A Classification Model for Detection of Chinese Phishing E-Business Websites

Hansi Jiang  
Beijing Institute of Technology, jianghansii@qq.com

Dongsong Zhang  
University of Maryland - Baltimore County, zhangd@umbc.edu

Zhijun Yan  
Beijing Institute of Technology, yangzhijun@bit.edu.cn

Follow this and additional works at: http://aisel.aisnet.org/pacis2013

Recommended Citation
http://aisel.aisnet.org/pacis2013/152
A CLASSIFICATION MODEL FOR DETECTION OF CHINESE PHISHING E-BUSINESS WEBSITES

Hansi Jiang, School of Management and Economics, Beijing Institute of Technology, Beijing, China, jianghansi@qq.com

Dongsong Zhang, Department of Information Systems, University of Maryland, Baltimore County, USA and Beijing Institute of Technology, China, zhangd@umbc.edu

Zhijun Yan, School of Management and Economics, Beijing Institute of Technology, Beijing, China, yangzhijun@bit.edu.cn

Abstract

There has been an increasing number of fake e-Business websites created and used, which have resulted in rising financial loss for online consumers and businesses. Therefore, developing effective approaches to detecting phishing websites is essential to mitigating the possibility of being victimized by those sites and minimizing financial loss and risks. In this research, we propose a novel classification model for automatically detecting Chinese phishing e-Business websites. By extending previous research and incorporating unique characteristics of Chinese e-Business websites, our model consists of feature vectors of both the URL and content of a Website. We have trained and evaluated the proposed model with roughly 900 Chinese e-Business websites using four different classification algorithms. Results show that among those four algorithms, the Sequential Minimal Optimization (SMO) algorithm performs the best. To examine the impact of individual features in the model on detection accuracy, we further conducted a sensitivity analysis to identify the most influential features, which helps make the classification model more parsimonious. The findings of this research provide several research and practical insights into the development of anti-phishing solutions.

Keywords: Phishing Websites, E-Business, Classification, Detection, Feature Vectors
1 INTRODUCTION

In the past decade, there has been remarkable growth of e-Business in China. According to the annual report of Anti-Phishing Alliance of China, it is estimated that about 193 million Chinese consumers shopped online in 2012. The convenience of e-Business, however, also bring consumers a variety of security and privacy concerns. Among them, phishing, a form of online identity theft associated with both social engineering and technical subterfuge, is a major threat. There has been an increasing number of phishing e-Business websites created and used to illegally acquire consumers’ personal and sensitive information for financial gains or mislead consumers to conduct business transactions that will never be fulfilled by masquerading a phishing website as a trustworthy e-Business website. Phishing websites have resulted in rising financial loss of online consumers and businesses (Herzberg and Jbara, 2008), posing a considerable threat to not only e-Business in specific, but the Internet security and personal privacy in general as well. According to Anti-Phishing Alliance of China, there were 60 million Chinese online users that became the victims of phishing websites and lost a total of more than 30 billion RMB (about 5 billion US$) between July 2011 and June 2012. Therefore, developing effective approaches to detecting phishing e-Business websites is essential to mitigating the possibility of being victimized by those sites and minimizing the financial loss.

There have been several types of phishing attacks in e-Business. The most common one is to create fake e-Business websites for identity theft. Criminals create fake Websites that look very similar to real, authentic e-Business websites in terms of domain names and Web content. Once the deceived consumers mistakenly log in a phishing site, their user name and password information will be stolen and used by criminals to log in the authentic website for illegal financial gains. Another type of phishing Websites are those without an authentic counterpart. Criminals post fake product sales information on those sites, then disappear after receiving consumers’ money for product purchase. Although the public awareness of phishing Websites has steadily increased over the years, the number of phishing websites and the resulted damage have grown even more at a stunning pace. According to APWG (Anti-Phishing Working Group)’s Phishing Activity Trends Report for the first quarter of 2012\(^1\), there were 56,859 unique phishing sites detected in February alone, which was an all-time high. China continues to be the most affected country and remains the only country with an infection ratio over 50 percent.

