International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-5, January 2020

Assessment of Machine Learning Classifiers for Malware Detection

K. Meghana, K. Satya Priya, T. V. V. L. Sruthi, T. Gunasekhar

Abstract: ***In our daily life, cell phones (e.g., cell phones and tablets) have met an expanding business achievement and have turned into an essential component of the regular daily existence for billions of individuals all around the globe. Day by day the advancements in technology is growing like an infinity thing .And the advancements in technology made everyone to use the smart phones and tablets regardless their professions .Everyday a big range of apps coming in to existence which made our lives very comfortable. While installing these apps without knowing we are allowing some malware in to our mobile which may leads to leakage of once private information. So in this paper we are going to analyze some machine learning techniques which will help in malware classification by taking the dataset. In this paper we calculated accuracy rate of malware classifiers such as KNN, Random Forest, SVM, and Gaussian Etc. Where we will be rating all these machine learning techniques according to their rate of accuracy. According to the experiments what we conducted Random forest stood as the best malware classifiers among all the other classifiers. We accept our study will be a reference work for specialists and experts in this examination field.***

Keywords: Mobiles and PCs, Malware detection, classification algorithms, random forest.

I. INTRODUCTION

THE most recent decade has been set apart by the ascent of cell phones which are these days broadly spread among individuals. The most diffused instances of such cell phones are mobiles and tablets. Cell phones and PCs have an immensely expanded computational power, increasingly accessible memory, a bigger showcase, and Internet availability through both Wi-Fi and cell systems. In addition, such gadgets run portable working frameworks which can encounter mixed media substance, just as to run versatile applications (likewise called applications). Consolidated together, these components make both cell phones and tablets to have similar functionalities ordinarily offered by workstations and personal computers.

Manuscript published on January 30, 2020.
* Correspondence Author
K.Meghana*, Department, Computer Science and Engineering, Koneru Lakshmaiah Educational Foundation situated at Vaddeswaram, Guntur District.
K.Satya Priya, Department, Computer Science and Engineering, Koneru Lakshmaiah Educational Foundation situated at Vaddeswaram, Guntur District.
T.V.V.L.Sruthi, Department, Computer Science and Engineering, Koneru Lakshmaiah Educational Foundation situated at Vaddeswaram, Guntur District.
T. Gunasekhar, Associate Professor, Department, Computer Science and Engineering, Koneru Lakshmaiah Educational Foundation situated at Vaddeswaram, Guntur District.

As indicated by the insights detailed, cell phone clients were 25.3% of the worldwide populace in 2015, and this rate is relied upon to develop until 37% in 2020. So also, the insights about tablets announced show a worldwide entrance of 13.8% in 2015, expected to arrive at 19.2% in 2020[8]. The main thrusts of this great achievement are the pervasive Internet availability, on account of the overall sending of cell and Wi-Fi systems,
and countless applications available in the official (and unofficial) marketplaces.A cell phone ordinarily has a great deal of delicate data about its proprietors, for example, contacts, photographs and recordings, and GPS position. Such data must be appropriately secured, particularly when it is transmitted to remote administrations.

The whole world is running behind the applications of Smartphone. Everyday some new apps were launching in to the market which are making everything to done on our finger tips.so while installing apps we never concentrate on terms and conditions what we are accepting and also what type of applications are intruding in to our phones. Sometime along with these applications there is a chance of getting malware in to our mobiles. This can damage our computer, server, client etc. It can also help in leakage of sensitive information.

A malware can simply be defined as a malicious program which the user unsuspectingly install on their machine and later these programs can begin to disrupt the proper operation of the machine or might continue unnoticed and carry out malicious actions without been noticed 9. When the attacker gains control of the machine, he can then have access to any information stored on the machine. Some of the deceptive approaches used to install malware on the computer system through the internet include repackaging the software, update attack 10 or desire for download. So in our paper we depict some major malware classifiers such as Random Forest, Gaussian, SVM, KNN etc. We took a dataset for this classification. Where we will be rating all these machine learning techniques according to their rate of accuracy.

II. RELATED WORKS

In this paper, we worked on different techniques that would detect malwares.Malware is a large term that focuses to an injection of harmful projects [1]. This word will define a few of the most popularly known sorts of malicious terms; adware, bots, bugs, rootkits, spyware, Trojan steeds, infections, and worms. Let us know different types of malwares.

