data analysis are useful. Besides that, various components of *orange* tool among them are known as widgets. *Orange* tool can operate in *Windows, Linux* and *macOS* [5].

B. Classification Techniques

Classification is data mining function that assign data in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. In data mining techniques, classification is one of the most popular techniques. There are seven classification techniques or algorithms: *Naïve Bayes (NB), Bayesian Network (BN), Support Vector Machine (SVM), Random Forest (RF), Adaboost, K-Nearest Neighbour (K-NN)* and *Neural Network (NN)*. However, in this research, only five classification techniques are applied which are *NB, RF, SVM, K-NN* and *Adaboost*.

Naïve Bayes (NB) depends on the Bayes theorem. It can use both binary and multi-class classification problems. NB method assesses the probability of each feature independently, regardless of any correlation, and predicts based on Bayes theorem [6]. Moreover, *NB* algorithm is a simple probabilistic classifier for calculating a set of probabilities where it calculates the frequency and combinations of values in a given data set [7].

Random forest (RF) is one of the most popular algorithms in machine learning and it has a collection of decision trees. RF can produce better prediction accuracy without needed almost any data preparation and modelling [8]. It also has a collection of unpruned CARTs and rule to combine individual tree decisions. *RF* has its own principle to encourage diversity among the tree [9].

Support Vector Machine (SVM) is a supervised learning approach. It is used for recognizing the pattern. Some researchers conclude that the accuracy value of 80% is good for *SVM* [10]. *SVM* can be defined as a system that uses hypothesis space linear. It is trained with learning algorithms from optimization of the theory that practice the learning bias which can be derived from statistical learning theory. *SVM* can guide the user to understand more on the ability of the algorithm [11].

K-Nearest Neighbour (K-NN) is one of the simple methods in data mining and machine learning. *K-NN* has practical facilities and efficiencies where it has proven that it has superior performance to classify several types of data and does not require modelling [12]. Furthermore, *K-NN* is used for pattern classification. Unlabelled test are categorized in *K-NN*. For example, it uses the majority of example labels among the *K-NN* that are most often in training set [13].

In machine learning algorithm, Adaptive Boosting known as *Adaboost*. It is formulated by Freund and Scapire. This algorithm is sensitive to noisy data. The executed classifiers are adjusted according to the examples wrongly classified with previous classifier [14]. Besides that, to improve performance, *Adaboost* can be used together with other algorithms [15].

C. Dataset

This research used three system call dataset with different size acquire from previous study [1]: dataset of 1 gram, 2 grams and 3 grams. The data are generated using four android tablets with GSM support which connected to fake DNs and web server. Captured log is processed into the log parser to obtain the total number of system calls executed by the application. This process is also required to differentiate between a malware and a valid system call, this is because not all system call can be considered as a malicious activity [16] [17].

D. Measurement Parameter

The measurement parameter used in this research are *Accuracy, Precision, Recall* and *F-Measure*. From the confusion matrix shown in Table 1, the following measurement are taken to calculate and use for the classifier.

- ***Confusion Matrix***

Tables 1 shows the confusion matrix. It is use as a guide on measurement of the *Accuracy, Precision, Recall* and *F-Measure*.

TABLE I
CONFUSION MATRIX

	Predicted	
Actual	0	1
0	TP	FN
1	FP	TN

In Table 1, *True Positive* (TP) in Actual and Predicted class will combine with *True Negative* (TN) in Actual and Predicted class to generate ***Correctly Classified Instances***. While, *False Negative* (FN) in Actual and Predicted class will combine *False Positive* in Actual and Predicted class to generate ***Incorrectly Classified Instances***.

Hence, the
Correctly Classified Instances = 00+11
Incorrectly Classified Instances = 01+10

i. ***Accuracy***
Accuracy is a parameter measure where the results close to the true value.

$$\text{Accuracy}=(TP+TN)/(TP+TN+FP+FN) \quad (1)$$

ii. ***Precision***
The proportion of *True Positive* classification from cases that are predicted as positive.

$$\text{Precision}=TP/(TP+FP) \quad (2)$$

iii. Recall
Recall is a ratio of correctly predicted positive observation to the all observation in actual class. It known as TPR (True Positive Rate).

$$\text{Recall}=TP/(TP+FN) \quad (3)$$

iv. ***F-Measure***
F-measure is a weighted harmonic mean of *Precision* and *Recall*.

$$\text{F-Measure} = 2 / (1 / \text{recall} + 1 / \text{precision}) \quad (4)$$

The higher value of Accuracy, Precision, Recall, F-Measure indicated show how precise the classifier classifies the instances. All the measurement parameters value will be used to determine the best algorithm in each of the machine learning tool.

