I. J. Computer Network and Information Security, **2015, 12, 70-77**
Published Online November 2015 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijcnis.2015.12.08

Performance Analysis of Anti-Phishing Tools and Study of Classification Data Mining Algorithms for a Novel Anti-Phishing System

Rajendra Gupta
BSSS Autonomous College, Barkatullah University, Bhopal - 462024, India
Email: rajendragupta1@yahoo.com

Piyush Kumar Shukla
University Institute of Technology, Rajiv Gandhi Technical University, Bhopal - 462026, India
Email: pphdwss@gmail.com

Abstract—The term Phishing is a kind of spoofing website which is used for stealing sensitive and important information of the web user such as online banking passwords, credit card information and user's password etc. In the phishing attack, the attacker generates the warning message to the user about the security issues, ask for confidential information through phishing emails, ask to update the user's account information etc. Several experimental design considerations have been proposed earlier to countermeasure the phishing attack. The earlier systems are not giving more than 90 percentage successful results. In some cases, the system tool gives only 50-60 percentage successful result. In this paper, a novel algorithm is developed to check the performance of the anti-phishing system and compared the received data set with the data set of existing anti-phishing tools. The performance evaluation of novel anti-phishing system is studied with four different classification data mining algorithms which are Class Imbalance Problem (CIP), Rule based Classifier (Sequential Covering Algorithm (SCA)), Nearest Neighbour Classification (NNC), Bayesian Classifier (BC) on the data set of phishing and legitimate websites. The proposed system shows less error rate and better performance as compared to other existing system tools.

Index Terms—Phishing, Anti-Phishing, Data Mining Algorithms, Add-on Anti-Phishing Tools.

I. Introduction

A number of governmental and private authorised agencies are doing study 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 keep all the information related to phishing and legitimate websites. The earlier anti-phishing systems do not show more than 90 percentage successful result [1-5]. According to information received from the records of APWG, the number of phishing attack reported around 24,370 from phishing websites in March 2014. During the second quarter of 2014, a total of 876 unique brands were targeted by phishing attacks [6]. In the report, even there is no economy loss mentioned but we can think if thousands of websites are declaring phishing in a month worldwide, how much loss could be possible. Based on a report given by Javelin Strategy and Research on April 2012, the economy loss reached to 21 billion [7]. 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 this thing in mind, the questions arises that how to identify the fraud and how to design and build a reliable and secure system environment for electronic business transactions. So the research study is very much important now-a-days to reduce the online transaction problems.

Several experimental design considerations have been proposed in earlier study to countermeasure the phishing attack. On the basis of the performance of earlier anti-phishing systems, some of the industries have applied these systems to protect their organisation from phishing attack. In phishing attack, the attacker creates the web pages that look like a mimic web page of the legitimate websites. Out of the proposed anti-phishing techniques, some are based on the machine learning or data mining algorithms. The data mining algorithms are applied on the set of possible features of phishing that can be extracted from the website. According to these features, the phishing problem can be solved by selecting proper/right set of features. In this study of anti-phishing, four different data mining algorithms have been applied on the data set of phishing and legitimate websites received in different days.

II. Previous Research Study On Phishing

A number of research papers have been presented and discussed in earlier study to countermeasure the phishing attack. Rosana J. Ferolin [8] suggested a Proactive Anti-Phishing Tool Using Fuzzy Logic; a Data Mining RIPPER algorithm which is used to characterize the Phishing e-mails and classify them based on both content-based and non-content based characteristics of Phishing emails. In addition, after the e-mail assessed and classified as a Phishing e-mail, the system proactively gets rid of the Phishing site by sending a notification message to the Administrator of the host server. The initial results of this system showed the RIPPER algorithm achieved 85.4% for correctly classified Phishing emails and 14.6% for wrongly classified Phishing emails based on publicly available datasets from Phistank. Colin Whittaker et.al. [9] presented a Large-Scale Automatic Classification of Phishing Pages in which the classifier analyzes millions of pages in a day, examining the URL and the contents of a page to determine whether or not a page is phishing. Despite the noise in the training data, its classifier learns a robust model for identifying phishing pages which correctly classifies more than 90% of phishing pages several weeks after training concludes. V. Shreeram et.al. [10] had given an idea of Anti-Phishing Detection of Phishing Attacks using Genetic Algorithm (GA) in which the algorithm generates a rule set that matches only the phishing links. This rule set is stored in a database and a link is reported as a phishing link if it matches any of the rules in the rule based system and thus it keeps safe from fake hackers. Preliminary experiments show that this approach is effective to detect phishing hyperlink with minimal false negatives at a speed adequate for online application. Tianyang Li et.al. [11] has proposed an offline phishing detection system which has given the name LARX (*L*arge-scale *A*nti-phishing by *R*etrospective data-e*X*ploration). It uses network traffic data which archived at a vantage point and analyzes the data for possible phishing attacks. LARX's phishing filter process uses cloud computing platforms which work parallel. Since LARX is an offline solution for phishing attack detection, it can be effectively scaled up to analyze a large volume of trace data when enough computing p+ower and storage capacity are provided. Huajun Huang et.al. [12] has explained a thorough overview of a deceptive phishing attack and its countermeasure techniques, which is called anti-phishing. In this paper, the technologies used by phishers and the definition, classification and future works of deceptive phishing attacks are discussed.