Trojan horse:

Trojan horse is a plan of action that is downloaded and introduced on a PC that seems innocuous, however, it is, truth be told, malevolent.

Astonishing changes to desktop fixings and unpredictable movement, in any event, when the PC prerequisite is to be not active, are rigid signs that a Trojan is dwelling on a PC. Conventionally, the Trojan horse is concealed in an exemplary looking email association or free download. Right when the customer taps on the email association or downloads the free program, the malware that is concealed inside is moved to the customer's handling device. Once inside, the malevolent code can execute whatever errand the assailant planned it to complete.

Adware:

Is malware whose solitary reason for existing is to demonstrate notices to the client. They are viewed as one of the least compromising classifications of malware. They will likely show on the influenced PC plugs which the client is probably going to be pulled in to, it records information from the PC, for example, program and web search tools narratives [12]. Adware is now and again delegated spyware subject to the earnestness of the chronicle. Adware, or publicizing bolstered programming, is any product bundle that naturally plays, shows, or downloads notices to a PC. These ads can be as a spring up. The object of the Adware is to create income for its author [12]. Adware, without anyone else's input, is innocuous; in any case, some adware may accompany coordinated spyware, for example, key loggers and other security obtrusive programming.

Spyware:

It is a sort of self-introducing malware that executes without the client's endorsement. It is utilized to assemble and follow data about the individual and the perusing history of a PC framework [6]. It is commonly bundled together with programming that is made accessible to clients at no cost. Spyware is additionally called rootkit as a result of the bundling with freeware [2]. Spyware is a code that empowers an outsider to keep an eye on a host. Spyware has been utilized for an assortment of purposes including wholesale fraud and robbery of individual information, keeping an eye on online exercises of people (for example life partners) and watching clients' online exercises[2]. It is a kind of malware introduced on PCs that gathers data about clients without their insight.

Worm:

Is a malware that doesn't append itself to another programming as it needn't bother with a host programming to attach itself. This is the thing that separates the worm from the infection [3] .A worm typically influences its injured individual through the zone of exposures that it can abuse. It utilizes different intends to spread, and degenerate other PC frameworks. Worms have the ability to wreck a similar degree of destruction an infection will cause to a contaminated PC framework [7] . Worms are not parasitic in conduct like the infections. They are free projects that can cause hurt alone. These worms might possibly have a payload yet the two kinds can be quite destructive.

III. METHODOLOGY

The whole technique of methodology will concentrates on classifying the malware classifiers some of them are Random forest, decision tree, gaussian naïve Bayes, knn, logistic regression.

Randomforest:

Random Forest are an outfit learning strategy for grouping, relapse and different errands that works by building a large number of choice trees at preparing time and yielding the class that is the method of the classes or mean expectation of the individual trees[4]. Irregular choice timberlands right for choice trees' propensity for overfitting to their preparation set. As part of their development, arbitrary woodland indicators normally lead to a uniqueness measure among the perceptions. One can likewise characterize an arbitrary timberland difference measure between unlabeled information: the thought is to develop an irregular woodland indicator that recognizes the "watched" information from appropriately produced engineered information. The watched information are the first unlabeled information and the manufactured information are drawn from a reference dispersion. An irregular timberland divergence can be alluring in light of the fact that it handles blended variable sorts well indeed, is invariant to monotonic changes of the information factors, and is vigorous to peripheral perceptions.

K Nearest Neighbor:

In design acknowledgment, the k-nearest neighbors calculation (k-NN) is a non-parametric technique utilized for arrangement and relapse. In the two cases, the info comprises of the k nearest preparing models in the element space. The yield relies upon whether k-NN is utilized for order or relapse:

In k-NN characterization, the yield is a class enrollment. An item is grouped by a majority vote of its neighbors, with the article being allotted to the class generally normal among its k closest neighbors (k is a positive whole number, commonly little). In the event that k = 1, at that point the item is essentially allocated to the class of that solitary nearest neighbor. In k-NN relapse, the yield is the property estimation for the article[9]. This worth is the normal of the estimations of k nearest neighbors. k-NN is a sort of occasion based learning, or languid realizing, where the capacity is just approximated locally and all calculation is conceded until classification. Both for characterization and relapse, a helpful method can be to allocate loads to the commitments of the neighbors, so that the closer neighbors contribute more to the normal than the more far off ones. For instance, a typical weighting plan comprises in giving each neighbor a load of 1/d, where d is the separation to the neighbor. The neighbors are taken from a lot of items for which the class (for k-NN grouping) or the article property estimation (for k-NN relapse) is known[9]. This can be thought of as the preparation set for the calculation, however no unequivocal preparing step is required. A quirk of the k-NN calculation is that it is delicate to the neighborhood structure of the data. The best decision of k relies on the information; by and large, bigger estimations of k diminishes impact of the commotion on the classification, yet make limits between classes less particular. A decent k can be chosen by different heuristic systems. The exceptional situation where the class is anticipated to be the class of the nearest preparing test (for example at the point when k = 1) is known as the nearest neighbor calculation.

