International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-4, February 2020

Machine Learning Based Malware Detection: A Boosting Methodology

Tejaswini Ghate, Chetan Pathade, Chaitanya Nirhali, Krunal Patil, Nilesh Korade

Abstract: ***Malware damages computers without user's consent; they cause various threats unknowingly, hence detection of these is very crucial. In this study, we proposed to detect the presence of malware by using the classification technique of Machine Learning. Classification type in Machine Learning requires the output variable to be of a categorical kind; it attempts to draw some conclusion from the ascertained values. In short, classification constructs a model based on the training set and values or predicts categorical class labels. In our work, we propose to classify the presence of malware by incorporating two chief classification algorithms, such as Support Vector Machine and Logistic Regression. The data set used for it was not satisfactory. Consequently, we tend to explore a data set that met our necessities and enforced Logistic Regression on the same moreover, we plotted a scatter-gram for the scope of visualization and incorporated XG-Boost for the performance enhancement. This study assists in analyzing the presence of malware by adopting a proper dataset and ascertaining pivotal attributes leading to this classification.***

Keywords: Machine learning, Cyber security, Visualization, Malware, Classification, XG Boost.

I. INTRODUCTION

A cyberattack is a malicious and deliberate effort by an individual or team to breach the information system of another individual or organization. Cyberspace and its underlying infrastructure are risky to a wide range of risks stemming from both physical and cyber threats and hazards. The malware attacks like spyware, ransomware, viruses, and worms cause great infliction. Thus, detecting these attacks is an influential component in this study of malware. The key roles in security organization like Chief Information Security Officer, Forensic Computer Analyst, Information Security Analyst, Penetration Tester, Security Architect, IT Security Engineer, Security Systems Administrator, IT Security Consultant are extensively affected which cripples the systems. Machine learning has determined to be productive in association with various fields, among this cyber security is the area in which it has proven to be extremely valuable [1]. Providentially, machine learning can assist in solving the most typical tasks including regression, prediction, and classification. In this era of extremely large amounts of data and cybersecurity expertise deficiency, ML seems to be the only solution. There are various machine learning tasks in cybersecurity, the regression (or prediction) is easy, the information of this present data is utilized to produce an opinion of the new data [2]. In cybersecurity, machine learning can be implemented for fraud detection [3]. The use of classification in terms of cybersecurity would be a spam filter separating spams from other messages. Spam filters were the first ML approach employed in cybersecurity tasks [4], in this article we emphasized more on the classification technique. The security tasks can be classified as prediction, prevention, detection, response, and monitoring and the ML method demonstrates to be productive. There are two classes of malware detection techniques, which are Static and Dynamic techniques [5]. The major difference between these two techniques is that the Static technique doesn't require the execution of the file, whereas Dynamic technique requires the execution of the file. In this study, we have considered the Static malware detection technique. The scope of the study is to detect the presence of malware. The application of Support Vector Machine (SVM) [6] and Logistic Regression Machine Learning algorithms assisted in the precise detection of malware. The study also includes a boosting technique to improve the accuracy of the algorithm on the huge dataset used further in the study. The remaining components of this article are defined as follows, section 2; presents related work which, will set the path for the subsequent sections. The detailed flow of the study is presented in section 3 and section 4 exhibits the specific description of the data-set along with the algorithms and techniques used in the study. Section 5; contains the evaluation of the model along with the results.

II. RELATED WORK

In Machine Learning, the classification technique is used widely to find the class to which the data elements belong; it is very efficient when output is finite and has discrete values. There are two main types of classification mainly Binary and Multi-Class. In Binary classification, there are only two classes taken into consideration the hyper plane is designed accordingly to map data into these two classes whereas; in Multi-class classification, there are more than three classes involved.

