Intelligent Phishing Possibility Detector

Rajeev Kumar Shah
PhD Scholar, School of Management and Economics, UESTC, Chengdu, 611731 China

Md. Altab Hossin
PhD Scholar, School of Management and Economics, UESTC, Chengdu, 611731 China

Asif Khan, PhD
UESTC, GICIT_CAS, China

ABSTRACT

Phishing techniques have not only grown in number, but also in sophistication. Phishers might have a lot of approaches and tactics to conduct a well-designed phishing attack. The on-line banking consumers and payment service providers, those are the main targets of the phishing attacks, are facing substantial financial loss and lack of trust in Internet-based services. In order to overcome these, there is an urgent need to find solutions to combat phishing attacks. Detecting the phishing website is a complex task which requires significant expert knowledge and experience. So far, various solutions have been proposed and developed to address these problems. Most of these approaches are not able to make a decision dynamically on whether the site is in fact phished, giving rise to a large number of false responses. This is mainly due to limitation of the previously proposed approaches, for example depending only on fixed black and white listing database, missing of human intelligence and experts, poor scalability and their timeliness. In this research the application of an intelligent fuzzy-based classification system for e-banking phishing website detection is investigated and developed. The main aim of the proposed system is to provide protection to the users from phisher's deception tricks, giving them the ability to detect the legitimacy of the websites. The proposed intelligent phishing detection system employed Fuzzy Logic (FL) model with association classification mining algorithms. The approach combined the capabilities of fuzzy reasoning in measuring imprecise and dynamic phishing features, with the capability to classify the phishing fuzzy rules.

Keywords

Phishing; e-banking; fuzzy logic; association; classification; machine- learning; internet security; data mining.

1. INTRODUCTION

Phishing websites is a kind of attack in which clients are trapped by entering his/her personal information like user names, passwords, bank account numbers, pins, social security numbers, mother's maiden names (or other secondary password), etc. The word phishing comes from the expression "website phishing" and it is similar to the word "fishing". It is akin to grabbing the phishers' bait that is thrown out hoping that a user will bite similar to a fish. As the number of Internet users and online transactions grow, the possibility of misuse also rises. The total number or phishing attacks launched in 2012 was 59% higher than 2011. It appears that phishing has been able to set yet another record year in attack volumes, with global losses from phishing estimated at $1.5 billion in 2012 [2]. To examine this threat, Let's take for example a typical phishing attack, which may be based on several techniques; including exploiting browser vulnerabilities or performing man-in-the middle attacks using a proxy. However, the most straightforward and widespread method consists of deploying a web page that looks and behaves like the one the user is familiar with. A number of software vendors and companies have brought out several different types of anti-phishing toolbars. Example, eBay offers a free toolbar that can positively identify eBay-owned sites, and Google offers a free toolbar aimed at identifying counterfeit sites' [3, 4]. That is why it was decided that there is a strong need for improved automated detection algorithms. Several sets of research has been conducted in the area of phishing detection, such as CANTINA, SpoofGuard, and Netcraft none to date have utilized the wonderful wealth of information found in social networks [6].

The purpose of this study is to study the problem of internet users' falling prey to fraudulent websites by entering important information such as account numbers', usernames, passwords, Personal Identification Numbers (PIN) and social security numbers, etc.

2. RELATED WORK

To understand phishing and find new methods to protect users from sophisticated attacks it is important to look at how and why people are susceptible to phishing attacks and how attackers design their attacks to make them 'appealing' for the intended victims'. Human Computer Interaction or more specific the field of Usable Security addresses security questions from a user perspective. Also phishers' setup the fake web sites. Generally speaking, past work regarding anti-phishing falls into different categories, such as studies to know why people fall for phishing attacks, educating people about phishing attacks, anti-phishing user interface, automated detection of phishing, setting up fake web sites, domain age, forms, owner information, page content information, community information, different algorithms have use to detect phishing detection, password hashing, CANTINA and spoofguard. Here all are described in details.

