IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 4, April 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204189 |

Websites Detection Using Optimal Feature Selection Neural Network and Random Forest Classifier

Baby Kalpana Y, Tharanipriya G

Associate Professor, Dept. of C.S.E., P.A College of Engineering and Technology, Pollachi, India

PG Student, Dept. of C.S.E., P.A College of Engineering and Technology, Pollachi, India

ABSTRACT: Phishing attack is now a big threat to people's daily life and networking environment. The illegal URLs as legitimate ones, attackers can induce users to visit the phishing URLs to get private information. An Effective methods of detecting phishing websites are needed to alleviate the threats posed by phishing attacks. The objective of our proposed approach is to develop a model for effective and accurate phishing website prediction and to avoid some useless or small impact features and falling into the problem of over-fitting. FVV, Feature Validity Value is firstly introduced to evaluate the impact of sensitive features. Optimal feature selection algorithm calculates the FVV values of all features of the input URLs and their relevant websites at first. Then, a threshold is set to select sensitive features to construct an optimal feature vector. Due to no disturbance from these redundant features, the over-fitting problem of the underlying neural network is alleviated. Meanwhile, this algorithm is also able to reduce the time cost of the process of phishing websites detection. The selected optimal features are used to train the underlying neural network and, finally, an optimal classifier (Random Forest Classifier) is constructed to detect the phishing websites.

KEYWORDS: FVV(Feature Validity Value); over fitting problem; optimal feature selection algorithm; random forest classifier.

I. INTRODUCTION

Phishing is one of the fastest growing cyber attacks. Nowadays phishing attacks frequently appear in the computers and mobile phones. Phishing attacks mainly attack user privacy information, in which the attacker tries to learn sensitive information such as login credentials or account information by sending as a reputable entity or person in email or other communication channels. It is easier trick to somebody clicking a malicious link which seems legitimate than trying to break through a computer defense systems. The Anti-Phishing Alliance of China (APAC) has reported that, at the last month of 2018, there is a total of 435193 phishing websites are detected. Effective methods of phishing websites detections are needed to alleviate the threats posed by phishing attacks. The automatic phishing detection methods are classified as four types, the blacklist method, white list method, heuristic method, visual similarity method and machine learning method. In Black and White lists method databases are used to construct based on the previously detected URLs, this method has difficult in newly emerged phishing attacks. The heuristic method is based on assigned signatures, for identified phishing attacks. Through scanning websites for assigned signatures the time-consuming is high, due to complicated nature of phishing attacks. The Visual similarity method also be the same accuracy level of blacklist and white list technique. The machine learning method is accurate in phishing detection websites. It has the ability to adapt newly emerged phishing websites. The main reason to success of this method is highly qualified features from phishing URLs and their related websites. This paper proposes AFNN-R: An Effective and Accurate Phishing Websites Detection using Optimal Feature selection Neural Network and Random Classifier. Under this model, FVV index is defined to evaluate the impact of sensitive features on phishing websites detection. Based on the FVV index, an algorithm is designed to select the optimal features from phishing URLs and their related websites. The selected features are used to train the underlying neural network, then finally random classifier to detect the phishing websites.

II. RELATED WORK

In [1] author used Detecting Phishing Websites through Deep Reinforcement Learning where an agent learns the value function from the given input URL and the classification task. Sequential decision making process for classification

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204189 |

using deep neural network. The reinforcement learning has been utilized to gain proficiency for optimal behavior. It is defined as the problem of an "agent", then to perform an action based on the "error" within an unknown environment which provides feedback in the form of numerical "rewards". The websites feature in four groups (i) Anomaly-based, (ii) Address bar-based, (iii) HTML and JavaScript-based and (iv) Domain-based. [2] author used An Effective Neural Network Phishing Detection Model Based on Optimal Feature Selection Under this model, an optimal feature selection algorithm that adapts to the sensitive features of phishing URLs (Uniform Resource Locators) is firstly proposed. Based on the calculation of the effective value of each feature, this algorithm sets a threshold to eliminate some useless features and selects an optimal feature set suitable for detecting phishing websites [3] author initialize Malicious URLs detection using Machine Learning Techniques. The main reason for malicious URL detection is that they provide an attack surface to the antagonist. In malicious URL detection, we used a novel classification method. In this classification model built on sophisticate machine learning methods that not only takes care about the synthetically nature of the URL, but also the semantic and lexical meaning of these dynamically changing URLs.

