SOCIETY FOR SCIENCE AND EDUCATION UNITED KINGDOM
TMLAI TRANSACTIONS ON MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE
VOLUME 3 ISSUE 4
ISSN 2054-7390

Phishing Websites Detection Using Data Mining Classification Model

[1]Riad Jabri and [2]Boran Ibrahim
Computer Science Department, University of Jordan, Amman, Jordan
jabri@ju.eu.jo; boranabed@yahoo.com

ABSTRACT

Phishing is a significant security threat to the Internet; it is an electronic online identity theft in which the attackers use spoofing techniques like fake websites that mimic legal websites to trick users into revealing their private information. Many of successful phishing attacks do exist and subsequently a considerable number of anti-phishing methods have been proposed. However, they vary in terms of their accuracy and error rate. This paper proposes an algorithm for phishing websites detection using data mining classification model. It is implemented and experimented using a dataset composed of 20 different webpage features and 1,000 instances. The experimental results showed that the proposed algorithm outperforms the original one in terms of the number of classification rules, accuracy (87%) and less error rate (0.1 %).

Keywords: Phishing detection; PRISM; Machine learning; Classification; Data mining.

1 Introduction

Recently, online services are widely deployed. For example, online banking over the web has become essential for customers as well as for banks. For more than a decade, the naive users have been targeted by phishing attacks to acquire their sensitive information such as username, passwords, and pass code through the fraudulent websites which is under the control of the attackers. Phishing attacks are increasing rapidly. According to the Anti-Phishing Working Group (APWG) [1], the phishing problem has grown significantly over the last years that affect economics, and the financial losses from phishing attacks have risen considerably. Although the most phishing targets are financial institutions like banks, the APWG reports that an increasing number of attacks against other organizations like government agencies and automotive associations [1]. The increasing number of scams (phishing) sites demands new methods to classify the websites with a high degree of confidence. Hence, in this paper, we will use data mining techniques and exploit the websites features to establish a large labeled training data, and then yield a classifier integrating these features in such a way that new coming websites could be classified correctly. We propose an algorithm for phishing websites detection using data mining classification model. The proposed algorithm is an enhanced version of a well-known data mining algorithm, called PRISM [2]. PRISM algorithm has been selected to learn the relationships between the selected phishing features. We have chosen this algorithm since the learnt classifiers are easily understood and used by a human. In addition, the learnt can easily be expanded to include features of newly emerging phishing web sites.

DOI: 10.14738/tmlai.34.1381
Publication Date: 3th September, 2015
URL: http://dx.doi.org/10.14738/tmlai.34.1381

This paper is organized as follows: in section 2 related works of Phishing Detection are presented, section 3 describes the classification model and the proposed algorithm, section 4 shows experimental evaluation. A discussion and conclusion are given in section 5 section 6 respectively.

2 Related Works

Although phishing is not new and well-known by Internet users, many people are still tricked into providing their confidential information. To counter the phishing threat, a number of anti-phishing solutions have been proposed. Representative ones are as follows:

