مجلة

كلية التـــراث الجامعة

مجلة علمية محكمة

متعددة التخصصات نصف سنوية

العدد التاسع والثلاثون

عدد خاص بوقائع المؤتمر العلمي السنوي السادس عشر (الدولي الخامس)

18 نيسان 2024

ISSN 2074-5621

رقم الايداع في دار الكتب والوثائق 719 لسنة 2011

مجلة كلية التراث الجامعة معترف بها من قبل وزارة التعليم العالي والبحث العلمي بكتابها المرقم
(ب 3059/4) والمؤرخ في (4/7 /2014)

ANNOUNCEMENT OF MACHINE LEARNING-BASED ANDROID MALWARE

SARI KHDHAER MUKHLIF

Al-Turath University

Abstract

Malware that uses the Internet to commit exploits is becoming more and more commonplace. Due to the widespread occurrence of malware, manual malware identification is no longer reliable or efficient. Therefore, it appears that autonomous behavior-based malware detection through machine learning approaches is a good way forward. Numerous research has demonstrated how well machine learning works to identify and categorize malware files. Several machine learning techniques have been studied in this study to identify malicious software applications such as Random Forest, SVM, KNN, Decision Tree, and Logistic Regression. Adware, Benign, Ransomware, SMS Malware, and Scareware are the five classes that have been used for classification. The trials' results demonstrated the value and efficacy of machine learning techniques in identifying malware, with these algorithms achieving 71% accuracy. Keywords: Machine learning, ensemble learning, static and dynamic analysis, multiple classifier systems, Android malware, and malware for the Android attack.

الملخص

أصبحت البرامج الضارة التي تستخدم الإنترنت لارتكاب عمليات استغلال أكثر شيوعًا. نظرًا لانتشار البرامج الضارة على نطاق واسع، لم يعد التعرف اليدوي على البرامج الضارة موثوقًا أو فعالاً. لذلك، يبدو أن اكتشاف البرامج الضارة القائم على السلوك بشكل مستقل من خلال أساليب التعلم الآلي يعد طريقة جيدة للمضي قدمًا. أظهرت العديد من الأبحاث مدى نجاح التعلم الآلي في تحديد ملفات البرامج الضارة وتصنيفها. تمت دراسة العديد من تقنيات التعلم الآلي في هذه الدراسة لتحديد التطبيقات البرمجية الضارة مثل Random Forest وSVM وKNN وDecision Tree والانحدار اللوجستي. Adware وBenign وRansomware وSMS Malware وScareware هي الفئات الخمس التي تم استخدامها للتصنيف. وأظهرت نتائج التجارب قيمة وفعالية تقنيات التعلم الآلي في تحديد البرامج الضارة، حيث حققت هذه الخوارزميات دقة بنسبة 71%. الكلمات الرئيسية: التعلم الآلي، التعلم الجماعي، التحليل الثابت والديناميكي، أنظمة التصنيف المتعددة، البرامج الضارة لنظام Android، والبرامج الضارة لهجوم Android.

1. Introduction

The malware is specifically made to attack the security policy of the phone system and prevent damage or unauthorized access. Malware falls into several categories, such as malware, B. Adware, smallware, and ransomware. Experts predict that by 2023, over 260 billion applications will have been downloaded, up from over 205 billion in 2018. Just the first half of 2021 saw the download of almost 57 billion apps [1]. Just 0.08% of Android devices running Google Play apps were affected by potentially hazardous applications (PHAs), but 0.68% of Android devices running non-Google Play apps had PHA infections [2]. The swift spread of malware on Android devices has posed significant obstacles for the anti-malware system. since the malware analysis system cannot keep up with the volume of malware samples. By grouping malware samples into discrete groups and using the common malware characteristics among them to recognize and search for malware, malware analysis can be expedited. However, the following two factors reduce the accuracy of the classification conclusions made today: As malicious components are typically inserted into popular applications to create malware, legitimate malware might first trick users into using categorization methods. Android malware that is polymorphic can avoid detection by changing up how it attacks [3]. Moreover, a new generation of signatures that use mobile test pilots for malware research are useless because malware is so widespread these days. Nonetheless, due to its capacity to recognize and categorize malware, dynamic analysis has garnered a great deal of attention. Malware detection can be greatly aided by machine learning-based dynamic analysis techniques like logistic regression, decision trees, random forests, knn, etc. Machine learning methods that provide high and low FPR accuracy must be used to the training data in order to address the issue of spotting Android malware from several perspectives. [4]; [5].

