A Comparative Study on Feature Selection Method for N-gram Mobile Malware Detection

Mohd Zaki Mas'ud, Shahrin Sahib, Mohd Faizal Abdollah, Siti Rahayu Selamat, Choo Yun Huoy
(Corresponding author: Mohd Zaki Mas'ud)

Faculty of Information Technology and Communication, Universiti Teknikal Malaysia Melaka
Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia
(Email: zaki.masud@utem.edu.my)
(Received Feb. 18, 2016; revised and accepted May 21 & July 19, 2016)

Abstract

In recent years, mobile device technology has become an important necessity in our community at large. The ability of the mobile technology today has become more similar to its desktop environment. Despite the advancement of the mobile devices technology provide, it has also exposes the mobile devices to the similar threat it predecessor possess. One of the anomaly based detection methods used in detecting mobile malware is the n-gram system call sequence. However, with the limited storage, memory and CPU processing power, mobile devices that provide this approach can exhaust the mobile device resources. This is due to the huge amount of system call to be collected and processed for the detection approach. To overcome the issues, this paper investigates the use of several different feature selection methods in optimizing the n-gram system call sequence feature in classifying benign and malicious mobile application. Several filter and wrapper feature selection methods are selected and their performance analyzed. The feature selection methods are evaluated based on the number of feature selected and the contribution it made to improve the True Positive Rate (TPR), False Positive Rate (FPR) and Accuracy of the Linear-SVM classifier in classifying benign and malicious mobile malware application.

Keywords: Feature Selection; Linear SVM; Mobile Malware; Mobile Malware Detection; N-gram

1 Introduction

The number of mobile malware has increased significantly within these recent years especially with the introduction of the android-based platform in the market. Android-based mobile device offers credibility; performance and ease of customizing has made it a preferable choice for most of mobile device users. Consequently, the high reputation of android-based mobile devices has invited the malware author to make it as a new target to exploit. This is shown in the 2013 Kaspersky's Lab report which reveals 98% of the mobile malware found in 2013 is targeting the Android platform [11]. Additionally, in 2015 new samples of mobile malware are still continuing to increase, this is based on the report done by the 2015 McAfee Labs Threats Report [20]. In order to overcome these issues, several researches had been done in finding the defense mechanism against the android-based mobile malware.

Previous research done in mobile malware detection showed that mobile malware detection can be classified into 3 different techniques which are signature-based, anomaly-based and specification-based [19]. Signature-based approach detects malware by comparing the mobile application activity signature with the database of known attack or threat. Even though it has been used in developing most of the antivirus software and successfully detects known malware with a high accuracy, the technique has a drawback in detecting unknown malware. On the other hand, the anomaly-based and specification-based detection techniques have - great reputation in detecting unknown or new malware but these two techniques tend to generate false alert or generating misclassification. Using the advantage of anomaly-based detection technique, this research has applied n-gram system call sequence as the feature to improve the classification accuracy and reduce the false alert. However, the n-gram system call sequence generates a huge number of features that can increase the processing time and complexity. Thus, in order to reduce the number of features and at the same time improves the classification accuracy, as well as minimizing the false alert, this paper investigates several existing feature selection approaches.

The aim of this research is to find the feature selection method that can provide an optimum n-gram system call sequence feature to be used in the classification of benign and malicious mobile application. Feature selection is one of the essential techniques in data mining especially during the data processing [3, 16]. The main objective of feature selection phase is to improve the mining performance as well as improving the detection accuracy by removing

irrelevant, redundant and noisy data from the dataset. Subsequently, as the number of irrelevant features is removed the data mining process become less complex to process and this can speed up the classification or clustering process.

The remainder of this paper is divided into four parts whereby section two and three review the background domain of n-gram system call sequence in mobile malware detection and the related feature selection approach investigated in this research. Section four explains the experimental setup used in evaluating the selection feature selection method. Section five presents and discusses the experimental result. Finally, the conclusion and future work is drawn in the last section.

