On the Selection of Key Features for Android Malware Characterization

Javier Sedano, Camelia Chira, Silvia González, Álvaro Herrero, Emilio Corchado and José Ramón Villar

Abstract Undoubtedly, mobile devices (mainly smartphones and tablets up to now) have become the new paradigm of user-computer interaction. The use of such gadgets is increasing to unexpected figures and, at the same time, the number of potential security risks. This paper focuses on the bad-intentioned Android apps, as it is still the most widely used operating systems for such devices. Accurate detection of this malware remains an open challenge, mainly due to the ever-changing nature of malware and the "open" distribution channel of Android apps through Google Play. Present work uses feature selection for the identification of those features that may help in characterizing mobile Android-based malware. Maximum Relevance Minimum Redundancy and genetic algorithms guided by information correlation

J. Sedano (✉) · S. González
Instituto Tecnológico de Castilla Y León, C/López Bravo 70, Pol. Ind. Villalonquejar,
09001 Burgos, Spain
e-mail: javier.sedano@itcl.es

S. González
e-mail: silvia.gonzalez@itcl.es

C. Chira
Department of Computer Science, University of Cluj-Napoca, Baritiu
26-28, Cluj-Napoca 400027, Romania
e-mail: camelia.chira@cs.utcluj.ro

Á. Herrero
Department of Civil Engineering, University of Burgos, Avenida de Cantabria S/N,
09006 Burgos, Spain
e-mail: ahcosio@ubu.es

E. Corchado
Department of Computer Science and Automation, University of Salamanca,
Plaza de La Merced, S/N, 37008 Salamanca, Spain
e-mail: escorchado@usal.es

J.R. Villar
Computer Science Department, ETSIMO, University of Oviedo, 33005 Oviedo, Spain
e-mail: villarjose@uniovi.es

Á. Herrero et al. (eds.), *International Joint Conference*, Advances in Intelligent
Systems and Computing 369, DOI 10.1007/978-3-319-19713-5_15

measures have been applied to the Android Malware Genome (Malgenome) dataset, attaining interesting results on the most informative features for the characterization of representative families of existing Android malware.

Keywords Feature selection • Max-Relevance Min-Redundancy criteria • Information correlation coefficient • Android • Malware

1 Introduction

Since the first smartphones came onto the market (late 90s), sales on that sector have increased constantly until present days. Among all the available operating systems, Google's Android actually is the most popular mobile platform, according to [1]. 250.06 million of Android-run units were sold in Q3 2014 worldwide, out of 301.01. Similarly, the number of apps available at Android's official store has increased constantly from the very beginning, up to around 1,541,500 apps [2] available nowadays. Moreover, Android became the top mobile malware platform as well. Recent news [3] confirm this trend as 99 % of the new threats that emerged in Q1 2014 were run on Android. This operating system is an appealing target for bad-intentioned people, reaching unexpected heights, as there are cases where PC malware is now being transfigured as Android malware [3].

To fight against such a problem, it is required to understand the malware and its nature. Otherwise, it will not be possible to practically develop an effective solution [4]. Thus, present study is focused on the characterization of Android malware families, trying to reduce the amount of app features needed to distinguish among all of them. To do so, a real-life publicly-available dataset [5] has been analyzed by means of several feature selection strategies. From the samples contained in such dataset, several alarming statistics were found [4], that motivate further research on Android malware:

- Around one third (36.7 %) of the collected samples leverage root-level exploits to fully compromise the Android security.
- More than 90 % turn the compromised phones into a botnet controlled through network or short messages.
- 45.3 % of the samples have the built-in support of sending out background short messages (to premium-rate numbers) or making phone calls without user awareness.
- 51.1 % of the samples harvested user's information, including user accounts and short messages stored on the phones.

To improve the characterization of the addressed malware, this paper proposes the use of feature selection. To more easily identify the malware family an app belongs to, authors address this feature selection problem using a genetic algorithm guided by information theory measures. Each individual encodes the subset of selected features using the binary representation. The evolutionary search process is guided by crossover and mutation operators specific to the binary encoding and a fitness function that evaluates the quality of the encoded feature subset. In the current study, this fitness function can be the mutual information or the information correlation coefficient.