Revised Manuscript Received on February 28, 2020.
***** Correspondence Author
Tejaswini Ghate*, Student, Department of Computer Engineering, Pimpri Chinchwad College of Engineering & Research, Ravet, India. E-mail: tejaswinighate9@gmail.com
Chetan Pathade, Student, Department of Computer Engineering, Pimpri Chinchwad College of Engineering & Research, Ravet, India. E-mail: chetanpathade1997@gmail.com
Chaitanya Nirhali, Student, Department of Computer Engineering, Pimpri Chinchwad College of Engineering & Research, Ravet, India. E-mail: chaitanyanirhali03@gmail.com
Krunal Patil, Student, Department of Computer Engineering, Pimpri Chinchwad College of Engineering & Research, Ravet, India. E-mail: pkrunalb888@gmail.com
Nilesh Korade, Assistant Professor, Department of Computer Engineering, Pimpri Chinchwad College of Engineering & Research, Ravet, India. E-mail: nilesh.korade@pccoer.in

The researchers practiced a linear Support Vector Machine (SVM) with ten-fold cross-validation to separate benign from malicious apps to track which features are suspicious. The researchers extracted intents filters and objects from each app and used an SVM with a ten-fold Cross-validation to detect malware. The static analysis points to the analysis of a program without running the program. In static analysis, researchers need to make decryption from the code in advance for static features. However, with the development of malware polymorphism techniques, malware authors began to avoid static analysis by using numerous new polymorphism methods. Dynamic analysis stands for running malicious code in a controllable simulation situation to observe its behavior and find its intent. Compared to static analysis, it does not require disassembly and does not require decryption or decompression. However, dynamic analysis needs a sandbox environment to run the malicious code, and it's hard for this kind of environment to simulate all kinds of situations. Therefore, it's almost impossible for dynamic analysis to traverse every execution branch of the sample [7].

PE (Portable executable) file format was developed by Microsoft and it was introduced with Windows NT 3.1 in the year 1993. It is metadata of executable, dlls and object files. Since its introduction, it has been extensively by researchers in different domains but PE has found its prime importance in Malware detection systems. Wang [8] designed a system, where he used Support Vector Machine to detect malware using PE. SVM was trained on a few features extracted from PE file format. Then it was used to classify PE file into legit or malicious files. Researchers have found different methods to implement malware detection system. Kateryna Chumachenko at el [9] used machine learning to train a model on data associated to APIs calls and return codes of APIs. The model accurately classified 9 malware families but it took 2 to 3 hours to obtain the result. Clemen Kolbitch at el [10] performed a dynamic analysis in a controlled environment; Clemen extracted dynamic system calls and used it to classify executable files into malware and legit files. Waqas Aman [11] again performed dynamic analysis but focused on function call monitoring and information flow tracking. Ijaz, Durad and Ismail at el [5] performed both static and dynamic analysis of malware and legit files and then trained Machine Learning models on collected. In dynamic analysis, they used different combinations of features consisting of Registry, DLLs, APIs and summary information. In static analysis, around 92 features were extracted from .exe files. Static approach proved to be proficient.

PE file format based approaches are effective for scanning executable files but PE cannot be generated for android app. Thus to classify an android app as malware or legit, different techniques are used. In [12] apk file is used to obtain manifest and various features are extracted from it to train the model. Qing-Fei Wang at el [13] describes static and dynamic analysis for android malware detection, where wang uses dex files along with apk file for feature extraction. Peirayian [14] used permission as a feature and combined it with APIs to create a feature vector. In [15] optimally selected feature subset is obtained from the genetic algorithm then used it for further processing. In [16] .dex files obtained were converted to grayscale image and then feed to different models.

III. PROPOSED MODEL

The main purpose of the study is to classify the presence of the malware and for this; we executed logistic regression a principal classification algorithm. It is slightly similar to the linear regression just the variation lies within the property of the output variable, in logistic regression; it is of categorical type (especially in binary form) whereas in linear regression it is of continuous type and mainly in numerical format.

$$y = e \wedge (b0 + b1 * x) / (1+ e \wedge (b0 + b1 * x)) \quad (1)$$

Regression analysis is a statistical technique for estimating the relationships among variables. From such an analysis, we can learn how a dependent variable changes when any one of several independent variables is varied, while the other independent variables are held fixed. On the contrary, logistic regression is used for estimating expected values of a categorical dependent variable in a qualitative response model, based on the explanatory variables. After setting a model, we can pick out the meaningful variables by some variable selection methods. In this model, there left only significant variables or features. In addition, it is considered a more efficient method than step-by-step detection as it has reduced dimensions.The data set used earlier did not have the last column records in proper proportion. The plot of histogram depicts the irregularities of the data set, hence we decided to search for a data set which had sufficient records in proper proportion. The size of the data set was huge enough so logistic regression had a reduced performance as compared to the earlier implementation. In order, to improve the performance we used the boosting technique and the accuracy score of the algorithm was quite astonishing as compared to the previous score.

