Published in IET Information Security
Received on 7th March 2013
Revised on 13th July 2013
Accepted on 21st July 2013
doi: 10.1049/iet-ifs.2013.0202

ISSN 1751-8709

Intelligent rule-based phishing websites classification

Rami M. Mohammad[1], *Fadi Thabtah*[2], *Lee McCluskey*[1]
[1]*School of Computing and Engineering, University of Huddersfield, Huddersfield, UK*
[2]*School of MIS, Philadelphia University, Amman, Jordan*
E-mail: rami.mohammad@hud.ac.uk

Abstract: Phishing is described as the art of echoing a website of a creditable firm intending to grab user's private information such as usernames, passwords and social security number. Phishing websites comprise a variety of cues within its content-parts as well as the browser-based security indicators provided along with the website. Several solutions have been proposed to tackle phishing. Nevertheless, there is no single magic bullet that can solve this threat radically. One of the promising techniques that can be employed in predicting phishing attacks is based on data mining, particularly the 'induction of classification rules' since anti-phishing solutions aim to predict the website class accurately and that exactly matches the data mining classification technique goals. In this study, the authors shed light on the important features that distinguish phishing websites from legitimate ones and assess how good rule-based data mining classification techniques are in predicting phishing websites and which classification technique is proven to be more reliable.

1 Introduction

Phishing attack classically starts by sending an email that seems to come from an honest enterprise to victims asking them to update or confirm their personal information by visiting a link within the email. Although phishers are now employing several techniques in creating phishing websites to fool and allure users, they all use a set of mutual features to create phishing websites because, without those features they lose the advantage of deception. This helps us to differentiate between honest and phishing websites based on the features extracted from the visited website.

Overall, two approaches are employed in identifying phishing websites. The first one is based on blacklists [1], in which the requested URL is compared with those in that list. The downside of this approach is that the blacklist usually cannot cover all phishing websites; since, within seconds, a new fraudulent website is expected to be launched. The second approach is known as heuristic-based method [2], where several features are collected from the website to classify it as either phishy or legitimate. In contrast to the blacklist method, a heuristic-based solution can recognise freshly created phishing websites. The accuracy of the heuristic-based method depends on picking a set of discriminative features that might help in distinguishing the website class [3]. The way in which the features are processed also plays an extensive role in classifying websites accurately. Data mining is one of the research fields that can make use of the features extracted from the websites to find patterns as well as relations among them [4]. Data mining is very important for decision-making since decisions may be made based on the patterns and rules achieved from a data-mining algorithm. Rules are a common representation of data because they are understood easily by humans [5]. Normally, the rule takes the form of the IF–THEN clause, for example, IF condition (s) A THEN class Ω where 'A' holds the value(s) of the feature(s) and it is called the rule-antecedent, and 'Ω' is the predicted class and it is called the rule-consequent. The process of detecting unseen knowledge in datasets is represented in terms of rules and is known as rule-induction. Rule-induction eases decision-making because it generates knowledge that ensures correctness, reliability and completeness [5], as well as reduces the time of knowledge achievement [4]. There are two kinds of rule-induction approaches in datamining; association-rule and classification-rule methods. The use of the classification-rule approach is of concern in this paper. The classification problem goal is to assign each dataset item to one of a predefined class. Several studies were conducted about phishing detection and the features that distinguish phishing websites from legitimate ones, but they all fail in defining precise rules to extract the features as well as defining rules to classify a website as either phishy or legitimate.

This paper differs from previous researches by proposing a group of features that can be extracted automatically using our own software tool. These features are examined in predicting phishing websites using rules derived from different rule-induction algorithms aiming to reduce the false-negative rate that is, classifying phishing websites as legitimate. Moreover, we showed that extracting features automatically is faster than manual extraction, which in turn increases the dataset size and allows us to conduct more experiments; and thus improve the prediction accuracy.

In this paper, we try to answer the following research questions:

1. What are the effective minimal sets of features that can be utilised in predicting phishing?
2. How good are rule-based datamining techniques in predicting phishing websites?
3. Which rule-based classification technique is proven to be more accurate in predicting phishing websites?

This paper is structured as follows: Section 2 defines the phishing problem and the harms it causes. Section 3 discusses related works and highlights different phishing detection methods presented in the literature. Section 4 introduces different phishing features, and grouping them into different categories. Finally, in Sections 7–9, we perform several experiments to measure the significance of the proposed features in detecting phishing websites and evaluate different rule-based classification algorithms for the same purpose. We conclude in Section 10.

2 Problem statement

Phishing websites are fake websites that are generated by dishonest people to impersonate honest websites. Users may be unable to access their emails or sometimes lose money because of phishing. Predicting and stopping this attack is a critical step towards protecting online trading. The accuracy of predicting the type of the website necessarily depends on the extracted features goodness. Since most users feel safe against phishing attacks if they utilise an anti-phishing tool, this throws a great responsibility on the anti-phishing tools to be accurate in predicting phishing.

In this context, we believe that developing rules of thumb to extracting features from websites then utilising them to predict the type of websites is the key to success in this issue.