Feature selection methods are normally used to reduce the number of features considered in a classification task by removing irrelevant or noisy features [6, 7]. Filter methods perform feature selection independently from the learning algorithm while wrapper models embed classifiers in the search model [8, 9]. Filter methods select features based on some measures that determine their relevance to the target class without any correlation to a learning method. The Minimum-Redundancy Maximum-Relevance (MRMR) feature selection framework [8] is a well-known filter method. Besides the maximal relevance criteria, MRMR requires selected features to further be maximally dissimilar to each other (the minimum redundancy criteria). On the other hand, wrapper models integrate learning algorithms in the selection process and determine the relevance of a feature based on the learning accuracy [10]. Population-based randomized heuristics are normally used to guide the search towards the optimal feature subset. Wrapper methods require a high computational time and present a high risk of overfitting [10] but they are able to model feature dependencies and the interaction of the search model with the classifier [11]. Although MRMR was previously applied to the detection of malware [12] and machine learning has also been applied to the detection of android Malware [13, 14], present study differentiates from previous work as feature selection is now applied from a new perspective, trying to ease the characterization of different Android malware families.

The MRMR method is further used in this study to compare or confirm the subsets of selected features related to Android malware. The results obtained for the considered problem are extensively analysed, describing their relevance that probes the positive aspects of gaining deep knowledge of malware nature.

The structure of the paper is as follows: the MRMR method and the proposed GA-based feature selection algorithm are described in Sect. 2, the experiments for the Android Malware Genome dataset are presented in Sect. 3, the results obtained are discussed in Sect. 3.1 and the conclusions of the study are drawn in Sect. 4.

2 Feature Selection Methods

Since the number of features to be analysed in present study is small (see Sect. 3), the various feature subsets can be extensively evaluated using different methods. The result of these methods can then be aggregated in a ranking scheme. It is proposed to determine an ordered list of selected features using (i) Minimum-Redundancy Maximum-Relevance criteria [8] and (ii) a genetic algorithm based on information theory measures as fitness function. The methods described in this section assume a matrix X of N feature values in M samples and an output value y for each sample.

2.1 Minimum-Redundancy Maximum-Relevance Criteria

The Minimum-Redundancy Maximum-Relevance (MRMR) [8] feature selection method aims to obtain maximum relevance to output and in the same time minimum redundancy between the selected features.

Defined by means of their probability distribution, the mutual information between two variables has a higher value for higher degrees of relevance between the two features. Let I(X, Y) be the mutual information between two features, given by:

$$I(X, Y) = \iint p(x, y) * \log\left(\frac{p(x, y)}{p(x) * p(y)}\right) dxdy \tag{1}$$

In the first step, the MRMR approach selects one feature out of the N input features in the set X which has the maximum value of $I(x, y)$. Let this feature be x_k. Next, one of the features in $X - x_k$ is chosen according to the MRMR criteria.

Let us suppose that we have $m - 1$ features selected already in the subset S_{m-1} and the task is to select the mth feature from $X - S_{m-1}$. This will be the feature that maximizes the following formula:

$$\max_{x_j} \in X - S_{m-1} \left[I(x_j, y) - \frac{1}{m-1} \sum_{x_i \in S_{m-1}} I(x_j, x_i) \right] \tag{2}$$

This MRMR scheme can be run for m = 1, 2, 3… resulting in different feature subsets.

2.2 A Genetic Algorithm Using Information Theory Measures for Feature Selection

The proposed Genetic Algorithm (GA) encodes in each individual the feature selection by using a binary representation of features. The size of each individual equals the number of features and the value of each position can be 0 or 1, where 1 means that the corresponding feature is selected (the number of features is N).

It is proposed to evaluate feature selection results using the following two measures from information theory [15] as fitness functions: mutual information(I) and information correlation coefficient (ICC).

Let $\mathrm{H}(\mathrm{X},\mathrm{Y})$ denote the joint entropy of the two features, and by $\mathrm{I}(\mathrm{X},\mathrm{Y})$ the mutual information between X and Y (see Eq. 1).The information correlation coefficient $\mathrm{ICC}(\mathrm{X},\mathrm{Y})$ is calculated based on the Eq. 3 for all the features selected in individual X and the output Y. ICC measures how independent two features are from each other (the higher the ICC value the more relevant the relationship is).

$$ICC(X,Y) = \frac{I(X,Y)}{H(X,Y)} \tag{3}$$

If $ICC(X,Y) = 1$ then the two variables X and Y are strictly dependent whereas a value of 0 indicates that they are completely irrelevant to each other.

The resulting genetic algorithm (called GA-INFO, where INFO can be either I or ICC) is outlined below. The population size is denoted by *N*, the maximum number of generations is denoted by G and *t* represents the current generation.

