ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

Machine Learning Techniques for Android Malware Detection and Categorization

Neha Sharma [1], Pooja Yadav [2]

P.G. Student, Department of Computer Science & Engineering, SRCEM, Palwal, Haryana, India[1]

Assistant Professor, Department of Computer Science & Engineering, SRCEM, Palwal, Haryana, India [2]

ABSTRACT: To understand the locus-standi of malware we hereby elaborating the channels which is responsible to distribute or dispense these harmful segments of program using underneath resources. Subsequently, by and large, any information transmission and correspondence channel can fill in as an assault and malware conveyance vector. Channels which are not implied for transmitting programming may at present be taken control of by regular methods for expecting command over any application's or administration's control stream because of mistaken programming. These vectors incorporate USB, Bluetooth, and NFS (Network File System) and their associations, scanner tags, QR codes, unbound remote associations which can be abused to infuse information, mistaken GSM/UMTS/LTE radio bundle dealing with, and some more. Normal correspondence channels which are intended to convey programming are the official Google Play Store and outsider application markets. In the accompanying, we will concentrate on the most imperative contamination channels for regular malware which can be found in the world today. For additional inside and out data on Android engendering channels, tenacious contamination, and malware approach all in all, allude to our best of techniques using machine learning specially using Support Vector Machine.

KEYWORDS: Machine Learning, Malware Detection, Android, Support Vector Machine, Binary Vector Generation Distributed Denial of Services, Spyware, Trojan.

I. INTRODUCTION

Malware identification is an essential factor in the security of the android frameworks. Nonetheless, as of now used mark based strategies can't give precise identification of zero-day assaults and polymorphic infections, malwares and DDOS (Distributed Denial of Services). That is the reason the requirement for machine learning-based identification emerges. The motivation behind this work was to decide the best component extraction, include portrayal, and characterization techniques that outcome in the best exactness when utilized on the highest point of Dalvik Sandbox (Android Framework). In particular, K-Nearest-Neighbors, Decision Trees, Support Vector Machines, Naive Bayes and Random Forest classifiers were assessed and studied to propose the new amalgamated technique for more accurate and effective results. This study presents suggested strategies for machine learning based malware grouping and location and in addition the rules for its execution. In addition, the investigation will be performed based on strategy for further valuable research in the field of malware examination for android framework with machine learning and its classifications.

The Machine Learning Techniques for Android Malware Detection and Categorization With the quick improvement and immense growth of the Internet, malware wound up one of the major digital dangers these days. Any product performing vindictive activities, including data taking, secret activities, and so on can be alluded to as malware. Kaspersky Labs (2017) characterize malware as a kind of algorithm or program intended to taint a real clients phone or computer and cause hurt on it in different ways. While the decent variety of malware is expanding, hostile to infection scanners which can not satisfy the necessities of insurance, bringing about a large number of hosts and smart phones being assaulted. As indicated by Kaspersky Labs (2016), 6563145 distinct hosts and Smartphone's were assaulted and 4000 one of a kind malware objects were recognized in 2015. Thusly, Juniper Research (2016) predicts the expense of information breaks to increment to $2.1 trillion all inclusive by this year i.e. 2019. Subsequently, malware attack

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

security to mobile frameworks is a standout amongst the most imperative cyber security undertakings for single client and groups too, since even a solitary assault can result in traded off information and adequate misfortunes. Mammoth venerable data loss and continuous assaults direct the requirement for exact and auspicious recognition strategies. Current static and dynamic techniques don't give productive discovery, particularly when administration zero-day attacks and assaults. Thus, machine learning-based strategies can be utilized. This scheme or scenarios talks about the primary concerns and worries of machine learning-based malware recognition, and additionally searches for the best element portrayal and grouping techniques.

