Science Q

JOURNAL OF ADVANCEMENT IN
ENGINEERING AND TECHNOLOGY

Journal homepage: http://scienceq.org/Journals/JAET.php

Research Article Open Access

Detection of Phishing Websites Using Hybrid Model

Ch. Chakradhara Rao[1], **A. V. Ramana**[2]

[1] Department of CSE, GMR Institute of Technology, Rajam, Andhra Pradesh, India.

[2] Department of IT, GMR Institute of Technology, Rajam, Andhra Pradesh, India.

*Corresponding author: Ch. Chakradhara Rao, E-mail: chakradhararao.ch@gmrit.org

Received: April 05, 2018, Accepted: May 19, 2018, Published: May 19, 2018.

ABSTRACT

Online technologies have revolutionized the modern computing world. There are number of users who purchase products online and make payment through various websites. There are multiple websites who ask user to provide sensitive data such as username, password or credit card details etc. often for malicious reasons. This type of website is known as phishing website. The phishing website can be detected based on some important characteristics like URL (Uniform Resource Locator) and Domain identity. Several approaches have been proposed for detection of phishing websites by extracting the phishing data sets criteria to classify their legitimacy. However, there is no such approach that can provide better results to the users from phishing attacks. This paper is an attempt to contribute in that area by presenting a hybrid model for classification to detect phishing websites with high accuracy and less error rate.

Keywords: *Anti-Phishing, Recall, F-Measure, Websites Detection*

1. INTRODUCTION

Over the years, there had been several phishing attacks and lots of people have lost massive sums of money via becoming a victim of a phishing attack. In a phishing attack, emails are sent to the person claiming to be a valid company, wherein the e-mail asks the person to enter details like usernames, passwords, social security number, etc. Phishing-site and their mails are sent to hundreds of thousands of persons each day and thus are still a big concern for cyber security. The phishing process starts with setting up counterfeited website by the phisher, which is very much similar to a legitimate website. Phisher frequently sends emails to target users with embedded hyperlinks directing to their fake website. As soon as the receiver clicks on the hyperlink, they are redirected to a bogus website. There it asks users for their confidential information like username, id, password, etc. When the users enter their personal information, phisher steal them and spoof the users.

Phishing is done through several ways, but all of them use a common set of features. Two techniques are used in classifying the phishing-site. The first technique is based on maintaining a blacklist and checking requested URL in that list. This technique is not effective because a new website can be launched within few seconds. Second one is heuristic based approaches which identifies features of phishing websites and then use these features to categorize requested URL as either phishy or legitimate. The efficiency of this technique depends on selecting features that can differentiate between phishy and legitimate website. Feature selection can be accomplished by analyzing the web pages and examining the patterns and properties that are used by phishing websites.

The principle steps that must be followed to resolve the website phishing problems, which are as follows:

Identity of the desired information: Given any sort of problem, we required a group of attributes, which pre-measured or predetermined and desired results of classifier. Consequently, a group of output and input features should be well-known.

Training dataset: training dataset contains input examples and preferred target attributes. Obtaining the phishing dataset, there are several ways, such as PhishTank; additionally dataset of phishing websites is obtainable on UCI repository.

Select the classification algorithm: Choosing the data mining algorithm is a very challenging phase. There are several data mining methods and techniques available in the literature where each method and techniques has its own advantage and disadvantage.

Three main essential points in choosing classification techniques are:

- The required data input features
- The performance of classifier measured through the accuracy rate
- The output results understandable.

Generally, there is no classifier separately meets expectations best with respect to all provided information and classifier efficiency and accuracy mostly depends on the training dataset features. For this reason, we present a hybrid classification model to categories phishing-sites using supervised learning algorithms. Our approach is to combine multiple weak classification models to classify and detect phishing sites attacks with improved accuracy and the understanding of output results.

Performance assessment of classifiers: The final step is to check the determined classifier overall efficiency and performance evaluation with respect to test data.

II. Proposed Algorithm

Existing System:

Decision Tree, IBK, Naïve Bayes and Bayes Net Algorithms are individually used for phishing websites. Fuzzy classification strategies are also used for detecting the phishing websites.

