Detecting Phishing-Sites using Hybrid Model

Mrs. Poonam Kumari
CSE Dept, ACSCE,
Bengaluru – 74, India

Apoorva H R Gowda
CSE Dept, ACSCE,
Bengaluru – 74, India

Bhandhavya K
CSE Dept, ACSCE,
Bengaluru – 74, India

Bhavya M U
CSE Dept,
ACSCE
Bengaluru – 74, India

Spurthi M N,
CSE Dept,
ACSCE,
Bengaluru – 74, India

***Abstract*—There is a drastic increase in number of online users due to highly enhanced online technologies. This led to increase in security threats as people now a days are using services from chatting to banking transactions through online. There are many security issues that people are facing from hackers/attackers. There are many types of attacks like keylogger, waterhole attacks, eavesdropping, phishing and many more. The attacker here is called the phisher, where phisher tricks the online users to reveal their confidential/sensitive information like bank account number, password, social network password etc., using phishing websites. There already exists many approaches and techniques to detect and filter out the phishing websites but still researches are going on to find a solution that provides best accuracy. Phishing website has a certain features/patterns through which it can be identified using data mining techniques and the phenomenon called as classification. An hybrid model for classification is presented in this paper to overcome phishing-sites problem. Through this approach we obtain higher accuracy as result.**

Keywords—Cyber threats, Phishing attack, Anti-phishing solutions, Hybrid model, SVM.

I. INTRODUCTION

Till now many people has become victim for phishing attack and lost huge amount of money. Phishers initiates the attack by sending phishing sites and e-mails, where users are asked to enter their confidential information such as bank account number, password, credit card pin, etc. Daily thousands of people are targeted for this kind of attack. Phishing attack is graphically represented in fig.1. Firstly the phisher creates a website which is much similar to a legal website. Then the phisher frequently sends emails to the targeted online users which are embedded with hyperlinks, that directs the users to a fake site when clicked on the link. The users without knowing will fill those details blindly thinking that it is a legal website.

There are many anti-phishing solutions like blacklist, whitelist, heuristic based methods, etc. Blacklist contains a set of malicious/fake URL's. In blacklist-based anti-phishing, web browser sends the URL that is visited by an user to the blacklist to check whether the website is already present in the blacklist. If the URL is present already in the blacklist the web browser notifies the user and warns them not to enter any details to the fake site. The whitelist contains trusted and legal website. This technique is not effective because a new website can be launched within few seconds. The other solution is heuristic based method where phishing attack is detected based on the features, then these features are further used to identify the URL that is requested by an user is whether a legal or a fake site. Web pages are analysed for patterns that are used by phishing websites and feature selection is done.

These are the steps that should be followed to solve the phishing problems:

- The Identity of the desired information.
- Training dataset: It is a set of data that contains input examples and target attributes. The phishing dataset is obtained from Phish Tank, Open Phish and phishing websites are also available on UCI repository.
- Choose a classification algorithm: The most difficult and challenging task is to choose a suitable data mining algorithm as there are many data mining techniques and methods available with their own advantage and disadvantage.

The points to be considered while choosing a classification techniques are as follows:

- Important input attribute.
- The execution of a classifier which is measured through accuracy.
- Normally, there is no single classifier that meets the expectations alone, i.e., in terms of accuracy. This is the reason we are presenting a hybrid. We use supervised learning algorithms as our main aim is to merge multiple weak classification models to classify and detect phishing attack with good accuracy as output.

- Evaluating the performance of classifiers: The final step of the process is to evaluate the performance of classifiers. To check the overall efficiency and performance with respect to data set.

The further sections of the paper is composed of Related work, Proposed model, Experiments, Results, Analysis and finally Conclusion.

