Machine Learning Approach to Efficient Phishing Detection

Prerna Kapse[1] Radhika Bangar[2] Mayuri Lohar[3] Snehal Kumavat[4] Rajendra Deshmukh[5]
[1,2,3,4,5]Department of Computer Science and Technology
[1,2,3,4,5]SSBT Collage of Engineering and Technology Jalgaon, India

Abstract— Phishing is a kind of cyber-attack in which perpetrators use spoofed emails and fraudulent web-sites to allure unsuspecting online users into giving up personal information. This project views phishing problem holistically by examining various research works and their countermeasures, and how to increase detection. It composes of studies which focus on dataset gathering, pre-processing, features extraction and dataset division in order to make the dataset suitable for the classification process. Phishing creates high rates of damage to internet user. The proposed system introduces a new end-host based anti-phishing algorithm, can also be called as Link Guard, by utilizing the generic characteristics of the hyperlinks in phishing attacks. These characteristics are derived by analyzing the phishing data archive provided by the Anti-Phishing Working Group. Various algorithm used for detection and defence are studied such as Naive Bayes, SVM and C0.5. The selection of the best algorithm to be implemented is based on the precision and accuracy of the algorithm in detecting the spoofed email/website effectively and protecting users privacy and liability. The propose system also discusses the effective use of Machine Learning approach in detecting the Phishing websites/ Mails that overcomes drawbacks of previously proposed And till Nonimplemented system.

Keywords: Exploit, Phishing Corpus, Legitimate Website

I. INTRODUCTION

Phishing is one of the different (and lucrative) types of fraud committed today In Criminal fraud I Defined as a deliberate deception made for the sole aim of personal gains or for smearing an individual's image. In General terms, fraud can be defined as An act of Deceiving people into revealing their personal information, basically for the purpose of financial or personal gains Phishing is an act that attempts to electronically obtaind elicate or confidential information from users (usually for. the purpose of theft) by creating a replica website of a legitimate organization. Phishing is usually perpetrated with the aid of an electronic device and a computer network they target the weakness Existing in various detection system caused by end user phishing attacker usually perpetrate their evil by Communication well composed message to user in order to persuade them revel their personal information Which will by fraudster to gain unauthorized access to the user account For example, a fraudulent email sent to a user might contain a malware (called man in the browser (MITB)), this malwa could be in form of web browser ActiveX components, plugins, or email attachments; if this user ignorantly download this attachment to his pc, the malware will install itself on the user's pc and would in turn transfer money to the fraudster's bank account whenever the user (i.e., the legitimate owner of the bank account) tries to perform an online transaction Fraudulent activities is on the increase daily; Individuals and companies who have been victims in the past now seek for ways to secure themselves from been attacked again. To achieve this, their defense mechanism has to be more secured to prevent them from falling prey again, which implies that the existing defense system (its designs and technology are be greatly improved Behdad et al. pointed out that improving th defense system is not enough stop fraudsters as some of them could still penetrate; the system should also be able to identify fraudulent activities and prevent them from occurring

II. RELATED WORK

Many researchers have analyzed the statistics of suspicious URLs in some way. Our approach borrows important ideas from previous studies. We review the previous work in the phishing site detection using URL features that motivated our own approach. The work by Garera et al. uses logistic regression over hand-selected features to classify phishing URLs. The features include the presence of red flag keywords in the URL, features based on Google's Page Rank, and Google's Web page quality guidelines. It is difficult to make a direct comparison with our approach without access to the same URLs and features. Another study has (Islam and Abawajy, 2013) proposed a multitier classification model for phishing email filtering based on a weighting of message content and message header and selects the features according to the priority ranking. The result from the experiments shows that the proposed algorithm reduced the false positive problems substantially with lower complexity. The features of URL such as transport layer security, unavailability of the top level domain in the URL and keyword withinthe path portion of the URL has been utilized in another study (Jeeva and Rajsingh, 2016). In addition to this, several slashes in the URL, dot in the host portion of the URL and the length of the URL are also the key factors for phishing URL. Then, they generated the rule using association rule mining. The result from the experiment shows that the apriori algorithm detected Approximate 93.00% of The phishing URL. A phishing webpage detection

III. FEATURE USED

There exist a number of different structural features that allow for the detection of phishing emails. In our approach, we make use of sixteen relevant features.
The features used in our approach are described below.

