Prediction of Phishing Websites Using Optimization Techniques

R.Sumathi[1] and Mr.R.Vidhya Prakash[2]
[1]M.E Computer Science and Engineering, Sri Shakthi Institute of Engineering and Technology, Coimbatore, India
[2]Assistant Professor, Department of Computer Science and Engineering, Sri Shakthi Institute of Engineering and Technology, Coimbatore, India

ABSTRACT

Phishing website is a fraudulent attempt usually made through email, to steal personal information. Phishing emails usually appear to come from a well-known organization and ask for personal information such as credit card number, social security number, account number or password. Often times phishing attempts appear to come from sites, services and companies with which do not even have an account. This paper presents a novel approach to overcome the difficulty and complexity in predicting and detecting phishing website. In existing system they proposed an intelligent resilient and effective model that is based on using association and classification Data Mining algorithms. They implemented PART classification algorithm and techniques to extract the phishing data training sets criteria to classify their legitimacy. In the proposed system, we implement the PSO algorithm for predicting Phishing Websites. In this project, we present novel approach to overcome the 'fuzziness' in the phishing website assessment and propose an intelligent resilient and effective model for phishing websites. The experimental results demonstrated the feasibility of using Association and Classification techniques and PSO real applications and its better performance.

KEYWORDS:- APRIORI, ASSOCIATION, CLASSIFICATION, DATA MINING, FUZZY LOGIC, PHISHING, PSO, RISK ASSESSMENT.

1. INTRODUCTION

Phishing is an e-mail fraud method in which the perpetrator sends out legitimate-looking email in an attempt to gather personal and financial information from recipients. Phishing is similar to fishing in a lake, but instead of trying to capture fish, phishers attempt to steal personal information. They send out e-mails that appear to come from legitimate websites such as eBay, PayPal, or other banking institutions. The e-mails state that information needs to be updated or validated and ask that enter username and password, after clicking a link included in the e-mail. Some e-mails will ask that to enter even more information, such as full name, address, phone number, social security number, and credit card number. However, even if we visit the false website and just enter username and password, the phisher may be able to gain access to more information by just logging in to account. The word phishing from the phrase "website phishing" is a variation on the word "fishing". The idea is that bait is thrown out with the hopes that a user will grab it and bite into it just like the fish. The motivation behind this study is to create a resilient and effective method that uses Data Mining algorithms and tools to detect phishing websites in an Artificial Intelligent technique. An Optimization Technique can be very useful in predicting phishing websites. It can give us answers about what are the most important phishing website characteristics and indicators and how they relate with each other. Comparing

between different Data Mining Optimization methods and techniques is also a goal of this investigation. The paper is organized as follows: Section A presents the literature review, Section B shows data mining phishing approach, Section C shows the theory and methodology of the research, Section D shows the utilization of the DM classification techniques, Section III reveals the conclusions and future work.

2. RELATED WORKS

2.1. Literature Review

A report by Gartner estimated the costs at $1,244 per victim, an increase over the $257 they cited in a 2004 report [1]. In 2007, Moore and Clayton estimated the number of phishing victims by examining web server logs. They estimated that 311,449 people fall for phishing scams annually, costing around 350 million dollars [2]. There are several promising defending approaches to this problem reported earlier.

One approach is to stop phishing at the email level [3], since most current phishing attacks use broadcast email (spam) to lure victims to a phishing website. Another approach is to use security toolbars. The phishing filter in IE7 [4] is a toolbar approach with more features such as blocking the user's activity with a detected phishing site. A third approach is to visually differentiate the phishing sites from the spoofed legitimate sites. Dynamic Security Skins [5] proposes to use a randomly generated visual hash to customize the browser window or web form elements to indicate the successfully authenticated sites. A fourth approach is two-factor authentication , which ensures that the user not only knows a secret but also presents a security token [6]. Many industrial anti phishing products use toolbars in Web browsers, but some Researchers have shown that security tool bars don't effectively prevent phishing attacks. Another approach is to employ certification, e.g., Microsoft spam privacy [7]. A variant of web credential is to use a database or list published by a trusted party, where known phishing web sites are blacklisted. The weaknesses of this approach are its poor scalability and its timeliness. The newest version of Microsoft's Internet Explorer supports Extended Validation (EV) certificates, coloring the URL bar green and displaying the name of the company. However, a recent study found that EV certificates did not make users less fall for phishing attacks [8].