III. PROPOSED ALGORITHM

3.1 Module Description

Preprocessing, FFV index Calculation, Optimal Feature Selection, Training the data, Classification of data

IV. SYSTEM ARCHITECTURE

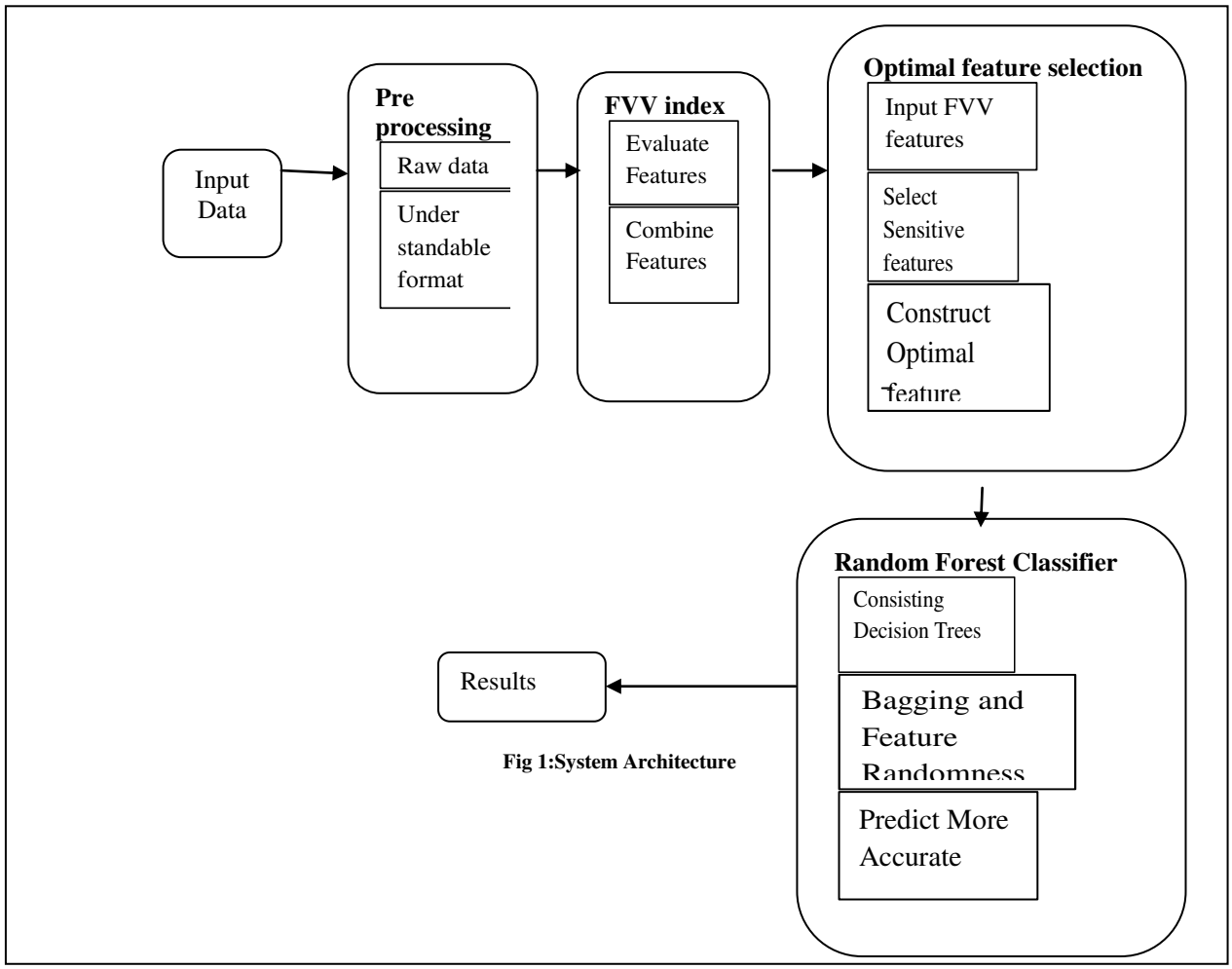

Fig 1:System Architecture

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204189 |

3.1.1 Preprocessing

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Data can have missing values for a number of reasons such as observations that were not recorded and data corruption. Handling missing data is important as many machine learning algorithms do not support data with missing values.

3.1.2 FVV index Calculation

Feature Validity Value is to evaluate the impact of sensitive features on phishing websites detection. In order to better evaluate the impact of a selected sensitive feature on detecting phishing attacks, presents the FVV index. The new FVV is defined by combining the positive and negative features of URLs.

3.1.3 Optimal Feature Selection

Feature extraction is the second class of methods for dimension reduction. This function is useful for reducing the dimensionality of high-dimensional data. (ie you get less columns). This module uses the FFV values of all features of the input URLs and their relevant websites at first. Then, a threshold is set to select sensitive features to construct an optimal feature vector. Through this algorithm, many useless and small influence features are pruned.

3.1.4 Training the data

The selected optimal features are used to train the underlying neural network. Neural network is used to train the data, which is resulted as optimal features. Neural network model is composed of 3 layers: the input, the hidden and the output layers. Algorithms utilized by neural network generally incorporate two phases: the forward propagation and the backward propagation. The forward propagation starts to work when it receives a positive propagation signal. The backward propagation is invoked when the model detects the occurrence of evident deviation between the calculation result of the output layer and the actual value.

