Detecting Phishing Websites Using Associative Classification

Moh'd Iqbal AL Ajlouni[1*], Wa'el Hadi[2], Jaber Alwedyan[3]

1. Dept of Business Administration, Al-Zaytoonah University, Jordan, m.alajlouny@zuj.edu.jo
2. Dept of MIS, University of Petra, whadi@uop.edu.jo
3. ITC Dept, Arab Open University, j.alwedyan@arabou.edu.sa

* E-mail of the corresponding author: m.alajlouny@zuj.edu.jo

Abstract

Phishing is a criminal technique employing both social engineering and technical subterfuge to steal consumer's personal identity data and financial account credential. The aim of the phishing website is to steal the victims' personal information by visiting and surfing a fake webpage that looks like a true one of a legitimate bank or company and asks the victim to enter personal information such as their username, account number, password, credit card number, …,etc. This paper main goal is to investigate the potential use of automated data mining techniques in detecting the complex problem of phishing Websites in order to help all users from being deceived or hacked by stealing their personal information and passwords leading to catastrophic consequences. Experimentations against phishing data sets and using different common associative classification algorithms (MCAR and CBA) and traditional learning approaches have been conducted with reference to classification accuracy. The results show that the MCAR and CBA algorithms outperformed SVM and algorithms.

Keywords: Phishing Websites, Data Mining, Associative Classification, Machine Learning

1. Introduction

During the last decade, most of the financial and government organizations have extended their online services to their clients. In 2011, 83% of Americans and 85% of Europeans regularly shopped online (Fortune Magazine, 2011). With the emerging use of smart phones, increasing number of people are depending on online services to shop, check their banking account, pay their bills, or even play with anonymous friends. While such activities had an important impact on the world economy, such large dependence on online financial services increases security risks for both customers and financial institutes (S. wedyan, 2013).

Phishing is a criminal technique employing both social engineering and technical subterfuge to steal consumer's personal identity data and financial account credential. Phishing is a new identity theft crime. The media reports stories almost on a daily basis about an organization that has customers targeted by a phishing attack. While financial organizations try always to improve their security techniques in order to protect their customers, phishers develop even more sophisticated attacking techniques.

Phishing websites are fake web pages that are created by malicious people to imitate web pages of real websites (Fortune Magazine, 2011). Phisher typically create web pages that are visually very similar to the real web pages in order to scam their victims. An unaware client might be easily deceived by this kind of scam. The Victims of a phishing Web page may expose their bank account, password, credit card number, or other important information to the phishing Web page owners. While phishing is a relatively new Internet crime when compared to other forms (e.g., viruses and hacking), a recognizable increase in the number and severity of phishing attacks is reported (Anti-Phishing Working Group, 2011). According to a recent study by Gartner (2011), 57 million US Internet users have identified the receipt of email linked to phishing, about 1.7 million of them are thought to have yielded to the convincing attacks and tricked them into revealing personal information. Studies by the Anti-Phishing Working Group (APWG) have concluded that Phishers are likely to succeed with as much as 5% of all message recipients (Anti-Phishing Working Group, 2011; S. wedyan, 2013).

The aim of the phishing website is to steal the victims' personal information by visiting and surfing a fake webpage that looks like a true one of a legitimate bank or company and asks the victim to enter personal information such as their username, account number, password, credit card number, …,etc. The impact is the break of information security through the compromise of confidential data and the victims may finally suffer losses of money or other kinds of assets. The attackers might also commit identity theft crimes using the victim's stolen information. Moreover, phishing attacks also damage the reputation of the attacked financial institutes since customers become less confident that they can securely access their accounts. Therefore, they might switch to other institutes (S. wedyan, 2013).

Phishing has a huge negative impact on organizations' revenues, customer relationships, marketing efforts, and overall corporate image. Phishing attacks may cost companies hundreds of thousands of dollars per attack in fraud-related losses and personnel time. Even worse, costs associated with the damage to brand image and consumer confidence can run in the millions of dollars (Brooks, 2006).

