ISSN No. 0976-5697

Volume 4, No. 3, March 2013 (Special Issue)

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Enhancement of Fake Website Detection Techniques Using Feature Selection and Filtering Algorithms

Hetal Rahul Rajpura
P.G Student,Department of Information Technology
L.D College of Engineering Ahmedabad, India
hetallangalia@gmail.com

Hiteishi Diwanji
Assoc. Prof., Department of Information Technology
L.D College of Engineering Ahmedabad, India
hiteishi@hotmail.com

Abstract: Phishing websites are malicious websites created by fraud people to mimic real websites. Phishing websites impersonates legitimate websites to lure users into visiting the fake websites.They may lure by mailing you suspicious links that appear to be legitimate.Phishing is a security threat to the Internet, which causes tremendous loss every year. It helps generate billions of revenue for fraudsters. Attackers may steal user private information, credit-card and other significant financial information. In this paper , a study on various phishing websites detection techniques such as WHOIS, Browser-Integrated Anti-phishing toolbars,attaching new top level domains to existing blacklisted urls, feature extraction and classification based approach, a hybrid phish detection method based on information extraction and information retrieval, CANTINA, a content-based approach, purely based on the TF-IDF(term frequency /inverse document frequency) used in information retrieval algorithm , a novel classification method that identifies malicious web pages based on static attributes is presented.On detailed study and analysis,it is found that there exists classification approach based on the decision tree algorithm, which has mean absolute error of 0.292 which is comparatively larger than other algorithms, whereas its prediction accuracy is 98.5%. Hence, improvements can be done to decrease the mean absolute error of decision tree by using feature selection and filtering techniques to classify a revised set of attributes and design a sound and a
robust phishing detection system **.**

Keywords: Phishing, Detection, Techniques, Classifiers, Feature reduction, Filters, Security

I. INTRODUCTION

Phishing is a malicious activity where attackers try to lure users into visiting their fraudulent websites . Even though the web users are aware of these types of phishing attacks, lot of users become victim to these attacks. The phishing attack include stealing user's confidential infor-mation such as login username,password and credit card details. Successful phishing detection system would dist-inguish any phishing websites from legitimate websites.

Phishing is a social engineering technique used to deceive users and exploits the poor usability of current web security technologies. In order to lure the victim into giving up sensitive information the message might include imper-atives such as "verify your account" or "confirm billing info-mation". Once the victim has revealed the password, the attacker could access and use the victim's account for fraud-ulent purposes or spamming. Only specialists can identify these types of phishing websites immediately. But all the web users are not specialist in computer engineering and hence they become victim by providing their personal and financial details to the phishing artist. Thus, an efficient mechanism is required to identify the phishing websites from the legitimate websites in order to save credential data.

II. STUDY OF EXISTING FAKE WEBSITE DETECTION TECHNIQUES

A. WHOIS:

It is a "query and response" protocol that is widely used for querying databases that store the registered users or assignees of an Internet resource, such as a domain name, an IP address block, or an autonomous system, but is also used for a wider range of other information. The protocol stores and delivers database content in a human-readable format. This can help provide insight into a domain's history and additional information. WhoIs lookup can be used to see who owns a domain name, how many pages from a site are listed with Google or even search WhoIs address listings for a website's owner.

B. Browser-Integrated Anti-phishing toolbars:

Google Safe Browsing is a service provided by Google that provides lists of URLs for web resources that contain malware or phishing content[1]. The disadvantage of the approach is that non blacklisted phishing sites are not recognized. NetCraft tool bar [2],Netcraft provides a browser toolbar to report and block phishing sites identified by the toolbar user community. It provides a display of hosting location of a given website.(e.g. the main online banking site of a large US bank cannot be hosted in the former Soviet Union). Once you report a phishing URL, it is blocked. Netcraft supervisor validation is used to contain the impact of any false reporting of urls.It displays browser navigational controls

(toolbar & address bar) in all windows, to defend against pop up windows which attempt to hide the navigational controls to disguise location.

EBay tool bar [3],the eBay solution is designed for eBay and PayPal and involves the use of a so-called "Account Guard" that changes color if the user is on a spoofed site. Verisign provides a commercial antiphishing service [5].