The objective of this study is to build up the confirmation of idea and strategy to produce the effective amalgamated malware detector for mobile phones specially using android framework using machine learning and extract the same with respect to android sandbox. This sandbox will be used for the extraction of the conduct of the malware tests, which will be utilized as a contribution to the machine learning calculations. The objective is to decide the best element portrayal strategy and how to highlights the category of malware and to establish the most precise calculation that can recognize the malware families with the least blunder rate. The precision will be estimated both for the instance of identification, whereas the record is malevolent and for the instance in respect to the malware family. The exactness or accuracy to be proposed with come with appropriate, effective outcomes which will likewise be evaluated in connection to current scoring actualized in existing android sandbox whereas the choice of proposed technique in future will performs better solution from previous developed solutions.

Different Types Of Malware

Adware or Admobs: The main motivation behind this malware type is showing promotions on the Android Phones or Smart Phones. Regularly adware can be viewed as a subclass of spyware and it will far-fetched lead to emotional outcomes.

Spyware: As it suggests from the name, the malware that performs reconnaissance can be alluded to spyware. Regular activities of spyware incorporate following pursuit you confidential information and venerable information to the outsiders (sniffers) along-with application program interface references.

Trojan or DDOS: The fundamental inspiration driving this malware type is appearing on the phones and machines is to embark the denial of services and hardware abstract layers. Consistently Trojan malware can be seen as a subclass of spyware and it will implausible This malware class is utilized to characterize the malware types that mean to show up as authentic programming. Along these lines, the general spreading vector used in this class is social designing, i.e. making individuals imagine that they are downloading the real programming (Moffie, et al. 2006).

Rootkits : Its usefulness empowers the aggressor to get to the information with higher authorizations than is permitted by the respected operating system or framework. For instance, it very well may be utilized to give an unapproved client managerial access. Rootkits dependably shroud its reality and frequently are unnoticeable on the framework, making the discovery and along these lines expulsion unimaginably hard. (Chuvakin 2003).

Ransomware: This kind of malware means to encode every one of the information on the machine and request that an injured individual exchange some cash to get the decoding key. More often than not, a machine contaminated by ransomware is "solidified" as the client can't open any document, and the work area picture is utilized to give data on aggressor's requests. (Savage, Coogan and Lau 2015).

Malware Detection Techniques

All malware recognition procedures can be isolated into mark based and demeanor based techniques. Prior to going into these techniques, it is fundamental to comprehend the nuts and bolts of two malware examination approaches: static and dynamic malware investigation. As it suggests from the name, static investigation is performed statically, i.e. without execution of the record. Conversely, dynamic examination is directed on the record while it is being executed for instance in the virtual machine i.e. dalvik virtual machine.

Fingerprinting: this incorporates cryptographic hash calculation, finding the ecological ancient rarities, for example, hardcoded username, filename, library strings.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

AV scanning: in the event that the examined record is a notable malware, doubtlessly all enemy of infection scanners will have the capacity to identify it. In spite of the fact that it may appear to be immaterial, along these lines of location is regularly utilized by AV sellers or sandboxes to "affirm" their outcomes.

String Extraction: this alludes to the examination of the product yield (e.g. status or mistake messages) and deriving data about the malware activity.

Disassembly: this refers to reversing the machine code to assembly language and inferring the software logic and intentions. This is the most common and reliable method of static analysis.

File Format Inspection: recorded metadata can give valuable data. For instance, file system or category of file or fire pertaining records (POSIX) can give much data on order time, imported and sent out capacities, and so on.

II. RELATED WORK

Justin Sahs and Latifur Khan [1] depicts that, the recent emergence of mobile platforms capable of executing increasingly complex software and the rising ubiquity of using mobile platforms in sensitive applications such as banking, there is a rising danger associated with malware targeted at mobile devices. The problem of detecting such malware presents unique challenges due to the limited resources avalible and limited privileges granted to the user, but also presents unique opportunity in the required metadata attached to each application. In this article, we present a machine learningbased system for the detection of malware on Android devices. Our system extracts a number of features and trains a OneClass Support Vector Machine in an offline (off-device) manner, in order to leverage the higher computing power of a server or cluster of servers.

