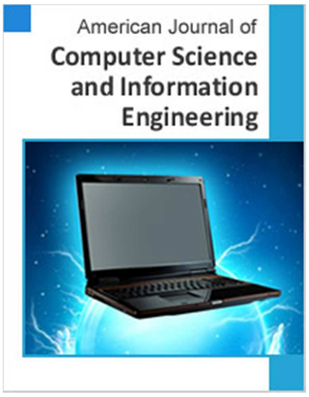
American Journal of Computer Science and Information Engineering
2017; 4(4): 31-35
http://www.aascit.org/journal/ajcsie
ISSN: 2381-1110 (Print); ISSN: 2381-1129 (Online)

American Association for Science and Technology



Website Phishing Detection Using Dom-Tree Structure and Cant-MinerPB Algorithm

Asma Al Sarhan, Riad Jabri, Ahmad Sharieh

Computers Sciences Department, KASIT, University of Jordan, Amman, Jordan

Email address
jabri@ju.edu.jo (R. Jabri)

Keywords
DOM-Tree Structure,
cAnt-MinerPB Algorithm,
Phishing Detection

Received: Spril 15, 2017
Accepted: September 6, 2017
Published: September 26, 2017

Citation
Asma Al Sarhan, Riad Jabri, Ahmad Sharieh. Website Phishing Detection Using Dom-Tree Structure and Cant-MinerPB Algorithm. *American Journal of Computer Science and Information Engineering*. Vol. 4, No. 4, 2017, pp. 31-35.

Abstract
Phishing is stealing users' confidential information by uploading a fake website that claims to be of another. Such a web site contains special features that aid an automatic classification of it as phishing one. However, there is no single feature that to be used to identify the phishing web-sites. Subsequently, the properties of phishing website are defined as a collection of features and then used to actively discover such sites in real-time. This research develops automatic classification of a web-site into a phishing or non-phishing one based on aggregation of a set of predetermined features related to the content of the site. A classifier is developed based on Ant-Colony optimization, known as cAnt-MinerPB. The features are combined and listed in a tree structure using document object model (DOM) representation to find the best level of detailed features that helps in capturing the phishing properties. Moreover, such hierarchical representation is used to capture strength and weakness of the classifier itself and other classifiers that are used for a comparative study. The proposed method has a fair accuracy as compared to well-known algorithms. For the tested data set, its obtained accuracy was close to that obtained by KNN and SVM. The proposed classier, for the conducted experiments, is better than other classifiers that are rule based. The cAnt-MinerPB has shown promising results compared to the well-known and well-established classification techniques.

1. Introduction

With the ever increasing in the applications over the World Wide Web (www), such as E-Payment applications, E-Banking and E-Businesses, hackers find more opportunities to violate users privacy and disclose their confidential information [1]. These applications are protected in a way that information being communicated with users are highly protected by disallowing unauthorized accessing/capturing of such data using network security technique (e.g. firewall). Even with such techniques for users' protection, there are still threats surrounding users as threat might not be embodied in stealing data in non-secure web-site but it can also be fraud or a phishing one. Thus, phishing is stealing users' confidential information by uploading a fake website that claims to be of another [1]. This is implemented by directing the users to fraud website by many ways such as hacking the Domain Name Server (DNS), from which the user is being served. Similarly, phishing can be implemented by sending the user an attractive email asking to update or validate his/her information (user-name, password, credit card number, etc.) claiming to be a legitimate party that users is attached to it. There are many other ways by which the user will end up in a phishing web-site that he/she thinks that is a legitimate one [2]. On the other hand, many phishing detection techniques have been proposed, for examples:

a. [3] used Support Vector Machine for classification of a dataset contains 100 legitimate web-site and 279 phishing web-sites, which achieve an 84% classification accuracy rate.
b. [4] proposed a Fuzzy Logic model that characterized the attributes of the web-sites in a fuzzy set of linguistic terms. The input features characterize the domain and URL of the web-sites.
c. [5] used distinct structural features with Support Vector Machine (SVM), which obtained a 95% accuracy rate on a small sample size of 400.
d. [6] used neural URL based features and achieved accuracy rate of 95%. But, both neural network and SVM are classifiers not easily interpretable (Otero, Freitas et al. 2013).
e. [7] presented enhance detecting phishing websites based on machine learning techniques of fuzzy logic with associative rules.

