Hindawi
Security and Communication Networks
Volume 2017, Article ID 9838169, 10 pages
https://doi.org/10.1155/2017/9838169

Research Article

New Hybrid Features Selection Method: A Case Study on Websites Phishing

Khairan D. Rajab

College of Computer Science and Information System, Najran University, Najran, Saudi Arabia

Correspondence should be addressed to Khairan D. Rajab; khairanr@gmail.com

Received 4 November 2016; Revised 26 February 2017; Accepted 8 March 2017; Published 19 March 2017

Academic Editor: Muhammad Khurram Khan

Phishing is one of the serious web threats that involves mimicking authenticated websites to deceive users in order to obtain their financial information. Phishing has caused financial damage to the different online stakeholders. It is massive in the magnitude of hundreds of millions; hence it is essential to minimize this risk. Classifying websites into "phishy" and legitimate types is a primary task in data mining that security experts and decision makers are hoping to improve particularly with respect to the detection rate and reliability of the results. One way to ensure the reliability of the results and to enhance performance is to identify a set of related features early on so the data dimensionality reduces and irrelevant features are discarded. To increase reliability of preprocessing, this article proposes a new feature selection method that combines the scores of multiple known methods to minimize discrepancies in feature selection results. The proposed method has been applied to the problem of website phishing classification to show its pros and cons in identifying relevant features. Results against a security dataset reveal that the proposed preprocessing method was able to derive new features datasets which when mined generate high competitive classifiers with reference to detection rate when compared to results obtained from other features selection methods.

1. Introduction

One of the crucial online risks is phishing, which can be defined as designing a fake website that is visually similar to a trusted website targeting online users in order to gain access to their credentials, that is, login details [1]. Common techniques of phishing are based on technical tricks that are integrated with social engineering methods via email to initialise an online attack. Often, phishing attacks are initiated by sending emails from legitimate sources asking users to validate certain information using a link embedded inside the email [2]. Hence, stopping phishing attacks during the preliminary stages is a crucial step toward creating a secure online space.

The Internet community has devoted tremendous efforts into creating defensive measures to fight phishing attacks. However, the problem is constantly progressing and becomes more sophisticated since new online tricks are appearing on a regular basis. Therefore, an intelligent antiphishing solution based around computational and machine learning techniques is needed to differentiate among websites types.

The number of available features that can be linked to a website or an email is massive [3, 4]. These features are associated with certain website's elements such as the URL, domain, and source code. One primary challenge in minimizing the phishing risk is to identify the smallest set of features before intelligently classifying the website as phishy or legitimate [5]. Not considering this challenge may cause a deterioration in the phishing detection rate especially when many irrelevant features are kept in the dataset. These irrelevant features increase the search space for the intelligent method during building the classifier and may also participate in several useless rules [6]. Furthermore, the search space for n features in the training dataset may reach $2^n - 1$ different nonempty subsets. This may trigger the intelligent algorithm to fail during feature processing phase.

This paper evaluates phishing features aiming to come up with a new method that combines scores of different feature selection methods hence increasing the reliability of the chosen features sets. This is since the current feature selection methods show discrepancies in their results especially in the

preprocessing security datasets related to website phishing [7]. The aim is to identify features that are significant in detecting phishing activities by the data mining classification algorithm. The research question is *"Can a new combined score per feature based on results improve the detection rate of data mining algorithm for the phishing problem?"*

It will be useful having a new normalized score per feature generated by two feature selection methods. This increases the decision maker's reliability of the chosen features that will be used by the antiphishing algorithm and may improve upon classification accuracy of the classifiers. Moreover, new significant features in the dataset may be identified by combining features' scores. These new website's features will serve as the new dataset where the antiphishing algorithm will mine rather the original complete dataset.

The features selection methods that have been selected to be investigated are Information Gain (IG) [8] and Chi-square (CHI) [9]. The choice of these two methods is based on their successful applicability in preprocessing large datasets from multiple domains including Bioinformatics, Banking, Text Classification, and Medical Diagnoses [3, 7, 10, 11]. Our ultimate aim is to combine the scores derived by IG and CHI from the phishing dataset in an attempt to develop an enhanced metric for feature selection. The IG and CHI normally compute different scores per feature. So they need to be normalized to make them comparable and then to develop a new metric.

To measure the impact of the new developed measure, we have adopted two classification algorithms from data mining in the experimental analysis of the features' goodness and these are C4.5 [8] and IRIP [12]. The choice of these classification algorithms has been based on two facts: (a) they utilize different learning techniques in deriving the classifiers and (b) they produce simple understandable classifiers that contain human interpretable rules.

The paper is structured as follows: feature selection methods particularly IG and CHI and their related works are surveyed in Section 2. In Section 3 the new method for measuring features scores is presented. Section 4 is devoted to the data and the experimental results. Results and analysis are described in Section 5. We conclude and highlight areas for future research in Section 6.

