Enhanced App Model for the Prediction and Prevention of Malware in Android System

Smikar S. Parab[1] Kalyani B. Pawar[2] Nikita N. Patil[3]
[1,2,3]Department of Computer Engineering
[1,2,3]Pillai HOC College of Engineering and Technology, Navi Mumbai, India

Abstract— Malware is termed as malicious software, its aim is to cause harm to the mobile devices, server and also computer network by either collapsing the system or leakage of some confidential data. It does the damage when it is some way implanted into the mobile. The effects of the malware is that your mobile becomes unstable, it will continuously get reboot or it may also get crash. To overcome this problem an enhanced app model is designed. This app is efficient enough to detect all sort of malware and prevent the cell phones from getting security breached. To remove this malware, an antivirus is required which has definition of all the malware types but in this app we have used a Droid Fusion framework which is used to classify the malware types and their versions. It also helps to improve the accuracy to predict the malware in the devices and also predicts at what limit it will do harm to the mobile. Data flow API features are extracted to check whether the sensible data leaves out the system. In the previous system the implementation is done using the k-nearest algorithm but in this system we have used an enhanced Quine mccluskey algorithm technique to track and differentiate the malicious data from the normal data. Because k map based technique breakdown in six variables, Quine mccluskey proposed an algorithmic based technique for simplifying Boolean logic functions and basically it has two advantages over the kmap method. Firstly it is systematic for producing a minimal function that is less dependent on visual pattern it is viable scheme for handling large no .of variables.

Keywords: Quine Mccluskey Algorithm Technique, Polynomial Regression

I. INTRODUCTION

In recent years, Android has become the leading mobile operating system with a substantially higher percentage of the global market share. Over 1 billion Android devices have been sold with an estimated 65 billion app downloads from Google Play alone. The growth in popularity of Android and the proliferation of third party app markets has also made it a popular target for malware. Last year, Mc A fee reported that there were more than 12 million Android malware samples with nearly 2.5 million new samples discovered every year. Android malware can be embedded in a variety of applications such as banking apps, gaming apps, lifestyle apps, educational apps, etc. These malware-infected apps can then compromise security and privacy by allowing unauthorized access to privacy- sensitive information, rooting devices, turning devices into remotely controlled bots, etc. Zero-day Android malware have the ability to evade traditional signature-based defences. Hence, there is an urgent need to develop more effective detection methods. Malware forms one of the core cyber-security threat landscape in distributed computing. In big data sphere, malware has been identified as one of the top security and privacy challenges to be tackled. The size and the variety of Android malware seen in the contemporary Android malware detection databases pertains to the domain of big data. Also, as smartphones are commonly used for accessing and storing big data, Android malware poses a serious challenge for big data security. Since it can impact the data machine learning based methods are increasingly being applied to Android malware detection. However, classifier fusion approaches have not been extensively explored as they have been in other domains like network intrusion detection. In this paper, we present and investigate a novel classifier fusion approach that utilizes a multilevel architecture to increase the predictive power of machine learning algorithms. The framework, called Droid Fusion, is designed to induce a classification model for Android malware detection by training a number of base classifiers at the lower level. A set of ranking-based algorithms are then utilized to derive combination schemes at the higher level, one of which is selected to build a final model. The framework is capable of leveraging not only traditional singular learning algorithms like Decision Trees or Naive Bayes, but also ensemble learning algorithms like Random Forest, Random Subspace, Boosting etc. for improved classification accuracy.

II. PURPOSE

III. CONSTRAINTS

A. Time:

Scope: This system deals with the malicious data of the system. It is basically an app model that scan any data entering in your system and check whether any malicious data is present if yes than it first predict it and then prevent it.

