Accuracy Enhancement of Malware Identification Using Machine Learning

[1]Jayateerth V Vadavi , [2]Priya S Patil, [1]Dr Umakant Kulkarni

[1]Associate Professor, [2]PG Student

Department of CSE, SDMCET Dharwad

Abstract: Malware detection is an important factor in the security of the computer systems. However, currently utilized signature-based methods cannot provide accurate detection of zero-day attacks and polymorphic viruses. That is why the need for machine learning-based detection arises. The purpose of this work was to determine the best feature extraction, feature representation, and classification methods that result in the best accuracy when used on the top of machine learning algorithms as follows Naive Bayes, Logistic Regression, Linear Discriminant Analysis, and Random Forest classifiers were evaluated. The dataset used for this study consist of the both malware and benign applications. This work presents recommended methods for machine learning based malware classification and detection, as well as the guidelines for its implementation. Moreover, the study performed can be useful as a base for further research in the field of malware analysis with machine learning methods.

Keywords: *Machine Learning, Python, Naive Bayes, Linear Discriminant Analysis, Logistic Regression, Random Forest Classifier.*

I INTRODUCTION

By fast advancement of the Internet, malicious ended up one of the main digital dangers these days. Any product performing pernicious activities, including data taking, surveillance, and so forth can be alluded to as malware. The malware can be characterized as "a kind of PC program intended to taint a true blue client's gadget and dispense hurt in different ways. "the variety of malware is expanding, against infection detection can't satisfy necessities in security, bringing about a great many hosts being assaulted. In this manner, malware insurance of gadgets is a standout amongst the most critical cybersecurity undertakings for single clients and organizations, since even a solitary assault can bring about traded off information and adequate misfortunes. Huge misfortunes and regular assaults direct the requirement for precise and convenient recognition strategies. Current static and dynamic strategies don't give effective recognition, particularly when managing zero-day assaults. Therefore, machine learning-based methods can be utilized for recognizing the malware. The objective is to decide the best

element representation strategy and how the highlights ought to be removed, the most exact calculation recognize the malicious patterns with the least mistake rate. The precision will be estimated both for the instance of discovery of whether the document is noxious and for the instance of grouping of the record to the malicious pattern.

II LITERATURE SURVEY

[1] Android OS is one of the broadly utilized versatile Operating Systems. The quantity of pernicious applications and adware are expanding continually comparable to the quantity of cell phones. An incredible number of business signature construct instruments are accessible with respect to the market which avoid to a degree the infiltration and conveyance of pernicious applications. Various looks into have been directed which guarantees that customary mark based discovery framework function admirably up to certain level and malware creators utilize various strategies to avoid these apparatuses.

So given this situation, there is an expanding requirement for an option, extremely intense malware discovery framework to supplement and amend the mark based framework. Late significant research concentrated on machine taking in calculations that investigate highlights from pernicious application and utilize those highlights to characterize and identify obscure vindictive applications. This examination abridges the advancement of malware recognition methods in view of machine learning calculations concentrated on the Android OS.

This investigation abridges late advancements in android malware location utilizing machine learning calculations. Location procedures and frameworks that utilizations static, dynamic and crossover approaches are examined and featured. A strategy that could prompt potential balancing the refresh assault is examined. The inaccessibility of a bigger android malware dataset remains an extraordinary issue in assessing different methodologies. With a legitimate dataset shared among analysts, a framework that takes in another malware and offer that information to all the cell phones, so they can shield themselves from future assaults, could be produced.

[2] Cell Phones have turned into an essential need of today. The term cell phone and advanced mobile phone are relatively indistinguishable now a days. Cell phone showcase is blasting with fast. Cell phones have increased such a colossal ubiquity because of extensive variety of abilities they offer. Right now

android stage is driving the cell phone showcase. Android has picked up an overnight fame and turned into the best OS among its rival OS. This greatness pulled in malware creators too. As android is an open source stage, it appears to be very simple for malware creators to satisfy their illegal expectations.

In this paper another method will be acquainted with distinguish malware. This method identifies malware in android applications through machine learning classifier by utilizing both static and dynamic investigation. This system does not depend on malware marks for static examination but rather android consent demonstrate is utilized. Under unique investigation, framework call following is performed. Utilizing both static and dynamic methods alongside machine learning gives across the board answer for malware location. The strategy utilized by us is tried on different amiable examples gathered from official android showcase and on different vindictive applications.