1.1 N-gram in Mobile Malware Detection

Mobile malware has become a lethal threat to mobile device users as the effects of mobile malware infection can be from stealing confidential information from the device, monitoring user activity and location, overcharge users by sending random SMS and MMS to contact, launching denial of services attack from user devices and overloading device resources such as memory, battery and storage [17]. One of the options in detecting these activities on mobile devices is by monitoring system call invoked by the mobile application [18]. Xi et al. [27], Crowdroid [5], Isohara et al. [12], AMDA [1] and MADAM [9] are among the works using system call as the features in classifying benign and malicious mobile application. From all these - works, only Isohara et al. use signature-based detection approach and the rest use anomaly-based detection approach that takes each single occurrence of the system call as the feature.

The use of anomaly-based detection approach in detecting mobile malware application can lead to the generation of false alert in which the benign application might be misclassify as a malicious mobile application or the other way around. This issue can be improved by using a feature of a sequence system call occurrence which has been used text classification and speech recognition domain [7, 13, 22, 24]. Known as n-gram analysis, n is the value of the number of sequence and it can represent the whole system call invoke to execute a malicious. For malware detection, n-gram analysis approach has been implemented in classifying malware using its byte level information [4, 21] and its opcode [6, 26] but this requires the malware application to be decompiled before the detection process take places.

Despite the improvement in reducing the false alert, the n-gram analysis can cause a huge number of features to be captured and processed. This is not an applicable option in a limited processing power and resources devices such as mobile device. The number of system call in an android 4.0.4 OS is 300 [25], yet only 111 system call is invoked during the experiment. Accordingly, as the n value increases, the total number of system call sequence used as features is also increases to the power of n. For example, for n=2 the total of system call sequence to be collected for this experiment is equal to 1112 or 12 321 and if n=3, the total of system call sequence to be collected is equal to 1113 or 1 367 631 which is quite a huge number to be processed in a classification problem. To reduce the number of relevance features in the classification, this research investigates several feature selection methods. The best feature selection method is evaluated based on the optimum number of features it generates and how good the features contribute in improving the classification accuracy while reducing the false alert.

1.2 Feature Selection

The n-gram system call sequence can generate a large number of features to be used in the classification and can contribute to the degradation of classification performance. This can be caused by the existence of useless features that might not be useful at all in classifying the problem. In order to overcome this matter, feature selection method is introduced in the framework for the purpose of finding the optimum features which can improve the classification performance and accuracy [2, 8, 10, 23]. In addition, the feature selection also contributes in reducing the number of selected features to be logged, thus less number of storage is used.

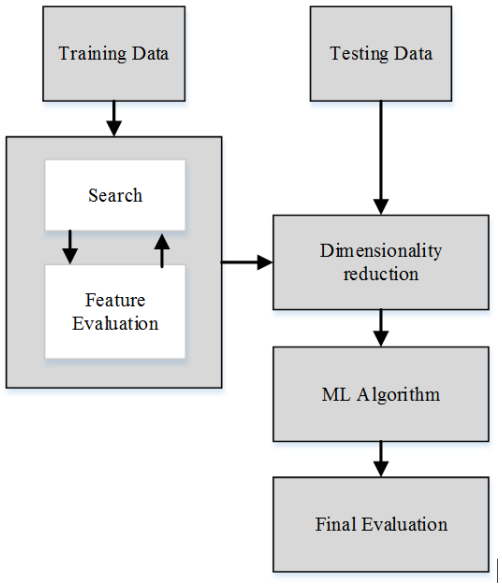

Figure 1: Filter method

Generally, there are 3 feature selection methods; filter method, wrapper method, and embedded method [23]. Filter method is illustrated in Figure 1. This method evaluates the significance of each feature using statistical approach that scored and ranked most relevance features. Features that obtained the highest ranked and scored are most likely to be chosen as the features in machine learning problem, whereas features with the lowest value are removed. Filter method is fast to compute and not affecting any of the classifier used; however this method ignores the feature dependencies and disregards the in-

teraction with the classifier. Furthermore, the threshold or the cut off value of the feature scored and ranked is not properly specified.

