ISSN- 2350-0530(O) ISSN- 2394-3629(P)
IF: 4.321 (CosmosImpactFactor), 2.532 (I2OR)

INTERNATIONAL JOURNAL OF RESEARCH – GRANTHAALAYAH
A knowledge Repository

Science

COMPARISON OF CLASSIFICATION ALGORITHMS TO DETECT PHISHING WEB PAGES USING FEATURE SELECTION AND EXTRACTION

Dr. Rajendra Gupta [*1]
[*1] Assistant Professor, BSSS Autonomous College, Barkatullah University, Bhopal - 462 024, INDIA

DOI: https://doi.org/10.29121/granthaalayah.v4.i8.2016.2570

ABSTRACT

The phishing is a kind of e-commerce lure which try to steal the confidential information of the web user by making identical website of legitimate one in which the contents and images almost remains similar to the legitimate website with small changes. Another way of phishing is to make minor changes in the URL or in the domain of the legitimate website. In this paper, a number of anti-phishing toolbars have been discussed and proposed a system model to tackle the phishing attack. The proposed anti-phishing system is based on the development of the Plug-in tool for the web browser. The performance of the proposed system is studied with three different data mining classification algorithms which are Random Forest, Nearest Neighbour Classification (NNC), Bayesian Classifier (BC). To evaluate the proposed anti-phishing system for the detection of phishing websites, 7690 legitimate websites and 2280 phishing websites have been collected from authorised sources like APWG database and PhishTank. After analyzing the data mining algorithms over phishing web pages, it is found that the Bayesian algorithm gives fast response and gives more accurate results than other algorithms.

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

Cite This Article: Dr. Rajendra Gupta, "COMPARISON OF CLASSIFICATION ALGORITHMS TO DETECT PHISHING WEB PAGES USING FEATURE SELECTION AND EXTRACTION" International Journal of Research – Granthaalayah, Vol. 4, No. 8 (2016): 118-135.

1. INTRODUCTION

A number of governmental and private authorised agencies are working on the topic of phishing and the countermeasure the phishing attack. The APWG (Advanced Phishing Working Group) and PhishTank are two most prominent agencies which keeps all the information related to phishing and legitimate websites. According to information received from the record of APWG,

ISSN- 2350-0530(O) ISSN- 2394-3629(P)
IF: 4.321 (CosmosImpactFactor), 2.532 (I2OR)

the total number of unique phishing web sites detected from quarter 1 to quarter 3 were 630,494 in the year 2015 [1].

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 [4]. Nevertheless, the phishing is seriously challenging and collapses the trust to electronic commerce and e-services security systems. 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 necessary to reduce the online transaction problems.

To solve the problem of phishing, the researchers are finding the solution at client side and server site systems. So far, slow progress has been noticed in the client and server side design testing. On the client side application, there have been around 110 types of user-centred applications developed. These application uses web browser toolbar and additional plug-in to install additionally with the web browser. It is found that the server site strong system designing is more important requirement 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 securities are proposing and applications are working at client site applications [5, 6].

2. METHODS OF PHISHING ATTACK

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

- ***Link manipulation:*** Several methods of phishing attack uses some kind of technical deception which is designed to make a link in an e-mail that appears to belong to the spoofed organization. Phishers try to misspell the URLs or the use of sub-domains to target the user. In an example of URL for http://www.mybank.services.com/, it appears that the URL is asking to login into 'mybank.services' section of the webpage, which is an phishing URL.
- ***Filter evasion:*** Here phisher uses images instead of text to make it harder for anti-phishing filters to detect text, commonly used in phishing e-mails. This type of phishing takes less time to prepare the spoof websites, and it uses very less coding statements to prepare the webpage.
- ***Website forgery:*** An attacker can even use flaws in a trusted website's own scripts against the victim. This type of attack (known as cross-site scripting) are particularly problematic because they direct the user to sign in at their bank or services section of web page, where everything from the web address to the security certificates appears correct.
- ***Phone phishing:*** Since the use of mobile and the internet access from mobile is increasing speedily, so it is seen that not all phishing attacks requires the use of fake website. The messages come from the mobile that claimed to be from a bank which ask user to dial a phone number regarding problems with their bank account information.

ISSN- 2350-0530(O) ISSN- 2394-3629(P)
IF: 4.321 (CosmosImpactFactor), 2.532 (I2OR)

- ***Tabnabbing:*** Tabnabbing is one another kind of phishing attack which directs the user to submit their login information and passwords to popular websites by impersonating those sites and convincing the user that the site is genuine [45].
- ***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 type of phishing, the phisher 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 users do not sure that the website in which they are entering their confidential information is controlled by phisher and is probably not even in the same country as the legitimate website [7].

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] 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 Google PageRank technique and it shows only 0.02 false positive rate and 0.02 false negative rate. Khonji M. et.al [8] proposed A Novel 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 the system on a mail server where email contents are checked passively without imposing a delay on client side applications. Wardman B. et.al. [9] 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 that contains 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 more than 90% by maintaining a low false positive rate.

