Extraction of Features and Classification on Phishing Websites using Web Mining Techniques

Nandhini.S [1], Dr.V.Vasanthi [2]
[1]M.Phil. Scholar, [2]Assistant Professor
[1]Department of Computer Science,
[1]Rathinam College of Arts and Science, Coimbatore, India

Abstract: Website Phishing is serious web security problem that involves mirroring genuine websites to deceive online users in order to steal their sensitive information. Phishing can be seen as a typical classification problem in data mining where the classifier is constructed from large number of website's features. There are high demands on identifying the best set of features that when mined the predictive accuracy of the classifiers is enhanced. This research work investigates features selection aiming to determine the effective set of features in terms of classification performance. We compare features selection and classification methods in order to determine the least set of features of phishing detection using data mining. Experimental tests on large number of features data set have been done using Information Gain and Correlation Features set methods. Further, five data mining algorithms Naïve Bayes, KNN, Random Forest, SVM and j48 have been used to classify the web phishing data set, analyse the results and identify the efficient technique to classify the web page phishing data set.

Keywords: Website Phishing, Classification, Feature Selection, Web Security, Web Mining

I. Introduction

Web Mining is the use of data mining techniques to automatically discover and extract information from Web documents. Figure 1.1 illustrates that Web Mining consists of three parts: Web Content Mining and Structural Mining and Web usage mining.

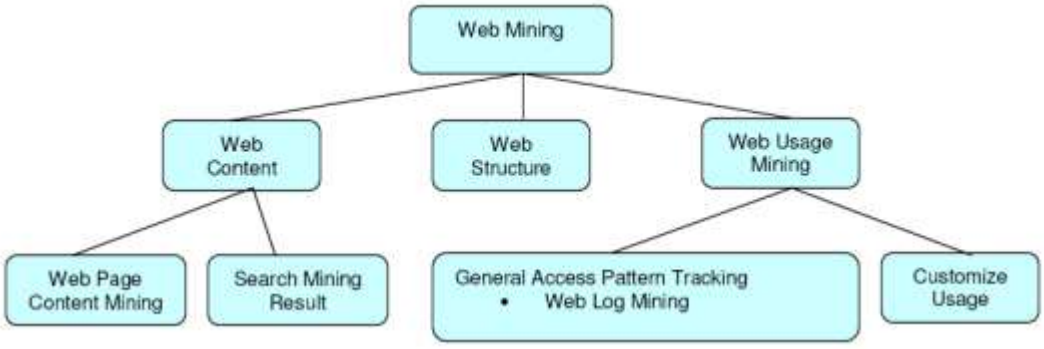


Fig. 1.1 Taxonomy of Web Mining

Web Content Mining

Documents in a website contain textual information which is highly unstructured and varies accordingly to the purpose of a website. Unlike a highly structured database, retrieving information from a collection of documents is a challenging task as the documents do not share a common content representation. Furthermore there is no standard range of attributes available that can be used to distinguish one document from another in a website. In order to extract information from documents, a suitable document representation model is required. One such approach is the bag-of-words (BOW), commonly used to represent a document as a large collection of words.

In BOW, A document's text is broken up into individual tokens or terms. Words that best distinguish a document is then determined by ranking terms using a frequency based method e.g., Term Frequency Inverse Document Frequency.

Web content mining involves applying data mining techniques on text contained in documents to extract useful knowledge. This particular knowledge is commonly used for document classification. Document classification involves creating a topic hierarchy based on all documents of a website by categorizing each document into a topic class it fits best. The topic hierarchy built can then be used for a variety of applications such as web browsing and web crawling.

The majority of current studies rely on a manual classification technique which is, requires feedback from users on each document's topic category and its importance to their information needs. However, manual classification of documents is a tedious and time consuming task. It requires a reasonable number of feedbacks to generate a topic hierarchy and many users do not have the time to tag every document they visit in a website. This suggests the need for a method to automatically discover a document's topic and its relevance to a user's need.

Web Structure Mining

A website consists of a collection of documents interconnected by hyperlinks. Each document may have a number of outgoing and incoming hyperlinks. A hyperlink from document A to document B may represent the continuity between units of information in a website. Figure 2 shows a graph representation of a website, where each node constitutes a document and the edges between the nodes are the hyperlinks connecting the documents of the website. An edge between any two nodes in the graph suggests that the documents may be related and may contain relevant information.