Hence, the accuracy phishing detection is largely dependent on feature selection. On the other hand, Ant-Miner based on pits burgh (Ant-MinerPB) approach has been recently used successfully in data classification with an accuracy rate 97.60% in annealing data set and 94.29% in breast-w data set. The results are very positive and in terms of predictive accuracy it achieved the best average rank [8]. To overcome the individuality of rules creation, cAnt-MinerPB algorithm has recently been proposed in [8] as an improved version of the original Ant-Miner. Therefore, this paper proposes Phishing Detection Model (PDM) that aims at accurately classifying websites as phishing or non-phishing based on applying cAnt-MinerPB on a proposed website object model (aggregation of features extracted from the inspected websites). This captures several levels of abstraction and achieves further conformity and scalability of rules creation. In addition, achieves less error rate and avoids rule iteration problem. This paper is extracted from theses [9]. It is organized in four sections. Section One introduces the research problem, research motivation, objectives and the proposed solution. Section Two proposes the PDM and its properties. Section three presents the results obtained by the implementation of the proposed model. Finally, Section Four gives a brief conclusion and summarizes the research findings.

2. Phishing Detection Model

The construction of PDM as given in Figure 1 proceeds as follows:

a. A website is represented as a set of features with their respective values. For example, {<subjreply {0, 1}>, …, <urlnoLinks{0,1,...}>}
b. 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.
c. A website object model is then constructed as described in section 2.1, where website features are categorized into levels and then are combined.
d. The website object model is divided into training and testing sets. The training data set is injected into cAnt-MinerPB algorithm, as described in Section 2.2. The output is set of rules that are then used in the testing phase to classify unknown web-sites into phishing or non-phishing one.

2.1. Website Object Model

Using document object model (DOM) [10], the website features are categorized into four levels. In the process, the values of aggregated feature (feature other than those in the bottom level) are combined without losing information by giving a unique value for each different input value of the underlying features. For example, given feature *x* with possible value 0 and 1 and feature *y* with similar possible values, the compound feature *z* will take different value for the combination of 00, 01, 10, 11. This leads to generating different feature levels for each input instance in the training set. Each of these levels will be tested to find the efficiency of the utilized algorithms with the same level of information but in different representation

Features are not combined randomly, but based on the similarity and association between the underlying features. A random combination will results in over-fitting, which is a known problem in data mining that refers to input data that describes random relationships and hypothesis that cannot be captured by any data mining techniques [10]. The way by which these features are combined is given in Table 1, for examples:

a. The features, from level four, "IP Address in the URL" up to "Having Prefix and Suffix" are aggregated into 4 compound features at level 3 and into a single one at levels 2 and 1, respectively.
b. The features "Using Shortening Service" up to "Having no Google" are aggregated into 5 compound features at level 3 and into a single one at levels 2 and 1, respectively.
c. The features "Having Low WEB Traffic" up to "Having no Statistical report" are aggregated into a single one at levels 3, 2 and 1, respectively.

The purpose of this combination is to reduce the size required to represent input instances and respectively the individuality, the redundancy and the size of the generated classification rules.

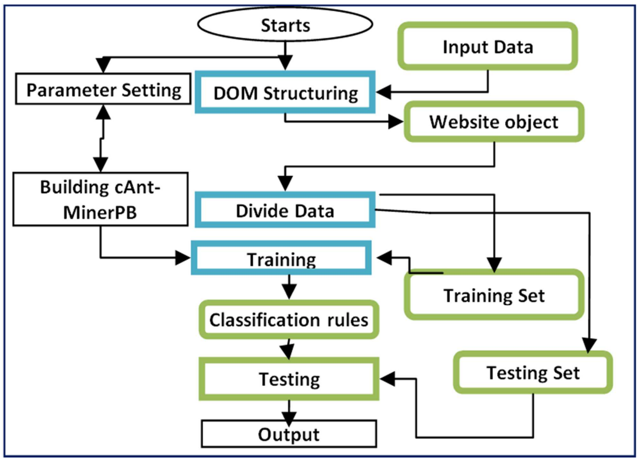


***Figure 1.** Phishing Detection Model.*

2.2. Training and Testing Phases

The training phase proceeds by applying Algorithm 1 (cAnt-MinerPB) on the training data set. As a result, classification rules are generated. The testing phase applies these rules on the testing data set to classify the considered websites as phishing and non-phishing ones.