A report published by 'Gartner' [6], which is a research and advisory company showed that phishing attacks continue to escalate. Gartner estimates that theft through phishing attacks costs US banks and credit card issuers an estimated $2.8 billion annually. The Director of Cisco's security-technology-business-unit said [7], 'Personalised and targeted attacks that focus on gaining access to more lucrative corporate bank accounts and valuable intellectual property are on the rise'.

3 Related work

Although quite a lot of anti-phishing solutions are offered nowadays, most of them are not able to make a high accurate decision thus the false-positive decisions raised intensely.

In this section, we review current anti-phishing methodologies and the features they utilise in developing anti-phishing solutions.

One approach employed in [8], is based on experimentally contrasting associative-classification algorithms, that is, Classification Based Association (CBA), and Multi-class Classification based on Association Rule (MCAR) with other traditional classification algorithms (C4.5, PART etc.). The authors have gathered 27 different features from various websites and then categorised them into 6 criteria as shown in Table 1. To evaluate the selected features, the authors conducted experiments using the following datamining techniques, MCAR [9], CBA [10], C4.5 [11], PRISM [12], PART [4] and JRip [4]. The results showed an important relation between 'Domain-Identity' and 'URL' features. There was insignificant effect of the 'Page Style' on 'Social Human Factor' related features.

Table 1 E-Banking phishing criteria

Category	Phishing factor indicator
URL and domain identity	using IP address
	request URL
	URL of anchor
	DNS record
	abnormal URL
security and encryption	SSL certificate
	certification authority
	abnormal cookie
	distinguished names certificate (DN)
source code and Java Script	redirect pages
	straddling attack
	pharming attack
	using onMouseOver
	server form handler
page style and contents	spelling errors
	copying website
	'submit' button
	using pop-ups windows
	disabling right-click
web address bar	long URL address
	replacing similar characters for URL
	adding prefix or suffix
	using the @ symbol to confuse
	using hexadecimal character codes
social human factor	much emphasis on security and response
	generic salutation
	buying time to access accounts

Later, in 2010 [13], the authors used 27 features to build a model based on fuzzy-logic. Although this is a promising solution, it fails to clarify how the features were extracted from the website, precisely features related to human-factors. Moreover, the rules were established based on human experience, which is one of the problems we aim to resolve in this paper. Furthermore, the website was classified into five different classes that is, (very legitimate, legitimate, suspicious, phishy and very phishy), but the authors did not clarify what is the fine line that differentiates between these classes.

Another method proposed in [14], suggested a new way to detect phishing websites by capturing abnormal behaviours demonstrated by these websites. The authors have selected six structural-features: Abnormal URL, Abnormal DNS record, Abnormal Anchors, Server-Form-Handler, Abnormal cookie, and Abnormal Secure Sockets Layer (SSL)-certificate. Once these features and the website-identity are known, support-vector-machine classifier 'Vapnik's' [15] is used to determine whether the website is phishy or not. The classification accuracy of this method was 84%, which is considered relatively low. However, this method snubs important features that can play a key role in determining the legitimacy of a website.

A method proposed in [16], suggested utilising CANTINA (Carnegie Mellon Anti-phishing and Network Analysis Tool) which is a content-based technique to detect phishing websites using the term-frequency-inverse-document-frequency (TF-IDF) information-retrieval measures [17]. TF-IDF produces weights that assess the term importance to a document, by counting its frequency. CANTINA works as follows:

doi: 10.1049/iet-ifs.2013.0202

1. Calculate the TF-IDF for a given webpage.
2. Take the five highest TF-IDF terms and find the lexical-signature.
3. The lexical-signature is fed into a search engine.

A webpage is considered legitimate if it is included in the N tops searching results. N was set to 30 in the experiments. If the search engine returns a zero result, then the website is labelled as phishy, this point was the main disadvantage of using such a technique since this would increase the false-positive rate. To overcome this weakness, the authors combined TF-IDF with some other features; those are; Age of Domain, Known-Images, Suspicious-URL, IP-Address, Dots in URL and Forms.

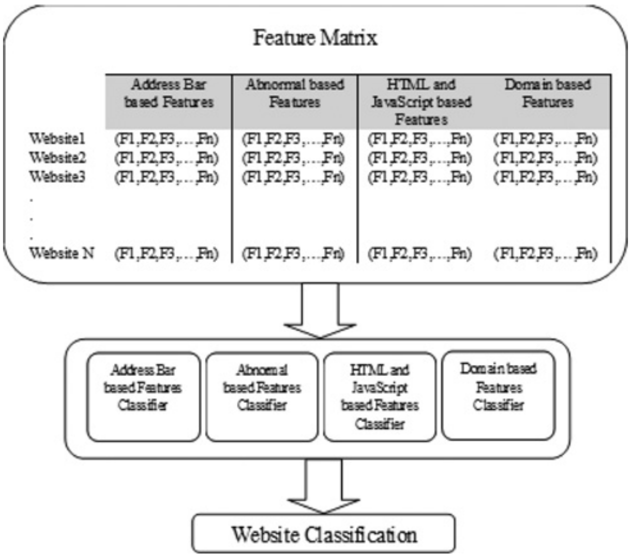
Another approach that utilises CANTINA with an additional attribute and uses different machine-learning algorithms was proposed in [1]. The authors have used 100 phishing websites and 100 legitimate ones in the experiments which is considered limited. The authors have performed three experiments; the first one evaluated a reduced CANTINA feature set that is (dots in the URL, IP-address, suspicious-URL and suspicious-link), and the second experiment involved testing whether the new feature that is (domain top-page similarity), are significant enough to play a key role in detecting website type. The third experiment evaluated the results after adding the new suggested feature to the reduced CANTINA features. The result showed that the new feature plays a key role in detecting the website type. The best accurate algorithm was neural network with error-rate equal to 7.50%, and Naïve Bayes gave the worst result with a 22.5% error-rate.

