A NOVEL APPROACH FOR GENETIC ALGORITHM BASED OPTIMIZED FEATURE SELECTION FOR DETECTING MALWARE IN ANDROID USING ML

PALADUGU VINOD, SURE VENKATA NAGA VISWESH, SYED AFZAL PARVEZ, YADLAPATI VENKATA SUBBARAO, Student, Department of CSE, NRI INSTITUTE OF TECHNOLOGY, Vijayawada, A.P., India.

Mrs. S. NAHIDA, Associate Professor, Department of CSE, NRI INSTITUTE OF TECHNOLOGY, Vijayawada, A.P., India.

ABSTRACT

In this paper proposes an Effectual machine-learning based approach for Android Malware Detection making use of evolutionary Genetic algorithm for Discriminatory feature selection. Selected features from Genetic Algorithm are used to train machine learning classifiers and their Capability in identification of Malware before and after feature Selection is compared. The experimentation results validate that Genetic algorithm gives most optimized feature subset helping in Reduction of feature dimension to less than half of the original Feature-set. Classification accuracy of more than 94% is Maintained post feature selection for the machine learning based Classifiers, while working on much reduced feature dimension, Thereby, having a positive impact on computational complexity of learning classifiers.

Keywords — Android malware analysis; feature selection; Genetic algorithm; machine learning; reverse-engineering

INTRODUCTION

Android platform due to open source characteristic And Google backing has the largest global market share. Being the world's most popular operating system, it has drawn the Attention of cyber criminals operating particularly through wide Distribution of malicious applications. Android Apps are freely available on Google Playstore, the official Android app store as well as third-party app stores for users to download. Due to its open source nature and popularity, malware writers are increasingly focusing on developing malicious applications for Android operating system. In spite of various attempts by Google Playstore to protect against malicious apps, they still find their way to mass market and cause harm to users by misusing personal information related to their phone book, mail accounts, GPS location information and others for misuse by third

parties or else take control of the phones remotely.

Therefore, there is need to perform malware analysis or reverse-engineering of such malicious applications which pose serious threat to Android platforms. Broadly speaking, Android Malware analysis is of two types: Static Analysis and Dynamic Analysis. Static analysis basically involves analyzing the code structure without executing it while dynamic analysis is examination of the runtime behavior of Android Apps in constrained environment. Given in to the ever-increasing variants of Android Malware posing zero-day threats, an efficient mechanism for detection of Android malwares is required. In contrast to signature-based approach which requires regular update of signature database.

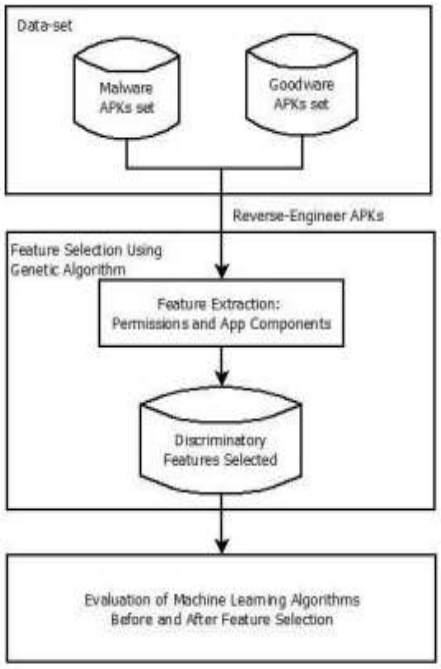


Fig. 1. Proposed Methodology

LITERATURE SURVEY

Mobile gadgets, such as smartphones, iPads, and computer tablets, have become daily needs to accomplish essential activities, including education, paying bills online, bank transactions, employment information, and pleasure. Based on the information from an online mobile device manufacturing website, Android is one of the prominent operating systems (OS) utilized by manufacturers (Rayner, 2019; Jkielty, 2019). (Rayner, 2019; Jkielty, 2019). The opensource framework of Android has helped the smartphone makers in creating Android devices of different sizes and kinds, such as smartphones, smart watches, smart TVs, and smart eyewear. In the most recent decades, the amount of amazing Android devices accessible globally has grown from 38 in 2009 to over 20,000 in 2016 (Android, 2019a) (Android, 2019a). As a consequence of the demand for this Android OS, the latest data from Statista showed that the number of Android malware rise to 26.6 million in March 2018 (Statista, 2019). (Statista, 2019). More over, McAfee identified a virus known as Grabos, which exploits the Android and breaks Google Play Store security (McAfee, 2019). (McAfee, 2019). It was also estimated that 17.5 million Android

devices have downloaded this Grabos mobile virus before they were taken down.