2.3. Illustration Example

The following is an example to illustrate the processing stages performed by the proposed Phishing Detection Model.

a. The input data set is considered as given by Table 2, where 11000 websites are represented by their attribute (At$_1$.. At$_{27}$) values and respective classes, for example, At1= IP Address in the URL, At2= Long URL,…, At27=Having URL of Anchor.

b. The web site object model is constructed and the training data set is extracted as shown by Table 3, where the websites are represented by values of their aggregated attributes (At$_{1'}$.. At$_{7'}$) and respective classes. At$_{1'}$.. At$_{7'}$ are compound features with values equal to the total binary value of their respective aggregated attributes, for example, the feature At1' { IP Address in the URL and Long URL and Having URL with @ Symbol } has the value 5.

c. Training proceeds using c-AntMinerPB, where multiple iteration with multiple ants are executed producing the global list of classification rules, as shown by Table 4.

Testing proceeds using the web site object model as shown by Table 5. As a result, the web sites are classified as phishing and none phishing ones, as shown by Table 6.

```
Algorithm 1: cAnt-MinerPB
Input: Training dataset
Output: Optimized List of  Classification Rules
Initialize Trail
Global_List := {}
While (Stopping Criteria == False)
    Rules_Best := {}
    Create a Set of Ants (Colony Size)
    Rules_List := {}
    For( Each Ant in the Set)
        Investigated-List := Training_Examples
        covered_Examples:= {}
        While (covered_Examples< #TobeCovered)
            ComputeHeuristicInformation(examples)
            rule:=CreateRule(examples)
            Prune(rule)
            Investigated-List:=←Investigated-List − Covered(rule, examples)
            Rules_List := Rules_List + rule
        End While
        IF (Quality(Rules_List) > Quality(Best_List))
            Best_List:=Rules_List
        End IF
    End For
    Update Trial (Best_List)
    IF (Quality(Best_List) > Quality(Global_List) )
        Global_List:= Best_List
    End IF
    iteration++
End While
return Global_List
```

***Table 1.** Levels and Aggregations of the Utilized Feature.*

Feature / Level Four	Category	Level Three	Level Two	Level One
IP Address in the URL	URL/ Lex	Com-Feature 3.1	Com-Feature 2.1	Com-Feature 1.1
Long URL	URL/Lex			
Having URL with @ Symbol	URL/Lex			
Having Double Slash	URL/Lex	Com-Feature 3.1		
Having Fav-icon	URL/Lex	Com-Feature 3.3		
Having HTTPS token	URL/Lex	Com-Feature 3.4		
Having Prefix and Suffix	URL/Lex	Com-Feature 3.2		
Using Shortening Service	URL/Host	Com-Feature 3.5	Com-Feature 2.2	Com-Feature 1.2
Having Subdomain Identity	URL/Host	Com-Feature 3.6		
Having SSL final State	URL/Host	Com-Feature 3.7		
Having short domain registration length	URL/Host	Com-Feature 3.8		
Having Non-Standard Port	URL/Host	Com-Feature 3.9		
Having Abnormal URL	URL/Host	Com-Feature 3.10		
Having Fresh Domain Age	URL/Host	Com-Feature 3.8		
Having no DNS Record	URL/Host	Com-Feature 3.5		
Low WEB Traffic	URL/Host	Com-Feature 3.11		
Having Low Page Rank	URL/Host			
Having no Google Index	URL/Host			
Having no Links pointing to page	URL/Host			
Having no Statistical report	URL/Host			

Table 2. *Input Data set.*

Website#	At_1	At_2	At_3	....	At_{25}	At_{26}	At_{27}	Class
1	1	0	1	....	-1	0	1	0
2	1	0	1	....	-1	0	-1	
3	1	0	0	....	0	0	1	
4	0	0	0	....	1	1	1	
11000	1	1	1	....	1	1	1	

Table 3. *Training Data Set.*

Website#	$At_{1'}$	....	$At_{7'}$	Class
1	5	....	33	0
2	5	....	35	0
3	4	....	1	1
4	0	....	20	0
11000	7	....	20	1

Table 4. *Global List of classification rules.*

Rule $_1$	IF ($At_{1'}$ ==4) &&.........&& IF ($At_{7'}$ ==1) THEN Class = 1
Rule $_2$	IF ($At_{1'}$ ==6) &&.........&& IF ($At_{7'}$ ==1) THEN Class = 1
Rule $_{36}$	IF ($At_{1'}$ ==0) &&.........&& IF ($At_{7'}$ ==20) THEN Class = 0