II. RELATED WORK

Phisher attacks the user to steal their personal information by creating fake websites that is called phishing attack.[1] Justin et al has proposed an approach where there is a list of fake/malicious URL's classified in blacklist database. When the online user wants to use a certain website and if this particular URL exists in the blacklist database, then this site is predicted as phishing site. This approach has low-false positive rates, but cannot detect malicious/fake sites that doesn't exist in the blacklist database. [2] Jin-Lee et al has proposed a

heuristic-based approach that identifies fake web pages that uses a classifier to recognize malicious web pages from valid ones. [3] The training data used for classifier is web pages. The feature extraction engine is used to process the training data and extract the features for the classifier. Best performance is attained from the classifier by using ML algorithms. This method detects fake sites contrast to other methods that are based on block blacklist.[4] Feng et al. presented a novel neural network to detect phishing sites. The Risk minimization principle is designed that upgrades the generalization ability of the network. Performance is assessed over a UCI repository of the presented network. 1 UCI repository contains 11,055 samples that are named as phishing/legitimate. The dataset contains 30 features for each site classified as Address bar-based, Abnormal-based, HTML/JavaScript-based, and Domain-based features.[5] Routhu Srinivasa Rao, Alwyn Roshan Pai has classified using ML algorithms. The model is tested with and without the third-party-based features to determine the efficacy of third-party services in the classification of doubtful websites. [6] Chunlin Liu, Lidong Wang, Bo Lang, Yuan Zhou attained 99.7% accuracy with a false positive rate of less than 0.4%. We also show that these features render better performance than the previously used features which combine lexical features and structural features and render similar results to the N-Gram or TF-IDF based features.[8] Sudhanshu Gautam, Kritika Rani and Bansidhar Joshi authors have as rule-based classification technique by which we can detect a phishy website and thereby identifying the better detection algorithm which has a higher accuracy detection rate for predicting the phising sites.

III. PROPOSED MODEL FOR PHISHING DETECTION

It became clear that an individual model cannot identify phishing site efficiently and hence a new approach came into existence. An Hybrid based model approach is proposed to resolve the issues that arises due to phishing web sites. An Hybrid based model is obtained by combining multiple models that improves the precision to detect phishing attack. The below diagram is a representation of the steps in the proposed model.

The dataset related to phishing is collected from the UCI repository. UCI repository is an assembly of databases, domain theories that is publicly available for analysis. 30 attributes are sorted out from phishing websites. Dataset is categorized into training and testing dataset. Training and testing dataset are supplied to several classifiers like Random Forest, Decision Tree, Sequential Minimal Optimization, Bayesian net, Naive Bayes, Fuzzy Unordered Rule Induction and Instance based learning to evaluate their accuracy. Firstly classifiers are analysed based on solitary performance, then the classifiers with good results i.e., better precision and less error rate are categorized. Then we fuse these finest classifiers one by one to obtain the Hybrid classification model.

- Dataset: The phishing dataset is collected from UCI repository i.e., publicly available. 11055 instances with 30 attributes are present in the dataset. The attributes selected are :
  - *IP Address with Hexadecimal.*
  - Hide the suspicious Part in long URLs.
  - *"@" Symbol in URL.*
  - Separated by using (-) in the Domain link with as a Prefix or Suffix.
  - *Having_Sub_Domain.*
  - *Domain_registeration_length.*
  - *Using Pop_up_Window.*
  - *URL_of_Anchor.*
  - *Server_Form_Handler.*
  - *Submit_to_Email.*
  - *DNS_Record.*
  - *Age_of_domain.*
  - *Favicon.*
  - *Request URL.*
  - *Website Forwarding.*
  - *Status Bar Customization.*
  - *Disabling Right Click.*
  - *Iframe Redirection.*
  - *Website Traffic.*
  - *PageRank algorithm.*
  - *Google Index.*
- Data Splitting Criteria: The dataset is split into 2:3 ratio, where 2/3 part of dataset is used for training the model, and 1/3 part of dataset is used for testing the model.
- Data Mining Classification Techniques: There are several Data Mining Techniques, methods and tools that exists. Here we use Random Forest, Decision tree, Sequential Minimal Optimization, Naive Bayes, Instance Based Learning, Fuzzy Unordered Rule Induction and Bayesian Net algorithms. These algorithms are used to analyse the relationship between the phishing classification features. The experiments are conducted in Net beans using the Java language with Weka API for performing our hybrid classification model experiment. Further these algorithms are selected based on our literature review.
- Ensemble Methods: fuses various estimators' base predictions. Improves the robustness and generality of estimators. Many effective ensemble methods are available, among them these are the three representative methods Boosting, Bagging, and Stacking
- Hybrid Model: A Hybrid Model is nothing but combination of two or more models. In this phase we combine multiple models for good accuracy. An Hybrid Model eliminates the drawbacks of individual model by combining the best attributes of various models and attains good precision. Bagging and Boosting or combination of both models. Here each model is a combination of learning model model1, to modeln to achieve a complex/composite/hybrid model with enhancement. A pseudo code for the hybrid model proposed by Pedro Domingo's is given below:
- *Performance Evaluation: this phase includes:*
  - MODEL EVALUATION: TO EVALUATE THE CLASSIFICATION MODEL WE USE PRECISION, RECALL, F-MEASURE, ERROR RATE AND ACCURACY. WE USE SPLIT OF DATA SET, OTHER STATISTICAL METHODS AND CONFUSION MATRIX TO EVALUATE THE PERFORMANCE. THE STATISTICAL

