Journal of Computer Science

Original Research Paper

Comparative Study of Data Mining Classification Techniques for Detection and Prediction of Phishing Websites

Luai Al-Shalabi

Information Technology and Computing, Arab Open University, Al-Ardia 92400, Kuwait

Article history
Received: 27-11-2018
Revised: 26-02-2019
Accepted: 18-03-2019

Email: lshalabi@aou.edu.kw

Abstract: Data mining is the process of discovering or extracting information from large amount of data that are stored in databases or datasets such as phishing dataset. Phishing is a vital web security problem that involves simulating legitimate websites to mislead online users in order to steal their sensitive information. This paper aims to detect and predict the type of the website to either legitimate or phishing class label. It investigates different data mining classifiers that are applied to the phishing dataset aiming to determine the effective ones in terms of classification performance. The comparison between nine classifiers with help of rapid miner software was conducted. Here, for comparing the result, five different metrics were used including accuracy, precision, recall, sensitivity and F-Measure. In this study, it has been able to identify the classifiers that precisely recognize fake websites especially with respect to the evolutionary nature of the information attacks.

Keywords: Classification, Website Security, Data Mining, Phishing

Introduction

Currently, phishing is one of the main problems in web security. According to Abdelhamid *et al*. (2014) it is the art of mimicking a legitimate website in order to deceive users by obtaining their sensitive information such as usernames, passwords, national insurance numbers and so on. Emails are the main way of phishing attacks. The attacker usually sends an email to the victim that includes information and a link together with the information. Once the victim clicks on the link, he/she will be forwarded to a forged website that looks like the real one. If the victim proceed with the website and then enter the username and password, this information will directly send to the attacker.

It is a priority of all users and organizations that send and receive information through the websites or any online environment to overcome this problem and minimize the online phishing in order to save their information. As matter fact, preventing this problem is still a big challenging task to the scientists and there is no real solution for that yet. There are always innovative ways that created regularly by phishing attackers to confuse the anti-phishing techniques. Hence, continues demands are essential to come up with intelligent anti-phishing methods that are based on data mining and machine learning (Zuhir *et al*., 2011). Qabajeh and Thabtah (2014) defined phishing detection in data mining context as a typical classification problem. The aim is to predict the type of the website in an automated manner to either accept (legitimate) or reject (phishing) class labels based on a classifier generated from the training data. The training data normally contains websites or website's features with a classification attribute. Some researchers proposed solutions for this problem based on data mining techniques (Abdelhamid *et al*., 2014; Uzun *et al*., 2013; Muhammad *et al*., 2014).

Classification data mining methods were used in the literature to build a satisfied classification model for specific problems such as in Learning (Al-Shalabi, 2016), online shopping (Al-Shalabi, 2018), crime (Al-Shalabi, 2017), medicine (Al-Shalabi, 2009) and so on. In this article, nine data mining classification methods were used for the problem of fishing websites. The comparison of the nine classifiers based on the accuracy, precision, recall, specificity and F-Score was conducted. Classifiers were chosen to represent different kind of modeling. autoMLP classifier belongs to neural net training, KNN classifier belongs to lazy modeling, naïve bayes classifier belongs to bayesian modeling, linear regression belongs to function fitting. Decision tree, random tree and random forest classifiers belong to tree induction and SVM and LibSVM belong to support vector modeling.

The article is organized as follows: next part presents the literature review, followed by the explanation of the dataset used and its features, the methodology which includes the explanation of the five evaluation measures and the nine classifiers used, the

experimental analysis and results of implementing the classification data mining techniques in the phishing training dataset and the comparison between them and then the conclusion of the work.

Phishing: An Overview

Different researchers have been proposed several techniques for solving the phishing problem. Junaid *et al*. (2016) explained that phishing can be reduced through a combination of user and corporate safeguards and server-side measures. User education remains the strongest and at the same time, the weakest link to phishing countermeasures. Kenneth *et al*. (2017) expressed that both perceived vulnerability and perceived net benefit significantly correlate with willingness to pay for an enhanced phishing filter. Kang *et al*. (2018) introduced an enhanced version of the favicon-based phishing attack detection with the introduction of the Domain Name Amplification feature and incorporation of addition features. They proved that additional features are very useful when the website being examined does not have a favicon. Choon and Kang (2017) proposed an anti-phishing technique based on a weighted URL tokens system, which extracts identity keywords from a query webpage. Using the identity keywords as search terms, a search engine is invoked to pinpoint the target domain name, which can be used to determine the legitimacy of the query webpage. Weider *et al*. (2009) presented a Phishing Detection Tool called PhishCatch. It detects phishing e-mails and alarms the user about phishing type e-mails by using heuristic. They tested the algorithm and determined that their proposed tool has a catch rate of 80% which gives an accuracy of 99%. A heuristics called PhishNet was proposed by Prakash *et al*. (2010). They used five heuristics to enumerate simple combinations of known phishing sites in order to discover new phishing URLs. High-Performance Content-Based Phishing Attack Detection was presented by Wardman *et al*. (2011). They implemented a cadre of file matching learning algorithm which is based on the websites content to detect phishing. They experiment their algorithm using a variety of different content-based approaches. They concluded that some can be achieved a detection rate more than 90% by maintaining a low false positive rate. A framework which is based on the Bayesian approach for content-based phishing web page detection was proposed by Jiang *et al*. (2013). They used fusion algorithm which outclasses the individual classifiers. A PageRank Based Detection Technique for Phishing Web Sites was proposed by Naga *et al*. (2012). They tested their proposed technique and showed that around 98% of the tested websites are correctly classified and only 0.02 false positive rate and 0.02 false negative rate are exist. Large-scale Anti-phishing by Retrospective data-eXploration (LARX) is a system proposed by Li *et al*. (2011). It is an offline phishing detection system that uses a network traffic data archived at a vantage point and analyzes the data for phishing detection. A Profiling Phishing E-mail Based on Clustering Approach in which an approach for profiling email-born phishing (ProEP) activities was proposed by Hamid and Abawajy (2013). Their algorithm determines favorable results with the Ratio Size rules for selecting the optimal number of clusters. Hadi *et al*. (2016) proposed a new associative classification algorithm called the Fast Associative Classification Algorithm (FACA). They investigate their proposed algorithm against four well-known AC algorithms (CBA, CMAR, MCAR and ECAR) on real-world phishing datasets. The results indicate that FACA is very successful with regard to the the accuracy and the F1 evaluation measures. Abdelhamid *et al*. (2014) investigated the problem of website phishing using a new proposed multi-label classifier-based associative classification, MCAC. The main goal of the MCAC algorithm developed is to recognize attributes or features that distinguish phishing websites from legitimate ones. The results showed that the MCAC algorithm forecasted phishing websites better than traditional data mining algorithms. Abdelhamid (2015) proposed an enhanced multi-label classifier based associative classification algorithm, eMCAC. This generates rules associated with a set of classes from single-label datasets using the transaction ID list (Tid-list) vertical mining approach. The algorithm employs a novel classifier building method that reduces the number of generated rules. The experiments indicated that the eMCAC algorithm outperformed other algorithms with regard to the accuracy evaluation measure. Zhang *et al*. (2011) presented a novel framework using a bayesian approach for content-based phishing web page detection. Their model takes into account textual and visual contents to measure the similarity between the protected web page and suspicious web pages. A text classifier, an image classifier and an algorithm fusing the results from classifiers were introduced in their model. Experimental results demonstrated that the text classifier and the image classifier they designed deliver promising results. Zhang *et al*. (2007) developed a content based approach known as Carnegie Mellon Anti-Phishing and Network Analysis Tool (CANTINA), for anti-phishing by employing the idea of robust hyperlinks. In this method first calculate the TF-IDF of each web page which is an algorithm usually used for information retrieval and generates a lexical signature by selecting a few terms. Signature supplies to search engines and then matches the domain name of current web page and several top search result to evaluate a current web page is legitimate or not. Liu *et al*. (2010) proposed the use of Semantic Link Network (SLN) to automatically identify the phishing target of a given webpage. The method works by first finding the associated web pages of the given webpage and then constructing a SLN from all those web pages. A mechanism of reasoning on the SLN is exploited to