Due to the wide variety of data being captured, efficient management and quick retrieval of information is very important for decision making. Data mining is the science of extracting meaningful information from these large data sets (Witten and Frank, 2000). Data mining and knowledge discovery techniques have been applied to several areas including market analysis, industrial retail, decision support and financial analysis.

According to the Anti-Phishing Working Group (APWG) reports for the fourth quarter on 2012 (Activity, 2012), The APGW Received reports of 28,195 unique phishing sites in December. During Q4, about 30 percent of personal computers worldwide were infected malware. Indeed, financial services found to be the most-targeted industry sector in the Q4 of 2012. Moreover, Payment Services eclipsed retail/services have the second-highest industry sector for targeted attacks (Activity, 2012).

This paper main goal is to investigate the potential use of automated data mining techniques in detecting the complex problem of phishing Websites. This type of prediction is closely related to classification problem in data mining where the class attribute in this case is the degree of phishing. The classification process will be based on the different characteristics such as spelling errors, long URLs, personalization, prefix and suffix, etc, collected from the input Websites using different online tools.

The literature review is introduced in Section 2. The MCAR algorithm main steps are presented in Section 3, and the experimental results are given in Section 4. Finally the conclusions are depicted in Section 5.

2. Literature Review

Phishing website is a recent problem, nevertheless due to its huge impact on the financial and on-line retailing sectors and since preventing such attacks is an important step towards defending against website phishing attacks, there are several promising approaches to this problem and a comprehensive collection of related works. In this section, we briefly survey existing anti-phishing solutions and list of the related works.

One approach proposed by (Aburrous, et al., 2010) is the integration of classification rule mining and association rule mining algorithm, the main goal of this algorithm is to categorize the key factors in detecting phishy website. The algorithm uses more than twenty phishing features and classifies these features into six groups. Then three fuzzy set values (“Genuine”, “Doubtful” and “Legitimate”) was used as input values. The output target attribute has the following a set of possible values (“Very Legitimate”, “Legitimate”, “Suspicious”, “Phishy” “Very Phishy”). Experimentations utilizing the different classification data mining algorithms have been conducted. In particular, PART, PRISM and C4.5, and associative classification (CBA, MCAR) have been contrasted against a collection of websites. The result revealed that there is a significant relation between (URL and Domain Identity) features. There was insignificant influence of the (Page Style and content) with (Social Human Factor) features. One common problem raised by the authors in the article that is associated by associative classification algorithms is the exponential growth of rules.

(Adida et al., 2005) suggested stopping fishing at the email level. Their motivation is that most of the phishing attacks use emails in order to fake their victims. According to this approach, the phishing problem can be considered as a spam filtering problem and therefore can be handled with effective spam filters.

(Dhamija and Tygar, 2005) suggested classifying phishing sites visually. The authors suggested using randomly generated visual hash to customize the browser window or web form elements to indicate the successfully authenticated sites. The problem with this approach is that it places the burden on users to notice the visual differences between a legitimate site and a phishing site and correctly conclude that a phishing attack is underway. This assumption requires user awareness and prior knowledge. However, if Internet users have high security concerns, simple URL checking will stop the attack.

(Nawafleh and Hadi, 2012) proposed new associative classification algorithm to detecting phishiy websites. The authors conduct a comparison between proposed algorithm, SVM, PRISM, RIPPER and NB. The results revel that the associative classification algorithm outperformed all other traditional methods.

The Federal Deposit Insurance Corporation (FDIC, 2004) suggested using two-factor authentication, similar to what is widely used today in connection with ATMs. To withdraw money from an ATM, the user must present both an ATM card and a password or PIN (something the person knows). A fraudster who succeeds in stealing one or the other will not be able to act as the legitimate account owner and access the ATM. However, the two-factor authentication approach is a server-side solution and cannot, due to legal issues, enforced on all websites. Moreover, the approach cannot prevent the revealing of sensitive information that is not related to a specific site (e.g., credit card information).

