Postprint of Phishing Website Detection Method Based on Feature Selection and Ensemble Learning

Authors: Zhou Chuanhua, Liu Zhicai, DING Jing' an, Zhou Jiayi

Date: 2018-05-02T00:00:00+00:00

Abstract

To address the issues of low detection accuracy and high false positive rates in most existing phishing website detection methods, this paper proposes a phishing website detection method based on feature selection and ensemble learning. First, the FSIGR algorithm is employed for feature selection. This algorithm combines the advantages of filter and wrapper modes, comprehensively measures features from the perspectives of information correlation and classification capability, and adopts a forward incremental backward recursive elimination strategy to select features. Classification accuracy is used as the evaluation metric to assess and select feature subsets, thereby obtaining the optimal feature subset. Subsequently, training is conducted based on the random forest ensemble learning classification algorithm using the selected optimal feature subset. Experiments on the UCI dataset demonstrate that the proposed method can effectively improve the accuracy of phishing website detection, reduce the false positive rate, and holds practical application significance.

Full Text

Method of Phishing Website Detection Based on Feature Selection and Ensemble Learning

Zhou Chuanhua[1],[2], Liu Zhicai[1]†, Ding Jing' an[1], Zhou Jiayi[3]

[1]School of Management Science & Engineering, Anhui University of Technology, Maanshan, Anhui 243002, China
[2]School of Computer Science & Technology, University of Science & Technology of China, Hefei 230026, China
[3]Graduate School of Information, Production and Systems, Waseda University, Tokyo, Japan

Abstract: Most existing phishing website detection methods suffer from low detection accuracy and high false positive rates. To address these issues, this paper proposes a phishing website detection method based on feature selection and ensemble learning. First, the FSIGR algorithm is employed for feature selection. This algorithm combines the advantages of filter and wrapper modes, comprehensively measuring features from two aspects: information correlation and classification capability. It adopts a forward-increasing backward-recursive elimination strategy for feature selection, using classification accuracy as the evaluation metric to assess and select feature subsets, thereby obtaining the optimal feature subset. Then, the selected optimal feature subset is used to train a classification model based on the random forest ensemble learning algorithm. Experiments on the UCI dataset demonstrate that the proposed method can effectively improve the accuracy of phishing website detection while reducing the false positive rate, making it suitable for practical applications.

Keywords: phishing website; random forest; information gain ratio; feature selection

Introduction

Phishing websites represent a significant threat to internet security. These malicious webpages impersonate legitimate sites and employ social engineering techniques to attack users and obtain sensitive information such as usernames and passwords for financial gain [1,2]. According to the Anti-Phishing Working Group (APWG) report, in the fourth quarter of 2016, APWG detected an average of 92,564 phishing attacks per month—a 5,753% increase over 2004. The total number of online fraud attacks in 2016 reached 1,220,523, a 65% increase from 2015. China was the most severely affected country, with 47.09% of machines infected [3].

As the internet continues to develop and gain popularity, particularly with the rapid growth of e-commerce, internet security has become increasingly critical. Although attackers use various techniques to create phishing websites to deceive users, they all employ a common set of features in their design. This provides anti-malware researchers with methods and ideas for solving the problem. Current phishing website detection methods primarily include user education [4-6], blacklist technology [7,8], and heuristic techniques [9-16]. Among these, heuristic techniques have been most widely researched and deployed. These methods extract relevant website features and then apply heuristic rules or machine learning algorithms to process these features for webpage classification (legitimate/phishing).

Literature [11] extracted features from webpage titles, keywords, and other elements, using NBC and SVM classification algorithms as base classifiers and employing an ensemble classification method to integrate detection results, proposing an effective intelligent phishing website detection system. Literature [12]

proposed a phishing detection system based on SVM classification algorithm for URL matching and classification recognition. While this system improved phishing website prediction accuracy, it was only suitable for low-dimensional small-sample data. Literature [13] used the K-means algorithm to process URL features or page features for phishing website prediction. Although this method improved classification model accuracy to some extent, its classification performance was limited. Literature [14] compared multilayer perceptron, decision tree, and Bayesian classification algorithms for phishing website prediction and found that the decision tree classification model exhibited superior performance relative to the other two algorithms.

