IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 12, December 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 12, December 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1012028 |

Behaviour Based Malware Detection Using Machine Learning

Mr. Pradeep Sachin, Mrs.S.Suganya M. E,

Department of Information Technology, College: KSR College of Engineering, Tiruchengode, India

ABSTRACT: During the last few years, several approaches have been proposed for detection of Android malware Apps, each usually using its own dataset. Generating a representative Android malware dataset to evaluate malware detection approaches is a challenging task. Recently, the Canadian Institute for Cybersecurity released the CICAndMal2017 dataset, which includes recent and sophisticated Android samples spanning between five distinct categories: Adware, Ransomware, SMS malware, Scareware, and Benign. The best classification result obtained for this dataset was with a Precision of 95.3%, achieved with the Random Forest algorithm, using Permissions and Intents as static features. In this paper, we investigate the usage of nine machine learning algorithms to classify malware in the above mentioned dataset. The comparison of the obtained results is performed with the ones obtained with Random Forest, including performance evaluation (in terms of Precision, Recall, F-Measure, and Accuracy) and resource usage (in terms of execution time and CPU and memory consumption). Besides, we also investigate the use of a non-sliding Bag of System Calls algorithm with the above mentioned machine learning algorithms. It is shown that the Adaboost algorithm, using the Random Forest as a base estimator, leads to the best classification results with an Accuracy of 98.24%, a Precision of 99.31% (for malware), and an F1-Measure of 95.05% (for malware), at the cost of a larger execution time than Random Forest.

I. INTRODUCTION

The problem to be examined involves the high spreading rate of computer malware (viruses, worms, Trojan horses, rootkits, botnets, backdoors, and other malicious software) and conventional signature matching-based antivirus systems fail to detect polymorphic and new, previously unseen malicious executables. Malware are spreading all over the world through the Internet and are increasing day by day, thus becoming a serious threat. The manual heuristic inspection of static malware analysis is no longer considered effective and efficient compared against the high spreading rate of malware. Nevertheless, researches are trying to develop various alternative approaches in combating and detecting malware. One proposed approach (solution) is by using automatic dynamic (behavior) malware analysis combined with data mining tasks, such as, machine learning (classification) techniques to achieve effectiveness and efficiency in detecting malware.

RELATED WORK

Malware has threatened the organizations for a long time and still have not made a lot of progress in detecting the malware on time. Malware can easily harm the system by executing the unnecessary services that will put the load on the system and hinder its smooth running. There are basically two methods to detect the malware, one being the old process of detecting the malware based on the signature and the other one being the Behavior based method. The behavior of the malware is defined by the task the malware performs when it gets activated in the machine, for example, running the Operating System services, downloading the infected files from the internet. The proposed algorithm detects the malware based on its behavior. In this paper, the proposed model is the combination of Support Vector Machine and Principle Component Analysis. This proposed model achieved an accuracy of 97.75% during validation with 97% precision, 99% recall and f1-score of.98 for actual Malwares.

II. LITERATURE SURVEY

1. A Machine Learning Technique to Detect Behavior Based Malware

Author- Shubham Chaudhary, Anchal Garg

Malware has threatened the organizations for a long time and still have not made a lot of progress in detecting the malware on time. Malware can easily harm the system by executing the unnecessary services that will put the load on the system and hinder its smooth running. There are basically two methods to detect the malware, one being the old process of detecting the malware based on the signature and the other one being the Behavior based method. The

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 12, December 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1012028 |

behavior of the malware is defined by the task the malware performs when it gets activated in the machine, for example, running the Operating System services, downloading the infected files from the internet. The proposed algorithm detects the malware based on its behavior. In this paper, the proposed model is the combination of Support Vector Machine and Principle Component Analysis. This proposed model achieved an accuracy of 97.75% during validation with 97% precision, 99% recall and f1-score of.98 for actual Malwares.

2. Analysis of Machine learning Techniques Used in Behavior-Based Malware Detection
Author - Ivan Firdausi,Charles Lim
The increase of malware that are exploiting the Internet daily has become a serious threat. The manual heuristic inspection of malware analysis is no longer considered effective and efficient compared against the high spreading rate of malware. Hence, automated behavior-based malware detection using machine learning techniques is considered a profound solution. The behavior of each malware on an emulated (sandbox) environment will be automatically analyzed and will generate behavior reports. These reports will be preprocessed into sparse vector models for further machine learning (classification). The classifiers used in this research are k-Nearest Neighbors (kNN), Naïve Bayes, J48 Decision Tree, Support Vector Machine (SVM), and Multilayer Perceptron Neural Network (MlP). Based on the analysis of the tests and experimental results of all the 5 classifiers, the overall best performance was achieved by J48 decision tree with a recall of 95.9%, a false positive rate of 2.4%, a precision of 97.3%, and an accuracy of 96.8%. In summary, it can be concluded that a proof-of-concept based on automatic behavior-based malware analysis and the use of machine learning techniques could detect malware quite effectively and efficiently.