We have run over different malware location systems that utilizations either static identification strategies that can be effortlessly jumbled or those that utilization just unique discovery methods, which are additionally not an entire arrangement and afterward influenced this model which to join highlights of both static investigation and dynamic examination and machine learning calculation. Every one of these procedures are consolidated so to get greatest precision in identifying vindictive examples.

III OBJECTIVES

1.To explore on the most proficient method to actualize machine learning for figuring out how to detect malware and benign application using classifier.

2.To develop up a malware detection software that actualize machine figuring out how to identify malware and benign application.

3. To approve that malware detection that execute machine learning will have the capacity to accomplish a high precision rate.

IV ALGORITHM DISCUSSION

1. Naive Bayes Algorithm

The Naive Bayes calculation is an instinctive technique that uses the probabilities of each property having a place with each class to make an expectation. It is the regulated learning approach you would create in the event that you needed to demonstrate a prescient displaying issue probabilistically. Innocent Bayes improves the figuring of probabilities by accepting that the likelihood of each credit having a place with a given class esteem is free of every single other characteristic. This is a solid suspicion however brings about a quick and compelling strategy.

The likelihood of a class esteem given an estimation of a property is known as the contingent likelihood. By duplicating the restrictive probabilities together for each characteristic for a given class esteem, we have a likelihood of an information case having a place with that class. To influence a forecast, we can ascertain probabilities of the case having a place with each class and select the class esteem with the most amazing likelihood.

2.Logistic Regression

Logistic Regression is a prescient demonstrating calculation that is utilized when the Y variable is in parallel. That is, it can take just two qualities like 1 or 0. The objective is to decide a scientific condition that can be utilized to foresee the likelihood of occasion 1. Once the condition is built up, it can be utilized to anticipate the Y when just the X's are known.

At the point when the reaction variable has just 2 conceivable qualities, it is attractive to have a model that predicts the esteem either as 0 or 1 or as a likelihood score that reaches in the vicinity of 0 and 1. Linear Regression does not have this ability. Since, if you utilize straight relapse to show a double reaction variable, the subsequent model may not limit the anticipated Y esteems inside 0 and 1.

3. Linear Discriminant Analysis

Fisher's linear discriminant is a hypothesis of Linear discriminant analysis(LDA) or discriminant function analysis, a technique used in statistics, pattern recognition and machine to find the linear combination of features that describes two or more classes of object. The resulting grouping can be used as a linear classifier.

LDA is firmly identified with analysis of difference (ANOVA) and regression check, which also additionally try to express one dependent variable as a linear combination of different highlights or measurements. However, ANOVA utilizes all categorical independent factors and a constant dependent variable, while discriminant investigation has persistent independent factors and a categorical independent variable.

Logistic regression and profit regression are more like LDA than ANOVA as they also explain a categorical variable by the values of continuous independent variables. These other methods are preferable in applications where it is not reasonable to assume that the independent variables are normally distributed, which is a fundamental assumption of the LDA method.

Both PCA (Principal Component Analysis) and LDA (Linear Discriminant Analysis) are using linear transformation technique for dimensionality reduction. But PCA make use of unsupervised algorithms that is it ignores the class labels but just find out the direction that maximize the variance in a dataset. Where in LDA works on supervised algorithms that evaluates the direction that represents the axes that maximize the separation between multiple classes. LDA is also used for multi class classification task.

LDA works when the measurements are made on independent variables for each observation are continuous quantities. When dealing with categorical independent variables, the equivalent technique is discriminant correspondence analysis. Discriminant analysis is used when groups are known a prior.

4. Random Forest Classifier

Random forest is a standout amongst the most prominent machine learning calculations. It requires no information planning and displaying yet for the most part brings about exact results. Random forests are the accumulations of choice trees, creating a superior forecast exactness. That is the reason it is known as a 'woodland' it is essentially an arrangement of choice trees. The fundamental thought is to develop numerous choice trees in light of the free subsets of the dataset. At every hub, n factors out of the list of capabilities are chosen arbitrarily and the best split on these factors is found.

1) Multiple trees are constructed generally on the two third of the training data 62.3% data is picked arbitrarily.

2) A few indicator factors are arbitrarily chosen out of all the indicator factors. At that point, the best split on these chose factors is utilized to part the hub. As a matter of course, the measure of the chose factors is the square foundation of the aggregate number of all indicators for arrangement, and it is consistent for all trees.