Through analysis of the above literature, we identify two key issues: (a) Features are typically extracted from HTML tags, URL addresses, encoding, page images, etc. [15,16], resulting in high-dimensional feature spaces with numerous redundant features that affect classification model accuracy. (b) Single classifier models have limited classification performance, with poor generalization capability and fault tolerance.

To address these problems, this paper proposes a phishing website detection method based on feature selection and ensemble learning algorithms. Feature selection can effectively reduce redundant features, thereby improving phishing website prediction accuracy [17-20] and reducing time overhead. Using ensemble learning algorithms to integrate classification results from various base classifiers for model construction can effectively improve fault tolerance and generalization capability, thereby reducing false positive rates in phishing website prediction. In the feature selection stage, we propose the FSIGR (Feature Selection based on Importance and Gain Ratio) algorithm. FSIGR combines the advantages of filter and wrapper modes. In the filter stage, features are selected based on their information correlation with class labels. In the wrapper stage, features are evaluated from both information correlation and classification capability dimensions to compute feature weight vectors and comprehensive weights, which are then ranked. A forward-increasing backward-recursive elimination strategy is employed for selection, with classification accuracy used to evaluate feature subsets, thereby selecting an optimal feature subset with strong relevance and low redundancy to improve phishing website prediction accuracy. In the classification stage, the random forest ensemble learning classification algorithm is used to train data and obtain the final classification model, reducing the false positive rate of phishing website prediction. Experimental results demonstrate that the proposed phishing website detection method can effectively improve prediction accuracy and reduce false positive rates.

1.1 Entropy and Information Gain Ratio

Shannon entropy is a fundamental concept in information theory, serving as a mathematical expression to measure the uncertainty of random variables and

the average information content of a variable or variable set, typically denoted as $H(X)$. Let $X = \{x_1, x_2, \dots, x_m\}$ and $Y = \{y_1, y_2, \dots, y_m\}$ be two random variables with probability density functions $p(x_i)$ and $p(y_i)$. The entropy of random variable X is defined as:

$$H(X) = -\sum p(x_i) \log_2 p(x_i)$$

The conditional entropy of random variables X and Y is defined as:

$$H(X|Y) = \sum p(y_i) H(X|Y = y_i)$$

Conditional entropy satisfies $H(X|Y) \leq H(X)$ and measures the correlation between variables X and Y. If X and Y are uncorrelated, then $H(X|Y) = H(X)$. If X and Y are correlated, then $H(X|Y) < H(X)$, and a larger value of $H(X) - H(X|Y)$ indicates stronger correlation between X and Y.

Information gain (IG) measures the amount of information one random variable provides about another random variable. IG is non-symmetrical and dimensionless, with larger values indicating stronger correlation between variables. The relationship between information gain, entropy, and conditional entropy is:

$$IG(X|Y) = H(X) - H(X|Y)$$

From this equation, a larger $IG(X|Y)$ value indicates stronger correlation between variables X and Y, where $IG(X|Y)$ represents the information gain of variable Y.

In information systems, information gain is commonly used to measure a feature' s contribution to classification and reduce sensitivity to noise in examples. However, information gain tends to prefer features with more branches, leading to overfitting. Therefore, a penalty factor is often introduced to penalize features with many branches, resulting in information gain ratio (Gain Ratio, GR):

$$GR(X|Y) = \frac{IG(X|Y)}{H(Y)}$$

Equation (4) shows that the information gain ratio of random variable Y is directly proportional to its information gain and inversely proportional to its information entropy. Consequently, when random variable Y has many values, $GR(X|Y)$ decreases as $H(Y)$ increases, mitigating selection bias to some extent.

1.2 Random Forest and Importance Measurement

Random forest (RF) is an ensemble learning algorithm that constructs multiple decision trees using random resampling and random node splitting techniques, with final results determined by a voting mechanism. Due to its robustness to noisy and missing data, fast learning speed, and ability to use variable importance measures as a feature selection tool for high-dimensional data, RF has been widely applied to various classification, prediction, feature selection, and anomaly detection problems [21,22].

