Detection and Prediction of Phishing Websites using Classification Mining Techniques

Mofleh Al-diabat
Department of Computer Science
Al Albayt University

ABSTRACT
Phishing is serious web security problem that involves mimicking legitimate websites to deceive online users in order to steal their sensitive information. Phishing can be seen as a typical classification problem in data mining where the classifier is constructed from large number of website's features.There are high demands on identifying the best set of features that when mined the predictive accuracy of the classifiers is enhanced. This paper investigates features selection aiming to determine the effective set of features in terms of classification performance. We compare two known featuresselectionmethod in order to determine the least set of features of phishing detection using data mining. Experimental tests on large number of features data set havebeen doneusing Information Gain and Correlation Features set methods. Further, two data mining algorithms namely PART and IREP have been trained on different sets of selected features to show the pros and cons of the feature selection process. We have been able to identify new knowledge in the forms of rules that show vital correlations among significant features.

Keywords
Classification Accuracy, Website Security, Data mining, Feature Assessment, Phishing

1. INTRODUCTION
One of the primary problems in web security nowadays is phishing. Phishing can be defined according to (Abdelhamid, et al., 2014) as the art of mimicking a legitimate website in order to deceive users by obtaining their sensitive information including usernames, passwords, accounts numbers, national insurance numbers, etc. Usually, the attackers proceed by conducting financial theft after phishing occurs. The primary mean of initiating a phishing attack is emails. The attacker sends an email to the candidate user that contains a link besides other information. When the user clicks on the link inside the email, he will be redirected to a fake website that looks like the authenticated website. Then, the user possibly input the username and password that on the fly will be redirected to the attacker.

It is vital to minimise online phishing activities due to the fatality of this problem on all involved stakeholders including online users, banks, businesses and government. As matter fact, preventing phishing activities early is imminent yet a challenging task due to the sophisticatedmethods used toattack users. 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., 2015).

Phishing detection can be defined in data mining context as a typical classification problem (Qabajeh anzd Thabtah, 2014). The aim is to predict the type of the website in an automated manner to either "legitimate" or "phishy" class labels based on a classifier generated from an input data set called the training data. The training data normally contains websites or more particularly website's features with a target attribute called the class. There have been several different anti-phishing solutions based on data mining such as (Uzun et al., 2013) (Mohammad, et al., 2014) (Abdelhamid, et al., 2014).

One promising direction to intelligently combat phishing threats is to identify the least set of website's features that can be utilised for phishing detection. Typically, there are huge numbers of features associated with phishing and non-phishing websites such as sourced code, URL, domains, etc. and capturing these features is not a straightforward task. Precisely, and for a website, there are x features connected with it and the search space in a dataset with n websites may reach $2n - 1$ different non-empty subsets. This creates huge burden on the process of choosing the relevant features (Abdelhamid, et al., 2013). Hence, there is a necessity of guiding the intelligent detection algorithm by a) reducing the dimensionality of the search space through pruning irrelevant features and b) group relevant features together. These advantages will indeed ease the automatic classification of the websites and minimise the use of computing resources. Moreover, keeping all website's features without differentiating between significant and insignificant ones may result in the production of redundant classifiers that may degrade the phishy prediction rate (Mohammad, et al., 2014).

This article investigates different feature selection methods with an aim to choose versions features sets that can derive well predictive classifiersusing data mining. The research question that we seek to answer is "Can small features sets be identified and used to generate high predictive classifiers?. We search for a small set of website's features that may not hinder the classifier's accuracy when compared with classifiers produced from the entire data set. The classifiers are learnt using two data mining algorithms named decision trees (C4.5 (Quinlan, 1993)) and rule induction (IREP (Cohen, 1995)). The feature selection methods used to select the features are Information Gain (IG) (Quinlan, 1986) and Symmetrical Uncertainty (SI) (Peng et al., 2005). These feature selection methods have been chosen since they have been used widely in different domains with a proven quality in filtering attributes. Furthermore, the data mining classification algorithms are selected since they utilise different learning mechanisms and produce simple understandable classifiers.