Table 5. *Testing Data.*

Website#	$At_{1'}$	....	$At_{7'}$
11001	1	....	33
11002	7	....	35
11003	4	....	1
11004	2	....	13
....			
11400	1	....	12

Table 6. *Classified Web Sites.*

Website#	$At_{1'}$	....	$At_{7'}$	Class
11001	1	....	33	0
11002	7	....	35	1
11003	4	....	1	1
11004	2	....	13	
11400	1	....	12	0

3. Experimental Results

3.1. Dataset

The Machine Learning Repository (UCI) dataset [11] was used for the experiments. The dataset contains a total of 11055 instances with extracted and normalized features. The dataset is not divided into training and testing set. Thus, in the experiments, part of the dataset was used for training and another part was used for testing. Table 1 shows a sub set of the data set that has been utilized in the proposed model.

3.2. Configuration

Initially, the data was split into 66% in the training set and 34% in the testing set. The generated training and testing sets were only used for parameter initialization. Pre-experiments were conducted to select the best values for the parameters. The list of experimented and finally selected values of each parameter is given in Table 7.

Table 7. *Parameters Settings.*

Parameter	Tested Values	Selected Value
Colony Size	5, 10, 15, 20	10
Abstraction Level	4 levels	Level three and Level four
Threshold	10, 50, 100, 500	50
Iteration	2000, 20000, 200000	20000

The results of using the different levels of aggregated features, as in Table 1 on a subset of 1000 instances, are illustrated in Figure 2. It is shown that, more feature details is better for phishing classification task. However, level three has given slightly better results compared to level two and four.

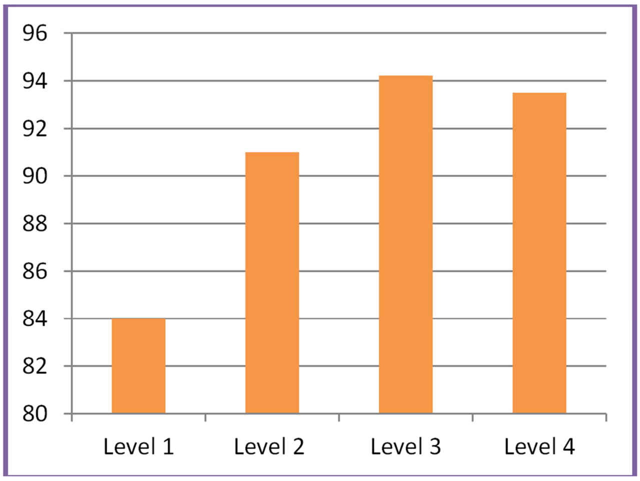


Figure 2. *Accuracy of the Proposed Model with Different Feature Levels.*

For further experiments, the data was divided into 10 folds equally. The experiments were conducted with 6 folds, 7 folds, 8 folds and 9 folds for training and the rest for testing. By other means, four generated models were obtained as follow: the data was divided into 60% training & 40% training in the first set of experiments, Then, data was divided into 70% training & 30% training in the second set of experiments, Then, data was divided into 80% training & 20% training in the third set of experiments, Then, data is divided into 90% training & 10% training in the fourth set of experiments. For each group (60%, 70%, 80% and 90%), 10 runs were conducted, each with different folds for the training and testing. The final result of each group is the average value of the 10 runs. Figure 3 shows the obtained results for level 3 and level 4, where level 3 is the result of applying slight aggregation between related features, while level 4 represents the data as it is.

To be able to evaluate the obtained results, the same set of experiments was conducted using main classification techniques, such as K-Nearest Neighborhood (KNN), Bayesian classifier, Decision Tree, Neural Network (NN) and SVM, as shown on Figure 4. The best results was given in [11] with 94.07% accuracy rate and with 95.25% accuracy rate. The proposed model, under different level of aggregation, achieved almost a similar result as the best ones reported in [11].

The results have shown that the aggregation levels are worth to be considered as many other pre-processing stages on different data mining applications and tasks.

