Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 30 (2012) 798 – 805

Procedia Engineering

www.elsevier.com/locate/procedia

International Conference on Communication Technology and System Design 2011

Efficient prediction of phishing websites using supervised learning algorithms

Santhana Lakshmi V[a], Vijaya MS[b], a*

[a] *P.S.G.R Krishnammal College for Women,Coimbatore-641004,India*
[b] *G.R.Govindarajalu School of Applied Computer Technology, Coimbatore-641004, India.*

Abstract

Phishing is one of the luring techniques used by phishing artist in the intention of exploiting the personal details of unsuspected users. Phishing website is a mock website that looks similar in appearance but different in destination. The unsuspected users post their data thinking that these websites come from trusted financial institutions. Several antiphishing techniques emerge continuously but phishers come with new technique by breaking all the antiphishing mechanisms. Hence there is a need for efficient mechanism for the prediction of phishing website. This paper employs Machine-learning technique for modelling the prediction task and supervised learning algorithms namely Multi layer perceptron, Decision tree induction and Naïve bayes classification are used for exploring the results. It has been observed that the decision tree classifier predicts the phishing website more accurately when comparing to other learning algorithms.

Keywords: Antiphishing; Classification; Machine learning; Phishing; Prediction; Supervised learning.

1. Introduction

The growth of the phishing websites seems to be astonishing. Even though the web users are aware of these types of phishing attacks, Lot of users become victim to these attacks. Numbers of attacks are launched with the aim of making web users believe that they are communicating with a trusted entity. Phishing is one among them. Communications from popular web sites, auction sites, online payment processors are commonly used as a source to lure the unsuspecting public. Phishing websites are mock websites that looks similar to legitimate. 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 details to the phishing artist. Phishing is continuously evolving since it

* Santhana Lakshmi.V,
Email address: sanlakmphil@gmail.com.

1877-7058 © 2011 Published by Elsevier Ltd. Open access under CC BY-NC-ND license.
doi:10.1016/j.proeng.2012.01.930

is easy to copy an entire website using the HTML source code. By making slight changes in the source code, it is possible to direct the victim to the phishing website. Phishers use lot of techniques to lure the unsuspected web user. They send generic greetings to the customers to check their account immediately. They also send threat messages indicating to update their account immediately otherwise their account will be cancelled. Thus an efficient mechanism is required to identify the phishing websites from the legitimate websites in order to save credential data.

Various methodologies are being adopted at present to identify phishing websites. Maher Aburous et, al. proposes an approach for intelligent phishing detection using fuzzy data mining. In [1], e-banking phishing website detection rate is performed based on six criteria: URL & Domain Identity, Security & Encryption, Source Code & Java script, Page Style & Contents, Web Address Bar, and Social Human Factor. Fuzzy logic and data mining algorithms are used to categorize e-banking phishing websites. Ram basnet et al. adopts machine learning approach for detecting phishing attacks [2]. Support vector machine, biased support vector machine and neural network are used for the efficient prediction of phishing e-mails. The main intention of this paper is to classify phishing emails by incorporating key structural features in phishing emails and employing different machine learning algorithms for the classification process.

Ying Pan and Xuhus Ding used discrepancies that exist in the website's identity, structural features and HTTP transactions to detect the mock website. It demands neither user expertise nor prior knowledge of the website. Support vector machine is used as page classifier. The main features of this approach includes: a) it does not rely on any prior knowledge of the server or users' security expertise; b) the adversary has much less adaptability since the detection is independent of any specific phishing strategy; c) it causes no changes on users' existing navigation behavior[3]. Anh Le, Athina Markopoulou, University of California used lexical features of the url to predict the phishing website . Classification accuracy of using lexical features is compared with accuracy of using automatically selected and hand selected features and compared with additional features. Machine learning algorithms used for prediction includes Support Vector Machine, Online Perceptron, etc [4].

In this paper, machine-learning algorithms have been used for modelling the prediction task. Training the features of phishing and legitimate websites creates the learning model. Third party services such as balcklist, search engine that contributes more for the accurate prediction of the phishing websites are included as one of the features that are used to identify the phishing websites. Supervised learning algorithms namely Multi layer perceptron(MLP), Decision tree induction(DT) and Naïve bayes(NB) classification are used for learning. The process of identity extraction and feature extraction are described in the following section and the various experiments carried out to discover the performance of the models are demonstrated in the rest of this paper.