identify the phishing target. Moghimi and Varjani (2016) present a new rule-based method to detect phishing attacks in internet banking. Their rule-based method used two novel feature sets, which have been proposed to determine the webpage identity. Their proposed feature sets include four features to evaluate the page resources identity and four features to identify the access protocol of page resource elements. They used approximate string matching algorithms to determine the relationship between the content and the URL of a page in their first proposed feature set. Evaluating of the implemented browser extension indicates that it can detect phishing attacks in internet banking with high accuracy and reliability. Akinyelu and Adewumi (2014) investigated and reported the use of random forest machine learning algorithm in classification of phishing attacks, with the major objective of developing an improved phishing email classifier with better prediction accuracy and fewer numbers of features. Results showed high accuracy classification of fake emails. Miyamoto *et al*. (2008) present the performance of machine learning-based methods for detection of phishing sites. They employed nine machine learning techniques. In their evaluation, they used f_1 measure, error rate and Area under the ROC Curve (AUC) as performance metrics along with their requirements for detection methods. The highest f 1 measure was 0.8581, the lowest error rate was 14.15% and the highest AUC is 0.9342. Mohammad *et al*. (2014a) 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.

The Dataset and the Website Features

In order to understand the comparisons between the classifiers, it is important to understand the dataset contents where those classifiers will be implemented on. This will simplify the understanding of the results as well.

Mohammad *et al*. (2012) shed light on the important features that have proved to be sound and effective in predicting phishing websites. In addition, they propose some new features.

There are vast numbers of features that can be used to recognize fraud websites from authentic ones. Some of these features are IP address, long URL, prefix, @ symbol, redirecting using "//" and others. Mohammad *et al*. (2014b) and Zuhir *et al*. (2011) have studied different webpage features in order to recognize the real from the fake ones. The followings are the 30 important features introduced by Mohammad *et al*. (2018) and the way they are coded in the dataset.

Using the IP Addres: If the domain part has an IP address then the website is phishing otherwise it is legitimate.

Long URL to Hide the Suspicious Part: Phishers can use long URL to hide the doubtful part in the address bar. The author showed that if the length of the URL is greater than or equal 54 characters and less than or equal 75 then the URL is classified as suspicious. If the legth is less than 54 then the URL is legitimate, otherwise it is phishing.

Using URL Shortening Services: URL may be made considerably smaller in length and still lead to the required webpage. It is called "Tiny URL and they are classified as phishing.

URL's having "@" Symbol: If the URL has "@" symbol then it is classified as phishing, otherwise it is legitimate.

Redirecting using "//": If the position of the last occurrence of "//" in the URL is greater than 7 then the URL is classified as phishing, otherwise it is legitimate.

Adding Prefix or Suffix Separated by (-) to the Domain: If the domain name part includes the "-"symbol then the URL is classified as phishing, otherwise it is legitimate.

Sub Domain and Multi Sub Domains: If the domain does not have sub-domain (has only one dot in the domain) then it is legitimate. If the domain has one sub-domain (has two dots in the domain) then suspicious, otherwise (the domain has multiple sub-domains) it is phishing.

HTTPS (Hyper Text Transfer Protocol with Secure Sockets Layer): If https is used and the issuer is trusted and the age of the certificate is greater or equal one year then it is legitimate. If https is used and the issuer is not trusted then it is suspicious, otherwise it is phishing.

Domain Registration Length: If the domain expires in one year or less then it is phishing, otherwise it is legitimate.

Favicon: A favicon is a graphic image associated with a specific webpage which represents a visual reminder of the website identity in the address bar. If the favicon is loaded from a domain other than that shown then it is phishing, otherwise it is legitimate.

Using Non-Standard Port: Several firewalls, Proxy and Network Address Translation (NAT) servers will, by default, block all or most of the ports and only open the ones selected. If all ports are open, phishers can run almost any service they want and as a result, user information is threatened. If only the selected port is opened then the website is legitimate, otherwise it is phishing.

The Existence of "HTTPS" Token in the Domain Part of the URL: If the HTTP token is used in the domain part of the URL then the URL is phishing, otherwise it is legitimate.