In [18], the authors compared a number of learning methods including support-vector-machine, rule-based techniques, decision-trees and Bayesian techniques in detecting phishing emails. A random forest algorithm was implemented in PILFER (Phishing Identification by Learning on Features of Email Received). PILFER detected 96% of the phishing emails correctly with a false-positive rate of 0.1%. Ten email's features displayed are used in the experiments, those are IP address URLs, Age of Domain, Non-matching URLs, 'Here' Link, HTML emails, Number of Links, Number of Domains, Number of Dots, Containing Javascript and Spam-filter Output.

4 Categorising phishing features

Going back to the approaches presented in Section 3, and after reviewing the surveys, we were able to derive many features of phishing websites and then categorise them into new groups as shown in Fig. 1.

The features have been categorised based on their effect within a website. Firstly, we examine if a webpage contains any text fields, because a phishing webpage asks users to disclose their personal information through these fields. If the webpage has at least one text field we continue to extract other features. Otherwise, the extraction process is terminated. To measure the feature significance in detecting phishing, we have collected 2500 datasets from the Phishtank [19] and Millersmiles archive [20] using our tool and computed each feature frequency within the dataset in order to reflect the feature importance. In the next section, every feature will be associated with a weight corresponding to the ratio of that feature in the dataset. These frequencies will give us an initial indication of how influential is the feature in a website.


Feature Matrix
Address Bar based Features Classifier
Abnormal based Features Classifier
HTML and JavaScript based Features Classifier
Domain based Features Classifier
Website Classification

	Address Bar based Features	Abnormal based Features	HTML and JavaScript based Features	Domain based Features
Website1	(F1,F2,F3,…,Fn)	(F1,F2,F3,…,Fn)	(F1,F2,F3,…,Fn)	(F1,F2,F3,…,Fn)
Website2	(F1,F2,F3,…,Fn)	(F1,F2,F3,…,Fn)	(F1,F2,F3,…,Fn)	(F1,F2,F3,…,Fn)
Website3	(F1,F2,F3,…,Fn)	(F1,F2,F3,…,Fn)	(F1,F2,F3,…,Fn)	(F1,F2,F3,…,Fn)
.				
.				
.				
Website N	(F1,F2,F3,…,Fn)	(F1,F2,F3,…,Fn)	(F1,F2,F3,…,Fn)	(F1,F2,F3,…,Fn)

Fig. 1 *Proposed phishing prediction hierarchical model*

4.1 Address bar-based features

4.1.1 IP address: If IP address is used as an alternative of a domain name in the URL for example, 125.98.3.123 or it can be transformed to hexadecimal representation for example, http://www.0x58.0xCC.0xCA.0x62, the user can almost be sure someone is trying to steal his personal information. By reviewing our dataset, we find 570 URLs having an IP address which constitutes 22.8% of the dataset. To produce a rule for extracting this feature, we examine the domain part of the URL which lies between '//' and '/', as shown in Fig. 2

$$\text{Proposed Rule: IF}\begin{cases}\text{IP address exists in URL} \rightarrow \text{Phishy}\\ \text{otherwise} \rightarrow \text{feature} = \text{Legitimate}\end{cases}$$

1. *Long URL:* Long URLs are commonly used to hide the doubtful part in the address bar. Scientifically, there is no reliable length which distinguishes phishing URLs from legitimate ones. As in [21], the proposed length of legitimate URLs is 75. However, the authors did not justify the reason behind their value. To ensure the accuracy of our study, we calculated the length of the URLs of the legitimate and phishing websites in our dataset and produced an average URL length. The results showed that if the length of the URL is less than or equal to 54 characters then the URL was classified as 'Legitimate'. On the other hand, if the URL length is greater than 74 characters then the website is 'Phishy'. In our dataset, we found 1220 URLs lengths greater than or equal to 54

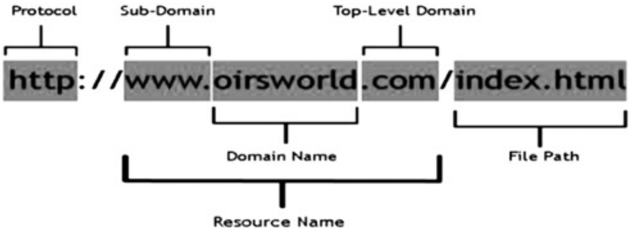


Fig. 2 *URL anatomy*

characters, which constitute 48.8%.

$$\text{Proposed Rule:} \begin{cases} \text{URL length} < 54 \rightarrow \text{Legitimate} \\ \text{URL length} \geq 54 \text{ and} \leq 75 \rightarrow \text{Suspicious} \\ \text{otherwise} \rightarrow \text{Phishy} \end{cases}$$