IV. METHODOLOGY

In general, conventional malware detection systems are a rule-based system, which has a massive database of malware signatures. For malware detection [17] such systems strongly rely on these databases to scan a file against all signatures they have respectively. The basic idea behind our paper is to use Machine Learning to construct a model and train it using a large data set containing windows PE sample, distributed evenly among malicious files and clean files. This eliminates the dependency on the database. Also, a new unknown file can be efficiently identified by the trained model as a malware.

A. Description of Data set

Overall we have used two data sets, results produced using first data set were not so satisfying, main attributes of first data set are as follows

Table I. Description Of The First Data Set

Column Name	Description	Type
hash	MD5 hash of the example	32 bytes string
size_of_data	The size of the section on disk	Integer

virtual_address	Memory address of the first byte of the section relative to the image base	Integer
entropy	Calculated entropy of the section	Float
virtual_size	The size of the section when loaded into memory	Integer
malware	Class	0(Goodware) or 1(Malware)

Second Data set used has 20000 windows Portable Executable (PE) sample evenly distributed among clean and malicious files.PE format is a binary format used in windows to represent executable files, object code and dll files.
The key facts about PE format are:

- It has header, which holds information related to addresses and various other properties
- It has sections, where each section carries data or code or both. Data usually consists of constant, global variables, etc.
- PE also has information of APIs which are used, along with the name of system libraries to which APIs belongs.

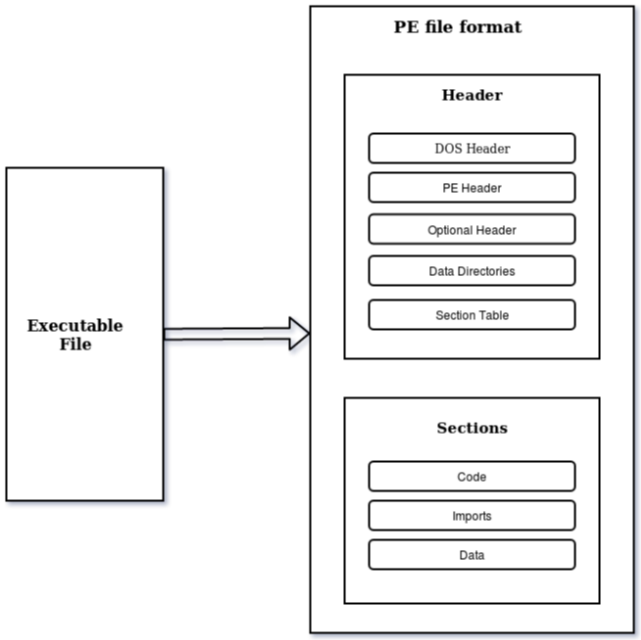

Fig. 1. PE Format

To create our dataset, we have parsed PE file format of each file using a library "LIEF" (Library to Instrument Executable Formats). First 11 attributes of our vector are binary properties, where attribute value '1' represent true and value '0' represent false. Table II describes these 11 scalars. Next 64 attributes represent the first 64 bytes of the PE entry point function, all of which are in the range of 0.0 to 1.0 obtained by normalizing data by dividing each attribute by 255. This approach helps the model to detect unique malware, even those which are slightly varying among several samples of the identical family.

Table II. Description Of The Second Data Set

Column Name	Description
pe.has_configuration	True if the PE has a Load Configuration.
pe.has_debug	True if the PE has a Debug section.
pe.has_exceptions	True if the PE is using exceptions.
pe.has_exports	True if the PE has any exported symbol.
pe.has_imports	True if the PE is importing any symbol.
pe.has_nx	True if the PE has the NX bit set.
pe.has_relocations	True if the PE has relocation entries.
pe.has_resource	True if the PE has any resource.
pe.has_rich_header	True if a rich header is present.
pe.has_signature	True if the PE is digitally signed.
pe.has_tls	True if the PE is using TLS.