Detecting phishing e-Business websites successfully is the major means of anti-phishing. However, it could be quite challenging. So far, there have been a variety of approaches to phishing website detection. Some approaches are based on the recognition of content and URL of a Web page (e.g., Ramanathan and Wechsler 2012; Huang et al. 2012; Gao-hui et al. 2011); some segment a website into images and then analyze those images (e.g., Cao et al., 2009; Zhang et al., 2010; Chen et al., 2010); others use the third-party search engines (e.g., Huh and Kim, 2012). However, those approaches have various limitations, such as reliance on prior knowledge about authentic websites and being domain specific. In addition, some of the features of websites used for phishing detection do not apply to Chinese e-Business websites. For example, some methods examine whether or not the URL of a website contains keywords such as eBay and PayPal. So far, there is little research reported in the literature on developing and empirically evaluating models for the detection of Chinese phishing e-Business websites.

In order to cope with the increasing problem of phishing e-Business websites in China and address some limitations of existing detection modelling approaches, we propose a new classification model for detecting those websites. There are two-fold novel contributions of this study. First, in the proposed model, we incorporate some new factors that reflect the unique characteristics of Chinese e-Business websites in addition to factors adopted or adapted from previous research. Our model neither requires user expertise and prior knowledge of websites, nor consults centralized white or blacklists to determine whether a website is a suspect. Second, we have trained and empirically evaluated the

\(^{1}\) http://anti-phishing.org/reports/apwg_trends_report_q1_2012.pdf
proposed model with websites collected from the real world using four different machine learning algorithms. Furthermore, we have conducted a sensitivity analysis on the influence of individual predictive factors in the model on the detection performance to identify the most influential factors, which has been rarely done in the literature. Such an evaluation approach offers different insights that can help improve the detection model.

The rest of the paper will be organized as follows. Section 2 introduces the related work on automated detection of phishing websites and the limitations of existing approaches. Then, we propose a new model for detection of Chinese phishing websites that integrates features of both the URL and content of Websites in Section 3. Section 4 describes the evaluation of the proposed model, followed by the major findings. Finally, the paper will discuss the implications of the study and conclude the paper in Section 5.

2 RELATED WORK

Phishing websites have become increasingly pervasive, generating billions of dollars in fraudulent revenue at the expense of unsuspecting Internet users. The design and appearance of these websites makes it difficult for users to manually identify them as fake. Automated detection systems have emerged as a mechanism for combating phishing websites. However, existing systems are susceptible to the myriad of obfuscation tactics used by fraudsters, resulting in ineffective detection performance.

Because the most phishing attacks steal users’ sensitive information by masquerading as trustworthy websites, the most basic and widely used detection techniques are based on website analysis. Researchers study the characteristics of the suspicious websites to determine whether the target is a phishing website. Generally speaking, there are a variety of fraud cues that have been examined in prior research on building tools for phishing websites detection. Those cues include those extracted from Web page text (e.g., concocted word phrases, lexical measures such as average sentence length and frequency of long words, selling), extracted from URLs (e.g., HTTPS, number of slashes), extracted from image metadata (e.g., file name, file extension and format) and image pixels (e.g., pixel colors), and hyperlinks (e.g., concocted links, number of in/out links) (Abass et al. 2010). We categorize the existing approaches to phishing Websites detection into four groups based on the focus of website analysis, including blacklist based, visual similarity based, website URL and text features based, and the third-party search engine based, which are described as follows:

- The blacklist based approach relies on a list of URLs of known phishing Websites. If the URL of a target website matches the URL of one of those known phishing websites in the blacklist, then it will be labelled as a phishing website. Although this is the simplest approach, it has severe problems in the lack of capability to detect any new phishing websites.