Web structure mining applies data mining techniques on the hyperlink structure and extracts information that can be used for a variety of purpose such as web crawling and ordering a web search result. Crawlers are automated programs that use the hyperlink structure to browse the Web in a systematic manner. A crawler starts with an initial list of URLs to visit and cycles through the list visiting each document in turn. While visiting a document it identifies all hyperlinks and adds them to the list of URLs to visit.

Web Usage Mining

Web usage mining is the application of data mining techniques on large web repositories to discover useful knowledge about users behavioral patterns and website usage statistics that can be used for various website design tasks. The main source of data for web usage mining consists of textual logs collected by numerous web servers all around the world. This is probably because web logs are the easiest and cheapest way of gathering information about the users and a website. Other sources of usage data may include proxy servers and client side logs.

Web logs i.e., web server logs are also commonly referred to as user access logs, user trails and click-stream data by the Web Mining community. Web logs are trails of past activities left behind by users of a website. These historical logs are embedded with significant information about the users and how a website is being used on a day to day basis.

Such information are deemed invaluable in today's world of customer-oriented businesses, especially for companies that rely on the web to advertise their services e-commerce and to web designers, who wish to maintain a constant stream of visitors to their website. Web usage mining provides these entities with the means of discovering such information that can be used to improve a business's performance or increase the effectiveness of a particular website.

Phishing website features

In this digital day and electronic world, Internet plays a vital role in day-to-day activities like communication, business, transactions, personal needs, marketing, e-commerce etc. Internet is a multifaceted facility which helps in completing many tasks readily and conveniently within few seconds. Almost everything is presently accessible over web in this period of progression of advances. Thus increasing usage of internet leads to cybercrime and other malware activities. The information divulged in online leaves digital imprint and if it happens to drop into the wrong hands, it will result in data theft, identity theft and monetary loss. Cybercrime includes many kinds of security issues over the internet and one of the most threatening problems is Phishing. Phishing is a fraudulent technique achieved by phishing web page. Phishing uses e-mails and websites, which are intended to look like from trusted organization, to hoodwink clients into unveiling their own or money related data. The threatening party then use these data for criminal purposes, such as, identity or data theft and extortion. Clients are deceived into revealing their data either by giving touchy data through a web shape or downloading and introducing unfriendly codes, which seek clients' PCs or checking clients' online actions to get data. Luring Internet users by making them click on rogue links that seem trustworthy is an easy task because of widespread credulity and unawareness.

It is important to prevent user's confidential data from unauthorized access. The procedure for the most part includes sending messages that then cause the beneficiary to either visit a deceitful site and enter their data or to visit an authentic site through a phishing intermediary attack or using spoofed website, which then gathers the details of user leads to several losses. The Phishing problem needs to be mitigated by anti-Phishing approaches. This research provides a solution that helps in detecting and preventing Phishing attacks using the features of phishing URLs and an automated real-time detection of phishing websites by machine learning approach.

Phishing attacks usually target user confidential information such as username, password and financial ID. Phishers would use their sophisticated attack vector such as emailing, or pop up window notification to lure the victim to visit the phishing website which has legitimate-looking layout. This will allow the phishers to harvest the victim credentials and sell them in the black market (as depicted in Figure 1).