Algorithm: GA-INFO Feature Selection

```
Require: X the input variables data set
Require: Y the output vector
P ← a vector of N Individual objects
t ← 0
Generate the initial population P(t): randomly initialize the value of each individual
while t < G do
    Evaluate each individual IND in P(t): calculate I(IND, Y) or ICC(IND, Y) value
    P(t +1) ← roulette wheel selection from P(t)
    for all individuals IND in P(t + 1) do
        Select mate J from P(t + 1)
        K ←two-point crossover (IND, J)
        if fitness(K) > fitness(IND) then
            IND ← K
        end if
        L ← mutation(IND)
        if fitness(L) > fitness(IND) then
            IND ← L
        end if
    end for
    t ← t+1
end while
Return Best Individual in P(t)
```

The GA follows a standard scheme in which roulette wheel selection, two-point crossover and swap mutation are used to guide the search. Each individual is evaluated based on the correlation between the current subset of selected features and the output. This correlation is given by either I or ICC used to evaluate the fitness. Therefore, depending on the fitness function used, two GA variants result:

GA-I is the GA using mutual information as fitness, while GA-ICC denotes the GA based on ICC fitness function.

3 Experiments

As previously mentioned, the Malgenome dataset [4], coming from the Android Malware Genome Project [5] has been analysed in preset study. It was the first large collection of Android malware (1,260 samples) that was split in 49 different malware families. It covered the majority of existing Android malware, collected from their debut in August 2010.

Data related to many different apps were accumulated over more than one year from a variety of Android Markets, and not only Google Play. Additionally, malware apps were thoroughly characterized based on their detailed behavior breakdown, including the installation, activation, and payloads.

Collected malware was split in 49 families, that were obtained by "carefully examining the related security announcements, threat reports, and blog contents from existing mobile antivirus companies and active researchers as exhaustively as possible and diligently requesting malware samples from them or actively crawling from existing official and alternative Android Markets" [4]. The defined families are: ADRD, AnserverBot, Asroot, BaseBridge, BeanBot, BgServ, CoinPirate, Crusewin, DogWars, DroidCoupon, DroidDeluxe, DroidDream, DroidDreamLight, DroidKungFu1, DroidKungFu2, DroidKungFu3, DroidKungFu4, DroidKungFu-Sapp, DoidKungFuUpdate, Endofday, FakeNetflix, FakePlayer, GamblerSMS, Geinimi, GGTracker, GingerMaster, GoldDream, Gone60, GPSSMSSpy, HippoSMS, Jifake, jSMSHider, Kmin, Lovetrap, NickyBot, Nickyspy, Pjapps, Plankton, RogueLemon, RogueSPPush, SMSReplicator, SndApps, Spitmo, TapSnake, Walkinwat, YZHC, zHash, Zitmo, and Zsone. Samples of 14 of the malware families were obtained from the official Android market, while samples of 44 of the families came from unofficial markets.

Bad-intentioned apps were then aggregated into 49 malware families [4], and information on those families is considered in present study. Thus, the analysed dataset consists of 49 samples (one for each family) and each sample has 26 different features. The features are divided into six categories; installation (repackaging, update, drive-by download, standalone), activation (BOOT, SMS, NET, CALL, USB, PKG, BATT, SYS, MAIN), privilege escalation (exploid, RATC/zimperlich, ginger break, asroot, encrypted), remote control (NET, SMS), financial charges (phone call, SMS, block SMS), and personal information stealing (SMS, phone number, user account). The values of those features are 0 (that feature is not present in that family) and 1 (the feature is present).

3.1 Results

MRMR, GA-ICC and GA-I algorithms were used for the selection of the best four features to characterize the above described Android malware families. The GA parameter setting used is the following:

- population size: 100.
- number of generations: 100
- number of runs for the algorithm in both cases (GA-ICC and GA-I): 50.

Firstly, the four first features have been selected by MRMR method, sorted by relevance and minimal redundancy. The selected features after the experiments are: Installation - Repackaging, Activation - SMS, Activation - BOOT, and Remote Control - NET.

Secondly, GA-INFO has been also applied to the same dataset for comparison purposes. Figure 1 displays the values obtained by ICC and mutual information (I) when running the GA for each one of the features. The Y axis shows the values of ICC and I respectively.

