Modelling Intelligent Phishing Detection System for e-Banking using Fuzzy Data Mining

Maher Aburrous
Dept. of Computing
University of Bradford
Bradford, UK
mrmaburr@bradford.ac.uk

M. A. Hossain
Dept. of Computing
University of Bradford
Bradford, UK
m.a.hossain1@bradford.ac.uk

Keshav Dahal
Dept. of Computing
University of Bradford
Bradford, UK
k.p.dahal@bradford.ac.uk

Fadi Thabatah
MIS Department
Philadelphia University
Amman, Jordan
ffayez@philadelphia.edu.jo

***Abstract*- Detecting and identifying any phishing websites in real-time, particularly for e-banking is really a complex and dynamic problem involving many factors and criteria. Because of the subjective considerations and the ambiguities involved in the detection, Fuzzy Data Mining (DM) Techniques can be an effective tool in assessing and identifying phishing websites for e-banking since it offers a more natural way of dealing with quality factors rather than exact values. In this paper, we present novel approach to overcome the 'fuzziness' in the e-banking phishing website assessment and propose an intelligent resilient and effective model for detecting e-banking phishing websites. The proposed model is based on Fuzzy logic (FL) combined with Data Mining algorithms to characterize the e-banking phishing website factors and to investigate its techniques by classifying there phishing types and defining six e-banking phishing website attack criteria's with a layer structure. The proposed e-banking phishing website model showed the significance importance of the phishing website two criteria's (URL & Domain Identity) and (Security & Encryption) in the final phishing detection rate result, taking into consideration its characteristic association and relationship with each others as showed from the fuzzy data mining classification and association rule algorithms. Our phishing model also showed the insignificant trivial influence of the (Page Style & Content) criteria along with (Social Human Factor) criteria in the phishing detection final rate result.**

Keywords- Phishing, Fuzzy Logic, data mining, classification, association, apriori, e-banking risk assessment

I. INTRODUCTION

E-banking Phishing websites are forged websites that are created by malicious people to mimic real e-banking websites. Most of these kinds of Web pages have high visual similarities to scam their victims. Some of these Web pages look exactly like the real ones. Unwary Internet users may be easily deceived by this kind of scam. Victims of e-banking phishing Websites may expose their bank account, password, credit card number, or other important information to the phishing Web page owners. The impact is the breach of information security through the compromise of confidential data and the victims may finally suffer losses of money or other kinds. Phishing is a relatively new Internet crime in comparison with other forms, e.g., virus and hacking. More and more phishing Web pages have been found in recent years in an accelerative way [7]**.** 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. In most cases, bait is either an e-mail or an instant messaging site, which will take the user to hostile phishing websites [10]. E-banking Phishing website is a very complex issue to understand and to analyze, since it is joining technical and social problem with each other for which there is no known single silver bullet to entirely solve it. The motivation behind this study is to create a resilient and effective method that uses Fuzzy Data Mining algorithms and tools to detect e-banking phishing websites in an automated manner. DM approaches such as neural networks, rule induction, and decision trees can be a useful addition to the fuzzy logic model. It can deliver answers to business questions that traditionally were too time consuming to resolve such as, "Which are most important e-banking Phishing website Characteristic Indicators and why?" by analyzing massive databases and historical data for training purposes.

The paper is organized as follows: Section 2 presents the literature review and related work, Section 3 shows the theory and methodology of the proposed fuzzy based data mining approach for the phishing website risk assessment model. Section 4 introduces the system design and implementation with the overall fuzzy data mining inference rules. Section 5 reveals the experiments and results of the fuzzy data mining e-banking phishing website risk assessment model and then conclusions and future work are given in Section 6.

II. LITERATURE REVIEW AND RELATED WORK

A. 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 e-banking phishing website attacks, there are several promising defending approaches to this problem reported earlier. In this section, we briefly survey existing anti-phishing solutions and list of the related works. 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 [21]. Another approach is to use security toolbars. The phishing filter in IE7 [19] 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

978-0-7695-3791-7/09 $26.00 © 2009 IEEE
DOI 10.1109/CW.2009.43

[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]. However, this approach is a server-side solution. Phishing can still happen at sites that do not support two-factor authentication. Sensitive information that is not related to a specific site, *e.g.*, credit card information and SSN (Social Security Number), cannot be protected by this approach either [22].