III. EXISTING SYSTEM

The static analysis consists in the analysis of an Android application, without executing the application, through reverse engineering, to verify their reliability and safety. In this way the application is scanned on a server and information such as hash, Permissions and source code are collected and processed by a classification algorithm that indicates whether it is or not a malware. The static features of Android applications extracted in this work are based on the information provided by the manifest file. The application file APK is the package file format used by the Android operating system on mobile platform, that may contain the files, classes.dex, AndroidManifest.xml, directories and resource files. The plaintext file AndroidManifest.xml obtained through decompilation is a manifest file that defines the running information of the entire application. Permissions and Intents were the characteristics chosen to define and characterize each of the applications. Permissions and Intents are defined in the manifest, Permissions refer to assigning or denying access to content.

Disadvantages

- A detection rate of 96% was achieved by training a support vector machine classifier through the optimization of a hinge loss function.
- The support vector machine classifier was trained on registry activity data generated by software from both classes.

PROPOSED SYSTEM

This section presents the workflow of the two experiments. The experimental workflow design of three stages is illustrated. The first stage is the collection of data that collects the network traffic and passes it on to the next phase. TCP packets shall be filtered during the second phase, the feature selection and extraction, and functions chosen from the various network features are extracted, labelled and stored in a next phase database. The machine learning classification entails the last phase, in which the information in a database forms the classification machine to produce a detection.

- Data Collection
- Feature Selection and Extraction
- Machine Learning Approach for Malware Detection

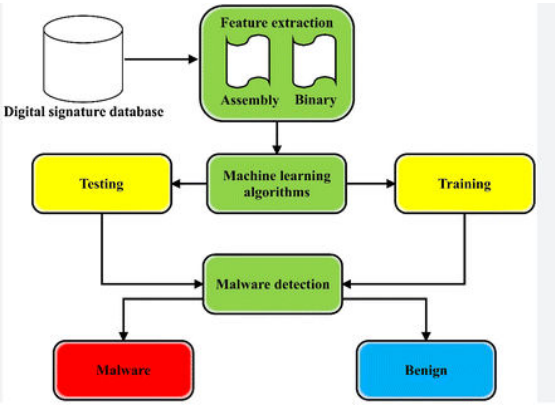
ADVANTAGES

The four Windows virtual machines were run concurrently, each with an instance of benign or malicious software. During the analysis task, a memory dump was produced of each Windows virtual machine along with a report with content and behavioral information about the sample.
Furthermore, VirusTotal was used to scan each sample to ensure that the sample was labeled correctly as benign or malicious, and then determine the malware family to which it belonged. The majority of the samples were Trojans, but worms, viruses, backdoors and adware were also encountered.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 12, December 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1012028 |



Data Collection

The selected classifiers were evaluated by public and private data bases. MalGenome is the public dataset of 1260 malware packages collected from2010to2011.It contains some of the most malicious mobile malware in the world. Surprisingly, there was a collection of 14 out of 49 families from the Google Play official Android market. A collection of the most recent mobile malwares is the private dataset. As mobile malware changes continue, which is the hacker's strategy to circumvent the current detection methods, as well as the rapids in the number of mobile malwares, our security team collected 30 malware samples and was monitored by a malware analyst. MalGenome Project (2013) includes 1,260 malware information samples in 49 households, 1,000 of which were selection for their network model from 49 malware households. It is because 93 per cent of all the samples are bots that 1000 samples are selected. Online dynamic analysis platforms, namely Anubis ISEC lab Anubis (2013) and Sand Droid (2013), were used for the samples analysis. In a regulated setting, dynamic internet analysis systems operate an algorithm and record networking traffic, which is used in studies. In the MalGenome sample set of 1,260 samples, 1,007 traffic samples were produced by Anubis isec Lab, and 164.Thesedynamic analytic platforms capture network traffic in a controlled setting. Given that malware samples do not require traffic installation on a physical device, this approach reduces the time required to generate traffic. 93 percent of the 1,260 MalGenome files are bots, which represent 1171 files. During our test, we verified for legitimate servers and deleted 171 files from unsuccessful servers.

