Feature selection for phishing detection: A review of research

Article *in* International Journal of Intelligent Systems Technologies and Applications · January 2016

DOI: 10.1504/IJISTA.2016.076495

CITATIONS
10

READS
1,038

3 authors:

Hiba Zuhair
Al-Nahrain University
17 PUBLICATIONS **71** CITATIONS
SEE PROFILE

Ali Selamat
Universiti Teknologi Malaysia
399 PUBLICATIONS **2,437** CITATIONS
SEE PROFILE

Mazleena Salleh
Universiti Teknologi Malaysia
109 PUBLICATIONS **682** CITATIONS
SEE PROFILE

Feature selection for phishing detection: a review of research

Hiba Zuhair*

Faculty of Computing,
Department of Computer Science,
Universiti Teknologi Malaysia (UTM),
Johor, 81310, Malaysia
Email: hiba.zuhair.pcs2013@gmail.com
*Corresponding author

Ali Selamat

Faculty of Computing,
Centre for Information and Communication Technologies,
and Software Engineering Department,
Universiti Teknologi Malaysia (UTM),
Johor, 81310, Malaysia
Email: aselamat@utm.my

Mazleena Salleh

Faculty of Computing,
Department of Computer Science,
Universiti Teknologi Malaysia (UTM),
Johor, 81310, Malaysia
Email: mazleena@utm.my

Abstract: Web services motivate phishers to evolve more deceptive websites as their never-ending threats to users. This intricate challenge enforces researchers to develop more proficient phishing detection approaches that incorporate hybrid features, machine learning classifiers, and feature selection methods. However, these detection approaches remain incompetent in classification performance over the vast web. This is attributed to the limited selection of the best features from the massive number of hybrid ones, and to the variant outcomes of applied feature selection methods in the realistic condition. In this topic, this paper surveys prominent researches, highlights their limitations, and emphasises on how they could be improved to escalate detection performance. This survey restates additional peculiarities to promote certain facets of the current research trend with the hope to help researchers on how to develop detection approaches and obtain the best quality outcomes of feature selection.

Keywords: feature selection; feature selection peculiarities; hybrid features; phishing detection.

Reference to this paper should be made as follows: Zuhair, H., Selamat, A. and Salleh, M. (2016) 'Feature selection for phishing detection: a review of research', *Int. J. Intelligent Systems Technologies and Applications*, Vol. 15, No. 2, pp.147–162.

Biographical notes: Hiba Zuhair is currently a PhD candidate in the Department of Computer Science at Faculty of Computing, Universiti Teknologi Malaysia, Johor, Malaysia. During this time, she works as a Senior Lecturer at Al-Nahrain University, Baghdad, Iraq. Prior to this, she received her Master's degree in Computer Science with high distinction from College of Science, Al-Nahrain University, Baghdad, Iraq. She holds a Bachelor of Science in Computer Science from Al-Nahrain University, Baghdad, Iraq. Her research interests and prior publications include: cyber-security, steganography, cryptography, biometrics, image processing and computer vision, wireless and computer networks and web development.

Ali Selamat is currently a Chief Information Officer & Director of Centre for Information and Communication Technologies, Universiti Teknologi Malaysia (UTM), Malaysia. He is also a Professor in Software Engineering Department at Faculty of Computing, UTM. He is nominated as the Vice Chair of IEEE Computer Society Malaysia since 2014. His research interests and publications include software engineering, software process improvement, software agents, web engineering, information retrievals, pattern recognition, genetic algorithms, neural networks and soft computing, computational collective intelligence, strategic management, key performance indicator and knowledge management.

Mazleena Salleh is currently an Associate Professor at Universiti Teknologi Malaysia, lecturing under Department of Computer Science, Faculty of Computing. She received her PhD in Computer Science at Universiti Teknologi Malaysia (UTM) in Computer Networking while her Master's from Virginia Polytechnic State University (USA) in Electrical Engineering. She has published several journal and conference papers related to her research works that include watermarking, steganography, chaos image encryption, networking analysis, e-learning and knowledge management. Her current research is on computer security-related issues namely data survivability and availability in cloud, privacy preserving in cloud environment, elliptic curve cryptography and detection of misuse in computer forensic.

This paper is a revised and expanded version of a paper entitled 'Feature selection for phishing detection: a review of research' presented at *1st ICRIL-International Conference on Innovation in Science and Technology (IICIST2015)*, Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia, 20 April, 2015.

1 Introduction

Phishing is a form of online fraud used to acquire the internet users' confidential and personal information through identification theft for financial gains. In this situation, phishers exploit spoofed technology to target software vulnerabilities and social engineering technology to deceive users on giving their own credentials via communication channels (He et al., 2011; Khonji et al., 2013; Wardman et al., 2014;