Edward Ferguson et.al. [13] have presented a research paper on "Cloud Based Content Fetching: Using Cloud Infrastructure to Obfuscate Phishing Scam Analysis", in which the proposed system presents different personas and user behavior to the phishing sites by using different IP addresses and different browsing configurations. By running a 10-day probe experiment against real phishing site, we show 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 [20-24]. Mohammed Mahmood Ali et.al. [14] presented a research 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. He has told that the online criminal's now-a-days adapted voice chatting technique along with text messages collaboratively or either of them in IM's and wraps out personal information leads to threat and hindrance for privacy. In order to focus on privacy preserving, they developed and experimented Anti Phishing Detection system (APD) in IM's to detect deceptive phishing for text and audio collaboratively. Abdullah Alnajim et.al. [15] has proposed an approach to the Implementation of the Anti-Phishing tool for Phishing Websites Detection to the implementation of the Anti-Phishing that uses Training Intervention for Phishing Websites Detection (APTIPWD), which also shows that the APTIPWD is feasible and can be implemented within any proxy-based network easily without writing a single line of a programming code and without undue disruption of the users system [16-19].

III. Research Criteria to Find Phishing

When a web user accesses the website, the user hit web address on URL or reached to the target webpage from any other website reference links. In this case, first of all the URL and its contents should be checked then the contents and existing images should be checked [28]. To check the various points of the website, it takes enough time to tally the website information with the database information stored in the database of the functional Add-on of the web browser. In the earlier study, browser-based client-side solutions have been proposed to mitigate the phishing attacks [29, 30]. Some techniques have also been developed which attempt to prevent phishing mails from being delivered [31, 32]. So we should have a system that can instantly check the fed information of the user with the database information of the system while user feed the confidential information in the website. To make the fast accessing system, we have defined the study points for the best possible solution. The studied criteria for the phishing have been collected from the previous study [25-27]. Following are the study points and the reason for taking these study points are discussed along with.

1. Number of Dots '. ' Present in the URL

When a website prepared, generally two dots '. ' are used for the separation of www and the domain type (For example www.yahoo.com). If more number of '.' are using in the website, it means the phishing attack is trying to redirect the website to another webpage. So if we find that the website is using more than 2 dots, we can keep the website in suspicious condition and the website can be checked by matching the associative features of the accessing webpage. The example of suspicious or

phishing website is http://www.myhomepage.co.in //yahoo.co/php or http://www.myhomepage.co.in/login.php etc.

2. Number of Dots '@' Present in the URL

Some of the phishing man uses '@' symbol to redirect the user to another website. Generally @ symbol is used in the FTP server to redirect the user. Since when user create his e-mail account, @ symbol is used. So the use of @ symbol in the URL is very good thinking of the attacker to spoof the internet user. The attacker can create the website like http://www.myhomepage.co@yahoo.com?login.com. In this case, the user can think that he is being directing from yahoo.com website.

3. Number of Slash ' // ' Present in the URL

When a website prepares, it is uploaded with either http or https protocols. 'http' protocol uses '//' symbol to redirect the webpage. So the attacker uses a number of '//' in the URL to spoof the web user. It is noticed that the legitimate websites do not use more than two '//' symbols during redirecting the webpage. So if an attacker uses more than two '//' symbols, we can identify whether the website is spoofing or not.

4. Existence of IP Address in the URL

To the functioning of any website, an IP address is provided to the domain of concerned URL. The sending and receiving of the data from the website functions with the use of this IP address. To spoof the user, generally attackers try to use IP address in URL instead of giving any alphabetic name. IPv4 addresses are separated in four different parts with the help of dot (.). For example, http://www.84.214.244.122 instead of http://www.mywebpage.com. In such situation, the internet user doesn't understand which website he is visiting.

