YIDDISH Vol 13 Issue 04,Dec 2023
ISSN NO: 0364-4308

SIMULATION AND CLASSIFICATION OF MOBILE COMMUNICATION, BASE STATION DOWNLINK OF SIGNALS

1.DR. RAGHVENDRA,2. S. TRISHA ,3. S. PRAVALIKA,4. SHAIKH SHABANA ,5. T. SNIGDHA

1.PROFESSOR,2,3,4&5 UG SCHOLAR

DEPARTMENT OF ECE, MALLA REDDY ENGINEERING COLLEGE FOR WOMEN, HYDERABAD

ABSTRACT

Propose a versatile framework in which one can employ different machine learning algorithms to successfully distinguish between malware files and clean files, while aiming to minimise the number of false positives. In this paper we present the ideas behind our framework by working firstly with cascade one-sided perceptrons and secondly with cascade kernelized one-sided perceptrons. After having been successfully tested on medium-size datasets of malware and clean files, the ideas behind this framework were submitted to a scaling-up Process that enable us to work with very large datasets of malware and clean files.

INTRODUCTION

In is defined as software designed to infiltrate or damage a computer system without the owner's in- formed consent. Malware is actually a generic definition for all kind of computer threats. A simple classification of malware consists of file infectors and stand-alone malware. Another way of classifying malware is based on their particular action: worms, backdoors, trojans, rootkits, spyware, adware etc. Malware detection through standard, signature based methods [1] is getting more and more difficult since all current malware applications tend to have multiple polymorphic layers to avoid detection or to use side mechanisms to automatically update themselves to a newer version at short periods of time in order to avoid detection by any antivirus software. For an example of dynamical file analysis for malware detection, via emulation in a virtual environment, the interested reader can see [2]. Classical methods for the detection of metamorphic viruses are described in [3]. An overview on different machine learning methods that were proposed for malware detection is given in [4]. Here we give a few references to exemplify such methods. - In [5], boosted decision trees working on n-grams are found to produce better results than both the Naive Bayes clasifier and Support Vector Machines. - [6] uses automatic extraction of association rules on Windows API execution sequences to distinguish between malware and clean program files. Also using association rules, but on honeytokens of known parameters, is [7]. - In [8] Hidden Markov Models are used to detect whether a given program file is (or is not) a variant of a previous program file. To reach a similar goal, [9] employs Profile Hidden Markov Models,

which have been previously used with great success for sequence analysis in bioinformatics. - The capacity of neural networks to detect polymorphic malware is explored in [10]. In [11], Self-Organizing Maps are

EXISTING SYSTEM

In existing system, The malware files in the training dataset have been taken from the Virus Heaven collection. The test dataset contains malware files from the WildList collection and clean files from different operating systems (other files that the ones usedin the first database). The malware collection in the training and test datasets consists of trojans, backdoors, hacktools, rootkits, worms and other types of malware. The first and third columns in Table II represent the percentage of those malware types from the total number of files of the training and respectively test datasets. The second column in Table II represents the corresponding percentage of malware unique combinations from the total number of unique combinations of feature values for the training dataset

DISADVANTAGES

- Doesn’t Efficient for handling large volume of data.
- Theoretical Limits
- Incorrect Classification Results.
- Less Prediction Accuracy.

PROPOSED SYSTEM

The proposed model is introduced to overcome all the disadvantages that arises in the existing system. This system will increase the accuracy of the classification results by classifying the data based on the software quality prediction dataset and others using SVM , Gradient Boosting ,Navie Bayes Random forest and decision Tree algorithms.It enhances the performance of the overall classification results.

ADVANTAGES

- High performance.
- Provide accurate prediction results.
- It avoid sparsity problems.

- Reduces the information Loss and the bias of the inference due to the multiple estimates.

LITERATURE REVIEW

Title: Cloud security architecture based on user authentication and symmetric key cryptographic techniques, 2020

Author: Abdul Raoof

Technologies and Algorithm Used:

The study is implemented on the Structure for cloud security with efficient security in communication system and AES based file encryption system. This security architecture can be easily applied on PaaS, IaaS and SaaS and one time password provides extra security in the authenticating users.

Advantages:

- Performance time and accuracy

Disadvantages:

- Training model prediction on Time is High
- It is based on Low Accuracy

Title: Analysis and Countermeasures for Security and Privacy Issues in Cloud Computing, 2019

Author: Q. P. Rana, Nitin Pandey

Technologies and Algorithm Used:

The cloud computing environment is adopted by a large number of organizations so the rapid transition toward the clouds has fuelled concerns about security perspective. There are numbers of risks and challenges that have emerged due to use of cloud computing. The aim of this paper is to identify security issues in cloud computing which will be helpful to both cloud service providers and users to resolve those issues. As a result, this paper will access cloud security by recognizing security requirements and attempt to present the feasible solution that can reduce these potential threats.