2. *Using the @ Symbol:* Using the '@' symbol in the URL leads the browser to ignore everything preceding the '@' symbol since the real address often follows the '@' symbol. After reviewing our dataset, we were able to find 90 URLs having '@' symbol, which constitute only 3.6%.

$$\text{Proposed Rule: IF} \begin{cases} \text{URL has @ symbol} \rightarrow \text{Phishy} \\ \text{otherwise} \rightarrow \text{Legitimate} \end{cases}$$

3. *Prefix or suffix separated by* '–'*:* Dashes are rarely used in legitimate domain-names. Phishers resort to add suffixes or prefixes separated by '−' to the domain names so that users feel they are dealing with a legitimate webpage. 661 URLs having the '−' symbol were found in our dataset which constitutes 26.4%.

$$\text{Proposed Rule: IF} \begin{cases} \text{domain part includes } '-' \text{ symbol} \rightarrow \text{Phishy} \\ \text{otherwise} \rightarrow \text{Legitimate} \end{cases}$$

4. *SubDomain and Multi SubDomains:* Assume that we have the following link http://www.hud.ac.uk/students/portal.com. A domain-name always includes the top-level domain, which in our example is 'uk.' The 'ac' part is shorthand for academic, '.ac.uk' is called the second-level domain, and 'hud' is the actual name of the domain. We note that the legitimate URL link has two dots in the URL since we can ignore typing 'www.'. If the number of dots is equal to three then the URL is classified as 'Suspicious' since it has one subdomain. However, if the dots are greater than three it is classified as 'Phishy' since it will have multiple subdomains. Our dataset contains 1109 URLs having three or more dots in the domain part, which constitutes 44.4%. (see equation at the bottom of the page)

5. *HTTPS 'Hyper Text Transfer Protocol with Secure Sockets Layer' and SSL 'Secure Sockets Layer':* Legitimate websites utilise secure domain-names every time sensitive information is transferred. The existence of https is important in giving the impression of website legitimacy, but it is not enough, since in 2005, more than 450 phishing URLs used https recognised by the Netcraft Toolbar Community [22]. Therefore we further check the certificate assigned with https including the extent of trust of the certificate issuer, and the certificate age. Certificate authorities that are consistently listed among the top names for trust include GeoTrust, GoDaddy, Network Solutions, Thawte and VeriSign. By reviewing our dataset, we find that the minimum certificate age for the URLs supporting HTTPs protocol was 2 years. In our dataset, we find 2321 URLs that do not support https or use a fake https, which constitute 92.8%. (see equation at the bottom of the page)

4.2 Abnormal-based features

1. *Request URL:* For legitimate websites, most of the objects within the webpage are linked to the same domain. For example, if the URL typed in the address bar was http://www.hud.ac.uk/students/portal.com we extract the keyword ⟨ src = ⟩ from the webpage source code and check whether the domain in the URL is different from that in the ⟨src⟩, if so, the website is classified as 'Phishy'. To develop a rule for this feature, we calculated the ratio of the URLs in the source code that have different domain than the domain typed in the address bar. By reviewing our dataset, we find that the legitimate websites have in the worst case 22% of its objects loaded from different domains, whereas for the phishing websites the ratio in the best case was 61%. Thus, we assumed that if the ratio is less than 22% then the website is considered 'Legitimate' else if the ratio is between 22 and 61% then the website is considered 'Suspicious'. Otherwise, the website is considered 'Phishy'. In this feature, we computed the rate of this feature existence not the number of feature existence; since the number of request URLs in the website varies. The dataset contains 2500 URLs having this feature, which constitutes 100%. (see equation at the bottom of the page)

2. *URL of anchor:* An anchor is an element defined by the ⟨a⟩ tag. This feature is treated exactly as a 'Request URL'. By reviewing our dataset, we find that the legitimate websites have in the worst case 31% of their anchor-tag connected to different domains, whereas for the phishing websites we find that the ratio was 67% in best case. Thus, we assumed that if the ratio is less than 31% then the website is considered 'Legitimate' else if the ratio is between 31 and 67% then the website is considered 'Suspicious'. Otherwise, the website is considered 'Phishy'. By reviewing our dataset, we find 581 URLs having this feature, which constitutes 23.2%. (see equation at the bottom of the page)

$$\text{Proposed Rule: IF} \begin{cases} \text{dots in the domain part} < 3 \rightarrow \text{Legitimate} \\ \text{else if dots in domain part} = 3 \rightarrow \text{Suspicious} \\ \text{otherwise} \rightarrow \text{feature} = \text{Phishy} \end{cases}$$

$$\text{Proposed Rule: IF} \begin{cases} \text{Use https and trusted issuer and age} \geq 2 \text{ years} \rightarrow \text{Legitimate} \\ \text{using https and issuer is not trusted} \rightarrow \text{Suspicious} \\ \text{otherwise} \rightarrow \text{Phishy} \end{cases}$$

$$\text{Proposed Rule:} \begin{cases} \text{request URL\%} < 22\% \rightarrow \text{Legitimate} \\ \text{request URL\%} \geq 22\% \text{ and } < 61\% \rightarrow \text{Suspicious} \\ \text{otherwise} \rightarrow \text{Phishy} \end{cases}$$