III. METHODOLOGY

This research implement Malware Administrator Analysis Methodology [17] as shown in Fig.1. It consists of five phases: Phase 1-Selecting tools and malware data, Phase II-Installing the tools, Phase III-information collection, Phase IV-information analysis and Phase V-Documenting the results.

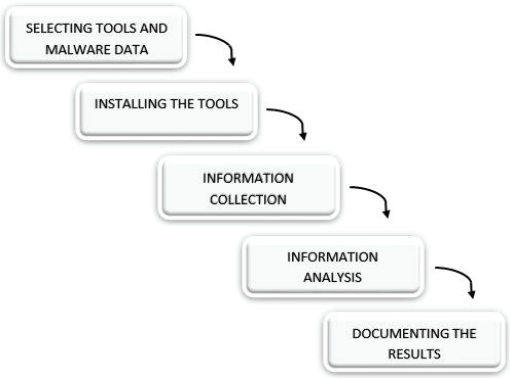

Fig. 1: Malware Administrator Analysis Methodology [17]

Phase I: In this phase, two tools are selected:*Weka* and *Orange*. The malware data of system call consists of three types of sizes: 1 gram, 2 gram and 3 gram obtained from the previous research [1].

Phase II: The tools *(Weka* and *Orange)* are installed in Windows 7.

Phase III: The selected techniques, algorithm and measurement parameter are applied during this phase and the results are collected.

Phase IV: The result is analyzed to identify the best algorithm for selected tools.

Phase V: The result is documented.

IV. RESULTS AND ANALYSIS

TABLE II and TABLE III is the *Accuracy* results after implementing these three dataset in *Weka* and *Orange* tools. Meanwhile, Fig. 2 and Fig. 3 are the graph generated from TABLE II and TABLE III respectively.

TABLE II
ACCURACY RESULT FOR WEKA

Dataset	Algorithm	Precision	Recall	F-Measure	Accuracy
1 gram	RF	0.520	0.512	0.495	51.22%
	NB	0.792	0.756	0.750	75.61%
	SVM	0.893	0.890	0.890	**89.02%**
	K-NN	0.827	0.817	0.816	81.70%
	Adaboost	0.819	0.817	0.817	81.70%
2 gram	RF	0.759	0.707	0.694	70.73%
	NB	0.806	0.732	0.717	73.17%
	SVM	0.936	0.927	0.927	**92.68%**
	K-NN	0.762	0.646	0.607	64.63%
	Adaboost	0.817	0.817	0.817	81.70%
3 gram	RF	0.796	0.744	0.734	74.39%
	NB	0.837	0.756	0.742	75.61%
	SVM	N/A	N/A	N/A	N/A
	K-NN	0.238	0.488	0.320	48.78%
	Adaboost	0.893	0.878	0.877	**87.80%**

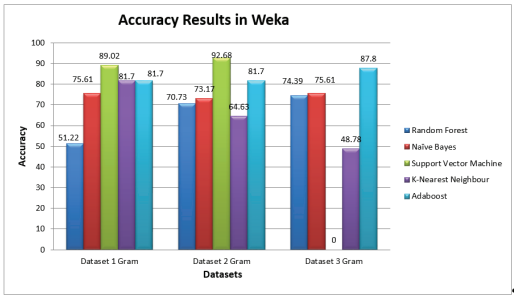

Fig. 2: Accuracy Result in Weka

Referring to TABLE II and Fig. 2, in *Weka* tool, the best algorithm selected is *Adaboost* since the accuracy for all dataset are above 80% (1 gram is 81.70%, 2 grams is 81.70% and 3 grams is 87.80%) compare to other algorithms. Eventhough SVM has higher result, still it is not considered as good since it cannot process the 3 gram's dataset.