2.System Overview

Phishing websites are replica of legitimate website. This is possible because of the HTML which is used for designing websites. Prior to capturing these websites, their source code is captured and parsed for Dom objects. Identity of these websites is extracted from the Dom objects. The main phase of this phishing website prediction system is identity extraction and feature extraction. Features that contribute much for the accurate prediction of phishing website are extracted from the url and HTML source code. In order to make the model more efficient, the page url is checked for the presence of more number of slashes. This paper seeks the usage of third party service named 'Blacklist' for predicting the website accurately. Blacklist contains the list of phishing and suspected websites. The page url is checked against 'Blacklist' to verify whether the url is present in the blacklist

2.1 Identity Extraction

The aim of identity extraction is to extract the identity of a web page. Identity of a web page is a set of words that uniquely identifies the ownership of the website. Even though phishing artist can create and design replica of website, there are some identity relevant features which cannot be exploited. The change in these features affects the similarity of the website. Therefore these features are useful to find the identity of the web page. Features extracted in identity extraction phase include META Title, META Description, META Keyword, HREF of <a> tag.

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.

The Meta description tag is a snippet of HTML code that comes in the Head section of a web page. It will be placed before the Meta keywords tag. The identity relevant object is the value of the content attribute in Meta tag. It consists of a description about the website.

The META Keyword Tag is where you list keywords and keyword phrases that you've targeted for that specific page.The value of the content attribute provides keywords related to the web page which may be the identity of a web page.

HREF Tag:

The href attribute specifies the destination of a link. When a hyperlink text is selected, it has to direct to the concerned web page. Phishers will not perform any change in the destination site address. So it points to the legitimate website. The value of the href attribute is a URL in which the domain name has high probability to be the identity of the website.

Once the identity relevant features are extracted, they are converted into individual terms by removing the stop words such as http, www, in, com, etc., and by removing the words with length less than three. tf-idf weight is evaluated for each of the keywords. The first five keywords that have high tf-idf value are selected for identity set. tf-idf value is calculated using the following formula.

$$tf_{ij} = \sqrt{\frac{n_{ij}}{\sum_k n_{kj}}} \qquad \text{(Eq.1)}$$

Where n_{ij} is the number of occurrence of t_i in document d_j and $\sum_k n_{kj}$ is the number of all terms in document d_j.

$$idf_i = ln\left(\frac{|D|}{|\{d_j : t_i \sum d_j\}| + 1}\right) \qquad \text{(Eq.2)}$$

Where |D| is the total number of documents in a dataset, and $\{|dj:t_i \in dj\}|$ is the number of documents.where term ti appears. To find the document frequency of a term, WebAsCorpus is used. It is a readymade frequency list. The total number of documents in which the term appears is the term that has the highest frequency. The highest frequency term is assumed to be present in all the documents. The tf-idf weight is calculated using the following formula

$$tf\text{-}idf = tf_{ij}.idf_i \qquad \text{(Eq.3)}$$

2.2 Feature extraction and Vector Generation

Feature extraction plays an eminent role for the efficient prediction of phishing websites. In a HTML source code there are many factors that can distinguish the original legitimate website from the forged websites. Those factors are extracted. The two features such as 'server form handler' and 'Whois lookup' are very much essential for detecting phishing websites but are not taken into account in [14]. The main aim of the phishing websites is to acquire the personal data from the user. Server form handler denotes the location where the personal data given by the user are transferred. So checking the value of action attribute is essential to know the destination of the user specified data. 'Whois' database provides all the information about the registered customers who owns the website. All the legitimate websites' details will be present in 'Whois' database. Since phishing websites are short-lived websites, they will not register and their details will not be available in 'Who is' database. So it is essential to check the 'Whois' database. Prediction accuracy shown in [14] is only 97.33%, which has been increased to 98.5% in this work by taking into consideration the above two features.