2. Feature Selection Methods and Literature Review

There have been several studies to reduce the risks associated with phishing attacks such as [13–16]. For instance, [13] reviewed different email phishing attacks and their potential solutions. In particular, the authors focused on machine learning techniques used to detect phishing activities and showed their advantages and disadvantages. Reference [15] extended the work of [13] by critically analyzing spoofing and website phishing risks. The others showed recent developments in deceptive techniques and surveyed different combat mechanisms. However, experimental analysis has been conducted to measure the success and failures of these solutions on real phishing datasets. A more comprehensive review on phishing attacks and their solutions were discussed in [16]. The authors illustrated the phishing problem, its history, and the different common antiphishing approaches. Moreover, challenges and future deceptive methods have been highlighted. Reference [14] discussed limitations associated with intelligent antiphishing solutions especially the slow time in generating warning reports to end-user. Then, the authors proposed a new white list approach that online warns user of any possible phishing attacks. The white list proposed model showed promising results in regard to access time over the machine learning solutions.

Feature selection methods have been proposed primarily for two reasons according to [2]:

(1) Reduce the search space by removing irrelevant variables so that
   (a) only interrelated variables with the class label are selected,
   (b) the intelligent algorithm is able to come up with results based on the available computing resources. This often happens when we filter out massive numbers of variables in the input dataset because of a necessity to minimize the data dimensionality.

(2) To enhance the predictive power of the classifiers in supervised learning.

Feature selection is preprocessing the input dataset to assess the available attributes so that only the relevant features are kept and irrelevant features are discarded [17]. Typically, the data owner wants to determine the subset of features that serves as a good sample of the entire data. In normal circumstances, this subset of features when mined generates similar performances to the entire dataset. Feature selection is useful in cases when the dataset dimensionality is large (where the numbers of attributes is numerous) which normally makes the search space wide and creates difficulties for the data mining algorithm [18]. In this section, two known feature selection methods (IG, CHI) are discussed besides their mathematical notations.

2.1. Information Gain. Information Gain (IG) is a frequently used metric in machine learning for evaluating the goodness of an attribute for classification datasets. Typically, IG is computed as the difference between the class entropy and the conditional entropy in the presence of the feature.

$$I(C, A) = H(C) - H(C \mid A), \tag{1}$$

where C is the class variable, A is the attribute variable, and $H(\,)$ is the entropy. In practice, given a training set $T = \{(x_1, x_2, \ldots, x_k, c)\}_i$, where x_a is the value of the ath attribute and c is the corresponding class label we can compute the IG of the ath attribute by

$$\begin{aligned} I(C, I) = &-\sum_{c \in C} p(c) \log p(c) \\ &+ p(a) \sum_{c \in C} p(c \mid a) \log p(c \mid a) \\ &+ p(\overline{a}) \sum_{c \in C} p(c \mid \overline{a}) \log p(c \mid \overline{a}). \end{aligned} \tag{2}$$

Features with higher IG score are ranked higher than features with lower scores.

It was made popular by Quinlan [19] who used it in an algorithm to build a decision tree from a dataset. In essence, IG evaluates the added value of an attribute by computing its IG with respect to the class. In particular and for decision tree algorithms like C4.5 [8], each attribute in the training dataset is evaluated to measure its gain with respect to the available class labels. The attribute with the largest gain gets chosen and placed as a root by the decision tree algorithm and a branch of that attribute's possible values are formed. The tree keeps growing in the same manner by placing the best attribute in IG until a stopping condition is met. Once this occurs, each path from the root node to the leaves represents knowledge (rule).

Many feature selection studies have been conducted on the basis of IG. In [20], the authors constructed an algorithm based on IG for selecting the relevant features from intrusion detection datasets. In particular, IG is used to construct a discernibility matrix that is used to select the optimal features from the dataset. They show that their approach leads to the selection of features whose classification accuracy is superior to the original full set of features.

Reference [21] used IG as part of a three-step algorithm to find the optimal subset of features to increase classification accuracy and scalability in credit risk scoring applications. The authors utilized IG along with other measures to build a feature ranking in the initial step of their algorithm. These rankings were then used to reduce the search space. They show that search space reduction can quickly remove most of the irrelevant features.

In [22], IG was employed to propose a greedy feature selection method. The performance of the method in terms of both classification accuracy and execution performance was found to be significantly high for the twelve real-life datasets with various characteristics.

2.2. Chi-Squared. Chi-squared (CHI) is another widely used metric in machine learning for evaluating the goodness of an attribute [9]. The CHI measures the degree of independence between a pair of categorical variables. In our context, the greater the CHI score of a feature is, the more independent that feature is from the class variable. To compute the CHI score let X be the number of times feature a and class c occur together, Y be the number of times feature a occurs without class c, W be the number of times class c occurs without feature a, Z be the number of times neither a or c occurs, and N be the total size of the training set. Then the CHI score is given by

$$\mathrm{CHI}(a,c) = \frac{N \times (XZ - YW)}{(X+W) \times (Y+Z) \times (X+Y) \times (W+Z)}. \quad (3)$$

In particular, CHI evaluates the worth of an attribute by computing the value of the Chi-squared statistic with respect to the class.

In [23], the authors evaluated various feature selection methods and showed that CHI performed very well under the "stability" criteria. They also revealed that CHI performs well under the "goodness" criteria. In addition, in [24], Support Vector Machine (SVM) techniques are utilized for sentiment analysis with many univariate and bivariate data analysis for feature selection, minimizing errors to 13%–15% after applying CHI in preprocessing the input texts. Moreover, [25] utilized CHI in the filtering process as part of their method for sentiment analysis. The level of accuracy achieved is on the level with topic categorization although the former is considered a harder problem in the literature.

