International Journal of Advanced Science and Technology
Vol.78 (2015), pp.29-40
http://dx.doi.org/10.14257/ijast.2015.78.03

System Design, Investigation and Countermeasure of Phishing Attacks using Data Mining Classification Methods and its Analysis

Rajendra Gupta[1] and Piyush Kumar Shukla[2]

[1]Assistant Professor, BSSS Autonomous College,
Barkatullahu University, Bhopal, India
[2]Assistant Professor, University Institute of Technology
Rajiv Gandhi Technical University, Bhopal, India
[1]rajendragupta1@yahoo.com, [2]pphdwss@gmail.com

Abstract

The phishing is a kind of e-commerce lure which is intended to steal the confidential information of the internet user by making identical website of legitimate one in which the contents and images most likely remains similar to the legitimate website. The other way of phishing website is to do minor changes in the URL or in the domain of the website. In this paper, an anti-phishing system is proposed which is based on the development of the Add-on tool for the web browser. The performance of the proposed system is studied with four different data mining classification algorithms which are Class Imbalance Problem (CIP), Rule based Classifier (Sequential Covering Algorithm (SCA)), Nearest Neighbour Classification (NNC), Bayesian Classifier (BC). To evaluate the performance of the proposed anti-phishing system for the detection of phishing websites, we have collected 7690 legitimate websites and 2280 phishing websites from the authorised sources like APWG database and PhishTank. After the analysis of the anti-phishing system, more than 90 percentage successful results achieved at different time periods.

Keywords: *Phishing, Anti-Phishing, Add-on for Web Browser, Data mining classification algorithms*

1. Introduction

A number of governmental and private authorised agencies are working on the topic of phishing and the countermeasure of phishing attack. The APWG (Advanced Phishing Working Group) and PhishTank are the most prominent agencies which keeps all the information related to phishing and legitimate websites. According to the information received from the records of APWG, the number of phishing attack reported around 21,480 from phishing websites in March 2012. During the second quarter of 2013, a total of 639 unique brands were targeted by phishing attacks [6].

In the report, even there is no economy loss mentioned but we can think if thousands website are declaring phishing in a month worldwide, how much loss could be possible. Based on the report given by Javelin Strategy and Research on April 2012, the economy loss reached to 21 billion [19]. Nevertheless, the phishing is seriously challenging and collapses the trust to electronic commerce and e-services security system. By watching the effect of less security in online transaction, many persons are stopping e-transactions facility. The peoples use convenient online services, since they are not sure whether their credentials are in danger or not. So to keep it in mind, the questions arises that how to identify the fraud, how to design and build a reliable and secure system environment for electronic business transactions. So the research study is very much important to reduce the online transaction problems.

ISSN: 2005-4238 IJAST

To solve the problem of phishing, the researchers are finding the solution at client side and server site system. So far, slow progress has been seen in the client and server side design analysis. On the client side application, there have been around 110 types of user-centred applications developed. These applications uses web browser toolbar and additional plug-in that can be installed additionally with the web browser. It is found that the server site strong system designing is most important to protect the user from phishing attack. During the study, it is seen that server side applications are not giving successful result, but the concept of server side security is proposed and applications are working at client site applications [9, 12].

2. Methods of Phishing Attack

The attacker can attack on any website in different ways. Some of them are as follows [2]:

- ***Link manipulation:*** This type of phishing is possible by making some changes in the link provided by the spoofed page. A number of phishing attacks use technical deception process which is designed to make a link in an e-mail that appears to the spoofed organization link. It is possible by doing misspell the URLs or by the use of sub-domains to target the web user. For example, in the URL http://www.mybank.services.com/, it appears that the URL is asking to login the 'mybank.services' part of the website, which is actually a phishing URL of the legitimate site.
- ***Website forgery:*** An phishing attack can use flaws in a trusted website's scripts tags against the web user. This type of phishing attack which is also known as cross-site scripting is very problematic, because they redirect the user to sign in at bank or services column of web page. In that page everything from the web address to the security certificates appears original and legitimate.
- ***Filter evasion:*** Images can also be used for the phishing attack. By the use of image in place of text, it is very difficult to trace the phishing webpage. The filter evasion technique uses this methodology while making the phishing webpage. This type of phishing web page takes less time to prepare the spoofing websites, and uses less number of coding tags on the webpage.
- ***Mobile phishing:*** Since the mobile users are increasing rapidly and the internet access from mobile is also increasing, so the phishing attacks are targeting the mobile user to steal the confidential information. In the mobile phishing, the messages looks link coming from the mobile that claimed to be from a bank which told users to dial a number regarding the problems with their bank account.
- ***Tabnabbing :*** Tabnabbing is one another attacking method, which directs the user to submit their login information and password to popular websites by impersonating those sites. These sites convince the user that the site is genuine.
- ***DNS-Based Phishing ("Pharming"):*** Pharming is the term given to hosts file modification. This type of phishing is also called DNS-based phishing. In this phishing, the hackers tamper with a company's host files or the DNS so that requests for URLs or name services return a bogus address and subsequent communications are directed to a fraudulent site. The targeted user do not sure that the website where they are entering their confidential information is controlled by hackers and is probably not even in the same country as the legitimate website [1].

