Practical Thinking on Neural Network Phishing Website Detection Research Based on Decision Tree and Optimal Feature Selection

Qi Zhang
Shandong University of Political Science and Law, 250014
20385489@dlvct.edu.cn

Abstract—This article analyzes the detection methods of neural network phishing websites. The research content of this paper includes naive Bayes method, decision tree method, support vector machine method, neural network technology. The author combines the key points of phishing website detection based on decision tree and optimal feature selection to study such as URL feature and HTML feature analysis, website application feature analysis, K-Medoids cluster analysis, feature set screening. The author uses simulation experiments to complete the website performance check. The purpose of this article is to optimize the performance of phishing website detection and improve the security of the website's operating environment.

1. INTRODUCTION
In the past detection activities of phishing websites, the main points of detection mainly focused on the basic mechanism of the phishing process, but the attention paid to the attack technology was poor. In the context of continuous optimization of various technologies, traditional phishing website detection methods are difficult to meet application needs, and we need to adopt a more high-quality detection system to complete the phishing work. Improving the detection content of neural phishing websites by using decision trees and optimal feature selection can not only improve the quality of phishing website detection work, but also have a positive meaning for improving the security of the network environment.

2. ANALYSIS OF DETECTION METHODS OF NEURAL NETWORK PHISHING WEBSITES

2.1. Naive Bayes Method
In the process of detecting phishing websites, the development system of Naive Bayes Method is relatively mature. It has good applications in fields such as mail screening, text hazard identification, and website security detection. In the application process of this technology, its application process is as follows. Suppose there is a sample data set $A = \{(a_1,b_1),(a_2,b_2),(a_3,b_3),...,(a_n,b_n)\}$, $x_i \in U$, $i = 1,2,3,..n$, represents the feature vector, and $y_i \in [1,-1]$, $i = 1,2,3,..n$, represents the dependent variable. At this point, we can calculate the parameters based on the relationship between the two. According to the relevant calculation conditions of the Naive Bayes method, on the basis of feature independence, the calculation conditions of the given category $Y$ in the application can be expressed as the following formula $P(X|Y=y)=\sum_{i=1}^{n} P(X_i |Y=y)$. According to the logical relationship between the parameters, we can also analyze the data parameters reasonably during the processing. In this way, the compliance of the website can be judged, and a reasonable evaluation result can be obtained.
2.2. Decision Tree Method

Decision tree is a decision analysis method that evaluates project risk and judges its feasibility by constructing a decision tree to obtain the probability that the expected value of the net present value is greater than or equal to zero based on the known probability of occurrence of various situations. This is a graphical method of intuitive use of probability analysis (refer to the diagram shown in Figure 1). Since this kind of decision-making branch is drawn into a graph like the branches of a tree, it is called a decision tree. In machine learning, a decision tree is a predictive model. It represents a mapping relationship between object attributes and object values. Among them, Entropy = the degree of disorder of the system, using algorithm ID3, C4.5 and C5.0 spanning tree algorithm using entropy. In actual drawing, the decision tree generally includes three parts. First, the root node. It is generally represented by a rectangle, which is the basic content of decision-making and delineates a rough analysis range to speed up the analysis of the optimal solution. Second, child nodes. It is generally represented by a circle or an ellipse, which is an extension of the decision-making base, and the degree of content refinement has also been greatly improved. Third, leaf nodes. It further refines the frame structure based on the child nodes, so as to facilitate the calculation of the optimal decision content.

2.3. Support Vector Machine

The support vector machine method in the application of this method is also a method often used in the current phishing website detection process. This method can also smoothly deal with the content of high-dimensional feature space in the application process, so as to obtain the most suitable application content and application parameters. In the specific application process, suppose there is a sample data set A, A={(a1,b1),(a2,b2),(a3,b3)...(an,bn)}, xi ∈ U, i=1,2,3...n, represent the feature vector, and yi ∈ {1,-1}, i=1,2,3...n, represent the dependent variable. Then in the calculation process, you can use the SVM application function (classification as shown in Figure 2) to complete the detailed analysis of the content, so as to obtain the optimal solution of the plane data analysis result in the entire function.
calculation process. In the detection of phishing websites, the accuracy rate can reach more than 95%, which meets the relevant requirements in the analysis process.