Few methods in the past combined the information from various feature selection methods into one score [26, 27]. Authors in [28] examined the effect of merging the scores obtained by a number of filtering methods (CHI, IG, GSS, NGL, and RS) on the predictive models of Arabic text. The authors have utilized two AND and union OR to merge the filtering methods' scores. The experiments showed minor accuracy enhancement on the text classifiers after preprocessing the textual data for three methods. There was no predictive power improvement in cases when the three filtering methods' scores are combined together.

In [26], the authors considered merging multiple scores derived by feature selection methods to seek more representative variables for predicting stock prices in the financial market. Three search methods were used by the authors: Principal Component Analysis (PCA), Genetic Algorithm (GA), and decision trees (CART). The combination methods to filter out undesirable variables were based on union, intersection, and multi-intersection strategies. It was shown that the intersection between PCA and GA and the multi-intersection of PCA, GA, and CART performed the best. In addition, these two combined feature selection methods filter out near 80% unrepresentative variables.

In [27], the authors examined the benefits of combining uncorrelated methods in the context of text classification. It was shown that in some cases the combined score enhanced the performance of the combined selection method. The combination of two methods was performed by normalizing the scores for each word and taking the maximum of the two scores essentially performing OR with equal weights.

3. The Proposed New Feature Score

In the research literature of feature analysis, there are quite few reliable methods that have been used by researchers to reduce the dimensionality of datasets. Some of which are purely statistical, others are probabilistic, and some are hybrid. Hence, these methods normally generate different results. For instance, Correlation Feature Set (CFS) [29] and Symmetrical Uncertainty (SU) feature selection methods are applied on Soybean and Vote datasets from University of California Irvine Repository [30] to seek whether there are obvious differences on the results. Soybean and Vote datasets consist of 36 and 17 attributes, respectively. The initial results of filtering features revealed that CFS have detected 22 and 4 attributes from Soybean and Vote datasets, respectively, whereas SU kept more attributes, that is, 34 from Soybean and 11 from Vote. These results, if limited, point out that there is a need to unify scores obtained by different feature

Table 1: The goodness rate of feature selection techniques on ten datasets.

Feature selection techniques	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
CHI (%)	67.36	92.68	95.94	85.55	48.12	84.92	95.51	76.50	90.99	89.32
IG (%)	87.78	88.96	95.95	83.06	50.36	86.75	94.99	89.70	87.71	48.40

selection methods. Hence, a better approach that combines the scores obtained from both SU and CFS will indeed take into account scores similarities and discrepancies and consequently increase confidence in the selected attributes.

The goal is to combine the scores from two known feature selection methods in an attempt to develop an enhanced metric for feature filtering. We can reduce the volatility of the new metric without sacrificing the accuracy. Our philosophy is akin to the idea of portfolio diversification in finance where it is recommended to own a basket of unrelated securities in order to minimize the risk. In particular, according the modern portfolio optimization theory in finance, one can maintain the same level of return but reduce the risk of the overall portfolio by combining uncorrelated assets. Thus, in the context of classification, combining uncorrelated features should stabilize the accuracy performance rate across various datasets while maintaining the overall average accuracy rate. In [23], it was shown that the goodness rates of IG and CHI are highly uncorrelated (see Table 1).

Therefore, to combine them, we first normalize both scores to make them comparable. Let $\mathrm{IG}_{\max}$ denote the maximum IG score among all the available features and then define the normalized score of the ath attribute by

$$\overline{\mathrm{IG}_a} = \frac{\mathrm{IG}_a}{\mathrm{IG}_{\max}}. \tag{4}$$

Likewise, we normalize the CHI scores by

$$\overline{\mathrm{CHI}_a} = \frac{\mathrm{CHI}_a}{\mathrm{CHI}_{\max}}. \tag{5}$$

Next score vector of feature a is defined to be

$$v_a = \begin{pmatrix} \mathrm{IG}_a \\ \mathrm{CHI}_a \end{pmatrix}. \tag{6}$$

The score vector thus contains information about both the IG and CHI scores. Recall that the magnitude of a vector is given by the square root of the sum of squares of its coordinates. Hence, the magnitude of the score vector can be used as a scalar metric of the vector.

$$|v_a| = \sqrt{(\mathrm{IG}_a)^2 + (\mathrm{CHI}_a)^2}. \tag{7}$$

The magnitude of the score vector can be used to compare feature to one another. Features with greater value of $|v_a|$ will be ranked higher. Unlike other ways of combining scores from different methods such as AND and OR our approach yields a true metric on the space of all pairs of scores. This allows for a mathematical structure for analyzing the space of combined scores.

Unifying and normalizing scores of CHI and IG measures are to come up with a new feature rank based on the computed scores. This indeed may place features within their true rank and enhance the chances of certain significant features to being detected during the preprocessing phase. These are the key feature(s) that we seek since they have a direct effect on determining phishing activities. All other nonsignificant features will be discarded and thus achieving the following advantages:

(1) Reducing data dimensionality since irrelevant features will be deleted at the preliminary stage.

(2) Minimizing the search space of the complete features dataset to enable efficient training for the data mining algorithm without hindering the overall predictive power of the derived classifiers.

