ISSN No:-2456-2165

Detection of Phishing Websites Using PSO and Machine Learning Frameworks

Kotturu Riteesh, Yarramaneni Maruthi Chowdary, Gudiseva Naga Sai Teja, Dr. A. Srisaila (Assistant professor)
Information Technology
Velagapudi Ramakrishna Siddhartha Engineering College
Vijayawada, Andhra Pradesh, India

Abstract— **In the last few years, the web phishing attacks have been constantly evolving, causing the customers to lose their trust in e-commerce, online services, trading platforms and new contacts. Various types of tools and systems based on a blacklist of phishing websites are applied to detect the phishing websites. Unfortunately, the daily continuous evolution of technology has led to the birth of more sophisticated methods when building websites to attract potential users. There is an increase in recent research studies they have been adopting machine learning techniques to identify phishing websites and utilizing them for early alarm methods to identify such threats. Phishing website detection is proposed using particle swarm optimization-based feature weighting is proposed to enhance the detection of phishing websites. The proposed approach of our work suggests the utilization of particle swarm optimization (PSO) to analyze various records of website features effectively to achieve higher accuracy while detecting phishing websites. The experimental results indicated that the proposed PSO-based feature analysis which achieved an outstanding improvement in the terms of classifying the accuracy and determining the best algorithm approach.**

Keywords - Random Forest, Decision Tree, Internet Identity, SVM, Machine learning, Logistic regression, Multilayer perceptron.

I. INTRODUCTION

In this today internet world, most of the people are communicating with each other either by a computer or by a digital device that has an active internet connection. The number of people who are using e-banking, online shopping and other online services has been increasing due to the availability of convenience, comfort, and assistance. Any hacker takes this as an opportunity to steal money and steal private information that are needed to access our bank accounts, emails and other passwords. The primary usage of phishing websites is to find a way to steal sensitive information from the users by hackers. It is carried out with a mimicked page of a legitimate site, directing online user into providing sensitive information. The term phishing is derived from the concept of 'fishing' of victims' personal details and other confidential details like passwords, banking details and private files which are known to the user only. The one who is attacking a target first sends a bait as mimicked webpage and waits for the outcome of sensitive information.

II. LITRATURE SURVEY

Jabri, Riad & Ibrahim, Boran published a work on Phishing Websites Detection Using Data Mining Classification Model by using PSO techniques. They used data mining for their research on phishing websites detection algorithms and code. Many of the phishing attackers use data mining in their approach of creating phishing websites. However, there may be many differences in their accuracies and their error rate. They have done their work using a dataset consist of tens of different webpage features and hundreds of instances. The experimental results they got showed them that the proposed algorithm they developed worked better and with an improvement in its accuracy value.

P. K. Sahoo had published a work by using Data mining to solve Phishing Attacks. As the Detection of phishing attack is a difficult process to detect it with high accuracy, it has become a challenging research issue. Generally Phishing sites are normally being detected by using an old approach by utilizing blacklist-based approach but this approach is unable to meet the demands and requests as the white listed phishing sites cannot be traced using this approach. The work done by them uses Data mining algorithms to study and analyze E-mails and the algorithms also helps in prevention of phishing attacks. He proposed an entirely unique architectural type model to help in differentiation of fake E-mails and real E-mails by finding a high accuracy value and use naive Bayesian classification in the algorithm. The algorithm proposed by him works in various levels in fake E-mail detection and thus try to protect the users information and confidential details from leaking from their devices.

W. Niu, X. Zhang, G. Yang, Z. Ma and Z. Zhuo proposed Phishing Emails Detection Using CS-SVM. Phishing attacks have become common nowadays, as they have resulted in an increase of financial losses in the past few years. Machine learning-based detection methods, of them particularly Support Vector Machine (SVM), which have been proved to be more effective when compared to all other machine learning methods. They have performed experiments on a self-gathered dataset consisting of 1,384 types of phishing emails and 20,071 types of non-phishing emails.

Giovane C. M. Moura & Aiko Pras had proposed work on Scalable Detection and Isolation of Phishing. The main ideas are to move the protection from end users towards the network provider and to employ the novel bad neighborhood concept. In addition, they proposed a development of a self-management architecture which uses a ISPs to protect their

users against phishing attacks, and they explained how evaluation of the architecture could be done.