Feature1: Foreign Anchor

An anchor tag contains href attribute whose value is an url to which the page is linked with. If the domain name in the url is not similar to the domain in page url then it is called as foreign anchor. A website can contain foreign anchor. But too many foreign anchor is a sign of phishing website. So all the <a> tags in the webpage are collected. And they are checked for foreign anchor. If the number of foreign domain exceeds, then the feature F_1 is assigned to -1 else F_1 is assigned as 1.

Feature2: Nil Anchor

Nil anchor denotes that the page is linked with none. The value of the href attribute of <a> tag will be null. The values that denote nil anchor are about: blank, javascript:; JavaScript: void(0),#. If these values exist then the feature F_2 is assigned the value of –1.Instead if the anchor is not a nil anchor F_2 is assigned as 1.

Feature3: IP Address

The main aim of phishers is to gain lot of money with no investment and they will not invest to buy domain names for their fake website. Most phishing websites contain IP address as their domain name. If the domain name in the page address is an IP Address then the value of the feature F_3 is –1 else the value of F_3 is 1.

Feature 4 and 5: Dots in Page Address and Dots in URL

The page address and url in the source code should not contain more number of dots. If they contains more number of dots then it is the sign of phishing website. If the page address contains more than five dots then the value of the feature F_4 is -1 or else the value of F_4 is 1.All the url's in the source code are checked for more number of dots if they contain F_5 is –1 or else F_5 is 1.

Feature 6 and 7: Slash in page address and url:

The page address and 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 and the value of F_6 is assigned as –1. If the page address contains less than 5 slashes, the value of F_6 is 1. Similarly for all the url's in the source code number of slashes are checked and if they contain more number of slashes F_7 is –1 else F_7 is 1.

Feature 8: Foreign Anchor in Identity Set

If the website is legitimate, then both the url and the page address will be similar and it will be present in the identity set. But while considering phishing website, the domain of the url and the page

address will not be same and domain name will not be contain in identity set. If the anchor is not a foreign anchor and is present in identity set then the value of F_8 is 1.If the anchor is a foreign anchor but present in the identity set then also the value of F_8 is 1.If the anchor is a foreign anchor and not present in the identity set then the value of F_8 is –1.

Feature 9: Using @ Symbol

Presence of @ symbol in page address indicates that, all text before @ is comment. So the page url should not contain @ symbol. If the page url contains @ symbol, the value of F_9 is –1 else F_9 is 1.

Feature10: Server Form Handler (SFH)

Forms are used to pass data to a server. Action is one of the attributes of form tag, which specifies the url to which the data should be transferred. In the case of phishing website, it specifies the domain name, which embezzles the credential data of the user. Even though some legitimate websites use third party service and hence contain foreign domain, it is not the case for all the websites. The value of the feature F_{10} is –1, if the following conditions hold. 1) The value of the action attribute of form tag comprise foreign domain, 2) value is empty, 3) value is #, 4) Value is void. If the value of the action attribute is its own domain then, F_{10}= 1.

Feature 11: Foreign Request

Websites request images, scripts, CSS files from other websites. Phishing websites to imitate the legitimate website request these objects from the same page as legitimate one. 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>, background attribute of body tag, 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 value of F_{11} is –1 else F_{11} is 1.

Feature 12: Foreign request urls in Identity set:

If the website is legitimate, the page url and url used for requesting the objects such as images, scripts etc., will be same and the domain name will be present in the identity set. Request urls are checked for their existence in identity set. If they exist the value of F_{12} is 1.If they does not exist in the identity set the value of F_{12} is –1.

Feature 13: Cookie

Web cookie is used by an origin website to send state information to a user's browser and for the browser to return the state information to the origin site. In simple it is used to store information. The domain attribute of cookie holds the server domain, which set the cookies. It will be a foreign domain for phishing website. If the value of the domain attribute of cookie is a foreign domain then F_{12} is -1 otherwise F_{13} is 1. Some websites do not use cookies. If no cookies found then F_{13} is 2.

Feature 14: SSL Certificate

SSL is an acronym of secure socket layer. It creates an encrypted connection between the web server and the user's web browser allowing for private information to be transmitted without the problems of eavesdropping. All legitimate websites will have SSL certificate. But phishing websites do not have SSL certificate. The SSL certificate of a website is extracted by providing the page address. If SSL certificate exists then the value of the feature F_{14} is 1. If there is no SSL certificate then the value of F_{14} is -1.