3. *Server form handler (SFH):* SFH that contains an empty string or 'about:blank' are considered doubtful since an action should be taken upon submitted information. In addition, if the domain-name in SFH-s is different from the domain of the webpage this gives an indication that the webpage is suspicious because the submitted information is rarely handled by external domains. In our dataset, we find 101 URLs having SFHs, which constitutes only 4.0%. (see equation at the bottom of the page)

4. *Abnormal URL:* This feature can be extracted from the WHOIS database [19]. For a legitimate website, identity is typically part of its URL. A 412 URLs having this feature were found in our dataset, which constitutes 16.4%. (see equation at the bottom of the page)

4.3 HTML and Javascript based features

1. *Redirect page:* Open redirects found on websites are liable to be exploited by phishers to create a link to their site. In our dataset, we find that the maximum number of redirect pages in the phishing websites was three, whereas this feature is rarely used in legitimate websites since we found only 21 legitimate websites having this feature and it is used for one time only in those websites. Thus, if the redirection number is less than 2 then we will assign 'Legitimate', else if the redirection number is greater than or equal to 2 and less than 4 then we will assign 'Suspicious', otherwise we will assign 'Phishy'. 249 URLs having a redirect-page were encountered in our phishing dataset, which constitutes 10%.

$$\text{Proposed Rule:}\begin{cases} \text{redirect page}\,\# \leq 1 \rightarrow \text{Legitimate} \\ \text{redirect page}\,\# > 1 \text{ and } < 4 \rightarrow \text{Suspicious} \\ \text{otherwise} \rightarrow \text{Phishy} \end{cases}$$

2. *Using onMouseOver to hide the link:* Phishers may use JavaScript to display a fake URL in the status bar to the users. To extract this feature we must explore the webpage source code particularly the 'onmouseover' event and check if it makes any changes to the status bar. A 496 URLs having this feature were found in our dataset, which constitutes 20%.

$$\text{Proposed Rule:}\begin{cases} \text{onmouseover change the status bar} \rightarrow \text{Phishy} \\ \text{it does not change status bar} \rightarrow \text{Suspicious} \\ \text{otherwise} \rightarrow \text{Legitimate} \end{cases}$$

3. *Disabling right click:* Phishers use JavaScript to disable the right-click function, so that the users cannot view and save the source code. This feature is treated exactly as 'Using onMouseOver to hide the Link'. However, for this feature, we will search for event 'event.button = = 2' in the source code and check if right click is disabled. We find this feature 40% times in our dataset, which constitutes 1.6%.

$$\text{Proposed Rule:}\begin{cases} \text{right click disabled} \rightarrow \text{Phishy} \\ \text{right click showing an alert} \rightarrow \text{Suspicious} \\ \text{otherwise} \rightarrow \text{Legitimate} \end{cases}$$

4. *Using pop-up window:* It is unusual to find a legitimate website asking users to submit their credentials through a popup window. A 227 URLs were found in our dataset in which the users credentials were submitted through a popup window, which constitutes 9.1%.

$$\text{Proposed Rule:}\begin{cases} \text{Not using popup} \rightarrow \text{Legitimate} \\ \text{otherwise} \rightarrow \text{feature} = \text{Phishy} \end{cases}$$

4.4 Domain based features

1. *Age of domain:* This feature can be extracted from the WHOIS database [23]. In our dataset, we find that some domains host several phishy URLs in several time slots. The blacklist may succeed in protecting the users if it works on the domain level not on the URL level that is, add the domain-name to the blacklist not the URL address. However, Rasmussen and Aaron [24] find that 78% of phishing domains were in fact hacked domains, which already serve a legitimate website. Thus, blacklisting those domains will in-turn add the legitimate websites to the blacklist as well. Even though the phishing website has moved from the domain, legitimate websites may be left on blacklists for a long time; causing the reputation of the legitimate website or organisation to be harmed. Some blacklists such as 'Google's Blacklist' need on average seven hours to be updated [25]. By reviewing our dataset, we find that the minimum age of the legitimate domain was 6 months. For this feature, if the domain was created in less than 6 months, it is classified as 'Phishy'; otherwise, the website is considered 'Legitimate'. In our dataset, 2392 URLs were created in less than 6 months, which constitute 95.6%.

$$\text{Proposed Rule:}\begin{cases} \text{age of domain is} \geq 6 \text{ months} \rightarrow \text{Legitimate} \\ \text{otherwise} \rightarrow \text{Phishy} \end{cases}$$

2. *DNS Record:* For phishing sites, either the claimed identity is not recognised by the WHOIS database [19] or

$$\text{Proposed Rule:}\begin{cases} \text{URL of anchor\%} < 31\% \rightarrow \text{Legitimate} \\ \text{URL of anchor\%} \geq 31\% \text{ and } \leq 67\% \rightarrow \text{Suspicious} \\ \text{otherwise} \rightarrow \text{Phishy} \end{cases}$$

$$\text{Proposed Rule:}\begin{cases} \text{SFH is } '\text{about: blank}' \text{ or an empty} \rightarrow \text{Phishy} \\ \text{SFH refers to a different domain} \rightarrow \text{Suspicious} \\ \text{otherwise} \rightarrow \text{Legitimate} \end{cases}$$

$$\text{Proposed Rule: IF}\begin{cases} \text{the host name is not included in the URL} \rightarrow \text{Phishy} \\ \text{otherwise} \rightarrow \text{Legitimate} \end{cases}$$

founded cord of the hostname is not found [14]. If the DNS record is empty or not found then the website is classified as 'Phishy', otherwise it is classified as 'Legitimate'. A 160 URLs were found in our dataset where the DNS record is not found, and that constitutes 6.4%.