2.2. Phishing Data Mining Approach

2.2.1. Phishing Characteristics and Indicators

There are many characteristics and indicators that can distinguish the original legitimate website from the phishing one. We managed to gather 27 phishing features and indicators and clustered them into six Criteria (URL & Domain Identity, Security & Encryption, Source Code & Java script, Page Style & Contents, Web Address Bar and Social Human Factor), and each criteria has its own phishing components. The full list is shown in table I which is used later on our analysis and methodology study.

2.2.2. Why use Data Mining?

DM is the process of searching through large amounts of data and picking out relevant information. It has been described as "the nontrivial extraction of implicit, previously unknown, and potentially useful information from large data sets [9]. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions [10].

3. THEORY AND METHODOLOGY

3.1. Data Mining Techniques

The approach described here is to apply data mining algorithms to assess e-banking phishing website risk on the 27 characteristics and factors which stamp the forged website. We utilized data mining classification and association rule approaches in our new phishing website detection model as shown in figure I to find significant patterns of phishing characteristic or factors in the e-banking phishing website archive data. Particularly, we used a number of different existing data mining association and classification techniques.Including JRip, PART [II], PRISM [12] and C4.5 [13], CBA [14], MCAR [15] algorithms to learn and to compare the relationships of the different phishing classification features and rules. The experiments of C4.5, RIPPER, PART and PRISM algorithms were conducted using the *WEKA* software 16]. CBA and MCAR experiments were conducted using an implementation provided by the authors of [14], [15]. We used two web access archives, one from APWG archive [17] and one from Phishtank archive [16]. We managed to extract the whole 27 phishing security features and clustered them to its 6 corresponding criteria as mentioned before in table 1.

3.2. Website Phishing Training Data Sets

Two publicly available datasets were used to test our implementation: the "phishtank" from the phishtank.com [16] which is considered one of the primary phishing-report collates both the 2007 and 2008 collections. The PhishTank database records the URL for the suspected website that has been reported, the time of that report, and sometimes further detail such as the screenshots of the website, and is publicly available. The Anti Phishing Working Group (APWG) which maintains a "Phishing Archive" describing phishing attacks dating back to September 2007 [3]. A data set of 1006 phishing, suspicious and legitimate websites is used in the study (412 row phishing websites, 288 rows suspicious and 306 row of real 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.

Table.1. Main Phishing Indicators with its Criteria

Criteria	N	Phishing Indicators
URL & Domain Identity	1	Using IP address
	2	Abnormal Request URL
	3	Abnormal URL of anchor
	4	Abnormal DNS record
	5	Abnormal URL
Security & Encryption	1	Using SSL certificate(Padlock Icon)
	2	Certificate Authority
	3	Abnormal Cookie
	4	Distinguished name certificate
Source code & Java Script	1	Redirect pages
	2	Straddling Attack
	3	Phanning Attack
	4	OnMouseOver to hide the link
	5	Server Form Handler (SFH)
Page style & Contents	1	Spelling errors
	2	Copying Websites
	3	Using forms with *Submit* button
	4	Using pop-ups Window
	5	Disabling right click
Web Address Bar	1	Long URL address
	2	Replacing similar char for URL
	3	Adding a Prefix or Suffix
	4	Using the @ symbol to confuse
	5	Using hexadecimal char codes
Social Human Factor	1	Emphasis on Security
	2	Public generic salutation
	3	Buying time to access accounts