Feature Selection and Extraction

During the data collection phase, wireshark and Java filtered benign and malware network traffic to remove unwanted packages from the collected packets. The routine Domain name scheme streams (DNS) were filtered out of the network traffic data; only TCP streams have then been used as a dataset since TCP streams transmit conversations between mobile malware and hackers. After that, all associated information was extracted using tshark (2013), a terminal version of wireshark that used to extract 11 functionality from the TCP packets (table 2). The characteristics of the intrusion data set TUIDS (Gogoi2013), have been selected from a wide array of network characteristics. The primary task with regard to the choice of characteristics was above all to find the most important characteristics leading to the greatest true positive rate. A wide range of data set features should be filtered and refined. Furthermore, certain characteristics are correlated and hamper the method of intrusion detection. In addition, certain functions may include redundant data from other characteristics. Redundant characteristics improve time and decrease IDS exactness. A technique named Weka computer training instrument called ClassifierSubsetEval was used to select the variable. After the application of the selection algorithm depending on the outcomes, we chosen six out of 11 functions. It should be noted that no characteristic can differ between harmless applications and evil applications, but a set of characteristics. A combination of all the features selected helped find anomalies, rather than just one feature. In China, particular IP addresses are linked to malicious activities, for example. Moreover, the length of the frame is larger than normal frames due to leaked data. This leads to an anomaly detection with the mixture of both characteristics. Classifiers examine the applications from a group of functions and create patterns of behavior to detect malware. Every model of conduct, which deviates from a smooth and ordinary model is considered malignant. The features chosen for our data set are specifically derived from four categories of features (Lee and Stolfo 2000).

The classifications are fundamental characteristics, depending on material, moment and relation. The most common network traffic function used by several malware tracking scientists is the fundamental function category. This section provides basic network traffic information such as the IP address of source, IP address of destination, host number, frame number of destination port, and frame numbers. The second category of content-based features is a common link record pattern. The time-base function in the third category records a connection with the same host as the current connection in the past 10s. The 4th cluster or link-based characteristics count the amount of data packets continuously

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 12, December 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1012028 |

from origin to target and vice versa. In this research, 11 chosen characteristics have been included in our malware tracking data set as stated above. The obtained characteristics were saved as a series of CSV files. Each record consists of 11 network features summarized data. The intrusion-detection experts have labelled the data set as' normal' or' infected' with a malwarebased dataset categorized as' infected' and a benign dataset as' normal,' finally using the malware-based data set with the normal dataset. By using the function selection method, we can select k from the extracted features for android application package files: Information Gain in equ 1. Gain (S,A) = Entrophy (S) - ∑ │SV│ Entrophy(SV) V€Values(A) │S│ (1) This method depends on the entropy of the characteristics and chooses the highest profit value as the best function. A feature A is gained from a set of S instances.

Machine Learning Approach for Malware Detection

The characteristics are colleited in the signature database and split into training data and test information and are used for the detection of Android malware apps through conventional machine learning methods. In the final phase, the phase of the classifying of machines, the result of the classifiers is generated. The finest learning classification machine for malware detection is determined by these phases. This section presents the concepts and descriptions used in the present experiment of our selected classifier.

IV. CONCLUSION

This study has presented an evaluation using machinelearning classifiers to effectively detect mobile malware by choosing the appropriate networking features for classifier inspections, as well as to find the ideal classifier based on TPR values. The findings and achievement of the classifiers were overwhelming. We evaluate several machine learning classification systems in this research, to improve the malware detection result of a wide range of samples and to obtain the best classification capable of detecting mobile malware. Random forest, multi-layer perceptron and proposed SVM are the classifiers selected. Experimental results indicate 96.89 percent accuracy of the detection rate with current classifiers for the Genome Malware dataset. It also proves that the latest malware can be identified by machine classifiers. The larger the dataset, the longer the processing time is needed to build the detection model and increase the accuracy, retraction, and f-measurement. In practice, we are proposing the development of real-time mobile malware identification through machine-learning classifiers in the cloud. Nevertheless, this strategy has demonstrated that machine learning is effective and efficient in true malware operations.