Decision Tree:

DTs pursue a tree-organized arrangement plot where the hubs speak to the info factors and the leaves relate to choice results. DTs are one of the soonest and most conspicuous ML techniques that have been generally applied for characterization purposes. In light of the engineering of the DTs, they are easy to translate and "brisk" to learn [11]. When navigating the tree for the arrangement of another example we can guess about its group. The choices came about because of their particular engineering take into account sufficient thinking which makes them an engaging system [11].

The underneath figure portrays a representation of a DT with its components and principles.

Naïve Bayes:

In AI, naive Bayes classifiers are a group of basic "probabilistic classifiers" in light of applying Bayes' hypothesis with solid freedom suppositions between the highlights. They are among the least difficult Bayesian system models . It was bring into the content recovery network and stays a prominent strategy for content arrangement, the issue of making a decision about archives as having a place with one classification or the other, for example, spam or genuine, sports or governmental issues, and so on with word frequencies as the highlights[7]. With suitable pre-processing, it is aggressive in this area with further developed techniques including bolster vector machines It additionally discovers application in programmed restorative determination[7]. Innocent Bayes classifiers are exceptionally adaptable, requiring various parameters straight in the quantity of variable in a learning issue. Most extreme probability preparing should be possible by assessing a shut structure expression , which takes direct time, as opposed to by costly iterative estimation as utilized for some, different kinds of classifiers. It is a straightforward strategy for building classifiers models that allot class marks to issue cases, spoke to as vectors of highlight esteems, where the class names are drawn from some limited set.

Logistic Regression:

Logistic Regression is fundamentally a regulated characterization calculation. In a characterization issue, the objective variable (or yield), y, can take just discrete qualities for a given arrangement of features (or inputs), X.
As opposed to prevalent thinking, calculated relapse IS a relapse model. The model forms a relapse model to anticipate the likelihood that a given information section has a place with the classification numbered as "1". Much the same as linear relapse expect that the information pursues a straight work, Logistic relapse models the information utilizing the sigmoid capacity. Logistic regression turns into an order method just when a choice limit is brought into the image. The setting of the limit esteem is a significant part of Logistic regression and is reliant on the grouping issue itself.

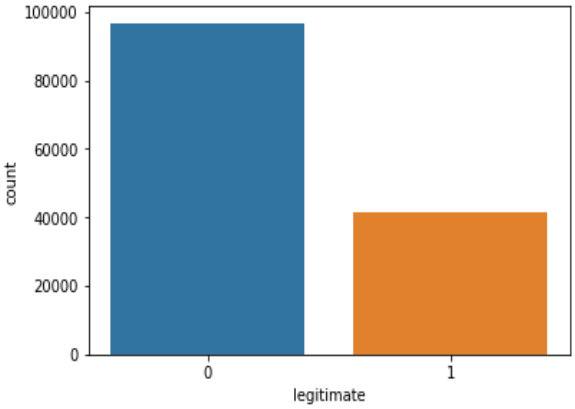

Fig 1: Plot that counts number of legitimate and malicious files in dataset.

This plot shows number of legitimate files when compared to malicious file. In our dataset there are 138048 files in total which consists of both legitimate and malicious. In X-axis 0 indicates number of malicious files and 1 indicates number of legitimate files in dataset.

Table 1: Analysis of different classifiers for malware detection

Machine learning algorithms	Accuracy (%)
Decision tree	99.07
Random Forest	99.4
GNB	70.3
KNN	98.9
Logistic regression	29.6

Above table is the table about accuracy of different classifiers used for malware detection. Random forest shows the highest accuracy among all the other accuracy.