Request URL: Request URL examines whether the external objects contained within a webpage such as images, videos and sounds are loaded from another domain. In legitimate webpages, the webpage address and most of objects embedded within the webpage are sharing the same domain. If the percentage of the request URL is between 22% and 61% inclusively then the URL is suspicious. If the percentage of the requested URL is less than 22% then the URL is legitimate, otherwise it is phishing.

URL of Anchor: An anchor is an element defined by the <a> tag. This feature is treated exactly as "Request URL". However, for this feature the followings were examined:

If the <a> tags and the website have different domain names. This is similar to request URL feature.
If the anchor does not link to any webpage, e.g.:

<a href="#">, <a href="#content">, <a href="#skip">, <a href="JavaScript ::void(0)">

If the percentage of the URL of anchor is between 31% and 67% inclusively then the URL is suspicious. If the percentage of the URL of anchor is less than 31% then the URL is legitimate, otherwise it is phishing.

Links in <Meta>, <Script> and <Link> tags: The owners' of the fishing dataset found that it is common for legitimate websites to use <Meta> tags to offer metadata about the HTML document; <Script> tags to create a client side script; and <Link> tags to retrieve other web resources. It is expected that these tags are linked to the same domain of the webpage. If the percentage of the Links in <Meta>, <Script> and <Link> tags is between 17% and 81% inclusively then the website is suspicious. If the percentage of the Links in <Meta>, <Script> and <Link> tags is less than 17% then the website is legitimate, otherwise it is phishing.

Server Form Handler (SFH): If SFH is "about: blank" or is empty then the website is phishing. If the SFH refers to a different domain then the website is suspicious, otherwise it is legitimate.

Submitting Information to Email: If the "mail() or mailto:" is used to submit user information then the website is phishing, otherwise it is legitimate.

Abnormal URL: If the host name is not included in the URL then it is phishing, otherwise it is legitimate.

Website Forwarding: If the number of the redirect pages is greater or equal to 2 and less than 4 then the website is suspicious. If the number of the redirect pages is less or equal to 1 then the website is legitimate, otherwise it is phishing.

Status Bar Customization: Phishers may use JavaScript to show a fake URL in the status bar to users. To extract this feature, the webpage source code must be dig-out, particularly the "onMouseOver" event and check if it makes any changes on the status bar. If onMouseOver even changes the status bar then the website is phishing, otherwise it is legitimate.

Disabling Right Click: Phishers use JavaScript to disable the right-click function, so that users cannot view and save the webpage source code. If right click function is disabled then the website is phishing, otherwise it is legitimate.

Using Pop-up Window: It is unusual to find a legitimate website asking users to submit their personal information through a pop-up window. On the other hand, this feature has been used in some legitimate websites and its main goal is to warn users about fraudulent activities or broadcast a welcome announcement, though no personal information was asked to be filled in through these pop-up windows. If the popup window contains text field then the website is phishing, otherwise it is legitimate.

IFrame Redirection: IFrame is an HTML tag used to display an additional web page into one that is currently shown. Phishers can make use of the "iframe" tag and make it invisible without frame borders. In this regard, phishers make use of the "frameBorder" attribute which causes the browser to render a visual delineation. If iframe is used then the website is phising, otherwise it is legitimate.

Age of Domain: Most phishing websites live for a short period of time. By reviewing the fishing dataset, the owners found that the minimum age of the legitimate domain is 6 months. If the age of the domain is greater or equal to 6 months then the website is legitimate, otherwise it is phishing.

DNS Record: If the DNS record is empty or not found then the website is classified as "Phishing", otherwise it is classified as "Legitimate".

Website Traffic: Phishing websites live for a short period of time. Domain traffic is represented by the number of visitors and the number of pages they visit. The owners of the phishing dataset found that if the traffic is among the top 100,000 then the website is classified as "legitimate". If the domain has no traffic then it is classified as "Phishing", otherwise, it is classified as "Suspicious".

PageRank: PageRank is a value ranging from 0 to 1. PageRank aims to measure how important a webpage is on the Internet. The greater the PageRank value the more important the webpage. In the fishing datasets, the owners found that about 95% of phishing webpages have no PageRank. Moreover, they found that the remaining 5% of phishing webpages may reach a PageRank value up to 0.2. If pageRank is less than 0.2 then the website is phishing, otherwise it is legitimate.

Google Index: This feature examines whether a website is in Google's index or not. When a site is indexed by Google, it is displayed on search results and it is considered legitimate, otherwise it is phishing. Usually, phishing webpages are only accessible for a short period and as a result, many phishing webpages may not be found on the Google index.

Number of Links Pointing to Page: The number of links pointing to the webpage indicates its legitimacy level, even if some links are of the same domain. In the fishing datasets and due to its short life span, the owners found that 98% of phishing dataset items have no links pointing to them so they are considered phishing. On the other hand, legitimate websites have at least 2 external links pointing to them, otherwise they are suspicious.

Statistical-Reports Based Feature: Several parties such as PhishTank and StopBadware formulate numerous

statistical reports on phishing websites at every given period of time. In the fishing dataset, the owners used 2 forms of the top ten statistics from PhishTank: "Top 10 Domains" and "Top 10 IPs". Whereas for "StopBadware", they used "Top 50" IP addresses. If the host belongs to the top phishing IPs or top phishing domains then it is phishing, otherwise it is legitimate.

Methodology

In this section, evaluation measures and classifiers used will be explained.

Important Evaluation Measures

Classifiers are evaluated by many means; one of them is the confusion matrix. Confusion matrix is a good way to show the prediction results clearly and unambiguity. It describes the performance of a classification model. If the researched dataset is a binary dataset with two classification values then the confusion matrix is the best choice. Table 1 shows the confusion matrix which contains information about actual and predicted results.

The abbreviation TP, FN, FP and TN of the confusion matrix cells refers to the following:

- TP (true positive): The number of positive cases that are correctly identified as positive
- FN (false negative): The number of positive cases that are misclassified as negative cases
- FP (false positive): The number of negative cases that are incorrectly identified as positive cases
- TN (true negative): The number of negative cases that are correctly identified as negative cases

It is important to decide which classifier is the best to solve the current problem. Classification accuracy alone is typically not enough information to make this decision. So, different performance metrics were used in this research article in order to test of the robustness of each classifier used. Below is the explanation of these metrics.