Future Enhancements

Feature selection was presented in this research using Best First search algorithm. By performing feature selection or feature reduction, the features were reduced drastically. Hence, the time taken to train and build the model becomes shorter at the cost of the performance decreases slightly. In some cases, the performance can also increase slightly. The performance comparison of 5 different classifiers was also presented. The overall best performance was achieved by J48 using the term frequency-weight without feature selection data set, with a recall (true positive rate) of 95.9%, a false positive rate of 2.4%, a precision (positive predictive value) of 97.3%, and an accuracy of 96.8%. The analysis of the tests and experimental results concluded that this proof-of-concept is quite effective and efficient in detecting malware.

REFERENCES

1 Amos, B., Turner, H., & White, J. (2013, July).Applying machine learning classifiers to dynamic android malware detection at scale.In 2013 9th international wireless communications and mobile computing conference (IWCMC) (pp. 1666-1671).IEEE.
2 Android (2013) Android 4.2, Jelly Bean. http://www.android.com/about/jelly-bean/.
3 Anuar, N. B., Sallehudin, H., Gani, A., &Zakaria, O. (2008).Identifying false alarm for network intrusion detection system using hybrid data mining and decision tree.Malaysian journal of computer science, 21(2), 101- 115.
4 Anubis (2013) Anubis: analyzing unknown binaries. http://anubis.iseclab.org/.
5 Arp, D., Spreitzenbarth, M., Hubner, M., Gascon, H., Rieck, K., & Siemens, C. E. R. T. (2014, February).Drebin: Effective and explainable detection of android malware in your pocket. In Ndss (Vol. 14, pp. 23-26).
6 Arstechnica (2013) More Bad News for android: new malicious apps found in google play. http://arstechnica.com/security/2013/ 04/more-bad newsfor-android-new-malicious-apps-found-in-go ogle-play/.
7 Bradley, A. P. (1997).The use of the area under the ROC curve in the evaluation of machine learning algorithms.Pattern recognition, 30(7), 1145-1159.
8 Breiman, L. (2001).Random forests.Machine learning, 45(1), 5-32.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 12, December 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1012028 |

9 Burguera, I., Zurutuza, U., &Nadjm-Tehrani, S. (2011, October).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). ACM

10 Curiac, D. I., &Volosencu, C. (2012). Ensemble based sensing anomaly detection in wireless sensor networks. Expert Systems with Applications, 39(10), 9087-9096.

11 Dini, G., Martinelli, F., Saracino, A., &Sgandurra, D. (2012, October). 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).Springer, Berlin, Heidelberg.

12 Egele, M., Scholte, T., Kirda, E., &Kruegel, C. (2012).A survey on automated dynamic malware-analysis techniques and tools. ACM computing surveys (CSUR), 44(2), 6.

13 Eskandari, M., &Hashemi, S. (2012).A graph mining approach for detecting unknown malwares. Journal of Visual Languages & Computing, 23(3), 154-162.

14 Felt, A. P., Finifter, M., Chin, E., Hanna, S., & Wagner, D. (2011, October). A survey of mobile malware in the wild.In Proceedings of the 1st ACM workshop on Security and privacy in smartphones and mobile devices (pp. 3-14).ACM.

15 F-Secure (2013) Android accounted for 79% of all mobile malware in 2012, 96% in Q4 alone. http://techcrunch.com/2013/03/07/f-secure-androidaccounted-for-79-of-all-mobile-malware-in-201296-inq4-alone/. Accessed 1st June 2013

16 Gogoi, P., Bhattacharyya, D. K., Borah, B., &Kalita, J. K. (2013). MLH-IDS: a multi-level hybrid intrusion detection method. The Computer Journal, 57(4), 602- 623.

17 Huang, C. Y., Tsai, Y. T., & Hsu, C. H. (2013).Performance evaluation on permission-based detection for android malware. In Advances in Intelligent Systems and Applications-Volume 2 (pp. 111-120). Springer, Berlin, Heidelberg.

18 Ibrahim, L., Salah, M., Rahman, A. A. E., Zeidan, A., &Ragb, M. (2013).Crucial role of CD4+ CD 25+ FOXP3+ T regulatory cell, interferon-γ and interleukin16 in malignant and tuberculous pleural effusions.Immunological investigations, 42(2), 122-136.

19 Liang, S., Keep, A. W., Might, M., Lyde, S., Gilray, T., Aldous, P., & Van Horn, D. (2013, November). Sound and precise malware analysis for android via pushdown reachability and entry-point saturation. In Proceedings of the Third ACM workshop on Security and privacy in smartphones & mobile devices (pp. 21-32).ACM.

20 Wyatt, T. (2010). Security alert: Geinimi, sophisticated new android trojan found in wild. Online] December,2010.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.165

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details