McAfee SiteAdvisor [4], SiteAdvisor is a service that reports on the safety of web sites by crawling the web and testing the sites it finds for malware and spam. It includes automated crawlers that browse websites, perform tests and create threat ratings for each visited site . One popular solution to address this problem is to add additional security features within an Internet browser that warns users whenever a phishing site is being accessed. Such browser security is often provided by a mechanism known as 'blacklisting', which matches a given URL with a list of URLs belonging to a blacklist.Large companies such as Microsoft, McAfee and Google, maintain blacklists of phishing web sites.

C. Related Work:

In [10],blacklists are easily evaded by attackers. They attach new top level domains to existing URLs and modify them so that they are not available in blacklists and can easily perform identity theft,obtain financial details of internet users.This approach generates new URLs by using various heuristics and uses a matching algorithm further to match new URL with entries in the blacklist.

In[12],a method based on feature extraction of websites and classification approach is used. The classifier is trained with features extracted from a training set of legitim-ate and non-legitimate websites .Then an unknown website is tested by the classifier which checks if a website is fake or real.

System Overview:

a. Source code of non-legitimate website is captured.
b. Features are extracted ,identity extraction and feature extraction.
c. More focus on URL and Source code.

D. Feature Extraction:

The features extracted are listed in Table1 :

Table 1:

FEATURE	*DESCRIPTION*
IP Address	Usually phishing websites contain IP address as URL
Starts with http or https	If the url starts with https ,it cannot be a phishing page.
Alexatop5000	If the website is listed in alexatop5000,then it is a safe URL.
Dots in URL	The url in the source code should not contain more number of dots. If it contains more number of dots then it pretends to be a phishing website.
Slash in page address	The page address should not contain more number of slashes. If they contains more than five slashes then the url is considered to be a phishing url
Slash in url	The URL should not contain more number of slashes. If they contains more than five slashes then the url is considered to be a phishing url.
Use of @ Symbol	Presence of @ symbol in page address indicates that, all text before @ is comment. So the page url should not contain @ symbol.
"Whois" Lookup	The details of phishing website will not be available in "whois database. "Whois database is checked for the existence of the data pertaining to a particular website.
META Tag	The <meta> tag provides metadata about the HTML document. Meta elements are typically used to specify page description, keywords, and author of the document, last modified and other metadata.. If there is no relevance between the URL address and contents of the META tag it can be a phishing website.
META Keyword Tag	The META Keyword Tag provides keywords related to the web page which may be the identity of a web page. If there is no relevance between the URL address and contents of the META Keyword tag then it can be a phish.
Foreign Anchor	An anchor tag contains href attribute whose value is an url to which the page is linked with. Foreign anchor occurs if the domain name in the url is not same as the domain in page url. A website can contain foreign anchor. If number of foreign anchors exceeds then itis a sign of phishing website. Check all the anchor <a> tags.
Server Form Handler (SFH)	Forms enables user to pass data to a server. Action is the attributes of form tag, which specifies the url to which the data should be transferred. In the case of phishing website, . 1) The value of the action attribute of form tag comprise foreign domain, 2) value is empty, 3) value is #, 4) Value is void.
Foreign Request	Many a time,websites may request images, scripts, CSS files etc. from other websites. Phishing websites imitating the legitimate website will request these objects from the same page as legitimate one.Then in such a case ,the domain name used for requesting will not be similar to page url. Request urls are collected from the src attribute of the tags <img> and <script>, href attribute of link tag and code base attribute of object and applet tag. If the domain in these urls is foreign domain then the domain then,it is a phishing website.
Blacklist	Blacklist contains list of suspected websites. It is a third party service. The page url is checked against the blacklist. If the page url is present in the blacklist, it means it is a phishing website .

The following, Table II[17] shows us the various parameters of the classification algorithms namely ,MultiLayer Perceptron and the Decision Tree(J48) and Naïve Bayesian. The prediction accuracy is measured as the ratio of number of correctly classified instances in the test dataset and the total number of test cases.