5. Port Number in the URL

Some of the phishing URL try to redirect the web user to different port addresses. To do this attacker uses the target port number in its phishing URL address. For example the phishing website http://www.191.102.34.09:8087/http://myhomepage.co.in /index.htm trying to send the myhomepage website contents to 8087 port of the server. Generally server has assigned 80 or 8080 port number. By tracing the port number from the URL address, we can find the website is trying to spoofing the web user or not.

6. The Websites Which are Having Https Protocol

It is noticed that phishing attacks try to make almost similar website to the legitimate website by ignoring the security. The phishing attacks give attention on the changing of URL address, website contents, images etc. Since the security certificate is required for safe transaction over the website. The website holder takes the prior permission from the authority concern. When the authority gives the security permission to the website holder, the protocols convert with HTTPs. The website which uses HTTPs protocol can transfer the data securely. The phisher create the spoofed URL address by ignoring the HTTPs. For example in place of https://www.google.com, the attacker may create the website http://www.goooogle.com

If attackers try to use fake security certificate in the website, web browser automatically detect the fake certificate and do not give the permission to the website to function.

7. Number of Phishing Keywords Present in the URL

It is seen that some attacker uses phishing keywords in place of legitimate website by changing, replacing, shifting or deleting the characters from the website. For example in place of http://www.google.com, phishing attack can create the website like http://www.goooogle.com, http://www.goog1e.com, http://www.gugle.com etc. In this case suppose a user hit the wrong URL, it means that the user is sending his confidential information to a spoofed website.

8. Country Code present in the URL

While checking the URL country code with the help of WorldIP plug-in of the Mozilla Firefox web browser, it is found that the URL web address doesn't match the exact country which is mentioned in the web URL. It has been seen in the report of Advanced Phishing Working Group that some targeted country codes are using for web URL to lure the user. By cross-checking the country code and the IP address of the website, it can be determined that the user accessing website is phishing or not.

9. Title Tag

In general, phishing websites do not give attention on the title of the website. Sometimes the web address and the title tags remain different. For example the website link http://www.derezo.com/kf06/ppl/paypal.html uses the legitimate website title tag and trying to redirect the user to paypel website as target.

10. Form Tags on the Web Page

The Form tags are commonly used while preparing the website. It can be used for requesting user to input the data. For example, the Form tag can be used to ask the information like login, password, credit card number etc. Mostly the phishing website developer uses the same form tags and fields to spoof the user. So by finding the number of Form tags and name of Form tags used into the website, we can identify the website category.

11. Image Tags on the Web Page

A phishing website can be created by using images instead of using text. The images can be used by taking the snapshot of the legitimate website. In this case, by using the web image matching algorithms [33] we can find the accessing website is using the same size or different size image of legitimate website. To insert the images in the webpage, <image> tags are used.

12. Href Tags on the Web Page

The <href> tag is used to create a link to another document or webpage. We can count the number of href tags of visiting website with the legitimate website and by using this tag, we can check the reference webpage is legitimate page or not. Some times href tag is used to make the link with legitimate webpage and sometimes it redirects the user to unauthorised webpage. If the href reference page matches the link with visiting webpage, the site will be legitimate otherwise phishing.

13. Login/Password Evaluation

The phishing websites uses login and password keywords in its webpage. The previous study of phishing has been done on the basis of these two keywords and found that generally banking and e-commerce websites uses these keywords to collect the username and password of the web user. The legitimate websites which ask the login and password information of the user takes the permission from the security authorities as a Security Certificate to protect the webpage. The HTTPs protocol is assigned for such websites. The phishy websites do not take permission of security authorities to secure the webpage, so we can check the login and password tags with HTTPs protocol of the website. On the basis of these tags, we can identify the accessing website is phishing or legitimate.

14. Script Tags on the Web Page

The phishing site uses the <script> tag to redirect the web user to client-side system. The <script> tag is used to define a client-side script, such as a JavaScript. The <script> element either contains scripting statements, or it points to an external script file through the src attribute. Common uses for JavaScript are image manipulation, form validation and dynamic changes of content. We can find the number of script tags in the accessing websites and can cross-check these tags with the legitimate website scripts tags. If the number of script tags of accessing and legitimate website matches, we can keep the accessing website for the observation as phishing.

15. Link Tags on the Web Page

While accessing the link tags of phishing website, it doesn't work or redirect the user to legitimate site which are not directly concerned with the visiting website. A number of link tags are possible like image tag, href tag, form tag, title tag etc. I have examined that while checking all the tags of the webpage, some links does not match the domain name or send the user to unconcerned webpage.