B. Boosting

The algorithms used in this investigation resulted in the formulation of more atomic accuracy. Consequently, there was a need to enhance the results thereby resulted in the use of boosting. Boosting is a meta-algorithm used for the supervised learning style for primarily lessening bias and variance [18]. The algorithm transformation from strong to a weak learner is performed using a boosting technique. The gradient boosting is a method that trains the model sequentially. The term y=ax+b+c which represents a loss function where c is the error variable injected. The algorithms used in the boosting procedure involves AdaBoost (Adaptive Boosting), Gradient Tree Boosting, XGBoost.

In this process of improving accuracy, the XGBoost supported in enhancing the speed and performance of the algorithm. It is an implementation of gradient boosted decision trees helping in Execution Speed, Model Performance.

C. Algorithm Used

1) *SVM:* Support Vector Machine is the of a supervised Machine Learning algorithm [19]. For a dataset consisting 'n' number of features, SVM plots all of its data items in an n-dimensional space, each item is rendered as a coordinate in n-dimensional space. To classify data items, SVM searches for a hyperplane that distinguishes two classes very well; to select the most desirable hyperplane, the hyperplane which has a maximum distance from the nearest data point and which appropriately classifies data point is considered. In the dataset used for study, the malware is designated as '1' and '0' otherwise. Hence, we utilized the SVM algorithm as it has a dominance when we have these type of scenarios. The entries of the dataset were mapped into the n-dimensional space after plotting all the entries a hyperplane is devised for more accurate classification.

2) *Logistic Regression:* Unlike linear regression, logistic regression is used to classify categorical attributes. Linear regression cannot be employed to the categorical variable. This is because no linear graph exist which can map 0s and 1s collectively. To overcome this barrier, logistic regression uses a sigmoid function, it takes any real number as input and maps it to values between 0 and 1. Equation of sigmoid function is as below.

$$S(x) = 1 / (1 + e \wedge (-x)) \quad (2)$$

Equation (1) in Proposed Model section is used to describe logistic regression, where y is output value to be predicted based on coefficients b0 and b1, b0 is a bias term and b1 is coefficient to x. In the training phase of the model, the algorithm determines the values of these coefficients which best befits the training dataset. There are mainly three types of logistic regression:

- Binary Logistic Regression: Used when attribute to be predicted has two possible values.
- Multinomial Logistic Regression: Used when the attribute has more than two possible categories not considering ordering.
- Ordinal Logistic Regression: Used when the attribute has more than two possible classes and ordering is important in it.

In our project, attribute to be predicted has two possible outcomes (either malicious or legit); hence we used binary logistic regression.

V. EVALUATION

The two main classification algorithms incorporated in this study were SVM and Logistic Regression. The initial step was to split the data set into the training and testing set, after that we applied these two algorithms and elected the one which produced the best classification accuracy score. The accuracy score was the chief factor for the evaluation purpose as it is determined from the confusion matrix which depicts the performance of the model.

Table III. Algorithm Used In This Study

Category	Algorithm(s)
Classification	Logistic Regression Support Vector Machine
Boosting	Logistic Regression XG-Boost

The accuracy score of logistic regression for the data set from IEEE data port was 94%, but after selecting a new huge data set the accuracy plummeted to 84% hence, there was a need to increase the performance of the model and so, we utilized boosting process and reached an accuracy score of 96%. To further improve the performance of the model we modified the first column which was a label (pe_legit & pe_malicious) into a categorical type which was binary. By repeating the entire procedure on the transformed data set we accomplished an accuracy score of 98% respectively.