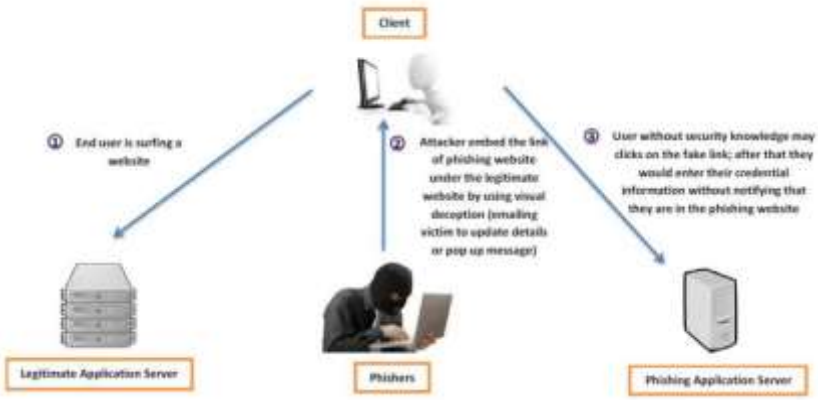


Fig. 1.2: Example of Phishing Attack Scenario

Anti-phishing refers to the method that is employed to prevent and defend phishing attacks. There are many techniques that offer protection at different domain. Some techniques work on emails, while others work on website attributes [2]. The proposed method works on the later. The contribution of this paper is twofold: first, it identifies and analyses attributes exhibited in phishing websites. Second, it proposes several new features and integrates to an existing method to enhance the overall detection performance. The works start by analysing and looking for abnormal attributes of a phishing website. The abnormal attributes that usually appear on phishing website include some uncommon symbols in the URL and some irregular HTML form and title elements. Therefore, extracting features from these attributes will enhance the phishing detection ability

Methods of phishing attack

The attacker can attack on any website in different ways. Some of methodologies are as follows [20]:

- ***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.
- ***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 [21].
- ***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 [22].

II. RELATED WORK

Blacklist and whitelist techniques are the most common and straightforward solutions. However, their effectiveness is determined by the completeness of the list. As a result, these techniques are not effective against new phishing websites [4]. Furthermore, majority of phishing websites are short-lived and the updated list is of less functional.

Page analysis inspects the properties of a webpage based on the features, which are extracted from the HTML source code or derived from a URL. For page source, the number of HTML form tags might provide an indicator to detect phishing website. In addition, the number of input fields such as user ID and password are also crucial and suitable to be used as an indicator [3]. Phishers may trick users to provide their credentials through these input elements. Unsuspecting users are also susceptible to visually deceptive text, images mask underlying text, images mimicking windows and windows mask underlying windows [2]. Hence, these elements are important for the analysis.

The URL string can be broken down into multiple tokens that constitutes of binary features. Examples of features include length of the URL, number of dots, existence of IP address in the URL and URL with HTTPS and SSL [1]. In order to get a more comprehensive analysis, Alexa database and WHOIS database are usually used to check the URL domain name, domain registrar, name server and age of domain.

In Phishing E-mail Detection Based on Structural Properties[5], the proposed approach explains to find phishing through appropriate identification and usage of structural properties of email. The experiment is done by SVM and classification technique to classify phishing e-mails. The technique used in this classification method is not large enough and it uses only one approach to identify phishing e-mails, which is low in efficiency and scalability. This is purely based on structural properties of e-mail and it has to extend more structural or content properties to reduce error results.

Discovering Phishing Target Based on Semantic Link Network[6], the paper proposes a novel approach to discover phishing website by calculating association relation among webpages that include malicious webpages and its associated webpages to measure the combination of link relation, search relation, and text relation. The semantic link network proposes a strategy based on four convergent situations to identify the suspicious webpage as phishing. The demerits in this approach are more kind of association has to be done, similarities between visual, layout and domain has to be related. This method is considered as a time consuming approach and also various sub-relations in the combined association relations be studied.

Evolving Fuzzy Neural Network for Phishing Emails Detection[7], deals with zero-day phishing email. It differentiates phishing email and ham email in online mode. It is adopted on feature fetching, rank fetching and grouping similar features of email. The technique is based on binary value 0 or 1 to produce the result for all features used in this method, where 1 denotes a phishing feature and 0 for non-phishing. This technique does not have more dynamic system so it is less in performance to produce accurate results.

Intelligent Phishing Website Detection and Prevention System by Using Link Guard Algorithm[8], proposed a system using link guard algorithm which works for hyperlinks. The algorithm performs certain tests like comparison of the DNS of actual and visual links, checks dotted decimal of IP address, checks encoded links and pattern matching. The drawbacks of this system

are, it produce the false positive results if any genuine site has IP address instead of domain name, and it considers some phishing site as normal one if the user does not visit the original site. This results in false negative conclusions.

In Said Afroz, Rachel Greenstadt - Phishzoo Approach[9], the algorithm detects current phishing sites by matching their content with genuine site. This will match images, contents and the structure of website with trusted one in order to avoid phishing. Drawbacks of this algorithm is, it requires matching image site and it is less robust for detecting phishing attacks.