Random forest-based importance measurement evaluates the impact of feature attributes on output variables through out-of-bag (OOB) data testing and random noise addition. Greater impact indicates higher feature importance [23-25]. The main steps are as follows:

Assume the random forest includes M classification and regression trees. To measure the importance of the j-th feature attribute to the output variable, process each classification tree in the random forest. For the i-th tree ($i = 1, 2, \ldots, M$):

a) Calculate the prediction error rate of the i-th tree based on out-of-bag observations, denoted as e_i.

b) Randomly shuffle the order of values for the j-th feature attribute on out-of-bag observations, rebuild the i-th tree, and make predictions on the out-of-bag observations.

c) Recalculate the prediction error of the i-th tree, denoted as e_i^j.

$\varepsilon_i^j = e_i - e_i^j$ represents the change in prediction error of the i-th tree caused by adding noise to the j-th feature attribute. Repeating the above steps yields M prediction error changes. $\varepsilon_j = \frac{1}{M}\sum_{i=1}^{M}\varepsilon_i^j$ represents the average change in overall prediction error of the random forest caused by adding noise to the j-th input variable, measuring the importance of the j-th input variable.

2.1 FSIGR Feature Selection Method

Feature selection aims to select a representative feature subset from an original feature set while maintaining classification performance, thereby reducing feature space dimensionality [26]. Based on dependency on machine learning algorithms, feature selection methods can be categorized into filter and wrapper approaches. Filter methods evaluate and select feature subsets using intrinsic data characteristics, independent of machine learning algorithms. These methods typically have high efficiency but poor classification performance. Wrapper methods rely on machine learning algorithm classification accuracy as the evaluation criterion for feature subset selection. These methods achieve better classification performance but lower efficiency.

Combining the advantages of filter and wrapper approaches, this paper proposes the FSIGR (Feature Selection based on Importance and Gain Ratio) algorithm based on information gain ratio and importance measurement. The FSIGR algorithm consists of filter and wrapper stages. The key steps are as follows:

Input: Dataset D, feature set $F = \{f_i | i = 1 \dots v\}$. Initialize $a_{max} = 0$, $F_{best} = \emptyset$.

a) Filter irrelevant features and comprehensively measure relevant features.

First, calculate the GR of each feature with respect to the class feature. If $GR = 0$, the feature is uncorrelated with the class feature and is deleted from the feature set. For remaining features in the subset, calculate comprehensive measurement values.

Calculate the GR value g_i of feature f_i with respect to the class feature. If $g_i = 0$, delete feature f_i from the set: $F = F - \{f_i\}$.

Use random forest to compute the importance value of feature f_i, denoted as m_i.

Apply equations (5) and (6) to standardize m_i and g_i, obtaining $\tilde{m}_i$ and $\tilde{g}_i$.

Calculate the comprehensive evaluation value c_i of feature f_i using equation (7).

Rank features in descending order according to their comprehensive evaluation value c_i.

b) Feature selection using forward-increasing backward-recursive elimination strategy.

According to the ranking, traverse the feature space using a forward selection strategy. Calculate the classifier accuracy a_i on feature subset F_i, where i represents the number of elements in the feature subset.

Set flag = false.

For a_i $(i = 1 \dots v)$ do: - If $a_i < a_{i-1}$ then: - Set flag = true - Remove feature f_i from set F and record the classifier accuracy after removing f_i as a_{temp} - If $a_{max} < a_{temp}$ then: - Set $a_{max} = a_{temp}$ - Set $F_{best} = F$ - End if - Break - End if End for

Repeat until flag == false (termination condition met).

Output: Optimal feature subset F_{best}.

The algorithm description shows that FSIGR includes both filter and wrapper stages. In the filter stage, features are filtered based on information correlation with class labels. In the wrapper stage, features are sorted by comprehensive measurement and selected using forward-increasing backward-recursive elimination strategy, with a classifier evaluating the feature subset to select the optimal

subset with strong relevance and low redundancy, thereby improving phishing website prediction accuracy.

The advantages of this selection strategy are: based on comprehensive evaluation, it uses classification accuracy to reassess each feature' s contribution to overall classification, which can reduce feature volatility and delete redundant attributes with low importance without sacrificing algorithm precision. After each feature deletion, the feature set is traversed again to generate new feature combinations, expanding the search space coverage of feature subsets to select the minimal redundancy and optimal performance feature set.