Anti-phishing education has brought attention to online instruction, testing, and situated learning. Online training materials have been published by government organizations [11, 12], non-profit organizations (NPO's) and businesses [13, 14]. These materials explain what phishing is and provide tips to prevent users from falling prey to phishing attacks. Testing is used to demonstrate how susceptible people are to phishing attacks and educate them on how to avoid it. For example, Mail Frontier [15] has a web site containing screenshots of potential phishing emails. Users are scored based on how well they can identify which emails are legitimate and which are not. A third approach uses situated learning, where users are sent phishing emails to test users' vulnerability of falling for attacks. At the end of the study, users are given materials that inform them about phishing attacks. This approach has been used in studies conducted by Indiana University in training students [16], West Point in instructing cadets [17, 18] and a New York State Office in educating employees [19]. The New York study showed an improvement in the participants' behavior in identifying phishing over those who were merely given a pamphlet containing the information on how to combat phishing. In previous work, an email-based approach was developed to train people how to identify and avoid phishing attacks, demonstrating that the existing practice of sending security notices is ineffective, while a story-based approach using a comic strip format was surprisingly effective in teaching people about phishing [20].

3. PROPOSED METHOD

3.1 The Phishing Website Detection Design Methodology

The technique of the model involves the fuzzification of input variables that is based on the 27 phishing website characteristics and factors (previously extracted from the implemented phishing website case-studies experiments, anti-phishing tools and surveys) which are mentioned and analyzed, rule evaluation, aggregation of the rule outputs, and defuzzification technique as shown in Figure 1.

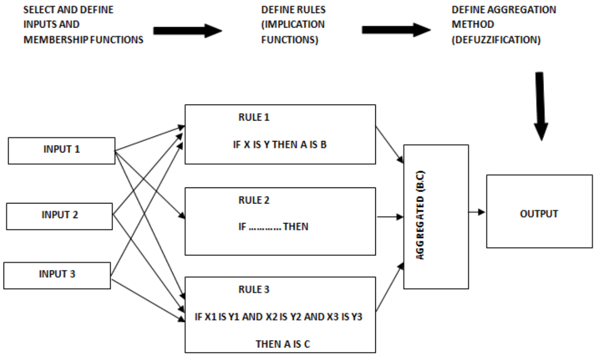

SELECT AND DEFINE INPUTS AND MEMBERSHIP FUNCTIONS
DEFINE RULES (IMPLICATION FUNCTIONS)
DEFINE AGGREGATION METHOD (DEFUZZIFICATION)
INPUT 1
INPUT 2
INPUT 3
RULE 1
IF X IS Y THEN A IS B
RULE 2
IF ………… THEN
RULE 3
IF X1 IS Y1 AND X2 IS Y2 AND X3 IS Y3
THEN A IS C
AGGREGATED (B.C)
OUTPUT

Figure 1: Design Layout

The model consists of four phases. Each phase will be explained in more detail to fully understand its function and output, and how it is connected with the other phases, to produce the final desired output.

3.2 Fuzzification

This is the process of generating membership values for a fuzzy variable using membership functions. The first step is to take the crisp inputs from the 27 characteristics and factors which stamp the forged phishing website and determine the degree to which these inputs belong to each appropriate fuzzy set. This crisp input is always a numeric value limited to the universe of discourse. Once the crisp inputs are obtained, they are fuzzified against the appropriate linguistic fuzzy sets. The fuzzy detection model provides more thorough definitions for each factor and its interactions with other factors. This approach will provide a decision tool for identifying phishing websites. The essential advantage offered by fuzzy logic techniques is the use of linguistic variables to represent key phishing characteristic indicators and the relation of phishing website probability. In this step, linguistic descriptors such as High, Low, and Medium are assigned to a range of values for each key phishing characteristic indicator. Since these descriptors will form the basis for capturing expert inputs based on the impact of Key Phishing Characteristic Indicators on the Phishing Website, it is important to calibrate them to how they are commonly interpreted by the experts providing input. An example of the linguistic descriptors used to represent one of the key phishing characteristic indicators (URL Address Length) and a plot of the fuzzy membership functions are shown in Figure 2 below. The x-axis in each plot represents the range of possible values for the corresponding key phishing characteristic indicators (Low, Moderate and High). The y-axis represents the degree to which a value for the key phishing characteristic indicators is represented by the linguistic descriptor. For example, in the plot of the membership function for URL Address Length, 4.5 cm is considered 'Low' with a membership of 30% and is also considered 'Moderate' with a membership of 65%. The fact that 4.5 cm URL Address Length is considered both Low and Moderate to varying degrees is a distinguishing feature of fuzzy logic, as opposed to binary logic which artificially imposes black-and white constraints. The fuzzy representation more closely matches human cognition, thereby facilitating expert input and more reliably representing experts' understanding of underlying dynamics (Bridges and Vaughn, 2001).