Advantages:

More effective and efficient**.**

Disadvantages:

Not give accurate prediction result.

Title: Using Firefly and Genetic Metaheuristics for Anomaly Detection based on Network Flows, 2014

Author: Faisal Hussain

Technologies and Algorithm Used:

- In this work, we proposed a Traffic monitoring is a challenging task which requires efficient ways to detect every deviation from the normal behavior on computer networks. In this paper, we present two models to detect network anomaly using flow data such as bits and packets per second based on: Firefly Algorithm and Genetic Algorithm. Both results were evaluated to measure their ability to detect network anomalies, and results were then compared. We experienced good results using data collected at the backbone of a university.

Advantages:

Efficiency measure and the accuracy

Disadvantages:

Not give accurate prediction result.

Title: A Multiple-Layer Representation Learning Model for Network-Based Attack Detection, 2018

- **Author:** Suresh M

Technologies and Algorithm Used:

- The proposed solutions are this ensures fine-grained detection of various attacks. The proposed framework has been compared with the existing deep learning models using three real datasets (a new dataset NBC, a combination of UNSW-NB15 and CICIDS2017 consisting of 101 classes).

Advantages:

It performs accurate classification of health state in comparison with other methods

Disadvantages:

It is low in efficiency.

Title: Detecting Distributed Denial of Service Attacks Using Data Mining Techniques, 2018
Author: Linga

Technologies and Algorithm Used:

In this study, we DDoS (Distributed Denial of Service) attack has affected many IoT networks in recent past that has resulted in huge losses. We have proposed deep learning models and evaluated those using latest CICIDS2017 datasets for DDoS attack detection which has provided highest accuracy as 97.16% also proposed models are compared with machine learning algorithms

Advantages:

- The proposed solution can successfully detect network intrusions and DDOS communication with high precision**.**
- More Reliable.

Disadvantages:

- It is less in efficiency and not give perfect result.
- This finding is disadvantageous to the organization experiencing such attack.
- The difficulty in identifying all articles that are related to this study

IMPLEMENTATION

- Data Selection and Loading
- Data Preprocessing
- Splitting Dataset into Train and Test Data
- Classification
- Prediction
- Result Generation

DATA SELECTION AND LOADING

- Data selection is the process of determining the appropriate data type and source, as well as suitable instruments to collect data.

- Data selection precedes the actual practice of data collection and it is the process where data relevant to the analysis is decided and retrieved from the data collection.
- In this project, the Malware dataset is used for detecting Malware type prediction.

DATA PREPROCESSING

- The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.
- **Missing Data:**
  This situation arises when some data is missing in the data. It can be handled in various ways.
  - ✓ Ignore the tuples:
    This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.
  - ✓ Fill the Missing values:
    There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.
- **Encoding Categorical data**: That categorical data is defined as variables with a finite set of label values. That most machine learning algorithms require numerical input and output variables. That an integer and one hot encoding is used to convert categorical data to integer data.
- **Count Vectorizer:** Scikit-learn's CountVectorizer is used to convert a collection of text documents to a vector of term/token **counts**. It also enables the pre-processing of text data prior to generating the vector representation. This functionality makes it a highly flexible feature representation module for text.

SPLITTING DATASET INTO TRAIN AND TEST DATA

- Data splitting is the act of partitioning available data into two portions, usually for cross-validator purposes.
- One Portion of the data is used to develop a predictive model and the other to evaluate the model's performance.

- Separating data into training and testing sets is an important part of evaluating data mining models.
- Typically, when you separate a data set into a training set and testing set, most of the data is used for training, and a smaller portion of the data is used for testing.
- To train any machine learning model irrespective what type of dataset is being used you have to split the dataset into training data and testing data.

CLASSIFICATION

Classification is the problem of identifying to which of a set of categories, a new observation belongs to, on the basis of a training set of data containing observations and whose categories membership is known.

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean/average prediction (regression) of the individual trees.

Decision Trees are a type of Supervised Machine Learning (that is you **explain** what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter. An **example** of a **decision tree** can be **explained** using above binary **tree**.

The **SVM** is one of the most powerful methods in machine learning algorithms. It can find a balance between model complexity and classification ability given limited sample information. Compared to other machine learning methods, the SVM has many advantages in that it can overcome the effects of noise and work without any prior knowledge. The SVM is a non-probabilistic binary linear classifier that predicts an input to one of two classes for each given input. It optimizes the linear analysis and classification of hyperplane formation techniques.