3. Overview of Previous Study on Phishing

On the basis of the above mentioned phishing methods, several anti-phishing techniques have been proposed by the researchers. Naga Venkata Sunil A. *et. al.*, [3] have proposed A PageRank Based Detection Technique for Phishing Web Sites, in which phishing web sites are detected using Google's PageRank method. He has collected a dataset of 100 phishing sites and 100 legitimate sites. According to Venkata Sunil, around 98 percentage websites are correctly classified by using this Google PageRank technique and it shows only 0.02 false positive rate and 0.02 false negative rate. Khonji M. *et. al.*, [18] has proposed a Phishing Classification Based system on URL Features. This approach is quite successful but this heuristic classification system might not be efficient on HTTP clients due to the delay with HTTP search queries, and therefore he has suggested implementing this system on a mail server where email content is checked passively without imposing a delay on client applications. Wardman B. *et. al.*, [30] presented a High-Performance Content-Based Phishing Attack Detection, in which a cadre of file matching learning algorithm is implemented which is based on the websites content to detect phishing. This is possible by employing a custom data set which is consisting of 17,992 phishing attacks targeting 159 different company brands. The results shown by Wardman for their experiments using a variety of different content-based approaches demonstrate that some can be achieved a detection rate of more than 90% while maintaining a low false positive rate.

Weider D.Yu *et. al.*, [33] presented an Phishing Detection Tool - PhishCatch in which the novel anti-phishing algorithm is developed to protect the user from phishing attack. This algorithm is based on the heuristic which can detect phishing e-mails and alert the users about the phishing type e-mails. The phishing filters used in the algorithm and rules are formulated after extensive research of phishing methodologies and tactics as presented in the research paper. After testing the algorithm, he has determined that this algorithm has a catch rate of 80% which gives an accuracy of 99%. Prakash P. *et. al.*, [28] presented a heuristics "PhishNet" in which five heuristics has been taken to specify combinations of known phishing sites to discover new phishing URLs. In the evaluation result with real-time blacklist feeds, he discovered near about 18,000 new suspicious URLs from a set of 6,000 new blacklisted entries. He has shown that approximate matching algorithm shows a very few false positives (3%) and negatives (5%) result. Isredza Rahmi A Hamid *et. al.*, [17] suggested an Profiling Phishing E-mail Based on Clustering Approach. In this approach, profiling e-mail - born phishing activities is proposed. These activities are useful in determining the activity of a particular person or a particular group of phishing attacker. By generating profiles of the user, phishing activities can be understood and observed very well. The proposed profiling email-born phishing algorithms (ProEP) express promising results with the RatioSize rules for selecting the optimal number of clusters. Zhang H. *et. al.*, [8] has presented a framework which is based on the Bayesian approach for content-based phishing web page detection. The effectiveness of the system is examined by taking a large-scale dataset, collected from real phishing cases of trusted sources. The experimental results of Zhang demonstrated that the text and image classifier which is designed to deliver promising results, the fusion algorithm outperforms the individual classifiers. His model can be adapted for the further study on phishing. Shuhua Wu *et. al.*, [34] said that if a protocol contains only one authentication factor, the password can be recovered through phishing attack.

Li T. *et. al.*, [22] has proposed an offline phishing detection system named *L*arge-scale *A*nti-phishing system by *R*etrospective data-e*X*ploration (LARX). This LARX checks the network traffic data archived at a vantage point. It analyzes the data for phishing countermeasure. The proposed phishing filter technique in the system uses cloud computing. Since the system is offline, it can be effective for the analysis of large volume of traced data when adequate computing power and storage capacity used. Huang H. *et. al.*, [23] has explained a detailed overview of a deceptive phishing attack and its all possible countermeasure techniques. In this study, the technologies used by phishing attacker with the definitions, classification and future works of deceptive phishing attacks have been discussed. Edward Ferguson *et. al.*, [15] presented "Cloud Based Content Fetching by using Cloud Infrastructure to Obfuscate the Phishing Scam Analysis", in which the system presents different personas and user behaviour to the phishing sites by using different IP addresses and different browsing configurations. By running a 10-day probe experiment against real phishing site, he showed the effectiveness of this approach in preventing detection and blocking of anti-phishing probes by the phishing site operators. The paper is based on the emerging phishing techniques [13, 26].