Table 2:

Evaluation Criteria	***MLP***	***J48***	***NB***
Kappa statistic	0.96	0.97	0.96
Mean Absolute Error	0.0397	0.292	0.0253
Root Mean Squared error	0.1487	0.1216	0.1285
Relative absolute error	7.9487	5.8302	5.0518
Root relative square error	29.7347	24.313	25.6924

Table 3:

Evaluation Criteria	***MLP***	***J48***	***NB***
Time taken to build model(secs)	0.87	0.03	0
Correctly classified Instances	194	197	187
Incorrectly classified Instances	6	3	13
Prediction accuracy	97%	98.5%	93.5%

A solution proposed by [15] describes a hybrid phish detection method .It is based on the information extraction (IE) and information retrieval (IR) techniques. The identity-based component of method detects phishing webpages by directly comparing the inconsistency between their identity and the identity they are claiming. The keywords-retrieval component uses the IR algorithms and exploits the power of search engines to detect phishing websites.This method requires no training data, no prior knowledge of phishing signatures,and thus is a robust method to detect new phishing patterns. CANTINA [13] is a content-based approach to detect phishing websites, It is purely based based on the TF-IDF(term frequency/inverse document frequency) used in information retrieval algorithm.It more specifically focuses on the Robust Hyperlinks algorithm previously developed for overcoming broken hyperlinks.

Robust Hyperlink:If a particular page is not found with its basic URL,then we form a lexical signature that is a composition of 5 words with highest tf-idf value and enter the lexical signature in a search engine to locate a robust hyperlink whose signature closely matches to our lexical signature .If no such link is found then the URL is of a fake website ,else a legitimate one.CANTINA looks for the content of a web page to determine whether it is legitimate or not, in contrast to other approaches that look at other characteristics of a webpage, for example the URL and its domain name.Results show that CANTINA is good at detecting phishing sites, detecting 94-97% of phishing sites. 1The solution described in [16] presents a novel classification method that identifies malicious web pages based on static attributes. It analyzes the underlying static attributes of the initial HTTP response and HTML code. Static attributes that characterize malicious actions can be used to identify a majority of malicious web pages. It makes use of a generic classifier ,high-interaction client honeypots and this new classification method into a hybrid system leads to significant performance improvements.

III. FEATURE REDUCTION AND FILTERING TECHNIQUES

To improve accuracy and reduce errors, feature selection algorithms are used to filter out relevant features by feature reduction and feature selection.The main idea of feature subset selection(FS) is to remove redundant or irrelev-ant features from the data set as they can lead to a reduction of the classification accuracy and to an unnecessary increase of computational cost. The advan-tage of FS is that no information about the importance of single features is lost. On the other hand , if a small set of features is required and the original features are very diverse, information may be lost as some of the features must be omitted.With dimensionality reduction techniques, the size of the attribute space can often be decreased strikingly without loosing a lot of information of the original attribute space. An important disadvantage of DR is the fact that the linear combinations of the original features are usually not interpretable and the information about how much an original attribute contributes is often lost.

A. Feature Reduction Techniques:

a. ***Correspondence Analysis:*** Multivariate statistical technique. Conceptually similar to principal component analysis, but applies to categorical rather than continuous data.It provides a means of displaying or summarizing a set of data in two-dimensional graphical form.

b. ***Canonical Discriminant Analysis:*** Is a dimension reduction technique related to principal component analysis and canonical correlation. Given a nominal classification variable and several interval variables, canonical discriminant analysis derives canonical variables (linear combinations of the interval variables) that summarize between-class variation in much the same way that principal components summarize total variation.

c. ***Principal Component Analysis:*** Is a dimensionality reduction technique which enables to visualize a dataset in a lower dimension without loss of information.

It is appropriate when obtaining measures on a number of observed variables and wish to develop a smaller number of artificial variables (called principal components) that will account for most of the variance in the observed variables.

B. *Filters:*

Filters are classifier agnostic pre-selection methods which are independent of the later applied machine learning algorithm. Besides some statistical filtering methods like Fisher score or Pearson correlation , information gain, originally used to compute splitting criteria for decision trees, is used to find out how well each single feature separates the given data set.

The overall entropy I of a given dataset S is defined as :

$$I(S) = -\sum_{i=1}^{C} Pi(logP\ i)$$

Where C denotes the total number of classes and Pi the portion of instances that belong to class i.