The above distinctive advantages benefit security and data mining scientists in enhancing the design process of antiphishing security models besides being useful knowledge that denotes crucial correlations among features.

Reducing the number of features based on the proposed method will lead to smaller phishing features sets. Therefore, security experts as well as novice users will be able to not only manage the security indicators (small sets of features) but also understand them easily. A next phase will possibly be building interactive visualization methods that "online" detect phishing attacks based on these security indicators chosen by the proposed filtering method. This visualization technique will be able to take advantage of the smallest yet effective features and then integrate these features within a web browser to empower not only security experts but also novice users in fighting phishing attacks.

Another benefit of the proposed feature selection method is the concise set of knowledge in the predictive models after data processing beside data processing efficiency. A data processing technique such as decision trees or JRIP that are built on the top of the chosen phishing features will definitely benefit from the reduction in the data dimensionality. This is since they will only process highly relevant small numbers of features and this will result normally in small yet predictive models. These models will contain new knowledge that security experts can control and use in minimizing external web threats such as phishing attacks. They can also adopt this knowledge as part of their IT security policy and antiphishing induction workshops.

4. Data and Experiments

The phishing dataset used in the experiment consists of over 11000 website examples and 30 different features (attributes). The dataset was recently developed by [31] and published in UCI repository [30]. Most of the dataset's attributes are binary $(0, 1)$ or ternary $(0, 1, -1)$. The dataset is categorized under classification in data mining since there is class label added

(target attribute) that has two possible values (phishy −1, legitimate 1). The primary sources of the data are Millersmiles [32], Phishtank [33], and Yahoo directory (yahoo.com). Each data example corresponds to a website that was collected using a PHP script which was plugged into the Firefox browser. Initially, all features considered have been assessed using frequency analysis. More details on the features names, types, possible values, and descriptions are given in [31].

All preprocessing experiments on the phishing dataset have been conducted using two primary feature selection methods: IG and CHI. The goal is to distinguish among the features by deriving correlated features sets and reducing the dimensionality of the original dataset without hindering classification accuracy. The correlated features sets discovered can be utilized by researchers to minimize phishing risks when users browse the World Wide Web (WWW). In other words, the new features sets are used to learn the classifiers using data mining that are able to classify websites into phishy or legitimate class labels. The classifiers may hold valuable knowledge about correlations among features and class values that when used may enhance the design of antiphishing detection algorithms.

The primary reason of choosing IG and CHI is the fact that they usually discard irrelevant features and keep related features. This has been proven in several business domains. In order to measure the effect of feature selection, two data mining algorithms JRIP [12] and C4.5 [8] have been employed. These algorithms generate If-Then classifiers and have been used to validate the proposed feature method's goodness. The choice of these classification algorithms is the simplicity of their classifiers which normally can easily be understood by users.

C4.5 is a known algorithm that constructs decision tree classifiers using entropy [34]. The algorithm tests the variables in the input dataset to determine the one that can split the data with respect to the class label in a way to maximize the information gained. This variable will be the tree root. Then C4.5 continues testing the remaining variables until all variables are tested or the training data examples are classified. Once the tree is constructed, each path in the tree to a leaf node corresponds to a rule. On the other hand, JRIP is a rule induction algorithm developed by [12] that employs separate and conquer learning in the way rules are induced from the training dataset. JRIP chooses a class (say C1) from the training dataset and then builds an empty rule, that is, if empty then C1. The algorithm computes the highest frequency item that appears with C1 and appends it to the empty rule's body. It keeps appending items until the rule gets 100% expected accuracy or cannot improve further. Once this occurs, the rule gets produced and JRIP moves on to the next rule for class C1 until all data linked with this class are covered. When this happens, JRIP moves to a new class and builds the rules in the same manner until the entire training dataset becomes empty or no more rules can be generated.

The data mining algorithms and feature selection methods experiments have been conducted using the WEKA tool [34]. WEKA is the acronym for Waikato Environment for Knowledge Analysis, which is a free Java platform system that was designed and implemented at Waikato University. It consists of different machine learning algorithms implementations as well as filtering methods. In all experiments of the C4.5 and JRIP, a cross validation technique was employed in the process of deriving the classifiers. Specifically, tenfold-cross validation was used during training in order to reduce overfitting. This has been accomplished by splitting the training datasets into 10 parts, learning the classifier from nine parts, testing it on the holdout part, and then repeating the same process ten times and averaging out all error rate results from the runs to come up with an average error rate of the classifier. Lastly, all experiments have been conducted on a personnel computer that has 2.5 Ghz processor.

The performance of the classifiers derived from different features sets using JRIP and C4.5 with respect to predictive accuracy is compared. This indeed will show the impact of the new selected features sets on detecting phishing activities besides determining the minimal numbers of features needed.

The objectives of the experiments are threefold:

(i) Assessing the entire features using CHI and IG. The results of this assessment are to determine common major features among these methods. Furthermore, features that correlate with each other and the class values can also be determined via statistical analysis on the scores computed by the features selection methods.

(ii) Evaluating the features sets generated in fold one's experiments using data mining algorithms. Basically, we look at the classifiers produced by the data mining algorithms before and after applying the feature selection methods.

(iii) Generating a new score method based on combining scores of CHI and IG. The new method should produce at least similar if not better detection accuracy when used with data mining classification algorithm.