Mahmood Ali M. *et.al.*, [24] presented a paper on 'Deceptive Phishing Detection System (From Audio and Text messages in Instant Messengers using Data Mining Approach)' in which, words are recognized from speech with the help of FFT spectrum analysis and LPC coefficients methodologies.

4. Anti-Phishing Toolbars

There is a number of anti-phishing toolbar approaches proposed in the early study that can be used to identify a web page whether it is phishing or not. We have taken observation to get a basic understanding of how each tool function. The earlier tools are trying to protect the user's confidential information but it is seen that these tools are not completely successful. The legitimate sites are defined as white lists which are known as safe sites and the fraudulent sites are defined as blacklists. The description of various anti-phishing tools are as follows [31]:

CallingID focuses on the site ownership details and real-time rating and confirms the user that the site is safe or not to provide information. The CallingID Toolbar checks 54 different verification tests to determine the legitimacy of a given site. Different visual indicators are given in the CallingID toolbar to check the type of website. These indicators shows different colours for differentiating the web page, like green colour represent a known-good site; yellow colour represent a site that is 'at low risk'; red colour represent a site that is 'at high risk' and therefore may be a phishing site. Some of the heuristics are included like examining the site's country of origin, length of registration, user reports, popularity of the website and the blacklisted data [32].

The Cloudmark Anti-Fraud toolbar is based on the user ratings [10]. When user visits the website, he has the right to report the site as the site should be accessible or not. On the basis of this feature, the toolbar display a coloured icon for each site visited by the user. The users themselves are rated according to their record of correctly identifying phishing sites. Each site's rating is computed by aggregating all ratings given for that site, with each user's rating of a site weighted according to that user's reputation.

The EarthLink Toolbar appears to rely on a combination of heuristics, user ratings and manual verification [13]. The toolbar allows users to report suspected phishing sites to EarthLink. These sites are then verified and added to a blacklist. The toolbar also appears to examine domain registration information such as the owner, age and country [13].

The eBay Tool uses a combination of heuristics and blacklists [14]. The Account Guard indicator has three modes: green, red, and grey. The icon is displayed with a green background when the user visits a site known to be operated by eBay (or PayPal), red background when the site is a known phishing site and grey background when the site is not operated by eBay and not known to be a phishing site. Known phishing sites are blocked and a pop-up appears, giving users the option to override the block. The toolbar also gives users the ability to report phishing sites.

Firefox includes a new feature designed to identify fraudulent web sites. Originally, this functionality was an optional extension for Firefox as part of the Google Safe Browsing Toolbar. URLs are checked against a blacklist, which Firefox downloads periodically [21]. The feature displays a popup if it suspects the visited site to be fraudulent and provides users with a choice of leaving the site or ignoring the warning. Optionally, the feature can send every URL to Google to determine the likelihood of it being a scam. According to the Google toolbar downloading site, the toolbar combines "advanced algorithms with reports about misleading pages from a number of sources" [16].

The Netcraft Anti-Phishing Toolbar uses several methods to determine the legitimacy of a web site [38]. The Netcraft web site explains that the toolbar traps the suspicious URLs which contains the characters that have no common purpose other than to deceive the user; enforces display of browser navigation controls (tool and address bar) in all the windows, to defend against pop-up windows that can be hide the navigational controls and the option ‘clearly displays sites’ which shows the hosting location, including country that help to evaluate fraudulent URLs.

The Netscape Navigator 8.1 web browser includes a built in phishing filter [27]. For the testing of this tool as well as the third party reviews, it appears that this functionality relies solely on a blacklist, which is maintained by AOL and updated frequently [11]. When a suspected phishing site is encountered, the user is redirected to a built-in warning page. Users are shown the original URL and are asked whether or not they would like to proceed.

SpoofGuard is a web browser based anti-phishing filter tool which helps to prevent a form of malicious attack called "web spoofing" [29]. Phishing attacks generally involve deceptive e-mail that looks like coming from a legitimate commercial website. The e-mail explains to the recipient that the user is having account problem, or some other account related serious problem to visit the commercial site and log in. However, the given link in the e-mail sends the user to a malicious "spoofing" web site that collects user’s confidential information such as account names, password and credit card numbers etc. Once the user information is collected by a spoofing web site, criminals may log into user’s account or cause other serious damage.

5. Disadvantages of the Existing Systems

On the basis of the previous study, we have seen some of the disadvantages which are needed to be solved for the future system. The earlier popular technologies have several drawbacks [8] which are found in the research, these are:

a. Applying Blacklist-based technique with low false alarm probability which cannot detect the websites that are not in the blacklist database.

b. Since the life cycle of phishing websites is too short, instant response in not coming in front of the user.

c. The accuracy of blacklist is not too high.

d. Heuristic-based anti-phishing technique is not successful which are having high probability of false and failed alarm.