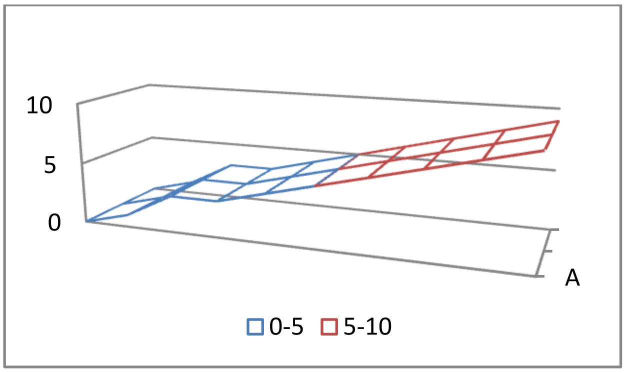


Figure 2: Input variable for URL Address Length component

URL Address Length – Low, Moderate, High.

Linguistic Variable: URL Address Length

Linguistic value Numerical Range

Low [0, 0, 3, 5]

Moderate [3, 5, 7]

High [5, 7, 10, 10]

Another example of the linguistic descriptors used to represent key phishing characteristic indicators is the Pop-Up Windows feature. If the website has two hyperlinks with pop-up windows asking for user credentials, then it is considered 'Low' with a membership of 50 % and is also considered 'Moderate' with a membership of 50%.

Pop-Up Windows – Low, Moderate, High.

Linguistic Variable: Pop-Up Windows

Linguistic value Numerical Range

Low [0, 0, 1, 3]

Moderate [1, 3, 5]

High [3, 5, 10, 10]

The ranges for this fuzzy variable were specified depending on the high risks that accompany this particular phishing feature. We cannot allow for too many pop-up windows asking for vital information that can be used for phishing purposes. That's why it is decided to put a very small fuzzy set range for fuzzy values "Low" and "Moderate" to mitigate these kinds of phishing risks. The same approach is used to calibrate all the other key phishing website characteristic indicators. The ranges of their fuzzy variables are derived and tuned from a series of phishing experiments with case-studies, surveys and expert knowledge.

3.3 Fuzzy Rule Evaluation

The fuzzy rule has multiple antecedents, the fuzzy operator (AND or OR) is used to obtain a single number that represents the result of the antecedent evaluation. The AND fuzzy operation is applied (intersection) to evaluate the conjunction of the rule antecedents. Having specified the risk associated with the phishing website and its key phishing characteristic indicators, the next logical step is to specify how the phishing website probability varies as a function of the Key Phishing Characteristic Indicators. Experts provide fuzzy rules in the form of if…then statements that relate phishing website probability to various levels of key phishing characteristic indicators based on their knowledge and experience. Phishing website experiments, anti-phishing tool analysis, web surveys, and detailed a phishing questionnaires were used to find and evaluate all factors and features of phishing websites, with all their relationships and associations with one another. This helped us greatly as experts in creating the phishing website fuzzy rules. The output is the phishing website risk rate and is defined in fuzzy sets like 'phishy' to 'legitimate'. The fuzzy output set is then defuzzified to arrive at a scalar value as shown in Figure 3.