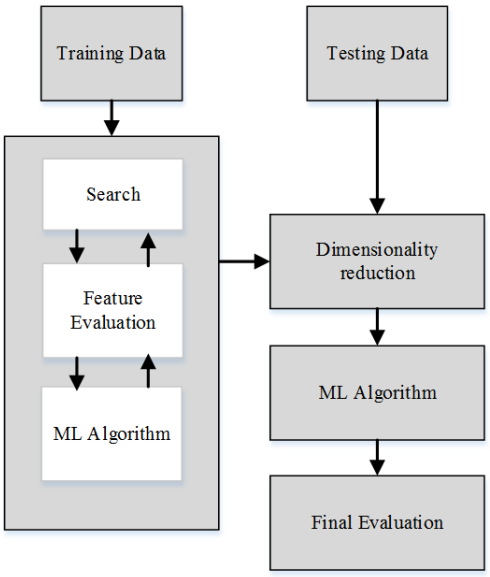

Figure 2: Wrapper method

Figure 2 illustrates wrapper method which evaluates subset of features using induction algorithm that incorporated the classifier as part of the evaluation. Thus, the features selected are specifically tailored and optimize to the classifier. Even though it is computationally intensive, wrapper method provides a possibility of interaction between features that can generate a more accurate classification. Meanwhile, embedded method is proposed to incorporate the advantages of the filter and wrapper method. The features are evaluated inside the induction algorithm itself and computationally intensive compared with wrapper methods. Nevertheless, for the purpose of finding the optimum selection method in the n-gram system call sequence feature, this paper only evaluates filter and wrapper selection methods. Four different filter methods, namely Correlation-based Feature Selection (CFS), Chi Square (CHI), Information Gain (IG), ReliefF (RF) and one wrapper method with a Linear SVM classifier (WR) are chosen to be evaluated in this paper.

CFS selects feature subset based on the maximal correlation of the subset to the class and the minimal correlation between the features. The features are ranked by using a correlation based heuristic evaluation function [14]. Meanwhile, CHI method evaluates feature subset with respect to the class labels based on the x^2-statistic function. The features are ranked and the higher the features ranked, the most likely it is chosen as the features. Similarly, IG also select features by ranking the feature based on the score generated on how much information about the class is gained when using the feature. RF assesses an attribute by repetitively sampling a feature and considering the value of the given attribute for the nearest features of the same and different class [14]. The wrapper method generates a subset of feature candidate using a search method and applied it to the Linear SVM classifier to be evaluated using the classification Accuracy and Root Mean Square Error (RMSE) [15]. The feature subset that produced the best accuracy and RMSE is used as the features. The next section describes the methodology and the experimental setup used in evaluating these feature selection methods.

2 Methodology

The objective of this study is to compare and suggest feature selection method to be used in selecting the optimum system call features in classifying benign and malicious android application. To achieve the objective stated, an extensive and rigorous empirical comparative study is designed and conducted. The experiment is conducted through several phases namely system call log phase, n-gram extraction phase, feature selection method comparison phase and followed by machine learning classifiers phase that is used for evaluating the feature selection method. The entire phase involved in this study is illustrated in Figure 3.