3. The MCAR Algorithm

In this section, we explain MCAR algorithm (Thabtah et al., 2005) in details since the proposed Arabic text categoriser is based on it. MCAR is an AC classification algorithm that was developed in 2005 by Thabtah, et al., and is considered the first AC algorithm that uses fast intersection method for rule discovery based on the concept of vertical mining. This algorithm constitutes multiple phases where the first phase is optional and the algorithm checks whether the input training data set contain continuous attribute and if so MCAR invokes Entropy based discretisation method. Once this done, MCAR utilises TID-List intersection method for frequent ruleitems discovery. A TID-List of a ruleitem (attribute value, class) contains the locations of ruleitem in the training data set along with the locations of its associated class labels. In other word, A TID-List is simply a data structure that stores the appearances of an attribute value and the class attribute in the input data set. This data structure is very useful when it comes to computing the support and confidence of the ruleitem and thus saves resources associated with time and memory usage (Thabtah et al., 2005).

Consider for example two ruleitems as follows (A1), Class1 and (K2), Class 1 and assume that these ruleitems have the following TIDLists (1,3,4,7,11,15,22) and (2,4,11,15,16,18,21,25) respectively. Further, assume that the minimum support is 3 meaning these two items are frequent ruleitems of size 1 since each of them contain a single attribute value in its antecedent (right-hand-side). Now, to validate whether the new candidate ruleitem (A1,K2), Class1 is frequent, MCAR algorithm simply intersects the TIDLists of frequent ruleitems of size 1, e.g. (A1), Class1 and (K2), Class 1 in order to determine whether the candidate ruleitem of size 2 is frequent. So, (1,3,4,7,11,15,22) gets intersected with (2,4,11,15,16,18,21,25) and the resulting TIDList (4,11,15) is actually the locations of ruleitem (A1,K2), Class1 in the training data set. Then taken the cardinality of this set we can determine that this ruleitem has support (3) which is greater than or equal to the minimum support (3) and thus this ruleitem is frequent.

The rule discovery method described earlier is very simple if compared with other AC mining method such as CBA that necessitates multiple training data set scans and consumes more time and memory. In fact, MCAR rule discovery method requires only one single data scan and then performs simple intersection between the TIDLists of ruleitems of size N-1 to generate candidate ruleitems of size N. Once all frequent ruleitems are discovered, MCAR algorithm generates the subset of those which hold larger confidence than the minimum confidence threshold as rules. When all rules are generated then the algorithm applies a ranking procedure to favour rules over each other. The basis of this rule favouring procedure is mainly the confidence value, and then support value and lastly the size of the rules (number of attributes values in the rule body). If two or more rules having similar confidence, support and rule size then the rank will be random.

Once all rules are sorted, then MCAR uses the database coverage pruning to remove redundant rules from taking any role in the prediction step. More details on the database coverage pruning can be found in (Thabtah et al., 2005). The output of the pruning are the subset of rules that are high predictive and those represent the classifier. Once the classifier is produced its predictive power is tested using cross validation or on test data set. The prediction procedure of MCAR works as follows: Given a test data case, the algorithm goes over the set of rules starting from the highest ranked rule and selects the rule that its body matches test data case and assigns its class to the test data case. The outcome of the prediction phase of MCAR is the error rate which is simply calculated by dividing the number of correct classification in the test data set by the size of the test data set.

4. Experimental Results

A data set of 1010 phishing, Phishing and legitimate e-banking websites is used in the study (562 rows phishing e-banking websites and 448 rows of real e-banking websites for the legitimate portion of the data set). In addition, 27 features are used to train and test the classifiers. We used a series of short scripts to programmatically extract the above features, and store these in an excel sheet for quick reference. Our goal is to gather information about classifying and categorizing of all different e-banking phishing attacks techniques. By thoroughly investigating these phishing attacks we’ve created a data set containing information regarding what different techniques have been used and how it can be predicted.