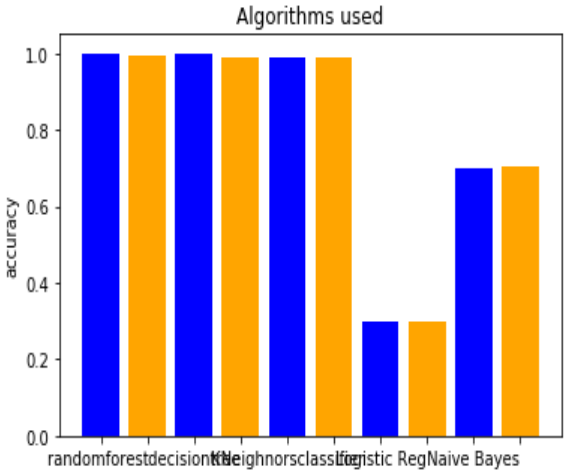

Fig 2: Plot for malware detection using different machine learning classifiers.

This graph plots accuracy for training and testing dataset of each and every malware classifier used.

IV. IMPLEMENTATION

1. **Import the libraries and dataset:**

   This is the basic step where we import libraries which are required in order to process our data and here we will also import our data set. Generally libraries' are a collection of precompiled routines that are used by the program while running. Some of the libraries that we used are numpy, panda, pickle etc. Among the libraries we used numpy and panda are default libraries.

2. **Feature selection using tree classifier:**

   In this step we will select a particular subset which contains relevant features. This is also called as variable selection .this step is done in order to reduce the complexity and to make the training process faster.

3. **Splitting dataset into training and testing data:**

   As the machine learning process done all the process by the previous experience .first it divides the entire data in two types one is the training data and the other is the testing data. Training set is the one on which we prepare and accommodate our model fundamentally to fit the parameters though testing data is utilized distinctly to assess the execution of the classifier.

4. **Apply decision tree, random forest [4], KNN, Gaussian naïve Bayes, logistic regression:**

   In order to find the accuracies of different classifiers first we need to apply those classifiers to the data like KNN, Random forest, decision tree etc. in this step we wrote codes for each and every classifiers.

5. **Evaluate results of accuracy for those classifiers:**

   This is the last step of our process which involves the finding the accuracies of each and every algorithm and comparing of every algorithm with the other. And at last we will have the classifier with highest accuracy rate.

   Thus these steps are implemented for detecting malware files against legitimate files.

Dataset

We needed dataset of legitimate and malware. Various attributes included in dataset are MD5 checksum, version information, optional header, size of heap commit, legitimate which are described in the Table 2. Dataset is divided into training and testing data. Classification algorithms are trained using training dataset and testing dataset is used to determine efficiency of algorithm. From the dataset used, 80% of both profiles (legitimate and malicious) are used to prepare a training dataset and 20% of both files are used to prepare a testing dataset.

Table 2: Description of different attributes in the dataset

Attribute	Explanation
Md5	A MD5 hash is made by taking a string of an any length and encoding it into a 128-piece unique mark. Encoding a similar string utilizing the MD5 calculation will consistently bring about the equivalent 128-piece hash yield. MD5 hashes are utilized to guarantee the information respectability of records.
Optional header	Each document has an optional header that gives data to the loader. This header is discretionary as in certain records (explicitly, object documents) don't have it. Note that the size of the optional header isn't fixed.
Checksum	A checksum is a little measured datum got from a square of advanced information to identify mistakes that may have been presented during its transmission or capacity.
Version Information	The version-information resource Contains the adaptation number of the document, its planned working framework, and the first record name.
Size of heap commit	In general the inbuilt size of the heap is equal to 1 MB. The rounds of the linker explains the specified value to the nearest 4 bytes.
legitimate	The safe file that is not malicious and can be the software that can be installed with out fear of malware is legitimate file.

IV. PERFOMANCE MEASURE

Confusion is a method for evaluating the capabilities of a classifier. The evaluation of the confusion matrix will help the user by giving a thought about which classifier will give better results among the classifiers we used and also about corrections that we need to do in order to correct the classifier we are using. We utilized four execution estimates like (1) True Positive Rate (TPR) expected positive and result is also positive, (2)True Negative Rate (TNR) expected negative and output is also negative (3)False Positive Rate (FPR) expected positive but output is negative and (4) False Negative Rate (FNR) expected negative but output is positive. Accuracy, all in all, tells about how normally is our calculation or model is right and it is the extent of accurately evaluated perceptions to the complete perceptions. (6) All the information like when the calculation gauges positive and how often it is correct will be examined by precision.

Efficiency/Accuracy = Number of correct predictions/ total number of predictions.