The existing approaches for anti-phishing are:

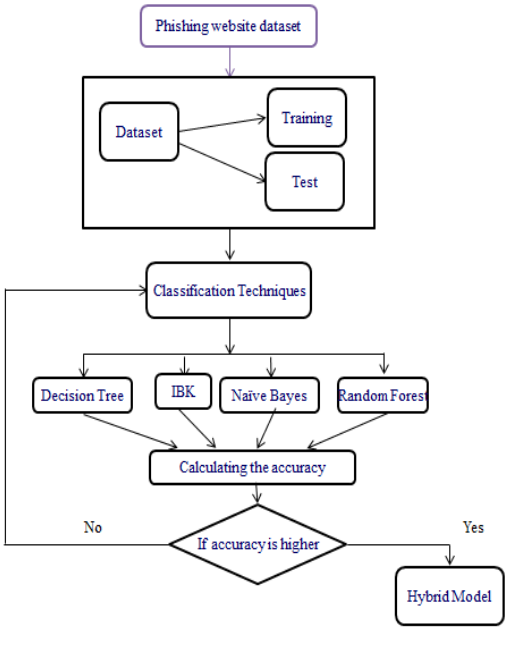
Detect and block the phishing Web sites manually in time, Enhance the security of the web sites at the time of developing, Block the phishing e-mails by various spam filter soft wares, Installing online anti-phishing software in user's computers.

Problem Statement:

Phishing is a technique used to steel personal information for the purposes of identity theft and using fake e-mail messages that appear to come from legitimate businesses. This is usually done by sending emails that seem to come from reliable source to gain access to person's confidential and private information. Phishing emails considers as the fastest rising online crime method used for stealing personal financial data and perpetrating identity theft. Individuals who respond to phishing e-mails, and input the requested financial or personal information into emails, websites, or pop-up windows put themselves and their institutions at risk. So, there is a need to keep on enhancing the accuracy of the detection techniques. Overall the problems carried out in this research are as following:

- How to determine the best set of features to be used with phishing detection.
- How to select the best classification algorithm to be used for phishing detection.
- How to enhance the performance of the best selected features and classifiers.
- How to integrate multiple classification algorithms for phishing detection and to evaluate such integration.

Proposed System:



Model Architecture

A Hybrid Model based approach has been proposed target to solve the phishing web sites problem. A single model cannot efficiently detect the phishing websites because there is needed to enhance and tuning the single model or second approach is to combine any two or three models for improving the accuracy for detecting the phishing-sites attack. We pursued to perform our experiments. 30 features are selected from phishing website, dataset which is publicly available on UCI repository. Dataset as training and testing are provided to various classifiers like Random Forest (RF), Decision Tree (J48), Naive Bayes (NB), Instance based learning (IBk) to evaluate their accuracy. Moreover, we firstly check the individual performance of a classifier and get the best classifier in term of high accuracy and less error rate. On the basis of best classifier we then combine the best classifier model with other classifiers, one by one and finally get a better hybrid classification model.

A. Dataset

In this, we have used a data from UCI repository that is publicly available for use. Dataset consists of 11055 instances and 30 attributes.

B. Data Splitting Criteria

In this, the processed data were categorized into training and testing categories for their respective purposes. The splitting criteria on data set are 2:3. 2/3 we used for training and 1/3 used for testing.

C. Data Mining Classification Techniques

In this, we used several data mining existing classification methods and techniques including Random Forest, Decision tree, Naive Bayes, Instance Based Learning algorithms. Further these algorithms are used to analyze and to examine the relationships of the different phishing classification features.

D. Hybrid Model

In this phase, multiple models combine to perform better performance accuracy. However a hybrid model combines the best features of two or more models, eliminating the drawbacks of individual model for achieving best accuracy by Bagging and boosting or combination of models. Ensemble methods are techniques that create multiple models and then combine them to produce improved results. These methods usually produce more accurate solutions than a single model would. The goal of ensemble methods is to combine the predictions of several base estimators built with a given learning algorithm in order to improve generalizability / robustness over a single estimator. There are several families of ensemble methods available; in this, we are using the voting classifier ensemble method. Voting is one of the simplest ways of combining the predictions from multiple machine learning algorithms. It works by first creating two or more standalone models from your training dataset. A Voting Classifier can then be used to wrap your models and average the predictions of the sub-models when asked to make predictions for new data. The EnsembleVoteClassifier is a meta-classifier for combining similar or conceptually different machine learning classifiers for classification via majority or plurality voting. The EnsembleVoteClassifier implements "hard" and "soft" voting. In hard voting, we predict the final class label as the class label that has been predicted most frequently by the classification models. In soft voting, we predict the class labels by averaging the class-probabilities. Each model is a combination of learning model model1, model2, . modeln etc.. modeln aiming to create a composite model with certain improvements. Both models are usable for categorization.