2.4. Neural Network Technology
Artificial neural networks, abbreviated as ANNs or NNs, are a way of simulating information processing similar to the human brain. Artificial neural network recognizes existing patterns and relationships through empirical learning. The main components include the connection pattern that constitutes its architecture, the learning algorithm of interconnection weights, and the transfer function. The neuron is the basic element of the neural network. The neural network is composed of multiple neurons and the weights of the neurons connected to them. The neurons in the back layer receive input from one or more other neurons in the front layer. These input information are passed through the weighted directed edge connection, and then processed by the "stimulus function" to generate the output of the neuron.

The neuron model established in the analysis process is shown in Figure 3. In the specific application process, it is assumed that there is a certain sample data set \( x = \{x_1, x_2, x_3...x_n\} \), which represents the type of input neuron. The connection weight is represented by \( W = \{w_1, w_2, w_3,...,w_n\} \), and \( f(\cdot) \) represents the activation function of the calculation process. The specific calculation process is as follows: \( y = f(\sum_{i=1}^{n} w_i x_i - \theta) \). Among them, \( \theta \) is the threshold. After calculation, the correlation between logics can be understood to meet application requirements.

3. Points of Phishing Website Detection Based on Decision Tree and Optimal Feature Selection

3.1. URL Feature and HTML Feature Analysis

3.1.1. URL Feature
During the operation of the network, all network visitors will use URLs to locate network resources uniformly, and complete uniform addressing as required in the application. Moreover, in the application process, the URL will be used as the only application address when the system accesses the server. In use, the address is also composed of protocols, service domain names, file application paths, and string query methods (as shown in Figure 4). Phishing attackers usually use the method of imitating the content of the URL structure to trick customers into visiting their fake websites and steal users' private information. When testing the website, this method will be used to verify the parameter information to meet the corresponding use requirements.
3.1.2. HTML Feature
During the operation of the established network, HTML files are also important application content used. As a hypertext document, many tags are used in the process of using it to meet the application requirements in different situations. The frequently used tag content includes html format, body format and so on. And in the format application process, a tree-like logic structure will also be used to improve the efficiency of logic analysis by 30%-50%. Taking internal web page analysis as an example, in the actual analysis process, the analysis tree that is often used is the DOM tree to complete the internal representation of the web page. The relevant schematic diagram is shown in Figure 5.

3.2. Analysis of Website Application Characteristics

3.2.1. Address Bar Feature Analysis
In the process of website application, the address bar feature mainly refers to the direct or indirect feature information extraction of the content based on the URL structure, and the character strings contained in the content include "@", "/", and ".". Sometimes content such as URL string length, geographic domain name, network IP address, system network short address, etc. will be added to the content to meet the security requirements of network operation. When criminals launch phishing attacks, they analyze part of the feature set content, such as hiding some strings and changing address numbers. For this type of situation, a backup record of the target source, character string, and SSL certificate number in the address bar of the legitimate website is used as an important reference for the detection process.

3.2.2. Anomaly Feature Analysis
When testing and analyzing the system, it is also necessary to do a good job of analyzing abnormal characteristic content to improve the reliability of content analysis results. In the specific analysis process, we need to analyze the number of requested links, the total amount of mailbox information, and the total amount of hyperlink labels in the network to obtain the corresponding analysis results. Moreover, the relevant staff can also check the SFH status information and whois information content to understand whether there is abnormal data information and determine the rationality of the parameter content. If in the inspection process, it is found that there is a lot of difference between the label field and the SFH content vacancy, which is quite different from the initial marked information. At this point,
the relevant staff can preliminarily judge that the website is a phishing website, and use other evaluation content to further determine the conjecture [1].