Percent Error = (1-Accuracy)*100

$$TPR=TP/(TP+FN) \quad (1)$$

$$FPR=FP/(FP+TN) \quad (3)$$

$$TNR=TN/(FP+TN) \quad (2)$$

$$FNR=1-TPR \quad (4)$$

Recall - How many of the true positives were called(found),i.e. what number of the right hits were likewise found.

$$Recall = TP/(TP+FN)$$

Precision- Precision is how many of the returned hits were true positive i.e. what number of the found were right hits.

$$Precision - TP/(TP+FP)$$

VI. CONCLUSION

This paper gives the comparative study of malware detection using five different machine learning classifiers. From the state-of-art models such as Decision Tree, Random Forest, K- Neural Networks, Logistic Regression, Gaussian Naïve Bayes are used in this study. The predictions indicate that the Random Forest algorithm produced 99.4% accuracy. From the experimental results obtained Random Forest is the best machine learning algorithm for malware detection.

REFERENCES

1. Mohamad Baset. Machine Learning for Malware Detection. MSc. Dissertation, School of Mathematical and Computer Sciences, Heriot-Watt University, 62 pages, 2016.
2. Symantec. What are malware, viruses, spyware, and cookies, and what differentiates them? 2009. Online. http://www.symantec.com/connect/articles/what-are-malware-viruses-spyware-and-cookies-and-what-differentiates-them.
3. Cisco (2015). What is the difference: Viruses, worms, trojans, and bots? Online. http://www.cisco.com/web/about/security/intelligence/virus-worm-diffs.html.
4. Mohammed K. Alzaylaee, Suleiman Y. Yerima and SakirSezer. EMULATOR vs REAL PHONE: Android Malware Detection Using Machine Learning, IWSPA 2017 Proceedings of the 3rd ACM International Workshop on Security and Privacy Analytics, co-located with CODASPY'17, pages 65-72, Scottsdale, Arizona, USA - March 24 - 24, 2017. DOI: 10.1145/3041008.3041010.
5. Moser, A., Kruegel, C., and Kirda, E. Limits of static analysis for malware detection. In Computer security
6. Brunton F. and Nissenbaum H. Political and Ethical Perspectives on Data Obfuscation, 2012, pp. 164-188.
7. Zeidanloo H. R., Tabatabaei S. F., Amoli P. V. and Tajpour A. All About Malwares (Malicious Codes), 2015.
8. Sanjay Chakrabortya and LopamudraDey. A rule-based probabilistic technique for malware code detection. Multiagent and Grid Systems – An International Journal, IOS Press, 12, 2016, pp. 271–286 271. DOI 10.3233/MGS-160254
9. Kateryna Chumachenko. Machine Learning Methods for Malware Detection and Classification. Bachelor's Thesis information Technology. University of Applied Sciences, 93 pages, 2017.
10. Mohammed K. Alzaylaee, Suleiman Y. Yerima and SakirSezer. EMULATOR vs REAL PHONE: Android Malware Detection Using Machine Learning, IWSPA 2017 Proceedings of the 3rd ACM International Workshop on Security and Privacy Analytics, co-located with CODASPY'17, pages 65-72, Scottsdale, Arizona, USA - March 24 - 24, 2017. DOI: 10.1145/3041008.3041010.
11. Baldangombo, U., Jambaljav, N. &Horng, S.J. A Static Malware Detection System Using Data Mining Methods. International Journal of Artificial Intelligence & Applications, 4(4), p.113, 2013. Available at:http://arxiv.org/abs/1308.2831
12. Radu-Stefan Pirscoveanu. Clustering Analysis of Malware Behaviour. Master Thesis Department of Electronic Systems at Aalborg University, 139 pages, 2015

AUTHORS PROFILE

K. Meghana, is a student of the Computer Science and Engineering Department at the Koneru Lakshmaiah Educational Foundation situated at Vaddeswaram, Guntur District.

K. Satya Priya, is a student of the Computer Science and Engineering Department at the Koneru Lakshmaiah Educational Foundation situated at Vaddeswaram, Guntur District.

T. V. V. L. Sruthi, is a student of the Computer Science and Engineering Department at the Koneru Lakshmaiah Educational Foundation situated at Vaddeswaram, Guntur District.

T. Gunasekhar, is an associate professor of the Computer Science and Engineering Department expertise in cloud computing at the Koneru Lakshmaiah Educational Foundation situated at Vaddeswaram, Guntur District.