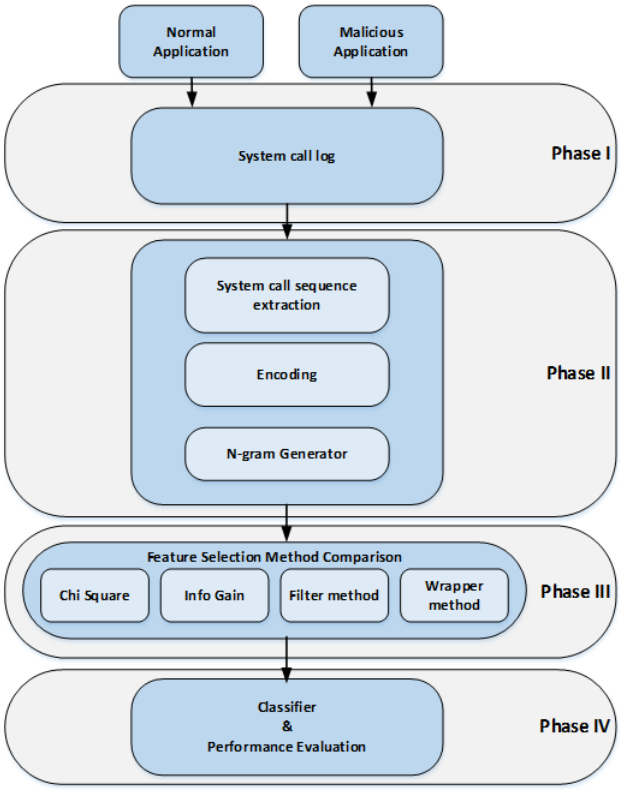

Figure 3: Research methodology

The four phases in the research methodology in Figure 3 begins with data collection phase where the system call invoked by the application is collected. Mobile applications used in the experiment are from 102 malware infected applications from the MalGenome Project [28] and 100 normal android application downloaded from Google

Play. In order to validate the benign and malicious application, the application used in this experiment has been verified by Bitdefender, eseT and VirusTotal for verification whether it is truly a malicious or benign application. Each android application is executed on a Samsung P6800 Galaxy tab 7.7 that is connected to a network experimental testbed via Wi-Fi. Each tablet is also provided with an active GSM service. Each application installed in the tablet is stimulated with user interaction as well as other mobile users common activities such as web browsing, sending and receiving SMS for duration of 10 minutes. During the execution and stimulation processes, a tool called strace is used to capture all the system calls invoked by the installed application. Once the android application has gone through these processes, the tablet is wiped out clean to its factory setting before the next application is executed.

The captured system call is then processed in the second phase where the output from the strace is transformed to a sequence of n-gram system call. Then in the third phase, the system call sequence is applied to several feature selection methods for generating the best feature. The next phase takes the best feature generated from each feature selection method and applied it to the classification method. This final phase also evaluated the number of feature selected and the classification performance in term of the Accuracy, True Positive Rate (TPR) and False Positive Rate (FPR).

3 Analysis and Discussion

The primary consideration in this study is to evaluate feature selection method that can reduce the number of features selected while improving the classification between the benign and malicious mobile application. The n value for the n-gram system call sequence used in this experiment is 3. This is based on the reason that if all the features are used during the classification, the optimum performance of the classification is only achieved when the n value of the n-gram system call sequence is 3. Table 1 shows the classification performance evaluation done on the dataset using all the features for n value of 1, 2, 3, 4, 5 and 6.

Table 1 shows all the classification performance evaluation results when all features are considered for each n-gram system call sequence. The highest accuracy in the classifier performance evaluation result is 96.19%, achieved when the n value is 3. Even though the TPR value is not the highest, yet the overall classification accuracy and FPR value is still higher than the other N-gram system call sequence. The results also show that the higher the number of sequence considered in generating the features does not affect the classification, instead it can produce a sparse vector resulting in lower accuracy. This caused by the existence of useless features that might not be useful at all in classifying the problem.

Despite of the higher classification accuracy produced by the 3-gram system call sequence, the 3-gram still use a large number of feature which is 41142. This large number of feature can degrade the classification performance. To improve the classification performance, each feature selection method discussed earlier is then applied to the 3-gram dataset for finding the most relevant feature and at the same time reducing the number of features. The result of each feature selection method is shown in Table 2.

Table 2 shows the number of features selected by each method with the TPR, FPR and Accuracy. Out of 10 methods, only WR+ES method reduced the number of features less than 50%. Six methods which are CFS+ES, CFS+GS, CFS+PSO, CHI+Ranker, IG+Ranker, RF+Ranker, WR+GA and WR+PSO successfully reduced the features to more than 50%. Two methods are able to reduce the features up to 99%, CFS+BFS reduce the features to only 83 features whereas WR+BF can reduce the features up to only 10.