A. HTML Email:

HTML-formatted emails are mainly used for phishing attacks, because plaintext emails do not provide for the scale of tricks afforded with HTML-formatted emails. Hyperlinks are active and clickable only in html formatted emails. Thus, a HTML-formatted email is flagged and is used as a binary feature.

B. IP-based URL:

One way to obscure a server's identity is achieved through the use of an IP address. Use of an IP address makes it difficult for users to know exactly where they are being

directed to when they click the link. A legitimate website usually has a domain name for its identification. Phishers usually use some zombie systems to host phishing sites. When a link in an email contains a link whose host is an IP address (for example, http://81.215.214.238/pp/) we flag the email and is used binary feature.

C. Age of Domain Name:

The domain names (if any) used by fraudsters are usually used for a limited time frame to avoid being caught. We can thus use this feature to flag emails as phishing based on the fact that the domain is newly registered and set a criteria of being new if it is less than 30 days old. This can be achieved by performing a WHOIS query on the domain name in the link. A WHOIS query provides other information such as the name or person to which the domain is registered to, address, domain's creation and expiration dates etc. This feature is a binary.

D. Number of Domains:

We make use of the domain names in the links that we extract and do a count of the number of domains. Two or more domain names are used in an URL address to forward address from one domain to the other. http://www.google.com/url?sa=t&ct=res&cd=3&url=http%3A%2F%2Fwww.antiphishing.org%2F&ei=0qHRbWHK4z6oQLTmBM&usg=uIZX_3aJvESkMveh4uItI5DDUzM=&sig2=AVrQFpFvihFnLjpnGHVs xQ for instance has two domain names where google.com forwards the click to URL antihphishing.org domain name. The number of domains we count is considered a Continuous feature.

E. Number of Sub-domains:

Fraudsters make use of sub domains to make the links look legitimate. Having sub domains means having an inordinately large number of dots in the URL. We can make use of this feature to flag emails as phishing emails. For instance, https://login.personal.wamu.com/verification.asp?d= has 2 sub domains. This is a continuous feature.

F. Presence of JavaScript:

JavaScript is usually employed in phishing emails, because it allows for deception on the client side using scripts to hide information or activate changes in the browser. Whenever an email contains the string "JavaScript", we flag it as a phishing email and use it as a binary feature.

G. Presence of Form Tag:

HTML forms are one of the techniques used to gather information from users. An example below shows the use of form tag in an email. An email supposedly from Paypal may contain a form tag which has the action attribute actually sending the information to http://www.paypalsite.com/ profile.php and not tohttp://www.paypal.com. The email used for collecting user's info has form tag <FORMaction=http://www.paypalsite.com/profile.php method=post>

H. Number of Links:

Most often phishing emails will exploit the use of links for redirection. The number of links in email is used as a feature. A link in an email is one that makes use of the "href" attribute of the anchor tag. This feature will be continuous.

I. URL Based Image Source:

To make the phishing emails look authentic, images and banner of real companies are used in the emails. Such images are usually linked from the real companies' web pages. Thus, if any of the emails make use of such URL based images we flag it as a phishing email.

J. Matching Domains (From & Body):

We make use of the information from the header of the email and match it with the domains in the body of the email. Most phishing emails will have different domains in the header and the body part. We will thus flag emails that have mismatching domain information. For example: The 'From' information in the header part of the email will show the email originating from "someone@paypal-site.com", while the body will have actual ("http://www.paypal.com") company's domain for an authentic look. This feature is binary.

K. Keywords:

Phishing emails contain number of frequently repeated keywords such as suspend, verify, username, etc. We use word frequency (Count of keyword divided by total number of words in an email) of a handful of most commonly used keywords by phishers. This feature is continuous. Some handful of keywords if present in emails are counted and normalized. Group of words with similar meaning or synonyms are used as a single feature. We use six groups of keywords as six separate features.