Many industrial antiphishing products use toolbars in Web browsers, but some researchers have shown that security tool bars don't effectively prevent phishing attacks. [4], [5] proposed a scheme that utilises a cryptographic identity-verification method that lets remote Web servers prove their identities. However, this proposal requires changes to the entire Web infrastructure (both servers and clients), so it can succeed only if the entire industry supports it. In [13], the authors proposed a tool to model and describe phishing by visualizing and quantifying a given site's threat, but this method still wouldn't provide an antiphishing solution. Another approach is to employ certification, e.g., Microsoft spam privacy [43], [14], [15], [17], [1]. A recent and particularly promising solution was proposed in [8], which combines the technique of standard certificates with a visual indication of correct certification; a site-dependent logo indicating that the certificate was valid would be displayed in a *trusted credentials area* of the browser. A variant of web credential is to use a database or list published by a trusted party, where known phishing web sites are blacklisted. For example Netcraft antiphishing toolbar [44], prevents phishing attacks by utilising a centralized blacklist of current phishing URLs. Other Examples include Websense, McAfee's anti–phishing filter, Netcraft anti-phishing system, Cloudmark SafetyBar, and Microsoft Phishing Filter [16]**.** The weaknesses of this approach are its poor scalability and its timeliness. Note that phishing sites are cheap and easy to build and their average lifetime is only a few days. APWG provides a solution directory at (Anti-Phishing Working Group) [2] which contains most of the major antiphishing companies in the world. However, an automatic antiphishing method is seldom reported. The typical technologies of antiphishing from the user interface aspect are done by [5] and [22]. They proposed methods that need Web page creators to follow certain rules to create Web pages, either by adding dynamic skin to Web pages or adding sensitive information location attributes to HTML code. However, it is difficult to convince all Web page creators to follow the rules [7]. In [12], [7], [13], [20], the visual similarity of Web pages is oriented, and the concept of visual approach to phishing detection was first introduced. Through this approach, a phishing Web page can be detected and reported in an automatic way rather than involving too many human efforts. Their method first decomposes the Web pages (in HTML) into salient (visually distinguishable) block regions. The visual similarity between two Web pages is then evaluated in three metrics: block level similarity, layout similarity, and overall style similarity, which are based on the matching of the salient block regions [7].

B. Main Characteristics of e-banking phishing websites.

Evolving with the antiphishing techniques, various phishing techniques and more complicated and hard-to-detect methods are used by phishers. The most straightforward way for a phisher to defraud people is to make the phishing Web pages similar to their targets. Actually, there are many characteristics and factors that can distinguish the original legitimate website from the forged e-banking phishing website like Spelling errors, Long URL address and Abnormal DNS record. The full list is shown in table I which is used later on our analysis and methodology study.

Table I. COMPONENTS AND LAYERS OF E-BANKING PHISHING WEBSITE CRITERIA.

Criteria	N	Component	Layer No.
URL & Domain Identity (Weight = 0.3)	1	Using the IP Address	Layer One Sub weight = 0.3
	2	Abnormal Request URL	
	3	Abnormal URL of Anchor	
	4	Abnormal DNS record	
	5	Abnormal URL	
Security & Encryption (Weight = 0.2)	1	Using SSL certificate	Layer Two
	2	Certification authority	
	3	Abnormal Cookie	
	4	Distinguished Names Certificate(DN)	
Source Code & Java script (Weight = 0.2)	1	Redirect pages	Sub weight = 0.4
	2	Straddling attack	
	3	Pharming Attack	
	4	OnMouseOver to hide the Link	
	5	Server Form Handler (SFH)	
Page Style & Contents (Weight =0.1)	1	Spelling errors	Layer Three
	2	Copying website	
	3	Using forms with *Submit* button	
	4	Using Pop-Ups windows	
	5	Disabling Right-Click	
Web Address Bar (Weight = 0.1)	1	Long URL address	Sub weight = 0.3
	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 (Weight = 0.1)	1	Emphasis on security	
	2	Public generic salutation	
	3	Buying time to access accounts	
Total Weight			**1**

C. Why using Fuzzy Logic and Data Mining?

FL has been used for decades in the engineering sciences to embed expert input into computer models for a broad range of applications. It offers a promising alternative for measuring operational risks [18]. The FL approach provides more information to help risk managers effectively manage assessing and ranking e-banking phishing website risks than the current qualitative approaches as the risks are quantified based on a combination of historical data and expert input. The advantage of the fuzzy approach is that it enables processing of vaguely defined variables, and variables

whose relationships cannot be defined by mathematical relationships. FL can incorporate expert human judgment to define those variable and their relationships.

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 [30], [31]. It is a powerful new technology with great potential to help researchers focus on the most important information in their data archive. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions [32].

III. The Proposed Fuzzy based Data Mining Approach

A. Fuzzy Data Mining Algorithms & Techniques

The approach described here is to apply fuzzy logic and data mining algorithms to assess e-banking phishing website risk on the 27 characteristics and factors which stamp the forged website. The essential advantage offered by fuzzy logic techniques is the use of linguistic variables to represent Key Phishing characteristic indicators and relating e-banking phishing website probability.

1) Fuzzification

In this step, linguistic descriptors such as High, Low, Medium, for example, are assigned to a range of values for each key phishing characteristic indicators. Valid ranges of the inputs are considered and divided into classes, or fuzzy sets. For example, length of URL address can range from 'low' to 'high' with other values in between. We cannot specify clear boundaries between classes. The degree of belongingness of the values of the variables to any selected class is called the degree of membership; Membership function is designed for each Phishing characteristic indicator, which is a curve that defines how each point in the input space is mapped to a membership value between [0, 1]. Linguistic values are assigned for each Phishing indicator as *Low, Moderate, and High* while for e-banking Phishing website risk rate as *Very legitimate, Legitimate, Suspicious, Phishy, and Very phishy* (*triangular and trapezoidal membership function*). For each input their values ranges from 0 to 10 while for output, ranges from 0 to 100. An example of the linguistic descriptors used to represent one of the key phishing characteristic indicators (URL Address Long) and a plot of the **fuzzy membership functions** are shown in figure 1.