Mobile malware is intended to disable a mobile device, enable malevolent actions to remotely control the device, or steal personal information (Beal, 2013). (Beal, 2013). Moreover, these harmful actions able to execute silently and circumvent permission if the Android kernel is infected by mobile malware (Ma & Sharbaf, 2013; Aubrey-Derrick Schmidt et al., 2009b) (Ma & Sharbaf, 2013; Aubrey-Derrick Schmidt et al., 2009b). In September 2019, a total of 172 harmful apps were discovered on Google Play Store, with roughly 330 million instals. According to experts, the harmful components were concealed within the functioning apps. After the apps are downloaded, it leads to the display of popup advertising, which continue visible even when the application was closed (O'Donnell, 2019). To identify this virus, security practitioners performing malware analysis, which attempts to analyse the malicious features and behaviours. There are dynamic, static, and hybrid analyses. To merge the features of the static and dynamic approach, three-layer detection model called SAMAdroid has been developed by Saba Arshad et al. (2018) which integrates static and dynamic properties. Mobile Sandbox by Spreitzenbarth et al. (2015) which suggested to utilise the findings of static analysis to drive the dynamic analysis and ultimately achieve categorization. The hybrid analysis method is excellent to assist in increasing the accuracy, but it also has a significant disadvantage such as the waste of time and space for the large number of malware samples to be identified and analysed (Fang et al., 2020; Alswaina & Elleithy, 2020). (Fang et al., 2020; Alswaina & Elleithy, 2020).

EXISTING SYSTEM

The main contribution of the work is reduction of feature dimension to less than half of original feature-set using Genetic Algorithm such that it can be fed as input to machine learning classifiers for training with reduced complexity while maintaining their accuracy in malware classification. In contrast to exhaustive method of feature selection which requires testing for 2N different combinations, where N is the number of features, Genetic Algorithm, a heuristic searching approach based on fitness function has been used for feature selection. The optimized feature set obtained using Genetic algorithm is used to train two machine learning algorithms: Support Vector Machine and Neural Network. It is observed that a decent classification accuracy of more than 94%

is maintained while working on a much lower feature dimension, thereby, reducing the training time complexity of classifiers.

PROPOSEDSYSTEM

- Two set of Android Apps or APKs: Malware/Good ware are reverse engineered to extract features such as permissions and count of App Components such as Activity, Services, Content Providers, etc. These features are used as feature vector with class labels as Malware and Good ware represented by 0 and 1respectivelyinCSVformat.
- To reduce dimensionality of feature-set, the CSV is fed to Genetic Algorithm to select the most optimized set of features. The optimized set of features obtained is used for training two machine learning classifiers: Support Vector Machine and Neural Network.
- In the proposed methodology, static features are obtained from AndroidManifest.xml which contains all the important information needed by any Android platform about the Apps. Andro guard tool has been used for disassembling of the APKs and getting the static features.

Advantages of proposed system:

- Security

- Proposed an overland efficient algorithm for feature selection on improve overall detection accuracy.

SAMPLE RESULTS

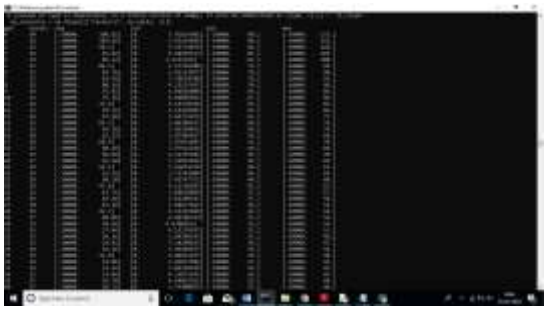
Here we are uploading 'AndroidDataset.csv' file and after upload will get below screen

Generate Train & Test Model' button to split dataset into train and test part. All machine learning algorithms will take 80% dataset for training and 20% dataset to test accuracy of trained model.

we can see there are total 3799 android app records are there and application using 3039 records for training and 760 records for testing. Now we have both train and test model and now

we got 98% accuracy for SVM and now click on 'Run SVM with Genetic Algorithm' button to choose optimize features and then run SVM on optimize features to get accuracy

screen SVM with Genetic algorithm got 93% accuracy. Genetic with SVM accuracy is less but its execution time will be less which we can see at the time of comparison graph.