Accuracy: is the number of True Positives (TP) and the number of True Negatives (TN) divided by the number of all cases. It represents how close a measurement comes to a true value:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Recall (also known as sensitivity): is the number of True Positives (TP) divided by the number of True Positives (TP) and the number of False Negatives (FN). It describes the accuracy of the positive cases:

$$\text{Re}call = \frac{TP}{TP + FN}$$

Table 1: The confusion matrix

Actual state	Predicted negative	Predicted positive
Negative	TN	FP
Positive	FN	TP

Precision: is the number of True Positives (TP) divided by the number of True Positives (TP) and the number of False Positives (FP):

$$Precision = \frac{TP}{TP + FP}$$

Specificity: is the number of True Negatives (TN) divided by the number of True Negatives (TN) and the number of False Positives (FP). It describes the accuracy of the negative examples:

$$Specificity = \frac{TN}{TN + FP}$$

F-Measure (F1 Score): is a measure of a test's accuracy. The F1 score can be interpreted as a weighted average of the precision and recall:

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision \times Recall}$$

The Classifiers

Classification as a famous data mining supervised learning techniques is used to extract meaningful information from large datasets and can be efficaciously used to predict unknown classes (Ngai *et al*., 2009). The predictive accuracy of the classifier is measured by using the training set and the accuracy of classifier on a given test set is the percentage of test set tuples which are classified correctly. If the accuracy is acceptable, the classifier can be used for future data tuples for which the class label is unknown (Han *et al*., 2012). In this part, the nine different classifiers that were used in this research are explained below.

Decision Trees: They are a supervised learning technique commonly used for tasks like classification, clustering and regression. Each node refers a test on an attribute value. The leaves symbolize classes or class distributions which predict classification models. The branches show coincidences of features, which go to classes. Input to a decision tree is the set of objects described by the set of properties and creates output as yes/no decision, or as one of several different classifications (Aitkenhead, 2008). Decision tree creation involves dividing the training data into root node and leaf node divisions until the entire data set has been analyzed.

Random Forest: It is an ensemble learning method for classification, regression and other tasks, that operate

by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees (Ho, 1995; 1998). Random decision forests correct for decision trees' habit of overfitting to their training set (Hastie *et al*., 2008).

Random Tree: Its operator works exactly like the decision tree operator with one exception: for each split only a random subset of attributes is available. The random tree operator works similar to Quinlan's C4.5 or CART but it selects a random subset of attributes before it is applied. The size of the subset is specified by the subset ratio parameter. Representation of the data as tree has the advantage compared with other approaches of being meaningful and easy to interpret. Mishra and Ratha (2016) studied random tree algorithm for microarray data analysis.

Support Vector Machines (SVM): It is a group of supervised learning methods that can be employed for classification or regression (Ivanciuc, 2007; Zhenzhou, 2012). In a two-class learning task, the SVM goal is to discover the best classification function to differentiate between members of the two classes in the training data. For that purpose, SVM construct a hyper plane or a set of hyper planes in a high or infinite dimensional space for separating dataset and SVM find the best function by maximizing the margin between the two classes.

LibSVM: It is a library for Support Vector Machines (SVMs). Chih-Chung Chang and Chih Jen Lin have been actively developing this package since the year 2000. The goal is to help users to easily apply SVM to their applications. LIBSVM has gained wide popularity in machine learning and many other areas. A typical use of LIBSVM involves two steps: training a data set to obtain a model and second, using the model to predict information of a testing data set (Chih-Chung and Chih-Jen, 2011).

AutoMLP: It is a simple algorithm for both learning rate and size adjustment of neural networks during training. The algorithm combines ideas from genetic algorithms and stochastic optimization. It maintains a small ensemble of networks that are trained in parallel with different rates and different numbers of hidden units. After a small, fixed number of epochs, the error rate is determined on a validation set and the worst performers are replaced with copies of the best networks, modified to have different numbers of hidden units and learning rates. Hidden unit numbers and learning rates are drawn according to probability distributions derived from successful rates and sizes. More information is explained in Breuel and Shafait (2010).

Naïve Bayes: A naive bayes classifier is a simple probabilistic classifier based on applying bayes' theorem with strong independence assumptions. It assumes that the presence or absence of a particular feature of a class is unrelated to the presence or absence of any other feature. The advantage of the naive bayes classifier is that it only requires a small amount of training data to estimate the means and variances of the variables necessary for classification. Because independent variables are assumed, only the variances of the variables for each "label" need to be determined and not the entire covariance matrix. For more information, you may refer to Patil and Pawar (2012).

KNN: The k-nearest neighbor algorithm is based on learning by analogy, that is, by comparing a given test example with training examples that are similar to it. The training examples are described by n attributes. Each example represents a point in an n-dimensional space. In this way, all of the training examples are stored in an n-dimensional pattern space. When given an unknown example, a k-nearest neighbor algorithm searches the pattern space for the k training examples that are closest to the unknown example. These k training examples are the k "nearest neighbors" of the unknown example. If k = 1, then the example is simply assigned to the class of its nearest neighbor. Cover and Hart (1967) explained nearest neighbor classification in more details.

Linear Regression: Regression is a statistical measure that attempts to determine the strength of the relationship between one dependent variable and a series of other changing variables known as independent variables. Linear regression attempts to model the relationship between a scalar variable and one or more explanatory variables by fitting a linear equation to observed data. You can refer to Fahrmeir *et al*. (2009).

Experimental Analysis and Results

The phishing dataset that was published at the University of Irvine by Mohammed *et al*. (2012) was used in this study. The data set represents 11055 different websites. Each row in the data set consists of 30 different features known as conditional attributes and one classification attribute. The classification attribute values are 1, 0, or -1 and are representing the legitimate website, the suspicious website and the phishing website respectively.

The nine discussed classifiers were applied to the phishing dataset. Different classifiers usually give different accuracies. The quality of the dataset (complete, representative, consistent, etc.) may affect the accuracy as well as the quality of the algorithm and its robustness. There is no suitable classifier for all datasets. One classifier may give high accuracy when it is applied to one dataset and may not when it is applied to other datasets. The proposed work is aimed to compare between the nine different classifiers in terms of accuracy, precision, recall, specificity and F-Measure in order to find the high predictive classifier for the phishing problem.