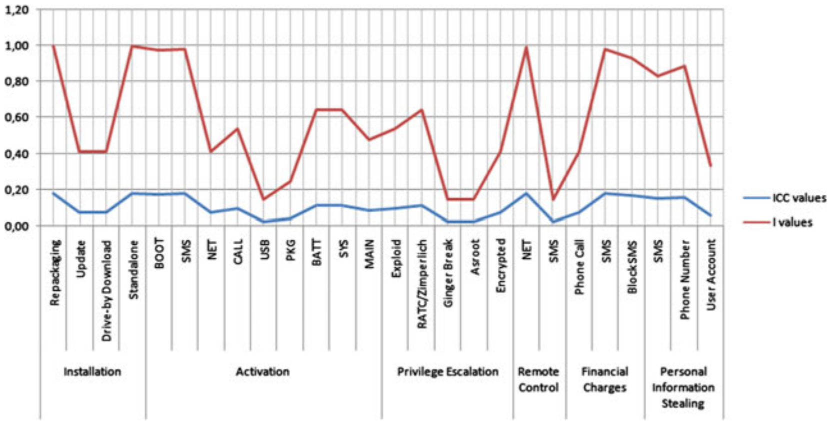

Fig. 1 ICC and I results for each feature

The GA-ICC and GA-I select two different individuals ([Installation - Repackaging, Installation - Standalone, Remote Control - NET, Activation - SMS] and [Installation - Repackaging, Installation –Standalone, Remote Control - NET, Financial Charges- SMS]) respectively, with the same fitness value for ICC and MI (0.18 and 0.99 respectively).

It should be noted that the GA methods were able to reach the optimum values in the population very early in the search process – around generation 11 (see Figs. 2 and 3). Each line represents a run of the algorithm, some lines overlap in some executions that were similar - and that is why 50 lines can not be identified. This is due to the relatively small number of features that had to be considered in the search, leading to an individual size easy to handle and quickly explore many feature subsets.

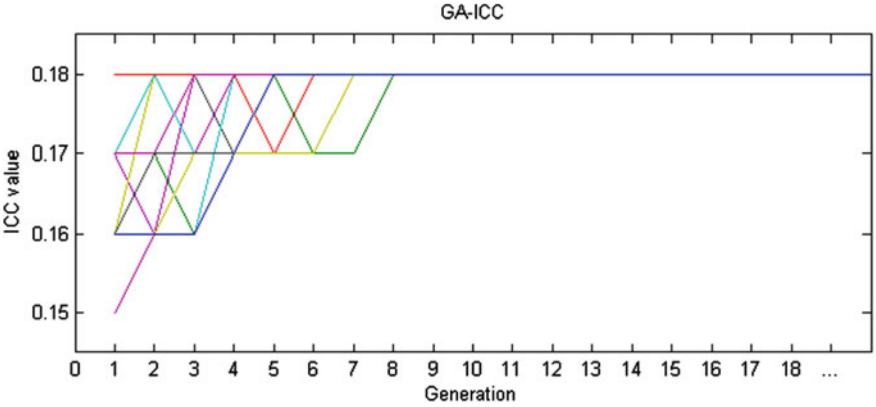

Fig. 2 Fitness ICC values in each generation of the 50 algorithm runs

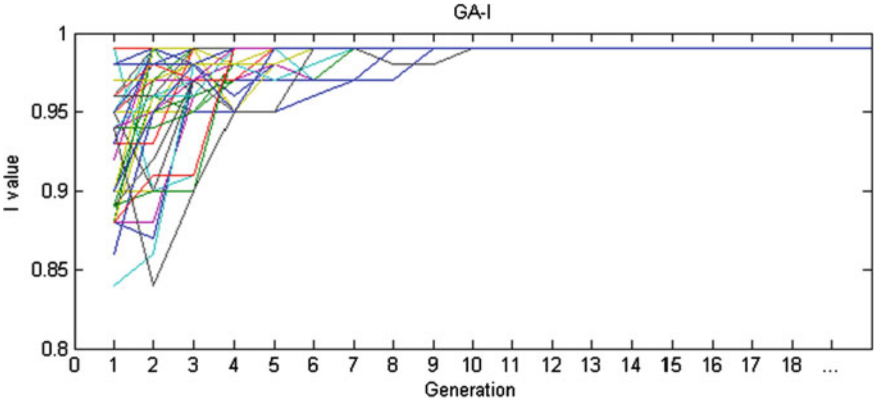

Fig. 3 Fitness I values in each generation of the 50 algorithm runs

The features selected by the three different algorithms are shown in Table 1 and, according to their relevance (being included in the selected subset), they are ordered in Table 2.

Table 1 Selected features by each one of the applied methods

Feature	MRMR	GA-I	GA-ICC
Installation - repackaging	√	√	√
Installation - standalone		√	√
Activation - SMS	√	√	
Activation - BOOT	√		
Remote control - NET	√	√	√
Financial charges - SMS			√

Table 2 Ordered list of selected features

Feature	Relevance
Installation - repackaging	100 %
Remote control - NET	100 %
Installation - standalone	66 %
Activation - SMS	66 %
Activation - BOOT	33 %
Financial charges - SMS	33 %

According to results shown in Tables 1 and 2, features 'Installation – Repackaging' and 'Remote Control – NET' have been selected by the three algorithms. Hence, it can be concluded that those features are the most relevant ones for the characterization of Android malware families.