André Bergholz , Gerhard Paaß , Frank Reichartz , Siehyun Strobel , Schloß Birlinghoven had worked on an Improved phishing detection using a model-based features and algorithms. As Filtering approaches are no longer working by employing blacklists checking as new phishing methods are being created. They investigated the phishing emails, where the classifier is being trained on the characteristic features of existing emails and subsequently it is to easily identify new phishing emails with different type of contents.

III. PROPOSED SYSTEMS

An attack on our devices by hacking people takes place through various types of phishing forms such as electronic mails, websites and malware. To perform an email phishing, fake emails are created by the hackers which will look as a one arrived from a trusted company.

The Phishing is a type of a unique process in which some people try to obtain or steal your personal information, such as your web passwords, credit card numbers, bank account details and other secrete information which you deal them as most important. Our proposed method is to differentiate a phished website with an original one and find which algorithm is best suited for finding them.

Given below flow diagram is the step-by-step representation of the process. Which shows how the process is done in steps.

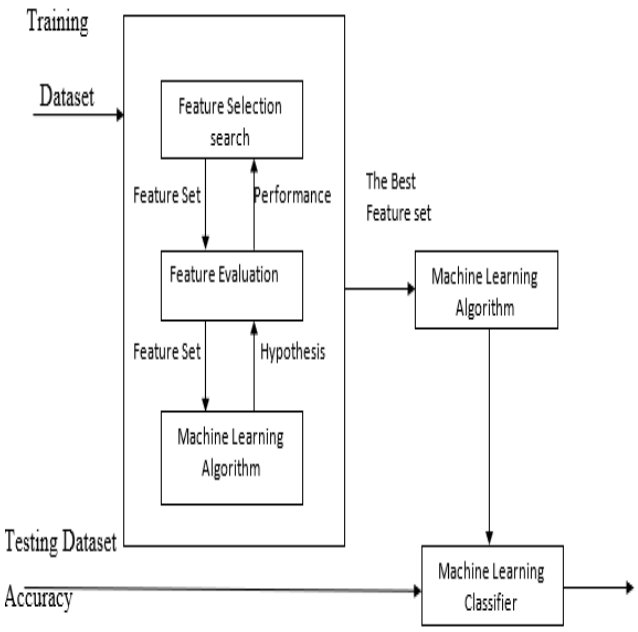

Fig. 1: Architecture Flow Daigram

IV. METHODOLOGY

A. *Data set Collection*

The data set we used for the application and algorithms is called a phishing websites dataset. The data set is made up of the collection of numerous phishing website details and also the details of the original websites. There are as many as "11,056" records in this phishing website dataset.

B. *Data Splitting*

The data set is separated into two parts for our application usage and algorithm functioning. The two separated parts of the dataset are named as a test dataset and a training dataset. The test data set to check the compatibility of algorithm and the training dataset for the getting of problem results.

C. *Data pre-processing*

The phishing dataset would undergo data pre-processing, the dataset would undergo data formatting, data cleaning, data sampling. The dataset will be reformed and the unnecessary data unfit for the algorithm will be filtered and compatible dataset is generated.

D. *Decision Tree*

We used decision tree algorithm for the running of phishing websites dataset to find the accuracy value for the generated training dataset. The accuracy value generated is compared with remaining algorithms accuracy to find which algorithm is best.

E. *Logistic Regression*

We used logistic regression algorithm for the running of phishing websites dataset to find the accuracy value for the generated training dataset. The accuracy value generated is compared with remaining algorithms accuracy to find which algorithm is best.

F. *Random Fores*

We used random forest algorithm for the running of phishing websites dataset to find the accuracy value for the generated training dataset. The accuracy value generated is compared with remaining algorithms accuracy to find which algorithm is best.

G. *Mean Square Error*

We have used mean square error a function to help in the calculation of accuracy value for all the algorithms we applied.

H. *Mean Average Error*

We have used mean average error a function to help in the calculation of accuracy value for all the algorithms we applied.

I. *R-Squared*

We use R-squared to help in the calculation of accuracy value for the algorithms we applied.