- The visual similarity based approach (e.g., Chen et al. 2010) treats phishing website detection as an image-matching problem. It normally divides a web page into a number of images. Then, it analyzes and compares the similarity between visual characteristics of those image blocks and those of actual sites registered with an anti-phishing system. This method is motivated by the perception that a website consists of blocks. The characteristics and distribution of those blocks in a website determine the visual characteristics of the website (Cao et al. 2009). Liu et al. (2006)’s system, for example, compares the potential phishing pages against actual pages and assesses block-level visual similarities between them in terms of key regions, page layouts, and overall styles. Chen et al. (2009) propose an effective image-based anti-phishing scheme based on discriminative key point features in Web pages. A point is considered a key point if it can still be detected after the image undergoes various changes, such as shifting, lighting variation, color transformation, or format conversion. Chen et al.’s invariant content descriptor, the Contrast Context Histogram (CCH), computes the similarity degree between suspicious and authentic pages. The obvious constraint with the visual similarity based approach lies in its reliance on comparing a target website against an authentic, real website, which may not always exist (e.g., http://www.51wangdian.net/) or known in advance.
The website URL and text feature based approach (e.g., Huang et al. 2012; He et al. 2011) focuses on the characteristics of the URL and text content of a target website. For example, CANTINA (Zhang et al. 2007) examines the content of a web page to determine whether it is legitimate or not. It makes use of the well-known TF-IDF (term frequency/inverse document frequency) algorithm used in information retrieval, and more specifically, the Robust Hyperlinks algorithm for overcoming broken hyperlinks. Roughly, CANTINA works as follows: Given a web page, it calculates the TF-IDF scores of each term on that web page. Then, it generates a lexical signature by taking five terms with the highest TF-IDF weights. Next, it feeds this lexical signature to a search engine, such as Google. If the domain name of the current web page matches the domain name of the N top search results, the system considers it to be a legitimate website. Otherwise, it will be considered as a phishing site. An assumption of CANTINA is that Google indexes the vast majority of legitimate websites, and those legitimate sites will be ranked higher than phishing sites. Huang et al. (2012) build an SVM (Support Vector Machine) classification model that uses a feature vector including IP address, the length of URL, the number of dots in URL, and the keywords such as eBay and Paypal contained in the URL, etc. Ramanathan and Wechsler (2012) use LDA (Latent Dirichlet Allocation) and AdaBoost to build a classifier for phishing websites detection with normal topics and fake topics identified from content of authentic and phishing websites. Although this is the most commonly explored approach, few existing methods take unique characteristics of URL and content of Chinese e-Business websites into consideration. The limitations of this approach include the difficulty in determining influential URL and text features that can be used as cues for detection. Some cues identified and used in prior research may not be applicable to Chinese e-Business Websites.

Some hybrid methods have attempted to combine the visual similarity based approach and the website URL and text feature based approach. Zhang et al. (2010), for example, synthesizes multiple textual and visual cues from a given web page to detect a phishing web page by using a text classifier, an image classifier, and a data fusion process. They divide the content representation into three categories: surface-level content (domain name, URL, and hyperlinks), textual content (terms), and visual content (e.g., overall style, the layout, and the block regions including the logos, images, and forms). They develop a Bayesian approach to integrate the classification results from textual and visual contents.

Another type of approach relies on the third-party search engines to search for relevant information about a URL and then uses the collected information to make a decision. In Huh and Kim (2012)’s study, the full URL of a target website is used as the search string. The number of results returned and ranking of the website will be used for classification. The assumption of this approach is that legitimate websites should return a large number of results and are ranked first, whereas phishing websites get back no result and/or are not ranked at all. A challenge faced by this type of detection approach is that the designers of phishing websites could use search engine optimization to make a phishing website ranked high among search results.

As discussed above, these existing general approaches to the detection of phishing websites have their strengths and weaknesses. Among them, the website URL and text feature based approach is the most commonly used approach.

3 A CLASSIFICATION MODEL FOR CHINESE PHISHING E-BUSINESS WEBSITE DETECTION

In this research, we aim to develop a new, generic classification model that can detect Chinese phishing e-Business websites by incorporating unique features of Chinese e-Business websites, without reliance on prior human reporting and prior knowledge on available authentic websites. Given the constraints of each detection method discussed in the preceding section, we decide to adopt the website URL and text feature based approach in this study. The model is trained with different machine learning algorithms, aiming to find a better classification model at the end.
3.1 The Feature Vector of the Classification Model

In China, e-Business websites must register their domain name and ICP (Internet Content Provider) Certificate at the Chinese Ministry of Industry and Information Technology. Many of them also receive e-Commerce website certificates from related authorities. By taking those unique features of Chinese e-Business websites into consideration (e.g., domain name inquiry and the examination of Chinese E-Commerce certificates) and incorporating some effective factors validated in the literature that are applicable to Chinese phishing websites (e.g., Zhang et al. 2011; He et al. 2011; Aburrous et al. 2008), we created a feature vector for the proposed model, which consists of two parts: URL features and Web content features, as shown in Figure 1.