The paper is structured as follows: a literature review on feature assessment methods related to a website phishing classification problem is presented in Section 2. The feature assessment methods and the classification algorithms are discussed in Section 3. The different websites features and their brief descriptions are given in Section 4. Sections 5 and

6 aredevoted to the data and the experimental results using the feature assessment methods and the classification algorithms. We conclude and highlight areas for future research in Section 7.

2. FEATURES SELECTION REVIEW

This section sheds the light on recent research literature of feature selection and their applications in classification particularly for the phishing problem.

Large number of features related to phishing has been investigated by (Basnet, et al., 2012) using Correlation Features Set and Wrapper methods. The authors have chosen 42 features among 177 features set after applying different machine learning algorithms including Naive Bayes and Random Forest algorithms. They contrasted the error rate generated against a security data from phishtank (https://www.phishtank.com) before applying the feature selection methods and after applying the feature selection methods on the classifiers produced by the machine learning algorithms. Experimental results initially showed that Wrapper feature selection method when utilised as a preprocessing on the security data picks the best features. This has been attributed to the low error rates generated against Wrapper sub set data when compared to that of CFS.

A number of features related to websites have been studied by (Mohammad, et al., 2012) to organise features into clusters. The data used in the experiment has been collected using a PHP script from different sources primarily Yahoo and Phishtank. The authors have employed simple statistical frequency analysis on over 2000 websites (phishy and legitimate) to seek the largest frequency features. Based on the frequency analysis, different human rules have been created to define each feature boundary and possible values. One noticeable shortcoming from their study is the fact that the data set was imbalanced which may create biased rules.

(Abdelhamid, et al., 2014) have investigated chi-square testing statistical method to filter our 16 different website's features. The authors have utilised over 1400 websites collected from Millersmiles (www.millersmiles.co.uk/) and Yahoo directory. Experimental results have been conducted using an associative classification algorithm developed by the same authors to evaluate the features effectiveness in predicting phishy websites. Two features sets have been identified based on the classifiers accuracy generated. This study can be criticized of only using one learning approach in the testing of classifiers.

A large number of phishing features was examined by (Qabajeh and Thabtah, 2014) using three feature selection methods. The aim was to find out ten or less features that substantially improve classification accuracy on classifiers produced. The authors have used a security data set with 47 different features and used rule based classification algorithms to produce classifiers from different sets of features. These sets of features are identified using chi-square and CFS methods. The results of the experiments showed two features sets were detected. One that contains twelve features and one that contains nine features. The twelve features set derived classifiers with the least error rate. One clear limitation of the analysis performed by the authors is that fact that no sharp lines have been defined to distinguishes between features' significance.

Recently, a study that has evaluated a number of features selection method on data collected from phishtank was published by (Zuhair, et al., 2015). The data sets used in the experiment are secondary data that have been used in other research articles (Uzun et al., 2013). Hence, there are no features analyses have been given by the authors or a proper justification on why these features have been chosen. The data sets contain 58 features. Four features selection methods have been compared using WEKA. Four machine learning methods (ID3, C4.5, Naïve Bays, Support Vector Machine) were tested on the data sets. The results revealed that there arefew features with high significance are identified by the feature selection methods. A recommendation was drawn that there are no golden filtering method that fits all classifications algorithms at least on the data sets used in the experiments.