6. System design and Development Model

The proposed system is based on the Add-on development in which the user can activate the Add-on of anti-phishing on his web browser. The concept behind the designing of the Anti-phishing tool is that when internet user hit the URL, a dialog box appear on the screen to inform the user about the type of the website whether it is phishing or not.

In the previous study, researchers suggested a number of anti-phishing models to find the solution [20, 33, 7]. The earlier proposed models are not giving more than 90 percentage successful result [4-5]. In some cases the tools gives only 40-50 percentage successful result. Since the techniques and tools are upgrading and changes are going on in website designing, the web site developers try to utilize advanced techniques to make the phishing websites. In this case, the existing tools are not finding the proper result. So a system should be developed that can manage and support the advanced tools of web development so that the better result could be achieved. A. Martin *et. al.*, [25] worked on the 27 phishing criteria using the concept of Neural Network. The same criteria have been taken by other researchers to find the solution of phishing attack. The functioning of the proposed system model for add-on based anti-phishing tool is as given below (Figure 1).

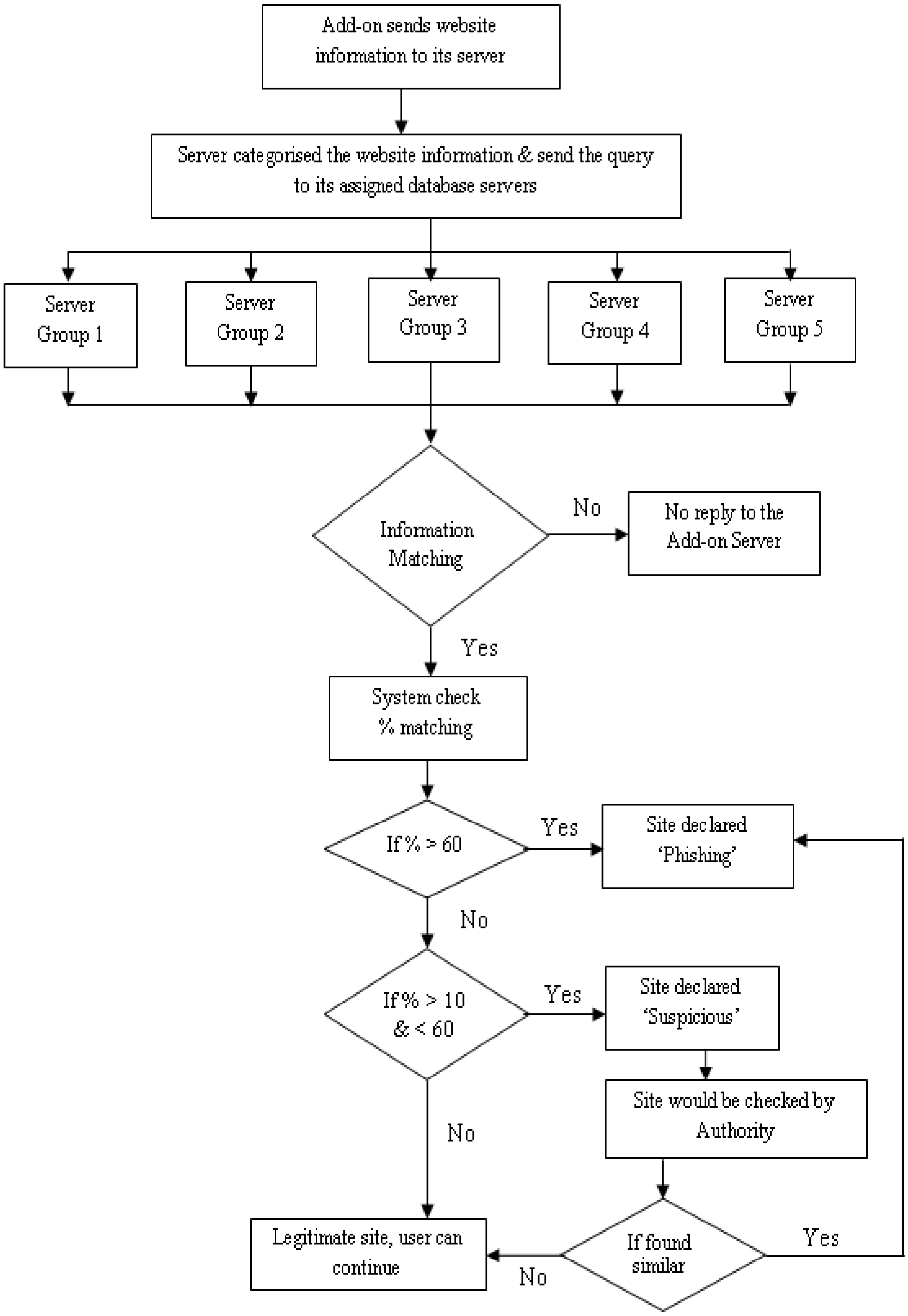
Add-on sends website information to its server
Server categorised the website information & send the query to its assigned database servers
Server Group 1
Server Group 2
Server Group 3
Server Group 4
Server Group 5
Information Matching
No
No reply to the Add-on Server
Yes
System check % matching
If % > 60
Yes
Site declared 'Phishing'
No
If % > 10 & < 60
Yes
Site declared 'Suspicious'
Site would be checked by Authority
No
Legitimate site, user can continue
No
If found similar
Yes