$$\text{Proposed Rule:}\begin{cases}\text{no DNS record for the domain} \rightarrow \text{Phishing}\\ \text{otherwise} \rightarrow \text{Legitimate}\end{cases}$$

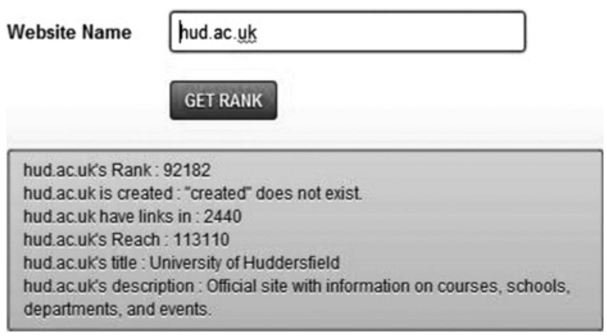
3. *Website traffic:* This feature measures the popularity of the website by determining the number of visitors and the number of pages they visit. However, since phishing websites live for a short period-of-time thus they may not be recognised by the Alexa database [26]. By reviewing our dataset, we find that in the worst-case, legitimate websites ranked among the top 100 000. Therefore if the domain has no traffic or is not being recognised by the Alexa database it is classified as 'Phishy' otherwise if the website ranked among the top 100 000 it is classified as 'Legitimate' else it is classified as 'Suspicious'. This feature constitutes 89.2% of the dataset since it appears 2231 times.

$$\text{Proposed Rule:}\begin{cases}\text{webpage rank} < 100\,000 \rightarrow \text{Legitimate}\\ \text{The Ranking} > 100\,000 \rightarrow \text{Suspicious}\\ \text{otherwise} \rightarrow \text{Phishy}\end{cases}$$

5 Preparing for experiments

To conduct experiments of producing new rules related to phishing, some preparatory steps must be taken as follows:

- *Dataset preparation:* A set of phishing websites was collected from Phishtank [19], which is a free community site where users can submit, verify, track and share phishing data. In addition, we utilised the Millersmiles [20], which is considered a prime source of information about spoof emails and phishing scams. The legitimate websites were collected from yahoo directory [27] and starting point directory [28]. We collected 2500 phishing URLs, and 450 legitimate ones.
- *Address bar features:* A JavaScript program was built to extract all the features related to the address bar.
- *Abnormal based features:* A simple PHP script was developed to extract those features since these features deal with servers and require a connection to external domains such as the WHOIS database [23].
- *HTML and JavaScript based features:* A JavaScript program was built to extract these features.



Website Name hud.ac.uk

hud.ac.uk's Rank : 92182
hud.ac.uk is created : "created" does not exist.
hud.ac.uk have links in : 2440
hud.ac.uk's Reach : 113110
hud.ac.uk's title : University of Huddersfield
hud.ac.uk's description : Official site with information on courses, schools, departments, and events.

Fig. 3 *Result of Alexa Query*

- *Domain based features:* These features can be extracted from the WHOIS database [23], and from Alexa.com [26]. Furthermore, we developed a PHP script to extract these features, as shown in Fig. 3.

6 Features popularity in designing phishing websites

To measure which feature is the most popular in designing phishing websites, we calculated the frequencies for each feature in our datasets, as shown in Table 2. The results showed that the 'request URL' feature is the most popular one since it is existing in all 2500 data elements, followed by 'age of domain' which presented in 2390 data elements, the next popular feature is 'HTTPS and SSL', with similar frequency. The lowest popular feature is 'disabling right click' feature, which appeared only four times, followed by 'URL having @ symbol' which constituted 3.6% of the dataset.

7 Compared rule-based classification algorithms

In this section, we compare different rule-based classification algorithms, each of which utilises a different methodology in producing knowledge. The first algorithm is C4.5 [11], which extracts a decision-tree from a dataset based on information theory. C4.5 utilises a divide-and-conquer methodology to develop decision-trees. The second algorithm is RIPPER [4], which adopts a separate-and-conquer technique. The third algorithm is PRISM [12], which is classified under the covering algorithms family. Finally, we utilise CBA algorithm [10], which is an implementation of the Apriori algorithm [29]. CBA is based on finding the frequent data items by passing over the dataset many times aiming to find a set of items with support value greater than the minimum support threshold. Then, after finding the frequent items, the CBA produces rules for each frequent item which passes the minimum confidence. The support of a rule indicates how frequently the items in the rule's body are inside the training dataset. The confidence of a rule represents its strength and is defined by the probability of both the rule's antecedent and the consequent together in

Table 2 Feature popularity in designing phishing websites

Feature	Frequency, Hz	Percentage, %
using the IP address	570	22.8
long URL	1220	48.8
URL's having @ symbol	90	3.6
adding prefix or suffix separated by (–) to domain	661	26.4
sub domain and multi sub domain	1109	44.4
HTTPS and SSL	2321	92.8
request URL	2500	100
URL of anchor	581	23.2
server form handler	101	4.0
abnormal URL	412	16.4
redirect page	249	10.0
using onMouseOver	496	20.0
disabling right-click	40	1.6
using pop-up window	227	9.1
age of domain	2392	95.6
DNS record	160	6.4
website traffic	2231	89.2

the training dataset divided by the frequency of the rule's antecedent.