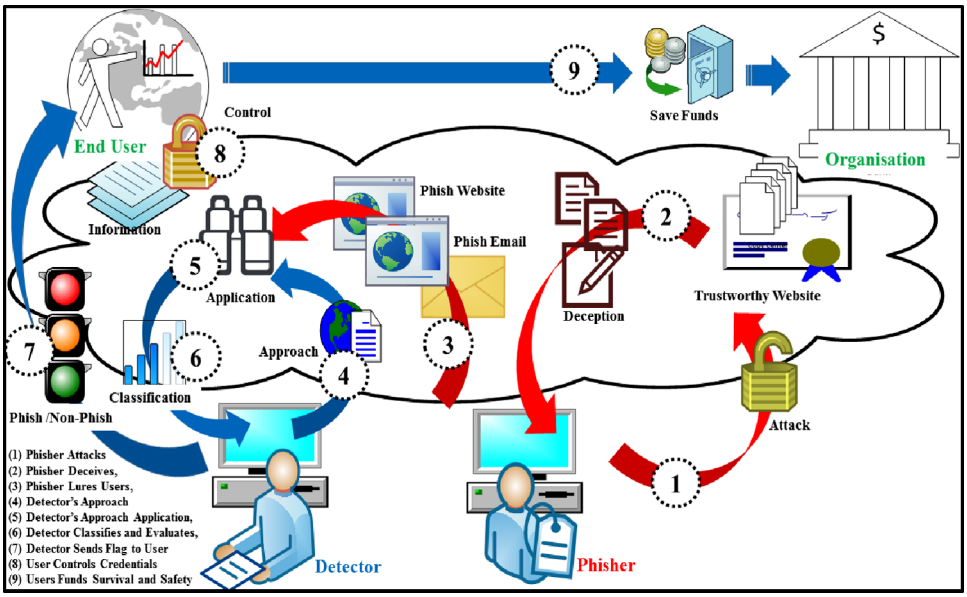
Whittaker et al., 2010). The phisher replicates a legitimate website to create the phishing one. Next, the phisher embeds a spoofed link in the visited website or delivers a phishing email to mislead the user. Then, the user catches the bait and submits his or her own credentials the phishing website. Lastly, the phisher exploits the acquired credentials to carry out digital identity theft and get financial profits (Huang et al., 2009; Mayuri and Tech, 2012). Motivating by the illegitimate gains, phishers continually evolve their attacks with more sophisticated deceptions to put users at more risks of identity theft, and the targeted industries at more reputation and monetary losses (Khonji et al., 2013; San Martino and Perramon, 2010). Undoubtedly, such swift increasing of phishing attacks and their consequences on security and economy has attracted the focused attention in both academia and industry (Almomani et al., 2013; Purkait, 2012). In academia, many efforts have been made by researchers towards obtaining an effective phishing mitigation (Almomani et al., 2013; Khonji et al., 2013; Purkait, 2012). Among them are hybrid detection approaches, which rely on hybrid features, classification techniques and features selection methods to achieve an accurate detection with least false detections in real world experience (Fahmy and Ghoneim, 2011; Gowtham and Krishnamurthi, 2014a, 2014b; Islam and Abawajy, 2013; Xiang et al., 2011). However, these approaches might be affected by the quality of feature selection output owing to the hybridity of features, the heterogeneity of feature values, and the irrelevant and redundant features that the dataset of either emails or webpages might contain (Basnet and Sung, 2012; Mohammad et al., 2012; Toolan and Carthy, 2010). So far, these issues have posed the needs to more storage, long time of execution, complicated computations, external resources and more effective classification techniques (Basnet and Sung, 2012; Fahmy and Ghoneim, 2011; Gowtham and Krishnamurthi, 2014a, 2014b; Islam and Abawajy, 2013; Mohammad et al., 2012; Toolan and Carthy, 2010; Xiang et al., 2011). Therefore, hybrid detection approaches still sub-optimally perform specifically whenever they detect evolving phishing attacks among hundreds billions of webpages and emails over the web. That, in turn, becomes a more intricate challenge because of the strongly inter-related emails and webpages with a large number of features (Uzun et al., 2013). In the light of hybrid phishing detection topic, this paper aims to highlight the current state of feature selection in the relevant research and reveal the causality between their outstanding limitations and phishing detection performance. To do so, a broad review of research and critical appraisal is conducted and introduced in this paper. On the basis of the insights gained from the conducted review, this study emphasises that the choice of the best features or best feature subset for phishing detection can be inspired by specific peculiarities significantly. Thus, it recommends additional peculiarities to serve as optimality criteria for feature selection's improvement and feature selection outcome's qualification through several facets.

To point out the aforesaid issues, this paper is organised as follows. Section 2 presents a global view on phishing detection in academia and industry. Section 3 introduces the preliminaries of feature selection methods in use. Whereas, Section 4 overviews and critically appraises the relevant literatures of phishing detection approaches assisted by feature selection methods. On the basis of the insights gained from both review and critical appraisal, Section 5 suggests specific peculiarities, summarises their related preliminaries and introduces the contributions that they could give to phishing detection through certain research facets. Lastly, Section 6 draws concluding remarks and discusses the future directions.

2 Phishing detection approaches

In phishing detection domain, various phishing detection approaches have been proposed, developed and achieved to mitigate the swift increase and the continuous evolution of phishing emails and websites. In general, such phishing detection approaches vary in terms of the features that they employ to characterise phishing attacks, and the scenarios that they achieve to keep track phishing activities. As adopted in Islam and Abawajy (2013), phishing detection approaches fall into non-classification and classification approaches owing to the use of machine learning classification techniques and features variations. For instance, non-classification approaches involve white lists of famous trustworthy URLs, black lists of phishing URLs, heuristics detection and information flow detection approaches. Whereas, classification approaches encompass hybridity of machine learning or data mining techniques along with the use of enormous features.

Figure 1 Global view on phishing detection based on filtering strategies (see online version for colours)



Source: Adopted in Cranor et al. (2007), Huang et al. (2009), San Martino and Perramon (2010), Sheng et al. (2009), Wardman et al. (2014)

Given that classification approaches outperform their competitors in phishing detection domain, most researchers and web developers have adopted them to develop intuitive phishing filters on the client side. These filters are referred to either independent web applications, or embedded settings in the popular anti-virus software, or plug-ins or toolbars integrated with web browsers (Cranor et al., 2007; Sheng et al., 2009). Moreover, these filters often keep track the users' interactions and their activities online. Then, those filters warn users intuitively against phish attacks whenever they encounter phishing deceptions on a received email or a visited website during browsing (Cranor et al., 2007; Huang et al., 2009; San Martino and Perramon, 2010; Sheng et al., 2009) as seen in Figure 1. As such initiative, phishing filters help to inspect phishing attacks and

phishers' activities on the cyberspace and mitigate their environmental effects on cyber security (Dhamija et al., 2006; San Martino and Perramon, 2010). Despite this, the evolution of phishing attacks and their deceptions are vividly escalating via the web recently. In addition, the existing phishing detection approaches and automotive phishing filters still represent the unideal solution. This is owing to the limited scenarios and operational parameters in use and the scarcity of several detective factors. Therefore, further optimisations are demanded for the survival of phishing detection approaches and automotive filters against escalation and evolution of phishing attacks (Alkhozae and Batarfi, 2011; Cranor et al., 2007; Dhamija et al., 2006; Sheng et al., 2009). In this context, our communication attempts to highlight the current state and the outstanding issues of some academic achievements on phishing detection domain; namely, those of hybrid detection approaches.