IV. OVERALL SYSTEM DESCRIPTION

A. Existing System

The growing popularity of Android based smartphone attracted the distribution of malicious applications developed by attackers which resulted the need for sophisticated malware detection techniques. Several techniques are proposed which use static or dynamic features extracted from android application to detect malware. In this project we develop an android app based on android studio using java code .This android application works as a feature to detect and prevent form melodious data. recent android malware detection work that employee machine learning with the static features including the following droidmat proposed applying the enhanced quine mccluskey algorithm technique. the machine learning based detection technique were based on API(application programming interface)calls.

In order to detect malicious mobile apps, several steps should be done. First, detection features such as user's operating behavior, API usage, and application network

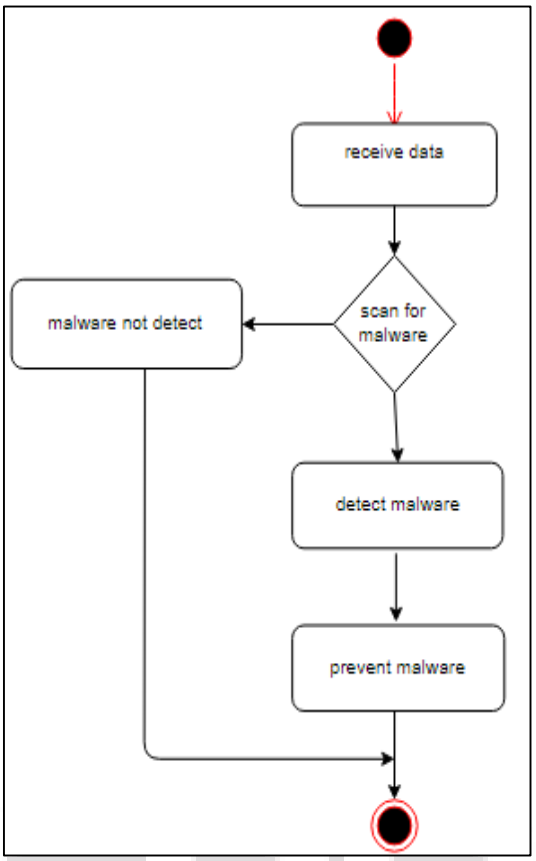
behavior should be defined and extracted. Then detection models are constructed and, finally, new apps are compared with the constructed models for mobile malware detection.
Depending on the locations where these steps are performed, current approaches can be generally categorized into two main groups: client-side and server-side detection. For client-side detection, these steps are all performed at mobile devices, while, for server-side detection, the main steps are carried out on remote servers. Even though server-side detection approaches are conducted remotely, their detection features are mainly collected and processed on mobile devices and later sent to remote servers for modeling and detection. Therefore, current approaches are required to install some kind of program on mobile devices or modify operating systems (e.g., modify Android source code) to collect detection feature information.

Obviously, this will increase energy consumption of mobile devices. Also, these methods will be difficult to be applied for mobile device protection in large organizations.it is hard to ensure that all mobile devices have installed information collection programs and it is impractical to manually audit every employee's personal device due to the privacy issue and also the large amount of mobile devices.

B. Proposed System

The system that we proposed solves most of the problems that we have with the existing system. In our system the accuracy of the prediction is mostly correct as compared to existing system. There are several factors that affect house prices. We divide these factors into three main groups, there are physical condition, concept and location. Physical conditions are properties possessed by a house that can be observed by human senses, including the size of the house, the number of bedrooms, the availability of kitchen and garage, the availability of the garden, the area of land and buildings, and the age of the house while the concept is an idea offered by developers who can attract potential buyers, for example, the concept of a minimalist home, healthy and green environment. Location is an important factor in shaping the price of a house. This is because the location determines the prevailing land price. In addition, the location also determines the ease of access to public facilities, such as schools, campus, hospitals and health centers, as well as family recreation facilities such as malls, culinary tours, or even offer a beautiful scenery. By using this all factors we optimized the result.