**Figure 1. The Feature Vector of the Proposed Model**

URL features refer to the following information items in the URL of a target website, including:

- **F1:** whether or not the URL contains an IP address. For example, a phishing website uses http://110.75.2.128 to replace the URL of official homepage of Taobao.com, the largest C2C website in China. If the URL of a target website contains an IP address instead of a domain name, then the value of this feature F1 is assigned with a 1. Otherwise, it is assigned with a 0.

- **F2:** whether or not the URL contains the symbol “@”. Phishing websites often insert the symbol “@” in the URL to take users to a website different from what users expect. For example, for the URL http://www.taobao.com@www.phishweb.com, the real site to which this address points is not www.taobao.com, but http://www.phishweb.com. In our model, if a URL contains the symbol “@”, F2 is assigned to a value of 1. Otherwise, it is 0.

- **F3:** whether the characters in a URL are coded in the UNICODE. A phishing website may use UNICODE in its URL, such as http://www.taobao.com@%77%77%77%2E%70%68%69%73%68, instead of using http://www.taobao.com@www.phishweb.com, to hide the URL of the truly intended website. In our model, F3 will be assigned a value of 1 if the domain name of the URL of a target website contains characters encoded in the UNICODE. Otherwise, its value will be 0.

- **F4:** the number of dots (“.”) in a URL. In general, the larger number of dots in a URL, the higher possibility that a website is a phishing website (Zhang et al. 2011).
F5: the number of suffixes (e.g., “.com” and “.cn”) in the domain name. http://www.boc.cn.1boc.com.cn is a phishing website URL, in which the number of domain name suffixes is 2. Users generally catch a glimpse of the first part of a URL, but likely miss the remaining part, which actually points to a phishing website.

F6: age of a domain name, which is represented by the number of days since the domain name was registered. The closer the date that a domain name was registered to the current date, the more likely that it is a phishing website.

F7: expiration of the domain name, which is represented by the number of days remaining before a domain name expires. The sooner that a domain name will expire (i.e., the smaller value of F7), the more likely that it is a phishing website.

F8: whether or not the address of DNS (Domain Name System) server is consistent with the URL. The DNS server address of a domain name can be obtained through whois domain name queries. For example, for www.taobao.com, we can get the address of the DNS server includes ns4.taobao.com, ns5.taobao.com, ns6.taobao.com, and ns7.taobao.com, which is consistent with the URL. In our model, if they match, the value of F8 will be 1. Otherwise, it will be 0.

F9: information about website registration. By searching a domain name at the Chinese MII (Ministry of Information Industry) website, a detection system can know whether the domain name is registered, registered by an individual or enterprise, and whether the registered site name and actual site pointed by the URL are consistent. For example, for www.taobao.com, the information retrieved from MII include the following information: “the name of register” is Zhejiang Taobao Network Limited; “register type” is enterprise; “license No.” is B2-200080224-1 Zhejiang; and “site name” is taobao. In our model, we use F9 to represent whether a domain name is registered and recorded (1) or not (0); F10 to represent whether the domain name applicant is an individual (0) or enterprise (1); and F11 to represent whether a recorded website name and actual pointed site are consistent (1) or not (0).

Web content features are extracted from the the source code of a Web page, which include the following:

F12: The ICP (Internet Content Provider) certificate number, which is referred to as the Telecommunications and Information Services business license number issued by the Chinese Regional Communication Administration. A formal Chinese e-Business website normally should provide its ICP certificate numbers at the bottom of the website. The ICP certificate is the license for online business cooperation. By law, any profit-driven online business must apply for and receive its ICP certificate before starting its business. Accordingly, an ICP certificate number is a unique identifier of an online business. For example, the ICP number of www.taobao.com listed on the website is B2-200080224-1 Zhejiang, which matches the archived information retrieved by a domain name query. In our model, we use F12 to represent whether or not an ICP certificate number listed on a website is consistent with the one retrieved by a domain name query. If the answer is yes, then F12 will be assigned with a value of 1. Otherwise, it will be 0.