The second measurement of this study is to measure the improvement of Accuracy, TPR and FPR when the features selected are applied to the classification method. The evaluation shows CFS+BF, WR+BF, WR+ES, WR+GA and WR+PSO have improved the classification accuracy to more than the original classification accuracy which is 96.2%. WR+BF selection and search methods improved the TPR, FPR and accuracy to 100%, 2% and 99% respectively making it the best features selection method for this evaluation. Although the result shows WR+BF has the smallest number of features selected, this selection method have better ability to support the classifier to accurately classify between the malicious and benign application compare to the other features suggested by the other selection methods. This shows that the WR+BF have the ability to find an optimum features that is optimally design for the classifier.

4 Conclusion

The rapid evolution in mobile device technology has triggered a sudden increase of mobile malware threat. The effects of mobile malware threat are devastating especially when communities nowadays are depending on mobile device to store crucial information. An anomaly-based detection using n-gram system call sequence is one option that can be used to mitigate the malicious application from exploiting vulnerabilities in mobile device. However, the approach can create a large number of features are as the n value increases and can degrade the classification performance. Based on this reason, this paper evaluates several feature selection methods by comparatively analyze the performance of each selection method. Each selection method is evaluated based on the number of feature selected and the contribution it made to improves the True Positive Rate (TPR), False Positive Rate (FPR) and Accuracy of the Linear-SVM classifier in classifying benign and malicious mobile

Table 1: The classifier performance evaluation result

N-gram Dataset	Number of features	TPR (%)	FPR (%)	Accuracy (%)
1-gram	111	96.07	78	87.08
2-gram	3631	96.07	91	93.51
3-gram	41142	97.06	5	96.19
4-gram	186610	100	85	92.33
5-gram	491782	100	71	85.59
6-gram	987263	100	58	79.16

Table 2: The feature selection method performance evaluation result

Feature Selection Method	Search Method	Number of Features Selected	Reduce Percentage	TPR (%)	FPR (%)	Accuracy (%)
None		41142		98.0	6.0	96.2
CFS	BF	83	99.80	99.0	4.0	97.5
	ES	12872	68.71	96.1	6.0	95.1
	GS	10950	73.38	96.1	12.0	92.1
	PSO	10495	74.49	96.1	14.0	91.1
CHI(50%)	Ranker	20572	50.00	98.0	6.0	96.0
IG(50%)	Ranker	20572	50.00	98.0	6.0	96.0
RF(50%)	Ranker	20572	50.00	99.0	6.0	95.0
WR	BF	10	99.98	100.0	2.0	99.0
	ES	23773	42.22	98.0	3.0	97.5
	GA	17490	57.49	99.0	3.0	98.0
	PSO	16874	58.99	99.0	3.0	98.0

malware application. The selection method evaluated in this paper are Correlation-based Feature Selection (CFS), Chi Square (CHI), Information Gain (IG), ReliefF (RF) and wrapper (WR) method with a Linear SVM classifier (WR). CFS and wrapper evaluator are match with four search method namely BestFirst (BF), Evolutionary Search (ES), Genetic Search (GS) and Particle Swarm Optimization (PSO) search whereas CHI, IG and RF are match with Ranker method. Each feature generated by the selection method is then applied to a Linear-SVM classifier for TPR, FPR and Accuracy. The evaluation shows an increase in accuracy for all the features generated by the feature selection algorithm and WR-BF generated the smallest number of feature while having an accuracy of 99% and small FPR of 2%. This shows that WR+BF have the ability to find and optimum features to be used in the classifier. Moreover, the result also shows that it is possible to improve the detection performance even though the features selection used in the classifier is reduced. This small number of feature can help reduce the size of log collection in the mobile device. In the near future, the method presented in this paper may be potentially applied to develop an android malware detection that can address the limitation and constrain of mobile devices environment especially on the storage, memory and power consumption usage.

Acknowledgments

The authors would like to thank INSFORNET Research Group of Universiti Teknikal Malaysia Melaka (UTeM) for the financial support under the Fundamental Research Grant Scheme with Project No. FRGS/1/2015/ICT04/UTeM/02/F00290.

References

[1] K. J. Abela, J. R. D. Alas, D. K. Angeles, R. J. Tolentino, and M. A. Gomez, "Automated malware detection for android: Amda," in *The Second International Conference on Cyber Security, Cyber Peacefare and Digital Forensic (CyberSec'13)*, pp. 180–188, 2013.