Apart from above finding criteria, we can also find the domain age from www.domaintools.com website. By the use of this website, we can find the website creation date. Some of the governmental authorities are also working on this topic of phishing to find the better solution to protect the user from internet fraud. These authorities have already declared many websites as phishing, so we have taken the help from these authorised sites to increase our database record. Wang Binjuna et.al. [36] has proposed Anti-phishing Authentication system for based on special custom image and client program.

IV. Contribution of the Research Study

Several phishing and anti-phishing techniques have been developed to protect the user from electronic fraud. In this research study, a novel anti-phishing solution is proposed by which the web user, website designer, network security provider, governmental and private sector persons can be aware of possible phishing attacks. This study is very useful to reduce the negative consequences of semantic attacks on the society by useful security information. The educational knowledge and key terminology which is used in the research study can potentially help and educate the increasing number of people who fall for phishing and other semantic attacks. The policy recommendations from this research could help various stakeholders to better prioritize their resources and manage their risks to fight for phishing and other semantic attacks over the net. The researcher and research scholar can study this research work for getting information about the phishing solutions. This study is also useful for persons who are associated with the fields of computer security and privacy, human computer interaction, financial and economic sectors. With the study of this research paper, people can much better distinguish the phishing and legitimate websites. After studying different anti-phishing system models, a novel system algorithm is developed to make fast and effective anti-phishing system.

V. Methodology

The proposed system works on the basis of the following key considerations:

1. Designing and implementing the anti-phishing Add-on for the web browser and its evaluation which gives the essential information about the phishing and blocking the phishing websites, so that the internet user can protect themselves by semantic attacks. This research shows that the computer user can be trained to take better decisions while feeding the confidential information on the website.
2. The performance of the proposed anti-phishing system is discussed with different peoples of different fields. The anti-phishing system is checked and compared with the other options available on the internet. Our analysis led to many key findings and recommendations to improve the phishing countermeasures.
3. The proposed system shows the details of false positive, false negative, true positive and true negative situations. It means the system model is differentiating different websites according to their behaviour and work. We have studied the effectiveness of popular phishing tools that is used

by major web browsers. It is found that no tool is giving accurate and timely result, so the use of proposed system can be beneficial for the internet user. It is found that blacklists were ineffective when protecting users initially. The tool that uses heuristics to complement blacklists caught significantly more phishing websites.

4. The system design is capable of storing the blacklisted and legitimate website information and showing how the websites are declaring as blacklisted or legitimated. By using this knowledgeable information, the internet user can be aware of the website. By viewing the website screenshot, the user can guess the website type and if it looks like phishing than the user can send the website information to the add-on tool to check it. After checking the website, the system tool will give the instant result to the user.

A. Algorithm for the Novel Anti-Phishing Tool

To differentiate the desired check point for phishing criteria which are defined on the main server that has to be sent to assigned server for cross checking the phishing indicator, the system works on the basis of Algorithm 1.1. In this algorithm, the variable ***K*** is initialized as the number of servers assigned for the proposed system and ***N*** be the check point for the training dataset. When user hit the web URL, ***K*** check and assign ***N*** for ***z***, where ***z*** is the result of response from cross-checking the phishing indicators. For every hitting websites, the values of ***z*** is stored in ***D*** and if the phishing indicators found in the database source, the selected phishing keywords are counted and replied to the main server. At the main server, the total number of check points are matches and assigned to a variable ***a***. The main server send the feedback to the user after analysing ***a*** by selected data mining algorithms for the study of the phishing analysis.

Algorithm 1.1

1. Let ***K*** be the number of servers assigned for the Add-on and ***N*** be the check point for the training dataset
2. **for** each test hit on URL, $z = \{x', x\}$, where z is the comparative variable
3. **do** compute z for every (x, y) ε D
4. Select $D_s \subseteq D$, the selected keywords for the check point of ***N***
5. if D_s ε K_i assign D_s to K_i, where i the assigned server's number
6. **end for**
7. **if** N ε D_s **then**
8. Compute the total number of check point matching from the data base source and assign to a, where a is an integer which gives the result
9. return a
10. **else**
11. return Null

VI. Performance Analysis of Proposed System & Its Comparison with the Existing Tools

In the proposed system, a web browser based anti-phishing add-on is installed at main server. All the phishing criteria are defined at different assigned servers to the main server. When a user hit any website, the add-on sends the webpage details to all assigned servers. Theses servers are categorised for different criteria of check points like *Character based, Coding based, Identity based, Contents based and Attribute based*. We have collected 8540 legitimate websites and 4480 phishing websites from APWG database [34] and PhishTank [35]. These websites are collected in 10 different days for the month of November and December, 2014. Since APWG and PhishTank are the trusted and reliable source, which keeps all the information about legitimate and phishing websites, so these are very helpful in the research study.

