A Survey on Different Classification Techniques for Detecting Phishing Websites

Maanasa Narayan[1]Gowri S. J.[2]Malavika.P.B[3] Meghana.R[4]Kushal Kumar.B.N[5]

Abstract*—Phishing is a fraudulent attempt to steal user's personal information and with the increase in the number of phishing websites in 2017, the users need to know what to look out for to stay safe online. Therefore there is a need to develop dynamic anti-phishing techniques. This can be achieved using data mining and machine learning techniques. In this paper, we have compared various Machine Learning algorithms on the basis of their accuracy, precision and recall. A dataset with 4500 URLs has been used in this experiment and we have observed that tree-based classifiers like Random Forest, Random Trees, J48 Tree and LMT gives the highest accuracy.*

Keywords—Phishing, Data Mining, Machine Learning, Classification, URL

I. INTRODUCTION

Phishing is an identity theft based internet fraud. It is a kind of attack in which phishers use spoofed emails and malicious websites to steal personal information of users. The effect is the rupture of data security. Phishing is a serious problem in the progressively limitless service of the Internet. There are many ways to trick the people to disclose the information from the users by using social engineering attack. In this, the attacker bait the users by sending mails such as prize winning, send message from fake account on social networking sites, hacking password, send emails to victims which seems like it is sent by banks to disclose the information for financial gain. For example, if the secret key of user's credit card has been entered incorrectly, the messages displayed will take the user to the self-made website page that demands the user to provide or change their account number and password through the hyperlink given in the e-mail. If the user submits the account number and secret key, the phishers then effectively gather the data at the server side, and can perform malicious activity with that data (e.g., pull back cash out from your record, changing of password etc.).

The existing approach is the blacklist method. In this method, the requested URL is compared with a predefined set of phishing URLs. Most of the browsers use this approach. The drawback of this method is that it typically doesn't deal with all phishing websites since newly launched fake website takes a large amount of time before it is being added to the blacklist. In this paper, a comparison of the heuristic methods is done. In contrast to the blacklist method, the heuristic-based method can identify newly created phishing websites in real-time. The success of an anti-phishing technique mainly depends on recognizing phishing websites within an acceptable time period.

This paper consists of six sections. First section consists of introduction; section 2 illustrates Literature Survey, Section 3 explains the different types of phishing attacks, Section 4 explains the different classification techniques that can be used for detection, Section 5 shows the experimental results after the comparison, followed by Conclusion in Section 6.

II. LITERATURE SURVEY

Data mining techniques are useful in real time prediction of phishing websites. Hence comparisons of classifiers are done.[6] In this paper they have compared algorithms to predict e-banking phishing website. Data mining Techniques used for this purpose are JRip, PART, PRISM, C4.5, CBA (Correlation Based Ensemble) and MCAR (Multi Classification Association Rule). After examining the phishing websites 27 characteristics are extracted from URL and domain identity, security and encryption, source code, page style and contents, web address bar and human factors related to sites. A phishing training data set has been created and six algorithms are applied on the training set. JRip gives the highest error rate and MCAR technique gives lowest error rate. In another paper [7], fuzzy data mining algorithm has been analyzed. Three different phishing types and six different criteria for detecting phishy websites with a layer structure has been defined in the paper. Fuzzy logic and RIPPER data mining algorithm have been applied in order to predict phishing emails.

III. TYPES OF PHISHING ATTACKS

A. Deceptive Phishing

The most widely recognized kind of phishing trick, deceptive phishing alludes to any assault by which fraudsters mimic a genuine organization and endeavor to take individuals' private information or login accreditations. Those messages utilize dangers and a feeling of earnestness to unnerve clients into doing the attackers' offering.

B. Spear Phishing

In spear phishing tricks, fraudsters redo their assault messages with the target's name, position, organization, work telephone number and other data trying to trap the beneficiary into trusting that they have an association with the sender. The objective is the same as deceptive phishing: draw the casualty into tapping on a noxious URL or email connection, with the goal that they will hand over their own information.

C. CEO Fraud

Phishers utilize an email deliver like that of a specialist figure to ask for installments or information from others inside the organization. The fundamental goal is: For the casualty to exchange cash straightforwardly to the cyber criminals.

D. Pharming

Fraudsters hijack a website's domain name and use it to redirect the visitors to an imposter site. The main objective is to intercept and steal online payments.

E. Dropbox Phishing

Realistic-looking emails claiming to come from drop box request the user to click through to "secure "their account or

download a shared account. the main objective is to install malware on the victims computer.

F. Google Docs Phishing

A message invites victims to view documents on Google docs. The landing page is indeed on Google drive so it seems convincing, but entering your credentials will send them straight to the scammers. The objective is access to your Google, including Gmail, Google play and android application.

IV. CLASSIFICATION ALGORITHMS

A. Naïve Bayes Classifier

Naive Bayes classifier applies Bayes' theorem and it is a simple technique for constructing classifiers. [1]Despite the naive design, these classifiers work well in many complex real-world situations. The benefit of using Naive Bayes is that, it needs only a small amount of training data to estimate the parameters necessary for classification.