F13: the number of void (null) links on a website. According to previous studies (e.g., He et al. 2011), a phishing website tends to have more void links than an authentic website.

F14: the number of out links on a website. It is typical that any website has some out links, but when there are too many, it increases the probability of a website being a phishing website.

F15: whether an e-Business website provides electronic commerce certificate information. In general, an authentic business would post images of its electronic commerce certificates at the bottom of its website (See Figure 2). The image is normally a hyperlink that is connected to the certificate information of the business archived by related authorities. At present, there are multiple e-Commerce certificates available, such as the Chinese electronic commerce website demonstration enterprise certificate issued by the Chinese Electronic Commerce Association Digital Service Center, the trusted sites certificate issued by the Chinese Internet Network Information Center, and the online transaction security certificate issued by Chinese Electronic Commerce Association Policy Law Commission. In this study, we take into consideration all the major certificates available. If a website does not provide any certificate link, the value of F15 will be set to 0. Otherwise, it will be 1.
Figure 2. An E-Commerce Certificate Link at the Bottom of An Online Business Website

3.2 The Classification Algorithms

In this study, we have used four different machine learning algorithms to train four classifiers for the detection of Chinese phishing e-Business websites. Those algorithms include Sequential minimal optimization (SMO), logistic regression, Naive Bayes classifier, and the Random Forest method, aiming to find a better classification model. They have been used in the previous studies on building classifiers for phishing website detection. In this subsection, we will briefly introduce these four machine learning approaches for classification.

**Sequential minimal optimization (SMO)** is an algorithm for efficiently solving the optimization problem that arises during the training of support vector machines (Platt 1999; Chang and Lin 2011; Zanni et al. 2006). Consider a binary classification problem with a dataset \((x_1, y_1), \ldots, (x_n, y_n)\), where \(x_i\) is an input vector and \(y_i\) is a binary label. A soft-margin support vector machine is trained by solving a quadratic programming problem, which is expressed in the dual form as follows:

\[
\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j K(x_i, x_j) \alpha_i \alpha_j
\]

subject to: 
- \(0 \leq \alpha_i \leq C\), for \(i = 1, 2, \ldots, n\),
- \(\sum_{i=1}^{n} y_i \alpha_i = 0\)

where \(C\) is an SVM hyperparameter and \(K(x_i, x_j)\) is the kernel function, both provided by the user; and the variables \(\alpha_i\) are Lagrange multipliers. SMO is an iterative algorithm for solving the optimization problem described above. It breaks this problem into a series of smallest possible sub-problems, which are then solved analytically. The algorithm proceeds as follows:
- Find a Lagrange multiplier \(\alpha_1\) that violates the Karush–Kuhn–Tucker (KKT) conditions for the optimization problem.
- Pick a second multiplier \(\alpha_2\) and optimize the pair \((\alpha_1, \alpha_2)\).
- Repeat steps 1 and 2 until convergence.

When all the Lagrange multipliers satisfy the KKT conditions (within a user-defined tolerance), the problem will be solved. Although this algorithm is guaranteed to converge, heuristics are used to choose the pair of multipliers so as to accelerate the rate of convergence.

**Logistic regression (LR)** measures the relationship between a categorical dependent variable and usually a continuous independent variable (or several) by converting the dependent variable to probability scores (Bhandari and Joensson 2009). An explanation of logistic regression begins with an explanation of the logistic function, which always takes on values between zero and one (Hosmer and Stanley 2000):

\[
\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{e^{\beta_0 + \beta_1 x} + 1} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}
\]

\[
g(x) = \ln \frac{\pi(x)}{1 - \pi(x)} = \beta_0 + \beta_1 x
\]

\[
\frac{\pi(x)}{1 - \pi(x)} = e^{(\beta_0 + \beta_1 x)}
\]

The logistic function is useful because it can take any value as an input, from negative infinity to positive infinity, whereas the output is confined to values between 0 and 1. In the above equations, \(g(X)\) refers to the logit function of some given predictor \(X\), \(\ln\) denotes the natural logarithm, \(\pi(x)\) is the probability of being a case, \(\beta_0\) is the intercept from the linear regression equation (the value of the
criterion when the predictor is equal to zero). $\beta_1 x$ is the regression coefficient multiplied by some value of the predictor. Base $e$ denotes the exponential function. The first formula illustrates that the probability of being a case is equal to the odds of the exponential function of the linear regression equation. The second equation illustrates that the logit (i.e., log-odds or natural logarithm of the odds) is equivalent to the linear regression equation. Likewise, the third equation illustrates that the odds of being a case is equivalent to the exponential function of the linear regression equation. This illustrates how the logit serves as a link function between the odds and the linear regression equation.