Ten-fold cross-validation was utilised to evaluate the classification models and to produce error rates in the experiments. Then rules are learned from 9 folds and evaluated on the remaining hold out fold. The process is repeated 10 times and the results are averaged and produced. The experiment was executed using 10 fold-cross validation (Yin et al., 2003.

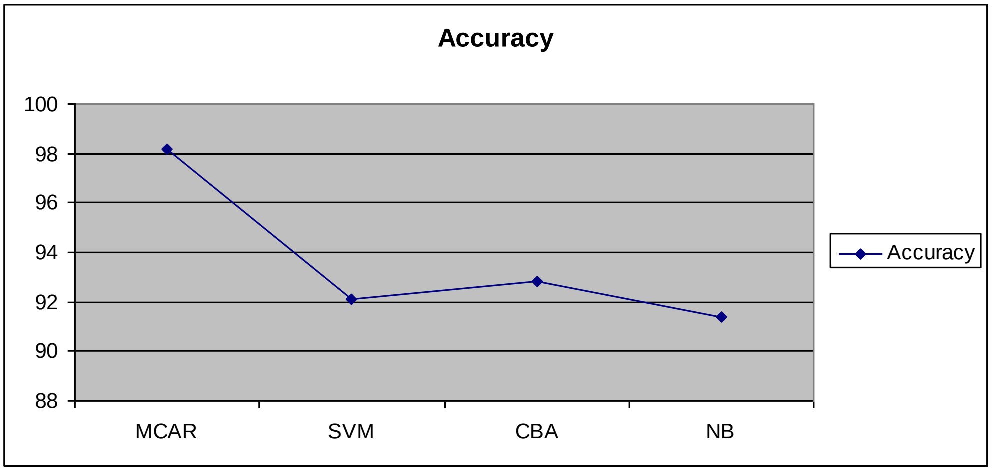

Figure 1 Accuracy of all classifiers

The following four known classification data mining approaches: statistical (SVM) (Vapnik, 1995), CBA (Liu, et al., 1998), probabilistic (NB) (Thabtah et al. 2009) and associative classification (MCAR) (Thabtah , et al., 2005).

Our selection of the above classification approaches is based on the different learning strategies they employ in discovering and producing the knowledge.

We used a Pentium IV 2.0 GH machine to run the experiments. The experiments of NB, SVM, and CBA were conducted using the Weka software system (www.weka.com), which is an open java source code for the common data mining and machine learning algorithms. Lastly, MCAR algorithm was implemented in Java.

The minsupp has been set to 2% since more extensive experiments reported in (Liu, et al., 1998, Thabtah, et al., 2010) suggested that it is one of the rates that achieve a good balance between accuracy and the size of the classifiers. The confidence threshold, on the other hand, has a smaller impact on the behaviour of any AC method and it has been set to 40% for MAC. The classification accuracy is used as the base of our experiments. The accuracy is computed by dividing the number of correctly classified cases by the total number cases in the testing data set.

After analysing Figure 1 we found out that the MCAR algorithm scales well if compared to common classification data mining algorithms. In particular, MCAR has achieved on average 6.8%, 6.1% and 5.4% higher accuracy than SVM, CBA and NB respectively. And CBA algorithm outperformed SVM and NB algorithms. Associative classifiers approch produce more accurate classification accuracy than other traditional classification approches such as statistcal , probabilistic.

Also the rules generated from our associative classifier (MCAR) indicates that URL and Domain Identity, and Security and Encryption features are consider important features to increase the final detection rate. The experiments demonstrate the feasibility of using Associative Classification techniques in real applications involving large datasets.

5. Conclusions

Phishing is a criminal technique employing both social engineering and technical subterfuge to steal consumer's personal identity data and financial account credential. Phishing is a new identity theft crime.