8 Experiments

We compare the algorithm's performance for each feature category in our model shown in Fig. 1. We conduct the experiments using the WEKA tool [30], which is an open source datamining application created in Java at the Waikato University [4]. As we mentioned earlier, we have 450 legitimate websites, furthermore, we randomly picked 450 URLs from the phishing dataset, thus, we have 450-phishing websites and 450-legitimate websites in our training dataset. Fig. 4 summarises the prediction error-rate produced by the considered algorithms for each dataset.

The results showed that the C4.5 algorithm outperforms RIPPER, PRISM and CBA in predicting the class for 'Abnormal based dataset', 'Address Bar based dataset' and 'HTML and JavaScript based dataset' where for 'Domain based dataset' C4.5 and RIPPER have the same error-rate. However, by computing the average error-rate for each algorithm, we noted that C4.5 outperforms all algorithms with 5.76% average error-rate, followed by RIPPER with 5.94% average error-rate whereas the highest average error-rate was achieved by PRISM with 21.24%. Overall, the prediction accuracy obtained from all the algorithms considered acceptable and that reflects the goodness of our features in predicting the website class.

9 Reduced features experiments

The work in this section aims to reduce the number of features in order to reduce the runtime of the datamining algorithm, as well as nominating the least number of features that can be used to predict phishing websites. In addition, selecting features may eliminate the noise in features, which occurs whenever there are irrelevant features presented within the training dataset, which in turn causes an increase in the classification errors [17]. A frequently used metric to evaluate features for relevance for the classification task is χ^2 [17]. Fortunately, WEKA facilitates this method. After evaluating the features using χ^2, we find that the best features that may be used to predict phishing websites are:

> 'request URL, age of domain, HTTPS and SSL, website traffic, long URL, subdomain and multi subdomain, adding prefix or suffix separated by (–) to domain, URL of anchor and using the IP address'.

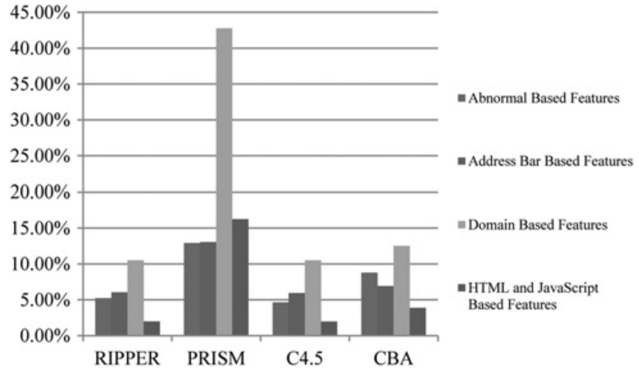


Fig. 4 *Prediction error rate summary*

We assess the prediction accuracy by means of the same classification algorithms used in Section 7. From the results, we noted that even PRISM has a good prediction-rate, knowing that no rule pruning had taken place while producing the rules, which reflects how good these features are in classifying the websites.

However, digging deeply into the rules produced by each algorithm, we noted that all the algorithms generated the following rule:

If HTTPS and SSL = High
And Age of Domain = High
Then Phishing

This reflects the importance of 'HTTPS and SSL' in predicting phishing websites, as well as 'Age of Domain'. Going back to the rules proposed to extract 'HTTPS and SSL', the modification on how to extract this feature was very effective.

10 Conclusions

This paper investigated the features that are effective in detecting phishing websites. These features extracted automatically without any intervention from the users using computerised developed tools. We managed to collect and analyse 17 different features that distinguish phishing websites from legitimate ones. Furthermore, we developed a new rule for each feature. These rules can be useful in applications related to discovering phishing websites based on their features. After performing frequency analysis for each feature, the results showed that 'Request URL' is the most popular feature in creating phishing websites since it appears in all dataset cases, followed by 'Age of Domain', which was presented in the 2392 datasets cased. The next popular feature is 'HTTPS and SSL' with a frequency-rate of 91%. Several experiments have been conducted using different rule-based classification algorithms to extract new hidden knowledge that can help in detecting phishing websites. The experiments showed that the C4.5 algorithm outperformed RIPPER, PRISM and CBA in terms of accuracy. Furthermore, an experiment was conducted after selecting the most effective features in predicting phishing websites. The results showed that we could improve the prediction accuracy relying only on nine features, those are: 'Request URL, Age of Domain, HTTPS and SSL, Website Traffic, Long URL, Sub Domain and Multi Sub Domain, Adding prefix or Suffix Separated by (−) to Domain, URL of Anchor and Using the IP Address'. After conducting the experiments on the nine chosen features, the error-rate has decreased for all the algorithms. Precisely, the CBA algorithm has the lowest error-rate with 4.75%.

In the near future, we will use the rules produced by different algorithms to build a tool that is integrated with a web browser to detect phishing websites in real time and warn the user of any possible attack.