Linguistic Variable: Phishing Website Risk Rate

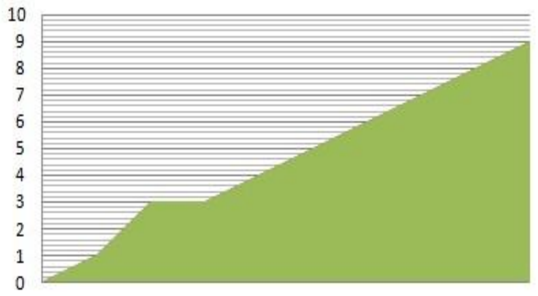


Figure 3: Output variable for phishing website rate

Linguistic value Numerical Range

Legitimate [0, 0, 30, 50]

Suspicious [30, 50, 70]

Phishy [50, 70, 100]

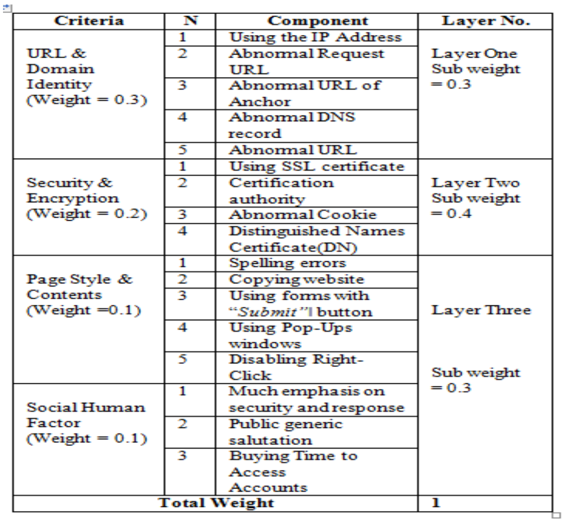
4. FUZZY LOGIC PHISHING DETECTION MODEL

In this phishing fuzzy model, the authors categorize the 27 phishing website characteristics and factors into six different criteria based on their attack type and source. After that the characteristic features were ranked and weighted in each criteria based on their importance, influence, effectiveness and complexity before considering those in the fuzzy learning process. The grouping process was undertaken to simplify the fuzzy model since dealing with the 27 website phishing features as a whole can make the fuzzy rule evaluation very complicated and time-consuming. These 27 phishing website features and factors were grouped and categorized into six criteria (URL & Domain Identity, Security & Encryption, Source Code & Java script, Page Style & Contents, Web Address Bar and Social Human Factor). Each criterion has its own fitted phishing feature criteria. A layering process was also implemented in these phishing website features to enhance and improve the final phishing website risk rate fuzzy output. Table 1 represents detailed information on grouping the phishing website features into specific criteria and their association-related layers based on the types of phishing source and nature. The weights assigned to those are according to their effectiveness and influence. The architecture of the fuzzy logic inference-based phishing website risk rate detection model can be shown from the structure figure, the final output website phishing result for this fuzzy model depends on evaluating the fuzzy outputs of the three layers and then combining those for the final result.

5. FUZZY RULE BASE

All fuzzy rules implemented in our proposed detection mode were derived based on our own phishing background experience and expert knowledge supported by a series of experimental phishing scenarios with case-studies. The following text will show all fuzzy rules for all phishing website criteria and layers.

Table 1. Components and layers of 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 Sub weight = 0.4
	2	Certification authority	
	3	Abnormal Cookie	
	4	Distinguished Names Certificate(DN)	
Page Style & Contents (Weight =0.1)	1	Spelling errors	Layer Three Sub weight = 0.3
	2	Copying website	
	3	Using forms with "*Submit*"‖ button	
	4	Using Pop-Ups windows	
	5	Disabling Right-Click	
Social Human Factor (Weight = 0.1)	1	Much emphasis on security and response	
	2	Public generic salutation	
	3	Buying Time to Access Accounts	
Total Weight			1

5.1 The Rule Base

The rule base has five input parameters and one output. The rule contains all the "IFTHEN" rules of the system. For each entry of the rule base, each component is assumed to be one of the three values and each criterion has five components. Therefore, the rule base 1-1 contains (35) = 243 entries. The output of rule base 1-1 is one of the phishing website risk 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 1-1 for layer 1 are shown in Table 2. The system structure for URL & Domain Identity criteria is the joining of its five components (Using the IP Address, Abnormal Request URL, Abnormal URL of Anchor, Abnormal DNS record and Abnormal URL), which produces the URL & Domain

Identity criteria (Layer one) as shown in Figure 4. Further, the three-dimensional plots of this system structure are shown in Figure 5.using MATLAB.