Experimentations against phishing data sets using different classification algorithms have been performed. The base of the experiments is accuracy measure. The results obtained reveal that the MCAR algorithm outperformed all other algorithms with respect to accuracy. In near future, we would like to extend our experiments to handle multi-label data sets and generated multiple labels classifiers.

References

Aburrous, M., Hossain, M. A., Dahal, K., & Thabtah F. (2010). Predicting Phishing Websites using Classification Mining Techniques with. Seventh International Conference on Information Technology (pp. 176-181). IEEE.

Activity, P., & Report, T. (2012). Phishing Activity Trends Report Q4.

Adida, B., Hohenberger, S., and Rivest, R., “Lightweight Encryption for Email”. USENIX Steps to Reducing Unwanted Traffic on the Internet Workshop (SRUTI), 2005.

Brooks J., “Anti-Phishing Best Practices: Keys to Aggressively and Effectively Protecting your organization from Phishing Attacks”, White Paper, Cyveillance 2006.

Dhamija R. and Tygar J.D., “The Battle Against Phishing: Dynamic Security Skins,” Proc. Symp. Usable Privacy and Security, 2005.

FDIC. “Putting an End to Account-Hijacking Identity Theft”. 2004. http://www.fdic.gov/consumers/consumer/idtheftstudy/identity_theft.pdf

Fortune Magazine, “Online Shopping Worldwide Ecommerce Statistics”, http://www.fortune3.com, October, 2011

Fu A., Liu Wenyin, Xiaotie Deng, “ Detecting Phishing Web Pages with Visual Similarity Assessment Based on Earth Mover’s Distance (EMD)”, IEEE transactions on dependable and secure computing, vol. 3, no. 4, 2006.

Guntar, O., “The Phishing Guide, Understanding and Preventing Phishing Attacks”, Technical Info, http://www.technicalinfo.net, 2011,march 21 2012

Lance J., “Phishing Exposed”, Tech Target Article sponsored by: Sunbelt software, searchexchange.com 2006.

Liu B., Hsu W. and Ma Y. (1998). Integrating classification and association rule mining.

Proceedings of the KDD, (pp. 80-86). New York, NY.

Ming Qi, Chaobo Yang, “Research and Design of Phishing Alarm System at Client Terminal”, Proceedings of the 2006 IEEE Asia-Pacific Conference on Services Computing (APSCC'06) 2006 IEEE.

S. Nawafleh, W. Hadi (2012). Multi-class associative classification to predicting phishing websites. International Journal of Academic Research Part A; 2012; 4(6), 302-306.

Seker R., "Protecting Users against Phishing Attacks with AntiPhish", Journal, Computer Software and Applications, Vol. 13, No. 8, 2006, pp. 517-524.

Shenoy P., Haritsa J., Sudarshan S., Bhalotia G., Bawa M. and Shah D. (2000). VIPER:

A vertical approach to mining association rules. Proceedings of the ACM SIGMOD International Conference on Management of Data, (pp 22-33).

S. wedyan, "Detecting Phishing Websites using Associative Classification Mining", Masters Thesis, Amman Arab University, Amman Jordan, April 2013.

Thabtah F., Hadi W., Issa A., Abdel-jaber H. (2010) Prediction Phase in Associative Classification. Journal of Knowledge Engineering and Software Engineering. WorldScinet.

Thabtah, F., Cowling, P., and Peng, Y. (2005) MCAR: Multi-class classification based on association rule approach. Proceeding of the 3rd IEEE International Conference on Computer Systems and Applications (pp. 1-7).Cairo, Egypt.

Witten, I., and Frank, E. (2000) Data mining: practical machine learning tools and techniques with Java implementations. San Francisco: Morgan Kaufmann.

Zaki M. and Gouda K. (2003). Fast vertical mining using diffsets. Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (pp. 326 – 335). Washington, D.C.