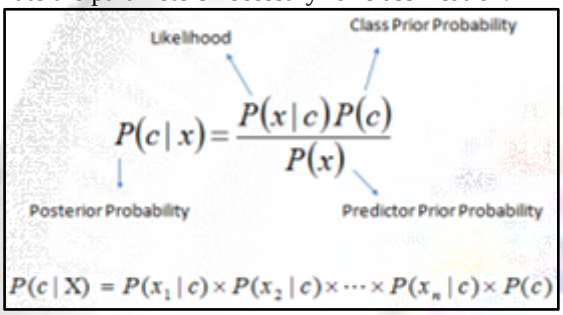

Fig.1: Naïve Bayes Theorem

B. J48 Algorithm

J48 is an improvement of ID3, which is a decision tree algorithm. The extra features present in J48 are accounting for missing values, decision trees pruning, continuous attribute value ranges, derivation of rules, and so on. In WEKA data mining tool, J48 is an open source Java implementation of the C4.5 algorithm. It uses an extension to information gain known as gain ratio, which attempts to overcome the bias for certain attributes.

C. Logistic Model Tree

A logistic model tree is a classification model with an associated supervised training algorithm, which is a combination of decision tree learning and logistic regression.[2] Logistic model trees are based on the idea of an earlier model tree. The idea is as follows: a decision tree that has linear regression models at the leaves to provide a piece wise linear regression model, whereas ordinary decision trees with constants present at the leaves would result in a piecewise constant model. The basic Logistic Model Tree induction algorithm utilizes cross-validation to find the number of Logit Boost iterations that does not over fit the training data.

D. Random Forest Algorithm

Random forests are a learning strategy for classification, regression and different errands, that work by building a large number of decision trees at training time and yielding the class that is the method of the classes (classification) or mean prediction (regression) of the individual trees.[3] Random decision forests amend for decision trees' propensity for over fitting to their training set. Specifically, trees that are become profound have a tendency to learn exceedingly sporadic patterns: they overfit their training sets, i.e. have low inclination, however high change. Random forests are a method for averaging numerous profound decision trees, prepared on various parts of a similar training set, with the objective of decreasing the change. This comes at the expense of a little increment in the inclination and some loss of interpretability, however largely improves performance in the final model.

E. Random Tree Algorithm

Random Tree is known as a supervised Classifier. Lot of individual learners are generated using this algorithm due to which it is an ensemble learning algorithm It employs a bagging idea to construct a random set of data for constructing a decision tree. In the standard tree every single node is split using the best split among all variables. In a random forest algorithm, every node is split after selecting the best among the subset of predicators which are randomly chosen at that node. Random trees have been introduced by Leo Breiman and Adele Cutler. The algorithm can deal with both classification and regression problems. Random trees are a group (ensemble) of tree predictors that is called forest. The classification mechanisms as follows: the random trees classifier gets the input feature vector, classifies it with every tree in the forest, and outputs the class label that received the majority of "votes". In case of a regression, the average of the responses over all the trees in the forest is called the classifier reply. Random Trees are essentially the combination of two existing algorithms in Machine Learning: single model trees are merged with Random Forest ideas.

F. C4.5 Algorithm

C4.5 algorithm was proposed by Quinlan in the year 1993[4]. C4.5 is an extension of ID3 algorithm which was also found by Quinlan.C4.5 is also called statistical classifier because decision tree generated by this algorithm can be used for classification. Description of c4.5 algorithm provided by the authors of WEKA machine learning tool authors is as follows: "A landmark decision tree program that is probably the machine learning workhorse most widely used in practice to date" [5].

G. ID3 (Iterative Dichotomiser 3)

ID3 is the forerunner to the C4.5 algorithm. It is used to create a decision tree from a given informational collection by utilizing a top down, voracious pursuit, to test each trait at each node of the tree. The subsequent tree is utilized for classification of future samples. The means in this algorithm are:

Estimating the entropy of each trait utilizing the data set ; Divide the set into subsets utilizing the trait for which entropy is least (or, information gain is most extreme) ; Create a decision tree node containing that trait. ; Recourse on subsets utilizing remaining traits.

H. C&RT Algorithm

Classification and Regression Trees(CART) which integrates the classification and regression trees for foreseeing persistent dependent factors (regression) and all

predictor factors (classification), is a recursive division method. The primary components of CART are:

Rules for dividing information at a node in light of the estimation of one variable; Stopping rules for choosing when a branch is terminal and cannot be divided anymore; and Finally, an expectation for the target variable in every terminal node.

1) K-Nearest Neighbor Algorithm

In design acknowledgment, the k-closest neighbor's calculation (k-NN) is a non-parametric method used for relapse and classification. In the two cases, the information comprises of the k nearest training cases in the component space. The yield relies upon whether k-NN is utilized for classification or regression.