TABLE III
ACCURACY RESULT FOR ORANGE

Dataset	Algorithm	Precision	Recall	F-Measure	Accuracy
1 gram	RF	0.871	0.868	0.868	**86.8%**
	NB	0.818	0.812	0.811	81.2%
	SVM	0.795	0.777	0.773	77.7%
	K-NN	0.841	0.832	0.831	83.2%
	Adaboost	0.828	0.827	0.827	82.7%
2 gram	RF	0.875	0.871	0.870	**89.5%**
	NB	0.790	0.776	0.773	77.6%
	SVM	0.728	0.723	0.722	72.3%
	K-NN	0.831	0.818	0.818	81.8%
	Adaboost	0.865	0.865	0.865	86.5%
3 gram	RF	0.881	0.879	0.879	**87.9%**
	NB	0.846	0.805	0.799	80.5%
	SVM	0.747	0.738	0.735	73.8%
	K-NN	0.858	0.841	0.840	84.1%
	Adaboost	0.877	0.877	0.877	87.7%

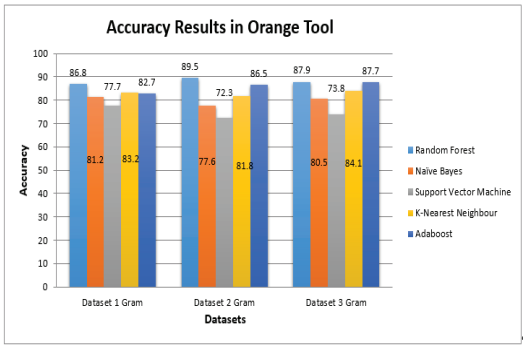

Fig. 3: Accuracy Result in Orange

Referring to TABLE III and Fig. 3, in *Orange* tool, the best algorithm chosen is *Random Forest* due to the accuracy for each of the dataset is also above 80% (1 gram is 86.8%, 2 grams is 89.5% and 3 grams is 87.9%).

Hence in this analysis, *Adaboost* is the best algorithm for *Weka* and *Random Forest* is the best algorithm for *Orange*.

V. Conclusion and Future Work

In conclusion both tools have acquired best technique or algorithm to be implemented on this dataset. Both algorithm manage to gain accuracy above 80%. It shows that *Adaboost* is the best algorithm for *Weka* tool and *Random Forest* is the best algorithm for *Orange* tool. This result provides an option for the researcher on applying technique or algorithm on selected tool when analyzing android malware data. In future, more machine learning tools and algorithm will be implemented on this dataset to identify the best algorithm or techniques to be used.

Acknowledgment

The authors would like to thank the INSFORNET, Center for Advanced Computing Technology (C-ACT), Fakulti Teknologi Maklumat dan Komunikasi (FTMK), Universiti Teknikal Malaysia Melaka.

References

[1] M.Z. Mas'ud, S. Sahib, M.F. Abdollah, S.R. Selamat, and R. Yusof, "Analysis of Features Selection and Machine Learning Classifier in Android Malware Detection", 2014 International Conference on Information Science & Applications (ICISA), 2014, pp. 1-5.

[2] K. Manchandia, N. Khare and M. Agrawal. "Weka As A Data Mining Tool To Analyze Students Academic performances Using Naïve bayes Classifier-A Survey". International Journal of Engineering Sciences & Research Technology, pp. 431-434, 2017.

[3] K. Rangra and K. L. Bansal. "Comparative Study of Data Mining Tools". International Journal of Advanced Technology in Computer Science and Software Engineering, vol. 4, Issue. 6, pp. 216-223, 2014. ISSN: 2277 128X.

[4] M.S. Kukasvadiya and N.H. Divecha, "Analysis of Data Using Data Mining tool Orange", International Journal of Engineering Development and Research, vol. 5, no. 2, pp. 1836-1840, 2017. ISSN: 2321-9939.

[5] M.Z. Mas'ud, S. Sahib, M.F. Abdollah, S.R. Selamat and R. Yusof. "Analysis of feature selection and machine learning classifier in android malware detection". 9th International Conference on Information Assurance and Security (IAS), 2013, pp. 1-5.