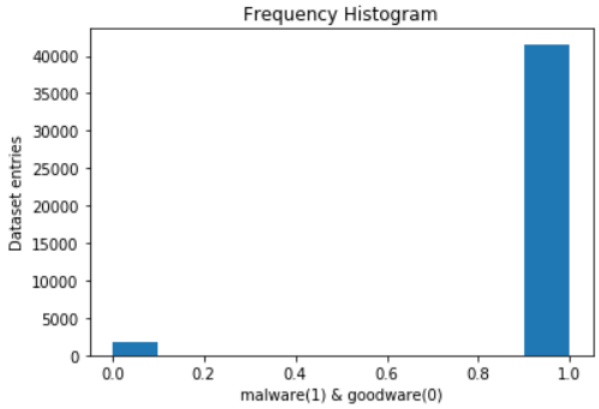

Fig. 2. First Dataset Histogram

The Fig. 2. is about the frequency distribution of Malware(1) and Goodware(0). Accordingly, the algorithm is applied to evaluate the accuracy of the pe-section headers dataset. This histogram specifies about the the disproportionality in the dataset values of attributes.

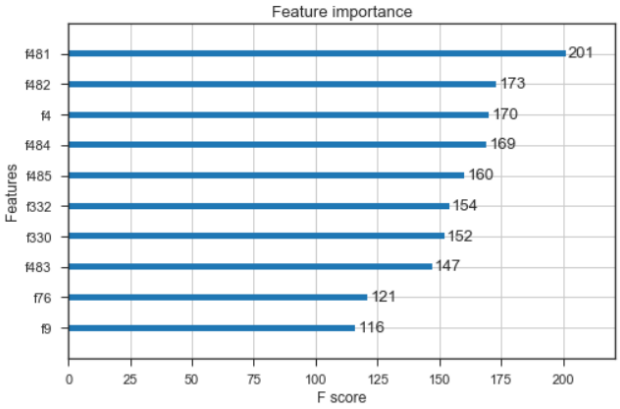

Fig. 3. Second Dataset Histogram

The graphical representation in Fig. 3. specifies about the features extracted from the malware dataset wherein the F-score plotted specifies the attribute values in it. Here the second dataset contains the values that are evenly distributed. The length of histogram describes about the importance of features.

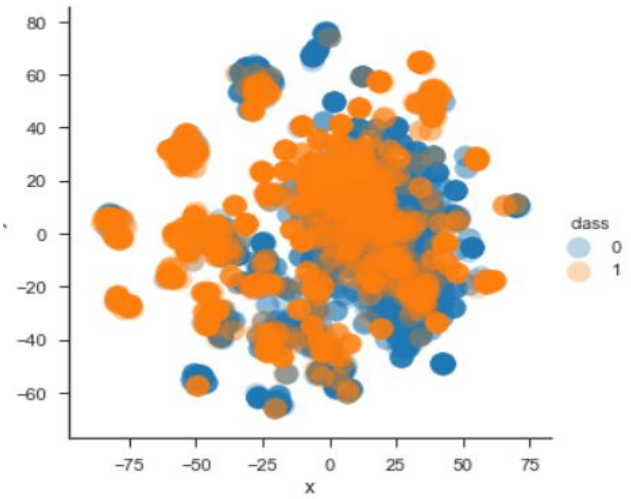

Fig. 4. Scatter gram from Second Dataset

Scatter-gram is utilised to interpret co-relation between variables. The plot below depicts the relation between various input variables with reference to the output variable. Each dot in the plot represents an individual data entry and a no correlation, for reference examine Fig. 4, it's the plot between the input and output variable.

VI. CONCLUSION

Performance of Support Vector Machine and Logistic Regression is good for a first dataset, but in the case of second dataset Logistic Regression is better than Support Vector Machine, when Boosting technique is applied along with Logistic Regression, it gives more accuracy. Hence, the approach of using Boosting techniques proves beneficial than trivial approaches.

ACKNOWLEDGMENT

We grant our special gratitude to the Department of Computer Engineering, Pimpri Chinchwad College Of Engineering & Research, Ravet, Pune for their supervision for the paper completion. We would like to offer our special thanks of gratitude to Prof. Nilesh Korade for escorting us in the accurate track in our research work.