A **Naive Bayes classifier (NB)** is a probabilistic model that applies Bayes' theorem with strong (naive) independence assumption. Abstractly, a probability model for a classifier is a conditional model over a dependent class variable $C$ with a small number of outcomes or classes, conditional on several feature variables $F_1$ through $F_n$ (Rish 2001), which can be represented as the following:

$$p(C|F_1, ..., F_n) = \frac{p(C)p(F_1, ..., F_n|C)}{p(F_1, ..., F_n)}$$

The "naive" conditional independence assumes that each feature $F_i$ is conditionally independent of every other feature $F_j$ ($j \neq i$) given the class $C$. This means that under the above independence assumptions, the conditional distribution over the class $C$ can be expressed like this:

$$p(C|F_1, ..., F_n) = \frac{1}{Z} p(C) \prod_{i=1}^{n} p(F_i|C)$$

where $Z$ (the evidence) is a scaling factor dependent only on $F_1 ... F_n$, i.e., a constant if the values of the feature variables are known.

**Random forest (RF)** is an ensemble classifier that consists of a number of decision trees and outputs the class that is the mode of the classes generated by individual trees. For classification, a new sample is pushed down the tree. It is assigned the label of the training sample in the terminal node that it ends up. This procedure is iterated over all trees in the ensemble, and the mode vote of all trees is used as the random forest prediction.

### 4 EVALUATION OF THE DETECTION MODEL

#### 4.1 Data Collection

To empirically evaluate the proposed model for the detection of Chinese phishing e-Business websites, we collected 413 phishing websites and 461 normal, authentic websites, all from www.315online.com.cn. Those phishing e-Business websites were originally reported by Chinese consumers to and then validated by 315online, which is a third-party service platform supported by the Policy and Law Committee of Chinese E-Business Association. 315online provides law, credit, security, and other value-added services to e-Business in China. After collecting those websites, we used a tool called WebZIP to download the source code of websites, then developed a crawler program to extract the proposed features from the source code.

We divided collected websites into a training data set and a testing data set for training the classification model. The former consisted of 320 authentic websites and 279 phishing websites, and the latter consisted of 141 authentic websites and 134 phishing websites. We used Weka (Waikato Environment for Knowledge Analysis, http://www.cs.waikato.ac.nz/ml/weka), a data mining tool that integrates a collection of machine learning algorithms, to train the model using the four different classification algorithms, which were introduced in the previous section. The reason that we used four different classification algorithms to train the proposed model is to compare their detection performance and select one that performs the best.
4.2 Evaluation Metrics

We use precision, recall, and F-measure to evaluate the performance of the detection model. For a binary classifier, there are four possible classification results, as shown in Table 1:

<table>
<thead>
<tr>
<th>Prediction</th>
<th>True results</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>True positive (TP)</td>
<td>False positive (FP)</td>
</tr>
<tr>
<td>0</td>
<td>False negative (FN)</td>
</tr>
<tr>
<td>True negative (TN)</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Classification Scenarios

Precision is the percentage of correct detections among all detections, which can be represented by \( \frac{TP}{TP + FP} \). Recall measures the proportion of actual positives in the population being tested, which can be represented by \( \frac{TP}{TP + FN} \). The F-Measure is the harmonic mean of precision and recall, which can be calculated as \( 2 \times \frac{Precision \times Recall}{Precision + Recall} \).

4.3 Comparisons of Detection Performance

Figure 3 shows the comparison of detection performance of SMO, logistic regression, Naive Bayes, and random forest algorithms along evaluation metrics. Among these algorithms, SMO performed the best in all three metrics. Therefore, we chose the model trained by the SMO algorithm, then compared its performance against that of the model recently proposed by Huang et al. (2012) for detecting Chinese phishing e-Business websites.