4. DM Classification Techniques

4.1. Associative Classification Algorithms

The practical part of this comparative study utilizes six different common OM classification algorithms (C4.5, JRip, PART, PRISM, CBA and MCAR). Our choice of these methods is based on the different strategies they used in learning rules from data sets. The C4.5 algorithm [13] employs divide and conquer approach, and the RIPPER algorithm uses separate and conquer approach. The choice of PART algorithm is based on the fact that it combines both approaches to generate a set of rules. PRISM is a classification rule which can only deal with nominal attributes and doesn't do any pruning. CBA algorithm employs association rule mining [14] to learn the classifier and then adds a pruning and prediction steps. Finally, MCAR algorithm consists of two phases: rules generation and a classifier builder. In the first phase, MCAR scans the training data set to discover frequent single items, and then recursively combines the items generated to produce items involving more attributes. MCAR then generates ranks and stores the rules. In the second phase, the rules are used to generate a classifier by considering their effectiveness on the training data set [IS].

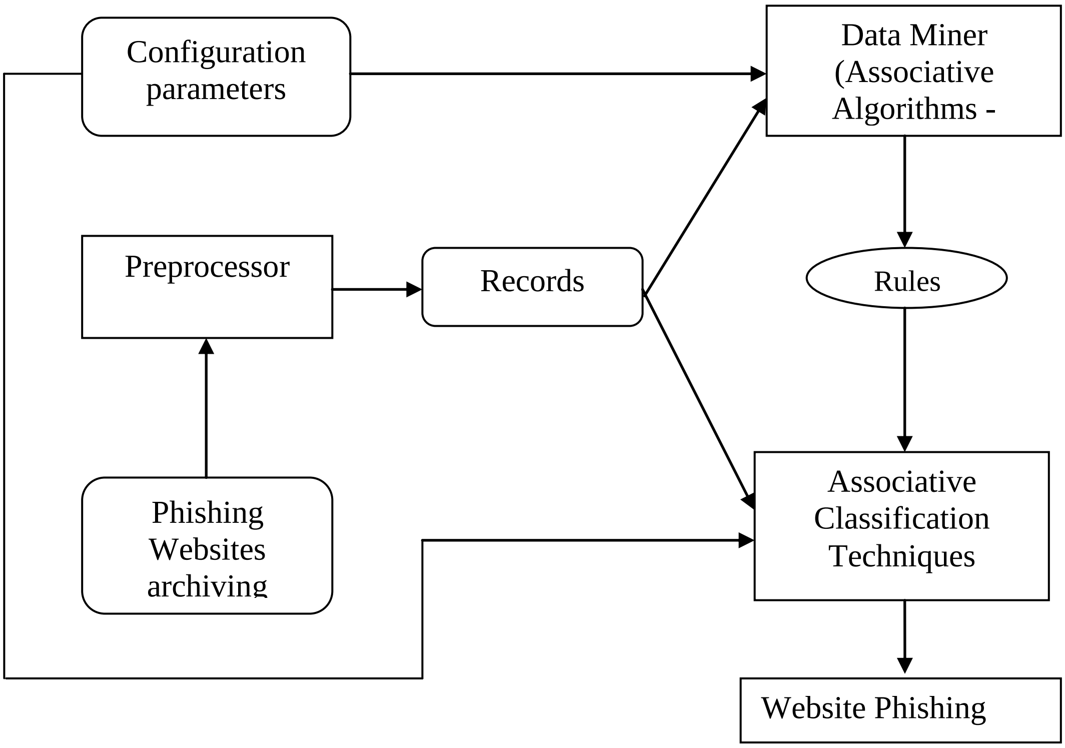
Configuration parameters
Data Miner (Associative Algorithms -
Preprocessor
Records
Rules
Phishing Websites archiving
Associative Classification Techniques
Website Phishing