C. *Feature Selection Algorithms:*

- ***a. Fisher filtering:*** Is a supervised feature selection algorithm which processes the selection independently from the learning algorithm. It follows univariate Fisher's ANOVA ranking which ranks the inputs attributes according to their relevance without considering the redundancy aspects of input attributes.
- ***b. ReliefF Algorithm:*** Detects conditional dependencies between attributes and provides a unified view on the attribute estimation in regression and classification. It is not limited to two class problems, is more robust and can deal with incomplete and noisy data.
- ***c. STEPDISC (Stepwise Discriminant Analysis):*** Procedure performs a stepwise discriminant analysis to select a subset of the quantitative variables for use in discriminating among the classes.The set of variables that make up each class is assumed to be multivariate normal with a common covariance matrix. The STEPDISC procedure can use forw-ard selection, backward elimination,or stepwise selection .
- ***d. Runs filtering:*** A non parametric test for predictive attribute evaluation.It is an univariate attribute ranking from runs test.It is a supervised feature selection algorithms based upon a filtering approach i.e. processes the selection independently from the learning algorithm. This component ranks the inputs attributes according to their relevance without considering redundancy aspect. A cutting rule enables to select a subset of these attributes.

After performing feature relevance analysis, various classification algorithms are applied over this training dataset on the relevant attributes after filtration.

D. *Dimensionality Reduction:*

Dimensionality reduction (DR) refers to algorithms and techniques "***which create new attributes as combinations of the original attributes in order to reduce the dimension-ality of a data set"*** .The most important DR technique is the principal component analysis (PCA), which produces new attributes as linear combinations of the original variables. In contrast, the goal of a factor analysis is to express the original attributes as linear combinations of a small number of hidden or latent attributes .The factor analysis searches for underlying (i. e. hidden or latent) attributes that summarize a group of highly correlated attributes.

Principal Component Analysis**:** The goal of PCA is to find a set of new attributes (PCs) which meets the following criteria:

The PCs are

a) linear combinations of the original attributes,
b) orthogonal to each other
c) capture the maximum amount of variation in the data.

Often the variability of the data can be captured by a relatively small number of PCs, and, as a result, PCA can achieve high dimensionality reduction with usually lower noise than the original patterns. The principle components are not always easy to interpret, and, in addition to that, PCA depends on the scaling of the data.

a. Mathematical background:

The covariance of two attributes is a measure how strongly the attributes vary together. The covariance of two random variables x and y of a sample with size n and mean x, y can be calculated as:

$$Cov(x, y) = \left(\frac{1}{n-1}\right) \Sigma \quad (x - x_m)\ (y\ - y_m)$$

When x and y are normalized by their standard deviations _x and _y, then the covariance of x and y is equal to the 1

Four main properties of the PCA:

(i) Each pair of attributes has covariance 0,
(ii) The attributes are ordered descendingly with respect of their variance,
(iii) The first attribute captures as much of the variance of the data as possible,
(iv) each successive attribute captures as much of the remaining data as possible.

b. Parameter to work on:

Mean absolute error (MAE): For each instance in the test set, Weka obtains a dist-ribution. This distribution is matched against the expe-cted distribution. For each class label the absolute error is calculated. Sum of the absolute error of all the labels gives absolute error of instance. The mean absolute error is "*the sum over all the instances and their absolute error instance divided by the number of instances in the test set"* with an actual class.In statistics, the mean absolute error (MAE) is a quantity used to measure how close forecasts or predictions are to the

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i|.$$

eventual outcomes. The mean absolute error is given by

As the name suggests, the mean absolute error is an average of the absolute errors $e_i = |f_i - y_i|$, where f_i is the prediction and y_i the true value. The mean absolute error is a common measure of forecast error in time series

analysis, where the terms "mean absolute deviation" is sometimes used in confusion with the more standard definition of mean absolute deviation. The same confusion exists more generally.Decreasing the mean absolute error,will decrease the absolute error of each instance ,further will lead to increase in prediction accuracy.Work will be done to decrease the MAE to increase the prediction accuracy.

IV. DECISION TREE INDUCTION(J48)

A decision tree is a predictive machine-learning model that decides the target value (dependent variable) of a new sample based on various attribute values of the available data. The internal nodes of a decision tree denote the different attributes, the branches between the nodes tell us the possible values that these attributes can have in the observed samples, while the terminal nodes tell us the final value (classification) of the dependent variable.