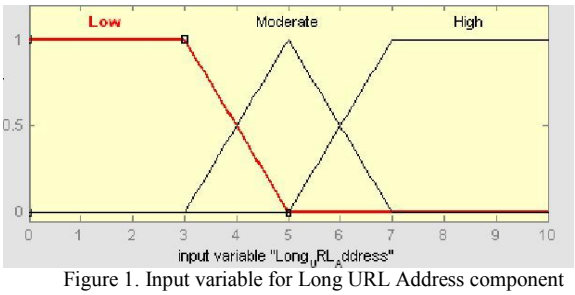

Figure 1. Input variable for Long URL Address component

The fuzzy representation more closely matches human cognition, thereby facilitating expert input and more reliably representing experts' understanding of underlying dynamics [4]. The same approach is used to calibrate the other 26 Key Phishing Characteristic Indicators.

2) Rule Generation using Classification Algorithms.

Having specified the risk of e-banking phishing website and its key phishing characteristic indicators, the next step is to specify how the e-banking phishing website probability varies. Experts provide fuzzy rules in the form of ***if…then*** statements that relate e-banking phishing website probability to various levels of key phishing characteristic indicators based on their knowledge and experience. On that matter and instead of employing an expert system, we utilised data mining classification and association rule approaches in our new e-banking phishing website risk assessment model as shown in figure 2 to automatically 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 classification techniques implemented within WEKA [27] and CBA packages [33]. JRip [34] WEKA's implementation of RIPPER, PART [34], Prism [35] and C4.5 [36] algorithms are selected to learn the relationships of the selected different phishing features. We have chosen these algorithms since the learnt classifiers are easily understood by human [29]. While for the association finding we have used the apriori [37] and predictive apriori algorithm [38] using WEKA.

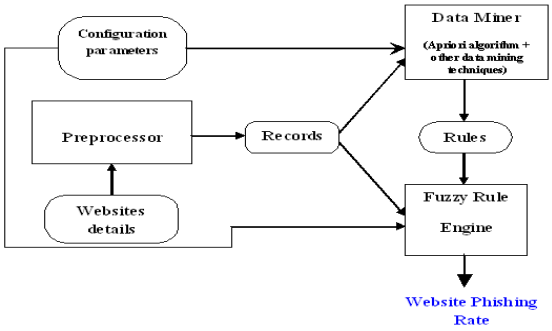

Figure 2. E-banking Phishing Website Risk Assessment Model

We used two web access archives, one from APWG archive [2] and one from Phishtank archive [39]. We managed to extract 6 different feature sets from the e-banking phishing website archives, and then derived many important rules which helped us in the fuzzy rule phase.

3) Aggregation of the rule outputs.

This is the process of unifying the outputs of all discovered rules. Combining the membership functions of all the rules consequents previously scaled into single fuzzy sets (output).

4) Defuzzification.

This is the process of transforming a fuzzy output of a fuzzy inference system into a crisp output. Fuzziness

helps to evaluate the rules, but the final output has to be a crisp number. The input for the defuzzification process is the aggregate output fuzzy set and the output is a number. This step was done using Centroid technique [40] since it is a commonly used method. The output is e-banking phishing website risk rate and is defined in fuzzy sets like '**very phishy**' to '**very legitimate'**. The fuzzy output set is then defuzzified to arrive at a scalar value.

B. Data Sets and Experimental Results

Two publicly available datasets were used to test our implementation: the "phishtank" from the phishtank.com [39] which is considered one of the primary phishing-report collators both the 2007 and 2008 collections, for a total of approximately 606 e-banking phishing websites. 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 [2]. We performed a cognitive walkthrough on 1006 sample attacks within this archive. 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 the strategies that are used by attackers and to formulate hypotheses 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 the usage of these techniques has changed over time. By investigating these information, we have found some interesting techniques depending on the main perception that phishers know that most users don't know how to check the security and often assumes that sites requesting sensitive information are secure which makes very difficult for them to see the difference between authentic security and mimicked security features [23]. We also found that some visual deception attacks can fool even the most sophisticated users. These results illustrate that standard security indicators are not effective for a substantial fraction of users, and suggest that alternative approaches are needed [24].

C. Mining e-banking Phishing Websites Challenges

There are a number of challenges posed by doing post-hoc classification of e-banking phishing websites. Most of these challenges only apply to the e-banking phishing websites data and materialize as a form of information, which has the net effect of increasing the false negative rate. The age of the dataset is the most significant problem, which is particularly relevant with the phishing corpus. E-banking Phishing websites are short-lived, often lasting only in the order of 48 hours. Some of our features can therefore not be extracted from older websites, making our tests difficult. The average phishing site stays live for approximately 2.25 days [25]. Furthermore, the process of transforming the original e-banking phishing website archives into record feature data sets is not without error. It requires the use of heuristics at several steps. Thus high accuracy from the data mining algorithms cannot be expected. However, the evidence supporting the golden nuggets comes from a number different algorithms and feature sets and we believe it is compelling [26].