Milosevic, N., Dehghantanha, A. and Choo, K.-K.R [2] depicts that, malware have been used as a means for conducting cyber attacks for decades. Wide adoption of smartphones, which store lots of private and confidential information, made them an important target for malware developers. Android as the dominant mobile operating system has always been an interesting platform for malware developers and lots of Android malware species are infecting vulnerable users every day which make manual malware investigation an impossible mission. Leveraging machine learning techniques for malware forensics would assist cyber forensic investigators in their fight against malicious programs. In this paper, we present two machine learning aided approaches for static analysis of the mobile applications: one based on permissions , while the other based on source code analysis that utilizes a bag of words representation model. Our source code based classification achieved F-score of 95.1%, while the approach that used permission names only performed with F-measure of 89%. Our approach provides a method for automated static code analysis and malware detection with high accuracy and reduces smartphone malware analysis time.

Alatwi, Huda Ali [3] depicts that, android malware growth has been increasing dramatically along with increasing the diversity and complicity of their developing techniques. Machine learning techniques are the current methods to model patterns of static features and dynamic behaviors of Android malware. Whereas the accuracy rates of the machine learning classifiers increase with increasing the quality of the features, we relate between the apps' features and the features that are needed to deliver its category's functionality. Differently, our classification approach defines legitimate static features for benign apps under a specific category as opposite to identifying malicious patterns. We utilize the features of the top rated apps in a specific category to train a malware detection classifier for that given category. Android apps stores organize apps into different categories, for instance, 26 categories on Google Play Store. Each category has its distinct functionalities which means the apps under a specific category are similar in their static and dynamic features. In general, benign apps under a certain category tend to share a common set of features. On the contrary, malicious apps tend to request abnormal features, less or more than what is common for the category that they belong to. This study proposes categorybased machine learning classifiers to enhance the performance of classification models at detecting malicious apps under a certain category. The intensive machine learning experiments proved that category-based classifiers report a remarkable higher average performance compared to non-category based.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

III. PROPOSED ALGORITHM

In this we present the malware detection formulation for SVM classification. This is a simple representation only.

SV classification:

$$\min_{f,\xi_i} \|f\|_K^2 + C\sum_{i=1}^{l}\xi_i \; y_i f(\mathbf{x}_i) \geq 1 - \xi_i, \text{ for all } i \quad \xi_i \geq 0$$

SVM classification, Dual formulation:

$$\min_{\alpha_i} \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j K(\mathbf{x}_i,\mathbf{x}_j)$$

$$0 \leq \alpha_i \leq C, \text{ for all } i; \qquad \sum_{i=1}^{l}\alpha_i y_i = 0$$

Variables ξ_i are called slack variables and they measure the error made at point $(\mathbf{x}_i,y_i)$. Training SVM becomes quite challenging when the number of training points is large. A number of methods for fast SVM training have been proposed to detect the malware.

IV. PSEUDO CODE

Malware Detection
input: – vector of L attributes to verify for normality; ; threshold
output: 0 - normal event; 1 - anomaly
1. finalProb = 1; maxProb = 0.999
2. FOR EACH feature WHERE i=0,…,L
3. .setClassAttribute( )
4. predictedVal = .classify( )
5. realVal = .getValue( )
6. IF (is NOMINAL)
7. IF (predictedVal== realVal) distance=maxProb
8. ELSE distance = 0
9. ELSE
10.[u]=**prbs_a**(L,nc,[predictedVal])
11. diff=|realVal-predictedVal|/ .getMean
12. distance =MIN(maxProb, diff)
13. prob=1-distance
14. END FOR
15. finalProb= finalProb * prob
16. IF (finalProb> threshold) RETURN 1

V. SIMULATION RESULTS

Support Vector Machines (SVM) tend to analyses, detect and match patterns of using non-probabilistic alongwith binary vector models that assigns training data into one category or more. It also can be used efficiently for nonlinear classification problems using Kernel Trick. Kernel Trick is a class of SVM algorithms that maps the input features into a very high dimensional output space in a simple and cheaper computational way. SVM is a representation of training data as points in the space that conglomerate based on their category in form of groups that are separated by a clear distinct gap called a hyperplane. In the training phase, SVM builds up a model of patterns from the training data which is used as a space for classification phase. In the classification phase, the new input points are mapped into the trained space and categorized based on which side of the gap they fall on. In the figure below a straight line separates between

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

two classes, the new data are mapped into the space if they up the line will be categorized into otherwise into. Hyperplane is a subspace less by one dimension than its ambient space; it is 2-dimensions in 3D space, and a 1-dimension in 2D space.