3 Feature selection

Mainly, the feature selection method aims at reducing the feature space dimensionality and enhancing the compactness of features by exploring the most contributing features in order to eliminate the less contributing ones. In the hybrid phishing detection, feature selection has been an active field of research owing to the curse of high dimensional web data (emails or websites), the existence of many irrelevant features and redundant in the examined web data, and less comprehensive and less effective machine learning classifiers against phishing evolution (Fahmy and Ghoneim, 2011; Gowtham and Krishnamurthi, 2014a, 2014b; Islam and Abawajy, 2013; Xiang et al., 2011). For such key challenges, different methods of feature selection have been employed in hybrid phishing detection approaches (Basnet and Sung, 2012; Basnet et al., 2012; Hamid and Abawajy, 2011; Olivo et al., 2013; Toolan and Carthy, 2010). In Table 1, examples of the most salient feature selection methods that frequently used in the domain of phishing detection, are briefly described with respect to their search procedure, selection concept, specifics and evaluation criteria.

It is noteworthy to mention that feature selection methods currently in use have shared the same process of selection involving search procedure and evaluation criterion (Chen et al., 2006; Molina et al., 2002). This means that the search procedure often discards or adds one feature against the evaluation criterion. Thus, feature selection methods in use broadly fall into two categories with respect to the search procedure: *filter* and *embedded with classifiers* (Guyon and Elisseeff, 2003; Zhao et al., 2010). Those of the former category, rely on evaluating the features of data without any learning classifier. Whilst, methods of the latter category incorporate a predetermined learning classifier and use its performance for the purpose of features evaluation. On the other hand, both categories of feature selection methods may result in either a selective subset of features or a subset of selective and weighted features owing to the concept of selection that they employ. Accordingly, they fall into two kinds of methods, namely feature subset selection methods, and features weighting methods. Both kinds of feature selection methods differ in the specifics that can be tuned for the search procedure and the evaluation criterion, whereby such specifics include *mutual information*, *dependency*, *consistency*, *distance* and *transformation* (Chen et al., 2006; Guyon and Elisseeff, 2003; Molina et al., 2002; Zhao et al., 2010).

Table 1 Characterisation of feature selection methods

Feature selection method	*Search procedure*	*Selection concept*	*Specifics*	*Evaluation criterion*
Information gain (IG)	Filter	Feature weighting	Information	$\mathrm{IG}(S,a)=\mathrm{Entropy}-\sum_{V\in a}\frac{\lvert S_V\rvert}{\lvert S\rvert}\times\mathrm{Entropy}(S_V)$ (1) where S, S_V, V and a are the collection of instances, a subset of instances with V of a, a relevant value and an attribute, respectively.
Correlation based feature selection (CFS)	Filter	Feature subset selection	Dependence	$p(C=c\mid V_i=v_i)\neq p(C=c)$ (2) where V_i is said to be relevant if there exists some v_i and c for which $p(V_i = v_i) > 0$.
Chi-squared (χ^2)	Filter	Feature weighting	Transformation	$\chi^2=\frac{N\times(AD-CB)^2}{(A+C')\times(B+D)\times(A+B)\times(C+D)}$ (3) where $A = \#(t, c)$, $B = \#(t, \neg c)$, $N = A + B + C + D = \#(\neg t, c)$, $D = \#(\neg t, \neg c)$, and t and $r$are independent parameters.
Wrapper feature selection	Embedded with classifier	Feature subset selection	Accuracy	Greedy search for feature subset in a forward selection and backward elimination of features
Principal component analysis (PCA)	Filter	Feature weighting	Distance	$x_i'=\sum_{k=1}^{d'}a_{k,i}e_k, d'\le d$ (4) where e_k, d', $a_{k,i}$ and x_i' are the eigenvectors corresponding to largest d' vectors for scatter matrix S, the projections of principal components of original vector x_i, and the new vector respectively.

Source: Adopted from Basnet et al. (2012), Chen et al. (2006), Guyon and Elisseeff (2003), Hamid, I.R.A. and Abawajy, J. (2011), Molina et al. (2002), Toolan and Carthy (2010), Zhao et al. (2010)

4 Application of feature selection for phishing detection

Numerous studies have been conducted with the aid of feature selection methods to assess the discriminative features from numerous hybrid features. However, not many of them look at the effects of the assisted feature selection methods and analyse the resultant feature subsets (Basnet and Sung, 2012; Basnet et al., 2012; Hamid and Abawajy, 2011; Olivo et al., 2013; Toolan and Carthy, 2010). Table 2 characterises the related studies briefly in terms of their proposed detective approaches, the assisted feature selection methods and the constructed classifiers. On the basis of Table 2, researchers in Pan and Ding (2006) proposed a phishing detector based on the supervised machine learning (SVM) classifier and extracted both textual and features of the document object model (DOM) from the examined webpages. They employed two major components in the detector, known as an information retrieval strategy to extract textual features and