The classification algorithms are employed to measure the increase or decrease of the error rate among using different sets of features.

5. Results and Analysis

Table 2 depicts the scores generated by the new feature method, CHI, and IG from the security dataset used. Each feature name, score, and rank are displayed according to the computed result. The last three columns of Table 2 show the normalized scores for both CHI and IG, the new score proposed by us, and the true score rank. For instance, "having_Sub_Domain" feature is ranked fourth among 30 features in both the proposed method and CHI and fifth in IG. The features' scores are computed by IG and CHI and the proposed method using the mathematical formulas described earlier.

Table 2 clearly shows consistency of the features scores calculated by IG, CHI, and the proposed method. Yet, for the top ranked features, there are some differences between our method and IG. For example, "Prefix_Suffex" feature is ranked #3 in IG whereas both CHI and our

Table 2: Features score(s) and rank of our new method, IG, and Chi-square on the phishing dataset.

Feature	Rank	Score	Normalized IG score	Rank	Score	Normalized Chi score	Combined IG and Chi score	New rank (IG + Chi)
having_IP_Address	13	0.006	0.012024048	13	98	0.014657493	0.018958371	13
URL_Length	17	0.003	0.006012024	17	57	0.008525277	0.010431911	17
Shortining_Service	18	0.003	0.006012024	18	51	0.007627879	0.00971231	18
having_At_Symbol	20	0.002	0.004008016	20	31	0.004636554	0.00612877	20
double_slash_redirecting	23	0.001	0.002004008	23	16.5	0.002467843	0.00317904	23
Prefix_Suffix*	3*	0.123	0.246492986	5*	1343	0.200867484	0.317972543	5*
having_Sub_Domain*	5*	0.109	0.218436874	4*	1595	0.238558181	0.323457375	4*
SSLfinal_State#	1#	0.499	1	1#	6686	1	1.414213562	1#
Domain_registeration_length	**9**	0.036	0.072144289	8	563	0.084205803	0.110884695	**8**
Favicon	29	0	0	29	0	0	0	29
port	24	0.0009	0.001803607	24	14.6	0.002183667	0.002832208	24
HTTPS_token	22	0.001	0.002004008	22	17.5	0.00261741	0.003296495	22
Request_URL	**7**	0.046	0.092184369	7	709	0.106042477	0.140509661	**7**
URL_of_Anchor#	2#	0.477	0.955911824	2#	5966	0.892312294	1.307665341	2#
Links_in_tags	**6**	0.047	0.094188377	6	712	0.106491176	0.142168283	**6**
SFH	**8**	0.037	0.074148297	9	542	0.081064912	0.10986123	**9**
Submitting_to_email	26	0.0002	0.000400802	26	3.7	0.000553395	0.000683292	26
Abnormal_URL	19	0.002	0.004008016	19	40	0.00598265	0.007201132	19
Redirect	25	0.0002	0.000400802	25	4.5	0.000673048	0.000783349	25
on_mouseover	21	0.002	0.004008016	21	19	0.002841759	0.004913226	21
RightClick	27	0	0	27	1.7	0.000254263	0.000254263	27
popUpWidnow	30	0	0	30	0	0	0	29
Iframe	28	0	0	28	0.1	$1.49566E-05$	$1.49566E-05$	28
age_of_domain	11	0.01	0.02004008	11	163	0.0243793	0.031558756	11
DNSRecord	16	0.004	0.008016032	16 63	0.009422674	0.012371078	16	
web_traffic*	4*	0.1145	0.229458918	3*	1712	0.256057433	0.343826707	3*
Page_Rank	12	0.008	0.016032064	12	121	0.018097517	0.024177411	12
Google_Index	10	0.011	0.022044088	10	183	0.027370625	0.035143889	10
Links_pointing_to_page	15	0.004	0.008016032	15	66	0.009871373	0.012716162	15
Statistical_report	14	0.004	0.008016032	14	70	0.010469638	0.013185981	14

Table 3: The decrement in % between two successive scores based on the results of Table 2 for IG, CHI, and the proposed method.

Features rank	IG drop %	CHI drop %	New method drop %	Average drop %
1 & 2	0.044088176	0.107687706	0.07534097	0.075705617
2 & 3	*0.742138365*	*0.713040563*	*0.737068273*	*0.730749067*
3 & 4	0.069105691	0.068341121	0.059243017	0.065563276
4 & 5	0.048034934	0.15799373	0.016956892	0.074328519
5 & 6	*0.568807339*	*0.469843634*	*0.552891323*	*0.530514099*
6 & 7	0.021276596	0.004213483	0.011666616	0.012385565
7 & 8	0.195652174	0.205923836	0.210839348	0.204138453
8 & 9	0.027027027	0.037300178	0.009229999	0.024519068
9 & 10	*0.694444444*	*0.662361624*	*0.680106538*	*0.678970869*
...	...	...	...	...

method place it on the third position in the features rank. The same thing applies to "Having_Sub_Doman" and "Domain_Registration_Length" features. Overall, the initial results consistently showed high correlations between the proposed method and CHI.