EQUATION OF PRECISION, RECALL, F MEASURE, ERROR RATE, ACCURACY AND CONFUSION MATRIX ARE AS FOLLOWS;

- RECALL = TP/(TP+FN). (1)
- PRECISION = TP/(TP+FP). (2)
- F MEASURE = 2 [(PRECISION*RECALL)/[(PRECISION+ RECALL)]] ……..(3)
- ERROR RATE = (FP+FN)/ (TP+TN+FP+FN) .....(4)
- CLASSIFICATION ACCURACY = (TP+TN)/ (TP+TN+FP+FN). (5)
- CONFUSION MATRIX: FOR BINARY CLASSIFICATION A CONFUSION MATRIX IS SHOWN IN TABLE 1

IV. DEMONSTRATION AND STUDY

In this experiment we are to find Accuracy of the training dataset, Accuracy of the testing dataset, Specification, False Positive rate, precision and recall by comparing algorithm using python code. The following Involvement steps shown in flow diagram

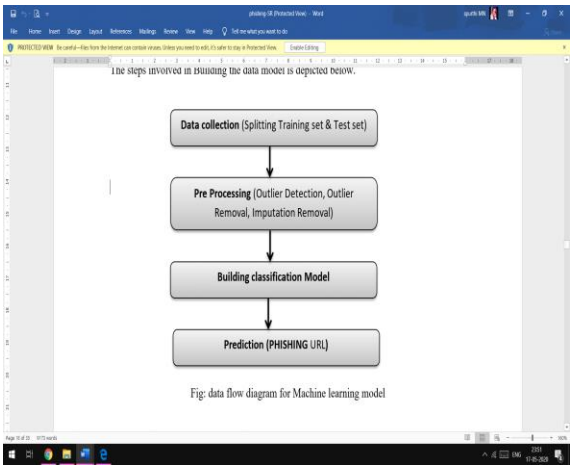

Fig:1. data flow diagram for Machine learning model

Data collection: The data set collected for phishing URL is split into dataset for Training and Testing in the ratio of 7:3.

Pre-Processing: In this stage the data preprocessing which the data collected because it might be contain with missing value noise data and error value are preprocessed and in end the data is reduced to some minimum amount of records. Initially the Attributes.

Data Validation: Importing the library packages with loading given data's analyzing the variable identification by data shape, data type and evaluating the missing values, duplicate values.

Data cleaning/preparing process: Data cleaning / preparing by rename the given dataset and drop the column etc. To analyze the uni-variate, bi-variate and multi-variate process.

Data visualization process:

To visualize the given dataset in the form of graphical representation like pair plot, heat map, bar chart, pi-chart from matplotlib.

Comparing Machine learning algorithms:

Scikit-Learn libraries. In this library package have to done pre-processing, linear model with logistic regression method, cross validating by KF old method, ensemble with random forest method and tree with decision tree classifier. Additionally, splitting the train set and test set. To predicting the result by comparing accuracy.

Prediction result by accuracy:

Logistic regression algorithm predicts a value by utilizing a linear equation with independent predictors. The estimated value may lie in between negative infinity to positive infinity. We need the output of the algorithm to be classified variable data. Higher accuracy predicting result is logistic regression or other model by comparing the best accuracy.