Chi square (χ^2) test to select the most effective features. Meanwhile, researchers of Ma et al. (2009) experimentally analysed seven webpages and page rank features with the aid of feature weighting method for the phish website classification and deployed two classifiers with varied classification accuracy due to the selected features. In another study (Toolan and Carthy, 2010), researchers evaluated 40 features that were mostly used for both phish and spam email filtering and ranked the most informative among the three datasets through the information gain (IG) analysis. The detection accuracy was different among all the three datasets owing to the selected set of features in the presence machine learning classifier. Alongside, high dimensional feature space including 177 features was adopted in Basnet and Sung (2012) and Basnet et al. (2012). This feature space was extracted from HTML document and URL of websites and analysed to select the best feature subset. Moreover, several subsets of features were selected by using wrapper feature selection (WFS) and correlation feature-based selection (CFS) methods, and they trained over dataset with the aid of two classifiers, namely logistic regression (LR) and random forest (RF). Besides that, the selection of contributing features varied among different feature selection methods and classifiers, causing different detection results. Likewise, researchers of Khonji et al. (2011) enhanced the classification performance by selecting the most effective subset of features among 47 commonly used features in the literature. Apart from that, all IG, WFS and CFS were deployed with classifiers to predict phish emails. However, the classification results differed owing to the used feature selection method and the number of selected features. On the other hand, researchers of Zhuang et al. (2012) developed a detection model consisting of several phases called feature extractor, training phase, ensemble classifier and cluster training. The proposed model relied on extracting hybrid features from webpages and training them by using 10 classifiers built as ensemble feature base classifiers to obtain better detection results. Then, clustering algorithm with the aid of maximum relevance criterion was used to select the most relevant feature subset. Meanwhile, some researchers developed automatic detection approach for Chinese e-business websites in Zhang et al. (2014). The proposed approach was incorporated with unique features extracted from URL and contents of website. Then, the extracted features were trained and tested by using four classifiers like sequential minimum optimisation (SMO), logistic regression (LR), Naïve Bays (NB) and Random Forest (RF). The study proceeded with evaluating features with the aid of Chi-squared statistic criterion with respect to the used classifier.

A study conducted in Hamid and Abawajy (2014), proposed a multi-tier detector for phish emails by using Adaboost and SMO classifiers, which was built in an ensemble design. Furthermore, they suggested clustering strategy to weigh features with the use of information gain and profiling the highest features for phish email filtering. When they tested their proposal profile among three large scale datasets, they encountered some critical limitations including large size dataset, the limit of cluster size and some error rates. Other researchers determined a minimal set of features to reduce the high dimensionality of feature space and improve the detection rate for phishing email filtering in Qabajeh and Thabtah (2014). They deployed information gain (IG), Chi square (χ^2) and CFS on a set of 47 features collected from previous studies. In addition, they utilised data decision tree (DT) and some rule-based classification algorithms for learning their constructed classifiers. On the other hand, an embedded feature selection algorithm was adopted in Hassan (2015) and it was applied on four groups of features extracted from phishing websites. Each group was learned and analysed separately to omit redundant and irrelevant features. The adopted selection algorithm was examined with four classifiers including SMO, SVM, Decision Tree (DT) and Naïve Bays (NB).

However, several problems like different selective outputs at a given time, and variation in classification performance were encountered and emphasised the classification dependency on the leaning classifiers and the datasets.

Table 2 Application of feature selection for phishing detection – a taxonomy

Citation	*Feature selection method*	*Classifier*	*Related limitations*
Pan and Ding (2006)	χ^2	SVM	• Heterogeneous features in their values • Dependent feature selection outputs on training datasets • High computational time and cost • Relevance and Redundancy problems
Ma et al. (2009)	IG	C4.5	• Heterogeneous features in their values • Low dimensional feature space (7 features) • Redundancy problem
Toolan and Carthy (2010)	IG	C5.0	• Large & imbalanced dataset contains spam, ham and phish • High dimensional hybrid feature space (40 features) • High computational cost • Dependent feature selection outputs on training datasets • Redundant and irrelevant features
Basnet and Sung (2012), Basnet et al. (2012)	CFS, WFS	LR, RF, C4.5	• High computational time and cost • High dimensional hybrid feature space (177 features) • Greedy selection of feature subset • Dependent selected subsets behaviour on classifiers • Redundant and irrelevant features
Khonji et al. (2011)	IG, WFS, CFS	RF	• Dependent classifier's performance on selected features • High dimensional hybrid feature space (47 features) • Problem of scalability to more challenging dataset
Zhuang et al. (2012)	Max relevance	SVM	• Complex computation • Problem of redundant features • Problem of resilient classification in real world application • Redundant features
Zhang et al. (2014)	χ^2	SMO, LR, NB	• Problem of redundant features • Problem of resilient classification in real world application • Problem of irrelevant and redundant features
Hamid and Abawajy (2014)	IG	AdaBoost, SMO	• Heterogeneous values of features • Problem of non-scalable training dataset • Time and resources consumption • Dependent feature selection outputs on training datasets and classifiers • Redundant and irrelevant features

Table 2 Application of feature selection for phishing detection – a taxonomy (continued)

Citation	*Feature selection method*	*Classifier*	*Related limitations*
Qabajeh and Thabtah (2014)	CFS, IG, χ^2	DT	• Dependent feature selection outputs on training datasets and classifiers • Redundant and irrelevant features
Hassan (2015)	Hybrid algorithm	DT, NB, SMO, SVM	• Heterogeneous values of features • Dependent feature selection outputs on training datasets and classifiers • Greedy selection of feature subset • Variations in classification performance • Problem of scalability to more challenging dataset

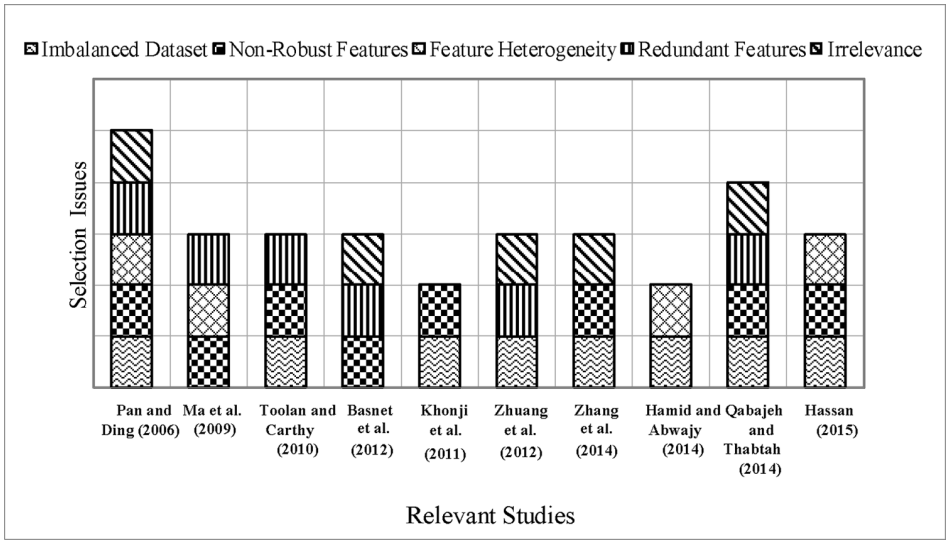
Where IG, WFS, CFS, χ^2, LR, RF, SVM, SMO, NB refer to information gain, wrapper feature selection, correlation based feature selection, chi-squared, logistic regression, random forest, supervised machine learning, sequential minimum optimisation and naïve Bayes.