D. Utilization of different DM Classification algorithms

The practical part of this study utilises five different common DM algorithms (C4.5, Ripper, Part, Prism, CBA). Our choice of these methods is based on the different strategies they used in learning rules from data sets [28]. The C4.5 algorithm [36] 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. It adapts separate-and-conquer to generate a set of rules and uses divide-and-conquer to build partial decision trees. The way PART builds and prunes partial decision tree is similar to the C4.5 implementation with a difference which can be explained as follows: C4.5 generates one decision tree and uses pruning techniques to simplify it; each path from the root node to one of the leaves in the tree represents a rule. On the other hand, PART avoids the simplification process by building up partial decision trees and choosing only one path in each one of them to derive a rule. Once the rule is generated, all instances associated with it, and the partial tree will be discarded. PRISM is a classification rule which can only deal with nominal attributes and doesn't do any pruning. It implements a top-down (general to specific) sequential-covering algorithm that employs a simple accuracy-based metric to pick an appropriate rule antecedent during rule construction. Finally, CBA algorithm employs association rule mining [33] to learn the classifier and then adds a pruning and prediction steps. This results in a classification approach named associative classification [41] [42].

We recorded the prediction accuracy of the considered classification approaches we used in this study in Table II and Table III. The overall summary output can be interpreted as: Web Address Bar and URL Domain Identity are the major important criteria for identifying and detecting e-banking phishing website. Such as if one or both of them is genuine then most likely the website is legitimate and on the other hand if it is Fraud, then the website is most likely phishing. The classification rules did not just showed the significance roll of the Web Address Bar criteria and URL Domain Identity criteria but showed also the magnitude value of some other e-banking phishing website criteria like Security & Encryption criteria comparing to the others. We have used ten-fold-cross-validation as a testing mode which evaluating the derived classifiers. Cross validation is a well-known testing method in DM and machine learning communities.

TABLE II. RESULTS FROM WEKA CLASSIFIER USING 4 METHODS APPLIED TO CLASSIFY PHISHING

	C4.5 Decision Tree	P.A.R.T.	JRip R.I.P.P.E. R.	PRISM
Test Mode	10 FOLD CROSS VALIDATION			
Attributes	URL DOMAIN IDENTITY SECURITY & ENCRYPTION SOURCE CODE & JAVA PAGE STYLE & CONTENTS WEB ADDRESS BAR SOCIAL HUMAN FACTOR CLASS			
Number of Rules	57	38	14	155
Correctly Classified	848 (84.2 %)	869 (86.3 %)	818 (81.3%)	855 (84.9%)
Incorrectly Classified	158 (15.7%)	137 (13.6%)	188 (18.6 %)	141 (14.0%)
Number of instances	1006	1006	1006	1006

TABLE III. RESULTS FROM CBA CLASSIFIER USING ASSOCIATION RULE MINING APPLIED TO CLASSIFY PHISHING

	Mine: Single Sup	Mine: Multi Sup
Num of Test Case	1006	1006
Correct Prediction	758	713
Error Rate	24.652%	29.125%
MinSup	20.000%	10.000%
MinConf	100.000%	100.000%
RuleLimit	80000	80000
LevelLimit	6	6
Number of rules	*22*	*15*

IV. SYSTEM DESIGN

In this paper, 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 as shown in Table I. This table also shows that there are different numbers of components for each criterion, five components for URL & Domain Identity, Source Code & Java script, Page Style & Contents, and Web Address Bar, respectively. Four components for Security & Encryption, and three components for Social Human Factor. Therefore, there are twenty seven components in total. There are three layers on this e-banking phishing website fuzzy data mining model as shown in Figure 3. The first layer contains only URL & Domain Identity criteria with a weight equal to 0.3 for its importance; the second layer contains Security & Encryption criteria and Source Code & Java script criteria with a weight equal to 0.2 each; the third layer contains Page Style & Contents criteria, Web Address Bar criteria and Social Human Factor criteria with a weight equal to 0.1 each. The six criteria have been classified and prioritized through mining the e-banking phishing website archive database using the classification and association algorithms mentioned earlier.

E-banking Phishing Website Rating = 0.3 * URL & Domain Identity crisp [First layer] + ((0.2 * Security & Encryption crisp)+(0.2 * Source Code & Java script crisp)) [Second layer] + ((0.1 * Page Style & Contents crisp) +(0.1 * Web Address Bar crisp) + (0.1 * Social Human Factor crisp)) [Third layer]

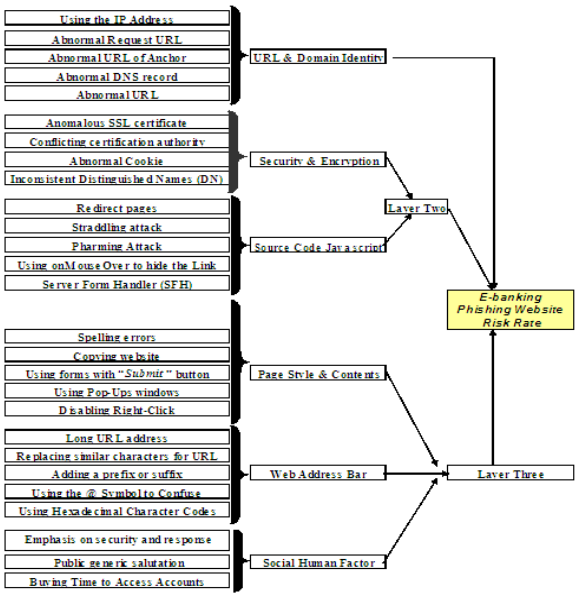

Figure 3. Structure of the fuzzy data mining inference overall system to evaluate e-banking phishing website risk rate.