Weider D.Yu et.al. [10] 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 user about 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 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. [11] presented a heuristics "PhishNet" in which five heuristics has been taken to enumerate simple combinations of known phishing sites to discover new phishing URLs. In its evaluation with real-time blacklist feeds discovered around 18,000 new phishing URLs from a set of 6,000 new blacklist entries. He showed that approximate matching algorithm leads to very few false positives (3%) and negatives (5%). Isredza Rahmi A Hamid et.al. [12] suggested an Profiling Phishing E-mail Based on Clustering Approach in which an approach for profiling email-born phishing activities is proposed. Profiling phishing activities are useful in determining the activity of an individual or a particular group of phishers. By generating profiles, phishing activities can be well understood and observed. His proposed profiling email-born phishing

ISSN- 2350-0530(O) ISSN- 2394-3629(P)
IF: 4.321 (CosmosImpactFactor), 2.532 (I2OR)

algorithm (ProEP) demonstrates promising results with the Ratio Size rules for selecting the optimal number of clusters. Zhang H. et.al. [13] 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 that collected from real phishing cases of trusted sources. The experimental results of Zhang demonstrated the text and image classifier that is designed to deliver promising results. They uses fusion algorithm that outperforms the individual classifiers. His model can be adapted for the further study on phishing.

Li T. et.al. [14] has proposed an offline phishing detection system named *L*arge-scale *A*nti-phishing by *R*etrospective data-e*X*ploration (LARX). This system uses a network traffic data archived at a vantage point and analyzes the data for phishing detection. The proposed phishing filter in the system uses cloud computing platform. Since the system is offline for the detection of phishing, LARX can be effective for the analysis of large volume of trace data when enough computing power and storage capacity is used. Huang H. et.al. [15] explained a thorough overview of a deceptive phishing attack and its countermeasure techniques. In his study, the technologies used by phishers with the definitions, classification and future works of deceptive phishing attacks have been discussed. Edward Ferguson et.al. [16] presented 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, they have shown 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 [17, 18].

Mahmood Ali M. et.al. [19] 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 are a number of anti-phishing approaches proposed in earlier study that can be used to identify a web page as a phishing or not. I have taken observations to get a basic understanding of how each tool function. The earlier tools are trying to protect 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 described below [20] :

CallingID focuses on the site ownership details and real-time rating and confirm user that the site is safe 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 show different colours for differentiating the web page. For example green colour shows 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 used include examining the site's country of origin, length of registration, user reports, popularity of the website and the blacklisted data [21].

ISSN- 2350-0530(O) ISSN- 2394-3629(P)
IF: 4.321 (CosmosImpactFactor), 2.532 (I2OR)

The Cloudmark Anti-Fraud toolbar is based on the user's ratings [22]. When user visits the website, he has the right to report the site as the site needs to be accessible or not. On the basis of this feature, the toolbar display a coloured icon for each site visited by the user. The user 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 [23]. The toolbar allows user 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.

The eBay tool uses a combination of heuristics and blacklists [24]. 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 user 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 [25]. 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 download site, the toolbar combines "advanced algorithms with reports about misleading pages from a number of sources [26]."

The Netcraft Anti-Phishing Toolbar uses several methods to determine the legitimacy of a web site [27]. The Netcraft web site explains that the toolbar traps the suspicious URL 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 [28]. 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. 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 tool to help preventing a form of malicious attack called "web spoofing" or "phishing" [29]. Phishing attacks usually involve deceptive e-mail that appears to come from a popular commercial site. The email explains that the recipient has an account problem, or some other reason to visit the commercial site and log in. However, the link in the email sends the user to a malicious "spoof" site that collects user's information such as account names, password and

ISSN- 2350-0530(O) ISSN- 2394-3629(P)
IF: 4.321 (CosmosImpactFactor), 2.532 (I2OR)

credit card number etc. Once the user information is collected by a "spoof:" site, criminals may log into the user's account or cause other damage.

5. AN APPROACH FOR THE EXPECTED OUTCOMES

The prior exposure of phishing knowledge is very much important to protect the user from phishing attack. When a user uses or access the website, a message should be appeared on the web browser window that shows the type of website whether it is suspicious, phishing or legitimate. By using this method, user can be informed about the type of website and can take a decision to use the website or not. The proposed add-on informs the user instantly when user hit the web address. Using this Add-on, user can learn the difference between legitimate and phishing websites. The following points should keep in mind when an anti-phishing tool is designed and prepared.

1) While preparing web-browser based Add-on anti-phishing tool, the concept of division of phishing keywords should take place to different assigned servers for achieving the fast and accurate result. As per the study of anti-phishing tool functioning, it is noticed that the tools are not giving timely and accurate results.
2) When a new phishing website is activated, various anti-phishing tools do not give proper message to the user or do not identify the website. In this situation, the new arriving websites should be stored in the anti-phishing tool's database when the user hit it. The website should be analysed at the time of execution and should display the result instantly.
3) Some anti-phishing tools do not support to web browser properly. In this case, the anti-phishing tool do not give the satisfactorily result. While making anti-phishing add-on tool, the tool should be compatible to all the web-browsers. Here in the proposed add-on designing system, an executable file is prepared which support all the web browsers.

It is noticed with the previous results, the anti-phishing tool give late response to web browser. In this situation, the user fed their confidential information in the suspicious website and get aware later about the type of website. It is very difficult task to inform to web user timely about the website category. It means the functioning of the anti-phishing tool should be fast enough which is only possible when the programming codes remain precise and easy to execute the tool.