Compared with forward and backward search strategies, our search strategy uses overall feature subset classification performance as the evaluation metric on the basis of ranking, recursively eliminating redundant features with minimal c_i values. This can reduce feature volatility without sacrificing algorithm precision. Compared with filter methods in literature [19,23], our approach adopts a filter+wrapper mode, improving feature subset classification performance.

2.2 FSIGR Algorithm Complexity Analysis

The algorithm' s time overhead consists of two main parts:

a) Filter stage: Filters features based on information correlation and comprehensively measures features based on classification capability.

b) Wrapper stage: Sorts features by comprehensive measurement, selects feature subsets using forward-increasing backward-recursive elimination search strategy, and evaluates feature subsets using a classifier.

The time overhead primarily manifests in the wrapper stage. According to literature [22], if the training dataset has m features and n training samples, and the random forest contains k base classifiers, then the time complexity of the random forest algorithm is approximately $O(kmn(\log n)^2)$, and the average time complexity of quicksort is $O(m(\log m))$.

Therefore, in our algorithm, the filter stage time complexity is $O(m + kmn(\log n)^2)$. The wrapper stage outer loop runs at most m times. Each loop uses forward-increasing strategy for feature selection with $(m, m-1, m-2, \ldots, 1)$ iterations, and backward-recursive elimination strategy with an average of $m/2$ iterations and at most $m-1$ iterations. Thus, the maximum time complexity of the FSIGR algorithm can be approximated as:

$$O(m + kmn(\log n)^2) + O(m(\log m)) + O(m + \frac{1}{2} \times m(m-1)) + O(m-1)$$

$$= O(m(\frac{1}{2} \times (m+5) + kn(\log n)^2 + \log m))$$

$$T(n) = O(m^2)$$

Since the temporary storage space occupied during algorithm execution is linearly proportional to the number of features, the space complexity can be expressed as:

$$S(n) = O(m)$$

As shown in equations (9) and (10), the maximum time complexity of the FSIGR algorithm is approximately quadratic with respect to feature dimensionality, while space complexity is linear with feature dimensionality. Therefore, the FSIGR algorithm demonstrates good processing capability for high-dimensional data and excellent scalability.

2.3 Phishing Website Detection Model

[Figure 1: see original paper] illustrates the phishing website detection model proposed in this paper, which consists of three main components:

a) Feature extraction: Parse webpage content and extract relevant features (the dataset used in our experiments).

b) Feature selection: Employ the FSIGR feature selection algorithm to evaluate and select features from both individual feature and feature subset perspectives, thereby selecting an optimal feature subset with high relevance and low redundancy.

c) Classification decision model: Construct a classification decision model using the RF ensemble learning algorithm to effectively improve classification accuracy for phishing website detection.

The model' s primary execution process is as follows: First, extract features from webpages based on HTML tags, URL addresses, encoding, page images, etc., and convert them into training and prediction data. Then, apply the FSIGR algorithm to the extracted feature data for feature selection to identify the optimal feature subset. Finally, train the RF classification decision model and perform prediction based on the selected optimal feature subset data.

3.1 Experimental Data

This experiment uses the phishing dataset from the UCI database [27], which includes 11,055 website instances. Among them, 4,898 instances (44%) are labeled as phishing pages (represented by -1), and 6,157 instances (56%) are labeled as legitimate pages (represented by 1). Each instance contains 30 features extracted based on address bar, abnormal signs, HTML and JavaScript,

and domain name. Feature values are binary (-1, 1) or ternary (0, 1, -1). More detailed information can be found in literature [27].

3.2 Experimental Setup

To fully validate the effectiveness of the proposed phishing website detection method, the experiment consists of two parts:

Experiment 1: Validate the effectiveness of the FSIGR algorithm.

In this experiment, we compare FSIGR with CFS (correlation-based feature selection), WFS (wrapper feature selection) algorithms, and the algorithm from literature [19]. The RF ensemble learning classification algorithm is used to validate experimental results from different feature selection algorithms, with 10-fold cross-validation employed to calculate classification accuracy. Experimental results are compared and analyzed to verify the effectiveness of our feature selection method.

Experiment 2: Validate the effectiveness of the proposed phishing detection method.