Figure.1 AC Model for Detecting Phishing

4.2. MCAR Phishing Model Approach Associative

Classification is a special case of association rule mining in which only the class attribute is considered in the rule's right-hand-side 11(consequent), for example A, B -) Y, Then A, B must be input items attributes and Y must be the output class attribute. The attribute values for all our input items which represent the six ebanking phishing features and criteria ranged between three fuzzy set values (Genuine, Doubtful and Legitimate) which we measured before in our previous paper using Fuzzy Logic [17] taking into consideration all the input fuzzy variables for all criteria different components as shown in Table I. The output class attribute of our ebanking phishing website rate is one of these values *(Very Legitimate, Legitimate, Suspicious, Phishy or Very Phishy).* Example of the training phishing data sets to be classified is shown in Table 2. To derive a set of class association rules from the training data set, it must satisfy certain user-constraints,i.e support and confidence thresholds. Generally, in association rule mining, any item that passes *MinSupp* is known as a frequent item. We recorded the prediction accuracy and the number of rules generated by the classification algorithms.

Table 2.Example of Training Phishing Data Sets

Row ID	**URL**	**Security**	**Java**	**Style**	**Address**	**Social**	**Class/Phishing**
1	G	G	D	G	G	G	**Very Legitimate**
2	D	G	G	D	G	D	**Legitimate**
3	D	D	G	F	D	G	**Suspicious**

4	F	D	G	D	F	D	**Phishy**
5	D	F	F	D	F	F	**Very Phishy**
*	**G=Geniue**			**D=Doubtful**			**F=Fraud**

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

Table 3 Results From Weka four Classifiers

	C4.5	P.A.R.T	JRip	PRISM
Test Mode	10 FOLD CROSS VALIDATION			
Attributes	URL Domain Identity Source Code & Java Web Address Bar Security & Encryption Page Style & Contents Social Human Factor			
No. of Rules	57	38	14	155
Correct Classified	848 (84.2 %)	869 (86.3%)	818 (81.3 %)	855 (84.9 %)
InCorrect Classified	158 (15.7%)	137 (13.6%)	188 (18.8%)	141 (14.0%)
No. of instances	1006	1006	1006	1006

Table 4.Results from CBA and MCAR Classifiers

	CBA	**MCAR**
Num of Test Case	1006	1006
Correct Prediction	873	886
Error rate	13.452%	12.622%
Min Sup	20.000%	20.000%
Min Conf	100.000%	100.000%
Number of rules	15	22

4.3. Particle Swarm Optimization

Particle Swarm Optimization is a population based heuristic optimization algorithm inspired by social behavior of birds flocking or fish schooling. Each particle is treated as a point in a D-dimensional space. Initially, N particles are uniformly distributed in the solution space. The particles in PSO fly through the search space with a certain velocity, and change their position dynamically in the hope of reaching the food source, the destination. Therefore, position and velocity are two important parameters in the PSO algorithm.

Each particle keeps track of the best position it has encountered during its travel, and the best position traveled by the swarm of particles. The best position traveled by a particle is called the *local best position*, and the best position traveled by the swarm is called the *global best position*. At the end of each iteration, the particles calculate their next velocity, and update their positions based on the calculated velocity.

Basic algorithm for PSO:

```
1. Initialize

        (a) Set constants $k_{max}$, $c_1$, $c_2$.

        (b) Randomly initialize particle positions $x^i_o \in D$ in $IR^n$ for i = 1, …, p.

        (c) Randomly initialize particle velocities $0 \le v^i_o \le v^{max}_o$ for i = 1, …, p.

        (d) Set k = 1

2. Optimize

        (a) Evaluate function value $f^i_k$ using design space coordinates $x^i_k$.

        (b) If $f^i_k \le f^i_{best}$ then $f^i_{best} = f^i_{k,} p^i_k = x^i_k$.

        (c) If $f^i_k \le f^g_{best}$ then $f^g_{best} = f^i_{k,} p^g_k = x^i_k$.

        (d) If stopping condition is satisfied then goto 3.

        (e) Update all particle velocities $v^i_k$ for i = 1, …, p

        (f) Update all particle positions $x^i_k$ for i = 1, …, p

        (g) Increment k.

        (h) Go to 2(a)

3. Terminate
```