6. PHISHING FEATURES

In the previous study, researchers have suggested a number of anti-phishing system models to find the solution of phishing [30-36]. These system models do not show more than 85 percentage successful result [37-41]. In some cases, the system tools are showing only 50-60 percentage successful result. A. Martin et.al. [42] worked on 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 [43-47].

The selection of phishing features is very important part of the research study of phishing. Here, 19 phishing features have been selected and categorized in five different groups on the basis of their nature. These phishing features are;

ISSN- 2350-0530(O) ISSN- 2394-3629(P)
IF: 4.321 (CosmosImpactFactor), 2.532 (I2OR)

URL & Domain Identity

1 Using IP address
2 Abnormal URL request
3 Abnormal URL of anchor
4 Abnormal DNS record
5 Abnormal URL

Security & Encryption

1 Using SSL certificate
2 Certificate authority
3 Abnormal cookies
4 Distinguished names certificate

Source code & java script

1 Redirect pages
2 Straddling attack
3 Pharming attack
4 OnMouseOver to hide the link
5 Server form handler

Page style & Contents

1 Spelling errors
2 Copying website
3 Using forms with Submit button
4 Using pop-up windows
5 Disabling right click

Web address bar

1 Long URL address
2 Replacing similar characters for URL
3 Adding a prefix or suffix
4 Using @ symbol in web address
5 Using the hexadecimal character codes

Social human factors

1 Emphasis on security
2 Public generation salutation
3 Buying time to access accounts

7. CRITERIA OF URL, CONTENT AND IMAGE MATCHING

When user wish to access webpage, a web URL is entered on web browser or user can directly reached to the target webpage from any other website referencing tags. In this case, first of all the URL and its contents should be checked then the contents and existing images should be checked [48]. To check various points of the website takes enough time to cross check the website information with the database source of the Add-on. In the earlier study, browser-based

ISSN- 2350-0530(O) ISSN- 2394-3629(P)
IF: 4.321 (CosmosImpactFactor), 2.532 (I2OR)

client-side solutions have been proposed to mitigate the phishing attacks [49, 50]. Some techniques have also been developed which attempt to prevent phishing mails which are being delivered [51, 52]. So we should have a system that can check fast and accurately the entered information of user with the database information. The phishing features has been selected from the previous study [53-55] and catorized as per their nature.

On the basis of different case conditions of a possible phishing webpage, the phishing features are defined at different group systems with different case conditions. The following Table 1 shows the evaluation criteria to find phishing in which the phishing criteria are defined at different assigned servers namely S1, S2, S3, S4 and S5.

Table 1: Evaluation criteria to find phishing defined at different assigned servers; S1, S2, S3, S4 and S5

Conditions	**S1**	**S2**	**S3**	**S4**	**S5**
Web URL matching	Same	Different	Same	Different	Different
No. of dots "." >	2	2	3	3	4
No. of "@" >	0	1	0	1	0
No. of "//" >	0	1	0	1	1
Port Number in the URL	Yes	No	Yes	No	No
Title Tag matching with URL	Yes	No	Yes	No	No
No. of Image Tags	Same	Different	Different	Different	Same
A Tag Data	Same	Different	Same	Different	Different
A Tag in URL	Same	Different	Different	Same	Same
Login/Password field	No	Yes	No	Yes	Yes
Website contents matching	Same	Different	Same	Same	Different
Client IP	Same	Different	Same	Same	Different
No. of Links functioning in webpage	Yes	Yes	No	No	Yes
CSS Class functioning	Yes	Yes	No	No	Yes
IDs of Control functioning	No	Yes	No	Yes	No
Approved Date by system	Yes	No	Yes	No	No
Approved IP	Yes	No	Yes	No	Yes
Is Online	Yes	No	No	Yes	No
NS1	Same	Different	Same	Same	Different

Apart from this selection of phishing features, the domain age can also be fined for the website from *www.domaintools.com*. By using this website, we can find the information about the website, like when it is created and how long it would be exist. Some of the governmental authorities are also working on this concept of finding phishing for achieving better solution to protect the user from electronic fraud. These authorities have already declared many websites as phishing, so in this study, the database source is increased with the help of these authorised sites.

ISSN- 2350-0530(O) ISSN- 2394-3629(P)
IF: 4.321 (CosmosImpactFactor), 2.532 (I2OR)

The phishing features are also studied with the previously defined anti-phishing system models [56].

8. EXPERIMENTAL ANALYSIS

The phishing features have been defined at five different assigned servers, so that when a user hit the website, the concerned server can send the reply to add-on system. In each of the rule based condition, every component is assumed to be one of the three situations (Low, Medium and High) and each situation has different component, which gives the result whether the accessed website is phishing or not. A hit is performed on the web URL '*www.login.yahoo.com*' and results obtained is as given in the Table 2. In the table, S-1, S-2, S-3, S-4 and S-5 are the classification systems assigned for the anti-phishing tool. At the end of the table, the result obtained from all the systems is mentioned in the form of percentage. The websites can be declared as *Phishing* if the percentage is higher than 60, if percentage is between 40 – 60, the website is declared as *highly risky* website and below 10 percentage, the website address would be stored in the anti-phishing tool's database for further checking. After checking 10 percentage suspicious conditions, the website would be declared as phishing or legitimate to alert the user for further accessing of the same website.