The repackaging way of installation was defined by the authors of the dataset [4] as "one of the most common techniques malware authors use to piggyback malicious payloads into popular applications (or simply apps). In essence, malware authors may locate and download popular apps, disassemble them, enclose malicious payloads, and then re-assemble and submit the new apps to official and/or alternative Android Markets." Furthermore, from the collected samples, dataset authors found that 1,083 of them (or 86.0 %) were repackaged versions of legitimate applications with malicious payloads.

Regarding the remote control feature, dataset authors stated [4] that 93.0 % of the samples turn the infected phones into bots for remote control. Moreover, 1, 171 of the samples use the HTTP-based web traffic to receive bot commands.

In a second order of importance, 'Installation – Standalone' and 'Activation – SMS', together with 'Activation – BOOT' and 'Financial Charges – SMS' have been identified as key features for characterizing malware families.

4 Conclusions and Future Work

This paper has proposed several methods for selecting those features that best characterize malware families in Malgenome dataset. A genetic algorithm using a binary representation and a fitness function based on information theory measures (mutual information and information correlation coefficient) has been developed for the selection of the optimal subset of features in the considered problem. Experimental results show that the applied methods agree on the selection of 3 of the 4 major features.

Future work will extend these methods to consider other measures as fitness functions in evolutionary search or other population-based search heuristics. A hybridization of such methods and MRMR-based approaches will also be investigated. Additionally, the applicability of these methods to more datasets for the characterization of Android malware families will be further explored.

Acknowledgments This research has been partially supported through the project of the Spanish Ministry of Economy and Competitiveness RTC-2014-3059-4. The authors would also like to thank the BIO/BU09/14 and the Spanish Ministry of Science and Innovation PID 560300-2009-11.

References

1. Statista - The Statistics Portal, http://www.statista.com/statistics/266219/global-smartphone-sales-since-1st-quarter-2009-by-operating-system/
2. AppBrain Stats, http://www.appbrain.com/stats/stats-index
3. F-Secure: Q1 2014 Mobile Threat Report (2015)
4. Yajin, Z., Xuxian, J.: Dissecting android malware: characterization and evolution. In: 2012 IEEE Symposium on Security and Privacy **5**, 95–109 (2012)
5. Malgenome Project, http://www.malgenomeproject.org/
6. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. **3**, 1157–1182 (2003)
7. Larrañaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., Lozano, J.A., Armañanzas, R., Santafé, G., Pérez, A.: Machine learning in bioinformatics. Brief. Bioinform **7**(1), 86–112 (2006)
8. Ding, C., Peng, H.: Minimum redundancy feature selection from microarray gene expression data. J. Bioinform. Comput. Biol. **3**(02), 185–205 (2005)
9. Liu, H., Liu, L., Zhang, H.: Ensemble gene selection by grouping for microarray data classification. J. Biomed. Inform. **43**(1), 81–87 (2010)
10. Saeys, Y., Inza, I., Larrañaga, P.: A review of feature selection techniques in bioinformatics. Bioinformatics **23**(19), 2507–2517 (2007)
11. Hatami, N., Chira, C.: Diverse accurate feature selection for microarray cancer diagnosis. Intell. Data Anal. **17**(4), 697–716 (2013)
12. Vinod, P., Laxmi, V., Gaur, M.S., Naval, S., Faruki, P.: MCF: MultiComponent Features for malware analysis. In: 27th International Conference on Advanced Information Networking and Applications Workshops (WAINA), 2013, pp. 1076–1081 (2013)
13. Sanz, B., Santos, I., Laorden, C., Ugarte-Pedrero, X., Bringas, P.G.: On the automatic categorisation of android applications. In: 2012 IEEE Consumer Communications and Networking Conference (CCNC), pp. 149–153 (2012)
14. Sanz, B., Santos, I., Laorden, C., Ugarte-Pedrero, X., Bringas, P., Álvarez, G.: PUMA: Permission Usage to Detect Malware in Android. In: Herrero Á., Snášel V., Abraham A., Zelinka I., Baruque B., Quintián H., Calvo J.L., Sedano J., Corchado E. (eds.) International Joint Conference CISIS’12-ICEUTE´12-SOCO´12 Special Sessions, vol. 189. Springer, Berlin, Heidelberg. pp. 289–298 (2013)
15. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley, New York (1991)