Figure 2. Flow Diagram of PSO-SVR forecasting model

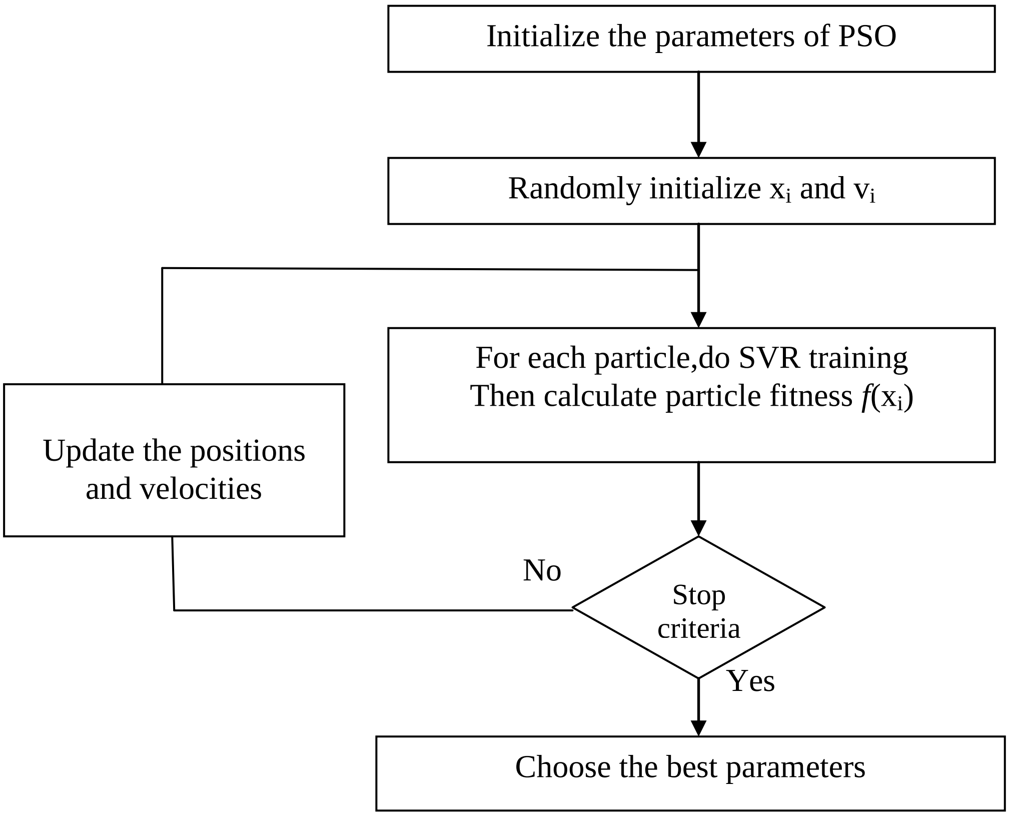
Initialize the parameters of PSO
Randomly initialize x_i and v_i
For each particle,do SVR training
Then calculate particle fitness $f(x_i)$
Update the positions and velocities
No
Stop criteria
Yes
Choose the best parameters

Table 5. Results From MCAR and PSO four Classifiers

	MCAR	PSO
Num of Test Case	1006	1006
Correct Prediction	886	934
Error rate	12.622%	9.544%
Min Sup	20.000%	20.000%
Min Conf	100.000%	100.000%
Number of rules	22	27

5. CONCLUSIONS

The Partial swarm optimization data mining phishing website model showed the significance importance of the phishing website two criteria's (URL & Domain Identity) and (Security & Encryption) with insignificant trivial influence of some other criteria like 'Page Style & content' and 'Social Human Factor' in the final phishing rate, which can help us in building website phishing detection system. The experiments indicate that Partical swarm optimization technique is highly competitive when compared with other traditional classifications in term of prediction accuracy and efficiency.