E. Performance Evaluation

Evaluation of the Model: Using Precision, Recall, F-measure, Error rate and Accuracy to evaluate the classification model. This can be performed by using split of data set, other statistical Methods and confusion matrix. The statistical equation of Precision, Recall, Fmeasure, Error rate, Accuracy and confusion matrix are as follows;

Recall: Also known as sensitivity, is the fraction of relevant instances that have been retrieved over the total amount of relevant instances.

$$\text{Recall} = TP/ (TP+FN)$$

Precision: Also called positive predictive value, is the fraction of relevant instances among the retrieved instances.

Precision = TP/ (TP+FP)

F-Measure: It is the weighted harmonic mean of the precision and recall of the test.

FMeasure=2[(Precision*Recall)/[(Precision+ Recall)]

Error Rate: It reflects the rate of errors made by predictive model. It is one minus the accuracy.

Error Rate = (FP+FN)/ (TP+TN+FP+FN)

Classification Accuracy: It reflects the number of times that the model is correct when applied to data.

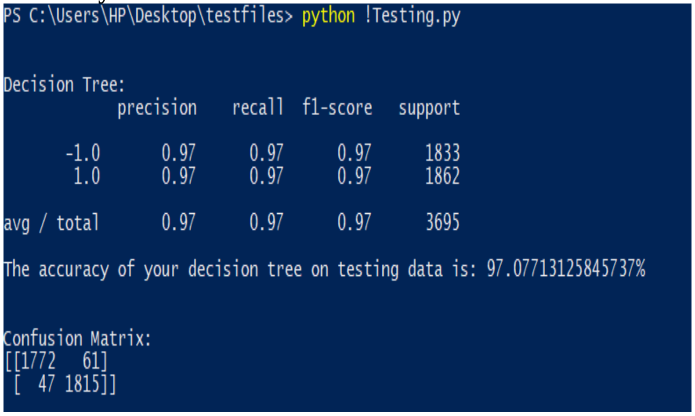
Classification accuracy = (TP+TN)/ (TP+TN+FP+FN)

Confusion Matrix: It sorts all cases from the model into categories, by determining whether the predicted value matched the actual value. All the cases in each category are then counted, and the totals are displayed in the matrix. The classification matrix is a standard tool for evaluation of statistical models and is sometimes referred to as a confusion matrix. It compares actual to predicted values for each predicted state that you specify. The rows in the matrix represent the predicted values for the model, whereas the columns represent the actual values. The categories used in analysis are false positive, true positive, false negative, and true negative.

Actual vs. Predicted	Positive(P)	Negative(N)
Positive(P)	TP	FN
Negative(N)	FP	TN

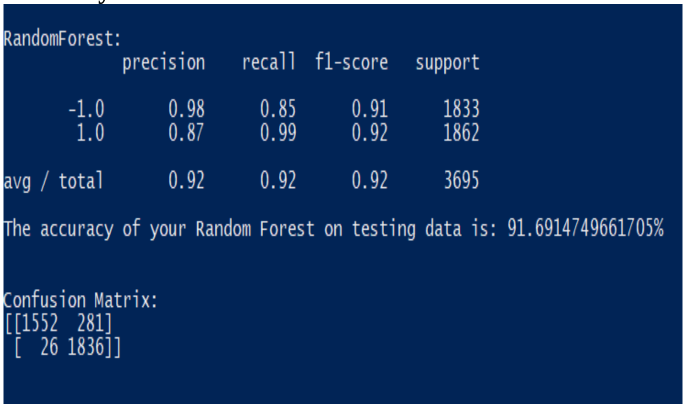
III. Experiment and Result

Accuracy measure for Decision Tree:



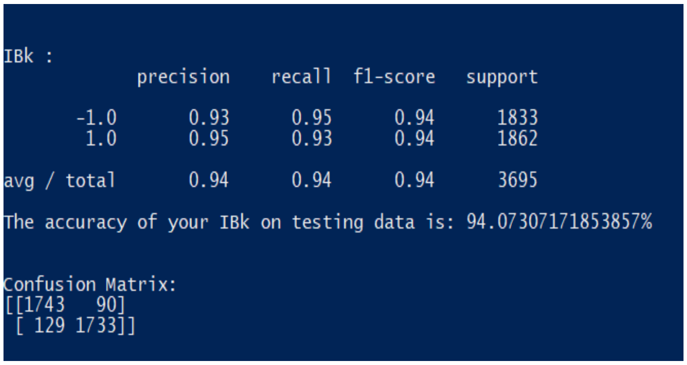
```
PS C:\Users\HP\Desktop\testfiles> python !Testing.py


Decision Tree:
             precision    recall  f1-score   support

       -1.0       0.97      0.97      0.97      1833
        1.0       0.97      0.97      0.97      1862

avg / total       0.97      0.97      0.97      3695

The accuracy of your decision tree on testing data is: 97.07713125845737%


Confusion Matrix:
[[1772   61]
 [  47 1815]]
```