C. System Architecture



D. Methodology

The system consists of Arduino Due, Wi-Fi module, motor drivers for DC motor which is used for the movement of the robot, night vision camera to capture images and ultrasonic sensor to find the distance of the object. The communication between the base station and the robot happens through the Wi-Fi module. The camera mounted on the robot is used to view the things happening around. The rotational angle of the camera is controlled by a stepper motor. Camera is used to take real time video and the captured images are compared with the images stored in the database. After verifying the captured images with the database, appropriate commands are issued to the robot using the android application. Android application is created in order to control the movement of the robot and to view the live surveillance to detect and recognize the faces which are captured in the camera. In the worst case, if there is no internet connectivity then the robot will store all the recordings in the external storage device which is already mounted on it. The robot also contains a Lithium polymer battery of 4000 mAh which would make it operable up to 20 hours.

1) Simple Linear Regression:

In simple linear regression, we predict scores on one variable from the scores on a second variable. The variable we are predicting is called the criterion variable and is referred to as Y. The variable we are basing our predictions on is called the predictor variable and is referred to as X. When there is only one predictor variable, the prediction method is called simple regression. In simple linear regression, the topic of this section, the predictions of Y when plotted as a function of X form a straight line.

The formula for a regression line is

$$Y' = bX + A$$

where Y' is the predicted score, b is the slope of the line, and A is the Y intercept.

2) Multiple Linear Regression

Multiple linear regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. Every value of the independent variable x is associated with a value of the dependent variable y. The population regression line for p explanatory variables $x_1, x_2, \ldots, x_p$ is defined to be $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$. This line describes how the mean response μ_y changes with the explanatory variables. The observed values for y vary about their means μ_y and are assumed to have the same standard deviation σ. The fitted values $b_0, b_1, \ldots. b_p$ estimate the parameters $\beta_0, \beta_1, \ldots, \beta_p$ of the population regression line.

Formally, the model for multiple linear regression, given n observations, is $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots \beta_p x_{ip} + \varepsilon_i$ for $i = 1,2, \ldots n$.

3) Polynomial Regression

This function fits a polynomial regression model to powers of a single predictor by the method of linear least squares. Interpolation and calculation of areas under the curve are also given.

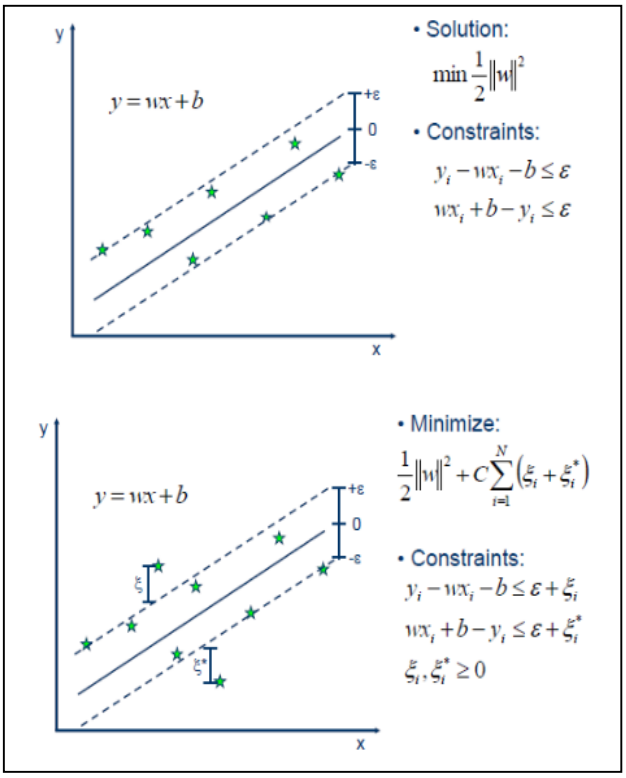
If a polynomial model is appropriate for your study then you may use this function to fit a k order/degree polynomial to your data:

$Y=\beta 0+\beta 1X+\beta 2X2+\ldots+\beta hXh+\epsilon$,

- where Y caret is the predicted outcome value for the polynomial model with regression coefficients β to X for each degree and Y intercept ϵ.