In k-NN classification, the yield is a class enrollment. An object is characterized by a greater part vote of its neighbors, with the assignment of the object to the class most regular among its k closest neighbors (k is a positive whole number, normally little). On the off chance that k = 1, at that point the object is basically alloted to the class of that solitary closest neighbor.

In k-NN regression, the yield is the property estimation for the object. This value is the normal of the estimations of its k closest neighbors.

V. Experimental Results

The classification of the data after extracting the relevant features was performed by using the WEKA tool with the following algorithms: Naive Bayes, J 48 Tree, LMT, Random Forest, Random Tree, C 4.5, ID3, C & RT and K-Nearest Neighbor.

Classification Algorithm	Naïve Bayes	J48 Tree	LMT	Random Forest	Random Tree	C4.5	ID3	C & RT	K-Nearest Neighbor
Training Accuracy %	89.73	97.3	98.16	99.5	99.6	97.97	94.17	97.7	95.6
Cross Validation (10 Fold) %	89.63	96.46	96.86	97.07	96.63	97.07	93.33	96.53	93.5
Cross Validation (3 Fold) %	88.16	96.83	97.13	97.17	96.67	96.3	93.87	96.47	92.47
Leave one out %	89.73	96.26	97.53	96.93	96.4	96.93	94.13	96.97	94.5

Table 1: Accuracy of Different Classifiers

From Table 1 and Figure 2 and 3, we observe that the classification accuracy, precision, and recall values are high for the tree based classification algorithms compared to the other frequently used algorithms. We can infer that the classification accuracy for the different categories of phishing URLs is around 98 to 99% in various sub domains, by using the tree-based classification algorithms. Tree-based classifiers are best suited for performing phishing URL classification.

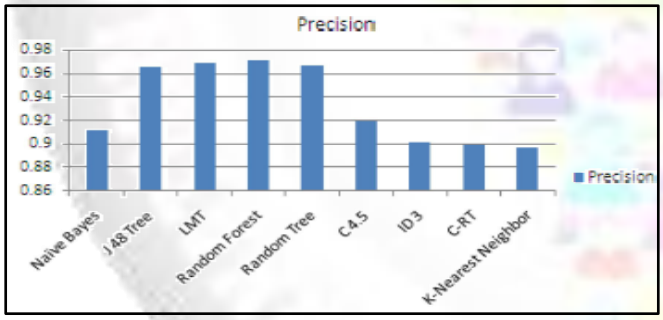

Fig. 2: Graph Comparing the Precision of Different Classification Algorithms

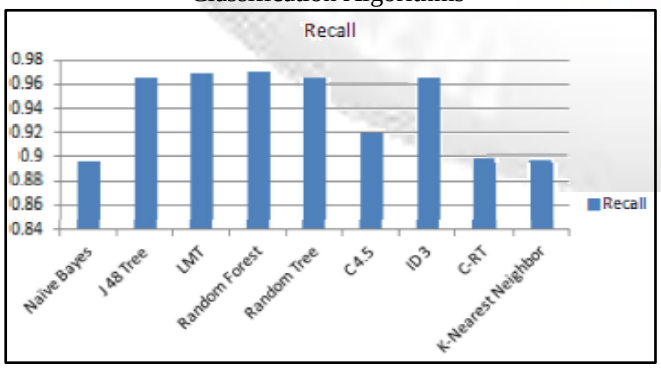

Fig. 3: Graph Comparing the Recall of Different Classification Algorithms.

VI. Conclusion

Internet security is an important aspect in today's world. Due to the increasing security attacks there is a necessity to provide efficient methods which help in staying safe online. In this study, we compare the different classification algorithms and have concluded that tree-based classifiers are best suitable for the task of phishing URL classification.

References

[1] Pradeep Singh, Niti Jain, AmbarMaini, "Investigating the Effect Of Feature Selection and Dimensionality Reduction On Phishing Website Classification Problem" in *2015 1st International Conference on Next Generation Computing Technologies.*

[2] Landwehr, N.; Hall, M.; Frank, E. (2005). *"Logistic Model Trees"* (PDF). Machine Learning.

[3] Ho, Tin Kam (1995). Random Decision Forests (PDF). Proceedings of the *3rd International Conference on Document Analysis and Recognition, Montreal.*

[4] G. K. Gupta, *Introduction to Data Mining with Case Studies,Prentice-Hall Of India Pvt. Limited*, 2011.

[5] Ian H. Witten; Eibe Frank; *Mark A. Hall (2011).* "Data Mining: Practical machine learning tools and techniques, 3rd Edition". Morgan Kaufmann, San Francisco.

[6] M. Aburrous, M. A. Hossain, K. Dahal and F. Thabtah, "Associative classification techniques for predicting e-banking phishing websites," *in MCIT'2010: International Conference on Multimedia Computing and Information Technology*, 2010.

[7] C. Thomas, J. James and L. Sandhya, "Detection of phishing URLs using machine learning techniques," in *InternationalConference on Control Communication and Computing (ICCC)*, 2013.