2.1 Feature Selection

Feature selection is filtering out a training data set in order to keep attributes / variables that have good representation of the entire training dataset. The selected subset attributes usually sever as a representative sample of the population and provide similar performances as the complete training dataset's attributes. Feature selection methods are extremely beneficial in cases when the dimensionality of the training dataset is large (very huge numbers of attributes). The dimensionality problem may limit the applicability of searching algorithms on the dataset and therefore dimensionality reduction becomes imminent. This section briefly reviews two feature selection methods that this article considers. Our choice was based on the popularity of these two methods, We have selected Information Gain (IG) (Quinlan, 1986) and Symmetrical Uncertainty (SU) (Peng et al., 2005) due to their successfully usage in different business data domains including Bioinformatics, medical analysis, text classification, email classification and many others (Uysal, 2016). In the next sub-sections, we describe the selected methods along with their related mathematical notations.

2.2 Information Gain

One of the effective statistical methods that have been originated from information theory to assess any attribute's significance is IG. IG has been used not only as filtering methods for variables but also as classifier building method in decision trees (Quinlan, 1993). In particular, IG has been used in different classification algorithms such as C4.5 and C5 to learn trees from input data sets in order to reduce the degree of uncertainty in predicting the value of the class labels in test data sets. Attributes are measured based on how they are informative in guessing the class values according to Equation (2).

Given training dataset D of P outcomes, for each available attribute such as X, is possible to calculate its information gain as:

$$E(D) = -\sum_{i} \mathrm{p}(x_i) \log_2 \mathrm{p}(x_i), \qquad (2)$$

where $\mathrm{p}(x_i)$ is the likelihood that xhave class c. The IG of attribute X in the input data (D) is

Gain (D, X) = Entropy (D) - ((|Dx| / | D |) * Entropy (Dx)) (3)

Where Dis the training dataset, Dx is subset of D for which X has value x, |Dx| = the size of the subset data having Dx from D , |D| = The training dataset size.

2.3 Symmetrical Uncertainty

In information theory, two attribute in a dataset can have a test of mutuality by computing the mutual information (MI)

between them. We try to measure how informative is one attribute via another attribute by computing the Entropy of the attributes with respect to the available class labels in the training dataset.

Symmetric Uncertainty (SU) is one way to find out how informative is an attribute in bits. Often high SU attributes are more important and useful in identifying the target class for classification datasets. For a variable /attribute (X) and a class (w), SU is defined as

$$SU(\mathbf{x}_k, \boldsymbol{\omega}) = 2\left(\frac{I(\mathbf{x}_k, \boldsymbol{\omega})}{H(\mathbf{x}_k) + H(\boldsymbol{\omega})}\right),$$

where, H(X) is the Entropy of variable X from calculated from the training dataset as : $H(X) = -\sum_i P(x_i)\log_2 P(x_i)$

and the mutual information of the Xth variable is calculated as

$$I(\mathbf{x}_k, \boldsymbol{\omega}) = \sum_{l=1}^{L}\sum_{j=1}^{C} P(\widetilde{x}_{ljk}, \omega_j)\log_2 \frac{P(\widetilde{x}_{ljk},}{P(\widetilde{x}_{ljk})P}$$

When the SU ends up with a value of "`1", this implies that there is a strong relationship between the attribute and the class and a value of "0" corresponds to a total independence between x and w.

3. THE WEBSITE FEATURES

There are huge numbers of features that are lined with a webpage and many of these can be utilised to identify fraudulent websites from legitimate ones including URL length, prefix, @ symbol, IP address, and others. Several scholars in the research area of computer security who are interested in phishing have studied the different webpage features. Examples of those are (Zuhir, et al., 2015), (Mohammad, et al., 2014), etc. This section describes the most common features that are used in the security research domain to distinguish between a legitimate and phishing websites. We have identified 30 features based on recent studies done, i.e. (Abdelhamid, et al., 2014).