The attribute that is to be predicted is known as the dependent variable, since its value depends upon, or is decided by, the values of all the other attributes. The other attributes, which help in predicting the value of the dependent variable, are known as the independent variables in the dataset.

The J48 Decision tree classifier follows the following simple algorithm. In order to classify a new item, it first needs to create a decision tree based on the attribute values of the available training data. So, whenever it encounters a set of items (training set) it identifies the attribute that discriminates the various instances most clearly. This feature that is able to tell us most about the data instances so that we can classify them the best is said to have the highest information gain. Now, among the possible values of this feature, if there is any value for which there is no ambiguity, that is, for which the data instances falling within its category have the same value for the target variable, then we terminate that branch and assign to it the target value that we have obtained. See Fig 1,an example of decision tree,

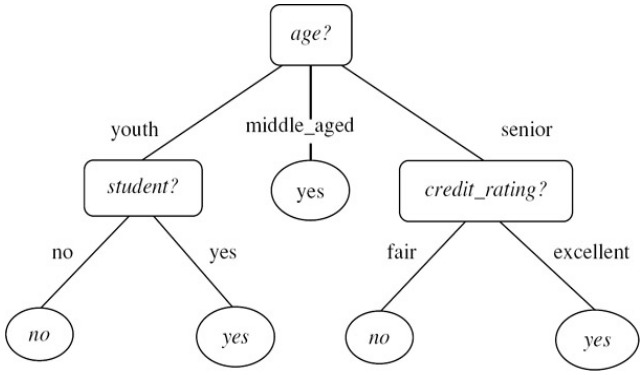

Figure 1:A decision tree

A. *Decision Tree Algorithm:*

a. Algorithm: Generate a decision tree from the training tuples of data partition D.

Input:

a) Data partition,D, which is a set of training tuples and their associated class labels.
b) Attribute_list , the set of candidate attributes;
c) Attribute_selection_method, a procedure to determine the splitting criterion that "best"partitions the data tuples into individual classes. This criterion consists of a splitting attribute and, possibly a split-point or splitting subnet.

Output:

a) A Decision tree

Method:

a) Create a node N;
b) f tuples in D are all of the same class C,then
c) return N as a leaf node labeled with the class C,
d) if attribute_list is empty then
e) return N as a leaf node labeled with the majority class in D;//majority voting
f) apply Attribute_selection_method(D,attribute_list)to find out the best Splitting_criterion
g) label node N with splitting_critrion,
h) if splitting_attribute is discrete-valued and Multiway splits allowed then//not restricted to binary trees
i) Attribute_list←Attirbute_list-splitting_criterion;
j) for each outcome j of splitting _criterion //partition the tuples and grow subtrees for each partition
k) let Dj be the set of data tuples in D satisfying the outcome j;//a partition
l) if Dj is empty then
m) attach a leaf labeled with the majority class in D to node N;
n) else attach the node returned by Generate.decision.tree(Dj,attribute_list) to node N; Endfor
o) return;

V. ANALYSIS

As it can be seen from the Table II , Decision tree algorithm's prediction accuracy is 98.5% whereas its mean absolute error is 0.292 which is substantially greater than Multilayer Perceptron and Naives Bayesian algorithms . Hence if improvements are made to decrease the mean abso-lute error of decision tree algorithm ,then we can develop a robust phishing system using Decision tree algorithm .It is found that applying feature selection and filtering techniques to extracted features can reduce the redundancy and highly correlated features can be used for classification.This new approach will be tested for a set of phishing and legitimate websites.