Accuracy measure for RandomForest:



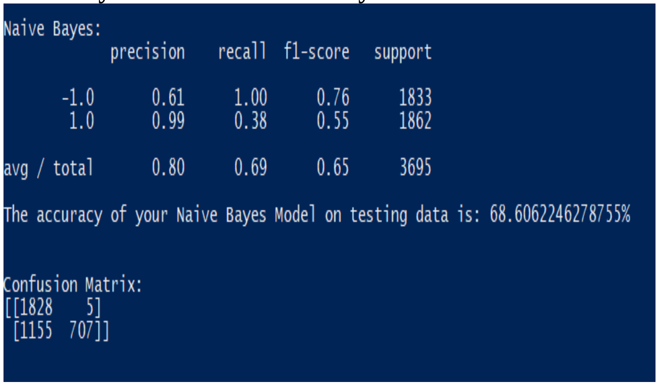
```
RandomForest:
             precision    recall  f1-score   support

       -1.0       0.98      0.85      0.91      1833
        1.0       0.87      0.99      0.92      1862

avg / total       0.92      0.92      0.92      3695

The accuracy of your Random Forest on testing data is: 91.6914749661705%


Confusion Matrix:
[[1552  281]
 [  26 1836]]
```

Accuracy measure for IBK:



```
IBk :
             precision    recall  f1-score   support

       -1.0       0.93      0.95      0.94      1833
        1.0       0.95      0.93      0.94      1862

avg / total       0.94      0.94      0.94      3695

The accuracy of your IBk on testing data is: 94.07307171853857%


Confusion Matrix:
[[1743   90]
 [ 129 1733]]
```

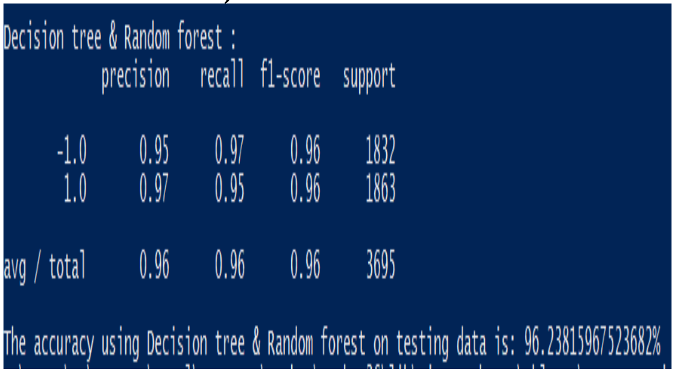
Accuracy measure for NaiveBayes:



```
Naive Bayes:
             precision    recall  f1-score   support

       -1.0       0.61      1.00      0.76      1833
        1.0       0.99      0.38      0.55      1862

avg / total       0.80      0.69      0.65      3695

The accuracy of your Naive Bayes Model on testing data is: 68.6062246278755%


Confusion Matrix:
[[1828    5]
 [1155  707]]
```

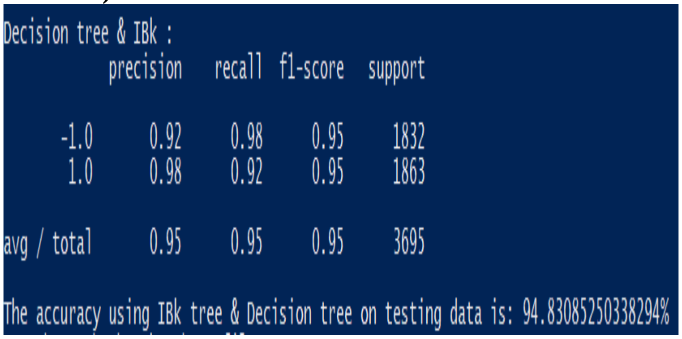
Since, the accuracy of NaiveBayes algorithm is very low when compared to other three algorithms, we exclude that in the combining process. Now the accuracy for combined models are calculated:

Accuracy measure for the combined model (DecisionTree and RandomForest):