According to the computed scores in Table 2, there are features that seem to be highly correlated based on the ratio computed as the difference between two successive features scores as shown in Table 3. This decrement is computed using the below equation. For any two features to be considered both must pass the minimum thresholds identified in CHI and IG. Cut-offs of 0.01 for IG and 10.83 for CHI are used in [7]. In other words, and for IG, any feature with a score less than 0.01 has been discarded and for CHI the cut-off score for feature goodness is 10.83.

Decrement (%) (D) for features scores (i, j)

$$D = \frac{\text{Score}_i - \text{Score}_j}{\text{Score}_i}, \tag{8}$$

where Score_i is the score of feature i and Score_j is the score of feature j and $i > j$.

Based on analyzing the results of Table 2, the "italic" rows of Table 3 are points corresponding to a drastic drop between two successive features scores for all the preprocessing methods used. These points are in fact fine lines that discriminate between sets of features. For instance, in Table 2, there are three obvious features sets:

(i) Set (A): this set is superscripted by # and represents two features: "SSLfinal_State" and "URL_Of_Anchor." These are the best two features in the security data which have the largest impact on phishing. In fact, all feature selection methods consistently rank these two features as first and second, respectively.

(ii) Set (B): this set contains all features of (A) plus the ones superscripted by $*$. So there are five features in total which are "SSLfinal_State," "URL_Of_Anchor," "Prefix_Suffix," "Having_Sub_Domain," and "Web_Traffic."

(iii) Set (C): this set consists of both sets (A) and (B) besides all bold features of Table 2 and it contains 9 features in total.

The analysis showed three major cut-off points that define three new clusters of features. Hence, any feature that appears after the third cut-off point has less impact on the phishing detection rate as seen soon in this section. According to our analysis, any decrement ratio between two successive features' can be qualified to become a cut-off point if it fulfils two conditions.

(1) The average ratio difference between the two successive features and for all preprocessing methods is > 50%.

(2) The minimum thresholds of IG, CHI, and the proposed method are fulfilled by the two successive features.

It should be noted that we have discarded the cut-off point between the "Port" and "Redirect" features since these two features are associated with scores < the minimum thresholds for IG and CHI.

The cut-offs determined in our analysis show a promising direction toward differentiating between features' goodness in phishing when using results obtained from IG and CHI. To validate our new derived features' sets, their impact is measured using data mining. Hence, data mining models generated from (a) the complete dataset, (b) Set A, (c) Set B, and (d) Set C, using two learning algorithms (JRIP, C4.5), are needed.

Figure 1 illustrates the classification accuracy in % for JRIP and C4.5 algorithms against

(1) feature set (A),

(2) feature set (B),

(3) feature set (C),

(4) the remaining 21 features' set that comes after the last cut-off point (see Table 3).

The accuracy results for both data mining algorithms show consistent performance on the three new derived features subsets. For example, the differences in accuracy generated from the complete dataset (30 features) and subset (A) (2 features) by JRIP and C4.5 algorithms are 3.75% and 4.61%, respectively. This is an evidence on the goodness of the subset (A) which only contains two features yet the accuracy of its classifiers is of high quality. The same analogy can be applied

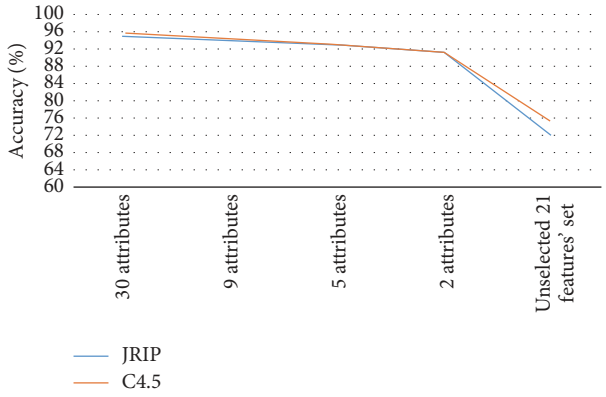

Figure 1: The accuracy (%) of JRIP and C4.5 algorithms on different features sets.

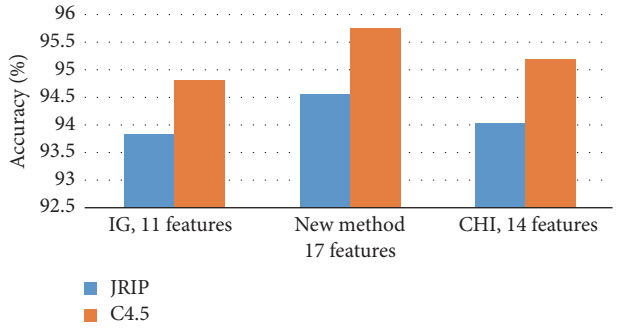

Figure 2: The accuracy (%) of JRIP and C4.5 obtained after applying the considered feature selection methods.