- A class of anti-phishing approaches aims to solve the phishing problem at the email level; the main idea is that when phishing emails reach its victims, they could fall for the scam. This line of research is closely related to anti-spam research. One reason for the abundance of spam and phishing emails is that the Simple Mail Transport Protocol (SMTP) does not contain any authentication mechanisms. The sender information in an email message can easily be spoofed [3]. In this context, in [4] a novel approach is proposed to detect phishing attacks by implementing a prototype web browser to processes the arriving email for phishing attacks, they integrate this method with the approach of using link based features. Such as a hyperlink to get personal details about users, such as usernames, passwords, account numbers, etc. Thus, this approach considers features, which have numerical values but with different ranges, these features will be extracted from the email body and according to a certain values, the user gets notified of phishing attacks to avoid opening the suspicious websites.
- To capture the patterns in phishing URLs, Afroz and Greenstadt [5] proposed a real time approach for web phishing detection, called PhishZoo. This approach uses profiles of trusted website's appearances and content, to detect targeted phishing attacks. They make use of computer vision techniques such as matching images, scene analysis (segmenting images into objects) and the URLs and (HTML) contents of a website to identify imitations.
- There exists a number of anti-phishing toolbars based on different techniques, many of which exploit blacklists to achieve close-to-zero false positive rate [6]. The most popular and widely-deployed anti-phishing techniques are based on the use of blacklists. These blacklists store a set of phishing domains that the browser blocks their visit [3]. On the other hand, other approaches exploit whitelists [5]. These approaches seek to detect known good sites. Some whitelisting approaches use server side validation to add additional authentication metrics to client browsers as a proof of its benign nature, for examples, dynamic security skins [7]; trust bar [8] and SRD ("Synchronized Random Dynamic Boundaries") [9].
- In an attempt to exploit visual similarity, Wenyin [10] proposed an approach to detect the phishing web pages based on visual similarity, if the visual similarity of these pages is higher than a predetermined threshold, this website is recorded as a suspicious and the owner is alerted.
- A content-based approach, called CANTINA, is proposed in [11] to detect phishing websites. CANTINA examines the content of a web page to determine whether it is legitimate based on the term frequency/inverse document frequency.

- In [12] authors proposed a server-side detection model; an automatic, and proactive identity theft detection model in online games utilizing rich features. The proposed model classifies a hacker's login from a genuine user's login using well-balanced features from connection information, user behavior, and economic variables.
- An interesting solution has been proposed in [13] to detect e-banking phishing websites using an artificial intelligent technique, authors propose a model based on using association and classification data mining algorithms and tools, which were used to classify the fishing websites and the relationship that correlate them with each other, six classification algorithm were implemented to extract the phishing training datasets criteria to classify their legitimacy.

Although there are many solutions available and act as a model for automatic phishing detection, the promising ones are based on machine learning and data mining techniques. Recently, an associative classification expert system was proposed in [14]. It has been shown that the rule induction and decision tree methods outperformed other ones [12, 14]. Hence, this paper constitutes an improvement of such techniques.

3 Classification Model

The Using machine learning techniques, a data mining classification model based on modified version of PRISM algorithm is proposed to categorize websites as phishing versus non-phishing. Such categorization is achieved by applying induction rules on selected websites features. The construction of such a model proceeds as follows:

- A website is represented as a set of features and their respective values. For example, {<subjreply {0, 1}>,…, <urlnoLinks{0,1,..}}
- A dataset is then constructed as a set of tuples of the values of the selected features of phishing and non-phishing websites. Hence, the dataset describes two classes (phishing versus non-phishing) and subsequently each tuple is assumed to belong to a predefined class, as determined by the class label. For example, the rows of Table 1 represent such tuples, where the first row represents site 1 that is classified as a phishing one. On the other hand, Site n is classified as non-phishing. The complete dataset is given in Table 2.

Table 1: Dataset example

	subjreply …….	attachedfile	class
site			
1	0	0	1
n	1	1	0

- Using distribution probabilities, the dataset is then divided into training and testing one. A classification model is then constructed as a set of induction rules. For example, {If invisiblelinks=1 and If subjforward=1 Then yes,…, If urlnoLinks=6 Then yes}. Such a set is obtained by applying the proposed algorithm on the training dataset. It is then tested and evaluated using the test set to achieve a high accuracy rate as well as less error rate, where the accuracy rate is percentage of test set samples that are correctly classified by the model, while the error rate is the opposite.