The experiments have been conducted using RapidMiner software tool. Al-Shalabi (2017) highlighted the importance of this software. RapidMiner, formerly known Yet Another Learning Environment (YALE), is

software widely used for machine learning, knowledge discovery and data mining. RapidMiner is being used in both research and also in practical data mining fields. It will be used here to discover useful relationships from phishing data.

Decision trees, random tree, random forest, SVM, LibSVM, autoMLP, naïve bayes, linear regression and KNN classifiers have been applied to the complete 30-features phishing dataset. The confusion matrices were generated by each classifier as shown in Tables 2-10.

Table 2: The confusion matrix of random tree

Actual state	Predicted negative	Predicted positive
Negative	TN 4432	FP 707
Positive	FN 466	TP 5450

Table 3: The confusion matrix of random forest

Actual state	Predicted negative	Predicted positive
Negative	TN 2448	FP 207
Positive	FN 2450	TP 5950

Table 4: The confusion matrix of decision tree

Actual state	Predicted negative	Predicted positive
Negative	TN 4419	FP 425
Positive	FN 479	TP 5732

Table 5: The confusion matrix of SVM

Actual state	Predicted negative	Predicted positive
Negative	TN 4331	FP 310
Positive	FN 567	TP 5847

Table 6: The confusion matrix of LibSVM

Actual state	Predicted negative	Predicted positive
Negative	TN 4226	FP 341
Positive	FN 672	TP 5816

Table 7: The confusion matrix of AutoMLP

Actual state	Predicted negative	Predicted positive
Negative	TN 4657	FP 162
Positive	FN 241	TP 5995

Table 8: The confusion matrix of KNN (k=5)

Actual state	Predicted negative	Predicted positive
Negative	TN 4691	FP 421
Positive	FN 207	TP 5736

The results of the classifies' performance with respect to classification accuracy, precision, recall, specificity and F-Measure generated by the nine classifiers are illustrated in Table 11.

Table 11 shows that AutoMLP is able to construct the highest accuracy with almost 2% higher than the KNN which comes in the second place. Linear Regression and SVM comes in the third and fourth places respectively with an advantage to Linear Regression which is slightly higher than SVM by 0.06%. Decision tree is in the fifth place with 0.98% higher than LibSVM which comes in the sixth place. Random forest comes after LibSVM with 1.45% less. The worst classifiers are the random forest and the naïve bayes with some advantage for the random forest. AutoMLP and KNN are the superior classifiers. SVM, linear regression, LibSVM and random forest are accepted since their accuracies are over 90% whereas naïve bayes and random forest are far from the optimal classifier (AutoMLP). Figure 1 represents the accuracy of each classifier.

Other metrics were used to extend the comparison between the classifiers in order to narrow the differences so we can pick up one of them to be the most suitable for predicting the phishing dataset. The accuracy is not the only metric used to determine the suitable classifier. Next paragraphs show these metrics.

Recall is another important metric used to compare the efficiency of the nine classifiers. Naïve bayes has the highest recall rate followed by KNN, autoMLP, random tree, linear regression, decision tree, LibSVM and random forest. Naïve bayes is superior in classifying the positive cases. (recall) as well as KNN and autoMLP. Random forest is the worst. Figure 2 represents the recall of each classifier.

Table 9: The confusion matrix of linear regression

Actual state	Predicted negative	Predicted positive
Negative	TN 4375	FP 347
Positive	FN 523	TP 5810

Table 10: The confusion matrix of naïve bays

Actual state	Predicted negative	Predicted positive
Negative	TN 4874	FP 3088
Positive	FN 24	TP 3069

Table 11: The results of the nine classifiers

Performance	Random Tree	Random Forest 496 (3)	SVM	LibSVM	Decision Tree 834 (17)	AutoMLP	KNN (k = 5)	Naïve Bayes	Linear Regression
Accuracy	89.39	75.96	92.07	90.84	91.82	96.35	94.32	71.85	92.13
Precision	88.52	69.86	94.97	94.46	91.76	97.37	93.16	49.85	94.36
Recall	92.12	73.30	91.16	89.64	91.66	96.14	96.52	99.22	91.74
F-Measure (F1 Score)	90.28	71.54	93.03	91.99	92.67	96.75	94.81	33.18	93.03
Specificity	86.24	92.20	93.32	92.53	91.19	96.64	91.76	61.22	92.65
Average	**88.26**	**81.87**	**93.18**	**92.26**	**91.93**	**96.70**	**93.29**	**47.2**	**92.84**