11 References

1 Sanglerdsinlapachai, N., Rungsawang, A.: 'Using domain top-page similarity feature in machine learning-based web'. Third Int. Conf. Knowledge Discovery and Data Mining, 2010, pp. 187–190

2 Sophie, G.P., Gustavo, G.G., Maryline, L.: 'Decisive heuristics to differentiate legitimate from phishing sites'. Proc. 2011 Conf. Network and Information Systems Security, 2011, pp. 1–9

3 Guang, X., Jason, o., Carolyn, P.R., Lorrie, C.: 'CANTINA + : a feature-rich machine learning framework for detecting phishing web sites', *ACM Trans. Inf. Syst. Secur.*, 2011, **14**, pp. 1–28

4 Witten, I.H., Frank, E.: 'Data mining: practical machine learning tools and techniques with Java implementations' (Morgan Kaufmann, New York, NY, USA, 2002)

5 Donald, J.H.: 'Rule induction-machine learning techniques', *Comput. Control Eng. J.*, 1994, **5**, pp. 249–255

6 Gartner, Inc. Available at: http://www.gartner.com/technology/home.jsp

7 Lennon, M. Security Week. Available at: http://www.securityweek.com/cisco-targeted-attacks-cost-organizations-129-billion-annually, 2011

8 Aburrous, M., Hossain, M.A., Dahal, K., Fadi, T.: 'Predicting phishing websites using classification mining techniques'. Proc. Seventh Int. Conf. Information Technology, Las Vegas, Nevada, USA, 2010, pp. 176–181

9 Thabtah, F., Peter, C., Peng, Y.: 'MCAR: multi-class classification based on association rule'. Proc. Third ACS/IEEE Int. Conf. Computer Systems and Applications, 2005, pp. 33

10 Hu, K., Lu, Y., Zhou, L., Shi, C.: 'Integrating classification and association rule mining'. Proc. Fourth Int. Conf. Knowledge Discovery and Data Mining (KDD-98, Plenary Presentation), New York, USA, 1998, pp. 443–447

11 Quinlan, J.R.: 'Improved use of continuous attributes in c4.5', *J. Artif. Intell. Res.*, 1996, **4**, pp. 77–90

12 Cendrowska, J.: 'PRISM: an algorithm for inducing modular rule', *Int. J. Man-Mach. Stud.*, 1987, **27**, pp. 349–370

13 Aburrous, M., Hossain, M.A., Dahal, K., Thabtah, F.: 'Intelligent phishing detection system for e-banking using fuzzy data mining. Expert systems with applications', *An Int. J.*, 2010, **37**, pp. 7913–7921

14 Pan, Y., Ding, X.: 'Anomaly based web phishing page detection'. Proc. 22nd Annual Computer Security Applications Conf. (ACSAC'06), December 2006, , pp. 381–392

15 Cortes, C, Vapnik, V.: 'Support-vector networks', *Mach. Learn.*, 1995, **20**, pp. 273–297

16 Zhang, Y., Hong, J., Cranor, L.: CANTINA: a content-based approach to detect phishing web sites'. Proc. 16th World Wide Web Conf., May, 2007

17 Manning, C., Raghavan, H., Schütze, H.: 'Introduction to Information Retrieval' (Cambridge University Press, 2008)

18 Sadeh, N., Tomasic, A., Fette, I.: 'Learning to detect phishing emails'. Proc. 16th Int. Conf. World Wide Web, 2007, pp. 649–656

19 PhishTank. [Cited 2011 November 25]. Available at: http://www.phishtank.com/, 2006

20 Millersmiles. Millersmiles. [Cited 2011 October]. Available at: http://www.millersmiles.co.uk/, 2011

21 Horng, S.J., Fan, P., Khan, M.K., *et al.*: 'An efficient phishing webpage detector', *Expert Syst. Appl., Int. J.*, 2011, **38**, (10), pp. 12018–12027

22 More than 450 Phishing Attacks Used SSL in 2005. [Cited 2012 March 8]. Available at: http://www.news.netcraft.com/archives/2005/12/28/more_than_450_phishing_attacks_used_ssl_in_2005.html

23 WhoIS. Available at: http://www.who.is/

24 Rasmussen, R., Aaron, G.: 'Global phishing survey: trends and domain name use 2H2009 [Survey]', Lexington, available at: http://www.anti-phishing.org/reports/APWG_GlobalPhishingSurvey_2H2009.pdf, 2010

25 Ask Sucuri. Security Blog. Available at: http://www.blog.sucuri.net/2011/12/ask-sucuri-how-long-it-takes-for-a-site-to-be-removed-from-googles-blacklist-updated.html, 2011

26 Alexa the Web Information Company. [Cited 2012 January 26]. Available at: http://www.alexa.com/

27 Yahoo Directory. Available at: http://www.dir.yahoo.com/

28 Starting Point Directory. Available at: http://www.stpt.com/directory/

29 Agrawal, R., Srikant, R.: 'Fast algorithms for mining association rules'. Proc. 20th Int. Conf. Very Large Data Bases (VLDB'94), 1994, pp. 487–499

30 Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: 'Waikato Environment for Knowledge Analysis', available from: http://www.cs.waikato.ac.nz/ml/weka/

doi: 10.1049/iet-ifs.2013.0202