2. Literature Review

2.1 Malware

Malware is software that ignores user preferences on how to use a computer or network. Malware is software designed to be used by criminals and for use in political, illegal, and risky actions [6]. Due to the large number of patterns, malware evaluation structures often rely on allocated computational resources to process all available data efficiently. Thus, the key component of those systems' general consistency and effectiveness is how the evaluation responsibilities are divided among the community nodes: We refer to this feature as Scheduled. Over the past few years, malware has emerged as the biggest threat to the records industry. An impartial IT security organization called AV-Test claims that the variety of malware is expanding annually at an unprecedented rate, notwithstanding the application of techniques for malware detection [7]. virus is typically created by groups of programmers who, more often than not, are genuinely attempting to gain money, either by disseminating the virus themselves or by selling it to the highest bidder on the foolish Internet. Malware may also be created for other purposes, such as testing security, serving as a tool for protest, or serving as a weapon of conflict between governments.

2.2 Malware Types

a. Adware: A security risk is malware. Usually, this is used to gather information about advertising or play ads so you can make money. This opportunity is not an easy extra-ordinary area like the traditional risk. But it also takes advantage of more potent tactics than those used in conventional malware. Software that goes beyond the reasonable advertisements one could anticipate from shared or open source software is known as adware. Spyware is usually installed separately from a computer at the same time or in a comparable manner. Spyware typically continues to produce advertisements even when the user isn't using the required software [8]. It's common, especially for mobile apps, for the program to be unresponsive while displaying an advertisement banner. Then, this demanding banner is removed upon purchasing the entire model. But since it's a component of the package, this cannot be regarded as marketing or marketing software.

b. Benign: Unintentional discovery may manifest as the records being posted on the business's open website or being sent by mail, fax, or email to the wrong party.

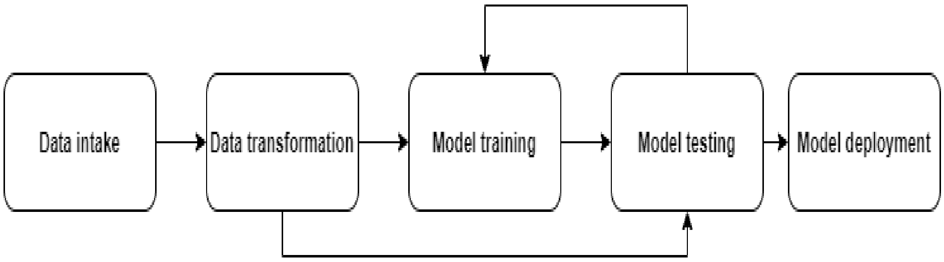
c. Ransomware: Users might suffer significant harm as a result of ransomware, a unique kind of software that can encrypt files and lock victims' displays in return for money. Researchers can reduce effort by grouping ransomware into families and using this information to identify variants of a known malware pattern [9].

d. SMS Malware: Spam, which includes SMS, WhatsApp, and other messaging services, is known to pose a risk to mobile device security by infecting mobile devices with malware. A mobile device may be used to send spam as a result of the security lapse [10].

3. Methodology

In order to investigate application collusion, this work suggests a two-stage classification model: A succession of linear vector machines (SVM) and the KMeans clustering method comprise the hybrid classifier, which is the initial stage. KMeans is an algorithm that divides the data set into several categories and is used in the first stages of both benign and harmful applications. creating a parameter vector using the parameters derived from the linear SVM training for every group. Parameter vectors and light discriminant functions are employed in the second step of the classification model to identify application collusion. Both single harmful applications and complicated application pairs are visible to this model. The model is simple to use and requires little processing power. An opcode-based technique for analyzing Android malware is presented in this paper. Malicious Android applications have been categorized using several machine learning methods. Based on the data, this technique has a 99.5% accuracy rate and a TPR of 0.995, indicating that it can more precisely and efficiently arrange malware. Machine learning techniques have been applied to swiftly browse through static analysis of malware on Android devices. Machine learning techniques are crucial for the investigation and categorization of Android malware Figure 2.6: Neural Network Layer 25. Their method of analyzing malware for Android is based on opcodes. Numerous data have been gathered in order to complete this study. The two main components of this study are the classification of ransomware and the discovery of behavioral variables that can be utilized to achieve the highest classification accuracy.

FIGURE 1: THE GENERAL WORKFLOW PROCESS[9]



3.1 Data Preparation

The first step in the research is to detect potential assaults using machine learning classification. Static analysis refers to malware analysis that is conducted on the system rather than without executing on it. If not, it's referred to as dynamic analysis. Although static analysis has a rapid low generation false positive rate (FPR), it is unable to detect undetectable malware due to its fixed functions, which are susceptible to obfuscation tactics. However, because numerous features will result in a decrease in inaccuracy, these methods require time-consuming learning and information-gathering procedures. Because of the high false-positive rate (FPR), floor installation is not viable. Further study is needed to create techniques that use less computing and cut down on redundancies.