[6] K. Chumachenko, "Machine Learning Methods for Malware Detection and Classification", University of Applied Science, pp. 1-93, 2017.

[7] T.R. Patil and S.S. Sherekar, "Performance Analysis of NAIVE BAYES and J48 Classification Algorithm for Data Classification". International Journal of Computer Science And Applications, vol. 6, no. 2, pp. 1-6, 2013.

[8] L. Breiman, "RANDOM FOREST". University of California Berkeley, pp. 1-33, 2001.

[9] A. Abdiansah and R. Wardoyo. "Time Complexity Analysis of SUPPORT VECTOR MACHINEs (SVM) in LibSVM", International Journal of Computer Applications, vol. 128, no. 3, pp. 28-34, 2015.

[10] S.M. Vinayak Choubey, S.K. Pandey and J.P. Shukla, "An Efficient Approach of SUPPORT VECTOR MACHINE for Runoff Forecasting". International Journal of Scientific & Engineering Research, vol 5, no. 3, pp. 158-166, 2014.

[11] A. KousarNikhath, K.Subrahmanyam and R.Vasavi. "Building a K-Nearest Neighbor Classifier for Text Categorization". International Journal of Computer Science and Information Technologies (IJCSIT), vol. 7, no. 1, pp. 254-256, 2016.

[12] M. Alkasassbeh, G.A. Altarawneh, A.B. Hassanat, "On Enhancing The Performance of Nearest Neighbour Classifiers using Hassanat Distance Metric". Canadian Journal of Pure and Applied Sciences (CJPAS), vol. 9, no. 1, pp. 1-6, 2015.

[13] B. Markoski, Z. Ivanković, L. Ratgeber, P. Pecev and D. Glušac, "Application of ADABOOST Algorithm in Basketball Player Detection". University of Health Sciences Pécs, vol. 12, no. 1, pp. 189-207, 2015.

[14] N.K. Korada, N.S.P. Kumar and Y.V.N.H Deekshitulu, "Implementation of NAIVE BAYESian Classifier and Ada-Boost Algorithm using Maize Expert System", International Journal of Information Sciences and Techniques (IJIST), vol. 2, no. 3, pp. 63-75, 2012.

[15] Dr. Sudhir B. Jagtap & Dr. Kodge B. G, "Census Data Mining and Data Analysis using WEKA." International Conference in Emerging Trends in Science, Technology and Management, 2013, pp. 3-40.

[16] M.Z. Mas'ud, S. Sahib, M.F. Abdollah, S.R. Selamat and R. Yusof, "Profiling mobile malware behaviour through hybrid malware analysis approach. 2013 9th International Conference on Information Assurance and Security (IAS), 2013, pp. 1-7.

[17] Masood.(2014).Malware Analysis for Administrator [online] Available at: https://www.symantec.com/connect/articles/malware-analysis-administrators [Accessed 10 April 2017].

[18] Laraib U. Memon, Narmeen Z. Bawany, Jawwad A. Shamsi, "A comparison of machine learning techniques for android malware detection using apache spark", Jounal of Engineering Science and Technology, vol. 14, no. 3, pp. 1572-1586, 2019.

Robiah Yusof is currently a senior lecturer at the Universiti Teknikal Malaysia Melaka, Malaysia. She received her PhD in Network Security from Universiti Teknikal Malaysia Melaka. Her Research Interest include intrusion detection, malware, network security and network forensic.

Nurul Syahirrah Adnan hold Bachelor of Computer Science (Networking) from Universiti Teknikal Malaysia Melaka, Malaysia. Her Research Interest include network security and malware intrusion.

Nurlaily Abd. Jalil hold Bachelor of Computer Science (Networking) from Universiti Teknikal Malaysia Melaka, Malaysia. Her Research Interest include network security and malware intrusion.

Raihana Syahirah Abdullah is currently a senior lecturer at the Universiti Teknikal Malaysia Melaka, Malaysia. She received her PhD in Network Security from Universiti Teknikal Malaysia Melaka. Her Research Interest include intrusion detection, malware, network security, and network forensic.