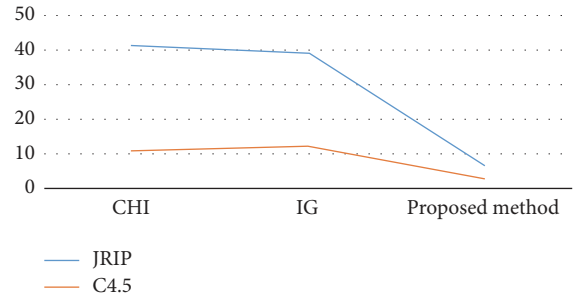

Figure 3: Time in ms of JRIP and C4.5 obtained after applying them on the features.

to the classifiers generated from subsets (B) and (C). These high accurate classifiers provide clear proof that the three cut-off points derived by us are distinguishing among features and creating a new promising research path toward clustering of features using preprocessing. Figure 1 demonstrated that all of the feature sets (A, B, C) were able to produce classifiers that classify phishing data examples with higher accuracy rate than the 21 features' set. As a matter of fact, the number of features in the subsets (A, B, and C) is now much less than 21 and therefore not only has the classification accuracy been substantially enhanced but also the search space and the dimensionality of the dataset have also been reduced.

The knowledge derived of JRIP and C4.5 against all of the subsets of the data features is investigated. The following two important knowledge items are revealed:

(i) (URL_of_Anchor = −1) and (SSLfinal_State = −1) ≥ Result = −1.

(ii) (URL_of_Anchor = −1) and (SSLfinal_State = 0) ≥ Result = −1.

These two knowledge items have no error and the proportion of the data examples covered by these two rules from the entire date set is 25% which is substantial for the decision maker. There was a notable result of the classifiers of the data mining algorithms which is no rules for JRIP were produced for the legitimate class label. In fact, all JRIP classifiers contain only rules that are linked with phishing class. This can be attributed to the fact that the features inside the security data have been defined using human knowledge that focuses more on phishing characteristics rather than legitimate ones. In fact, most security experts are studying phishing scenarios from a narrow angle without spending considerable time on investigating features related to authenticated websites.

Figure 2 investigates the accuracy figures derived from the IG, CHI, and the new combined scores method. We have taken into account the features chosen by the preprocessing methods based on the minimum threshold defined in each respective method. After normalizing the scores, a cut-off score of ≥0.01 for all methods is chosen for a fair deal. Figure 2 clearly shows the goodness in choosing features by the proposed method. The fact that the new combined score for each feature helped in improving classification accuracy is justified in Figure 2. CHI and IG were able to detect 11 and 14 features, respectively. Our method was able to select 17 features that have contributed in slightly improving the phishing detection rate for both JRIP and C4.5 algorithms. These figures obviously reveal the benefit of combining scores of multiple feature selection methods at least in security domains such as phishing detection.

Figure 3 shows the runtime in milliseconds of the data mining algorithms considered on the features selected by CHI, IG, and the proposed method. It is obvious in the figure that the predictive models generated when using the proposed method are faster. This is due to the fact that the proposed method normally tends to generate less number of features since it combines scores of both IG and CHI at least for the security data obtained from phishing websites. In other words, the reduction in the data dimensionality by removing correlated features has speeded up the data processing phase. This will contribute to minimizing even the antiphishing models size as indicated earlier and therefore more efficient antiphishing solutions.

6. Conclusions and Future Work

Determining the set of relevant features in phishing classification is one of major challenges faced by security experts and data minters. There have been a number of preprocessing methods from statistics and information theory such as Information Gain (IG), Correlation Features Set (CFS), and

Chi-square (CHI) that have been applied to phishing classification. Yet, these methods produce discrepant results which makes identifying the right features a hard task. This paper has developed a new feature score based on combining scores from two effective feature selection methods (IG, CHI). The new feature method computes a new normalized score in the preprocessing stage of the phishing dataset. The results obtained from applying our method, CHI, and IG against 30 features set security data revealed that the new method is able to pick relevant features that impact on the phishing detection rate. In particular, after applying data mining algorithms on the features identified by the new method, IG, and CHI, the accuracy of the classifiers derived from the set of features that was chosen by our method outperformed those of CHI and IG. Moreover, the analysis of the results of all preprocessing methods showed new promising clusters of relevant features that when mined detect phishing activities and produce vital correlations among features that may help decision makers minimizing the risk of phishing.

In the near future, a new feature selection method based on feature correlations will be investigated.

Conflicts of Interest

The author declares that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This project is funded under Najran University Research Grant no. NU/ESCI/14/035.

References

[1] R. M. Mohammad, F. Thabtah, and L. Mc-Cluskey, "Predicting phishing websites based on self-structuring neural network," *Neural Computing and Applications*, vol. 25, no. 2, pp. 443–458, 2014.

[2] N. Abdelhamid, A. Ayesh, and F. Thabtah, "Phishing detection based associative classification data mining," *Expert Systems with Applications*, vol. 41, no. 13, pp. 5948–5959, 2014.

[3] I. Qabajeh and F. Thabtah, "An experimental study for assessing email classification attributes using feature selection methods," in *Proceedings of the 3rd International Conference on Advanced Computer Science Applications and Technologies (ACSAT '14)*, pp. 125–132, Amman, Jordan, December 2014.

[4] R. B. Basnet, A. H. Sung, and Q. Liu, "Feature selection for improved phishing detection," in *Proceedings of the 25th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (IEA/AIE '12)*, pp. 252–261, Springer, Dalian, China, 2012.

[5] H. Zuhair, A. Selmat, and M. Salleh, "The effect of feature selection on phish website detection," *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 10, pp. 221–232, 2015.

[6] N. Abdelhamid, A. Ayesh, and F. Thabtah, "Phishing detection using associative classification data mining," in *Proceedings of the International Conference on Artificial Intelligence (ICAI '13)*, pp. 491–499, 2013.

