International Journal of Scientific Research in Civil Engineering
© 2018 IJSRCE | Volume 2 | Issue 4 | ISSN : 2456-6667

Phishing Web Pages detection Using Feature Selection and Extraction Method

Ritika Arora[1], Ashok Kumar Arora[2]
[1]Assistant Professor Panjab University SSG Regional Centre Hoshiarpur, Punjab, India
[2]Superintending Engineer, Water Resources Department (Pb. Irrigation),Govt of Punjab, Mohali, Punjab, India

ABSTRACT

Phishing is a security attack that involves obtaining sensitive information as a trustworthy entity. The user tries to steal the confidential information of the web user such as online banking passwords, credit card number and other financial data by making identical website of legitimate one in which the contents and images almost remains similar to the legitimate website with small changes. In this paper, a number of anti-phishing toolbars have been discussed and proposed a system model to tackle the phishing attack. 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. The motivation of our study is to propose a safer framework for detecting phishing websites with high accuracy in less time.

Keywords : Phishing, Anti-Phishing , Add-on For Web Browser, Data Mining Classification Algorithms.

I. INTRODUCTION

A number of government and private authorised agencies are working on the topic of phishing and the countermeasure the phishing attacks.The APWG(Advanced Phishing Working Group) and PhishTank are two prominent agencies which keeps all the information related to phishing and legitimate websites. 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 secure system environment for electronic business transactions. So the research study is very much necessary to reduce the online transaction problems. With due to rapid increase in the use of internet technology for communication different kind of attacks can be possible on the network such as DOS (denial of service attack), masquerade, replay and phishing etc. It is one of the most serious attacks which steals our personal information or hackthe website.[1] 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

II. METHODS OF PHISHING ATTACKS

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 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.[2]

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 number regarding problems with their bank account information.

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.

DNS-Based Phishing ("Pharming"): Pharming is the term given to hosts file modification. This type of phishing is alsocalled DNS-based phishing. In this type of phishing, the phisher tamper with a company's host files or the DNS so that requests for URLsor name services return a bogus address and subsequentcommunications 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.

Hosts File Poisoning When user enters the URL to visit the website, hackers will look up the host names and transmit the bogus address that look alike an original website and their information will be stolen.

Content-Injection Phishing In this type of attack hackers will replace the original content with the fake content in the website which misdirects the user to give their sensitive information.

Web Trojans They collect the users information and transmit them to the phisher. This will happen at the time of login by the user.

Man-in-the-Middle:In this hacker will be in between the user and the website. Whenever user enters their information hackers will take the information without causing interruption to the users. Later on hackers will use this information when the user is not active on the system.

Data Theft Sensitive data's will be stored in Pcs. These data's will be taken by the victims without knowing to the user. Commonly, this information is user information such as passwords, social security numbers, credit card information, other personal information, or other confidential corporate information By stealing confidential communications, design documents, legal opinions, employee related records, etc., thieves profit from selling to those who may want to embarrass or cause economic damage or to competitors.

III. ANTI PHISHING TECHNIQUES

Various anti-phishing techniques have been evolved to protect our website/ link and personal information against phishing attacks.

A) List Based Approach

This is possibly the most straightforward solution for anti-phishing. A white list contains URL's of known legitimate sites. Many current anti-phishing techniques rely on the combination of white list and blacklist. The representative blacklist/white list based systems include Phish Tank Site Checker, Google Safe Browsing, Fire Phish and Calling ID Link Advisor. This anti-phishing result would generally deploy similarly as toolbars or extension of web browsers should remind those clients if they would scan a sheltered websites. Blacklist undergo from a window of vulnerability between the time a phishing site is launched and the site's addition to the blacklist as it requires frequent updating which is the case for white list also.

B) Ant Colony Optimization

The Ant Colony System or the basic idea of an ant food searching system is illustrated in Fig. 2. In the left picture, the ants shift in a straight row to the food. The subsequent picture illustrates the circumstances rapidly after an obstacle is inserted among the nest and the food. To evade the obstacle, initial each ant selected to turn right or left at random. Let us presuppose that ants shift at the identical speed depositing pheromone in the trail equivalently. Though, the ants that, by possibility, prefer to turn right will reach the food sooner, although the ants that go around the obstruction turning right will pursue a longer path, and hence will take long time to circumvent the impediment. As a consequence, pheromone gathered quicker in the shorter path around the impediment. Ever since ants desire to pursue tracks with better amounts of pheromone, eventually all the ants congregate to the shorter path around the impediment .