The performance of the proposed system is tested with four different types of data mining classification algorithms; these are Class Imbalance Problem (CIP), Sequential Covering Algorithm (SCA), Nearest Neighbour Classification (NNC) and Bayesian Classifier (BC). Since, all these algorithms works differently and cover almost all areas of data mining problems, so the study of these algorithms have been successful while doing the study of anti-phishing system performance.

In the earlier study, the research has been done on the concept of black list (the websites which are black listed by anti-phishing tool), while list (the websites which can be safely accessing permission by anti-phishing tool), legitimate and spoofing websites. To achieve the accuracy, a series of experiments are needed. To investigate the system performance, a study of earlier tool's functioning is important. In this study, we have studied the functioning of 7 anti-phishing tools which are CallingID, Earthlink, NetCraft, NetScape, SpoofGuard, eBay and Cloudmark Anti Fraud Toolbars and installed these tools in 7 different machines to check the performance of these tools. We have conducted a series of experiments in different days on these tools to get the accuracy.

During the study of anti-phishing tools, we found some effectiveness and limitations of the tool's functioning. Some tools gives very fast result, some display late result and some of these tools are not identifying the phishing websites. During the study, it is found that some tools are declaring legitimate websites as phishing in some cases. The Table 1 is showing the result obtained from various anti-phishing tools on the basis of data mining techniques. To get the result we have conducted a number of tests hit on the web browser URL. To hit the website, we have collected around 1240 legitimate and 1480 declared phishing websites by APWG and PhishTank during the month of November-December, 2014. To find the better result, we have taken 8 computer systems having same configuration and same time period. The configuration of the system was, Windows 7 Operating System, Core I3 2.4 GHz processor, 8 GB RAM and the same version of web-browser (on each system).

A group of 8 computer experts have accessed these declared websites at the same time. Same assignment has been given to all the computer experts. After hitting the websites, the anti-phishing tools get active on the system and give the output result on the web browser screen. According to the obtained result, the data set is applied on WEKA data mining tool to analyse the result. The WEKA software has given the result in the form of percentage for accuracy, which is displayed in the Table 1.

The above results are not showing 100 percentage accuracy of finding phishing or legitimate websites. In above tools, some are not identifying the website correctly whether it is phishing or legitimate. According to the Anti-Phishing Working Group, the phishing websites remain active for a short period of time from 1 day to 4-5 days.

Table 1. Comparison of Anti-Phishing Tools that correctly identify the phishing /legitimate web sites using different data mining algorithms

Anti-Phishing Tools		Data Mining Algorithms							
		CIP		SCA		NNC		BC	
		TPR	FNR	TPR	FNR	TPR	FNR	TPR	FNR
CallingID	L	78	10	72	09	84	12	70	08
	P	94	04	92	05	82	08	95	04
Earth link	L	87	08	81	10	NA	NA	NA	NA
	P	91	06	93	06	NA	NA	NA	NA
Net Craft	L	82	06	83	06	88	06	84	05
	P	94	03	91	02	94	03	95	04
Net Scape	L	28	18	58	12	74	12	82	08
	P	94	05	94	06	92	05	95	04
SpoofGuard	L	82	07	72	08	84	06	92	04
	P	87	03	92	02	94	05	96	03
Cloudmark Anti Fraud	L	72	12	78	09	NA	NA	NA	NA
	P	90	06	92	08	NA	NA	NA	NA
eBay	L	58	14	76	10	74	08	82	09
	P	94	06	93	05	92	04	94	06
ePhishNet (Proposed Add-on)	L	88	07	87	06	92	06	96	03
	P	96	03	95	03	96	03	98	02

* TPR – True Positive Rate, TNR – False Negative Rate, L – Legitimate, P - Phishing

If tools do not give the result instantly, so it is worthless. The above results are displaying around 80-90 percentage accuracy for True Positive (TP) and 90-98 percentage accuracy for False Negative (FN). In the received result, Earthlink and CloudMark Anti Fraud toolbars are not showing any result with the study of Nearly Neighbour Classification (NNC) and Bayesian Classification (BC) because their result is not obtained during the testing time. The above results can be changed if a number of tests are performed at another time period, but the result would be differing slightly. Therefore regular study is needed to get the better result from anti-phishing tool.

The result shows that the proposed anti-phishing system (ePhishNet) in which servers are classified, is showing almost better result and in some cases equivalent result by using data mining algorithms. So the concept of dividing the main server in different assigned serves shows the better and instant result. The accuracy of the proposed anti-phishing system is around 96 percentages to find the legitimate website, around 98 percentages to find phishing websites and only 2-3 percentage websites are falsely identifying.