4. Conclusions

The conducted research has focused on website phishing detection by investigating and subsequently aggregating features related to the URL and content of the web sites. As a result, a web site object model with different levels of details for these features is proposed. cAnt-MinerPB algorithm is then used to classify input web-sites into phishing or non-phishing based ones. The proposed method has a fair accuracy as compared to well-known algorithms, where Neural Network classifier outperforms the others. The proposed method was second in order, close to the accuracy obtained by KNN and SVM. Surprisingly, it is better than classifiers that are rule based. cAnt-MinerPB has shown promising results compared to the well-known and well-established classification techniques. Furthermore, the results have shown that studied the aggregation levels is worth to be considered as many other pre-processing stages on different data mining applications and tasks. As a future work, the effect of abstraction and aggregation on the accuracy of different phishing detection models will be further investigated.

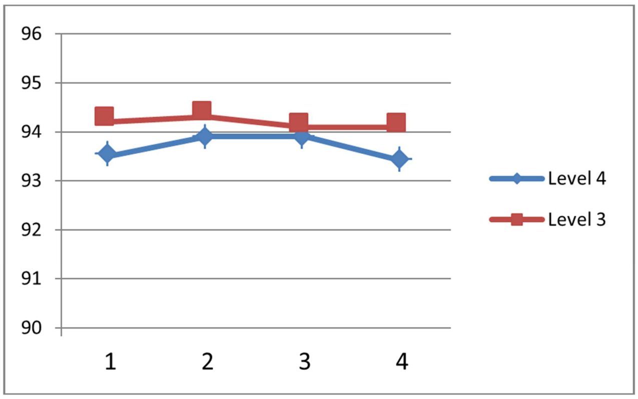


Figure 3. *Accuracy of Proposed Model (Level 4 and Level 3).*

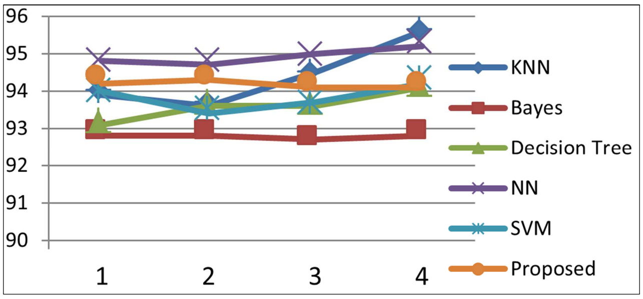


Figure 4. *Accuracy of Proposed Model and other Classification.*

References

[1] Khonji, M., Y. Iraqi, et al. (2013). Phishing Detection: a Literature Survey. Communications Surveys & Tutorials, IEEE 15 (4), 2091-2121.

[2] Anti-Phishing Working G. (2004). Phishing Activity Trends Report, Anti-Phishing Working Group.

[3] Pan, Y. and X. Ding (2006). Anomaly Based Web Phishing Page Detection. In Computer Security Applications Conference, 2006, ACSAC'06. 22nd Annual, IEEE.

[4] Aburrous, M., M. A. Hossain, et al. (2010). Intelligent Phishing Detection System for E-banking Using Fuzzy Data Mining, Expert Systems with Applications, 37 (12), 7913-7921.

[5] Chandrasekaran, M., K. Narayanan, et al. (2006). Phishing Email Detection Based on Structural Properties. In NYS Cyber Security Conference, U.S.A, Buffalo, NY-14260.

[6] Zhang, N. and Y. Yuan (2013). Phishing Detection Using Neural Network. CS229 lecture notes. http://cs229.stanford.edu/proj2012/ZhangYuan

[7] Riaty, S, Sharieh, A., Jabri, R., Al Bdour, H.(2017). Enhance Detecting Phishing Websites Based on Machine Learning Techniques of Fuzzy Logic with Associative Rules, Kasmera Journal, Vol. 45 (1), pp 63-75.

[8] Otero, F. E. B., A. Freitas, et al. (2013). A New Sequential Covering Strategy for Inducing Classification Rules with An Colony Algorithms. Evolutionary Computation, IEEE.

[9] Al Sarhan, A., Jabri, R. (2016). Security Phishing Detection using DOM-Tree Structure and cAnt-MinerPB Algorithm, University Of Jordan, Thesis, Master in Computer Science.

[10] Mohammad, R. M., F. Thabtah, et al. (2012). An Assessment of Features Related to Phishing Websites Using an Automated Technique. In International Conference for Internet Technology and Secured Transactions, IEEE: 492-497.

[11] Mohammad, R. M., F. Thabtah, et al. (2014). Intelligent Rule-Based Phishing Websites Classification. Information Security, IET8 (3): 153-160.