3.1.5 Classification of data

Random Forest Classifier is used to classify the data of phishing website detection. The random forest is a classification algorithm consisting of many decisions trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.

V. RESULT AND DISCUSSION

Finally, we utilized Python to simulate the machine learning. In this section, we present our scheme, which consists of five phases: Preprocessing, FVV index Calculation, Optimal Feature Selection, Training the data, Classification of data. The optimal feature selection algorithm can properly deal with problems of big number of phishing sensitivity features and the continuous change of features. This algorithm can reduce the over-fitting problem of the neural network classifier to some extent. This paper to collect more features to perform optimal feature selection. and choose better accuracy of the random forest algorithm. Finally, we utilized Python to simulate the machine learning. In this section, we present our scheme, which consists of five phases: Preprocessing, FVV index Calculation, Optimal Feature Selection, Training the data, Classification of data. The optimal feature selection algorithm can properly deal with problems of big number of phishing sensitivity features and the continuous change of features. This algorithm can reduce the over-fitting problem of the neural network classifier to some extent. This paper to collect more features to perform optimal feature selection. and choose better accuracy of the random forest algorithm.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204189 |

Fig: Performance comparison chart

VI. CONCLUSION AND FUTURE WORK

In practice, a feasible phishing website detection model OFS-NN, which has proved to be highly accurate and has low false negative rate and low false positive rate. In addition, the optimal feature selection algorithm is combined with the neural network algorithm to select the optimal feature value set for the input of the neural network. OFS-NN model uses the neural network algorithm it has strong ability of independent learning. The optimal feature selection algorithm can properly deal with problems of big number of phishing sensitivity features and the continuous change of other features. This algorithm can reduce the over-fitting problem of the neural network classifier to some extent In the future, it is necessary to collect more features to perform optimal feature selection and choose better accuracy algorithm to beat the random forest algorithm.

REFERENCES

1. Joshua Saxe, Richard Harang, Cody Wild, Hillary Sanders. A DeepLearning Approach to Fast, Format-Agnostic Detection of MaliciousWeb Content. In: Proceedings of the 2018 IEEE Symposium on Security and Privacy Workshops (SPW 2018), San Francisco, CA , USA, August2, 2018, pp.8-14.
2. Erzhou Zhu, Chengcheng Ye, Dong Liu, Feng Liu, (2018),"An Effective Neural Network Phishing Detection Model Based on Optimal Feature Selection". In: Proceedings of the 16th IEEE International Symposium on Parallel and Distributed Processing , Melbourne, Australia, Vol:11-13,pp.781-787.
3. Frank Vanhoenshoven, Gonzalo Napoles,(2016), "Detecting malicious URLs using machine learning techniques", IEEE Symposium Series on Computational Intelligence, December 6-9.
4. Cheng Huang, Shuang Hao, Luca Invernizzi, Yong Fang, (2017) "Gossip:Automatically Identifying Malicious Domains from Mailing List Discussions". In: Proceedings of the 2017ACM on Asia Conference on Computer and Communications Security, Abu Dhabi, United Arab Emirates, ISSN NO 2-6,pp.494-505.
5. El-Sayed M. El-Alfy.(2017) "Detection of Phishing Websites Based on Probabilistic Neural Networks and K-Medoids Clustering". The Computer Journal, ISSN NO 60(12), pp.1745-1759. [6] Guang Xiang, Jason I. Hong, Carolyn Penstein Rose,(2011), "A Feature-rich Machine Learning Framework forDetecting Phishing Web Sites". ACM Transactions on Information and System Security, 14(2), Vol No. 21.
6. R.Gowtham,IlangoKrishnamurthi,(2014), "A comprehensive and efficacious architecture for detecting phishing webpages",Computers & Security,Vol.40, pp.23-37.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204189 |

7. Mahmoud Khonji, Youssef Iraqi, Senior Member, Andrew Jones. Phishing Detection: A Literature Survey. IEEE Communications Surveys and Tutorials, 15(4), 2013, pp.2091-2121.
8. María M. Moreno-Fernández, Fernando Blanco, Pablo Garaizar, Helena Matute. Fishing for phishers. Improving Internet users' sensitivity to visual deception cues to prevent electronic fraud. Computers in Human Behavior, Vol.69, 2017, pp.421-436.
9. M. Junger, L. Montoya, F.-J. Overink. Priming and warnings are not effective to prevent social engineering attacks. Computers in Human Behavior, Vol.66, 2017, pp.75-87.
10. Alessandro Acquisti, Idris Adjerid,(2017) Rebecca Balebako, Laura Brandimarte, Lorrie Faith Cranor, Saranga Komanduri, Pedro Giovanni Leon, Norman Sadeh, Florian Schaub, Manya Sleeper, Yang Wang, Shomir Wilson. Nudges for Privacy and Security: Understanding and Assisting Users‟ Choices Online. ACM Computing Surveys Issue No. 44,50(3)

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details