The anti-phishing tools performance can also be calculated by getting the False Positive cases while hitting the websites. If the percentage of falsely identifying website is higher means the anti-phishing tool is not giving satisfactorily result. The error rate of the system can be determined with Number of wrong predictions divided by the Total number of predictions.

$$Error\ Rate = \frac{Number\ of\ wrong\ predictions}{Total\ number\ of\ predictions}$$

The above formula is applied on different data mining algorithms which we have taken in our study and found the following outcomes.

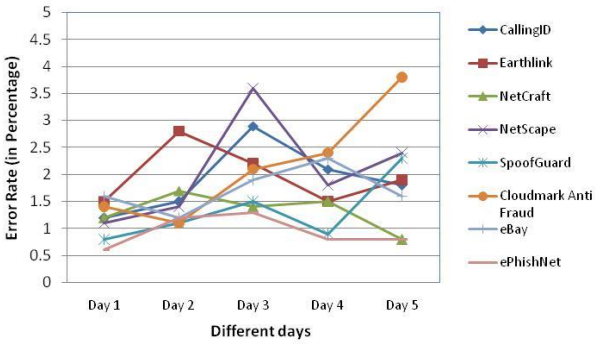

Fig. 1. Error Rate Analysis of Different Anti-Phishing Tools using Class Imbalance Problem (CIP), when hitting the websites in different days

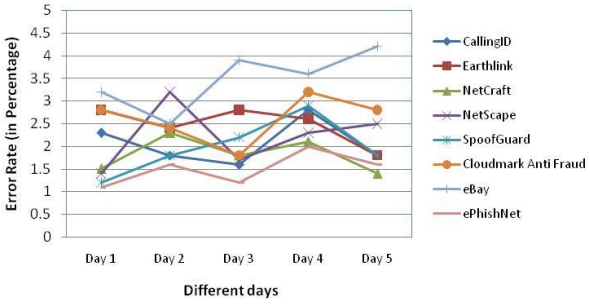

Fig. 2. Error Rate Analysis of Different Anti-Phishing Tools using Sequential Covering Algorithm (SCA), when hitting the websites in different days

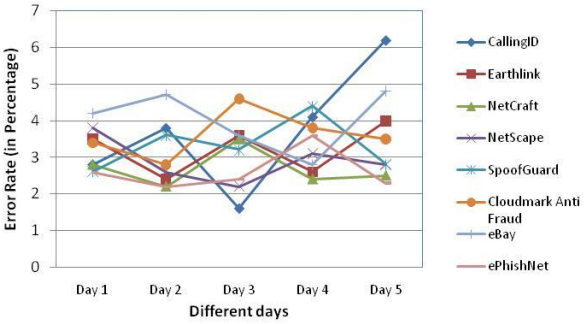

Fig. 3. Error Rate Analysis of Different Anti-Phishing Tools using Nearest Neighbour Classification (NNC), when hitting the websites in different days.

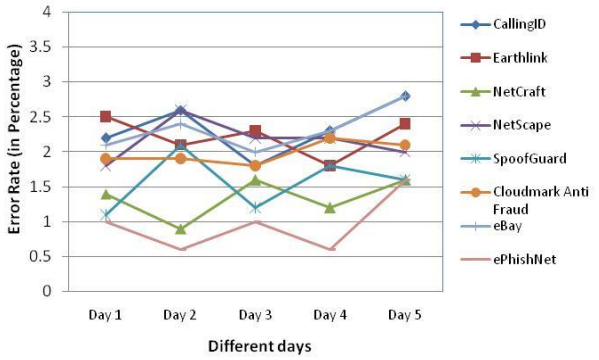

Fig. 4. Error Rate Analysis of Different Anti-Phishing Tools using Bayesian Classification (BC), when hitting the websites in different days.

The above result shows that the anti-phishing tools are giving different performance when applying different types of data mining classification algorithms. In all the result, the proposed system is showing almost better performance than other anti-phishing tools. Since the database information in the proposed anti-phishing system is divided on different assigned servers, so the server which is getting the suspicious website information, send the feedback to the main server. During this period, other servers do not send any response. In this case, the system gives fast result and the error rate goes down and the system shows better result.