3.2.3. Java Script Feature Analysis
When analyzing this kind of characteristic content, it mainly focuses on the analysis of the relevant information of the webpage, such as the set Java Script source code information and HTML code information. According to the extracted evaluation criteria, such information is scientifically evaluated, and the recognition accuracy of more than 90% can meet the relevant functional requirements during the operation of the system. For example, a phishing website imitates a regular website by changing the content of the URL status bar, and then analyzes the content of the Java Script source code when identifying it. At this time, it is found that the content of "on Mouse Over" is quite different from the original running state, which can be used to identify the website as a phishing website [2].

3.3. K-Medoids Cluster Analysis
Based on previous analysis experience, it can be understood that for the collected data set, how to perform de-noising analysis on it is also the content that needs to be paid attention to in the analysis process. At present, K-medoids clustering algorithm is mostly used to complete the noise reduction process, reducing the detection workload by 30%-50%. Compared with the previous K-medoids algorithm, in the context of the continuous development of Internet technology and intelligent technology, the clustering center of the analysis process has also undergone major changes. Moreover, its basic search capabilities have also been greatly improved, which has increased the amount of existing calculations by 10%-15% on the original basis, and the accuracy of the calculation results has reached more than 98%. In the meantime, the incremental method is also incorporated in the algorithm calculation process to optimize the existing cluster centers, which can further improve the optimization degree of the classification calculation results. This speeds up the detection speed of phishing websites and meets related application requirements [3].

3.4. Feature Set Screening

3.4.1. Information Gain Feature Analysis
Relevant staff need to strengthen the analysis process of information gain features when performing feature set screening processing. In this process, the Gini coefficient is also used to complete the optimized definition of the system content, thereby completing the reliable f_Value index. This can increase the work efficiency by more than 30% and meet the application requirements in different situations. In the specific analysis, the related applications of the Gini coefficient are as follows. Suppose there is a sample data set A, \( A=\{(a_1,b_1),(a_2,b_2),(a_3,b_3)\ldots(a_n,b_n)\} \), \( x_i \in U, i=1,2,3\ldots n \), represents the feature vector, and \( y_i \in \{1,-1\} \), \( i=1,2,3\ldots n \), represents the dependent variable. At this time, in the calculation process, the following calculation formula can be obtained according to the definition of Gini coefficient: \( G=1-\sum_{|y_i|=1}p_i^2 \). Among them, \( G \) represents the Gini coefficient; \( p_i \) represents the sample set value of the i-th sample data. The required reference coefficient calculated according to the formula can complete the establishment of the new \( f \_Value \) index. Coupled with the decision tree algorithm, reliable data analysis results can be obtained [4].

3.4.2. Feature Evaluation Index Analysis
The characteristic index established according to the characteristic a parameter needs to be expanded around the content of the \( f \_Value \) index during the application process. According to the above content, the calculation formula is as follows: \( Q(a) = (saG(a)-slGl-srGr)/|a|. \) Among them, the index a is the application parameter of the entire decision tree system; \( Q(a) \) represents the evaluation result of the characteristic index parameter of a. sa represents the number of sampling points corresponding to the decision tree root node a in the application process. sl represents the number of sampling points corresponding to the left subset of the decision tree root node a in the application process. The number
of sampling points corresponding to the right subset of the sr decision tree root node a in the application process. G(a) represents the Gini coefficient corresponding to the root node a. Gl represents the Gini coefficient corresponding to the left child of the root node a. Gr represents the Gini coefficient corresponding to the left and right children of the root node a. Based on this to complete the feature evaluation index calculation and do a good job of comparing the standard website index parameters can get a reliable application basis [5].