![Figure 3. Performance Comparison of Four Algorithms](image-url)
Because our main goal is to assess the predictive power of the feature set that we developed, in order to minimize the potential effect of machine learning algorithms on detection performance, we compared the performance of our model against that of Huang et al. (2012)’s model trained by the SMO algorithm. Their model consists of the following predictive features: whether to use IP address in the URL; whether the length of the URL is longer than 22 characters; whether the number of “.” In the URL is more than two; whether the number of dashes in the URL is more than two; whether the URL contains tokens such as HTTP, confirm, banking, secure, ebayisapi, webscr, log in, and sign in, as well as brand names such as PayPal and Master Card. The result shown in Figure 4 demonstrate that using the same machine learning algorithm (i.e., SMO), our model significantly outperforms Huang et al.’s model, indicating that our proposed feature set is considerably more effective for detecting Chinese phishing e-Business websites in comparison to the feature set proposed in Huang et al.’s recent model.

![Figure 4. Performance Comparison of Two Models](Note: CBML refers to our proposed model)

### 4.4 Sensitivity Analysis

We also performed a sensitivity analysis on the developed model, aiming to identify relatively less important features so as to make the model more parsimonious. This has rarely been done in previous studies. This sensitivity analysis was conducted by using ChiSquaredAttributeEval, an attribute evaluation algorithm offered in Weka that evaluates the worth of individual features by computing the value of the chi-squared statistic with respect to the classification.

By applying ChiSquaredAttributeEval and RankSearch, the result shows that features F9 (information about website registration), F15 (e-Commerce certificate), F10 (the type of domain name applicant), F7 (expiration of the domain name), F6 (age of domain name), F14 (the number of out links), and F1 (containing IP address) in our model (in a decreasing order of importance) have significant influence on the detection performance of the model. We then created and tested a new model with the above identified seven important features using the same training and testing website collection described earlier. The testing results demonstrate that although the pruned model only consists of half of the features that the original model has, it achieves equivalent precision, recall, and F-measure values, as shown in Table 2. Among these seven features, the top 3 most important ones (i.e., F9, F15, and F10) have never been explored before in the related literature.

<table>
<thead>
<tr>
<th>Class</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phishing websites</td>
<td>0.969</td>
<td>0.94</td>
<td>0.955</td>
</tr>
<tr>
<td>Real websites</td>
<td>0.945</td>
<td>0.972</td>
<td>0.958</td>
</tr>
</tbody>
</table>

*Table 2. The Performance of the Pruned Model*
5 CONCLUSION

With the rapid growth of e-Business in China, we have witnessed fast emergence of phishing e-Business websites, which have resulted in a huge amount of financial loss to consumers and businesses. Developing effective technological solutions to detecting those websites in a timely manner has become essential. However, there has been scarce research on developing models for detecting Chinese phishing e-Business websites, and existing models developed for the detection of generic phishing websites are not necessarily effective because they do not take specific characteristics of Chinese e-Business websites into consideration.

In this research, we propose a new classification model for detecting Chinese phishing e-Business websites. The model takes into consideration some unique features of Chinese e-Business websites that have never been studied before, combined with a few other generic features that have been validated by previous studies. We build multiple models with four different machine learning algorithms and selected the SMO-based model that produced the best precision, recall, and F-measures. In addition, we have analyzed the significance of individual features of the model to the detection performance, and then pruned the model by removing less important features while achieving the similar performance.

This research provides several research and practical contributions. First, it proposes a new model for phishing website detection with features that have never been studied before. Second, the sensitivity analysis of the model further demonstrates the importance of new features that we integrated in the model, and enables us to simplify the model. It also shows that some features we adopted from the literature are not very effective for detecting Chinese phishing e-Business websites. These results imply that when developing solutions for the detection of specific phishing websites such as e-Business websites, we must take the context and characteristics of those websites into consideration. Third, although the model is built for Chinese phishing e-Business websites detection and may not be directly applicable to other phishing websites, the insights gained and ideas behind the approach should be helpful to researchers and practitioners.

ACKNOWLEDGEMENTS

This research is supported by the National Natural Science Foundation of China (Award #: 71128003, 70972006, 71102111) and Program for New Century Excellent Talents in University (Award #: NCET-11-0792).

REFERENCES