IV. DATASET

To implement and test our approach, we have used one publicly available datasets i.e.,the data set of url for project as legitimate emails and the url from PhishingCorpus as phishing emails (Phishing 2006, Spam 2006). The total number of emails used in our approach is 50. Out of which 35 are used as phishing emails and 15 as legitimate (ham) emails. The entire dataset is divided into two parts for testing and training purpose. Total website used for traning and testing dataset contain 11,000 format url type .In that data set 9000 pattern used for traning and next 9000 url pattern used for testing. According that provide accurate result by applying various algorithm

1) Random forest algorithm
2) SVM
3) Desigion tree algorithm
4) Knn algorithm
5) NaiveBayes

Feature	Example Link (Source: PhishTank Archive)
Redirect Using //	http://www.tasteofthewest.co.uk/images//wsecure/ap5c/
Extremely Long URL	https://docs.google.com/a/valpo.edu/forms/d/17zrMsBmbTzz4tvu3VqcXM3huxNwnxfeyuU0Bc9iTKZc/viewform?usp=send_form
@ Symbol in the URL	http://imessage-audits.org/profile/?email=abuse@example.com

HTTPS (Hypertext Transfer Protocol with Secure Sockets Layer)	https://accounts.google.com/ServiceLogin?continue=https://drive.google.com/st/auth/host/0Bz9pzRUAjfXaT3RXengxQXV3dlU/
- Separator	http://irstax.wap-ka.com/index.xhtml
Sub/Multisub Domains	http://www.grandimperial.com.my/v2/en/
Nonstandard Port	http://www.belcotech.com:32000/mail/wait.html
IP Address in the URL	http://194.78.154.195/CFIDE/services/labanquepostale.html
HTTPS within URL	http://www.roma.md/templates/system/https:/www2.itau.com.br/atendimento/

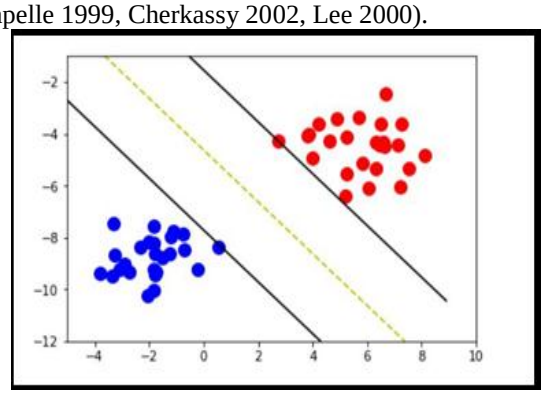
V. ALGORITHM

A. *Random Forest Algorithm*

The random forest algorithm is supervised learning algorithm. From name of algorithm suggest that it should create the no of desig on tree for comparing various outcomes result. Large no tree are to be generated so that accuracy of that algorithm should be large for predict accurate result. Random Forest (RF) is a classifier that consists of many decision trees and outputs the class that is the mode of the classes output by individual trees. In building each decision tree model based on a different random subset of the training dataset, a random subset of the available variables is used to choose how best to partition the dataset at each node. Each decision tree is built to its maximum size, with no pruning performed. The basic idea is similar to Bagging. The main difference between Bagging and RF is that RF uses a random subset of the available variables whereas Bagging uses all available variables. So, RF is suitable for handling a very large number input variables.

B. *SVM*

In any predictive learning task, such as classification, both a model and a parameter estimation method should be selected in order to achieve a high level of performance of the learning machine. Recent approaches allow a wide class of models of varying complexity to be chosen. Then the task of learning amounts to selecting the sought after model of optimal complexity and estimating parameters from training data (Chapelle 1999, Cherkassy 2002, Lee 2000).



Within the SVMs approach, usually parameters to be chosen are (i) the penalty term C which determines the trade-off between the complexity of the decision function. Support Vector Machines (SVM) [10] is also one of the typical machine learning methods for classification and regression. The key idea of SVM is to map data from the input space into a higher dimensional feature space, and to find the optimal separating hyperplane between two classes by maximizing the margin between the classes' closest points.

C. *KMAP*

K-means clustering is an unsupervised non-hierarchal clustering. This attempts to improve the estimate of the mean of each cluster and re-classifies each sample to the cluster with nearest mean. Practical approaches to clustering use an iterative procedure, which converges to one of numerous local points. These iterative techniques are sensitive to initial starting conditions. The refined initial starting condition allows the iterative algorithm to converge to a "better" local point. The procedure is being used in k-means clustering algorithm which being used for both discrete and continuous data points. Let us consider a n example feature vectors x1, x2,..., xn all from the same class, and we know that they fall into k compact clusters, k < n. Let mi be the mean of the vectors in Cluster I. If the clusters are well separated, we can use a minimum-distance classifier to separate them. That is, we can say that x is in Cluster i if || x - mi || is the minimum of all the k distances (Witten 2005). Each cluster then creates a centroid frequency distribution. Each instance is then iteratively reassigned to the cluster with the closest centroid. When instances stop moving between clusters, the iteration process also stops.