A. *Overall Fuzzy Data Mining Inference Rules*

1) *The Rule Base1 for layer 1.*

The rule base has five input parameters and one output and contains all the "IF-THEN" rules of the system. For each entry of the rule base, each component is assumed to be one of three values and each criterion has five components. Therefore, the rule base 1 contains $(3^5) = 243$ entries. The output of rule base 1 is one of the e-banking phishing website rate fuzzy sets (Genuine, Doubtful or Fraud) representing URL & Domain Identity criteria phishing risk rate. A sample of the structure and the entries of the rule base 1 for layer 1 are shown in Table IV. The system structure for URL & Domain Identity criteria is the joining of its five components, which produces the URL & Domain Identity criteria (Layer one).

TABLE IV. SAMPLE OF THE RULE BASE 1 STRUCTURE AND ENTRIES FOR URL & DOMAIN IDENTITY CRITERIA

Rule #	(comp. 1) Using the IP Address	(comp. 2) Abnormal Req. URL	(comp. 3) Abnormal URL	(comp. 4) Abnormal DNS record	(comp. 5) Abnormal URL	URL & Domain Identity Criteria
1	Low	Low	Low	Low	Low	Genuine
2	Low	Low	Low	Low	Mod.	Genuine
3	Low	Low	Mod.	Mod.	Mod.	Doubtful
4	Low	Low	Low	Mod.	high	Doubtful
5	Low	Low	Mod.	Mod.	high	Fraud
6	Mod.	Mod.	Mod.	Low	high	Fraud
7	Mod.	Low	high	Mod.	high	Fraud
8	high	Mod.	Low	Mod.	Low	Doubtful
9	Low	Mod.	Low	Low	Mod.	Fraud
10	high	Mod.	high	high	Low	Fraud

2) *The Rule Base for layer 2.*

In Layer 2, there are two inputs, which are (Security & Encryption and Source Code & Java script) and one output. The system structure for Security & Encryption criteria is the joining of its four components (Using SSL certificate, Certification authority, Abnormal Cookie and Distinguished Names Certificate(DN)) using Rule base 1,

which produces Security & Encryption criteria. The system structure for Source Code & Java script criteria is the joining of its five components (Redirect pages, Straddling attack, Pharming Attack, Using onMouseOver to hide the Link and Server Form Handler (SFH)) using Rule base 1, which produces Source Code & Java script criteria. The structure and the entries of the rule base for layer 2 are illustrated in Table V. The system structure for layer 2 is the combination of two e-banking phishing website criteria (Security & Encryption and Source Code & Java script), which produces rule base 2. The rule base contains (3^2) = 9 entries and the output of rule base 2 is one of the e-banking phishing website rate fuzzy sets (Legal, Uncertain or Fake) representing Layer Two criteria phishing risk rate.

TABLE V. THE RULE BASE 2 STRUCTURE AND ENTRIES FOR LAYER TWO

Rule	Security & Encryption	Source Code & Java script	Phishing Risk (Layer Two)
1	Genuine	Genuine	Legal
2	Genuine	Doubtful	Legal
3	Genuine	Fraud	Uncertain
4	Doubtful	Genuine	Uncertain
5	Doubtful	Doubtful	Uncertain
6	Doubtful	Fraud	Uncertain
7	Fraud	Genuine	Uncertain
8	Fraud	Doubtful	Fake
9	Fraud	Fraud	Fake

3) The Rule Base for layer 3.

In Layer 3, there are three inputs, which are: the Page Style & Contents, Web Address Bar and Social Human Factor which is the output from layer 3, and one output. The system structure for Page Style & Contents criteria is the joining of its five components (Spelling errors, Copying website, Using forms with "*Submit*" button, Using Pop-Ups windows and Disabling Right-Click) using Rule base 1, which produces Page Style & Contents criteria. The system structure for Web Address Bar criteria is the joining of its five components (Long URL address, Replacing similar characters for URL, Adding a prefix or suffix, Using the @ Symbol to Confuse and Using Hexadecimal Character Codes) using Rule base 1, which produces Web Address Bar criteria. The system structure for Social Human Factor criteria is the joining of its three components (Much emphasis on security and response, Public generic salutation and Buying Time to Access Accounts) using Rule base 1, which produces Social Human Factor criteria.