Figure 1. Functioning of the Proposed System

The reason of categorising the system in different groups is to find the spoofed website as quickly as possible when user hit the target URL. The functioning of the add-on is defined on the main server. The main server sends the instructions to 5 different assigned servers. The assigned servers are defined in five different categories; these are *Character based, Coding based, Identity based, Contents based and Attribute based*. When a user hit the website, instantly the concerned information about the webpage goes to these servers by main server. On the assigned servers, the information is cross checked with the database information. If it is found that the webpage information matches with the database information, the server replies to the user system about the matching status.

7. Working Procedure of Novel Anti-Phishing System

In the proposed system model, the expected criteria are divided in five different groups to speed up the response system. The system works with the following steps:

a. When user hits the target website, the system activates and sends the URL information to its database source. The database is categorised at five different servers.
b. If the matches are found in any of the server, the server get reply the matching information instantly to the Add-on system. An algorithm is defined at Add-on system.
c. When the result comes from all the servers, the system takes the decision and sends the reply on the web browser’s window with the confirmation message whether the website is phishing or legitimate.
d. To test the system, we have collected a number of legitimate website information and previously declared phishing website information from various sources. To develop and analyse the proposed system, we have used data mining techniques that is an integration of the collection of machine learning algorithms, to function the model with the use of five different classification algorithms.
e. If the user accessed website found to be phishing, the Add-on store the website information to its database so that no user can further enter the confidential information into such websites.

To evaluate the proposed anti-phishing model for the detection of phishing websites, we have collected 7690 legitimate websites and 2280 phishing websites. All websites collected from APWG database and PhishTank. After collecting these websites, we have applied the classification server logic on these websites. We have tried to hit these websites with the activation of the proposed tool on the web browser. However, when the website gets hit by the user, instantly the tool displays the result on the web browser window about the type of website. After hitting the website, the website related information gets stored into the system’s database. The database information than crosschecked with the hitting website contents and gives the result whether the system tool gives right or wrong result. The database information is defined at 5 different servers. When user hit the website, the system sends the URL and website content information to all the servers. If the information is matched with one of the server, it replies to the Add-on tool. Add-on tools show the result on the user’s screen.

8. Data Mining Analysis

8.1. Study of the Anti-Phishing Tool with Class Imbalance Problem

In the proposed system, we have divided collected websites into a training data set and a testing data set for training the different classification model. The reason for taking different classification algorithms is to train the proposed model, to compare their detection performance and select one of that to perform the best result. Table 1 shows the classification result of the hitting websites for all the assigned servers.

Table 1. Classification Results of Hitting Websites

	Legitimate /Phishing Sites	TP	TN	FP	FN
Server-1	2250 / 580	2248	579	01	02
Server-2	1560 / 470	1554	466	04	06
Server-3	1480 / 410	1478	409	01	02
Server-4	1150 / 540	1142	536	04	08
Server-5	1250 / 280	1249	275	05	01
Average:	**7690 / 2280**	**7671**	**2268**	**12**	**19**

Where TP – True Positive, TN – True Negative, FP – False Positive and FN – False Negative.

In the above result, the value of Precision, Recall and F1-measures can be calculated by using the equations of Class Imbalance Problem. The results are as follows:

Table 2. Comparison of Precision, Recall and F1-Measure

	Precision	Recall	F1-measure (%)
Server-1	0.999555	0.999111	99.93332
Server-2	0.997433	0.996154	99.67928
Server-3	0.999324	0.998649	99.89861
Server-4	0.99651	0.993043	99.47735
Server-5	0.996013	0.9992	99.76038
Average:	**0.99776**	**0.99723**	**99.75**

The reason for taking 6 digit values after decimal to show the accuracy of the result. The above result shows the proposed system is producing around 99.75% success result at a particular time period.