Feature 15: Search Engine

If the website is legitimate and if the page url is given to any search engine, the first 10 results produced will be about the concerned website. If the page url is fake, the results will not be related to the concerned website. If the first 5 results from the search engine is similar to the page url then F_{15} is 1 or else F_{15} is -1.

Feature 16: 'Whois' Lookup

'Whois' is a request response protocol used to fetch the registered customer details from the database. The database contains the information about the registered users such as registration date, duration, expiry date etc. The legitimate site owners are the registered users of 'whois' database. 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. If exists then the value of F_{16} is 1 or else the value is –1.

Feature 17: 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 is considered to be a phishing website and the value of F_{17} is assigned as –1 or else the value is 1.Thus a set of 17 features are extracted from the HTML source code and url of a website by developing PHP code and the feature vectors are generated for all the websites.

3 Supervised Learning Algorithms

Supervised learning is the machine learning task of inferring a function from supervised training data. The training data consist of a set of training examples. In supervised learning, each example is a pair consisting of an input object and a desired output value called the supervisory signal. A supervised learning algorithm analyzes the training data and produces an inferred function, which is called a classifier. The classifier is then used for predicting the accurate output value for any valid unseen input object. The three classification algorithms used for learning the website data namely Multilayer perceptron, Decision tree induction, Naive Bayes are briefed below.

3.1 Multi Layer Perceptron

Multilayer Perceptron network is the most widely used neural network classifier. MLP networks are general purpose, nonlinear models consisting of a number of units organized into multiple layers. The complexity of the MLP network can be changed by varying the number of layers and the number of units in each layer. Given enough hidden units and enough data, it has been shown that MLPs can approximate virtually any function to any desired accuracy.

3.2 Decision Tree Induction

Decision Tree Classification generates the output as a binary tree like structure called a decision tree. A Decision Tree model contains rules to predict the target variable. This algorithm scales well, even where there are varying numbers of training examples and considerable numbers of attributes in large databases. J48 algorithm is an implementation of the C4.5 decision tree learner. The algorithm uses the greedy technique to induce decision trees for classification [12]. A decision-tree model is built by analysing training data and the model is used to classify unseen data.

3.3 Naïve Bayes

The Naive Bayes classifier is designed for use when features are independent of one another within each class, but it appears to work well in practice even when that independence assumption is not valid. It classifies data in two steps (a) Using the training samples, the method estimates the parameters of a probability distribution, assuming features are conditionally independent given the class.(b) For any unseen test sample, the method computes the posterior probability of that sample belonging to each class. The method then classifies the test sample according the largest posterior probability.

4. Experiments and Results

The phishing website prediction model is generated by implementing supervised learning algorithms. The dataset used for learning is collected from PHISHTANK [6]. It is an archive consisting of collection of phishing websites. The dataset with 100 phishing websites and 100 legitimate websites is developed for implementation. The features describing the properties of websites are extracted as described in section 2 and the size of each feature vector is 17. The feature vector corresponding to phishing website is assigned a class label -1 and +1 is assigned to legitimate website.

The classification algorithms, Multi Layer Perceptron(MLP), Decision tree Induction(J48) and Naïve Bayes(NB) are implemented and trained using WEKA. The Weka, Open Source, Portable, GUI-based workbench is a collection of state-of-the-art machine learning algorithms and data pre processing tools [7] [8]. The robustness of the classifiers is evaluated using 10–fold cross validation. Predictive accuracy is used as a primary performance measure for predicting the phishing website and is measured as the ratio of number of correctly classified instances in the test dataset and the total number of test cases. The performances of the trained models are evaluated based on the two criteria, the prediction accuracy and the training time. The prediction accuracy of the models is compared. The 10-fold cross validation results of the three classifiers MLP, J48 and NB are summarized in Table 1 and Table 2.