TABLE IIII. THE RULE BASE 3 STRUCTURE AND ENTRIES FOR LAYER THREE

Rule	Page Style & Contents	Web Address Bar	Social Human Factor	Phishing Risk (Layer 3)
1	Genuine	Genuine	Doubtful	Legal
2	Genuine	Doubtful	Fraud	Uncertain
3	Genuine	Fraud	Doubtful	Uncertain
4	Doubtful	Doubtful	Genuine	Uncertain
5	Doubtful	Doubtful	Doubtful	Uncertain
6	Doubtful	Fraud	Doubtful	Fake
7	Doubtful	Genuine	Genuine	Legal
8	Fraud	Doubtful	Doubtful	Uncertain
9	Fraud	Fraud	Fraud	Fake

A sample of the structure and the entries of the rule base for layer 3 are shown in Table VI. The system structure for layer 3 is the combination of Page Style & Contents, Web Address Bar and Social Human Factor, which produces rule base 3. The rule base contains (3^3) = 27 entries and the output of rule base 3 is one of the e-banking phishing website rate fuzzy sets (Legal, Uncertain or Fake) representing Layer Three criteria phishing risk rate.

4) The Rule Base for final e-banking phishing rate.

In the e-banking phishing website rule base last phase, there are three inputs, which are: layer one, layer two and layer three, and one output which is the rate of the e-banking phishing website. The structure and the entries of the rule base for e-banking phishing website rate are shown in Table VII.

TABLE IVI. THE E-BANKING PHISHING WEBSITE RATE RULE BASE STRUCTURE AND ENTRIES FOR FINAL PHISHING RATE

Rule	URL & Domain Identity	Layer Two	Layer Three	Final e-banking phishing website Rate
1	Genuine	Legal	Legal	Very Legitimate
2	Genuine	Legal	Uncertain	Legitimate
3	Genuine	Legal	Fake	Suspicious
4	Genuine	Uncertain	Legal	Suspicious
5	Genuine	Uncertain	Uncertain	Suspicious
6	Genuine	Uncertain	Fake	Phishy
7	Genuine	Fake	Legal	Suspicious
8	Genuine	Fake	Uncertain	Suspicious
9	Genuine	Fake	Fake	Phishy
10	Doubtful	Legal	Legal	Legitimate
11	Doubtful	Legal	Uncertain	Suspicious
12	Doubtful	Legal	Fake	Suspicious
13	Doubtful	Uncertain	Legal	Suspicious
14	Doubtful	Uncertain	Uncertain	Suspicious
15	Doubtful	Uncertain	Fake	Phishy
16	Doubtful	Fake	Legal	Phishy
17	Doubtful	Fake	Uncertain	Phishy
18	Doubtful	Fake	Fake	Very Phishy
19	Fraud	Legal	Legal	Suspicious
20	Fraud	Legal	Uncertain	Suspicious
21	Fraud	Legal	Fake	Phishy
22	Fraud	Uncertain	Legal	Suspicious
23	Fraud	Uncertain	Uncertain	Phishy
24	Fraud	Uncertain	Fake	Phishy
25	Fraud	Fake	Legal	Phishy
26	Fraud	Fake	Uncertain	Very Phishy
27	Fraud	Fake	Fake	Very Phishy

The system structure for is the combination of layer one, layer two and layer three, which produces final e-banking phishing website rule base. The rule base contains (3^3) = 27 entries and the output of final e-banking phishing website rule base is one of the final output fuzzy sets (Very Legitimate, Legitimate, Suspicious, Phishy or Very Phishy) representing final e-banking phishing website rate.

V. EXPERIMENTS AND RESULTS

Clipping method [9] is used in aggregating the consequences and the aggregated surface of the rule evaluation is defuzzified using Mamdani method [11] to find the Center Of Gravity (COG). Centroid

defuzzification technique shown in equation (1) can be expressed as where x^* is the defuzzified output, $\mu i(x)$ is the aggregated membership function and x is the output variable.

$$x^* = \frac{\int \mu_i(x)\, x\, dx}{\int \mu_i(x)\, dx} \qquad \text{Equation (1)}$$

The proposed intelligent e-banking Phishing website detection system has been implemented in MATLAB 6.5. The results of some input combinations are listed in Tables VIII, IX and X. The final e-banking phishing website risk rating will be balanced (54%) representing a [***suspicious website***], when the Layer one (URL & Domain Identity) of the e-banking phishing website risk criteria has 10 input values which indicate *High* phishing indicator and all other layers have the value of zero inputs as shown in Table VIII. Same result can be made when all e-banking phishing website risk criteria's representing by the three layers have middle (5) input values which indicate *Mod.* phishing indicator. These results shows the significance and importance of the e-banking phishing website criteria (URL & Domain Identity) represented by layer one especially when compared to the other criteria's and layers. Table IX shows that when the Layer one and Layer two of the e-banking phishing website risk criteria has middle (5) input values which indicate *Mod.* phishing indicator and other third Layer has the value of 10 input values which indicate *High* phishing indicator, the final e-banking phishing website risk rating will be reasonably high (72%) representing a [***phishy website***], which means that there is a Good guarantee that the website is forged phishy website. This result clearly shows that even if some of the e-banking phishing website characteristics or layers are not very clear or not definite, the website can still be phishy and forged, and users should be aware when dealing with it especially when other phishing characteristics or layers are obvious and clear.