On the phishing dataset, we first use the FSIGR feature selection algorithm to select the optimal feature subset (using the corresponding classification algorithm as the feature subset evaluator). Then, we train classification models using C4.5, KNN, Naive Bayes, REP Tree, and RF algorithms, respectively, and calculate classification model accuracy using 10-fold cross-validation. Experimental results are compared to verify the effectiveness of our phishing detection method.

The experimental platform is WEKA with Java language, running on an Intel® Core™ i5-6300HQ @ 2.3 GHz processor with 8 GB of memory.

3.3 Evaluation Metrics

Two metrics are generally used to evaluate classification algorithm performance:

1) Accuracy: Also called precision, calculated as:

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

2) Recall: Also called sensitivity, calculated as:

$$\text{recall} = \frac{TP}{TP + FN}$$

Where: TP (true positive) is the number of positive samples correctly classified; FP (false positive) is the number of negative samples incorrectly classified as positive; TN (true negative) is the number of negative samples correctly classified; FN (false negative) is the number of positive samples incorrectly classified as negative.

3.4 Results Analysis

Experiment 1 results:

Table 2 shows experimental results on the phishing dataset using different feature selection algorithms to build RF classification prediction models with 10-fold cross-validation. In literature [19], features are selected from GR and RF importance-ranked lists using a threshold of 0.01. SF represents the number of selected features, Acc represents classification accuracy, M-error represents mean absolute error, and AUC represents the area under the ROC curve.

Table 2: Experimental results of RF classification prediction models based on different feature selection algorithms/%

Feature Selection Algorithm	SF	Accuracy	Recall	M-error	AUC
WFS(BF)	28	95.215±3.448	95.2	0.095	0.987
WFS(GS)	29	96.002±2.769	96.0	0.080	0.991
CFS(BF/GS)	9	94.772±3.873	94.8	0.105	0.985
Literature [21] Algorithm	20	96.834±2.292	96.8	0.0487	0.994
FSIGR	23	97.341±2.053	97.3	0.048	0.996

Note: The data before and after ± represent the mean classification accuracy and variance of 10 test runs, respectively.

As shown in Table 2, the FSIGR feature selection method achieves a classification accuracy of 97.341%, recall of 97.3%, and mean absolute error of 0.048, all superior to other feature selection methods. The algorithm from literature [21] achieves 96.834% accuracy, 96.8% recall, and 0.0487 mean absolute error, with classification model performance significantly lower than the FSIGR method. CFS, GR, and RF feature selection methods perform well in dimensionality reduction, selecting feature subsets of sizes 9, 11, and 13, respectively, but with lower classification accuracy. WFS feature selection methods with two search strategies select feature subsets of sizes 28 and 29, showing lower dimensionality reduction performance than other methods, with classification accuracies of 97.205% and 97.286%, respectively. While these outperform CFS and other methods, their overall performance is inferior to our method and they incur greater time costs.

Experimental results demonstrate that the FSIGR feature selection method can select feature subsets with lower dimensionality and optimal classification performance, meeting practical application requirements and proving its effectiveness.

Experiment 2 results:

Table 3 lists experimental results on the phishing dataset using different classification algorithms with FSIGR feature selection compared with our method. All results are generated using 10-fold cross-validation.

Table 3: Experimental results of classification prediction models based on FSIGR feature selection algorithm/%

Classification Algorithm	Accuracy	Recall	M-error	AUC
C4.5	96.056±3.312	96.1	0.079	0.991
KNN	96.834±2.292	96.8	0.328	0.994
Naive Bayes	92.999±5.308	93.0	0.140	0.976
REP Tree	97.205±2.091	97.2	0.056	0.995
RF	97.341±2.053	97.3	0.048	0.996

Note: The data before and after ± represent the mean classification accuracy and variance of 10 test runs, respectively.

As shown in Table 3, our method achieves 97.341% classification accuracy, 97.3% recall, 0.048 mean absolute error, and a feature subset dimension of 23, with comprehensive classification performance significantly superior to C4.5, REPTree, and NaiveBayes algorithms. Compared with the KNN algorithm, although its mean absolute error is higher than KNN' s 0.328, its overall performance is superior. The experimental results demonstrate that our phishing website detection method significantly outperforms C4.5, KNN, REPTree, and NaiveBayes algorithms, validating the effectiveness of our approach.