VII. Conclusion

In the phishing attack, the user sends their confidential information on mimic websites, so the user should be informed immediately about the category of website. For this, a browser based add-on is prepared by categorizing the system with different assigned servers. The functioning of all the servers is defined. When the web user hit any webpage, the add-on send the website information to the database server and it send the request to assigned servers. The assigned server sends the feedback to the add-on tool and instantly user informed. To make awareness among web users about phishing or legitimate website, the web browser should provide the security tools. In the proposed system, the outcome from different servers is tested by using four data mining algorithms. The performance evaluation of the proposed anti-phishing system is compared with the existing anti-phishing tools and found that if the task is divided into different groups, it can give better results. The WEKA data mining tool is used in the study to analyse the machine learning algorithms. The proposed anti-phishing system 'ePhishNet' is compared with the CallingID, EarthLink, NetCraft, NetScape, SpoofGuard, CloudMark and ebay anti-phishing systems. The four data mining algorithms Class Imbalance Problem (CIP), Rule based Classifier (Sequential Covering Algorithm (SCA)), Nearest Neighbour Classification (NNC), Bayesian Classifier (BC) have been applied on these systems. This study would be helpful for the further study of anti-phishing system development.

Acknowledgment

I thank my guide Dr. Piyush Shukla for giving me valuable support to prepare the manuscript and the BSSS College provided me the working environment to do the research work. I am also thankful to my programmer who have developed me tools for the result analysis. The college has given me additional support for my research design, data collection, analysis and interpretation.

References

[1] Ahmed Abbasi, Fatemeh "Mariam" Zahedi and Yan Chen, "Impact of Anti-Phishing Tool Performance on Attack Success Rates", *10th IEEE International Conference on Intelligence and Security Informatics (ISI)* Washington, D.C., USA, June 11-14, 2012.

[2] A. Abbasi and H. Chen, "A Comparison of Fraud Cues and Classification Methods for Fake Escrow Website Detection," *Information Technology and Management*, Vol. 10(2), pp. 83-101, 2009.

[3] G. Bansal, F. M. Zahedi, and D. Gefen, "The Impact of Personal Dispositions on Information Sensitivity, Privacy Concern and Trust in Disclosing Health Information Online," *Decision Support Systems*, Vol. 49(2), pp. 138-150, 2010.

[4] Y. Chen, F. M. Zahedi, and A. Abbasi, "Interface Design Elements for Anti-phishing Systems," *In Proc. Intl. Conf. Design Science Research in Information Systems and Technology*, pp. 253- 265, 2011.

[5] S. Grazioli and S. L. Jarvenpaa, "Perils of Internet Fraud: An Empirical Investigation of Deception and Trust with Experienced Internet Consumers," *IEEE Trans. Systems, Man, and Cybernetics Part A*, vol. 20(4), pp. 395-410, 2000.

[6] APWG 2nd Quarter 2014 Phishing Activity Trends Report from www.antiphishing.org

[7] Javelin Strategy and Research. http://www.javelinstrategy.com, 2012

[8] Rosana J. Ferolin, "A Proactive Anti-Phishing Tool Using Fuzzy Logic and RIPPER Data Mining Classification Algorithm", pp. 292-304, 2012.

[9] Colin Whittaker, Brian Ryner, Marria Nazif, "Large-Scale Automatic Classification of Phishing Pages", *The Internet Society, 2010*.

[10] V.Shreeram, M.Suban, P.Shanthi, K.Manjula, "Anti-Phishing Detection of Phishing Attacks using Genetic Algorithm", *Communication Control and Computing Technologies (ICCCCT)*, IEEE International Conference, 7-9 October 2010.

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

[12] Huajun Huang, Shaohong Zhong, Junshan Tan, "Browser-side Countermeasures for Deceptive Phishing Attack", *Fifth International Conference on Information Assurance and Security*, IEEE Computer Society, pp. 352-355, 2009.

[13] Edward Ferguson, Joseph Weber, and Ragib Hasan, "Cloud Based Content Fetching: Using Cloud Infrastructure to Obfuscate Phishing Scam Analysis", *IEEE Eighth World Congress on Services, IEEE Computer Society*, pp. 255-261, 2012.

[14] Mohammed Mahmood Ali, Dr. Lakshmi 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)*, March 21-23, 2012.

[15] Abdullah Alnajim, Malcolm Munro, "An Approach to the Implementation of the Anti-Phishing Tool for Phishing Websites Detection", *International Conference on Intelligent Networking and Collaborative Systems*, IEEE Computer Society, pp. 105-112, 2009.

[16] J. S. Downs, M. B. Holbrook and L. F. Cranor, "Decision strategies and susceptibility to phishing". *Proc. the 2nd Symposium on usable Privacy and Security*. New York, USA: ACM Press, 2006, pp. 79–90.

[17] M. Chandrasekaran, R. Chinchani and S. Upadhyaya, "PHONEY: Mimicking User Response to Detect Phishing Attacks". *Proc. International Symposium on a World of Wireless, Mobile and Multimedia Networks*. Washington DC: IEEE Computer Society, 2006, pp. 668-672.