Table 2: Result received from proposed system tool for the website www.login.yahoo.com

Group 1	**Details**
Web URL	https://login.yahoo.com/
No. of DOT(.) in URL	2
No. of @ in URL	0
No. of // in URL	1
PORT in URL ?	NO
Group 2	**Details**
Title Tag	Yahoo -
No. of Image tag	2
a Tag Data	231
Group 3	**Details**
Login/Password	Yes
Site Contents Matching	Yes
Client IP	122.168.204.201
No. of Links on the page	233
Group 4	**Details**
CSS Class functioning	Yes
IDs of Control functioning	Yes
Group 5	**Details**
Approved Date by System	12 December 2014
Approved IP	Host IP
Is Online	False
NS1	NS1.YAHOO.COM
NS2	NS2.YAHOO.COM

ISSN- 2350-0530(O) ISSN- 2394-3629(P)
IF: 4.321 (CosmosImpactFactor), 2.532 (I2OR)

On the basis of above outcomes, the rules are defined according to the feature present on the webpage. The following Table 3 shows the risk status of the accessing website '*www.login.yahoo.com*'.

Table 3: Phishing Characteristics and the Risk Status of the Accessed Website (www.login.yahoo.com)

Phishing Characteristics	**Assigned System**	**Features Present**				
		S-1	**S-2**	**S-3**	**S-4**	**S-5**
Web URL matching	1	0	-1	-1	-1	-1
Number of dots ' . ' present in the URL		1	1	1	1	1
Number of '@' present in the URL		0	1	0	1	0
Number of ' // ' present in the URL		1	1	1	1	1
Port Number in the URL		-1	1	-1	1	1
Title Tag of the URL matching	2	1	-1	1	-1	-1
Image Tags on the web page		1	1	1	1	-1
Anchor Tag data		-1	1	-1	1	1
Anchor Tags in URL		-1	1	-1	-1	-1
Login/Password field present or not	3	1	-1	1	-1	-1
Webpage contents matching		1	-1	1	1	-1
Client IP of the webpage		1	-1	1	1	-1
No. of links functioning in web page		1	1	-1	-1	1
CSS Class functioning or not	4	1	1	-1	-1	1
Whether the IDs of control functioning or not		1	-1	1	-1	1
Approval date by anti-phishing system	5	1	-1	1	-1	-1
Approved IP		1	-1	1	-1	-1
Website is online or not		1	-1	-1	1	1
NS1		1	1	1	1	-1
NS2		1	-1	1	-1	-1
Rating (in %)		**16**	**53**	**37**	**47**	**64**
Website Class		**R**	**HR**	**R**	**HR**	**P**

Where L – Legitimate, HR – Highly Risky, R – Risky, P – Phishing

9. PERFORMANCE ANALYSIS OF THE PROPOSED SYSTEM

There is a number of anti-phishing tools have been proposed in earlier study to protect the user from phishing attack. The previous study is based on the functioning of anti-phishing tools with a number of data mining techniques which are analysed to solve the phishing problem [57-59]. Earlier study shows that the performance of classification techniques is affected by the type of data sets used and the way in which the classification algorithms have been implemented in the toolkit. The WEKA (Waikato Environment for Knowledge Analysis) data mining toolkit shows the better performance as compared to other data mining comparing tools [60]. The WEKA is

ISSN- 2350-0530(O) ISSN- 2394-3629(P)
IF: 4.321 (CosmosImpactFactor), 2.532 (I2OR)

designed to solve the data mining algorithm problems, which is an open Java source code that includes implementations of different methods for several different types of data mining tasks such as clustering, classification, association rules and regression analysis. Here, three data mining algorithms have been discussed under WEKA Version 3.6.

The database contents that Weka support is .ARFF (Attribute Related File Format) which is given below for the website *www.login.yahoo.com*. In Weka, initially attributes have been defined, so 19 attributes (based on phishing features) have been taken in the study and the last one is the attribute taken for the result. The analyser calculates the result only in the form of 0, 1 and -1. Here in the study of phishing, 1 is assigned to the 'positive' result, 0 denote 'no relation' and -1 show the 'negative' result.

```
@relation phishing

@attribute Web_URL  { 1,0,-1 }
@attribute No_of_._in_URL { 1,-1 }
@attribute No_of_@_URL { 1,0,-1 }
@attribute No_of_//_URL { 1,-1 }
@attribute Port_Number_URL { 1,-1 }
@attribute Title_Tag_matching {1,-1}
@attribute No_of_Image_Tags {-1,0,1}
@attribute A_Tag_Data {1,0,-1}
@attribute A_Tag_URL {1,-1}
@attribute Login_Password_field {1,-1}
@attribute Website_contents_matching {1,-1}
@attribute No_Links_functioning_webpage {1,-1}
@attribute CSS_Class_functioning {1,-1}
@attribute IDs_Control_functioning {-1,0,1}
@attribute Approved_Date_system {1,-1}
@attribute Approved_IP {-1,1}
@attribute Is_Online {-1,1}
@attribute NS1 {-1,1}
@attribute NS2 {1,-1}
@attribute Result {1,-1}

@data
0,1,0,1,-1,1,1,-1,-1,1,1,1,1,1,1,1,1,1,1,1
-1,1,1,1,1,-1,1,1,1,-1,-1,-1,1,1,-1,-1,-1,-1,1,-1
-1,1,0,1,-1,1,1,-1,-1,1,1,1,-1,-1,1,1,1,-1,1,1
1,1,1,1,1,1,1,-1,1,1,-1,-1,1,1,-1,-1,-1,-1,-1,1,-1
-1,1,0,1,1,-1,-1,1,-1,-1,-1,-1,1,1,1,-1,-1,1,-1,-1
1,1,0,1,-1,1,1,-1,1,1,1,1,-1,-1,1,1,1,1,1,1
-1,1,1,1,1,-1,1,1,1,1,-1,-1,1,1,1,-1,-1,-1,1,-1
-1,1,0,1,-1,1,1,-1,-1,1,1,1,-1,1,1,1,1,-1,1,1
-1,1,1,1,1,1,1,-1,1,1,-1,-1,1,1,-1,1,-1,1,-1,1,-1
-1,1,0,1,1,-1,-1,1,-1,-1,-1,-1,1,1,1,1,1,1,-1,-1
```