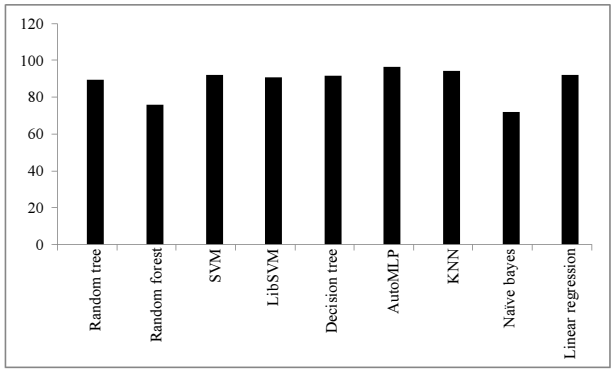

Fig. 1: The accuracy of each classifier

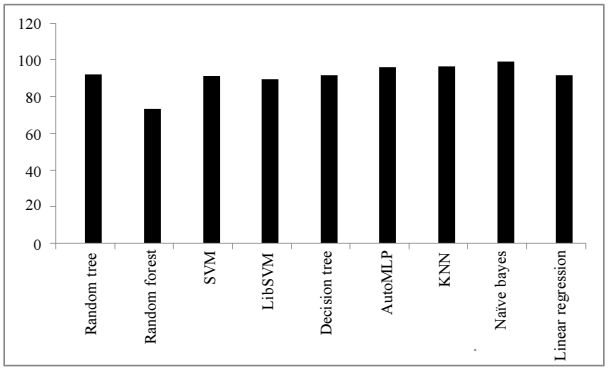

Fig. 2: The recall of each classifier

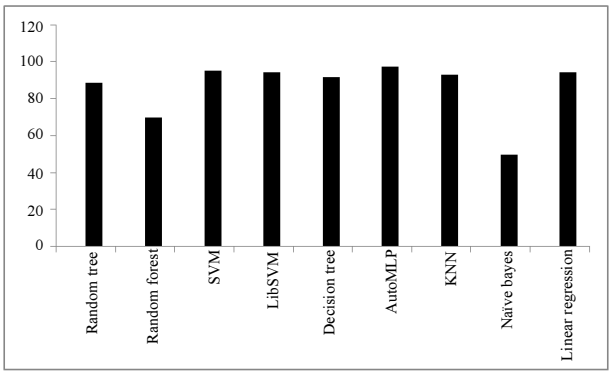

Fig. 3: The precision of each classifier

The third metric is precision. AutoMLP has the highest precision rate followed by linear regression, KNN, SVM, decision tree, LibSVM, random tree, random forest and naïve bayes. AutoMLP is superior in telling what proportion of websites that it diagnosed as phishing are actually phishing whereas naïve bayes the worst. Figure 3 represents the precision of each classifier.

The forth metric is the F-measure metric or what is called F1 score which gives the following results from highest percentage rate to the lowest: AutoMLP, KNN, SVM, linear regression, decision tree, LibSVM, random tree, random forest and naïve bayes. Figure 4 represents the F-measure of each classifier.

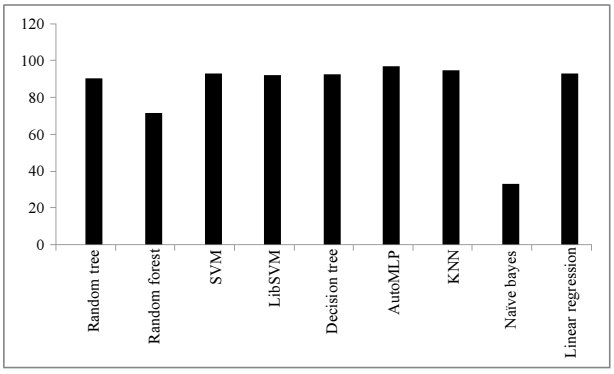

Fig. 4: The F-measure of each classifier

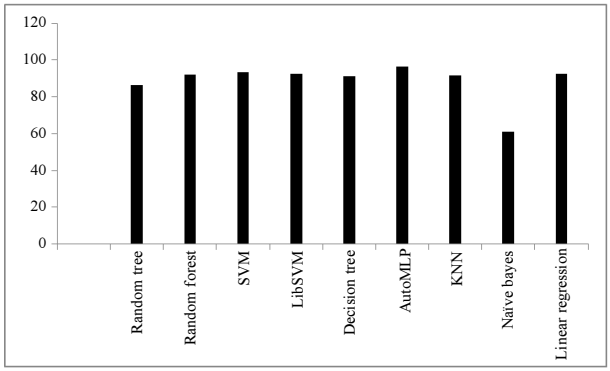

Fig. 5: The specificity of each classifier

Finally, the Specificity metric was applied to the fishing dataset. Specificity is the probability to predict an example as negative when it is truly negative. The highest specificity percentage is for the autoMLP classifier which means that it is the best classifier to classify the negative examples correctly. The next higher specificity percentage is for SVM followed by linear regression, LibSVM, random forest, KNN, decision tree, random tree and finally the naïve bayes. Figure 5 represents the specificity of each classifier.

Since we are interested in the best result of each metric that describes the performance of different portions of the dataset, the average of all measures was found. The best average is for autoMLP which means that it is the best classifier for predicting the legitimate and the phishing websites. KNN and SVM are also interested classifier which comes in the second and third places followed by the linear regression, LibSVM, decision tree, random tree, random forest and naïve bayes. Figure 6 represents the average of all metrics for each classifier.

If we are interested in negative region then the model of maximum specificity performance ratio is the most suitable one. In this case, the number of wrongly classified examples in the negative region is low.

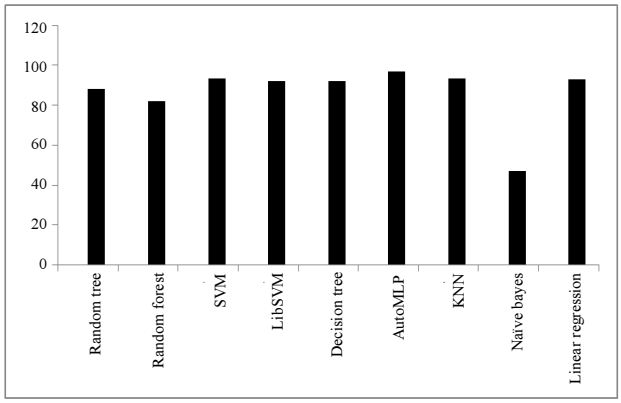

Fig. 6: The average of each classifier

Table 12: The classification error rate of FP websites given by each classifier

The classifier	# of FP websites	Classification error rate (%)
Random tree	707	13.76
Random forest	207	7.80
Decision tree	425	8.81
SVM	310	6.68
LibSVM	341	7.47
AutoMLP	162	3.36
KNN	421	8.24
Linear Regression	347	7.35
Naïve Bayes	3088	38.78

Table 13: The classification error rate of FN websites given by each classifier

The classifier	# of FN websites	Classification error rate (%)
Random tree	466	7.88
Random forest	2450	26.70
Decision tree	479	8.34
SVM	567	8.84
LibSVM	672	10.36
AutoMLP	241	3.86
KNN	207	3.48
Linear Regression	523	8.26
Naïve Bayes	24	0.78

Table 14: The average classification error rate of FP and FN websites given by each classifier.

The classifier	Average error rate (%)
Random tree	10.82
Random forest	17.25
Decision tree	8.58
SVM	7.76
LibSVM	8.92
AutoMLP	3.61
KNN	5.86
Linear Regression	7.81
Naïve Bayes	19.78

For the phishing dataset, it is important to look carefully at negative examples that represent the phishing cases. If the classifier predicts the truly phishing cases as false positive then it will be risky to use such cases (website).