The NN algorithm is mainly used for classification and regression in machine learning. To determine the category of an unknown sample, all training samples are used as representative points, the distances between the unknown sample and all training sample points are calculated, and the NN is used. The category is the sole basis for determining the unknown

sample category. Because the NN algorithm is particularly sensitive to noise data, the K-nearest neighbour algorithm (KNN) is introduced. The main concept of the KNN is that when the data and tags in the training set are known, the test data are input, the characteristics of the test data are compared with the features corresponding to the training set, and the most similar K in the training set is found.

PREDICTION

Predictive analytics algorithms try to achieve the lowest error possible by either using "boosting" or "bagging".

Accuracy − Accuracy of classifier refers to the ability of classifier. It predict the class label correctly and the accuracy of the predictor refers to how well a given predictor can guess the value of predicted attribute for a new data.

Speed − Refers to the computational cost in generating and using the classifier or predictor.

Robustness − It refers to the ability of classifier or predictor to make correct predictions from given noisy data.

Scalability − Scalability refers to the ability to construct the classifier or predictor efficiently; given large amount of data.

Interpretability − It refers to what extent the classifier or predictor understand.

RESULT GENERATION

The Final Result will get generated based on the overall classification and prediction. The performance of this proposed approach is evaluated using some measures like,

- Accuracy

  Accuracy of classifier refers to the ability of classifier. It predicts the class label correctly and the accuracy of the predictor refers to how well a given predictor can guess the value of predicted attribute for a new data.

$$AC= \frac{TP+TN}{TP+TN+FP+FN}$$

- Precision

Precision is defined as the number of true positives divided by the number of true positives plus the number of false positives.

$$\text{Precision}=\frac{TP}{TP+FP}$$

- Recall

Recall is the number of correct results divided by the number of results that should have been returned. In binary classification, recall is called sensitivity. It can be viewed as the probability that a relevant document is retrieved by the query.

- ROC

ROC curves are frequently used to show in a graphical way the connection/trade-off between clinical sensitivity and specificity for every possible cut-off for a test or a combination of tests. In addition the area under the ROC curve gives an idea about the benefit of using the test(s) in question.

- Confusion matrix

A **confusion matrix** is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing.

CONCLUSION

We reviewed several influential algorithms for malware prediction based on various machine learning techniques. Characteristics of ML techniques makes it possible to design IDS that have high prediction rates and low false positive rates while the system quickly adapts itself. We divided these algorithms into three types of ML-based classifiers: Random Forest (RF), Support vector machine(SVM), and Decision Tree (DT). Although these two algorithms share many similarities, several features of techniques, such as adaptation, high computational speed and error resilience in the face of noisy information, conform the requirement of building efficient software quality prediction.

REFERENCES

- A. A. Khan, M. H. Rehmani, and M. Reisslein, ‘‘Cognitive radio for smart grids: Survey of architectures, spectrum sensing mechanisms, and networking protocols,’’ IEEE Commun. Surveys Tuts., vol. 18, no. 1, pp. 860–898, 1st Quart., 2016.
- Y. Lin, X. Zhu, Z. Zheng, Z. Dou, and R. Zhou, ‘‘The individual identification method of wireless device based on dimensionality reduction,’’ J. Supercomput., vol. 75, no. 6, pp. 3010–3027, Jun. 2019.
- T. Liu, Y. Guan, and Y. Lin, ‘‘Research on modulation recognition with ensemble learning,’’ EURASIP J. Wireless Commun. Netw., vol. 2017, no. 1, p. 179, 2017.
- Y. Tu, Y. Lin, J. Wang, and J.-U. Kim, ‘‘Semi-supervised learning with generative adversarial networks on digital signal modulation classification,’’ Comput. Mater. Continua, vol. 55, no. 2, pp. 243–254, 2018.
- C. Shi, Z. Dou, Y. Lin, and W. Li, ‘‘Dynamic threshold-setting for RFpowered cognitive radio networks in non-Gaussian noise,’’ EURASIP J. Wireless Commun. Netw., vol. 2017, no. 1, p. 192, Nov. 2017.
- Z. Zhang, X. Guo, and Y. Lin, ‘‘Trust management method of D2D communication based on RF fingerprint identification,’’ IEEE Access, vol. 6, pp. 66082–66087, 2018.
- H. Wang, J. Li, L. Guo, Z. Dou, Y. Lin, and R. Zhou, ‘‘cFractal complexitybased feature extraction algorithm of communication signals,’’ Fractals, vol. 25, no. 4, pp. 1740008-1–1740008-3, Jun. 2017.
- J. Zhang, S. Chen, X. Mu, and L. Hanzo, ‘‘Evolutionary-algorithm-assisted joint channel estimation and turbo multiuser detection/decoding for OFDM/SDMA,’’ IEEE Trans. Veh. Technol., vol. 63, no. 3, pp. 1204–1222, Mar. 2014.