The aforementioned studies acknowledged that more effective detection approaches can be attained by feature selection methods to rank the most predictive features or select the most predictive feature subsets among a massive number of available ones. However, they suffered from several limitations as those depicted in Table 2 which cause an intricate challenge to deal with the web data in realistic application. In greater detail, these limitations are categorised and discussed as follows:

- *Large and imbalanced datasets*. As presented in Table 2, large and imbalanced datasets may contain different categories of data to be classified including ham, spam, phish and legitimate instances. Besides, they may contain different amount and abundance of common features, which yield complex computations and non-optimistic classification accuracy. Consequently, learning such data in real-world conditions yields that a substantial proportion of one-class samples could be relative to other classes such that it could be non-reflective to their abundance (Chen et al., 2006; Gowtham and Krishnamurthi, 2014a; He et al., 2011; Islam and Abawajy, 2013; Mandal and Mukhopadhyay, 2013; Olivo et al., 2013; Toolan and Carthy, 2010). This problem, in turn, causes the misclassification of less important instances by the classification model. Thus, it leads to the selection of unsuitable features for predicting the less abundant class.
- *Heterogeneity of feature values*. Heterogeneity of feature values. Heterogeneous values of features include those of categorical, continuous and mixed values. Such features are varied among classes over training and testing datasets. To deal with heterogeneous values, proper strategies of clustering, learning, hybrid feature selection, discretisation and normalisation; are required for evaluating the discriminating power of such features among given data (Komiyama et al., 2010; Kudo and Sklansky, 2000; Kwon and Sim, 2013).

- *Resilient features selection outcomes*. The outcomes of feature selection methods against emerging streams of web data could contribute minor misclassification cost and give rise to the classification accuracy if they are being resilient (Fahad et al., 2013; Kudo and Sklansky, 2000; Kwon and Sim, 2013; Xiang et al., 2011).
- *Relevance and redundancy*. Capturing, processing and searching on hybrid feature space for selection must be undertaken to avoid irrelevance and redundancy problems. For instance, irrelevance and redundancy are caused by the large amount of overlapping features between the examined classes; for example, emails of ham, spam and phishing classes (Mandal and Mukhopadhyay, 2013; Zhuang et al., 2012).
- *Computational cost*. The high-dimensional data have a substantial amount of irrelevant and redundant features, which require high computational cost selection strategy to reduce. Such strategy potentially causes inefficient phishing classification, leading to potential instability of classification performance, false detections and errors which may be highly escalated (Guyon and Elisseeff, 2003; Kwon and Sim, 2013; Miyamoto et al., 2008; Molina et al., 2002).
- *Classification performance*. The variations in classification performance across different feature selection methods could be monitored with heavy dependence on different selected features, together with large scale training and testing datasets in realistic situations. Specifically, the applied classification models in phishing detection might produce different outcomes on the same given dataset in the presence of different feature selection methods, which lead to unsatisfactory decisions of classification. Furthermore, the best classification model is that performing well in the presence of other chosen subsets of features and having minimum error rates over a period of time (Basnet and Sung, 2012; Komiyama et al., 2010; Kwon and Sim, 2013; Miyamoto et al., 2008; Olivo et al., 2013; Toolan and Carthy, 2010).

Figure 2 Reviewed studies categorised into their relevant issues of feature selection



On the basis of their absence and their presence in the reviewed researches as plotted in Figure 2, these limitations raise two research questions:

1 How to optimise the existing feature selection methods to resolve the limitations at hands?

2 How to achieve feature selection without deteriorating phishing detection performance?

In the next section, several peculiarities are recommended through two facets of research to answer these two questions.

5 Additional peculiarities

From the insights of review, we can infer that the current state of hybrid phishing detection approaches with respect to feature selection is flourishing; however, these approaches still need more investigations and optimisations in specific facets. Such optimisations could be undertaken with additional peculiarities as those depicted in Table 3. Below, the research facets alongside their related promising peculiarities are discussed:

- *Contribution and quality of feature selection outcomes*. Variation of classification performance could be monitored with the quality of the selected features or feature subset on a large scale datasets in realistic situations. Phishing classification in reality could not be set with minimum error rates when the selected features are relatively non-robust against emerging phishing attacks at a given time. In this context, a set of promoting peculiarities helping to qualify the most useful feature selection method amongst the others for the problem at hand. Peculiarities such as goodness, stability, similarity and performance weighting that presented in Table 3, are recommended to testify empirically the significance that the chosen features or feature subset will give rise to phishing detection. Besides this, large feature space extracted from examined websites or emails are inconsistent in their constituent features in terms of the type of features and their values, their relevance to phishing class, and their redundancy in phishing exploits. With such feature space, training and testing a machine learning classifier effectively requires deploying the most contributing and advantageous feature subset in order to discriminate phishing attacks with minimal false detections. To do so, peculiarities such as maximum relevance and minimum redundancy that briefly described in Table 3, would serve significant criteria to handle the most relevant and least redundant features to phishing class.
- Performance overhead of phishing detection in realistic condition. The feature selection method must result in a representative output that well denoting the phishing attacks and their behaviours (Hamid and Abawajy, 2014; Olivo et al., 2013). To do so, subsequent validations of the selected features or feature subsets must be implemented frequently whenever the examined dataset is modified.