VI. RESULT

By using above algorithm use to detect given url is phishing or not compare result by each an every algorithm according that we understand Random forest algorithm provide highly accurate result for given website is phishing or not. And Naïve bayes provide worst result for identify given website if original or fake.

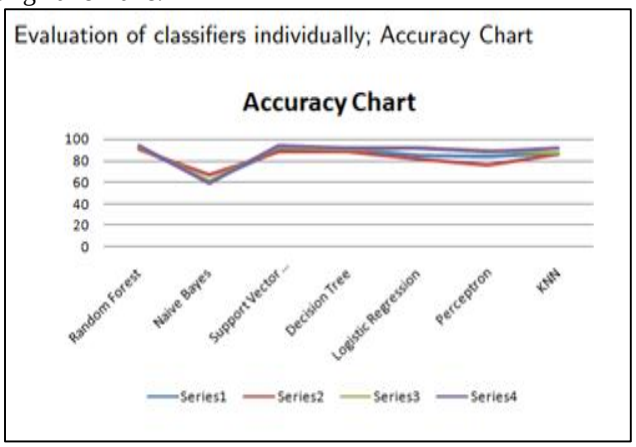


Fig. 1: Accuracy Results: Graphical Representation

According above calculation for performance for each and every algorithm plot an graph .graphically show which algorithm provide best accuracy provide result.

Evaluation of classifiers individually; Accuracy Outcome

Algorithms	Dataset Ranges (Accuracy in %)			
	2000	4000	6000	9000
Random Forest	92.1	90.7	92.5	94.2
Naive Bayes	61.2	67	60.9	58.9
Support Vector Classification	90.2	88.9	92.3	93.7
Decision Tree	90.7	88.6	90	91.4
Logistic Regression	84.9	82	90.9	92.2
Perceptron	84.7	76.4	88.7	88
KNN	87.5	86.4	86.4	91.9

Fig. 2: Accuracy Results

VII. CONCLUSION

The most important way to protect the user From Phishing attack is the education awareness. Internet users must be aware of all security tips which are given by experts. Every user should also be trained not to blindly follow the links websites where they have to enter their sensitive information. It is essential to check the URL before entering the website. In Future System can upgrade to automatic Detect the web page and the compatibility of the Application with the web browser. Additional work also can be done by adding some other characteristics to distinguishing the fake web pages from the legitimate web pages. PhishChecker application also can be upgraded into the web phone application in detecting phishing on the mobile platform.

REFERENCES

[1] Mohammad, R. M.; F. Thabtah; L. McCluskey; "Predicting Phishing Websites Based on Selfstructuring Neural Network," Neural Computing and Applications, vol. 25, iss. 2, 2014
[2] Ammar Almomani, BB Gupta, Samer Atawneh, A Meulenberg, and Eman Almomani. A survey of phishing email filtering techniques. IEEE communications surveys & tutorials, 15(4):2070– 2090, 2013.
[3] Enrico Blanzieri and Anton Bryl. A survey of learning-based techniques of email spam filtering. Artificial Intelligence Review, 29(1):63–92, 2008.
[4] Shriya Se, R Vinayakumar, M Anand Kumar, and KP Soman. Amrita-cen@ sail2015: sentiment analysis in indian languages. In International Conference on Mining Intelligence and Knowledge Exploration, pages 703–710. Springer, 2015
[5] Witten, I., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)
[6] R Vinayakumar, KP Soman, and Prabaharan Poornachandran. Deep encrypted text categorization. In Advances in Computing, Communications and Informatics (ICACCI), 2017 International Conference on, pages 364–370. IEEE, 2017.
[7] R Vinayakumar, KP Soman, and Prabaharan Poornachandran. Evaluating deep learning approaches to characterize and classify malicious urls. Journal of Intelligent & Fuzzy Systems, vol 6 2018.
[8] R Vinayakumar, KP Soman, Prabaharan Poornachandran, and S Sachin Kumar. Evaluating deep learning approaches to characterize and classify the dgas at scale. Journal of Intelligent & Fuzzy Systems, vol 2 2018