Table-1 Comparison of Estimates

Evaluation Criteria	Classifiers		
	MLP	J48	NB
Kappa ststistic	0.96	0.97	0.96
Mean Absolute Error	0.0397	0.292	0.0253
Root Mean Squarred 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-2 Performance comparison of classifiers

Evaluation Criteria	Classifiers		
	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%

The performance evaluation based on kappa statistics, mean absolute error, root mean squared error, relative absolute error and root relative squared error is shown in table-1. Kappa is a chance-corrected measure of agreement between the classifications and the true classes. It is calculated by taking the agreement expected by chance away from the observed agreement and dividing by the maximum possible agreement. For each instance in the test set, Weka obtains a distribution. This distribution is matched against the expected 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.

The root mean squared error is the difference between forecast and corresponding observed values. Each values are squared and then averaged over the sample. Finally, the square root of the average

is taken. Since the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors. The relative absolute error takes the total absolute error and normalizes it by dividing by the total absolute error of the simple predictor. The root relative squared error is relative to what it would have been if a simple predictor had been used. More specifically, this simple predictor is just the average of the actual values. Thus, the relative squared error takes the total squared error and normalizes it by dividing by the total squared error of the simple predictor. By taking the square root of the relative squared error one reduces the error to the same dimensions as the quantity being predicted.

From Table-2 it is found that the time taken to build the model and the prediction accuracy is high in case of decision tree induction when compared to other two algorithms

5. Conclusion

This work models the phishing website prediction as a classification task and demonstrates the machine learning approach for predicting whether the given website is legitimate website or phishing. Naïve Bayes classifier, Decision tree classifier, Multilayer perceptron have been applied for training the prediction model. Features have been extracted from a set of 200 url and the corresponding HTML source code of phishing and legitimate websites and the training dataset has been prepared in order to facilitate training and implementation. The performance of the models has been evaluated using 10-fold cross validation and two performance criteria, predictive accuracy and ease of learning. From the results it has been found that the decision tree classifier performs well than the other two models. It is hoped that more interesting results will follow on further exploration of data.

References

[1] Hossain M.A, Keshav Dahal, Maher Aburrous, “Modelling Intelligent Phishing Detection System for e-Banking using Fuzzy Data Mining”.

[2] Andrew H.Sung, Ram Basenet, Srinivas Mukkamala, “Detection of Phishing Attacks: A machine Learning Approach”.

[3] Ying Pan, Xuhus Ding “Anomaly Based Phishing page Detection”.

[4] Anh Le,Athina Markopoulou,Michalis Faloutsos “PhishDef: URL Names Say it All”.

[5] Troy Ronda ,Stefan Sarolu,Alec Wolman “iTrustpage:A User-Assisted Anti-Phishing Tool”.

[6] www.phishtank.com

[7] Ian H. Witten, Eibe Frank, “Data Mining – Practical Machine Learning Tools and Techniques”, 2005, Elsevier.

[8] lan H. Witten, Eibe Frank, Len Trigg, Mark Hall, Geoffrey Holmes, Sally Jo Cunningham, “Weka: Practival Machine Learning Tools and Techniques with Java Implementations” , Working Paper 99/11, Department of Computer Science, The University of Waikato, Hamilton, 1999.

[9] T. Mitchell ”Machine learning” Ed. Mc Graw-Hill International edition.

[10] http://webascorpus.org/WebCorpus2006/WebCorpus2006_min100.html.

[11] Georg Dorffner,Horst Bischof,Kurt Hornik(EDs.), “Artificial Neural Networks-ICANN 2001” International Conference Vienna,Austria,August 2001 proceedings.

[12] M. Chen, A. X. Zheng, J. Lloyd, M. I. Jordan, and E. Brewer, “Failure diagnosis using decision trees”, In Proc. IEEE ICAC, 2004.

[13] Jason Hong, Lorrie Cranor, Yue Zhang, “CANTINA: A Content Based Approach to Detecting Phishing Web Sites.

[14] Adi Sutanto , Jui-Lin Lai, Muhammad Khurram Khan, Mingxing He, Pingzhi, Rong-Jian Chen, , Ray-Shine Run, Shi-Jinn Horng, ”An efficient phishing webpage detector”.

[15] Andrew Donkin, Geoffrey Homes, lan H.Witten, “WEKA: a machine learning workbench”.