To have some deep look at those FP websites which are truly phishing but predicted as legitimate (so much risky) and the FN websites that are legitimate but predicted as risky (not fair), the numbers of those websites were highlighted and the error rate was calculated based on the classifier used. Table 12 shows that the lowest risky classifier (the best to choose for the current dataset) is autoMLP with minimum error rate and the worst is naïve bayes with the highest error rate. Table 13 shows that the highest rational classifier (the best one to choose for the current dataset) that predicts legitimate websites as phishing with lower error rate is naïve bayes whereas the worst is random forest with the highest error rate.

We may find the average of the classification error rate of FP and FN which arranges the classifiers from best to worst based on the number of websites they wrongly classified regardless phishing or legitimate. The formula is as follows:

$$Average = \frac{FP + FN}{2}$$

Table 14 shows that the best classifier for predicting the phishing dataset which is used in this research is the autoMLP and the worst is naïve bayes.

Conclusion

In this paper, the performance measure of SVM, libSVM, decision tree, random forest, random tree, autoMLP, linear regression, KNN and naïve bayes classifiers was found by applying each of them to the phishing dataset. Results obtained by comparing output of confusion matrix and summary statistic. The classification performance of all classifiers was investigated by using the five statistical performance measures: accuracy, precision, recall, sensitivity and F-Measure. From the experimental results the average accuracy of each classifier was found.

As a conclusion, this research has met its objective which is to evaluate and investigate the nine selected classifiers based on RapidMiner. According to the results of the above classifiers, the best technique for the classification of the phishing dataset is autoMLP and the worst is naive bayes. Also, more promising result can be achieved by applying the SVM and KNN classifiers. AutoMLP classifier has achieved a remarkable performance with accuracy of 96.70% which is a competitive classifier for prediction the phishing websites from the dataset.

Acknowledgements

The author would like to thank Arab Open University for their support. He also likes to thank many anonymous people for their efforts in improving the readability of this paper including the patience of my wife and kids.

Funding Information

This research was supported and funded by the research sector, Arab Open University-Kuwait Branch under decision number 18143.

Ethics

This research article is original and has not been published elsewhere. The corresponding author confirms that there are no ethical issues involved.

References

Abdelhamid, N., 2015. Multi-label rules for phishing classification. Appl. Comput. Inf., 11: 29-46.

Abdelhamid, N., A. Ayesh and F. Thabtah, 2014. Phishing detection based associative classification data mining. Expert Systems Applications, 41: 5948-5959. DOI: 10.1016/j.eswa.2014.03.019

Aitkenhead, M.J., 2008. A co-evolving decision tree classification method. Expert Systems Applications, 34: 18-25. DOI: 10.1016/j.eswa.2006.08.008

Akinyelu, A.A. and A.O. Adewumi, 2014. Classification of Phishing Email Using Random Forest Machine Learning Technique. J. Applied Math. DOI: 10.1155/2014/425731

Al-Shalabi, L., 2009. Improving accuracy and coverage of data mining systems that are built from noisy datasets: A new approach. J. Computer Sci., 5: 131-135. DOI: 10.3844/jcssp.2009.131.135

Al-Shalabi, L., 2016. Data mining application: predicting students' performance of ITC program in the Arab Open University in Kuwait – The blended Learning. Int. J. Comp. Sci. Inform. Security, 14: 827-833.

Al-Shalabi, L., 2017. Perceptions of crime behavior and relationships: Rough set based approach. Int. J. Computer Science Information Security, 15: 413-420.

Al-Shalabi, L., 2018. Online shopping adoption factors in kuwait market based on data mining rough set approach. Int. J. Computer Application, 180: 10-17. DOI: 10.5120/ijca2018916832

Breuel, T.M. and F. Shafait, 2010. Automlp: Simple, effective, fully automated learning rate and size adjustment. In the Learning Workshop, Snowbird, Utah.

Chih-Chung, C. and L. Chih-Jen, 2011. LIBSVM: A library for support vector machines. ACM Trans. Intelligent Syst. Technol., 2: 1-27.

Choon, L.T. and L.C. Kang, 2017. Phishing webpage detection using weighted URL tokens for identity keywords retrieval. Proceedings of the 9th International Conference on Robotic, Vision, Signal Processing and Power Applications, Publisher: Springer, Singapore, pp: 133-139.

Cover, T.M. and P.E. Hart, 1967. Nearest neighbor pattern classification. IEEE Trans. Inf. Theory, 13: 21-27.

Fahrmeir, L., T. Kneib and S. Lang, 2009. Regression-Modelle, Methoden und Anwendungen. 2nd Edn., Berlin, Heidelberg.

Hadi, W., F. Aburuba and S. Alhawarib, 2016. A new fast associative classification algorithm for detecting phishing websites. Applied Soft Computing, 48: 729-734. DOI: 10.1016/j.asoc.2016.08.005

Hamid, I.R.A. and J.H. Abawajy, 2013. Profiling phishing email based on clustering approach. Proceedings of the 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, Jul. 16-18, IEEE Xplore press, Melbourne, VIC, Australia, pp: 629-635. DOI: 10.1109/TrustCom.2013.76

Han, J., M. Kamber and J. Pei, 2012. Data Mining: Concepts and Techniques. 3rd Edn., Morgan Kaufmann Publishers, USA, ISBN-10: 9780123814791

Hastie, T., R. Tibshirani and J. Friedman, 2008. The Elements of Statistical Learning. 2nd Edn., Springer, ISBN10: 0-387-95284-5.

Ho, T.K., 1995. Random Decision Forests. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Aug. 14-16, IEEE Xplore press, Montreal, QC, pp: 278-282. DOI: 10.1109/ICDAR.1995.598994

Ho, T.K., 1998. The random subspace method for constructing decision forests. IEEE Trans. Pattern Analysis Machine Intelligence, 20: 832-844. DOI: 10.1109/34.709601

Ivanciuc, O., 2007. Applications of support vector machines in chemistry. Reviews Computational Chemistry, 23: 291-400.

Jiang, H., D. Zhang and Z. Yan, 2013. A classification model for detection of Chinese phishing e-business websites. PACIS2013 Proceedings, pp: 152.