3.2 Research Questions

THIS STUDY AIMS TO ADDRESS THE FOLLOWING QUESTIONS: CAN MACHINE LEARNING TOOLS AND METHODS BE USED TO CREATE MODELS FOR THE CATEGORIZATION OF MALWARE ATTACKS THAT CAN DISTINGUISH BETWEEN ATTACKS FROM VARIOUS MALWARE FAMILIES? HOW ACCURATE AND SUCCESSFUL ARE THESE MODELS, AND CAN WE RELY ON THEM TO FORESEE FUTURE ATTACKS?

4. Experimental Results

To help you fully grasp the research's findings, the following words have been defined:

a. Accuracy: The total amount of apps that are appropriately classified as harmful or benign.

$$\mathbf{Accuracy} = \frac{\mathbf{tp + tn}}{\mathbf{tp + tn + fp + fn}} \qquad (\mathbf{1})$$

b. **The F-Score:** The F1-score, also called the model's accuracy, is a measure of a model's accuracy on a dataset. It is used to assess binary classification techniques, which categorize examples into "positive" or "negative" categories. Integrating the model's precision and recall may be possible with the F-score, which is defined as the concordant combination of the model's accuracy and recall. In standard dialect preparation and data recovery frameworks such as look motors, the F-score is widely used to evaluate different types of machine learning models [11]. Formula One Points:

$$\mathbf{F_1} = \frac{\mathbf{2}}{\frac{\mathbf{1}}{\mathbf{recall}} \times \frac{\mathbf{1}}{\mathbf{precision}}} = \mathbf{2} \times \frac{\mathbf{precision\ \times\ recall}}{\mathbf{precision + recall}} = \frac{\mathbf{tp}}{\mathbf{tp} + \frac{\mathbf{1}}{\mathbf{2}}(\mathbf{fp + fn})} \quad (\mathbf{2})$$

c. **Precision** is how the situations that the show deems positive are distributed with simple, positive examples. Stated otherwise, the total of true positives and false positives equals the number of actual positives [11].

$$\mathbf{precision} = \frac{\mathbf{tp}}{\mathbf{tp + fp}} \quad (\mathbf{3})$$

d. **Recall:** Affectability, the number of true positives is divided from the number of false negatives by the number of true positives plus false negatives [11]. This is the division of cases classified as positive among all positive illustrations, as well as the division of patients classified as positive among all positive specimens.

$$\mathbf{recall} = \frac{\mathbf{tp}}{\mathbf{tp + fn}} \quad (\mathbf{4})$$

4.1 Binary Classification:

RESULTS OF BINARY CLASSIFICATION: AS THE NAME IMPLIES, BINARY CLASSIFICATION ONLY PRODUCES TWO CLASSES: 0 AND 1. TABLE 4-1 SHOWS THAT ONLY THE BENIGN AND RANSOMWARE CHARGE CLASSIFICATIONS HAVE BEEN APPLIED TO THE DATA.

TABLE 4-1 THE BINARY CLASSIFICATION'S ACCURACY, F1, PRECISION, AND RECALL SCORE.

MODEL\MEASURE	ACCURACY	F1_SCORE	PRECISION SCORE	RECALL SCORE
Binary Classification	92.18 %	78.34 %	87.65 %	74.65 %

The preceding table shows that the model can typically predict a virus's class with high accuracy. This is expected when learning and making predictions because Binary Classification has only two classifications. This is evident from the accuracy result, which is calculated by dividing the total number of predictions by the number of accurate forecasts.

4.2 Multi-Class Classification:

The model's output is shown below, with over 60% accuracy being the highest. As previously stated, we can remove the classifier's primary characteristic and discard the other characteristics by segmenting the features. Table 4-2 shows that the findings of the multi-class classification were not as accurate as those of the binary classification.

TABLE 4-2 THE RECALL SCORE, ACCURACY, F1, AND PRECISION BASED ON THE MULTI-CLASS CATEGORIZATION.

MODEL\MEASURE	ACCURACY	F1_SCORE	PRECISION SCORE	RECALL SCORE
KNN	62.89 %	56.86 %	60.86 %	54.50 %
Random Forest	71.28 %	66.05 %	71.92 %	62.83 %
Logistic Regression	59.07 %	31.56 %	53.18 %	36.41 %
Decision Tree	68.87 %	63.57 %	63.44 %	64.50 %
SVM	58.13 %	43.45 %	57.53 %	46.74 %

Table 4-2 shows that models have increasing difficulty when it comes to differentiating between multiple classes. It is actually possible to conclude from the results of most models that the predictions are more akin to guesswork with accuracy rates close to 50%, with the exception of the random forest model, which is able to provide better predictions because it passes over the data multiple times during multiple Decision Trees and is therefore more capable of the prediction because it learns patterns in the data better. The table's conclusion highlights the need for both an adjustment to the prediction method and additional data preparation steps. As was already said, the voting group approach is used to rate the data with the greatest vote score, which further increases accuracy even further, reaching 71.67%, as Table 4-3 below shows.