III. RESEARCH METHODOLOGY

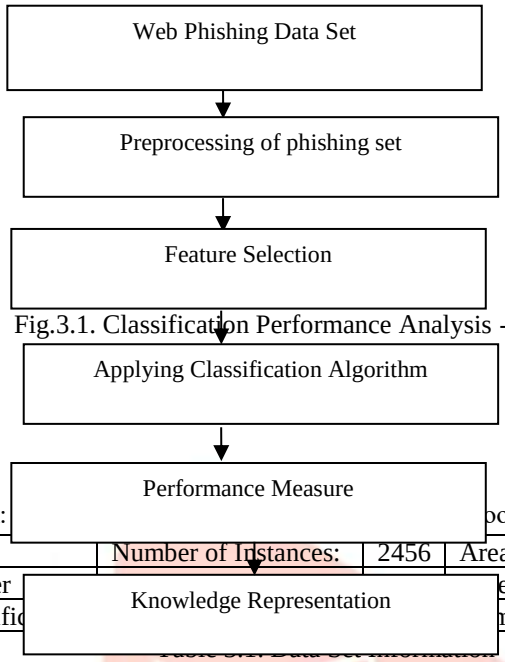


Fig.3.1. Classification Performance Analysis - Framework

Web page phishing data set

This dataset collected mainly from: [illegible]oogleâ€™s searching operators.

Data Set Characteristics:	N/A	Number of Instances:	2456	Area:	Computer Security
Attribute Characteristics:	Integer	[illegible]	[illegible]	e Donated	2015-03-26
Associated Tasks:	Classific	[illegible]	[illegible]	mber of Web Hits:	57465

ATTRIBUTES

- having_IP_Address
- URL_Length
- Shortening_Service
- having_At_Symbol
- double_slash_redirecting
- Prefix_Suffix
- having_Sub_Domain
- SSLfinal_State
- Domain_registeration_length
- Favicon
- port
- HTTPS_token
- Request_URL
- URL_of_Anchor
- Links_in_tags
- SFH
- Submitting_to_email
- Abnormal_URL
- Redirect
- on_mouseover
- RightClick
- popUpWidnow
- Iframe
- age_of_domain
- DNSRecord
- web_traffic
- Page_Rank
- Google_Index
- Links_pointing_to_page
- Statistical_report
- Result

The following classifiers are used to classify this data set with 10 folds cross validation

- Naïve Bayes
- Random Forest
- K Nearest Neighbour
- Support Vector Machine
- J48