The Receiver Operating Characteristic (ROC) curve reflects the comprehensive generalization performance of classification models across different tasks. The area under the ROC curve (AUC) indicates stronger generalization capability when larger. Table 2 shows that our FSIGR algorithm achieves an AUC of 0.996 under the same classifier, outperforming other feature selection algorithms and proving its applicability. Table 3 shows that under the same feature selection algorithm, the RF classification model achieves an AUC of 0.996, outperforming other classification models and proving that the RF ensemble learning model has strong fault tolerance. Therefore, our phishing website detection model exhibits strong generalization capability.

[Figure 2: see original paper] shows the ROC curve of the RF classification decision model based on the FSIGR algorithm.

[Figure 3: see original paper] and [Figure 4: see original paper] depict line charts of classification accuracy changes for the C4.5 algorithm across different feature

dimensions on the phishing dataset and optimal feature subset, respectively. In Figure 3, blue triangles represent cases where classification accuracy remains unchanged after adding the current feature to the subset, while red triangles represent cases where accuracy decreases after adding the feature.

[Figure 5: see original paper] depicts the line chart of classification accuracy changes for the RF ensemble learning algorithm across different feature dimensions on the optimal feature subset.

Comparing Figures 3 and 4 reveals that in the original data, classification accuracy remains unchanged (feature 18 in Figure 3) or even decreases (features 2, 21, 22, 23, 28, and 30 in Figure 3) as the number of features increases. In Figure 4, however, classification accuracy continuously improves as feature dimensionality increases, reaching a maximum accuracy of 96.056% at 25 features—superior to the original dataset' s 95.920% at 30 features. Time overhead is reduced from 0.17s to 0.1s. This occurs because the original dataset contains irrelevant and redundant features that degrade classifier performance. The FSIGR method can comprehensively measure features from both information correlation and classification capability perspectives, selecting an optimal feature subset with strong relevance and low redundancy to improve classifier accuracy.

Figures 4 and 5 show that the RF ensemble learning algorithm achieves 97.341% classification accuracy with 23 features, significantly outperforming the C4.5 single classifier' s 96.056% at 25 features. This is because ensemble learning algorithms can enhance fault tolerance and generalization capability by integrating classification results from different base classifier models, thereby improving classification accuracy and reducing classification errors. Our experiments demonstrate the effectiveness of ensemble learning algorithms for phishing website detection, thereby validating the effectiveness of our phishing website detection method.

---

4 Conclusion

This paper proposes a phishing website detection method based on feature selection and ensemble learning. The method first employs the FSIGR algorithm to select an optimal feature subset with strong relevance and low redundancy, then uses this optimal feature subset to train a classification model based on the RF ensemble learning classification algorithm to improve prediction accuracy.

Experimental results on the phishing dataset comparing FSIGR with CFS, WFS, and literature [19] algorithms demonstrate that FSIGR exhibits excellent performance in both dimensionality reduction and classification accuracy improvement, proving its effectiveness. Complexity analysis of FSIGR reveals its good processing capability and scalability for high-dimensional data. Experiments comparing the RF ensemble learning algorithm with C4.5, KNN, REPTree, and NaiveBayes algorithms on the phishing dataset show that RF significantly out-

performs other single classifier models, with advantages of high classification accuracy, low classification error, and high recall rate. Based on the above, the effectiveness and practical applicability of our phishing website detection method are proven.

Future work will focus on using associative information entropy to rank the correlation of combined features, selecting optimal feature subsets, and constructing phishing website detection models to improve prediction accuracy.

References

[1] Almomani A, Gupta B B, Atawneh S, et al. A survey of phishing email filtering techniques [J]. IEEE Communications Surveys & Tutorials, 2013, 15(4): 2070-2090.

[2] Mishra A, Gupta B B. Hybrid solution to detect and filter zero-day phishing attacks [C]//Proc of the 2nd International Conference on Emerging Research in Computing, Information, Communication and Applications. 2014: 373-382.

[3] Anti-Phishing Working Group. Phishing activity trends report of 4th quarter of 2016 [R]. 2016.

[4] Sheng S, Holbrook M, Kumaraguru P, et al. Who falls for phish?: a demographic analysis of phishing susceptibility and effectiveness of interventions [C]//Proc of Sigchi Conference on Human Factors in Computing Systems. 2010: 373-382.