4) Support Vector Regression

Support Vector Machine can also be used as a regression method, maintaining all the main features that characterize the algorithm (maximal margin). The Support Vector Regression (SVR) uses the same principles as the SVM for classification, with only a few minor differences. First of all, because output is a real number it becomes very difficult to predict the information at hand, which has infinite possibilities. In the case of regression, a margin of tolerance (epsilon) is set in approximation to the SVM which would have already requested from the problem. But besides this fact, there is also a more complicated reason, the algorithm is more complicated therefore to be taken in consideration. However, the main idea is always the same: to minimize error, individualizing the hyper plane which maximizes the margin, keeping in mind that part of the error is tolerated.



5) Decision Tree Regression

Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node (e.g., Outlook) has two or more branches (e.g., Sunny, Overcast and Rainy), each representing values for the attribute tested. Leaf node (e.g., Hours Played) represents a decision on the numerical target. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

6) Decision Tree Algorithm

The core algorithm for building decision trees called ID3 by J. R. Quinlan which employs a top-down, greedy search through the space of possible branches with no backtracking. The ID3 algorithm can be used to construct a decision tree for regression by replacing Information Gain with Standard Deviation Reduction.

7) Standard Deviation

A decision tree is built top-down from a root node and involves partitioning the data into subsets that contain instances with similar values (homogenous). We use standard deviation to calculate the homogeneity of a numerical sample. If the numerical sample is completely homogeneous its standard deviation is zero.

V. CONCLUSION

So we conclude that in the existing system there are many solution for house's sales price prediction problem for one of the Kaggle competitions, in which they combine standard machine learning algorithms with their original ideas like residual regressor, logit transform and neural network machine. But during data analysis the results show that the

house price variation prediction results is not accurate enough. Sometimes the Standard deviation of the results is very high because of small dataset size. So, the system that we proposed solves most of the problems that we have with the existing system. Therefore, the outcome of our project will be predicting house prices with good accuracy which can help the customer as well as developer.

ACKNOWLEDGMENT

We remain immensely obliged to Prof. Shamna Sadanand for providing us with the moral and technical support and guiding us. We would also like to thank our guide for providing us with her expert opinion and valuable suggestions at every stage of project.

We would like to take this opportunity to thank Prof. Monisha Mohan, Head of Information Technology for her motivation and valuable support. This acknowledgement is incomplete without thanking teaching and non-teaching staff of department of their kind support.

We would also like to thank Dr.Chelpa Lingam, Principal of Pillai HOC college of Engineering and Technology, Rasayani for providing the infrastructure and resources required for project.

REFERENCES

[1] Tarunpreet Kaur, Dilip Kumar, "Wireless Multifunctional Robot for Military Applications", IEEE Proceedings of 2015 RAECS, 21- 22nd December 2015.
[2] Benedict Ebinesar.J, Vijay Nagaraj, "Surveillance and Target Engagement using Robots", IOSR Journal of Electronics and Communication Engineering (IOSR-JECE). e- ISSN: 2278- 2834, p- ISSN: 2278- 8735.
[3] K.AnilBablu Louis, K.M.S.R.Tarun, T.Teja and B.Santhi Kiran, "Intelligence Spy Robot with Wireless Night Vision Camera Using Android Application", International Journal for Modern Trends in Science and Technology, Vol. 03, Special Issue 02, 2017.
[4] Mr. P. Surendra Kumar, Ms. S. Geetha Priyanka, Mr. V. Venkatesh, Mr. Sk. Tausif Ahamed, "Video Surveillance Robot with Multi Mode Operation", International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181.
[5] Moon Sun SHIN, Byung Cheol KIM, Seon Min HWANG, Myeong Cheol KO, "Design and Implementation of IoT-based Intelligent Surveillance Robot", Studies in Informatics and Control, Vol. 25, No. 4, December 2016.