LOAD DATA SET

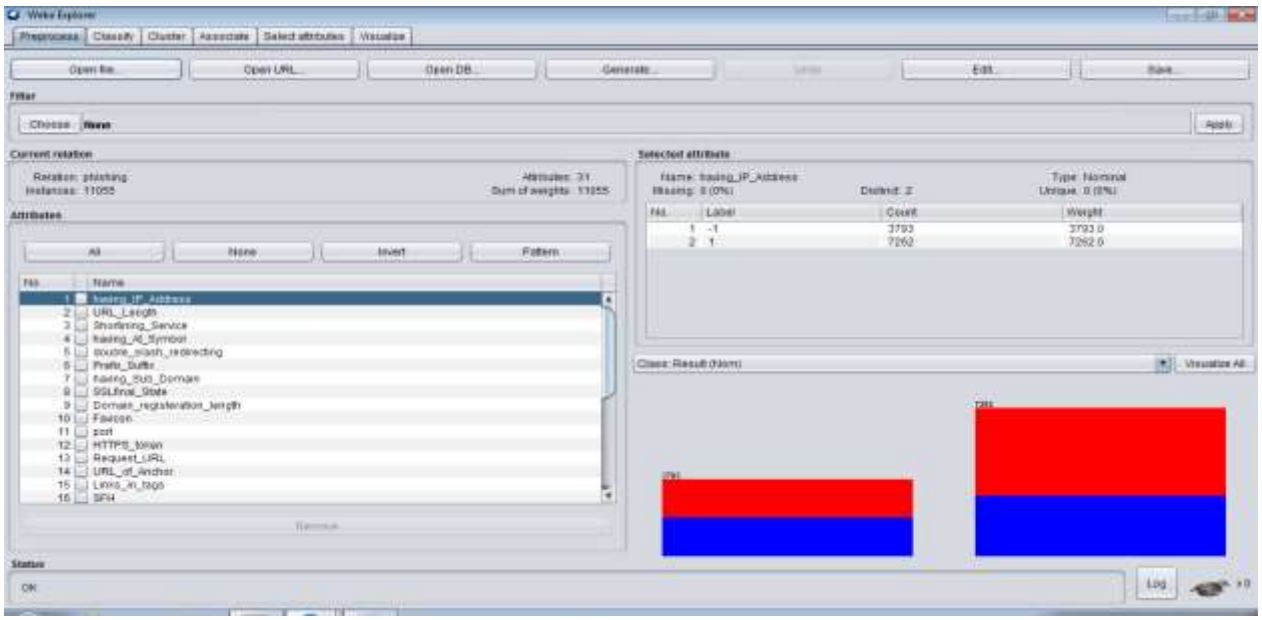


Fig.3.2. Load Data Set

IV. CLASSIFICATION, RESULTS AND DISCUSSION

4.1. CLASSIFICATION RESULTS

4.1.1. NAÏVE BAYES CLASSIFICATION

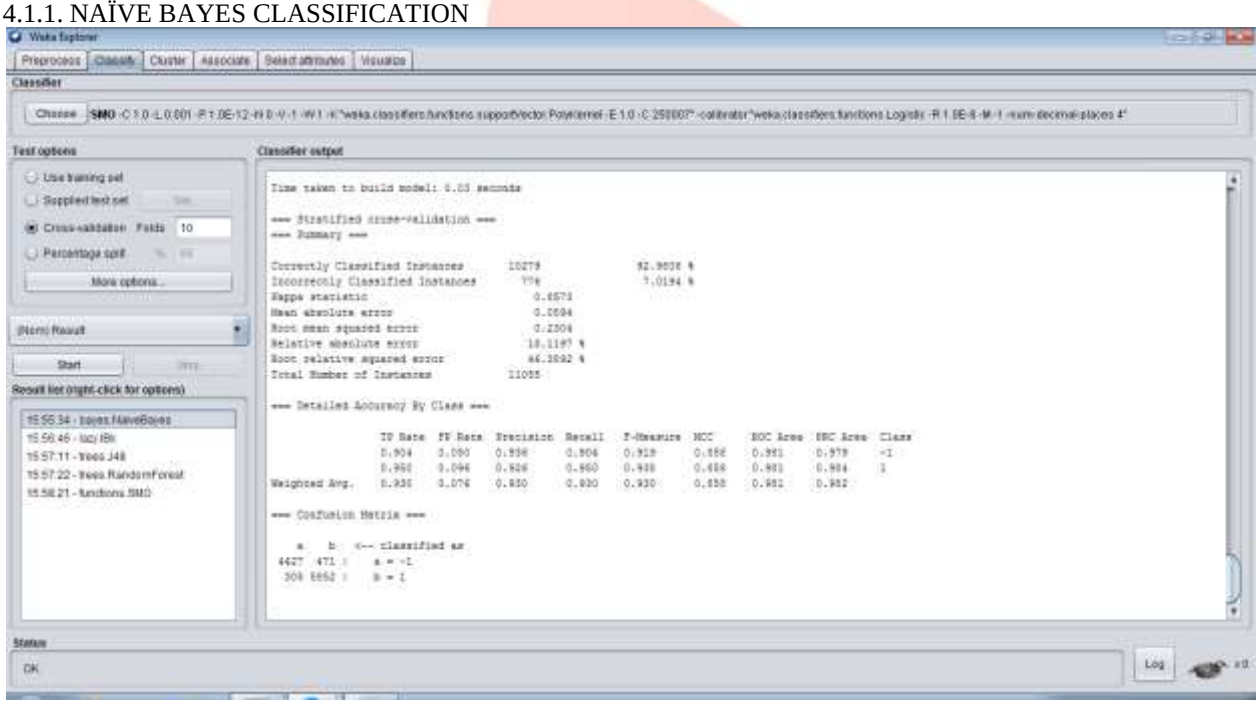


Fig.4.1 Naïve Bayes Classification Result

92.9806% of web page phishing data instances are correctly classified and remaining 7.0194% of instances are incorrectly classified. The percentage of correctly classified instances is often called accuracy or sample accuracy. So this data set consists of 92.98% accurate instances.