8.2. Study of the Anti-Phishing tool with Rule based Classifier Method

The Sequential Covering Algorithm check the relationships of selected various phishing features. It uses multiple if-then statements which are based on the phishing sites probability with the different phishing characteristics. On the basis of the result received from the developed Anti-phishing tool, the system is classified in five input parameters (which are defined in five different servers) for the rule base and it has one output (which is mentioned as Risk Status). All the five servers are defined with the phishing parameters, so that when a user hit the website, corresponding server send the response. The servers contain all the defined if-then rules of the system. In each of the rule base, every component is assumed to be one of three situations (Low, Medium, High) and each criterion has different components, which gives the result about the website. The result received from Anti-phishing tool is as given in the Table 3. In the table, S-1, S-2, S-3, S-4 and S-5 are servers assigned for the anti-phishing tool. At the end of the table, the result obtained from all the servers is mentioned in percentage. The websites can be declared as *Phishing* if the percentage is higher than 40, if percentage is between 10 – 40, the website can be declared as *highly risky* and below 10 percentage, the website address would be stored in the anti-phishing tool's database for further checking. After checking the 10 percentage suspicious condition, the website would be declared as phishing or legitimate for further accessing of the website by the user.

Table 3. Phishing Characteristics and the Risk Status of the Accessed Website

Phishing Characteristics	Assigned Server	Feature Present					Risk Status
		S-1	S-2	S-3	S-4	S-5	
No. of dots '.' in the URL	1	Yes	No	No	No	No	Low
No. of '@' in the URL		Yes	No	No	No	No	Low
No. of '//' in the URL		No	No	No	No	No	No
Existence of IP address in the URL		No	No	No	No	No	No
Port Number in the URL		No	No	No	No	No	No
Title Tag	2	No	Yes	No	No	No	Low
Form Tags on the web page		No	Yes	No	No	Yes	Mode-rate
Image Tags on the web page		No	Yes	No	No	No	Low
href Tags on the web page		No	Yes	No	No	No	Low
Country Code present in the URL	3	No	Yes	No	No	No	Low
Login/Password evaluation		No	Yes	No	No	No	Low
Script Tags on the web page		No	No	No	No	No	No
Link Tags on the web page		No	Yes	No	No	No	Low
Website having HTTPs protocol	4	No	No	No	Yes	No	Legiti-mate
No. of Phishing Keywords in the URL		No	No	No	No	No	No
Domain Age < 1 yr	5	No	No	No	Yes	No	Low
First Webpage Creation Date Identified		Yes	Yes	Yes	Yes	Yes	No
Snap of Web matching		No	No	No	No	Yes	Low
Declared phishing websites from other authorities		Yes	No	No	No	Yes	High
Long URL Address		Yes	No	No	No	Yes	High
Rating		**25%**	**40%**	**5%**	**15%**	**25%**	
Website Class		**Risky**	**Highly Risky**	**Legi.**	**Legi.**	**Risky**	

8.3. Study of the Anti-Phishing tool with Nearest Neighbour Classification

According to the Nearest Neighbour Classification, the class label is determined for text example. The classification algorithm can be used for 2 or k number of times data sets. Here in the study of phishing website identification, a number of keywords can be taken which are very common to the website designer for making the similar website. Table 4 shows the different case condition of this classification.

Table 4. Different Case Conditions of Nearest Neighbour Rules

	Y_n^1	Y_n^2	Y_n^3
Case 1	0	0	1
Case 2	1	1	1
Case 3	0	0	0
Case 4	0	1	1
Case 5	1	0	0
Case 6	1	1	0
Case 7	1	0	1
Case 8	1	1	0

There are 4 conditions when hitting 2 website and while hitting 3 websites, there will be 8 conditions. If the result in all the conditions is 1, the website can be decided as phishing websites, in case of two 0's than the website is decided as 'less risky' and in the case of all 0's, the website is declared as 'legitimate'. In the case of two 1's and one 0, the website is treated as 'suspicious'.

The Nearest Neighbour classification technique gives better and accurate result when the checking conditions are less. In the case of a number of checking conditions, the algorithm performs slow and not gives accurate result. So at the initial state of testing of the website, the algorithm can be applied.

8.4. Study of the Anti-Phishing tool with Bayesian Classifier

The Bayesian theorem is generally used to solve the prediction problem. According to this classification algorithm, two websites can be cross checked on the basis of probability. To perform the calculation for a Bayesian filter, only two datasets are required; a dataset of phishing and a dataset for legitimate websites. A large number of datasets are required in Bayesian filter to find the most effective result. The following Table 5 shows the test performed on the phishing and legitimate websites and its result in the form of False Positive (FP), False Negative (FN).

Table 5. Test Performance of Websites by Bayesian Classification Method

	Phishing	**Legitimate**	**No Result**
Hitting Websites	985	1223	12
False Positive	14 (1.43%)	23 (1.89%)	8
False Negative	8 (0.8%)	10 (0.08%)	4

The above results shows that the False Positive rate is around 1.43% which means that the in the Bayesian Classification, the number of phishing websites can be calculated around 98.5% while the legitimate website finding percentage is around 98%. During the experiment, 8 websites are not found as phishing or legitimate because of non availability of the web contents of those websites. The Bayesian Classification is given satisfactorily result, but it is not suitable for the filter's effectiveness. Since the Bayesian Classification based on the probability of the situation, the results varies by hitting the target website repeatedly.