Table 4-3 the recall score, accuracy, F1, and precision following the ensemble.

Model\Measure	Accuracy	F1_score	Precision score	Recall score
After Ensemble	72.67 %	66.31 %	72.48 %	63.52 %

After the data preparation and ensemble procedure, the final result is shown in Table 4-3, and it is evident that this yields better results than merely speculating about the multi-class classification predictions. The accuracy scores still require improvement, in particular, the previously discussed Recall score. This takes us to the topic of discussion, future initiatives, and steps that may be taken to enhance the results.

5. Conclusion

In conclusion, a variety of data transactions, uploads, downloads, and other internet activities lead to a daily influx of malware attacks on Android and other devices. Because of this, assaults like Adware, Benign, Ransomware, SMSmaleware, and Scareware cannot be defended against using conventional techniques or static protection. Both the attacks and their users will gain from this. Rather, dynamic Rather than being specifically and precisely programmed, dynamic Artificial Intelligence and Machine Learning solutions offer a superior method of assessing and categorising malware attacks and forecasting a new attack mission. In order to create a model for forecasting recent attacks, machine learning classification methods from previous malware attacks were employed in this work. Supervised and unsupervised learning are the two main functions of machine learning. Unsupervised learning refers to techniques that make use of unlabeled databases, whereas supervised learning refers to algorithms that use labeled databases, such as classification. This is the situation in this study, which makes advantage of previously labeled records to describe malware infections. On the other hand, unsupervised learning proposes calculations like clustering. Important steps in the machine learning process include data extraction and gathering, data analysis and reprocessing, model training, testing, and evaluation. Before going on to the following stage, which was getting the data ready for input into the training mode, data processing included obtaining the data, sorting it, removing outliers and extraneous information, filling in the blanks, rearranging the data, scaling it, and extracting its features. Many classification models, such as SVM, Decision Tree, Random Forest, Logistic Regression, and K-nearest neighbors, demonstrated binary and multi-classification throughout the training phase. The study's binary classification results were initially quite encouraging, but since this is not how things actually function in the real world, extra classifications have been added.

References

[1] Imtiaz, S. I. et al. (2021) ‘Deep AMD: Detection and identification of Android malware using high-efficient Deep Artificial Neural Network’, Future Generation Computer Systems, 115, pp. 844–856. doi: https://doi.org/10.1016/j.future.2020.10.008.

[2] Marlene. (2018) ‘Global Android devices with potentially harmful apps (PHAs).’

[3] Fan, M. et al. (2016) ‘Frequent Subgraph Based Familial Classification of Android Malware’, in 2016 IEEE 27th International Symposium on Software Reliability Engineering (ISSRE), pp. 24–35. doi: 10.1109/ISSRE.2016.14.

[4] Utku, A. and Doğru, İ. A. (2017) ‘Malware detection system based on machine learning methods for Android operating systems’, in 2017 25th Signal Processing and Communications Applications Conference (SIU), pp. 1–4. doi: 10.1109/SIU.2017.7960231.

[5] Joshi, S., Upadhyay, H., Lagos, L., Akkipeddi, N.S. and Guerra, V., (2018) ‘Machine learning approach for malware detection using random forest classifier on process list data structure.’ In Proceedings of the 2nd International Conference on Information System and Data Mining (pp. 98-102).

[6] Tyler, M. and Marie V,. (2011) “Stop Badware Project.’ Reference: http://www.stopbadware.org. [online] Available: https://www.av-test.org/en/statistics/malware/.

[7] Galal, H.S., Mahdy, Y.B. and Atiea, M.A. (2016) ‘Behavior-based features model for malware detection.’ J Comput Virol Hack Tech 12, 59–67 https://doi.org/10.1007/s11416-015-0244-0.

[8] Ndagi, J. Y. and Alhassan, J. K. (2019) ‘Machine Learning Classification Algorithms for Adware in Android Devices: A Comparative Evaluation and Analysis’, in 2019 15th International Conference on Electronics, Computer and Computation (ICECCO), pp. 1–6. doi: 10.1109/ICECCO48375.2019.9043288.

[9] Zhang, H. et al. (2019) ‘Classification of ransomware families with machine learning based onN-gram of opcodes’, Future Generation Computer Systems, 90, pp. 211–221. doi:

[10] Khan, J., Abbas, H. and Al-Muhtadi, J. (2015) 'Survey on Mobile User's Data Privacy Threats and Defense Mechanisms', Procedia Computer Science, 56, pp. 376–383. doi: https://doi.org/10.1016/j.procs.2015.07.223.
[11] Wood, T. (2019). "F-score." Retrieved February 17, 2021.