[7] E. Uzun, H. V. Agun, and T. Yerlikaya, "A hybrid approach for extracting informative content from web pages," *Information Processing and Management*, vol. 49, no. 4, pp. 928–944, 2013.

[8] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kauffmann, 1993.

[9] H. Liu and R. Setiono, "Chi2: feature selection and discretization of numeric attributes," in *Proceedings of the IEEE 7th International Conference on Tools with Artificial Intelligence (ICTAI '95)*, pp. 388–391, November 1995.

[10] R. M. Mohammad, F. Thabtah, and L. McCluskey, "An assessment of features related to phishing websites using an automated technique," in *Proceedings of the 7th International Conference for Internet Technology and Secured Transactions (ICITST '12)*, pp. 492–497, IEEE, London ,UK, December 2012.

[11] A. K. Uysal, "An improved global feature selection scheme for text classification," *Expert Systems with Applications*, vol. 43, pp. 82–92, 2016.

[12] W. W. Cohen, "Fast effective rule induction," in *Proceedings of the 12th International Conference on Machine Learning*, pp. 115–123, Tahoe City, Calif, USA, July 1995.

[13] A. Tewari, A. K. Jain, and B. B. Gupta, "Recent survey of various defense mechanisms against phishing attacks," *Journal of Information Privacy and Security*, vol. 12, no. 1, pp. 3–13, 2016.

[14] A. K. Jain and B. B. Gupta, "A novel approach to protect against phishing attacks at client side using auto-updated white-list," *EURASIP Journal on Information Security*, vol. 2016, no. 1, article 9, pp. 1–11, 2016.

[15] A. Almomani, B. B. Gupta, S. Atawneh, A. Meulenberg, and E. Almomani, "A survey of phishing email filtering techniques," *IEEE Communications Surveys and Tutorials*, vol. 15, no. 4, pp. 2070–2090, 2013.

[16] B. B. Gupta, A. Tewari, A. K. Jain, and D. P. Agrawal, "Fighting against phishing attacks: state of the art and future challenges," *Neural Computing and Applications*, 2016.

[17] F. Thabtah, W. Hadi, N. Abdelhamid, and A. Issa, "Prediction phase in associative classification mining," *International Journal of Software Engineering and Knowledge Engineering*, vol. 21, no. 6, pp. 855–876, 2011.

[18] F. Thabtah and S. Hammoud, "MR-ARM: a map-reduce association rule mining framework," *Parallel Processing Letters*, vol. 23, no. 3, Article ID 1350012, 2013.

[19] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.

[20] B. Azhagusundari and A. S. Thanamani, "Feature selection based on information gain," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 2, no. 2, pp. 18–21, 2013.

[21] S. Oreski and G. Oreski, "Genetic algorithm-based heuristic for feature selection in credit risk assessment," *Expert Systems with Applications*, vol. 41, no. 4, pp. 2052–2064, 2014.

[22] N. Hoque, D. K. Bhattacharyya, and J. K. Kalita, "MIFS-ND: a mutual information-based feature selection method," *Expert Systems with Applications*, vol. 41, no. 14, pp. 6371–6385, 2014.

[23] A. Fahad, Z. Tari, I. Khalil, I. Habib, and H. Alnuweiri, "Toward an efficient and scalable feature selection approach for internet traffic classification," *Computer Networks*, vol. 57, no. 9, pp. 2040–2057, 2013.

[24] A. Abbasi, S. France, Z. Zhang, and H. Chen, "Selecting attributes for sentiment classification using feature relation networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 3, pp. 447–462, 2011.

[25] E. Haddi, X. Liu, and Y. Shi, "The role of text pre-processing in sentiment analysis," *Procedia Computer Science*, vol. 17, pp. 26–32, 2013.

[26] C.-F. Tsai and Y.-C. Hsiao, "Combining multiple feature selection methods for stock prediction: union, intersection, and multi-intersection approaches," *Decision Support Systems*, vol. 50, no. 1, pp. 258–269, 2010.

[27] X.-X. Zhou, Y. Zhang, G. Ji et al., "Detection of abnormal MR brains based on wavelet entropy and feature selection," *IEEJ Transactions on Electrical and Electronic Engineering*, vol. 11, no. 3, pp. 364–373, 2016.

[28] A. Al-Thubaity, N. Abanumay, S. Al-Jerayyed, A. Alrukban, and Z. Mannaa, "The effect of combining different feature selection methods on arabic text classification," in *Proceedings of the 14th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD '13)*, pp. 211–216, July 2013.

[29] M. A. Hall, *Correlation-based feature selection for machine learning [Ph.D. thesis]*, The University of Waikato, 1999.

[30] M. Lichman, *UCI Machine Learning Repository, 2013*, vol. 114, 2015, http://archive.ics.uci.edu/ml.

[31] R. Mohammad, F. A. Thabtah, and T. McCluskey, "Phishing websites dataset," 2015, https://archive.ics.uci.edu/ml/datasets/Phishing+Websites.

[32] Millersmiles, 2015, http://www.millersmiles.co.uk/.

[33] "Phishtank. (2015). phishtank: Phising on the internet," http://www.phishtank.com/.

[34] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, 2005.