Table X shows that when the Layer one of the e-banking phishing website risk criteria (URL & Domain Identity) has middle (5) input values which indicate *Mod.* phishing indicator and all other Layers has the value of zero input values which indicate *Low* phishing indicator, the final e-banking phishing website risk rating will be reasonably low (39%) representing a [***legitimate website***], which means that there is Good guarantee that the website is legitimate website. This result clearly shows that even if some of the e-banking phishing website characteristics or layers are noticed or observed, that does not mean at all that the website is phishy or forged, but it can be safe and secured especially when other phishing characteristics or layers are not noticeable, visible or detectable. The results also indicates that the worst e-banking phishing website rate (all three layers have 10 input value) equals 83.7% representing [Very Phishy Website] and the best e-banking phishing website rate (all three layers have 0 input value) is 16.4% representing [Very Legitimate Website] rather than a full range, i.e. 0 to 100, because of the fuzzification process

TABLE VI. FIVE HIGHEST (10) FOR LAYER ONE AND ALL OTHERS LOWEST (0).

Comp	Layer One URL & Domain	Layer Two		Layer Three			% ebanking phishing Rating
		Security & Encrypt	Source Code & Java	Page Style & Contents	Web Address Bar	Social Human Factor	
1	10	0	0	0	0	0	**54%**
2	10	0	0	0	0	0	
3	10	0	0	0	0	0	
4	10	0	0	0	0		
5	10		0	0	0		

TABLE IX. FIVE MIDDLE (5) INPUTS FOR LAYER ONE AND LAYER TWO AND HIGHEST (10) INPUTS FOR LAYER THREE.

Comp	Layer One URL & Domain	Layer Two		Layer Three			% ebanking phishing Rating
		Security & Encrypt	Source Code & Java	Page Style & Contents	Web Address Bar	Social Human Factor	
1	5	5	5	10	10	10	**72%**
2	5	5	5	10	10	10	
3	5	5	5	10	10	10	
4	5	5	5	10	10		
5	5		5	10	10		

TABLE VI. FIVE MIDDLE (5) INPUTS FOR LAYER ONE AND ALL OTHERS LOWEST (0) INPUTS.

Comp	Layer One URL & Domain	Layer Two		Layer Three			% ebanking phishing Rating
		Security & Encrypt	Source Code & Java	Page Style & Contents	Web Address Bar	Social Human Factor	
1	5	0	0	0	0	0	**39%**
2	5	0	0	0	0	0	
3	5	0	0	0	0	0	
4	5	0	0	0	0		
5	5		0	0	0		

VI. CONCLUSION AND FUTURE WORK

The fuzzy data mining e-banking phishing website model showed the significance importance of the phishing website two criteria's (URL & Domain Identity) and (Security & Encryption) in the final phishing detection rate result. The model showed the correlation or relationship between some of their characteristics like the conflict of using SSL certificate with the abnormal URL request as showed from the fuzzy data mining classification and association rule algorithms. Our phishing model also showed the insignificant trivial influence of the 'Page Style & content' criteria along with 'Social Human Factor' criteria in the final rate result of detecting phishing e-banking websites, taking into consideration its characteristic association with each other like Spelling errors and Public generic salutation. Nevertheless, it is worth noting that our phishing model proved and confirmed that even if some of the e-banking phishing website characteristics or layers are not very clear or not definite, the website can still be phishy especially when other phishing characteristics or layers are obvious. On the other hand even if some of the e-banking phishing website characteristics or layers are noticed or observed, it does not mean at all that the website is phishy, but it can be safe and secured especially when other phishing characteristics or layers are not detectable. Our first goal was to determine whether we could find any golden nuggets in the e-banking phishing website archive data using classification algorithms. In this, major rules discovered were inserted into the fuzzy rule engine to help giving exact phishing rate output. A

major issue in using data mining algorithms is the preparation of the feature sets to be used. Finding the "right" feature set is a difficult problem and requires some intuition regarding the goal of data mining exercise. We are not convinced that we have used the best feature sets and we think that there is more work to be done in this area. Moreover, there are number of emerging technologies that could greatly assist phishing classification that we have not considered. However, we believe that using features such as those presented here can significantly help with detecting this class of e-banking phishing websites.

REFERENCES

[1] WholeSecurity Web Caller-ID, www.wholesecurity.com

[2] Anti-Phishing Working Group. Phishing Activity Trends Report, http://antiphishing.org/reports/apwg_report_sep2007_final.pdf. September 2007.

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

[4] S. M. Bridges and R. B. Vaughn, "fuzzy data mining and genetic algorithms applied to intrusion detection," Department of Computer Science Mississippi State University, White Paper, 2001.

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

[6] FDIC., "Putting an End to Account-Hijacking Identity Theft," http://www.fdic.gov/consumers/consumer/idtheftstudy/identity_theft.pdf, 2004.

[7] A. Y. Fu, L. Wenyin and X. 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.

[8] A. Herzberg and A. Gbara, "Protecting Naive Web Users," Draft of July 18, 2004.

[9] C. Y. Ho, B. W. Ling and J. D. Reiss, "Fuzzy Impulsive Control of High-Order Interpolative Low-Pass Sigma–Delta Modulators," IEEE Transactions on Circuits and Systems—I: Regular Papers, Vol. 53, No. 10, October 2006.

[10] L. James**,** "Phishing Exposed," Tech Target Article sponsored by: Sunbelt software, searchexchange.com, 2006.

[11] M. Liu, D. Chen and C. Wu., "The continuity of Mamdani method," International Conference on Machine Learning and Cybernetics, Page(s): 1680 - 1682 vol.3, 2002.