3.1 The Proposed Algorithm

PRISM is attribute-value-oriented; it measures the attribute-value pair in determination of the classification and selects the attribute that contributes the maximum information gain. If this attribute is perfect (the attribute accuracy is equal 1), the corresponding rule will be generated and added to the rule set to represent the classification model for the prediction purposes. PRISM generates one rule at a time, in a particular iteration if more than one attribute is perfect this algorithm chooses the attribute with the highest coverage to generate a rule, otherwise a random attribute will be chosen to generate a specific rule. In contrast to PRISM algorithm, the proposed one derives all the perfect rules at the same time instead of learning one rule at a time. As a result, the number of derived rules and their coverage rate were increased. Hence, the accuracy and the expressiveness of the proposed model and subsequently the prediction process were improved. The pseudo code of the proposed algorithm is given below as Algorithm 1.

Algorrithm1

```
Input: labeled training dataset D (%)
Output:  rule set R that covers all instances in D
Method:
Initialize R as the empty set
for each class C {
        while D is nonempty {
            Construct all perfect rules r that correctly classifies some
          instances in D that belong to class C and do not incorrectly
          classify any non-C instances;
            Add rules r to rule set R
            Remove from D all instances which are correctly classified by
          all r
        }
return R
```

4 Experimental Evaluation

In this section, we present the experiments designed to build and evaluate the proposed model. This includes a description of the constructed dataset, metrics to evaluate the effectiveness of our approach. The experiments were conducted over 1000 different websites (70% are phishing ones and 30% are non-phishing) were used in the conducted experiments.

4.1 Dataset

For the purpose of this research, a large number of phishing pages with their respective features were explored. As a result, 20 feature- value pairs were selected based on the frequency analysis as suggested in [15] and as shown by the following set

{The set of features and it their possible values:

Attribute subjreply {0, 1},

Attribute subjforward {0, 1},

Attribute subjnoWords {0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16},

Attribute subjverify {0, 1},

Attribute subjbank {0, 1},

Attribute bodyhtml {0, 1},

Attribute bodyFunctionWords {0, 1},

Attribute bodysuspension {0, 1},

Attribute bodyconfirm/verifyYourAccount {0, 1},

Attribute urlnoIpAddresses {0, 1},

Attribute urlatSymbol {0, 1},

Attributeur lnoLinks {0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,..},

Attribute urlnoDomains {0,1,2,3,4,5,6,7,9,11,12,14,15,17,29,73},

Attribute urlmaxNoPeriods {0,1,2,3,4,5,6,7,8,10,11,12,13,16},

Attribute urllinkText {0, 1, 2},

Attribute OnClickEvents {0, 1, 10},

Attribute Invisiblelinks {0, 1},

Attribute Unmatchingurl {0, 1},

Attribute Longurladdresses {0, 1},

Attribute attachedfile {0, 1}**}**

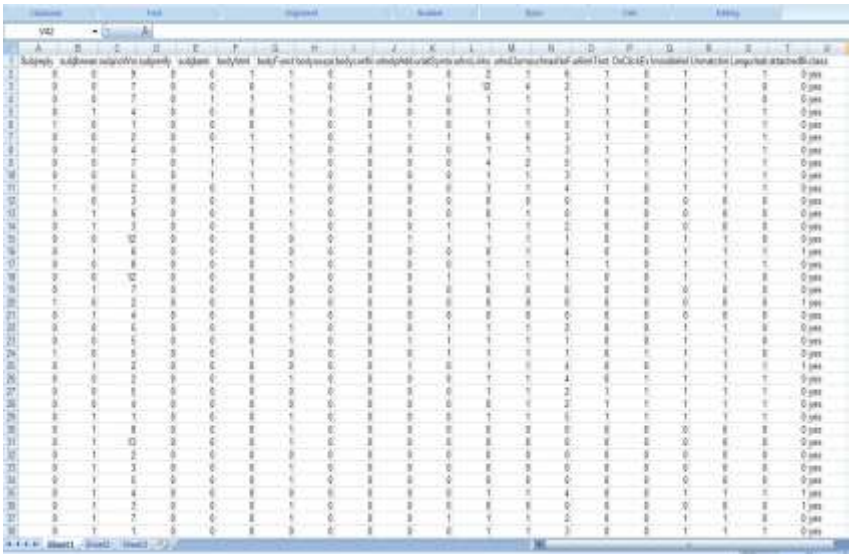
The dataset is then constructed by extracting feature – values pairs of phishing cases from one source, and non-phishing web pages from another source. We chose PhishTank [15] as a source of phishing websites. The information from this site is freely available and the amount of reported phishing sites is very large (approximately five hundred new phishing reports every day). The attributes of the non-phishing web sites were collected by a web data extractor. Such extractor is a user defined software that is plugged with a browser to automatically extract features from a user connection [16]. Thus, the constructed dataset contains a total of 700 phishing websites and 300 legitimate ones, as shown in table 2.