TP Rate: rate of true positives (instances correctly classified as a given class). Weighted average TP Rate of this data set is 0.930. FP Rate: rate of false positives (instances falsely classified as a given class). Weighted average FP Rate of this data set is 0.076.

Precision: proportion of instances that are truly of a class divided by the total instances classified as that class. Weighted average Precision value of this data set is 0.930.

Recall: proportion of instances classified as a given class divided by the actual total in that class (equivalent to TP rate). Weighted average Recall of this data set is 0.930.

F-Measure: A combined measure for precision and recall calculated as 2 * Precision * Recall / (Precision + Recall). Weighted F-Measure value is 0.930.

4.1.2. K-NEAREST NEIGHBOUR CLASSIFICATION

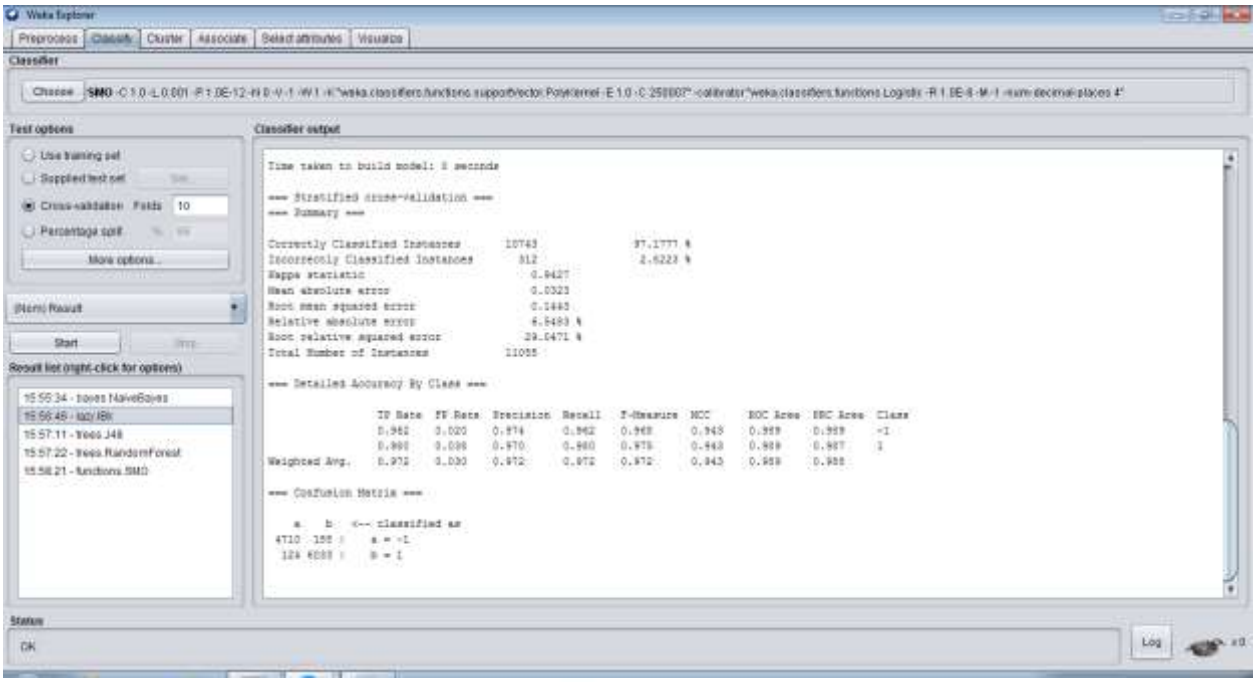


Fig.4.2. KNN Classification Result

97.1777% of web page phishing data instances are correctly classified and remaining 2.823% of instances are incorrectly classified. The percentage of correctly classified instances is often called accuracy or sample accuracy. So this data set consists of 97.18% accurate instances.

TP Rate: rate of true positives (instances correctly classified as a given class). Weighted average TP Rate of this data set is 0.972.