9. Result and Conclusion

In this paper, four different data mining algorithms have been discussed for the analysis of anti-phishing website data. Theses algorithms are Class Imbalance Problem (CIP), Rule based Classifier (Sequential Covering Algorithm (SCA)), Nearest Neighbour Classification (NNC), Bayesian Classifier (BC). How these algorithms can be applied on the given data set is also discussed. The performance evaluation and finding of the phishing websites by using these algorithms, an experimental study is performed. The *Class Imbalance Problem* is showing around 99 percentage of successful result in phishing website detection, if the database is already available but in case of finding the on-spot result, this algorithm is not successful. The *Rule based Classifier Method* is declaring the website as phishing or legitimate by using the percentage matching which doesn't shows the accurate result. The *Nearest Neighbour Classification* technique gives better and accurate result when the checking conditions are less. The result of *Bayesian Classification* shows that the False Positive rate is around 1.43% which means that in the Bayesian Classification, the number of phishing websites can be found around 98.5% while the legitimate website finding percentage is around 98. With the comparison of all data mining algorithms which are studied for phishing website detection, the Bayesian classification is more accurate and showing fast response to the system.

Acknowledgments

I thank Dr. Piyush Shukla, Assistant Professor for giving me valuable support and guidance to prepare the manuscript and also to the Principal, BSSS College to provide me the working environment for the research work. I am also thankful to my programmer Mr. Shankar Sarkar who have developed me tools for the result analysis. The college colleagues have given me additional moral support for my research design, data collection, analysis and interpretation.

References

[1] A Report from 'Computer Associate Internationals Inc.', **(2012)** September.

[2] A research report from http://securityresearch.in/?ubiquitous_id=88, **(2013)** January.

[3] A. Naga Venkata Sunil and A. Sardana, "A PageRank Based Detection Technique for Phishing Web Sites", 2012 IEEE Symposium on Computers & Informatics, **(2012)**, pp. 58-63.

[4] A. Abbasi and H. Chen, "A Comparison of Fraud Cues and Classification Methods for Fake Escrow Website Detection", Information Technology and Management, vol. 10, no. 2, **(2009)**, pp. 83-101.

[5] A. Abbasi, M. Z. Fatemeh and Y. Chen, "Impact of Anti-Phishing Tool Performance on Attack Success Rates", 10th IEEE International Conf. on Intelligence and Security Informatics, Washington, D.C., **(2012)** June 11-14.

[6] APWG 2nd Quarter Phishing Activity Trends Report from www.antiphishing.org, **(2013)**.

[7] T. Balamuralikrishna, N. Raghavendrasai and M. Satya Sukumar, "Mitigating Online Fraud by Anti phishing Model with URL & Image based Webpage Matching", International Journal of Scientific & Engineering Research, vol. 3, no. 3, **(2012)** March, pp. 1-6.

[8] CallingID Ltd. Report from http://www.callingid.com/DesktopSolutions/ CallingIDToolbar.aspx. Accessed: December 1 **(2008)**.

[9] N. Chou, R. Ledesma, Y. Teraguchi and C. Mitchell John, "Client-Side Defense Against Web-Based Identity Theft" 11th Annual Network and Distributed System Security Symposium, San Diego, **(2004)** February.

[10] Cloudmark, Inc. Report from http://www.cloudmark.com/desktop/download Accessed: **(2008)** September 5.

[11] Computer Crime Research Center. "Netscape: Anti-Phishing Bundled." February 2, 2010. Accessed: November 9, 2011. http://www.crimeresearch. org/news/02.02.2005/1024/ **(2010)**.

[12] R. Dhamija and J. D. Tygar, "The Battle against phishing: Dynamic Security Skins", Proc. of ACM Symposium on Usable Security and Privacy, **(2005)**, pp. 77-88.

[13] EarthLink, Inc. EarthLink Tool from http://www.earthlink.net/ Accessed: **(2010)** November 9.

[14] eBay, Inc. Using eBay Tool's Account Guard from http://pages. eBay.com/help/confidence/accountguard.html, Accessed: **(2010)** June 13.

[15] E. Ferguson, J. Weber and R. Hasan, "Cloud Based Content Fetching: Using Cloud Infrastructure to Obfuscate Phishing Scam Analysis", Eighth World Congress on Services, IEEE Computer Society, **(2012)**, pp. 255-261.

[16] Google, Inc. Google Safe Browsing for Firefox from http:// www.google.com/tools/firefox/safebrowsing/. Accessed: **(2010)** June 13.