J. *RMSE*

We use RMSE values to help in the calculation of accuracy value for the algorithms we applied.

ISSN No:-2456-2165

V. OBSERVATIONS AND RESULTS

	DECISION TREE	LOGISTIC REGRESSION	RANDOM FOREST
MSE	0.602026	0.523878	0.448625
MAE	0.301013	0.261939	0.224312
R-Squared	0.395633	0.470998	0.543893
RMSE	0.775903	0.723794	0.669794
ACCURACY	0.849493	0.869030	0.887843

Table 1

The above is one of the many entries of collected data of three algorithm we used, we selected a entry amongst them. The table consists of accuracy value for a trainning data set for the all three algorithms.

The graph below is the compilation of our algorithms results into a single graph. From the result graph we can see that random forest provides the best accuracy value when compared to other algorithms. The results used for comparison have been gathered randomly from multiple results of each algorithms.

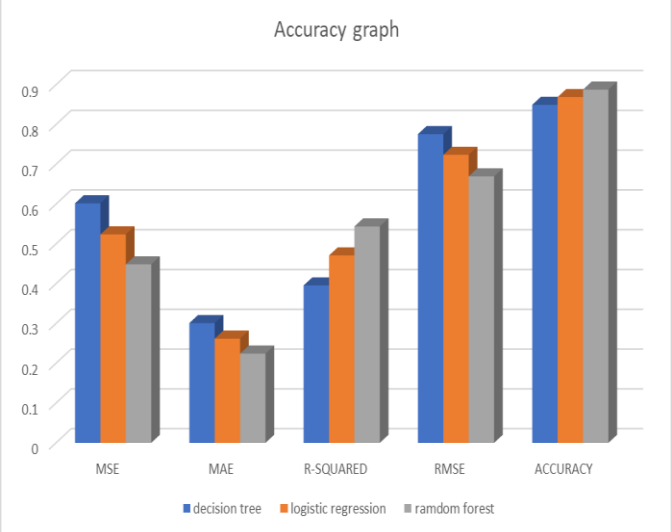

Fig. 2: Final Accuracy Comparision graph

VI. CONCLUSION

Phishing is one of the many cyber-crime procedures that utilizes both social building and specialized deception to take individual sensitive data. Besides, Phishing of the systems is considered as one type of extensive frauds. The proposed system implements a type of unique methodology on phishing website detection based on the PSO-based feature weighting was suggested for the work. In the proposed PSO-based feature weighting. The website features were weighted help to enhance the detection of phishing websites. We implemented three different types of machine learning algorithms and found will find out the best one among them.

REFERENCES

[1.] S. Nawafleh, W. Hadi (2012). Multi-class associative classification to predicting phishing websites. International Journal of Academic Research Part A; 2012;4(6), 302-306J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2.] Sadeh N, Tomasic A, Fette I. Learning to detect phishing emails. Proceedings of the 16th international conference on World Wide Web. 2007: p. 649-656.

[3.] Andr Bergholz, Gerhard Paa, Frank Reichartz, Siehyun Strobel, and Schlo Birlinghoven. Improved phishing detection usingmodel-based features. In Fifth Conference on Email and Anti-Spam, CEAS, 2008

[4.] P. Tiwari, R. Singh International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181Vol. 4 Issue 12, December-2015.

[5.] UCI Machine Learning Repository." http://archive.ics.uci.edu/ml/, 2012.

[6.] H. A. Chipman, E. I. George, and R. E. McCulloch.BART: Bayesian Additive Regression Trees. Journal of the Royal Statistical Society, 2006. Ser.B,Revised.

[7.] J. P. Marques de Sa. Pattern Recognition: Concepts,Methods and Applications. Springer, 2001.

[8.] D. Michie, D. J. Spiegelhalter, and C. C. Taylor. Machine Learning, Neural and StatisticalClassi_cation. Ellis Horwood, 1994.

[9.] L.Breiman.Random forests. Machine Learning,45(1):5{32, October2001

[10.] Mrs. Sayantani Ghosh, Mr. Sudipta Roy, Prof. Samir K.Bandyopadhyay, "A tutorial review on Text Mining Algorithms".