[18] S. A. Robila and J. W. Ragucci, "Don't be a Phish: Steps in User Education". *Proceeding 11th annual SIGCSE Conference on Innovation and Technology in Computer Science Education*. New York: ACM Press, 2006, pp. 237 – 241.

[19] A. Alnajim and M. Munro, "An Anti-Phishing Approach that Uses Training Intervention for Phishing Websites Detection". *Proc. 6th IEEE International Conference on Information Technology - New Generations (ITNG)*. Las Vegas, IEEE Computer Society, 2009, pp. 405-410.

[20] R. Weaver and M. Collins, "Fishing for phishes: applying capture-recapture methods to estimate phishing populations," in *Proc. of the Anti-phishing Working Groups, 2nd Annual eCrime Researchers Summit*. ACM, 2007, pp. 14–25.

[21] S. Sheng, B. Wardman, G. Warner, L. Cranor, J. Hong, and C. Zhang, "An empirical analysis of phishing blacklists,", *Proceeding of CEAS*, 2009.

[22] M. Cova, C. Kruegel, and G. Vigna, "There is no free phish: an analysis of free and live phishing kits", *Proceeding of USENIX WOOT. USENIX Association*, p. 4, 2008.

[23] B. Wardman, T. Stallings, G. Warner, and A. Skjellum, "High-performance content-based phishing attack detection", *Proceeding of eCrime*. IEEE, pp. 1–9, 2011.

[24] C. Whittaker, B. Ryner, and M. Nazif, "Large-scale automatic classification of phishing pages", *Proceeding of NDSS*, 2010.

[25] M. Hara, A. Yamada, and Y. Miyake, "Visual similarity-based phishing detection without victim site information" Nashville, Tennessee, *USA: IEEE*, pp. 30–36, Apr. 2009.

[26] Y. Zhang, S. Egelman, L. Cranor, and J. Hong, "Phinding phish: Evaluating Anti-Phishing tools," in *Proceedings of the 14th Annual Network & Distributed System Security Symposium*, San Diego, California, USA, Mar. 2007.

[27] Y. Zhang, J. Hong, and L. Cranor, "CANTINA: A Content-Based approach to detecting phishing web sites," in *Proceedings of the 16th international conference on WorldWideWeb*. Banff, Alberta, Canada: ACM, May 2007, pp. 639–648.

[28] Matthew Dunlop, Stephen Groat, and David Shelly "GoldPhish: Using Images for Content-Based Phishing Analysis", *The Fifth International Conference on Internet Monitoring and Protection*, IEEE Computer Society, pg. 123-128, 2010

[29] N. Chou, R. Ledesma, Y. Teraguchi, D. Boneh, and J. Mitchell. Client-side defense against web-based identity theft. In *11th Network and Distributed System Security Symposium (NDSS)*, 2004.

[30] B. Ross, C. Jackson, N. Miyake, D. Boneh, and J. Mitchell "Stronger Password Authentication Using Browser Extensions", in *14th Usenix Security Symposium*, 2005.

[31] Microsoft. Sender ID Framework Overview. http://www.microsoft.com, 2005

[32] Yahoo. Yahoo! Anti-Spam Resource Center. http://antispam.yahoo.com, 2006.

[33] Matthew Dunlop, Stephen Groat, and David Shelly, "GoldPhish: Using Images for Content-Based Phishing Analysis", *The Fifth International Conference on Internet Monitoring and Protection*, IEEE Computer Society, 2010

[34] APWG 2nd Quarter 2014 Phishing Activity Trends Report from www.antiphishing.org

[35] Phishing website list from http://www.phishtank.com/, November 2013.

[36] Wang Binjuna, Wei Yangb, Yang Yanyanc, Han JiaI. J., "Design and Implementation of Anti-phishing Authentication System Wireless and Microwave Technologies", Published Online at IJWMT-MECS, pp. 38-45, December 2011

Authors' Profiles

Mr. Rajendra Gupta has completed Master degree in Information Technology, M.Phil and pursing Ph.D. (Computer Science). He has published 8 research papers in International Journals, 3 research papers in National Conferences 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 eight years and Member of the various Academic Bodies.

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.

How to cite this paper: Rajendra Gupta, Piyush Kumar Shukla,"Performance Analysis of Anti-Phishing Tools and Study of Classification Data Mining Algorithms for a Novel Anti-Phishing System", IJCNIS, vol.7, no.12, pp. 70-77, 2015.DOI: 10.5815/ijcnis.2015.12.08