REFERENCES

1. Hafiz M. Farooq,Naif M. Otaibi "Optimal Machine Learning Algorithms for Cyber Threat Detection".
2. Mohammed Lalou,Hamamache Kheddouci,Salim Hariri " Identifying the Cyber Attack Origin with Partial Observation: A Linear Regression Based Approach",November 1918.
3. Christine Hines,Abdou Youssef "Machine Learning Applied to Rotating Check Fraud Detection' ',April 2018
4. Md. Rafiqul Islam, Wanlei Zhou, Morshed U. Choudhury "Dynamic Feature Selection for Spam Filtering Using Support Vector Machine",July 2007.
5. Muhammad Ijaz, Muhammad Hanif Durad, Maliha Ismail, "Static and Dynamic Malware Analysis Using Machine Learning", March 2019.
6. Yaokai Feng, Hitoshi Akiyama,Liang Lu, Kouichi Sakurai "Feature Selection For Machine Learning-Based Early Detection of Distributed Cyber Attacks",Aug. 2018.
7. Qu Wei, Shi Xiao, Li Dongbao, "Malware Classification System Based on Machine Learning", September 2019.
8. Wang, T. Y., Wu, C. H., Hsieh, C. C. "Detecting unknown malicious executables using portable executable headers", IEEE, Aug. 2019.
9. Chumachenko, K. "Machine Learning Methods for Malware Detection and Classification.", 2017.
10. Kolbitsch, C., Comparetti, P. M., Kruegel, C., Kirda, E., Zhou, X. Y., Wang, X. " Effective and Efficient Malware Detection", Aug. 2009.
11. Aman, W. "A framework for analysis and comparison of dynamic malware analysis tools.", 2017.
12. Monica Kumaran, Wenjia Li, "Lightweight malware detection based on machine learning algorithms and the android manifest file", February 2018.
13. Wang Qing-Fei, Fang Xiang, "Android Malware Detection Based on Machine Learning", April 2018.
14. Peiravian, Naser, and X. Zhu. "Machine learning for Android malware detection using permission and API calls." IEEE, International Conference on TOOLS with Artificial Intelligence IEEE, 2014:300-305.
15. Anam Fatima, Ritesh Maurya, Malay Kishore Dutta, Radim Burget, Jan Masek, "Android Malware Detection Using Genetic Algorithm based Optimized Feature Selection and Machine Learning", July 2019.
16. Fauzi Mohd Darus, Salleh Noor Azurati Ahmad, Aswami Fadillah Mohd Ariffin, "Android Malware Detection Using Machine Learning on Image Patterns", January 2019.
17. Muhammad Ejaz Ahmed, Surya Nepal, Hyoungshick Kim "MEDUSA: Malware detection using statistical analysis of system's behavior", Aug. 2018.
18. Vina Ayumi "Pose-based Human Action Recognition with Extreme Gradient Boosting", January 2017.
19. Wenjia Li*, Jigang Ge, Guqian Dai "Detecting Malware for Android Platform: An SVM-based Approach", Nov. 2015 .
20. Ivan Firdausi, Charles lim, Alva Erwin, Anto Satriyo Nugroho, "Analysis of Machine learning Techniques Used in Behavior-Based Malware Detection", December 2010.

AUTHORS PROFILE

Tejaswini Ghate, She is currently studying Computer Engineering from PCCOE&R. Her field of interest is Machine Learning and Software Development. She worked on a project entitled Churn Modeling. She looks ahead to explore more in the field of Machine Learning.

Chetan Pathade, He is currently studying Computer Engineering from PCCOE&R. His field of interest is Cyber Security. He is a Certified Ethical Hacker. He is planning to pursue his Master's, where he looks forward to more research opportunities in the field of Cyber Security.

Chaitanya Nirhali, He is currently studying Computer Engineering from PCCOE&R. His field of interest is Machine Learning. He has gained experience through his internship as an Machine Learning intern & will be pursuing his Master's in the field of Computer Science.

Krunal Patil, He is currently studying Computer Engineering from PCCOE&R. His field of interest is Machine Learning and Web Development. He has gained experience through an internship as a developer. He wants to explore more in the domain of Machine Learning.

Nilesh Korade, He is currently an Assistant professor of the Computer Engineering Department at Pimpri Chinchwad College of Engineering & Research. He has more than 7 years of teaching experience. He has expertise in Machine Learning, IOT, and Cyber Security. He is currently pursuing his PhD.