Preparatory experiments were conducted on this dataset using the well-known data mining software WEKA to show the classification effectiveness of the machine learning algorithms and to justify the proposed algorithm. The performance of three algorithms (JRip, J48, Naïve Bayes (NB)) has been compared to the one of PRISM, as shown Table 3. Such comparison has been performed using 70% of the dataset for training and the remainder for testing. In addition, the following evaluation metrics have been applied:

- Accuracy and error rate
- Number of derived rules

It is clear that the J48 algorithm is the best one, because it has the highest accuracy in the websites dataset, while NB algorithm is a weak algorithm since it has the lowest accuracy with the higher error rate, and it doesn't produce rules at all. PRISM algorithm has a minimum accuracy with some unclassified instances in both datasets.

Table 2: The constructed dataset



Hence, we set an objective to improve this algorithm PRISM algorithm in terms of discovered knowledge, accuracy, error rate and other criteria which appear useful for phishing detection. PRISM is "weak" algorithm; the aim is to change it into an accurate "strong" typical algorithm for our detecting purposes. We selected the PRISM classifier for the following reasons. First, we believe that a PRISM provides intuitive insight into which features are important in classifying a dataset. Second, the PRISM algorithm is known to work well for a wide range of classification problems.

Table 3: Results of WEKAs' algorithms

	No. of Instances	WEKA Algorithms	No. of Rules	Accuracy Rate %	Error Rate %	Unclassified Rate %
Websites Data	1000	JRip	7	91	9	-
		J48	26	94	5	-
		NB	-	88	11	-
		PRISM	189	86	5	8

4.2 Experimental results

Experiments were conducted to evaluate the performance of PRISM and the proposed algorithm. Figure 1 and Figure 2 show samples of their run using the collected dataset.