[17] I. R. A. Hamid and H. Abawajy Jemal, "Profiling Phishing Email Based on Clustering Approach" 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, **(2013)**, pp. 629-635.

[18] H. Huang, S. Zhong and J. Tan, "Browser-side Countermeasures for Deceptive Phishing Attack", 2009 Fifth International Conference on Information Assurance and Security, IEEE Computer Society, **(2009)** pp. 352-355.

[19] Javelin Strategy and Research from http://www.javelinstrategy.com, **(2012)**.

[20] H. Jiang, D. Zhang and Z. Yan, "A Classification Model for Detection of Chinese Phishing E-Business Websites", PACIS 2013 Proceedings, Paper 152, **(2013)**.

[21] S. Michael Kerner, "Firefox 2.0 Bakes in Anti-Phish Antidote", Internet News. http://www.internetnews.com/devnews/ article.php/3609816, **(2006)**.

[22] M. Khonji, A. Jones and Y. Iraqi, "A Novel Phishing Classification based on URL Features", 2011 IEEE GCC Conference and Exhibition (GCC), Dubai, United Arab Emirates, **(2011)** February 19-22, pp. 221-224.

[23] T. Li, F. Han, S. Ding and Z. Chen, "LARX: Large-scale Anti-phishing by Retrospective Data-Exploring Based on a Cloud Computing Platform", Computer Communications and Networks, Proceedings of 20th International Conference, **(2011)** July 31-August 4, pp. 1-5.

[24] M. Mahmood Ali and L. Rajamani, "Deceptive Phishing Detection System (From Audio and Text messages in Instant Messengers using Data Mining Approach)", Proceedings of the International Conference on Pattern Recognition, Informatics and Medical Engineering (IEEE), **(2012)** March 21-23.

[25] A. Martin, N. Ba. Anutthamaa, M. Sathyavathy, M. Manjari Saint Francois and Dr. P. Venkatesan, "A Framework for Predicting Phishing Websites Using Neural Networks", IJCSI International Journal of Computer Science Issues, vol. 8, no. 2, **(2011)** March, pp. 330-336.

[26] Netcraft Anti-Phishing Tool. Accessed: June, 13, 2010. http://tool.netcraft.com/.

[27] Netscape Communications Corporation "Security Center" from http://browser.netscape.com/ns8/product /security.jsp Accessed: **(2006)** November 9.

[28] P. Prakash, K. Manish, R. R. Kompella and M. Gupta, "PhishNet: Predictive Blacklisting to Detect Phishing Attacks", presented as part of the Mini-Conference at IEEE INFOCOM, **(2010)**.

[29] Quick Start: Spoof Guard from http://crypto. stanford.edu/SpoofGuard/, **(2010)**.

[30] B. Wardman, T. Stallings, G. Warner and A. Skjellum, "High-Performance Content-Based Phishing Attack Detection", published in IEEE conference on eCrime Researchers Summit (eCrime), **(2011)**, pp. 1-9.

[31] D. Weider Yu, S. Nargundkar and N. Tiruthani, "PhishCatch – A Phishing Detection Tool", presented in 33rd Annual IEEE International Computer Software and Applications Conference, IEEE Computer Society, **(2009)** pp. 451-456.

[32] H. Zhang, G. Liu and W. S. Chow Tommy, "Textual and Visual Content-Based Anti-Phishing: A Bayesian Approach", IEEE Transactions on Neural Networks, vol. 22, no. 10, **(2011)** October.

[33] Y. Zhang, S. Egelman, L. Cranor and J. Hong, "Phinding Phish: Evaluating Anti-Phishing Tools", NDSS '07: Proceedings of the 14th Annual Network and Distributed System Security Symposium, **(2007)** February.

[34] S. Wu and Y. Zhu, "Improved Two-Factor Authenticated Key Exchange Protocol", International Arab Journal of Information Technology, vol. 8, no. 4, **(2011)**, pp. 430-439.

Authors

Mr. Rajendra Gupta has completed Master degree in Information Technology, M.Phil and pursing Ph.D. (Computer Science). He has published 6 research papers in International Journals, 2 research papers in National Conference and completed one Research Project. At present he is working as an Assistant Professor in Department of Computer Applications, BSSS Autonomous College, Bhopal for last seven years.

Dr. Piyush K. Shukla received his Bachelor's degree in Electronics & Communication Engineering, LNCT in 2001, Bhopal, M. Tech (Computer Science & Engineering) in 2005 from SATI, Vidisha, Ph.D. (Computer Science & Engineering) in 2013 from RGPV, Bhopal. M.P. India. He is a member of IACSIT. He has published more than 15 papers in reputed International Journals and 10 papers in International Conferences. At present, he is working as an Assistant Prof. in Department of Computer Science & Engineering, UIT-RGPV, Bhopal Since July 2007.