Websites' Features Description

1. Lengthy URL: Often attackers hide the mistrustful portion of the URL to catch a user's submitted data. They also may redirect the uploaded webpage to a doubtful domain. Normally, there is no measure for the URL length but recent studies identified that an accepted URL length is often less than 56 characters.
2. Anchor URL: This feature is similar to the URL feature yet links within the webpage could redirect users to a different domain from the one inserted inside the URL address bar.
3. Pop-up Window: When a pop-up window asks a user to insert certain data usually this is an indication of a fraudulent activity.
4. WebPages with "//": Whenever a link within a webpage transmits a user's data input to a mistrustful webpage this is a sign of phishing. Often attackers utilise the redirecting links to conceal the legitimate link to fake users.
5. IP address: When the URL contains an IP address this can be a sign of phishing.
6. Server Form Handler: Whenever the user transmits data on a webpage, the data gets loaded on a server for processing on the same domain of the webpage. Normally attackers tend to deliberately leave the server form handler blank, which often indicates phishing.
7. Suffix and Prefix: Attackers usually trick users by changing URLs so they can feel secure by adding a suffix or a prefix to the original authentic URL.
8. Subdomain: When subdomains are inserted into the URL this may indicate a suspicious URL though online users may not notice. Therefore, when we have multiple subdomains (usually greater than two) this can be an indication of a phishing website.
9. DNS: This website feature gives information associated with the current live domain whereas discarded domains are often unavailable on the DNS. Often, phishers seek user's information promptly simply since the lifespan of a phishing website is less than 72 hours.
10. HTTPs: When there is a HTTPs protocol linked with a website this reflects security for the online users in particular regarding the presence of financial transactions. Nevertheless, attackers utilise false HTTPs to lure online users. Many online organisations offer an examination of the HTTPs using programmes such asVerisign and GoDaddy. If a security certificate within a URL has been there more than a year this can be a sign of legitimacy.
11. @ Symbol: One of the signs of fake websites is the use of the "@" symbol within the URL address. This may leadusers to neglect all characters before the @ so attackers can guide users to fake websites.
12. Request URL: There are different objects inside a webpage including text, picture, videos, etc. In cases where the current webpage's objects are loaded from a server that is different to the URL's then there is a possibility that this webpage is fake.
13. Irregular URLs: A test to examine whether the current browsed website is inside the WHO-IS database can determine the legitimacy of the website.
14. Right Click Disable: A known technique, which attackers utilisein order to hide the legitimate links and display fake ones to deceive online users. This technique can be implemented by chasing the mouse cursor movements and once it arrives to the fake link the status bar content is altered. When the property "Right Click" is disabled this is often a sign of phishing.
15. Domain Age: If a website has been in place for more than a year this is a sign of good security.
16. Website Traffic: When a website has high traffic then it is indeed secure and users can feel safe browsing it. Phishing websites normally have low browsing traffic and this can be checked through the rank inside Alexa-database.
17. Short URL: Sometimes the URL length can be shortened using HTTP Redirect.
18. Domain Length: Fraudulent websites often have a domain, which has been recently registered, and their lifespan is short. Therefore, when the domain expires in less than a year it can be suspicious.

19. Favicon: A favicon is an icon with a graphical image linked with a webpage. Web surfers display favicon inside the address bar. When the favicon is loaded from a domain, which is differenttothe one shown in the address bar, this is an indication of phishing.
20. HTTPS in URL's Domain: This feature can be used by phishes to deceive users by inserting the "HTTPs" within the URL's domain. For instance, http://https-www-barclays.co.uk.
21. Submitting Information to Email: Sometimes phishers intend to use PHP function such as mail () in order to redirect web surfers into a desired email. A phisher might redirect the user's information to his personal email. This technique is based on a server-side scripting language.
22. Website Forwarding: The number of times a webpage redirects the users to certain destinations can be a sign of phishing. Usually, legitimate websites have at most one redirect page.
23. Status Bar: In the source code, if the "onMouseOver" event changes on the status bar this is a sign of phishing.
24. Meta, Script and Link tags: When Meta, script and link tags are associated with the same domain of the webpage this is an indication of legitimacy.
25. IFrame Redirection: IFrame is a tag to display an extra webpage inside the current webpage. Users are often deceived when the "iframe" is invisible. This may allow the browser to render a visual delineation.
26. Ranking of a Webpage: This feature measures the significance of a webpage in the WWW. Fraudulent webpages often have little rank(< 0.2) or no rank.
27. Google Index: When a website is indexed by Google this indicates security.
28. Non-Standard Port: This feature can be used in checking a service status such as HTTP to manage penetrations. The security administrator controls servers by opening or blocking ports such as Proxy. When ports are not blocked, phishers can run services as they wish, which may risk user's data. .
29. Number of Links: A good indication for website legitimacy is counting the number of links pointing to it .Normally, phishing websites have one or no link directing to them because of their short life.
30. StatisticalReports Phishing forums and communities like Phishtank usually generate statistical reports about phishing activities. So when the "Host" belongs to top ranked IPs or domains in the annual statistical report produced by Phishtank, this is a sign of phishing.