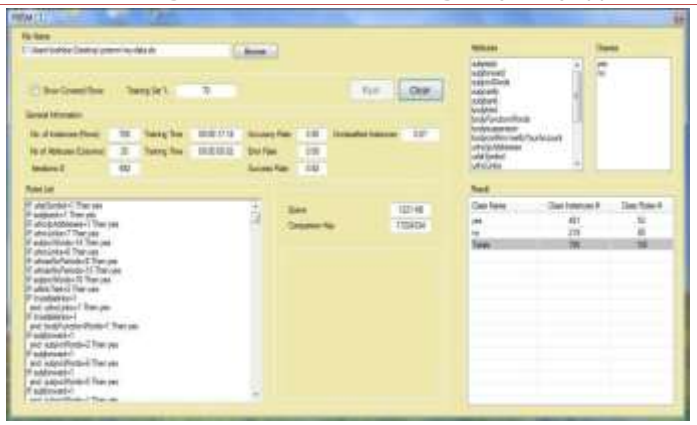


Figure 1: A sample run of PRISM

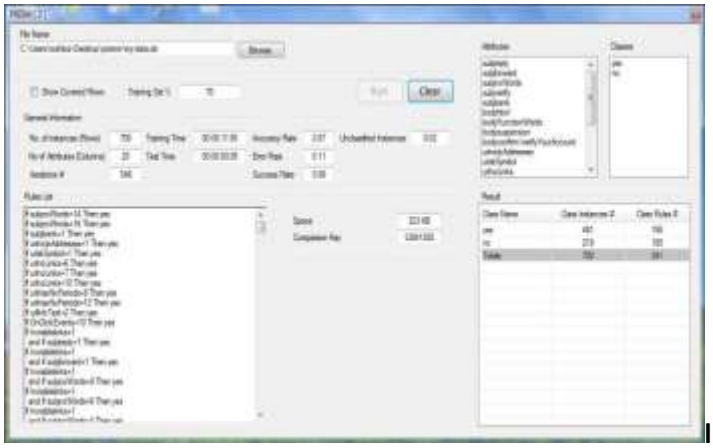


Figure 2: A sample run of the proposed algorithm

Several runs of PRISM and the proposed algorithm have performed using different ratios between training, and testing datasets show the results of these runs as follows

- Table 4 shows the achieved accuracy and error rate by both algorithms
- Table 5 shows the phishing and legitimate rules generated by both algorithms.
- Table 6 shows the number of iterations performed during the run of both algorithms

Table 4: Accuracy and error rates for the considered models

	PRISM				Modified PRISM			
Training Data % %	40	60	70	80	40	60	70	80
No. of Phishing Instances	267	407	481	555	267	407	481	555
No. of Legitimate Instances	133	193	219	245	133	193	219	245
Accuracy Rate %	84	84	85	88	84	85	87	86
Error Rate	12	9	8	6	15	13	11	12

Table 5: Total number of rules genrated by the considered models

	PRISM				Modified PRISM			
Training data %	40	60	70	80	40	60	70	80
No of phishing instances	267	407	481	555	267	407	481	555
No of legitimate instances	133	193	219	245	133	193	219	245
No. of Rules	67	117	138	150	121	270	341	373
no. of phishing rules	25	49	53	51	45	123	156	161
no. of legitimate rules	42	68	85	99	76	147	185	212

The obtained results demonstrate that the constructed classifiers have 341 rules, with accuracy of (87%). Thus, the proposed algorithm achieves higher accuracy and generates more rules. Furthermore, it is more efficient, as demonstrated by the number of iterations performed during its run.

Table 6: Total number of iterations performed by the considered models

	PRISM				Modified PRISM			
Training data Percentage %	40	60	70	80	40	60	70	80
No of instances	400	600	700	800	400	600	700	800
No of phishing instances	133	407	481	555	267	407	481	555
No of legitimate instances	267	193	219	245	133	193	219	245
No of Iterations	513	652	692	713	444	515	546	563

5 Discussion

Considering the results achieved by the proposed algorithm, we can conclude that this approach is quite successful in protecting users, especially when predicting the phishing websites. The proposed algorithm has correctly recognized almost 87% of the malicious websites. It appears that this algorithm outperforms the original PRISM and has a comparable performance to the one of several algorithms (CBA, PART, C4.5, MCAC, JRip, and MCAR) as published in recent research [14], where the accuracy of these algorithms ranges from 92% to 94.5% , while and the number of rules ranges from 10 to 180. Hence, the results the proposed algorithm is encouraging. Although the results for the larger datasets are rich, it takes more time to compute. It is worth mentioning that the proposed phishing detection system represents a good indicator for the phishing websites since this model produced a large number of significant rules. In addition, it is flexible and capable of capturing new emerging features of the phishing web sites.

6 Conclusion

In this paper, we have proposed a new phishing websites detection model based on PRISM algorithm, the prediction of phishing websites is essential, and this can be done using data mining classification algorithms, our classification system automatically classify phishing pages. We have tested the phishing detection model in terms of accuracy and error rate. The experimental results confirm the increasing number of rules of the proposed method for phishing detection and show that we maintain a false positive and false negative rate about (0.1%) which is reasonably low. The experiments have demonstrated the feasibility, and effectiveness of using rule induction classification techniques in real applications involving large databases. It has better performance as compared to other traditional classification algorithms with accuracy rate of (87%). Due to the variations in the existing approaches, a generalized and formalized one constitutes a future work.