The performance of the algorithms can be measured with the use of classification accuracy metric. The accuracy of the data set can be calculated by the percentage of correctly classified websites from the given data set.

In the proposed system, the system tool sends the website information to 5 different assigned servers to check the status/category of the websites. Theses servers are categorised by different

ISSN- 2350-0530(O) ISSN- 2394-3629(P)
IF: 4.321 (CosmosImpactFactor), 2.532 (I2OR)

phishing features which are *Character based, Coding based, Identity based, Contents based and Attribute based*. The applied data mining algorithms shows the result for the proposed system in which 8540 legitimate and 4480 phishing websites has been checked. The database of phishing and legitimate websites is collected from APWG [61] and PhishTank [62]. These websites are collected in 10 different days for the month of November and December, 2015. Since APWG and PhishTank are the trusted and reliable source, which keeps all the information about legitimate and phishing websites, are very helpful in the research study.

10. ALGORITHMS for FEATURE SELECTION

The performance of the proposed system is tested with three different data mining classification algorithms; Random Forest (RF), Nearest Neighbour Classification (NNC) and Bayesian Classifier (BC). Since, all these algorithms work differently and cover almost all the areas of data mining problems, so the study of these algorithms for checking the performance of anti-phishing tool gives better result. Following is the brief description of these algorithms;

1) *Random Forest*, It is one of the best algorithm for classification problems which is able to classify large amount of datasets with accuracy. The algorithm is a combination of tree predictors in which each tree depends on the values of a random vector sampled independently. The basic concept of this algorithm is that a group of "weak learners" can come together to form a "strong learner".
2) *Nearest Neighbour Classification (NNC)*, It is one of the data mining algorithms that stores all available cases of the problem and classifies new cases based on a similarity measure. The classes are defined with numeric value which is called K.
3) *Bayesian Classifier (BC)*, It is a well know algorithm for studying the matter of phishing. To apply the Bayesian filter to find phishing websites, two datasets are required; legitimate website details and phishing website information. A large data set of legitimate transactional website is needed because the set of websites mostly resembles just like phishing websites and the filter must have numerous examples of legitimate transactional websites to achieve a low false positive rate.

11. RESULTS AND DISCUSSION

To study the performance of above mentioned data mining algorithms, consecutive hits has been done on the web browser for a number of legitimate and phishing websites which are collected from different authentic sources. After hitting websites, the Add-on system sends the response to assigned servers. The assigned servers cross check the website details with the database source and send the response to the main server. All the assigned servers keep the record of hitting websites. Figure 1 shows the snap shot of WEKA Explorer in which all the phishing features has been taken in the study. The figure shows pre-process configuration of classification algorithm filter that are showing 20 attributes and 10 instances for any outcome.

ISSN- 2350-0530(O) ISSN- 2394-3629(P)
IF: 4.321 (CosmosImpactFactor), 2.532 (I2OR)