V. ARCHITECTURE

In our project we are collecting the datasets of 10000 containing 30 attributes which are features of URL are pre-processed and validated and analysis made into hybrid model when user is enter the URL. The feature are exacted from the URL entered by user .the dataset are split into test and training data on classifier building model and all algorithm like SVM (support vector system), Forest decision, Logistic regression having their own test and train dataset are classifies the URL given by the user whether legitimate site or illegitimate site. Algorithm's output should be a classified variable data. Higher accuracy predicting result is logistic regression or other model by comparing the best accuracy.

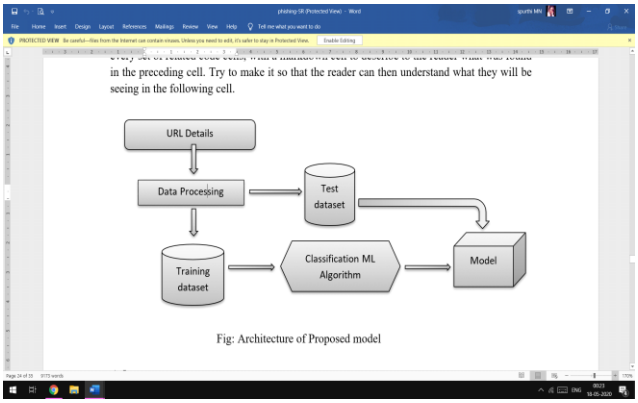

Fig:2. Architecture of Proposed model

VI. CONCLUSION & FUTURE WORK

Phishing attack is the most difficult cyber threats that people are facing. Another challenging phase is to identify the website and classify them as safe/legitimate/legal or fake/Phishy websites. This the reason we conduct our experiment in two different stages. In first stage we individually perform classification techniques, i.e., RF, SMO, J48, BN, NB and IBK model. Then we select the best 3 models based on performance and high accuracy. We then combine weak models and can see the difference in results as we obtain good accuracy than before. In second stage we fuse the models to obtain Hybrid model that overcomes the limitations/drawbacks that we face in individual models.

Later through a deep research and analysis we can wind up that union of classification model as IBK with other classification models individually gives better result when compared to individual classification model in term of enhanced accuracy. The maximum accuracy we have obtained is 97.75% in testing data as combination of BN with IBk model and also equal maximum accuracy is obtained when ensemble of J48 with IBk model and less error rate 0.225 on both hybrid models.

The solutions are focused to enhance the further work to append a feature selection mechanism to decrease the attributes which takes more time to develop a hybrid model. Our future work is to develop an automatic detection of phishy websites with the trials of our Hybrid classification model.

REFERENCES

[1] Hossein Shirazi, Kyle Haefner, Indrakshi Ray Department of Computer Science Colorado State University Fort Collins, USA Email: {shirazi, kyle.haefner , iray}@colostate.edu.

[2] Srushti Patil Department of ComputerEngineering Sardar Patel Institute of Technology Mumbai, India srushti.patil@spit.ac.in, Sudhir Dhage Department of Computer Engineering Sardar Patel Institute of Technology Mumbai, India sudhir_dhage@spit.ac.in.

[3] S. Gautam (✉) · K. Rani · B. Joshi Department of Computer Science and Engineering, Jaypee Institute of Information Technology, Noida, India e-mail: isudhanshugautam@gmail.com K. Rani e-mail: kritika.rani17@gmail.com B. Joshi e-mail: bansidhar.joshi@jiit.ac.in

[4] B B. B. Gupta gupta.brij@gmail.com 1 National Institute of Technology Kurukshetra, Kurukshetra, India.

[5] Chunlin Liu State Key Lab of Software Development Environment School of Computer Science and Engineering, Beihang University Beijing, China jackliu@buaa.edu.cn,Bo Lang State Key Lab of Software Development Environment School of Computer Science andEngineering, Beihang University Beijing, China langbo@buaa.edu.cn.

[6] Soon Fatt Choo, Jeffrey§, Kang Leng Chiew† and San Nah Sze‡Faculty of Computer Science and Information Technology Universiti Malaysia Sarawak94300 Kota Samarahan Sarawak,Malaysia§sfchoo.jeffrey@gmail.com†klchiew@fit.unimas.my‡snsze@fit.unimas.my