Table 2. Sample of rule base1-1 entries for URL & Domain Identity criteria

Rule #	(comp. 1) Using the IP Address	(comp. 2) Abnormal Request URL	(comp. 3) Abnormal URL Anchor	(comp. 4) Abnormal DNS record	(comp. 5) Abnormal URL	URL & Domain Identity Criteria Phishing Risk (Layer one)
1	Low	Low	Low	Low	Low	Genuine
2	Low	Low	Low	Low	Moderate	Genuine
3	Low	Low	Low	Moderate	Moderate	Doubtful
4	Low	Low	Low	Moderate	High	Doubtful
5	Low	Low	Moderate	Moderate	High	Fraud
6	Low	Moderate	Moderate	Low	High	Fraud
7	Moderate	Low	High	Moderate	High	Fraud
8	High	Moderate	Low	Low	Low	Doubtful
9	Low	High	Low	Low	Moderate	Doubtful
10	High	Moderate	High	High	Low	Fraud

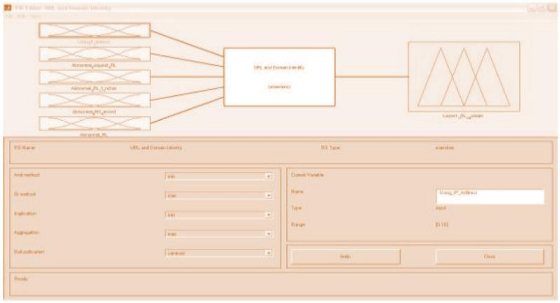


Figure 4: System structure for URL & Domain Identity criteria

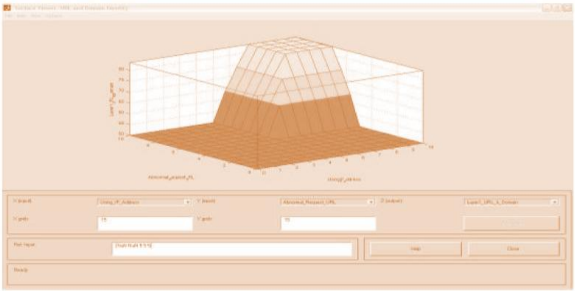


Figure 5: Three-dimensional plot for URL & Domain Identity criteria

6. TESTING AND VALIDATION

While there is no mature technology that defends against phishing web sites yet, there is currently no anti-phishing benchmark set of expectation or standardized set of data for phishing detection products evaluation. Most of the claims made by vendors of available products are based on proprietary test data and testing methodology. In this research, a test framework has been constructed which can evaluate a generic antiphishing technology against the latest existing phishing sites. This framework has been used to evaluate the effectiveness of the intelligent plug-ins phishing detection toolbar. The PhishTank data was selected as the public benchmark for comparing the phishing detection. Details of this experimentation framework and findings are presented below. Using testing sample of 120 different e-banking website that was used previously on our fuzzy logic phishing website detection model, our intelligent web-based plug-ins toolbar were further tested to prove its validation and high phishing detection precision. The dataset sample was taken from the public benchmark Phishtank archive data (Phishtank, 2008), consisting of 60 phishing websites: 35 suspicious websites and 25 legitimate websites. The proposed toolbar managed to detect the phishing e-banking websites that were found in the testing sample with a very small miss-classification rate. The results indicate clearly the high precision of phishing classification with very small false positive and false negative rates, as specified in the confusion matrix shown in Table 3.