VI. PROPOSED MODEL

PHP will be used to do feature extraction. Extracted features will be input to the Feature Selection and and filtering module.Revised features will now be fed to the classifier which will take the decision stating whether the given instance is a fake or a real one .See Fig. 2

Model Designed:

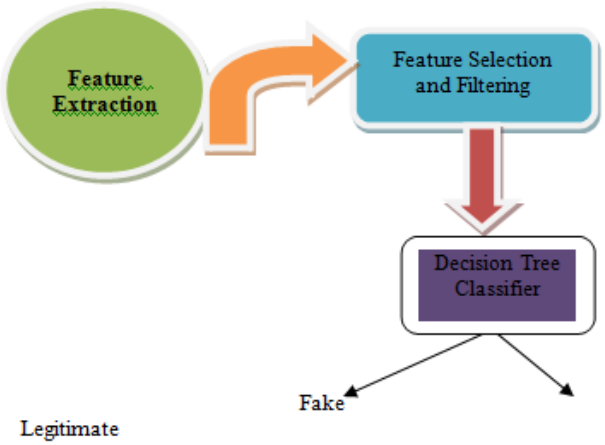

Figure 2:Proposed Model

VII. CONCLUSION

In this paper, different techniques to detect fake websites have been presented. As we have seen, there exists a huge number of techniques to detect fake websites, but still there are false positives present. Analysis shows that improve-ements are needed to decrease the mean absolute error of Decision Tree algorithm. It is found that feature selection and reduction algorithms can be used to filter out a subset of attributes which can be fed to the classifier to further decrease the mean absolute error and thereby increase the prediction accuracy of the decision tree algorithm.

VIII. ACKNOWLEDGMENT

Mrs.Hetal Rajpura wishes to acknowledge Prof. Hiteishi Diwanji for showing her the right path to carry out her research work and for her constant support and guidelines and to all the staff members of Computer Department, L. D. College of Engineering for extending their kind support throughout the work. Prof.Hiteishi Diwanji wishes to acknowledge her family, and all the staff members at L. D. College of Engineering .

IX. REFERENCES

[1]. Google. Google Toolbar for Firefox, http//www.google.com/tools/firefox/toolbar/FT3/intl/en/, 2006

[2]. NetCraft. Netcraft anti-phishing tool bar. http://toolbar.netcraft.com, 2007.

[3]. eBay. eBay tool bar. http://pages.ebay.com/ebaytoolbar/, 2007.

[4]. McAfee. McAfee SiteAdvisor. http://www.siteadvisor.com, 2007.

[5]. Verisign. Anti-Phishing Solution. http://www. .verisign.com/verisign-business-solutions/anti-phishing-solutions/, 2005.

[6]. Y. Cao, W. Han, and Y. Le, Anti-phishing based on automated individual white-list, In 4th ACM Workshop on Digital Identity Management (DIM 2008), October 2008.

[7]. E. Ukkonen, Approximate string-matching with q-grams and maximal matches, In Theoretical Computer Science 92, 1992.

[8]. Ian H. Witten and Eibe Frank. ”Data Mining: Practical machine learning tools and techniques”. Morgan Kaufmann, 2nd edition edition, 2005.

[9]. Weka, http://www.cs.waikato.ac.nz/ml/weka/

[10]. P. Prakash, M. Kumar, R. R. Kompella, and M. Gupta, Phishnet:Predictive blacklisting to detect phishing attacks, In IEEE Infocom Mini-Conference, 2010

[11]. http://www.statsoft.com/textbook/support-vector-machines/

[12]. Insoon Jo, Eunjin (EJ) Jung, Heon Y. Yeom, “ You’re not who you claim to be: website identity check for phishing detection”

[13]. Y. Zhang, J. Hong, and L. Cranor, Cantina: A content-based approach to detecting phishing web sites, In the 16th International Conference on World Wide Web, May 2007

[14]. Wikipedia, Phishing, http://en.wikipedia.org/wiki/Phishing.

[15]. G. Xiang and J. I. Hong, A hybrid phish detection approach by identity discovery and keywords retrieval, In In Proceedings of 18th International World Wide Web Conference, 2009

[16]. C. Seifert, I. Welch, and P. Komisarczuk. Identification of malicious webpages with static heuristics. In Telecommunication Networks and Applications Conference, 2008. ATNAC 2008. Australasian, pages 91 {96, 2008.

[17]. Santhana Lakshmi V, Vijaya MS, Efficient prediction of phishing websites using supervised learning algorithms, International Conference on Communication Technology and System Design 2011

[18]. Nir Friedman, Moises Goldszmidt,Building Classifiers using Bayesian Networks, AAAI-96 Proceedings

[19]. http://www.d.umn.edu/~padhy005/Chapter5.html