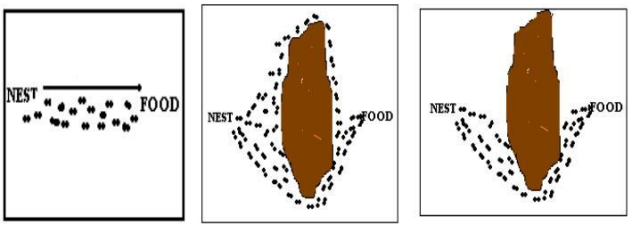

Figure 2 . Depicts the behavior of real ant movement

This novel heuristic known as Ant Colony Optimization (ACO) has been found to be mutually vigorous and multipurpose in handling an extensive range of combinatorial optimization problems. The major suggestion of ACO is to model a predicament as the search for a least cost path in a graph. Artificial ants as if walk on this graph, gazing for cheaper paths. Each ant has a somewhat uncomplicated behavior accomplished of finding comparatively costlier paths.

Cheaper pathways are found as the growing consequence of the universal cooperation among ants in the colony. The behavior of artificial ants is stimulated from real ants: they put down pheromone trails (noticeably in a mathematical outline) on the graph edges and prefer their path with reverence to probabilities that depend on pheromone tracks. These pheromone tracks progressively abridged by evaporation. In addition, artificial ants have a few superfluous attributes not seen in their counterpart in real ants. In meticulous, they subsist in a discrete world (a graph) and their progresses consist of conversions from nodes to nodes[3].

The ACO fluctuates from the conventional ant system in the intellect that here the pheromone tracks are updated in two ways. Initially, while ants build an excursion they nearby transform the quantity of pheromone on the visited edges by a narrow updating role. Subsequently, after all the ants have fabricates their personage tours, a global updating rule is applied to transform the pheromone level on the boundaries that belong to the preeminent ant tour found so far .

C) PhishZoo

It can detect current phishing sites if they look like legetimate sites by matching their content against a saved profile. In order to avoid detection, a phishing site must gaze fundamentally unique in relation to a genuine website. Our working assumption is that such different-looking sites have a better chance of catching users' attention about their phishiness. Branding is an issue that is well-studied in the marketing literature, and, with PhishZoo, it can be used to improve security as opposed to the current case, when this branding is co-opted by attackers to mis use client trust [11].

D) K-NearestNeighbor (k-NN)

This Classifier proposed for phishing email filtering. Using this classifier, the decision is made as follows: based on k-nearest training input, samples are chosen using a pre-defined similarity function; after that, the email x is labeled as belonging to the same class as the bulk among this set of k [13].

E) Information-flow-based approaches

PwdHash is a well-known anti-phishing solution in literature It generates domain-specific passwords that are rendered unusable if they are submitted to another domain (e.g., a password for www.hotmail.com will be different if submitted to www.phisher.com).In comparison, Antiphish takes an alternate methodology and stay with track about the place sensitive data is, no doubt submitted [8]. That is, if it detects that confidential information such as a password is being entered into a form on a fake web site, a warning is generated and the pending operation is canceled. The main disadvantage of AntiPhish is that it requires user interaction to specify which sensitive information should be captured and monitored. Later, the author significantly improves the original idea of AntiPhish by eliminating the necessary user interaction with an extra comparison step that analyzes the DOM structure of the pages [9]. They present an extension of AntiPhish, called DOMAntiPhish, which leverages design similitude majority of the data should recognize between pernicious furthermore favorable pages.

F) Attribute Based Anti-Phishing Techniques

Attribute-based anti-phishing strategy uses both reactive and proactive anti-phishing. This technique has been implemented in Phish Bouncer [15] tool. The Image Attribution technique does a comparison of images of accessing site and the sites already being registered with phish bouncer. The HTML Crosslink checks and looks at the responses coming from nonregistered sites and counts the number of links the page has to any of the registered sites. A high number of cross-links indicate that it is a phishing site. In false info feeder checker, false information is provided and if that information is accepted by the site ,then probably that link is phished one. It checks for suspicious certificates and validates site certificates presented during SSL handshake and extends the typical Usage by looking for Certification Authority (CA).As multiple checks are performed to authenticate the site this results in slow response time .