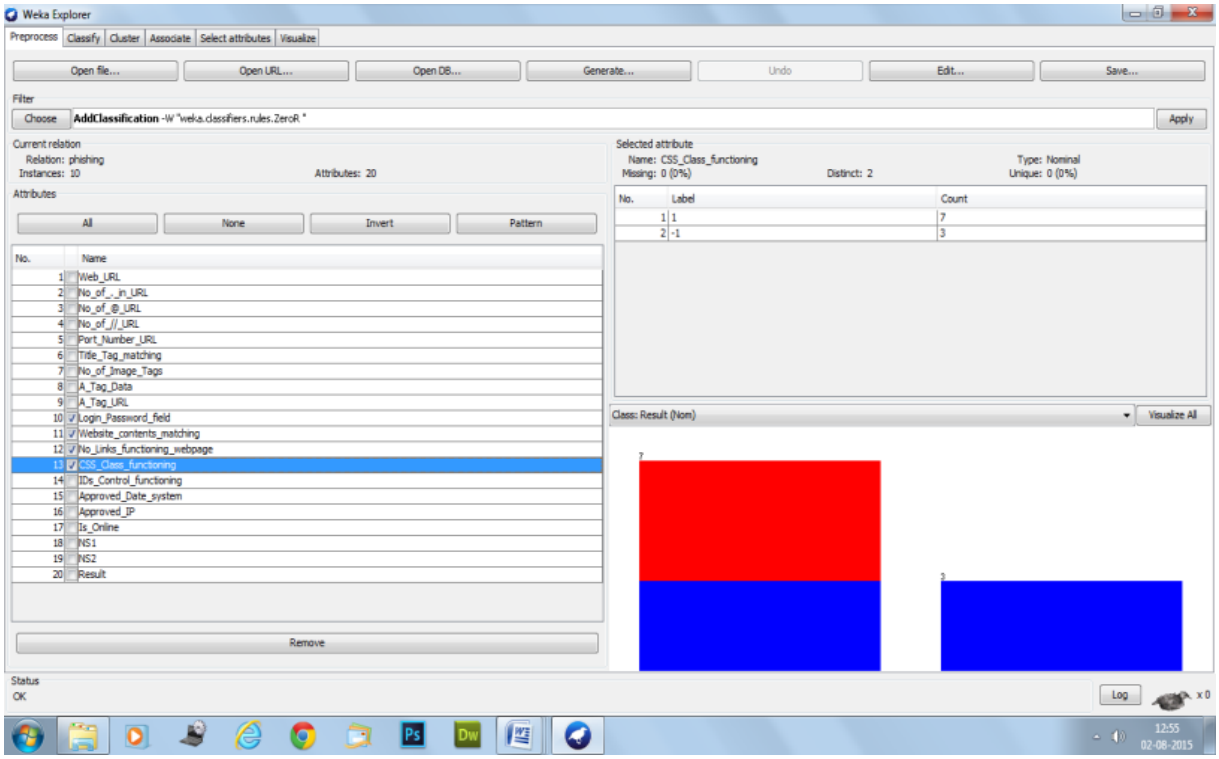

Figure 1: WEKA Explorer showing all the 19 phishing features taken in the study

At 2, 5 and 10 fold, the algorithms have been tested with 75 and 66 percentage split of data. The testing option 10-fold validation shows better performance than other percentage split cases. When the training data size is small, the system tool functions well. For larger data sets, this result slightly decreases. By using pruning method in a classification algorithm, results achieved with higher accuracy and get better performance as compared to mining the data without pruning. If we test the large dataset, a large decision tree needs to prepare which result in longer computation time. Table 4 and 5 shows the phishing training data set tested with Weka software with 75 and 66 percentages split test condition.

Table 4: Data mining algorithm's performance at 75 percentage split of training data set

Algorithms	**No. of Folds**	**Percentage Split**	**Accuracy Rate (%)**	**Error Rate (%)**	**Unclassified Rate (%)**
Random Forest	2	75	68	15	17
Nearest Neighbour	5	75	74	12	14
Bayesian	10	75	88	6	6

ISSN- 2350-0530(O) ISSN- 2394-3629(P)
IF: 4.321 (CosmosImpactFactor), 2.532 (I2OR)

Table 5: Data mining algorithm's performance at 66 percentage split of training data set

Algorithms	**No. of Folds**	**Percentage Split**	**Accuracy Rate (%)**	**Error Rate (%)**	**Unclassified Rate (%)**
Random Forest	2	66	48	22	30
Nearest Neighbour	5	66	62	21	17
Bayesian	10	66	82	9	9

The performance of the Random Forest and Nearest Neighbour algorithms were almost similar on all kinds of data sets, whereas the Bayesian algorithm is slightly better in different case conditions. In almost all the conditions, the cross-validation data test method has a better performance.

If the data set is defined for more than 500 instances that is treated as a large data set, then we can say that the large data sets perform better result. In the study of Random Forest for large dataset, it is found that it builds the largest trees, which causes lowest overall performance. Out of three, two algorithms uses reduced-error pruning method that build approximately equally sized trees which is large enough. The Bayesian algorithm builds the smallest trees. This indicate that the cost-complexity pruning reduce to smaller trees than reduced error pruning. The Bayesian algorithm performs better result on data sets having many numerical attributes. It is also noticed for achieving better performance for all the three algorithms, the data sets with few numerical attributes shows better performance.

12. CONCLUSION

In this paper, three different data mining algorithms have been discussed for the analysis of anti-phishing website data sets. Theses algorithms are Random Forest (RF), Nearest Neighbour Classification (NNC), Bayesian Classifier (BC). The *Random Forest* shows around 68 percentage of successful result when the training data is split to 75 percentage. If the database is already available for testing, the algorithm shows better result but in case of finding on-spot hitting, this algorithm is not well suited. The *Nearest Neighbour Classification* technique gives better and accurate result when the checking conditions are less. The result of *Bayesian Classification* shows the accuracy rate is around 88 percentage for finding the phishing websites. With the comparison of all these algorithms, the Bayesian classification is more accurate and shows fast response to the system.

13. ACKNOWLEDGMENT

We would like to acknowledge the Anti-Phishing Working Group for providing database of phishing and anti-phishing websites and my colleague who encourage me to find more and more data on this topic. My sincere profound gratitude to the research committee for providing me the

ISSN- 2350-0530(O) ISSN- 2394-3629(P)
IF: 4.321 (CosmosImpactFactor), 2.532 (I2OR)

guidance for data collection, valuable guidance, suggestions and encouragement throughout the work.