This strategy called profiling as explained in Table 3. Hence, adaptable selection of features will be conducted whenever the current learning dataset is changed either by removing or adding an instance (a change at instance level), or by appending the currently learning feature set with more, new and noisy features (a change at feature level) (Kwon and Sim, 2013; Yin et al., 2013). On the other hand, validating the applied detection approach to restate whether its detection rate of accuracy fits the required rate, was often used in the literature by using cross validation peculiarity (Hamid and Abawajy, 2014; Hassan, 2015; Khonji et al., 2011; Li et al., 2013; Ma et al., 2009; Miyamoto et al., 2008; Pan and Ding, 2006; Qabajeh and Thabtah, 2014; Zhang et al., 2014; Zhuang et al., 2012) which is presented in Table 3. However, the evolutionary and escalating of phishing deceptions and activities over the web require adaptable and effective detection over the past and current streams of web data (Hamid and Abawajy, 2014; Miyamoto et al., 2008). That implies selecting and learning features that extracted from the past data is needed to be relearned over the present aggregated dataset. Alongside cross validation, the chronological assessment peculiarity (Table 3) could help as a validation scenario to testify the performance of phishing detection approach against evolutionary phishing attacks and on an up-to-date data streams (Moskovitch et al., 2009).

Table 3 Suggested peculiarities for feature selection

Metrics	*Advantage*	*Evaluation criterion*
Minimal redundancy (Chen et al., 2006; Mandal and Mukhopadhyay, 2013; Zhao et al., 2010)	It eliminates duplicate features that having another one replicate them in the dataset	$\text{Min } R(S) = \frac{1}{\lvert S\rvert^2} \sum_{x_i, x_j \in S} I(x_i, x_j)$ (5) where $R(S)$ is the set of highest mutually exclusive features that selected between x_i and x_j.
Maximal relevance (Chen et al., 2006; Mandal and Mukhopadhyay, 2013; Zhao et al., 2010)	It selects most relevant features to the target class and highly affecting the classification output	$\text{Max } D(S,c) = \frac{1}{\lvert S\rvert} \sum_{x_i \in S} I(x_i, c)$ (6) where $D(S, c)$ is the mean value of all mutually informative features x_i with respect to class c.
Profiling (Hamid and Abawajy, 2014; Olivo et al., 2013)	Re-evaluating the extracted features to adjust a minimal and effective subset of features that being expected to well denote activities of emerging phishes at any given time	By applying feature subset selection that convolved with search strategy, the best combination of features is learned by a classifier to build "phish profile" for better phishing prediction with accurate detection results. In regard, feature ranking algorithms, feature subset search methods and clustering approach are commonly used for this purpose.
Goodness (Fahad et al., 2013; Kudo and Sklansky, 2000; Kwon and Sim, 2013; Li et al., 2013; Miyamoto et al., 2008; Yin et al., 2013)	It measures how well the selected feature subset can accurately classify extremely imbalanced datasets	$Goodness(S_i) = \frac{1}{Y} \sum_{i=1}^{Y} \frac{N_i^{tp}}{N_i}$ (7) where Y, N_i^{tp} and N_i are the number of classes in the dataset, the number of true positive of each class and the total number of instances for class i respectively.

Table 3 Suggested peculiarities for feature selection (continued)

Metrics	*Advantage*	*Evaluation criterion*
Stability (Fahad et al., 2013; Kudo and Sklansky, 2000; Kwon and Sim, 2013; Li et al., 2013; Miyamoto et al., 2008; Yin et al., 2013)	It quantifiably proves whether the selected features are relatively stable against variations of real world datasets over a period of time	$Stab(S) = \sum_{f_i \in X} \frac{F_{f_i}}{N} \times \frac{F_{f_i} - 1}{\lvert D \rvert - 1}$ (8) where $f_i \in X$ and F_{f_i}/N are all features in a collection dataset S and the relative frequency of each feature in a subset. If all subsets are identical then *Stab*(*S*) is close to 1; otherwise is close to 0.
Similarity (Fahad et al., 2013; Kudo and Sklansky, 2000; Kwon and Sim, 2013; Li et al., 2013; Miyamoto et al., 2008; Yin et al., 2013)	It compares the behaviour of multiple feature selection methods and their selected features on the same data	$Sim(t_1, t_2) = 1 - \frac{1}{2} \sum \left\lvert \frac{F_{f_i}^{t_1}}{N^{t_1}} - \frac{F_{f_i}^{t_2}}{N^{t_2}} \right\rvert$ (9) where $F_{f_i}^{t_1}$ and $F_{f_i}^{t_2}$ denoting the number of frequencies of feature f_i in two candidate feature selection methods t_1 and t_2, respectively. Similarity takes values within [0, 1]."
Performance weighting (Baumann, 2003)	It evaluates whether a feature selection method's results matches the expected solution correctly	$S(x, y) : X \times X \to [0,1]$ s.t. $s(x, y) = 1 \leftrightarrow x = y$ and $s(x, y) = s(y, x)$ (10) where $s(x, y) > s(x, z)$ and y is more similar to x than z. X is the feature set contains relevant, irrelevant and redundant features such that $X = X_R \cup X_I \cup X_{R'}$ and $X' \subseteq X$ is the optimal solution.
Cross validation (Baumann, 2003)	It provides the expected level of accuracy that a classifier fits using selected feature subset over a dataset	$ACC_{CV} = \frac{1}{n} \sum_{(v_i, y_i) \in D} \delta\left(\mathcal{L}\left(D \setminus D_i, v_i\right), y_i\right)$ (11) where D_i indicates the testing dataset of instances $x_i(v_i, y_i)$, and CV is the random number that depends on the division of dataset into folds.
Chronological assessment (Moskovitch et al., 2009)	It experimentally assesses the detection capability of a feature base classifier that learned on old dataset against novel attacks with realistic dataset	This assessment depends on the release date of attacks and the consequently evolved detective approaches for both training and testing tasks. The results indicate whether the detective approach copes with unknown novel attacks and newly aggregated datasets in realistic situation.