4. EXPERIMENTAL RESULTS

We have used a real data related to web security domain and published at the University of Irvine Repository (Mohammed, et al., 2015). The phishing dataset contains around 11000data examples (websites) that have been collected from Phishtank (https://www.phishtank.com) and Yahoo Directory (Yahoo.com). The dataset consists of thirty website's features (Described earlier) plus a lass attribute which corresponds to the type of the website. Hence, this dataset is considered binary classification since the class has two distinct values : Legitimate (1) and Phishy (-1). The majority of the features in the dataset has two possible values (0, 1) and some features are linked with three different values (0,1,-1). Sample of ten data examples for none features and the class are displayed in Table 1. The data features values are created based on rules proposed in the literature mainly by (Abdelhamid, et al., 2014) (Mohammad, et al., 2012) .

In this section, three feature selection methods, i.e. IG, SU, CFS, that were reviewed in Section III are applied on the security dataset. We aim to identify significant features that when mined by data mining algorithms will a) Generate high predictive classifiers for the phishing problem and b) Derive new hidden correlations among the features that decision makers may make use in minimising the risk of phishing. Hence, the experiments in this section are divided into two main scenarios:

1) Applying the data mining algorithms on the complete security dataset without feature selection
2) Applying the data mining algorithms after pre-processing the features using the considered features election methods.

In both scenarios, the experiments conducted have deeply evaluated the classifiers derived by the data mining algorithms before and after feature selection methods, have been applied.

The key to success in measuring the significance of the features are based on testing chosen features utilizing classification algorithms from data mining. These algorithms will produce classifiers that are associated with hidden correlation (rules) and phishing detection success rate (Classification accuracy). Therefore, we have chosen to classification algorithms named IREP (Cohen, 1995) and C4.5 (Quinlan, 1993). These two algorithms employ different classifier building techniques. IREP is a greedy algorithm that constructs classifiers containing simple If-Then rules using search methods and excessive rule pruning techniques. Whereas, C4.5 utilises Entropy mathematical approach to build trees. Each path in the tree from the node to the leaf denotes an If-Then rule. C4.5 also implements pruning technique to cut down unnecessary branches while building the classifier. Both IREP and C4.5 have been applied in many +

-

The feature selection and data mining experiments have been conducted using WEKA (www.cs.waikato.ac.nz/ml/weka) software tool. WEKA was developed at the Waikato Environment for Knowledge Analysis research center in Waikato University, New Zealand. It is free Java platform that contains several algorithms implementations for classification, clustering, association rule, regression, feature selection, etc. We used ten-fold-cross validation method to learn the classifiers in WEKA..Finally, all experimental runs were performed using 2.8Ghz processor computer machine.

5. RESULTS ANALYSIS

C4.5 and IREP have been applied in the complete 30-features dataset without pre-processing phase. The results of the classifies' performance with respect to classification accuracy generated by both data mining algorithms are depicted in Figure 1. The figure states that decision tree algorithm is able to construct classifiers with slightly higher predictive rate than greedy IREP at least for the phishing classification problem. The different in accuracy % between C4.5 and IREP classifiers is under 1% which shows superiority of both algorithms and this indeed goes along with previous findings in other business domains.