Junaid, A.C., A.C. Shafique and G.R. Robert, 2016. Phishing attacks and defenses. Int. J. Security Applications, 10: 247-256. DOI: 10.14257/ijsia.2016.10.1.23

Kang, L.C., S.C. Jeffrey, N.S. San and S.C.Y. Kelvin, 2018. Leverage website Favicon to detect phishing websites. Security Communication Netw., 2018: 1-11. DOI: 10.1155/2018/7251750

Kenneth, D.N., R. Heather and S.J. Richard, 2017. Valuing information security from a phishing attack. J. Cybersecurity, 3: 159-171. DOI: 10.1093/cybsec/tyx006

Li, T., F. Han, S. Ding and Z. Chen, 2011. LARX: Large-scale anti-phishing by retrospective data-exploring based on a cloud computing platform. Proceedings of the 20th International Conference Computer Communications and Networks, Jul. 31 to Aug. 4, IEEE Xplore press, Maui, USA, pp: 1-5. DOI: 10.1109/ICCCN.2011.6005822

Liu, W., N. Fang, X. Quan, B. Qiu and G. Liu, 2010. Discovering phishing target based on semantic link network. Future Generat. Comput. Syst., 26: 381-388. DOI: 10.1109/ICCCN.2011.6005822

Mishra, A.K. and B.K. Ratha, 2016. Study of random tree and random forest data mining algorithms for microarray data analysis. Int. J. Advanced Electrical Computer Eng., 3: 5-7.

Miyamoto, D., H. Hazeyama and Y. Kadobayashi, 2008. An evaluation of machine learning-based methods for detection of phishing sites. Proceedings of the International Conference on Neural Information Processing ICONIP 2008: Advances in Neuro-Information Processing (ICONIP'08), Springer, Berlin, Heidelberg, pp: 539-546.

Moghimi, M. and A.Y. Varjani, 2016. New rule-based phishing detection method. Expert Systems Applications, 53: 231-242. DOI: 10.1016/j.eswa.2016.01.028

Mohammad, R., F. Thabtah and L. McCluskey, 2012. Phishing websites dataset.

Mohammad, R., F. Thabtah and L. McCluskey, 2014a. Predicting Phishing websites based on self-structuring neural network. J. Neural Computing Applications, 3: 1-16. DOI: 10.1007/s00521-013-1490-z

Mohammad, R., T.L. McCluskey and F. Thabtah, 2018. An assessment of features related to phishing websites using an automated technique. Proceedings of the International Conference for Internet Technology And Secured Transactions, Dec. 10-12, IEEE Xplore press, London, UK, pp: 492-497.

Mohammad, R.M., F. Thabtah and L. McCluskey, 2014b. Intelligent rule-based phishing websites classification. IET Inf. Secur., 8: 153-160. DOI: 10.1049/iet-ifs.2013.0202

Naga, A., S. Venkata and A. Sardana, 2012. A pagerank based detection technique for phishing web sites. Proceedings of the IEEE Symposium on Computers & Informatics, Mar. 18-20, IEEE Xplore press, pp: 58-63. DOI: 10.1109/ISCI.2012.6222667

Ngai, E.W.T., L. Xiu and D.C.K. Chau, 2009. Application of data mining techniques in customer relationship management: A literature review and classification. Expert Systems Applications Elsevier, 36: 2592-2602.

Patil, A.S. and B.V. Pawar, 2012. Automated Classification of Web Sites using Naïve Bayesian Algorithm. Proceedings of the International Multi-Conference of Engineers and Computer Scientists, Mar. 14-16, Hong Kong.

Prakash, P., K. Manish, R.R. Kompella and M. Gupta, 2010. PhishNet: Predictive blacklisting to detect phishing attacks. Proceedings of the IEEE INFOCOM, Mar. 14-19, IEEE Xplore press, San Diego, USA. DOI: 10.1109/INFCOM.2010.5462216

Qabajeh, I. and F. Thabtah, 2014. An experimental study for assessing email classification attributes using feature selection methods. Proceedings of the 3rd IEEE Conference on Advanced Computer Science Applications and Technologies, Dec. 29-30, IEEE Xplore press, Amman, Jordan, pp: 125-132. DOI: 10.1109/ACSAT.2014.29

Uzun, E, H.V. Agun and T.A. Yerlikaya, 2013. A hybrid approach for extracting informative content from web pages. Inform. Processing Management, 49: 928-944. DOI: 10.1016/j.ipm.2013.02.005

Wardman, B., T. Stallings, G. Warner and A. Skjellum, 2011. High-Performance Content-Based Phishing Attack Detection. Proceedings of the IEEE Conference on eCrime Researchers Summit, Nov. 7-9, IEEE Xplore press, San Diego, USA, pp: 1-9. DOI: 10.1109/eCrime.2011.6151977

Weider, D.Y., S. Nargundkar and N. Tiruthani, 2009. PhishCatch - A phishing detection tool. Proceedings of the 33rd Annual IEEE International Computer Software and Applications Conference, Jul. 20-24, IEEE Xplore press, Seattle, USA, Computer Society, pp: 451-456. DOI: 10.1109/COMPSAC.2009.175

Zhang, H., G. Liu, T. W. S. Chow and W. Liu, 2011. Textual and visual content-based anti-phishing: A Bayesian approach. IEEE Trans. Neural Netw., 22: 1532-1546. DOI: 10.1109/TNN.2011.2161999

Zhang, Y., J. Hong, L. Cranor, 2007. Cantina: A content-based approach to detecting phishing web sites. Proceedings of the 16th international conference on World Wide Web, May, 08-12, ACM, Banff, Alberta, Canada, pp: 639-648. DOI: 10.1145/1242572.1242659

Zhenzhou, C., 2012. Local support vector machines with clustering for multimodal data. Advances Inform. Sci. Service Sciences, 4: 266-275. DOI: 10.4156/AISS.vol4.issue17.30

Zuhir, H., A. Selmat and M. Salleh, 2011. The effect of feature selection on phish website detection, an empirical study on robust feature subset selection for effective classification. Int. J. Advanced Computer Science Applications, 6: 221-232.