14. *REFERENCES*

[1] *APWG 1 to 3rd Quarter 2015 Phishing Activity Trends Report from www.antiphishing.org*
[2] *A research report from http://securityresearch.in/ ?ubiquitous_id=88, January 2013*
[3] *A.Naga Venkata Sunil, Sardana A., "A PageRank Based Detection Technique for Phishing Web Sites", 2012 IEEE Symposium on Computers & Informatics, 2012, pp. 58-63*
[4] *Javelin Strategy and Research. http://www.javelinstrategy.com, 2012*
[5] *Chou N., LedesmaR., Teraguchi Y. and Mitchell John C. "Client-Side Defense Against Web-Based Identity Theft" in 11th Annual Network and Distributed System Security Symposium, San Diego, February, 2004*
[6] *Dhamija R., Tygar J.D., "The Battle against phishing: Dynamic Security Skins. In: Proc. of ACM Symposium on Usable Security and Privacy, 2005, pp.77-88*
[7] *A Report from 'Computer Associate Internationals Inc.', September 2012*
[8] *Khonji M., JonesA., IraqiY., "A Novel Phishing Classification based on URL Features", 2011 IEEE GCC Conference and Exhibition (GCC), February 19-22, 2011, Dubai, United Arab Emirates, 2011, pp. 221-224*
[9] *Wardman B., Stallings T., Warner G., Skjellum A., "High-Performance Content-Based Phishing Attack Detection", published in IEEE conference on eCrime Researchers Summit (eCrime), 2011, pp. 1-9*
[10] *Weider D. Yu, Nargundkar S.,Tiruthani N., "PhishCatch – A Phishing Detection Tool", presented in 33rd Annual IEEE International Computer Software and Applications Conference, IEEE Computer Society, 2009, pp. 451-456*
[11] *Prakash P., Manish K., Kompella R.R., Gupta M., "PhishNet: Predictive Blacklisting to Detect Phishing Attacks", presented as part of the Mini-Conference at IEEE INFOCOM 2010*
[12] *IsredzaRahmi A Hamid and Abawajy Jemal H., "Profiling Phishing Email Based on Clustering Approach" 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, 2013, pp. 629-635*
[13] *Jiang H., ZhangD., Yan Z., "A Classification Model for Detection of Chinese Phishing E-Business Websites", PACIS2013Proceedings. 2013, Paper 152*
[14] *Li T., HanF., Ding S.and ChenZ., "LARX: Large-scale Anti-phishing by Retrospective Data-Exploring Based on a Cloud Computing Platform", Computer Communications and Networks, Proceedings of 20th International Conference on, July 31-August 4, , 2011, pp. 1-5*
[15] *Huang H., Zhong S., TanJ., "Browser-side Countermeasures for Deceptive Phishing Attack", 2009 Fifth International Conference on Information Assurance and Security, IEEE Computer Society, 2009, pp. 352-355*
[16] *Ferguson Edward, Weber Joseph, and Hasan Ragib, "Cloud Based Content Fetching: Using Cloud Infrastructure to Obfuscate Phishing Scam Analysis", IEEE Eighth World Congress on Services, IEEE Computer Society, 2012, pp. 255-261*

ISSN- 2350-0530(O) ISSN- 2394-3629(P)
IF: 4.321 (CosmosImpactFactor), 2.532 (I2OR)

[17] *Microsoft Corporation. Internet Explorer 7. http://www. microsoft.com/windows/ie/default.mspx, Accessed: November 9, 2010*
[18] *Aburrous Maher, Khelifi Adel, "Phishing Detection Plug-In Toolbar Using Intelligent Fuzzy-Classification Mining Techniques", The International Journal of Soft Computing and Software Engineering [JSCSE], Vol. 3, No. 3, pp. 54-61, March 2013*
[19] *Mahmood Ali M., Dr. Rajamani L., "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*
[20] *Chou N., Ledesma R., Teraguchi Y., Boneh D., and Mitchell J.C., "Client-side Defense Against Web-based Identity Theft" In Proc. Network and Distributed System Security Symposium, San Diego, CA., 2004*
[21] *CallingID, Ltd. http://www.callingid.com/DesktopSolutions/ CallingIDToolbar.aspx, Accessed: December 1, 2008*
[22] *Cloudmark, Inc. http://www.cloudmark.com/desktop/ download, Accessed: September 5, 2008*
[23] *EarthLink, Inc. EarthLink Tool. http://www.earthlink.net /software/free/tool/, Accessed: November 9, 2010*
[24] *eBay, Inc. Using eBay Tool's Account Guard, Accessed: June 13, 2010, http://pages.eBay.com/help/confidence/accountguard.html*
[25] *Kerner, Michael S., Firefox 2.0 Bakes in Anti-Phish Antidote. Internet News. http://www.internetnews.com/devnews/ article.php/3609816.2006*
[26] *Google, Inc. Google Safe Browsing for Firefox. http://www.google.com/tools/firefox/safebrowsing/, Accessed: June 13, 2010*
[27] *Netcraft. Netcraft Anti-Phishing Tool. http://tool.netcraft.com/, Accessed: June, 13, 2010*
[28] *Netscape Communications Corp. "Security Center" Accessed: November 9, 2006. http://browser.netscape.com/ns8/product /security.jsp*
[29] *Quick Start : Spoof Guard, A http://crypto. stanford.edu/SpoofGuard/, October 10, 2011*
[30] *Jiang Hansi, Zhang Dongsong, Yan Zhijun, "A Classification Model for Detection of Chinese Phishing E-Business Websites", PACIS 2013 Proceedings. Paper 152, 2013.*
[31] *Zhuang Weiwei, Jiang Qingshan, Xiong Tengke, "An Intelligent Anti-phishing Strategy Model for Phishing Website Detection", IEEE Computer Society, 32nd International Conference on Distributed Computing Systems Workshops, 2012.*
[32] *Balamuralikrishna T., Raghavendrasai N., Satya Sukumar M., "Mitigating Online Fraud by Ant phishing Model with URL & Image based Webpage Matching", International Journal of Scientific & Engineering Research, Vol. 3, Issue 3, March-2012, pp.1-6*
[33] *Madhuri S. Arade, Bhaskar P.C., Kamat R.K., "Antiphishing Model with URL & Image based Webpage Matching", International Conference & Workshop on Recent Trends in Technology (TCET), Proceedings published in International Journal of Computer Applications® (IJCA), 2012, pp 18-23*
[34] *Aburrous Maher, Hossain M.A., Dahal Keshav, Thabatah Fadi, "Modelling Intelligent Phishing Detection System for e-Banking using Fuzzy Data Mining", IEEE Computer Society, International Conference on CyberWorlds, pp. 265-272, 2009*
[35] *Zhuang W., Ye Y., Li T., Jiang Q. "Intelligent phishing website detection using classification ensemble Systems" Engineering Theory & Practice, Volume 31(10), 2011, P2008-2020*