[2] B. Arslan, S. Gunduz, and S. Sagiroglu, "A review on mobile threats and machine learning based detection approaches," in *4th IEEE International Symposium on Digital Forensic and Security (ISDFS'16)*, pp. 7–13, 2016.

[3] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial Intelligence*, no. 97, pp. 245–271, 1997.

[4] T. B. Assaleh, V. Keselj, and R. Sweidan, "N-gram based detection of new malicious code," in *Proceedings of The 28th IEEE Annual International Computer Software and Applications*, pp. 41–42, Hong Kong, Sept. 2004.

[5] I. Burguera, U. Zurutuza, and S. Nadjm-Tehrani, "Crowdroid: Behavior-based malware detection system for android," in *Proceedings of the 1st ACM Workshop on Security and Privacy in Smartphones and Mobile Devices*, pp. 15–26, 2011.

[6] G. Canfora, A. D. Lorenzo, E. Medvet, F. Mercaldo, and C. A. Visaggio, "Effectiveness of opcode n-grams for detection of multi family android malware," in *10th IEEE International Conference on Availability, Reliability and Security (ARES'15)*, pp. 333–340, 2015.

[7] W. B. Cavnar and M. T. John, "N-gram-based text categorization," *Ann Arbor MI 48113.2*, vol. 48113, no. 2, pp. 161–175, 1994.

[8] M. Dash and H. Liu, "Feature selection for classification: Intelligent data analysis," *Intelligent Data Analysis*, vol. 1, no. 1, pp. 131–156, 1997.

[9] G. Dini, F. Martinelli, A. Saracino, and D. Sgandurra, "Madam: A multi-level anomaly detector for android malware," in *International Conference on Mathematical Methods, Models, and Architectures for Computer Network Security*, pp. 240–253, 2012.

[10] F. Fernández-Gutiérrez, J. I. Kennedy, S. Zhou, R. Cooksey, M. Atkinson, and S. Brophy, "Comparing feature selection methods for high-dimensional imbalanced data: Identifying rheumatoid arthritis cohorts from routine data," in *IEEE International Conference on Industrial Engineering and Systems Management (IESM'15)*, pp. 236–241, 2015.

[11] C. Funk and M. Garnaeva, *Kaspersky Security Bulletin 2013. Overall Statistics for 2013*, Technical Report, Kaspersky, 2013.

[12] T. Isohara, T. Keisuke, and K. Ayumu, "Kernel-based behavior analysis for android malware detection," in *Seventh IEEE International Conference on Computational Intelligence and Security (CIS'11)*, pp. 1011–1015, 2011.

[13] D. Jurafsky and H. M. James, *Speech & Language Processing*, India: Pearson Education, 2000.

[14] K. Kira and A. R. Larry, "A practical approach to feature selection," in *Proceedings of The Ninth International Workshop on Machine Learning*, pp. 249–256, Scotland, UK, July 1992.

[15] R. Kohavi and H. J. George, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1, pp. 273–324, 1997.

[16] H. Liu and H. Motoda, *Feature Extraction, Construction and Selection: A Data Mining Perspective (2nd printing)*, Boston: Kluwer Academic Publishers, 2001.

[17] L. P. Mariantonietta, F. Martinelli, and D. Sgandurra, "A survey on security f or mobile devices," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 1, pp. 446–471, 2013.

[18] M. Z. Mas'ud, S. Sahib, M. F. Abdollah, S. R. Selamat, R. Yusof, and R. Ahmad, "Profiling mobile malware behaviour through hybrid malware analysis approach," in *9th International Conference on Information Assurance and Security (IAS'13)*, pp. 78–84, 2013.

[19] M. Z. Masud, S. Sahib, M. F. Abdollah, S. R. Selamat, and R. Yusof, "Android malware detection system classification," *Research Journal of Information Technology*, vol. 6, no. 4, pp. 325–341, 2014.