4. SIMULATION EXPERIMENT ANALYSIS

4.1. Set up the Experimental Environment

In order to verify the rationality of the content of the detection system established this time, in specific applications, programmers need to establish a corresponding simulation experiment environment. The main points of its specific establishment environment application are as follows. (1) The operating system should choose the Mac OS series to meet the needs of phishing website detection. (2) The Eclipse series is selected as the development platform, and the Python language is used to establish a corresponding simulation experiment environment to meet specific application requirements. (3) In the established detection system, its CPU cannot be lower than 2.9 GHz. At the same time, the running memory of the detection system needs to exceed 16G with a higher running speed to meet the requirements of rapid detection and identification of phishing websites [6].

4.2. Build Experimental Data Set

In order to improve the reliability and persuasiveness of the experimental results, programmers need to collect the URL sample parameters of compliant websites and the URL sample parameters of phishing websites to improve the reliability of the analysis results. In order to improve the richness and timeliness of the established experimental data sets, programmers can use the scrapy crawler to complete the rapid extraction of relevant information. Among them, the sample data of the normal website comes from the basic data collected in the Common Crawl system. The phishing sample data comes from the commonly used sample data included in the Phish Tank system. It can group the experimental data sets that have been summarized, and the number of groups is 5-8 groups. Programmers can reduce the error tolerance rate of data analysis results and lay the foundation for the smooth progress of subsequent analysis activities [7].

4.3. Analysis of Results

Using the new detection system and the previous detection system to process the sorted data set, the following data analysis results are obtained. (1) From the perspective of classification effect, the accuracy of the new detection system can reach 97.3%, which is higher than the traditional detection system (88.6%). The classification results obtained are more conducive to the subsequent analysis work. (2) From the perspective of feature selection effects, 15 groups of features are extracted from each group for analysis, and the accuracy difference between the feature vectors obtained by the new detection system before and after selection is less than 0.1%. This can meet the analysis needs of the experimental process. (3) From the perspective of the feature contribution effect, when using the new detection system for data analysis, the average time taken for each type of index extraction is 2.35s-2.40s. It has a higher work efficiency [8].

5. CONCLUSION

In summary, with the continuous improvement of phishing website detection technology, there will be more and more sensitive features of phishing websites. It is very effective to use the best feature selection after studying, analyzing and extracting features. At the same time, useless or negative features may play a key role in a specific phishing website detection model. In future research, we can conduct more in-depth analysis of these features and fuse useless or negative features to improve the reliability of detection results.
REFERENCES

[1] Li Xue. Research on phishing website detection technology based on neural network [J]. Journal of Anyang Normal University, 2020(05): 28-31.

[2] Bi Qingsong, Liang Xuechun, Chen Shuqi. Phishing website detection based on mRMR-RF feature selection and XGBoost model[J]. Computer Applications and Software, 2020, 37(09): 296-301.

[3] Ju Yinyin. Research on Neural Network Phishing Website Detection Based on Decision Tree and Best Feature Selection [D]. Anhui University, 2020.

[4] Ren Mingxia, Liu Junqing, Song Jin, Chen Peng. Phishing website detection method based on RAKE algorithm [J]. Information and Communication, 2019(10): 18-20.

[5] Zhu Shiqi, Nuerbuli. Quantitative analysis of research status and development trend of phishing website detection[J]. Computer Engineering and Applications, 2020, 56(15): 92-100.

[6] Zhu Qi, Lin Guoyuan. Research on phishing website detection method based on improved random forest algorithm[J]. Microelectronics and Computer, 2019, 36(04): 43-46+51.

[7] Yang Peng, Zeng Peng, Zhao Guangzhen, Lv Peipei. Phishing website detection method based on Logistic regression and XGBoost[J]. Journal of Southeast University (Natural Science Edition), 2019, 49(02): 207-212.

[8] Ma Gang, Liu Feng, Zhu Erzhou. An Android Phishing Website Detection Scheme Based on Improved Naive Bayes Algorithm[J]. Computer Engineering and Science, 2018, 40(08): 1420-1428.