F) Fuzzy Logic

Fuzzy logic has been exercised for decades in the engineering sciences to entrench specialist input into computer models for a wide range of applications. It suggests a promising unusual for measuring operational risks [7]. The fuzzy logic techniques presents more information to help risk managers successfully manage assessing and ranking website phishing risks than the existing qualitative approaches as the risks are quantified based on a amalgamation of historical data and practiced input. The benefit of the fuzzy system is that it enables processing of indistinctly defined variables, and variables whose relationships cannot be defined by mathematical relationships. Fuzzy logic can integrate expert human judgment to describe those variable and their relationships.[10]

G) Genetic Algorithm

Genetic algorithms can be used to develop simple rules for preventing phishing attacks. These rules are used to differentiate normal website from anomalous website. These anomalous websites refer to events with probability of phishing attacks [8]. The rules saved in the rule base are usually in the following form:

if { condition }
then
{ act}

For the problems we presented above, the condition generally refers to a match between the URL of the current website link in the e-mail and the rules in PADPS (Phishing Attack Detection and Prevention System), which indicates the probability of phishing attack. The act field usually refers to an action defined by the security policy such as reporting an alert to the browser, through the status field. For example, a rule can be defined as:

if
{
The IP address of the URL in the received e-mail finds any match in the Ruleset
}
then
{
Phishing e-mail
}

This rule can be explained as follows: if there exists an IP address of the URL in e-mail and it does not match the defined Rule Set for White List then the received mail is a phishing mail; so the status is phishing e-mail. The final objective of applying GA is to generate rules that match only the anomalous URLs of websites. These rules are tested on historical URLs and are used to filter new URLs to find suspicious phishing attacks.

IV. 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 algorithm (ProEP) demonstrates promising results with the Ratio Size rules for selecting the optimal number of clusters. Zhang H. et.al. presented a framework which 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. [12] has proposed an offline phishing detection system named Large-scale Anti-phishing by Retrospective data-eXploration (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. 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. 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 [11].

Mahmood Ali M. et.al. 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.

V. 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 [13]

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 .

The Cloudmark Anti-Fraud toolbar is based on the user's ratings . 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 . 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]. 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 . 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 ."

The Netcraft Anti-Phishing Toolbar uses several methods to determine the legitimacy of a web site . 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 . 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" . 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 th email sends the user to a malicious "spoof" site that collects user's information such as account names, password and 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.

VI. 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 . 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 client-side solutions have been proposed to mitigate the phishing attacks . Some techniques have also been developed which attempt to prevent phishing mails which are being delivered . 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 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.

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 authorized sites.

VII. 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 . 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 . The WEKA 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
-1,1,0,1,1,-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
-1,1,0,1,1,-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.

phishing features which are Character based, Coding based, Identity based, Contents based and Attribute based. Thapplied 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 and PhishTank . 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.

VIII. 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.

IX. 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.

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 shows the phishing training data set tested with Weka software with 75 and 66 percentages split test condition

Table 1

Algorithm	No.of Folds	Percentage Split	Accuracy Rate (%)
Random Forest	2	75	68
Nearest Neighbour	5	75	74
Bayesian	10	75	88

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.[14]

X. 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.

XI. 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 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

ABOUT AUTHORS

Mrs.Ritika Arora received her B.Tech. degree in Computer Science Engineering from Kurukshetra University, in 2009,M.Tech. degree in Computer Science Engineering from Shaheed Bhagat Singh State Technical Campus,Ferozepur, in 2012 and is Currently working as Assisstant Professor in Panjab University Regional Centre Hoshiarpur with an experience of 5 Years . Her research interests include Network Security. She has Presented and Published various Research Papers National and International Journals and Conferencs . She is currently member of ISTE AND IAENG

Mr.Ashok Kumar Arora received his B.Sc. Engg. degree in Civil Engineering from Punjab University, Chandigarh in 1981 and Master of Engineering in Irrigation and Hydraulics from PEC , Chandigarh in 1985 and is currently working as Superintending Engineer in Water Resources Deptt (Pb Irrigation) Pb.Govt. Department in Mohali and has experience of 35 years. His research Area includes in Cyber Security.He is fellow of the Institution of Engineers. (India)