The feature selection methods results are shown in Table 2. The scores given in Table 2 have been calculated by SU and IG methods in WEKA using the mathematical notations described in Section 3.To be more specific, Table 2 illustrates the best features that have been detected by both SU and IG measures and fulfilled the 0.01 score requirement. In other words, we only presented the features that have computed scores above 1% for both SU and IG.

Table 1.Sample of nine features plus the class and ten data examples (websites features) of the dataset

having_At_Symbol	URL_Length	Shortining_Service	having_At_Symbol	double_slash_redirecting	Prefix_Suffix	having_Sub_Domain	SSLfinal_State	Favicon	Class
-1	1	1	1	-1	-1	-1	-1	1	-1
1	1	1	1	1	-1	0	1	1	-1
1	0	1	1	1	-1	-1	-1	1	-1
1	0	1	1	1	-1	-1	-1	1	-1
1	0	-1	1	1	-1	1	1	1	1
-1	0	-1	1	-1	-1	1	1	1	1
1	0	-1	1	1	-1	-1	-1	1	-1
1	0	1	1	1	-1	-1	-1	1	-1
1	0	-1	1	1	-1	1	1	1	1
1	1	-1	1	1	-1	-1	1	1	-1

The results clearly indicate consistency in the scores produced by SU and IG and against the complete features set. In fact, thetop 11 features are common between both feature selection methods which obviously show performance consistency in pre-processing of

datasets. The highest three score features identified by SU and IG were"SSL_Final_state" "URL_of_Anchor" and "Prefix_Suffix". One notable results has been detected in Table 2 is that the score of third feature has been significantly deteriorated when compared to the score of the second top feature. To be specific, the score of the "Prefix_Suffix" feature was substantially dropped by 74% and 58% respectively in IG and SU feature selection methods when compared to the scores obtained for the "URL_of_Anchor" feature. This can be attributed to the high correlation between the top two ranked features. Table 1 also show some scores deterioration between "Web_Traffic" and "having_Sub_Domain" features.

We have evaluated the classifiers' predictive accuracy for both data mining algorithms and against the top 11-features set resulted after pre-processing. Figure 2 demonstrates the accuracy obtained against the 12-features data set by IREP and C4.5 algorithms along with the previous results of Figure 1. It is clear from Figure 2 results that feature selection has an impact on the classification accuracy derived from the security dataset. Yet, this impact between the complete features set and 11-features is minimal. In particular, the classification accuracy figures have dropped by 1.02% and 1.23% by IREP and C4.5 algorithms. We believe that reducing the dimensionality of the dataset from 30 features to 11 features with an exchange with 1% error rate is tolerable . There should be a tradeoff between the search space and the classification accuracy derived.

We further investigate the common 11 features significance by looking at 19 remaining features set. This set represents all

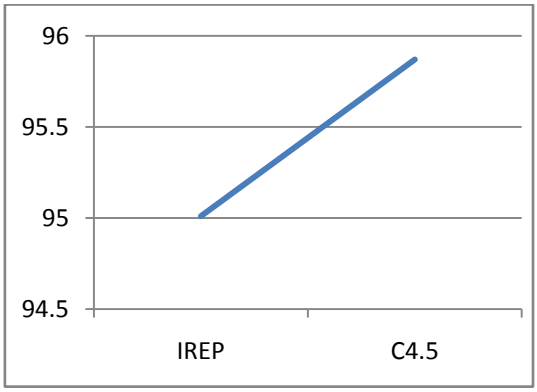

Fig. 1: Classification accuracy % generated by C4.5 and IREP from the complete set of features