[5] Arachchilage N A G, Love S. A game design framework for avoiding phishing attacks [J]. Computers in Human Behavior, 2013, 29(3): 706-714.

[6] Zhuang W, Jiang Q. Intelligent anti-phishing framework using multiple classifiers combination [J]. Journal of Computational Information Systems, 2012, 8(17): 7267-7281.

[7] Zhang J, Porras P, Ullrich J. Highly predictive blacklisting [C]//Proc of Conference on Security Symposium. USENIX Association. 2008: 107-122.

[8] Sharifi M, Siadati S H. A phishing sites blacklist generator [C]//Proc of IEEE//ACS International Conference on Computer Systems and Applications. 2008: 840-843.

[9] Zhang Y, Hong J I, Cranor L F. CANTINA: a content-based approach to detecting phishing web sites [C]//Proc of International Conference on World Wide Web. 2007: 639-648.

[10] Xiang G, Hong J, Rose C P, et al. CANTINA+: a feature-rich machine learning framework for detecting phishing web sites [J]. ACM Trans on Information & System Security, 2011, 14(2): 21.

[11] Zhuang W, Ye Y, Li T, et al. Intelligent detection system for phishing websites based on classification ensemble [J]. Systems Engineering—Theory & Practice, 2011, 31(10): 2008-2020.

[12] He G, Zou F, Tan D, et al. Phishing detection system based on SVM active learning algorithm [J]. Computer Engineering, 2011, 37(19): 126-128.

[13] Sahu K K, Shrivastava S. Kernel K-means clustering for phishing website and malware categorization [J]. International Journal of Computer Applications, 2015, 111(9): 20-25.

[14] Lakshmi V S, Vijaya M S. Efficient prediction of phishing websites using supervised learning algorithms [J]. Procedia Engineering, 2012, 30(9): 798-805.

[15] Pan Y, Ding X. Anomaly based Web phishing page detection [C]//Proc of the 22nd Computer Security Applications Conference. 2006: 381-392.

[16] Basnet R B, Sung A H, Liu Q. Rule-based phishing attack detection [C]//Proc of International Conference on Security and Management. 2011.

[17] Zuhair H, Selmat A, Salleh M. The effect of feature selection on phishing website detection [J]. International Journal of Advanced Computer Science & Applications, 2015, 6(10): 221-232.

[18] Zhang W, Ren H, Jiang Q. Application of feature engineering for phishing detection [J]. IEICE Trans on Information & Systems, 2016, 99(4): 1062-1065.

[19] Rajab K D. New hybrid features selection method: a case study on websites phishing [J]. Security & Communication Networks, 2017, 2017(2): 1-10.

[20] Basnet R B, Sung A H, Liu Q. Feature selection for improved phishing detection [C]//Proc of the 25th International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems. 2012, 7345: 252-261.

[21] Hideko K, Hiroaki Y. Rapid feature selection based on random forests for high-dimensional data [J]. Ipsj Sig Notes, 2012, 2012: 1-7.

[22] Yao D, Yang J, Zhan X. Feature selection algorithm based on random forest [J]. Journal of Jilin University: Engineering and Technology Edition, 2014, 44(1): 137-141.

[23] Han H, Guo X, Yu H. Variable selection using mean decrease accuracy and mean decrease gini based on random forest [C]//Proc of IEEE International Conference on Software Engineering and Service Science. 2017: 219-224.

[24] Wang H, Lin C, Peng Y, et al. Application of improved random forest variables importance measure to traditional Chinese chronic gastritis diagnosis [C]//Proc of IEEE International Symposium on It in Medicine and Education. 2008: 84-89.

[25] Nicodemus K K. Letter to the editor: on the stability and ranking of predictors from random forest variable importance measures [J]. Briefings in Bioinformatics, 2011, 12(4): 369-373.

[26] Guyon, Isabelle, Elisseeff, et al. An introduction to variable and feature selection [J]. Journal of Machine Learning Research, 2003, 3(6): 1157-1182.

[27] Mohammad R M, Thabtah F, McCluskey L. Phishing websites features [EB/OL]. (2015) http://eprints.hud.ac.uk/24330/6/RamiPhishing_Websites_Features.pdf.

Note: Figure translations are in progress. See original paper for figures.