6 Concluding remarks and future perspectives

In this review paper, relevant research is critically appraised with the perspective of feature selection methods in hybrid phishing detection. Besides that, reviewed research are characterised and discussed in terms of the machine learning classifiers they used, and the feature selection methods they assisted with. Furthermore, their limitations are emphasised and the effects are categorised in terms of hybridity of features, heterogeneity of values, the tolerance to substantial amount of irrelevant and redundant features, and high dimensional and imbalanced data as well as the detection performance against evolving phishing attacks and the web data. On the other hand, this review revealed that the aforesaid issues could be attained by using some additional peculiarities to enrich the

selection of the best quality features or subset of features. In addition, recommended peculiarities promote the detection with the most relevant and least redundant features, robust selection outcomes, less prone selection to imbalanced datasets and fair dataset dimensionality. In effect, a proficient detection will be achieved with accurate classification, least costs of false detections and errors as well as short runtime, less complicated computations and less storage amount. On the basis of this observation, it is hoped that an overlook to the suggested peculiarities and their significant gains in hybrid phishing detection domain will be taken into account for future research and applications.

Acknowledgement

The authors wish to thank Universiti Teknologi Malaysia (UTM) under Research University Grant Vot-02G31 and Ministry of Higher Education Malaysia (MOHE) under the Fundamental Research Grant Scheme (FRGS Vot-4F551) for the completion of the research.

References

Alkhozae, M.G. and Batarfi, O.A. (2011) 'Phishing websites detection based on phishing characteristics in the webpage source code', *International Journal of Information and Communication Technology Research*, Vol. 1, No. 6, pp.283–291.

Almomani, A., Gupta, B., Atawneh, S., Meulenberg, A. And Almomani, E. (2013) 'A survey of phishing email filtering techniques', *Communications Surveys & Tutorials, IEEE*, Vol. 15, No. 4, pp.2070–2090.

Basnet, R.B. and Sung, A.H. (2012) 'Mining web to detect phishing URLs', *11th International Conference on Machine Learning and Applications (ICMLA)*, Boca Raton, FL, USA, pp.568–573.

Basnet, R.B., Sung, A.H. and Liu, Q. (2012) 'Feature selection for improved phishing detection', *Advanced Research in Applied Artificial Intelligence*, Vol.7345, Lecture Notes in Computer Science, Springer, pp.252–261.

Baumann, K. (2003) 'Cross-validation as the objective function for variable-selection techniques', *TrAC Trends in Analytical Chemistry*, Vol. 22, No. 6, pp.395–406.

Chen, Y., Li, Y., Cheng, X.Q. and Guo, L. (2006) 'Survey and taxonomy of feature selection algorithms in intrusion detection system', *Information Security and Cryptology*, Lecture Notes in Computer Science, Springer Berlin Heidelberg, Vol. 4318, January, pp.153–167.

Cranor, L.F., Egelman, S., Hong, J.I. and Zhang, Y. (2007) 'Phinding Phish: an evaluation of anti-phishing toolbars', *NDSS'04*, San Diego, CA, pp.1–19.

Dhamija, R., Tygar, J.D. and Hearst, M. (2006) 'Why phishing works', *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, University of California, Berkeley, pp.581–590.

Fahad, A., Tari, Z., Khalil, I., Habib, I. and Alnuweiri, H. (2013) 'Toward an efficient and scalable feature selection approach for internet traffic classification', *Computer Networks*, Vol. 57, No. 9, pp.2040–2057.

Fahmy, H.M. and Ghoneim, S.A. (2011) 'PhishBlock: a hybrid anti-phishing tool', Paper presented at the *2011 International Conference on Communications, Computing and Control Applications (CCCA)*, IEEE, pp.1–5.

Gowtham, R. and Krishnamurthi, I. (2014a) 'A comprehensive and efficacious architecture for detecting phishing webpages', *Computers & Security*, Vol. 40, pp.23–37.

Gowtham, R. and Krishnamurthi, I. (2014b) 'PhishTackle – a web services architecture for anti-phishing', *Cluster Computing*, Vol. 17, No. 3, pp.1051–1068.

Guyon, I. and Elisseeff, A. (2003) 'An introduction to variable and feature selection', *The Journal of Machine Learning Research*, Vol. 3, pp.1157–1182.

Hamid, I.R.A. and Abawajy, J. (2011) 'Hybrid feature selection for phishing email detection', *Algorithms and Architectures for Parallel Processing*, Lecture Notes in Computer Science; Vol. 7017, Springer, Berlin, Germany, pp.266–275.

Hamid, I.R.A. and Abawajy, J.H. (2014) 'An approach for profiling phishing activities', *Computers & Security*, Vol. 45, pp.27–41.

Hassan, D. (2015) 'On determining the most effective subset of features for detecting phishing websites', *International Journal of Computer Applications*, Vol. 122, No. 20, pp.1–7.

He, M., Horng, S-J., Fan, P., Khan, M.K., Run, R-S., Lai, J-L. and Sutanto, A. (2011) 'An efficient phishing webpage detector', *Expert Systems with Applications*, Vol. 38, No. 10, pp.12018–12027.

Huang, H., Tan, J. and Liu, L. (2009) 'Countermeasure techniques for deceptive phishing attack', *International Conference on New Trends in Information and Service Science (NISS'09)*, 30 June–02 July, 2009, China, pp.636–641.

Islam, R. and Abawajy, J. (2013) 'A multi-tier phishing detection and filtering approach', *Journal of Network and Computer Applications*, Vol. 36, No. 1, pp.324–335.

Khonji, M., Iraqi, Y. and Jones, A. (2013) 'Phishing detection: a literature survey', *Communications Surveys & Tutorials, IEEE*, Vol. 15, No. 4, pp.2091–2121.

Khonji, M., Jones, A. and Iraqi, Y. (2011) 'A study of feature subset evaluators and feature subset searching methods for phishing classification', *Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS 2011)*, 1–2 September, 2011, ACM, Venice, pp.135–144.