features that are associated with scores below 1% in both SU and IG methods. Figure 3 depicts the classification accuracy measures generated by IREP and C4.5 algorithms against this dataset. we also included in the figure the results of the 30-features and 11-features sets for comparison purpose. Figure 3 not surprisingly showing that when using the 19-features set not detected by feature selection methods the accuracy of the data mining algorithm has been hindered. As matter fact, the classifiers generated by IREP and C4.5 from the 19-features dataset have drastically dropped by 23.58% and 212.84% by IREP and C 4.5 algorithms. This is a clear evidence of feature selection in phishing classification application in which 11-features set have pretty well phishing detection rate when compared with 19-features or even the complete features set.

Lastly, we investigated the classifiers in terms of the rules and correlations that have been derived by the data mining algorithms from the 11-features set and the complete features set. Figure 4 displays the classifiers size built by the data mining algorithms against the features sets specified above. There was an interesting findings which indicates that IREP produced the same size classifiers from the complete features set and the 11-features set. This means, there is no gain obtained in terms of correlations if we go beyond mining the 11-features detected by SU and IG feature selection methods.

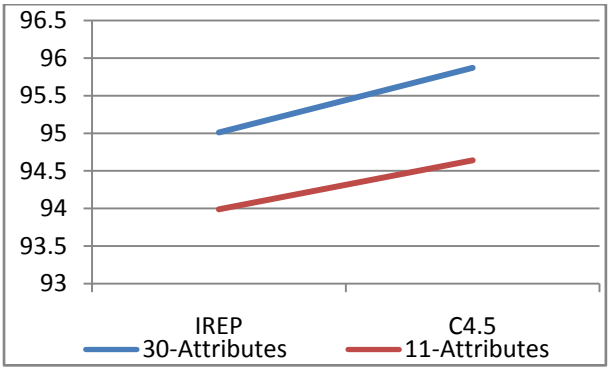

Fig. 2: Classification accuracy % generated by C4.5 and IREP from the 30-features set and 11-features set

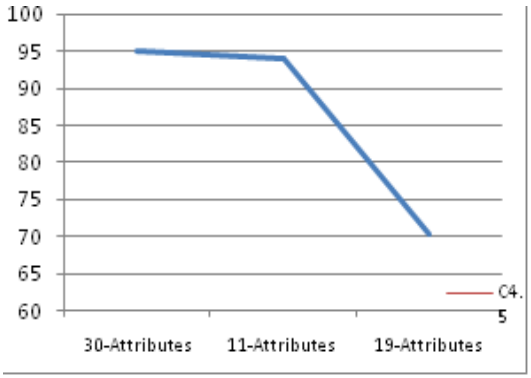

Fig. 3: Classification accuracy % generated by C4.5 and IREP from the 30-features set, top 11-features set and bottom 21-features set

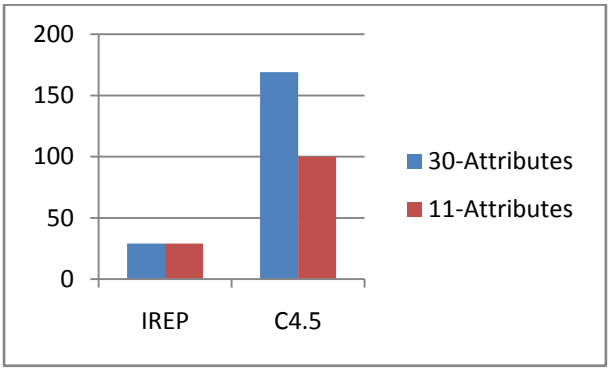

Fig. 4: Classifier size (# of rules) generated by C4.5 and IREP from the 30-features set, and the top 11-features set

6. CONCLUSIONS

Identifying the most significant features for website's phishing problem is a major task for both security and data mining scholars. In this article, we investigated two common feature selection methods namely Symmetrical Uncertainty (SU) and Information Gain (IG) hoping to differentiate among features and detect a small sets of correlation among features. This is vital for minimising the risks associated with phishing and may help in designing new anti-phishing solutions. We have utilised two common data mining approaches to measure the significance of features on two criteria: phishing detecttion rate and classifier size. In other words, we applied greedy and decision tree algorithms on different versions of a real security dataset related to phishing. After experimentations, the following promising results have been revealed:

1) SU and IG showed competitive performance and high similarity in detecting vital features. There was eleven common features detected in the pre-processing phase by both methods. This features set when mined guide the data mining algorithm in classifying phishy websites with high accuracy.