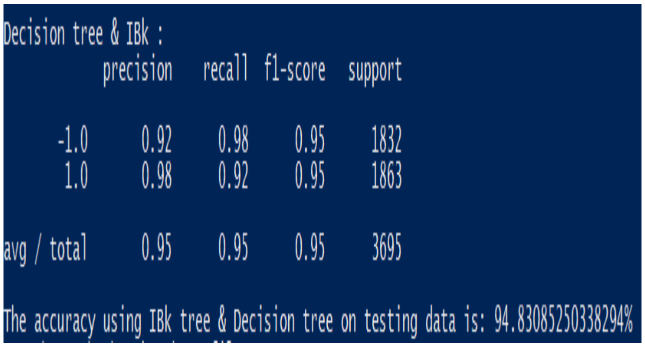
```
Decision tree & Random forest :
             precision    recall  f1-score   support


       -1.0       0.95      0.97      0.96      1832
        1.0       0.97      0.95      0.96      1863


avg / total       0.96      0.96      0.96      3695


The accuracy using Decision tree & Random forest on testing data is: 96.23815967523682%
```

Accuracy measure for the combined model (RandomForest and IBK):



```
Decision tree & IBk :
             precision    recall  f1-score   support


       -1.0       0.92      0.98      0.95      1832
        1.0       0.98      0.92      0.95      1863


avg / total       0.95      0.95      0.95      3695


The accuracy using IBk tree & Decision tree on testing data is: 94.83085250338294%
```

Accuracy measure for the combined model (DecisionTree and IBK):



```
Decision tree & IBk :
             precision    recall  f1-score   support

       -1.0       0.92      0.98      0.95      1832
        1.0       0.98      0.92      0.95      1863

avg / total       0.95      0.95      0.95      3695

The accuracy using IBk tree & Decision tree on testing data is: 94.83085250338294%
```

IV.CONCLUSION

Phishing website is one of the worldwide challenging security problems since last few decades, detection of web sites as a legitimate and phishy as one of the challenging aspects. For this reason we carry out experiments in two phases. In phase 1 we individually perform classification techniques, i.e., RF, J48, NB and IBk model and select the best 3 models on criteria of performance and high accuracy. We can further combine with the weak model as a result we can see that combine models can increase the accuracy. So, In phase II, we further combine each model with our best 3 individual models to make a hybrid model.

REFERENCES

1. G. AARON "Phishing Activity Trends Report, 4th Quarter 2015",http://docs.apwg.org/reports/apwg_trends_report_q4_2015.pdf, Anti-Phishing Working Group, 2016.
2. Neil Chou, Robert Ledesma, Yuka Teraguchi, Dan Boneh, John C. Mitchell, Stanford Ca, "Client-side defense against web-based identity theft". In Proceedings of the 11th Annual Network and Distributed System Security Symposium (NDSS), 2004.
3. Colin Whittaker, Brian Ryner, Marria Nazif, "Large-scale automatic classification of phishing pages". In Proceedings of the 17th Annual Network and Distributed Security Symposium (NDSS), 2010.
4. L. Cranor, S. Egelman, J. Hong, Y. Zhang, "Phinding phish: An evaluation of anti-phishing toolbars". In Proceedings of the 14th Annual Network and Distributed System Security Symposium (NDSS), 2007.
5. Steve Sheng, Brad Wardman, Gary Warner, Lorrie Faith Cranor, Jason Hong, "An Empirical Analysis of Phishing Blacklists". In Proceedings of the Sixth Conference on Email and AntiSpam (CEAS), July 1617, 2009.
6. Sujata Garera, Niels Provos, Monica Chew, Aviel D. Rubin, "A Framework for Detection and Measurement of Phishing Attacks". In Proceedings of 2007 ACM Workshop On Recurring Malcode, pp. 1-8, 2007.
7. A. Blum, B. Wardman, T. Solorio, G. Warner, "Lexical feature based phishing URL detection using online learning". In Proceedings of the 3rd ACM Workshop on Artificial Intelligence and Security, pp. 54-60, 2010.
8. J. Ma, L.K. Saul, S. Savage, G.M. Voelker, "Learning to detect malicious URLs", ACM Transactions on Intelligent Systems and Technology (TIST), 2, pp.30, 2011.
9. Justin Ma, Lawrence K. Saul, Stefan Savage, Geoffrey M. Voelker, "Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious URLs". In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1245-1254, 2009.
10. D.K. McGrath, M. Gupta, "Behind Phishing: An Examination of Phisher Modi Operandi". In Proceedings of Proceedings of the 1st Workshop on Large-Scale Exploits and Emergent Threats, pp. 1-8, 2008.

Citation: Ch.Chakradhara Rao *et al.* (2018). Detection of Phishing Websites Using Hybrid Model, J. of Advancement in Engineering and Technology, V6I3.06. DOI: 10.5281/zenodo.1250270.