Metric	personaCateg-SVM	allCateg-SVM
Accuracy	0.9855	0.8947
Precision	0.9833	1
F-Measure	0.9736	0.9318
Recall	0.9651	0.8726
FPR	0.0006	0
TPR	0.9651	0.8726
FNR	0.0348	0.1273
TNR	0.9934	1
Specificity	0.9934	1
Sensitivity	0.9651	0.8726

Table 1: Results of Personalization Apps using SVM Classifier

VI. CONCLUSION AND FUTURE WORK

SVM Based Classification: SVM is a strategy for information characterization; it can create a nonlinear hyper plane and orders information which has non-customary dissemination based on binary vector utilization. It stays away from properties with more prominent numeric reaches commanding those with littler numeric extents and it maintains a strategic distance from numerical challenges amid the count as piece esteems more often than not rely upon the internal results of highlight vectors. SVM works in two stages preparing stage and testing stage. Amid the preparation stage a SVM takes a lot of info focuses as Attribute Relation File Format, every one of which is set apart as having a place with one of two classifications, and assembles a model speaking to the information focuses in such way that the purposes of various classifications are isolated by a reasonable hole that is as wide as could be expected under the circumstances. From that point, another information point is mapped into a similar space and anticipated to have a place with a class dependent on which side of the hole it falls on. A straight SVM show isolates information having a place with various classifications by utilizing a hyper-plane so the separation from its closest information point on each side is expanded. The part trap enables the SVM calculation to end up nonlinear to isolate focuses by a hyper-plane in a changed component space. The SVM is designed and prepared to cross through two kinds of highlights. At first SVM features those records whose framework calls are having digressing conduct than typical conduct of amiable documents and other one is SVM pinpoints those records having codes that are having positive effect on the grouping of amiable and noxious programming. At long last amid testing stage SVM approves the dataset separating the documents into sets of kind and malignant records using dynamic sort mechanism, the below diagram depicts the workflow diagram of proposed scheme

REFERENCES

1. Justin Sahs and Latifur Khan, A Machine Learning Approach to Android Malware Detection, 2012 European Intelligence and Security Informatics Conference
2. Milosevic, N., Dehghantanha, A. and Choo, K.-K.R. (2017) Machine learning aided Android malware classification. Computers & Electrical Engineering, 61. pp. 266-274. ISSN 0045-7906
3. Alatwi, Huda Ali, "Android Malware Detection Using Category-Based Machine Learning Classifiers" (2016). Thesis. Rochester Institute of Technology.
4. Ivan Firdausi ; Charles lim ; Alva Erwin ; Anto Satriyo Nugroho Analysis of Machine learning Techniques Used in Behavior-Based Malware Detection 2010 Second International Conference on Advances in Computing, Control, and Telecommunication Technologies.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

5. Wen-Chieh Wu, Shih-Hao Hung : DroidDolphin: a dynamic Android malware detection framework using big data and machine learning ACM New York, NY, USA ©2014
6. Ying-Chih Shen1, Roger Chien1, and Shih-Hao Hung1, Toward Efficient Dynamic Analysis and Testing for Android Malware 1Academia Sinica 2016, Taiwan, National Taiwan University, Taiwan
7. D. Arp, M. Spreitzenbarth, M. Hübner, H. Gascon, K. Rieck, and C. Siemens. Drebin: Effective and explainable detection of android malware in your pocket. 2014.
8. S.-J. Chang. Ape: A smart automatic testing environment for android malware. 2013.
9. S. Y. Yerima, S. Sezer and G. McWilliams. "Analysis of Bayesian Classifcation Aprroaches for Android Malware Detection," IET Information Security, Vol 8, Issue 1, January 2014.