2) The most common two features identified were "SSL_Final_state" "URL_of_Anchor". These two features are linked with the top two scores by IG and SU methods.

3) IREP algorithm was able to generate controllable classifiers when compared with decision trees. In fact, IREP consistently produced only 29 correlations from both the complete and top-11-features sets.

4) Very interesting drops in scores have been noticed in the features selection methods results. There were two major drop points that may correspond to features sets that have high correlations. A deeper mathematical investigation on this issue is recommended.

In near future, we will explore the possibility of merging scores of known feature selection methods to increase reliability of the results of pre-processing phase.

7. ACKNOWLEDGMENTS

Our thanks to Al al-Bayt University which has contributed towards the development of this research through a sabbatical leave.

8. REFERENCES

[1] Abdelhamid N., Ayesh A., Thabtah F. (2014) Phishing detection based associative classification data mining. Expert Systems with Applications 41 (13) Pages 5948–5959, Oct 2014.

[2] Abdelhamid N, Ayesh A., Thabtah F. (2013) Phishing Detection using Associative Classification Data Mining. ICAI'13 - The 2013 International Conference on Artificial Intelligence, pp. (491-499). USA.

[3] R. Basnet, A. Sung, and Q. Liu, "Feature selection for improved phishing detection," Advanced Research in Applied Artificial Intelligence, pp. 252–261, 2012.

[4] Cohen W. (1995) Fast effective rule induction. In machine learning: Proceedings of the 12th International conference, pp. 115-123. Lake Tahoe, California. Morgan Kaufmann.

[5] Mohammad R. Thabtah F. McCluskey L., (2015) Phishing websites dataset. Available: https://archive.ics.uci.edu/ml/datasets/Phishing+Websites Accessed January 2016.

[6] Muhammad R., Thabtah F., McCluskey L., (2014) Predicting Phishing Websites based on Self-Structuring Neural Network. Journal of Neural Computing and Applications, (3)1-16. Springer.

[7] Mohammad R., Thabtah F, McCluskey L (2012) An Assessment of Features Related to Phishing Websites using an Automated Technique. In The 7th International Conference for Internet Technology and Secured Transactions (ICITST-2012); 2012; London: ICITST.

[8] Peng, H.C., Long, F., and Ding, C. (2005). Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (8): 1226–1238. doi:10.1109/tpami.2005.159. PMID 16119262.

[9] Qabajeh I, Thabtah F. (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 (ACSAT), pp. 125-132, 2014.

[10] Quinlan, J. (1993) C4.5: Programs for machine learning. San Mateo, CA: Morgan Kaufmann.

[11] Quinlan J. (1986). Induction of Decision Trees, Machine Learning, (1), 81-106.

[12] Uysal A. K. (2016) An improved global feature selection scheme for text classification. Expert systems with Applications, Vol. 43, pp. 82-92.

[13] Uzun E., Agun H. V., and Yerlikaya T. A. (2013) A hybrid approach for extracting informative content from web pages. Information Processing & Management, (49), 928-944, 2013.

[14] Zuhir H., Selmat A., Salleh M. (2015) The Effect of Feature Selection on Phish Website Detection An Empirical Study on Robust Feature Subset Selection for Effective Classification. International Journal of Advanced Computer Science and Applications, Vol. 6.,pp 221-232. 2011.