REFERENCES

[I] GARTNE R, INC. Gartner Says Number of Phishing Emails Sent to U.S. Adults Nearly Doubles in Just Two Years, http ://www.gartner.com/it/page.jsp?id=498245.November 9 2006 .

[2] T. Moore and R. Clayton, "An empirical analysis of the current state of phishing attack and defense", In Proceedings of the Workshop on the Economics of Information Security (WEIS2007).

[3] B. Adida, S. Hohenberger and R. Rivest , "Lightweight Encryption for Email," USENIX Steps to Reducing Unwanted Traffic on the Internet (SRUTI), 2005 .

[4] T. Sharif, "Phishing Filter in IE7," http://blogs,msdn.com/ie/archive/2005 /09/09/463204,as px,,2006,

[5] R. Dhamija and J.D. Tygar, "The Battle against Phishing: Dynamic Security Skins," Proc , Syrnp. Usable Privacy and Security, 2005.

[6] FDIC., "Putting an End to Account-Hijacking Identity Theft," fdic .gov/idtheftstudy/identity_theft.pdf, 2004 , [7] Microsoft, "microsoft,com/twc/privacv/spam", 2004

[8] C. Jackson, D. Simon, D. Tan, and A. Barth, "An evaluation of extended validation and picture-in-picture phishing attacks". In Proceedings of the 2007 Usable Security. www. usablesecurity.orgipapers/jackson.pdf

[9] Kantardzic and Mehmed. *"Data Mining: Concepts, Models, Methods, and Algorithms ,"*, John Wiley & Sons. ISBN 0471228524. OCLC 50055336,2003.[10] U,M, Fayyad, "Mining Databases: Towards Algorithms for Discovery," Data Eng, Bull., vol. 21, no, 1, pp, 3948,1998.

[II] I.H. Witten and E. Frank, "Data Mining : Practical machine learning tools and techniques", 2nd Edition,Morgan Kaufmann, San Francisco, CA, 2005 ,

[12] J. Cendrowska., *"PRISM: An algorithm for inducingmodular rule"*, International Journal of Man-Machine Studies (1987), Vo1.27, No.4, pp.349-370.

[13] J. R. Quinlan, "Improved use of continuous attributes in c4.5", Journal of Artificial Intelligence Research, 4:7790,1996,

[14] Bing Liu, Wynne Hsu, Yiming Ma, "Integrating Classification and Association Rule Mining."*Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98,Plenary Presentation),* New York, USA, 1998.

[15] '1', Fadi, c.Peter and Y. Peng, *"MCAR: Multi-classClassification based on Association Rule"*, IEEE International Conference on Computer Systems and Applications ,2005, pp. 127-133.

[16] WEKA - University of Waikato, New Zealand, EN,2006: "Weka -Data Mining with Open Source Machine Learning Software in Java", 2006 ,

AUTHORS

Ms.R.Sumathi received B.E degree in Computer Science and Engineering from Avinashilingam University, Coimbatore and Currently pursuing M.E degree in Computer Science and Engineering in Sri Shakthi Institute of Engineering and Technology, under Anna University of Technology, Coimbatore. Her research interest includes Data mining and Computer Networks.

Mr.R.Vidhya Prakash received the M.E degree in Software Engineering from P.S.G College of Technology, Coimbatore and received the B.E degree in Computer Science and Engineering from M.P.Nachimuthu M.Jaganathan Engineering College under Anna University Chennai. He is currently working as Assistant Professor in Department of Computer Science and Engineering in Sri Shakthi Institute of Engineering and Technology,Coimbatore. He has presented papers in conferences. His main research interest is Software Engineering and Computer Networks.