[20] McAfee, *Mcafee Labs Threats Report 2015*, Technical Report, McAfee, 2015.

[21] R. Moskovitch, D. Stopel, C. Feher, N. Nissim, N. Japkowicz, and Y. Elovici, "Unknown malcode detection and the imbalance problem," *Journal in Computer Virology*, vol. 5, no. 4, pp. 295–308, 2009.

[22] S. Nagaprasad, T. R. Reddy, P. V. Reddy, A. V. Babu, and B. VishnuVardhan. "Empirical evaluations using character and word n-grams on authorship attribution for telugu text,". in *Intelligent Computing and Applications*, pp. 613–623, 2015.

[23] S. F. Pratama, A. K. Muda, Y. H. Choo, and N. A. Muda, "A comparative study of feature selection methods for authorship invarianceness in writer identification," *International Journal of Computer Information Systems and Industrial Management Applications*, vol. 4, pp. 467–476, 2012.

[24] M. Ren and S. Kang, "Document classification using n-gram and word semantic similarity," *International Journal of u-and e-Service, Science and Technology*, vol. 8, no. 8, pp. 111–118, 2015.

[25] T. Robot, *Android Platform Bionic*, 2014. (`https://github.com/android/platform_bionic/blob/master/libc/SYSCALLS.TXT`)

[26] A. Shabtai, M. Robert, F. Clint, D. Shlomi, and E. Yuval, "Detecting unknown malicious code by applying classification techniques on opcode patterns," *Security Informatics*, vol. 1, no. 1, pp. 1–22, 2012.

[27] X. Xi, F. Peng, X. Xianni, J. Yong, L. Qing, and L. Runiu, "Two effective methods to detect mobile malware," in *4th International Conference on Computer Science and Network Technology (ICCSNT'15)*, vol. 1, pp. 1041–1045, 2015.

[28] Y. Zhou and J. Xuxian, "Dissecting android malware: Characterization and evolution," in *Proceedings of The IEEE Symposium on Security and Privacy (SP'12)*, pp. 95–109, San Francisco, California, May 2012.

Biography

Mohd Zaki Mas′ud is a lecturer at the Universiti Teknikal Malaysia Melaka, Malaysia and current pursuing his PhD study in Malware Analysis. His research interest include network forensic, cyber terrorism, intrusion detection, network security , network management and penetration testing.

Shahrin Sahib received the Bachelor of Science in Engineering, Computer Systems and Master of Science in Engineering, System Software in Purdue University

in 1989 and 1991 respectively. He received the Doctor of Philosophy, Parallel Processing from University of Sheffield in 1995. He is a professor and the Vice Chancellor of Universiti Teknikal Malaysia Melaka. His research interests include network security, computer system security, network administration and network design. He is a member panel of Experts National ICT Security and Emergency Response Center and also Member of Technical Working Group: Policy and Implementation Plan, National Open Source Policy.

Mohd Faizal Abdollah is currently an associate professor at the Universiti Teknikal Malaysia Melaka, Malaysia. He received his Doctor of Philosophy in Computer Science. His research interests include network forensic, cyber terrorism, intrusion detection, network security , network management and penetration testing.

Siti Rahayu Selamat is currently a lecturer at the Universiti Teknikal Malaysia Melaka, Malaysia. She received her Doctor of Philosophy in Computer Science. Her research interests include network forensic, cyber terrorism, intrusion detection, network security and penetration testing.

Yun-Huoy Choo was born in Johor, Malaysia, in 1977. She received the B.Sc. and M.Sc. degree from the University of Technology Malaysia, in 2000 and 2002, respectively. In 2008, she was awarded the PhD in System Management and Science from the National University of Malaysia specializing in Data Mining. Since June 2002, she has been with the Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka (UTeM), Malaysia, where she was a Lecturer, became a Senior Lecturer in 2009, and an Associate Professor in 2015. Her current research interests include the fundamental studies of rough set theory, fuzzy sets theory, association rules mining, and feature selection, besides the application of data science and data mining in different domains includes person authentication using bio signal, muscle endurance analysis, machine failure analysis, personalized itinerary and route planning, etc.