ISSN- 2350-0530(O) ISSN- 2394-3629(P)
IF: 4.321 (CosmosImpactFactor), 2.532 (I2OR)

[36] *Kang JungMin, DoHoon Lee. "Advanced White List Approach for Preventing Access to Phishing Sites", International Conference on Convergence Information Technology (ICCIT 2007), 2007, pp.491–496*

[37] *Abbasi Ahmed, "Mariam" Zahedi Fatemeh and Chen Yan, "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.*

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

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

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

[41] *Grazioli S. and Jarvenpaa S.L., "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), 2000, pp. 395-410*

[42] *Martin A., Anutthamaa Na.Ba., Sathyavathy M., Marie Manjari Saint Francois, Dr. Venkatesan Prasanna, "A Framework for Predicting Phishing Websites Using Neural Networks", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 2, 2011, pp. 330-336*

[43] *Aburrous Maher, Hossain M.A., Dahal Keshav, Thabtah Fadi, "Intelligent phishing detection system for e-banking using fuzzy data mining", Expert Systems with Applications: An International Journal, Vol. 37 Issue 12, December, 2010.*

[44] *Zhang, H., Liu, G., Chow, T., and Liu. W., "Textual and Visual Content-Based Anti-Phishing: A Bayesian Approach", IEEE Transactions on Neural Networks, 22(10), 2011, 1532–1546*

[45] *Herzberg A. and Jbara A. "Security and identification indicators for browsers against spoofing and phishing attacks", ACM Transactions on Internet Technology, 8(4), 2008, pp.1-36*

[46] *Prakash P., Kumar M., Kompella R.R., and Gupta M., "Phish-Net: predictive blacklisting to detect phishing attacks" in IEEE INFOCOM Proceedings. San Diego, California, USA: IEEE, March, 2010, pp. 1–5*

[47] *Garera S., Provos N., Chew M. and Rubin A.D., "A framework for detection and measurement of phishing attacks" Alexandria, Viriginia, USA: ACM, 2007, pp. 1–8*

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

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

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

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

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

ISSN- 2350-0530(O) ISSN- 2394-3629(P)
IF: 4.321 (CosmosImpactFactor), 2.532 (I2OR)

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

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

[55] *Zhang Y., Hong J., and Cranor L., "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*

[56] *Garera S., Provos N., Chew M., "A Framework for Detection and Measurement of Phishing Attacks", In: Proc. of the 5th ACM Workshop on Recurring Malcode, 2007, pp.1-8*

[57] *Raffetseder Thomas, Kirda Engin, and Kruegel Christopher, "Building Anti-Phishing Browser Plug-Ins: An Experience Report", SESS '07 Proceedings of the Third International Workshop on Software Engineering for Secure Systems, IEEE Computer Society Washington, DC, USA ©2007, p.6*

[58] *Aburrous Maher, Hossain M.A., Dahal Keshav, Thabtah Fadi, "Predicting Phishing Websites using Classification Mining Techniques with Experimental Case Studies", Seventh International Conference on Information Technology, IEEE Computer Society, 2010, pp. 176-184*

[59] *Wedyan Suzan, Wedyan Fadi, "An Associative Classification Data Mining Approach for Detecting Phishing Websites", Journal of Emerging Trends in Computing and Information Sciences, Vol. 4, No. 12, 2013, pp. 888-899*

[60] *H. Wahbeh Abdullah, A. Al-Radaideh Qasem, Mohammed N. Al-Kabi, and Emad M. Al-Shawakfa, "A Comparison Study between Data Mining Tools over some Classification Methods", International Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence, 2012, pp. 19-26*

[61] *APWG 4th Quarter 2015 Phishing Activity Trends Report from www.antiphishing.org, 2015*

[62] *Phishing website list from http://www.phishtank.com/, November 2015*