3)Utilizing whatever is left of the information, the misclassification rate is computed. The aggregate mistake rate is calculated as the overall out of bag error rate.

4)Each prepared tree gives its own particular grouping result, giving its own vote the class that got the most votes is picked as the outcome.

Random forest make use of the advantages of decision tree algorithms which are applicable for both regression and classification problems are easy to compute and generate accurate results. In the decision tree by examining the resulting data can get the valuable information about the features which are important and how they affect the result. This is not possible in random forest because random forest classifier is more stable than the decision tree because if we modify the data little the decision tree will change and reduce the accuracy hence combination of many decision tree and random forest yields the stable result.

V PROBLEM DEFINITION

1.Identifying the total number of malware and benign application in the given dataset based on the knowledge acquired by the machine learning classifiers.

2.Measuring the accuracy of malware detection in the given dataset using machine learning classifiers.

VI IMPLEMENTATION

The proposed system aims to get better performance compare to previous approach using various machine learning algorithms, the proposed model is specified with two phases training and testing data.

The malware data set is used as input for the machine and it is arranged in structured form in the pre-processing stage than the features of input data are extracted and these extracted features are used to train the machine and later with the help of classifiers the accuracy is obtained for finding out the total number of malware and begin application are present in the given dataset.

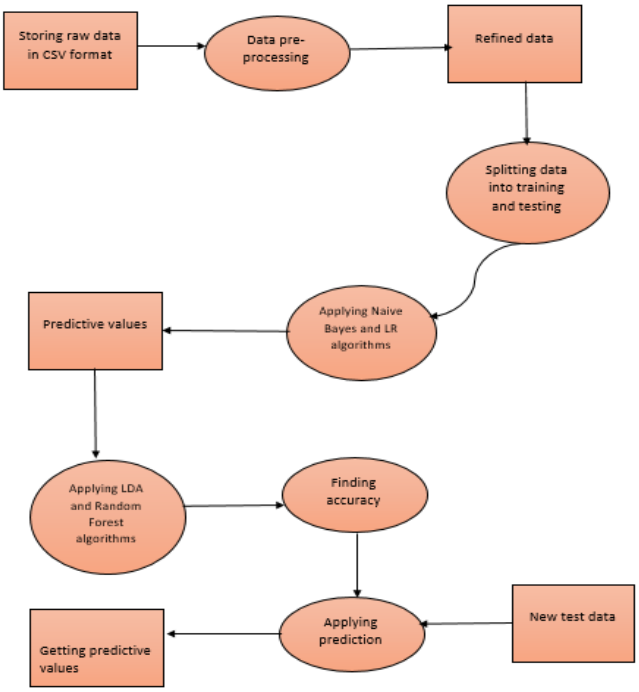

Figure 1: Flow Diagram of Proposed Model

1)Supervised Learning

In supervised learning it has presence of supervisor i.e., we can train or teach the machine using data which is labeled that means some data is already tagged with a correct answer. Once the machine is trained it is provided with the new set of data so that using some supervisor learning algorithms analyses the training data and provides the correct result from labelled data.

Consider the example a basket consists of set of vegetables and fruits we train the machine in such a way that if the shape of object is round and red in color it is labelled as apple and similarly we label rest of the items present in basket once we fed new data to the machine based on the labelled data we have trained to the machine it classifies the item we passed to it.

Classification: Where the outcome result is category values like red, black, malware, benign so on.

Regression: Where the outcome result is real values like dollars, weights so on.

2)Unsupervised Learning

Unsupervised learning it works opposite to supervised learning where does not consists of any supervisor to train the machine hence each of the machine is to group unsorted information according to the similarities, patterns without any prior training of data

Consider the example suppose it is given an image having dogs and cat which is not seen nor machine has idea about the features of animals hence it has to category the dogs and cat by itself according to their similarities and pattern and classify them.

Clustering: where the machine need to discover the inherent grouping in the given data set. Example grouping customers by their purchasing behavior.

Algorithm: Proposed system

Input: Malware dataset

Output: Categories the total number of malware and benign application

Step1: Start

Step2: Train the malware dataset

Step3: Test one single dataset

Step4: Pre-processing the dataset by cleaning and identifying

Step5: Extract the features by using static and dynamic method

Step6: Classify the total number of malware and benign application using classifiers

Step7: Output result: Total number of malware and benign application are found in the given dataset

Step8: Stop

3)Malware Dataset

Raw malware dataset is extracted directly from the internet need to be converted into standard CSV format. The malware dataset contains both benign and malware application along with the various feature values that can be useful to predict the total number of malware and benign application present in the given malware dataset using various machine learning classifiers.