[12] W. Liu, G. Huang, X. Liu, M. Zhang, and X. Deng, "Phishing Web Page Detection," Proc. Eighth Int'l Conf. Documents Analysis and Recognition, pp. 560-564, 2005.

[13] W. Liu, X. Deng, G. Huang and A. Y. Fu, "An Antiphishing Strategy Based on Visual Similarity Assessment," Published by the IEEE Computer Society 1089-7801/06 IEEE , INTERNET COMPUTING IEEE, 2006.

[14] Microsoft Corp, "Microsoft Phishing Filter: A New Approach to Building Trust in E-Commerce Content," White Paper, 2005.

[15] S. Olsen, "AOL tests caller ID for e-mail," CNET News.com, January 22, 2004.

[16] Y. Pan and X. Ding, "Anomaly BasedWeb Phishing Page Detection," Proceedings of the 22nd Annual Computer Security Applications Conference (ACSAC'06), Computer Society, 2006.

[17] J. C. Perez, "Yahoo airs antispam initiative," ComputerWeekly.com, December 8, 2003.

[18] S. Shah, "Measuring Operational Risks using Fuzzy Logic Modeling, "Article, Towers Perrin, JULY 2003.

[19] T. Sharif, "Phishing Filter in IE7," http://blogs.msdn.com/ie/archive/2005/09/09/463204.aspx, September 9, 2006.

[20] L. Wood, "Document Object Model Level 1 Specification," http://www.w3.org, 2005.

[21] M. Wu, R. C. Miller and S. L. Garfinkel , "*Do Security Toolbars Actually Prevent Phishing Attacks?,"* CHI April 2006.

[22] M. Wu, R. C. Miller and G. Little, "Web Wallet: Preventing Phishing Attacks by Revealing User Intentions," MIT Computer Science and Artificial Intelligence Lab, 2006.

[23] Anders Persson, "Exploring Phishing Attacks and Countermeasures", *Master Thesis in Computer Science,* Thesis No: MCS-2007:18, September 2007.

[24] Rachna Dhamija, J. D. Tygar and Marti Hearst, "Why Phishing Works", Harvard University, White Paper, 2006.

[25] "Putting an end to account-hijacking identity theft," FDIC, Tech. Rep., Dec. 2004. [Online]. Available: http://www.fdic.gov/consumers/consumer/idtheftstudy/identity theft.pdf.

[26] Ian Fette, Norman Sadeh and Anthony Tomasic, "Learning to Detect Phishing Emails", Institute for Software Research International, CMU-ISRI-06-112, June 2006.

[27] WEKA - University of Waikato, New Zealand, EN, 2006: "Weka - Data Mining with Open Source Machine Learning Software in Java";http://www.cs.waikato.ac.nz/ ml/weka (2006/01/31).

[28] Sebastian Misch, "Content Negotiatio in Internet Mail", Diploma Thesis, University of Applied Sciences Cologne, Mat.No.: 7042524, February 2006.

[29] Vic Ciesielski and Anand Lalani, "Data Mining of Web Access Logs From an Academic Web Site", Proceedings of the Third International Conference on Hybrid Intelligent Systems (HIS'03): Design and Application of Hybrid Intelligent Systems, Pages 1034-1043, December 2003, IOS Press.

[30] Lyman, Peter; Hal R. Varian (2003). "How Much Information". http://www.sims.berkeley.edu/how-much-info-2003. Retrieved on 2008-12-17.

[31] Kantardzic and Mehmed. "*Data Mining: Concepts, Models, Methods, and Algorithms*.", John Wiley & Sons. ISBN 0471228524. OCLC 50055336, 2003.

[32] U.M. Fayyad, "Mining Databases: Towards Algorithms for Knowledge Discovery," Data Eng. Bull., vol. 21, no. 1, pp. 39-48, 1998.

[33] 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.

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

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

[36] J. R. Quinlan, "Improved use of continuous attributes in c4.5", Journal of Artificial Intelligence Research, 4:77-90, 1996.

[37] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases", Proc International Conference on Very Large Databases, pp. 478-499. Santiage, Chile: Morgan Kaufmann, Los Altos, CA, 1994.

[38] T. Scheffer, "Finding Association Rules That Trade Support Optimally against Confidence", Proc of the 5th European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD'01), pp. 424-435, Germany: Springer-Verlag, (2001).

[39] http://www.phishtank.com/phish_archive.php

[40] E. Han and G. Karypis, "Centroid-Based Document Classification: Analysis and Experimental Results, Principles of Data Mining and Knowledge Discovery", p. 424–431, 2000.

[41] Fadi T.,Peter C. & Peng Y, "*MCAR: Multi-class Classification based on Association Rule*", IEEE International Conference on Computer Systems and Applications ,2005, pp. 127-133.

[42] F.Thabtah, P. Cowling and Y. Peng, "*A new multi-class,multi-label associative classification approach*".The 4th International Conference on Data Mining(ICDM'04), Brighton, UK, 2004.

[43] Microsoft, "http://www.microsoft.com/mscorp/twc/privacy/spam", April 2004.

[44] Netcraft, "http://toolbar.netcraft.com/", Dec 2004.