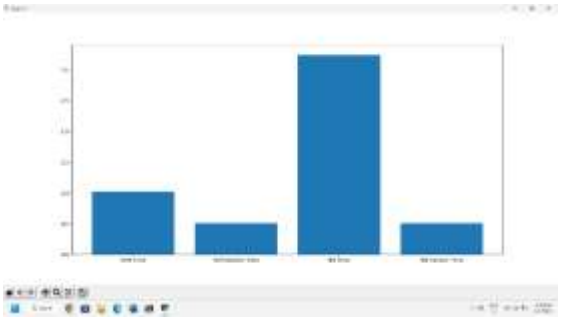


In above console we can see genetic algorithm chooses 40 features from all dataset features

In above screen neural network also gave 98.64% accuracy. Now click on Run Neural Network with Genetic Algorithm 'button to get NN accuracy with genetic algorithm.

In above screen NN with genetic got 95.39% accuracy. Now click on Accuracy Graph 'button to see all algorithms accuracy in graph

In above graph x-axis represents algorithm name and y-axis represents accuracy and in all SVM got high accuracy. Now click on _Execution Time Graph' button to get execution time of all algorithm.



In above graph x-axis represents algorithm name and y-axis represents execution time. From above graph we can conclude that with genetic algorithm machine learning algorithms taking less time to build model

CONCLUSION

As the number of threats posed to Android platforms is increasing day to day, spreading mainly through malicious applications or malwares, therefore it is very important to design a framework which can detectsuch malwares with accurate results. Where signature-based approach fails to detect new variants of malware posing zero-day threats, machine learning based approaches are being used. The proposed methodology attempts to make use of evolutionary Genetic Algorithm to get most optimized feature subset which can be used to train machine learning algorithms in most efficient way.

FUTURE SCOPE FOR FURTHER DEVELOPMENT

From experimentations, it can be seen that adecent classification accuracy of more than 94% is maintained using Support Vector Machine and Neural Network classifiers while working on lower dimension feature-set, thereby reducing the training complexity of the classifiers Further work can been hanced using larger datasets for improved results and analyzing the effect on other machine learning algorithms when used in conjunction with Genetic Algorithm.

REFERENCES

[1] N. Milosevic, A. Dehghantanha, and K. K. R. Choo, –Machine learning aided Android malware classification, ‖Comput.Electr.Eng.,vol.61,pp.266–274,2017.

[2] J. Li, L. Sun, Q.Yan, Z. Li, W. Srisa-An, and H. Ye, –Significant Permission Identification for Machine-Learning-Based Android Malware Detection,‖ IEEE Trans.

Ind. Informatics, vol. 14, no. 7, pp.3216–3225,2018.

[3] A.Saracino,D.Sgandurra,G.Dini,andF.Martinelli,–MADAM:Effectiveand EfficientBehavior- based Android Malware Detection and Prevention,‖ IEEE Trans. Dependable Secur. Comput., vol. 15,no.1,pp.83–97,2018.

[4] S.Arshad,M.A.Shah,A.Wahid,A. Mehmood,H.Song,andH.Yu,–SAMADroid:A Novel3- Level Hybrid Malware Detection Model for Android Operating System,‖ IEEE Access, vol. 6, pp.4321–4339,2018.

[5] Lim, K.; Kim, N.Y.; Jeong, Y.; Cho, S.j.; Han, S.; Park, M. Protecting Android Applications with Multiple DEX Files Against Static Reverse Engineering Attacks. Intell. Autom. Soft Comput. 2019, 25, 143–154.

[6] Meimandi, A.; Seyfari, Y.; Lotfi, S. Android malware detection using feature selection with hybrid genetic algorithm and simulated annealing. In Proceedings of the 2020 IEEE 5th Conference on Technology In Electrical and Computer Engineering (ETECH 2020) Information and Communication Technology (ICT), Tehran, Iran, 22 October 2020.

[7] Wang, L.; Gao, Y.; Gao, S.; Yong, X. A New Feature Selection Method Based on a Self-Variant Genetic Algorithm Applied to Android Malware Detection. Symmetry 2021, 13, 1290

[8] Sahin, D.Ö.; Kural, O.E.; Akleylek, S.; Kılıç, E. A novel Android malware detection system: adaption of filter-based feature selection methods. J. Ambient. Intell. Humaniz. Comput. 2021, 15, 1–15.