VII ACCURACY ENHANCEMENT

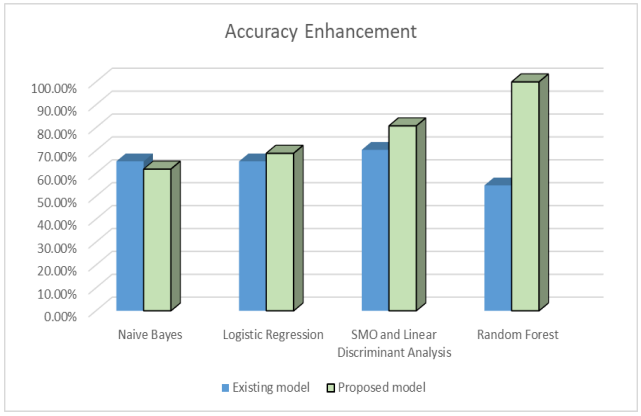

Figure 2: Bar Graph of Accuracy Between Existing and Proposed Model

VIII CONCLUSION AND FUTURE WORK

Using the various machine learning classifiers, the accuracy obtained by Naive Bayes Algorithms was 65.29% over the malware dataset using static and dynamic analysis. Similarly, with respect to Logistic Regression was 65.29%, SMO was 70.31% and Random Tree was 54.79% in the existing model. The accuracy rate was 61.91% in the proposed model using Naive Bayes similarly it was increased with Logistic Regression by 68.81%, Linear Discriminant Analysis by 80.79%, Random Forest classifier by 100% over the malware dataset. Malicious detection technique need to be improved by using the various machine learning algorithms as per current situation there are many malicious applications hence need to be identified and secure the user from any harmful causes towards the user sensitive data. As future work for this model is to identify the different patterns of malware by taking different samples and larger number of dataset.

REFERENCES

[1] Hahnsang Kim, Joshua Smith, Kang G. Shin "Detecting energy greedy anomalies and mobile malware variants" in MobiSys'08

[2] Aubrey-Derrick Schmidt, Rainer Bye, Hans-Gunther Schmidt, Jan Clausen, Osman Kirazy, Kamer Ali Y¨uksely, Seyit Ahmet Camtepe, and Sahin Albayrak "Static analysis of executables for collaborative malware detection on android" in Communications,2009. ICC'09. IEEE International Conference

[3] Burguera, I., Zurutuza, U., & Nadjm-Tehrani, S. (2011). "Crowdroid: Behavior-based malware detection system for Android" in 2011 ACM CCS Workshops on Security and Privacy in Smartphones and Mobile Devices (SPSM'11), 17-21 October 2011, Chicago, Illinois, USA.

[4] Grace, M., Zhou, Y., Zhang, Q., Zou, S., & Jiang, X. (2012). "RiskRanker: scalable Aliyev, Vusal. 2010. Using honeypots to study skill level of attackers based on the exploited vulnerabilities in the network. Chalmers University of Technology. and accurate zero-day Android malware detection." in The 10th International Conference on Mobile Systems,

Applications, and Services (MobiSys'12), Low Wood Bay, Lake District, United Kingdom

[5] Portokalidis, G Homburg P, Anagnostakis K., and Bos, H.: "Paranoid Android: Versatile protection for smartphones" in ACSAC'10, Dec. 2010.

[6] Su, X., Chuah, M., Tan, G."Smartphone dual defense protection framework: Detecting malicious applications in android markets" in: Mobile Ad-hoc and Sensor Networks (MSN), 2012 Eighth International Conference on, pp. 153-160 (2012).

[7] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian [8] H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

[9] Aquino, Maharlito. 2014. Fake BACS Remittance Emails Delivers Dridex Malware. WWW document. Available at: https://blog.cyren.com/articles/fake- bacs-remittance-emails-delivers-dridex-malware.html. [Accessed 15 February 2017].

[10] Jing, Ranzhe, and Yong Zhang. 2010. A View of Support Vector Machines Algorithm on Classification Problems. International Conference on Multimedia Communications.