REFERENCES

[1]. APWG. *Phishing Activity Trends*: Technical report, Anti Phishing Working Group, [online], http://www.antiphishing.org/reports/apwg_trends, 2013.

[2]. Cendrowska, Jadzia, *PRISM: An algorithm for inducing modular rules*. International Journal of Man-Machine Studies, 1987. 27(4): P. 349-370.

[3]. Medvet, E., Kirda, E., and Kruegel, C., *Visual-Similarity-Based Phishing Detection*. In Proceedings of the 4th international conference on Security and privacy in communication networks, ACM 22, Istanbul, Turkey, 22 – 25, September, 2008.

[4]. Jain, A., and Richariya, V., *Implementing a Web Browser with Phishing Detection Techniques*. World of Computer Science and Information Technology Journal (WCSIT), 2011. 1 (7):p. 289-291.

[5]. Afroz, S. and Greenstadt, R., *PhishZoo: Detecting Phishing Websites By Looking at Them*. In Proceedings of the Semantic Computing (ICSC), Fifth IEEE International Conference , 18-21 September, 2011, 368 – 375.

[6]. Xiang, G. and Hong, J.I., *A Hybrid Phish Detection Approach by Identity Discovery and Keywords Retrieval,* In Proceedings of the 18th International World Wide Web Conference (IW3C2), ACM, Madrid, Spain, 2009, p. 571- 580.

[7]. Kumaraguru, P. Rhee, Y. Acquisti, A. Cranor, L. F. Hong, J. and Nunge, E., *Protecting people from phishing: The design and evaluation of an embedded training email system*. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07), ACM, New York, NY, USA, 2007, p. 905-914.

[8]. Herzberg A. and Gbara, *A Security and Identification Indicators for Browsers against Spoofing and Phishing Attacks*, Journal of ACM Transactions on Internet Technology (TOIT), 8(4): p. 1-36.

[9]. Ye Z. and Smith, S. *Trusted paths for browsers*, In Proceedings of the 11th Usenix Security Symposium, ACM NY, USA, 2005, p. 263-279.

[10]. Wenyin, L. Huang, G. Xiaoyue, L. Min, Z. and Deng, X., *Detection of Phishing Webpages based on Visual Similarity*, In Proceedings WWW '05 Special interest tracks and posters of the 14th international conference on World Wide Web, ACM, May, 2005, p. 1060-1061.

[11]. Zhang, Y. Hong, J. and Cranor, L., *CANTINA: A Content-Based Approach to Detecting Phishing Web Sites*. In Proceedings of the 16th international conference on World Wide Web, , **ACM**, Banff, Alberta, Canada, 8-12, 2007, p. 639-648.

[12]. Woo, J. Choi, H.J. and Kim, H.K., *An automatic and proactive identity theft detection model in MMORPGs*, .Applied Mathematics and Information Sciences, 2012, 6(1S) : p. 291S-302S.

[13]. Aburrous, M. Hossain, M. A. Thabtah, F. and Dahal, K., *Intelligent Detection System for e-banking Phishing websites using Fuzzy Data Mining*, International Conference on CyberWorlds, IEEE Conference Publications **,** 2009**,** 37(12)**:** p. 265-272.

[14]. Abdelhamid,N., Ayash, A., Tabatah, F., *Phishing Detection Based Associative Classification Data Mining*, Expert System with Application, 2014, 41: p., 5948-5959.

[15]. PhishTank, *Out of the Net, into the Tank*, http://www.phishtank.com/developer_info.php, 2012

[16]. Aburrous, M. Hossain, M. A. Dahal, K. Thabtah, F., *Predicting Phishing Websites using Classification Mining Techniques with Experimental Case Studies*, In Proceedings of the 2010 Seventh International Conference on Information Technology: New Generations (ITNG '10), **IEEE**, Las Vegas, Nevada, USA, April 2010, p. 176-181