Table 3. Results of website legitimacy decision using the intelligent fuzzy-based classification detection model

Decision Website Legitimacy	Legitimate	Suspicious	Physhi
Legitimate Website	22	2	2
Suspicious Website	1	32	2
Phishing Website	1	3	56

As shown in Table 3, there were just 3 legitimate websites miss-classified as suspicious or phishy websites, and only 4 phishing websites were miss-classified as legitimate or suspicious website. These results demonstrate very clearly how effective and reliable detecting phishing website can be when applying an intelligent heuristic search using association classification mining algorithms combined with a fuzzy logic model approach. The obvious enhancement that happened to the final results can be justified by using an approach not only depending on the human expert knowledge alone, but also on integrating and combining an intelligent supervised machine learning approach, using specific mining associative classification algorithms. When comparing the new intelligent web browser plug-ins toolbar with other famous anti-phishing toolbars like Netcraft (Netcraft, 2006) and Spoofstick (Spoofstick, 2005) toolbars, it was found that our toolbar outperformed the other detection toolbars regarding the accuracy, efficiency and the speed of classifying and detecting phishing websites. It managed to classify correctly approximately 92% of all tested websites, beating all other anti-phishing toolbars, which depend mainly on using black-list and white-list databases in classifying phishing websites. Figure 6, shows the comparative performance of all tested anti-phishing toolbars for the accuracy phishing classification rate.

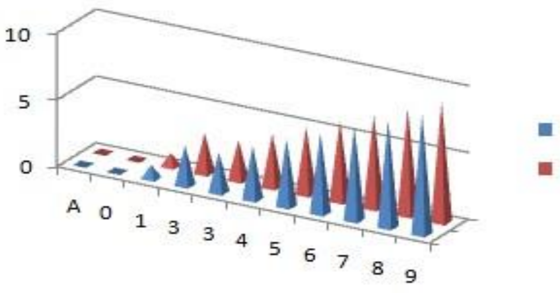


Figure 6: Phishing classification precision comparing chart

It is noted that the proposed tool offered best performance among the tested tools, being about 11% better compared to Netcraft and 6% better compared to Spoofstick. The authors are of the view that this solution is better since it uses a novel AI heuristic search on all phishing features that can be found on the websites, grouping them into specific criteria and layers depending on their type, and then by using specific fuzzy-based classification rules, the final phishing detection rate appears.

7. CONCLUSIONS

An AI-based hybrid system has been proposed for phishing website detection systems. Fuzzy logic has been combined with association classification data mining algorithms to provide efficient techniques for building intelligent models to detect phishing websites. Empirical phishing experimental case studies have been implemented to gather and analyze range of different phishing website features and patterns, with all its relations. The experimental case-studies point to the need for extensive educational campaigns about phishing and other security threats. People can become less vulnerable with a heightened awareness of the dangers of phishing and our experimental case-studies also suggest that a new approach is needed to design a usable model for detecting e banking phishing websites, taking into consideration the user's knowledge, understanding, awareness and consideration of the phishing pointers located outside the user’s centre of interest.

8. REFERENCES

[1] Intelligent phishing detection system for e-banking using fuzzy data mining by Maher Aburrous, M.A. Hossain, Keshav Dahal, Fadi Thabtah in 2010.

[2] Fette Ian, Sadeh Norman, & Tomasic Anthony (2006). Learning to detect phishing emails. Institute for Software Research International. Behavioral response to phishing risk by Julie S. Downs, Mandy Holbrook and Lorrie Faith Cranor.

[3] Ahmed Abbasi, Fatemeh “Mariam” Zahedi, Yan Chen “Impact of Anti-Phishing Tool Performance on Attack Success Rates”.

[4] Weiwei Zhuang, Qingshan Jiang, Tengke Xiong,“An Intelligent Anti-phishing Strategy Model for Phishing Website Detection”.

[5] Mallikka Rajalingam, Saleh Ali Alomari, Putra Sumari “Prevention of Phishing Attacks Based on Discriminative

[6] Key Point Features of WebPages”, International Journal of Computer Science and Security (IJCSS), Vol. 6, 2012.

[7] M. J. et al., “What instills trust? a qualitative study of phishing,” inProceeding of first Int’l Workshop on Usable Security, Springer-Verlag,2007.

[8] ASIF KHAN, and JIAN-PING LI, "Vision Based Geo Navigation Information Retrieval”, (IJACSA), Vol. 7, No. 1, 2016

[9] P. Kumaraguru, Y. Rhee, A. Acquisti, L. F. Cranor, J. Hong, andE. Nunge, “Protecting people from phishing: The design and evaluationof an embedded training email system,” inCHI2007: Proceedings of the Conference on Human Factors in Computing Systems, 2007.