Komiyama, K., Seko, T., Ichinose, Y., Kato, K., Kawano, K. and Yoshiura, H. (2010) 'In-depth evaluation of content-based phishing detection to clarify its strengths and limitations u-and e-service', *Proceedings of 2nd International Conference on e- Service, Science and Technology 2010*, Springer, pp.95–106.

Kudo, M. and Sklansky, J. (2000) 'Comparison of algorithms that select features for pattern classifiers', *Pattern Recognition*, Vol. 33, No. 1, pp.25–41.

Kwon, O. and Sim, J.M. (2013) 'Effects of data set features on the performances of classification algorithms', *Expert Systems with Applications*, Vol. 40, No. 5, pp.1847–1857.

Li, Y., Xiao, R., Feng, J. and Zhao, L. (2013) 'A semi-supervised learning approach for detection of phishing webpages', *Optik-International Journal for Light and Electron Optics*, Vol. 124, No. 23, pp.6027–6033.

Ma, L., Ofoghi, B., Watters, P. and Brown, S. (2009) 'Detecting phishing emails using hybrid features', Paper presented at the *Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing (UIC-ATC'09)*, IEEE, July, Brisbane, Queensland, Australia, pp.493–497.

Mandal, M. and Mukhopadhyay, A. (2013) 'An improved minimum redundancy maximum relevance approach for feature selection in gene expression data', *Procedia Technology*, Vol. 10, pp.20–27.

Mayuri, A. and Tech, M. (2012) 'Phishing detection based on visual-similarity', *International Journal of Scientific and Engineering Research (IJSER)*, Vol. 3, No. 3, March, pp.1–5.

Miyamoto, D., Hazeyama, H. and Kadobayashi, Y. (2008) 'An evaluation of machine learning-based methods for detection of phishing sites', *Aus. J. Intell. Inf. Process Syst*., Vol. 10, No. 2, pp.54–63.

Mohammad, R.M., Thabtah, F. and McCluskey, L. (2012) 'An assessment of features related to phishing websites using an automated technique', *Proceedings of International Conference for Internet Technology and Secured Transactions*, IEEE, pp.492–497.

Molina, L.C., Belanche, L. and Nebot, À. (2002) 'Feature selection algorithms: a survey and experimental evaluation', *Proceedings of Data Mining, ICDM*, IEEE, pp.306–313.

Moskovitch, R., Feher, C. and Elovici, Y. (2009) 'A chronological evaluation of unknown malcode detection', *Pacific Asia Workshop on Intelligence and Security Informatics*, Bangkok, Thailand, Springer, pp.112–117.

Olivo, C.K., Santin, A.O. and Oliveira, L.S. (2013) 'Obtaining the threat model for e-mail phishing', *Applied Soft Computing*, Vol. 13, No. 12, pp.4841–4848.

Pan, Y. and Ding, X. (2006) 'Anomaly based web phishing page detection', *Proceedings of the 22nd Annual Computer Security Applications Conference (ACSAC '06)*, LNCS, Vol. 4186, Springer, Heidelberg, pp.381–392.

Purkait, S. (2012) 'Phishing counter measures and their effectiveness-literature review', *Information Management & Computer Security*, Vol. 20, No. 5, pp.382–420.

Qabajeh, I. and Thabtah, F. (2014) 'An experimental study for assessing email classification attributes using feature selection methods', *3rd International Conference in Advanced Computer Science Applications and Technologies (ACSAT)*, IEEE, December, Amman, Jordan, pp.125–132.

San Martino, A. and Perramon, X. (2010) 'Phishing secrets: history, effects, countermeasures', *IJ Network Security*, Vol. 11, No. 3, pp.163–171.

Sheng, S., Wardman, B., Warner, G., Cranor, L., Hong, J. and Zhang, C. (2009) 'An empirical analysis of phishing blacklists', Paper presented at the *Sixth Conference on Email and Anti-Spam (CEAS 2009)*, Mountain View, California USA, pp.94–100.

Toolan, F. and Carthy, J. (2010) 'Feature selection for spam and phishing detection', Paper presented at *eCrime Researchers Summit (eCrime)*, Dallas, TX, pp.1–9.

Uzun, E., Agun, H.V. and Yerlikaya, T. (2013) 'A hybrid approach for extracting informative content from web pages', *Information Processing and Management: An International Journal*, Vol. 49, No. 4, pp.928–944.

Wardman, B., Britt, J. and Warner, G. (2014) 'New tackle to catch a phisher', *International Journal of Electronic Security and Digital Forensics*, Vol. 6, No. 1, pp.62–80.

Whittaker, C., Ryner, B. and Nazif, M. (2010) 'Large-scale automatic classification of phishing pages', Paper presented at *17th Annual Networks and Distributed System Security Symposium (NDSS2010)*, March, The Internet Society, San Diego, California, USA.

Xiang, G., Hong, J., Rose, C.P. and Cranor, L. (2011) 'Cantina+: a feature-rich machine learning framework for detecting phishing web sites', *ACM Transactions on Information and System Security (TISSEC)*, Vol. 14, No. 2, p.21.

Yin, L., Ge, Y., Xiao, K., Wang, X. and Quan, X. (2013) 'Feature selection for high-dimensional imbalanced data', *Neurocomputing*, Vol. 105, pp.3–11.

Zhang, D., Yan, Z., Jiang, H. and Kim, T. (2014) 'A domain-feature enhanced classification model for the detection of Chinese phishing e-Business websites', *Information & Management*, Vol. 51, No. 7, pp.845–853.

Zhao, Z., Morstatter, F., Sharma, S., Alelyani, S., Anand, A. and Liu, H. (2010) *Advancing Feature Selection Research-ASU Feature Selection Repository*, Technical Report, Computer Science & Engineering, Arizona State University.

Zhuang, W., Jiang, Q. and Xiong, T. (2012) 'An intelligent anti-phishing strategy model for phishing website detection', *Proceedings of 32nd International Conference on Distributed Computing Systems Workshops (ICDCSW 2012)*, IEEE, Macau, June, pp.51–56.