FP Rate: rate of false positives (instances falsely classified as a given class). Weighted average FP Rate of this data set is 0.030.

Precision: proportion of instances that are truly of a class divided by the total instances classified as that class. Weighted average Precision value of this data set is 0.972.

Recall: proportion of instances classified as a given class divided by the actual total in that class (equivalent to TP rate). Weighted average Recall of this data set is 0.972.

F-Measure: A combined measure for precision and recall calculated as 2 * Precision * Recall / (Precision + Recall). Weighted F-Measure value is 0.972.

4.1.3. J48 CLASSIFICATION

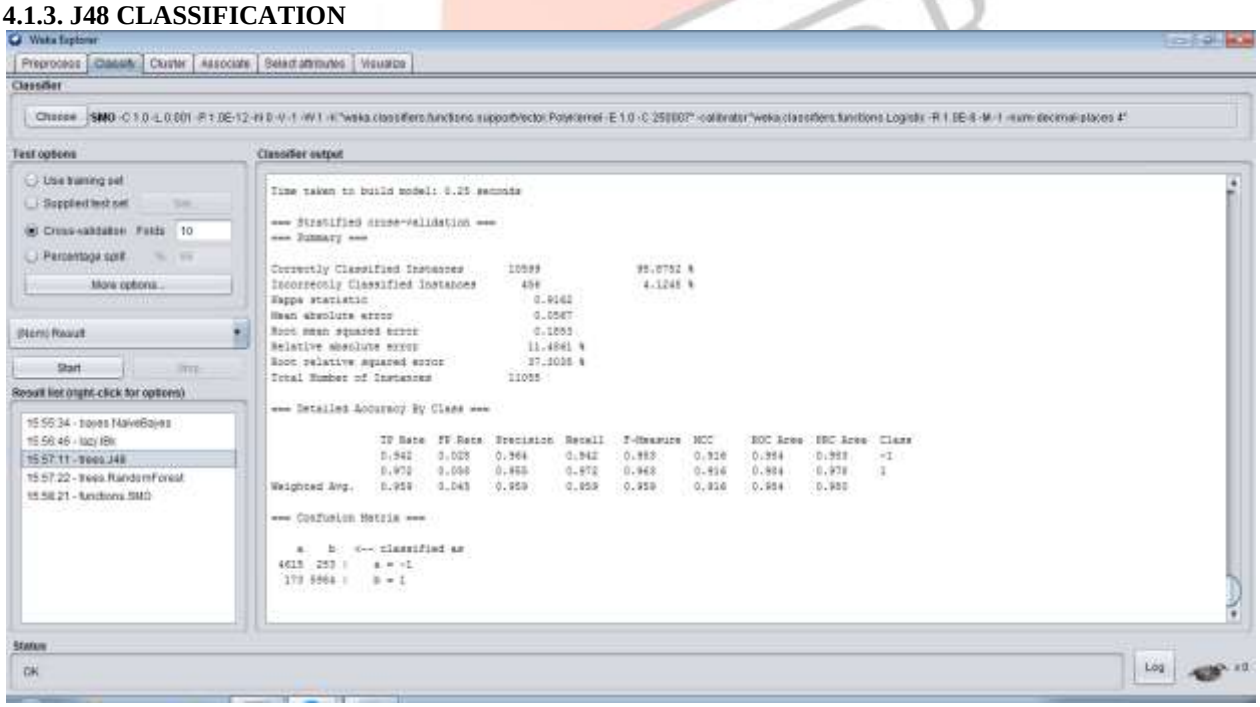


Fig.4.3 J48 Classification Result

95.8752% of web page phishing data instances are correctly classified and remaining 4.1248% of instances are incorrectly classified. The percentage of correctly classified instances is often called accuracy or sample accuracy. So this data set consists of 97.18% accurate instances.

TP Rate: rate of true positives (instances correctly classified as a given class). Weighted average TP Rate of this data set is 0.959.

FP Rate: rate of false positives (instances falsely classified as a given class). Weighted average FP Rate of this data set is 0.045.

Precision: proportion of instances that are truly of a class divided by the total instances classified as that class. Weighted average Precision value of this data set is 0.959.

Recall: proportion of instances classified as a given class divided by the actual total in that class (equivalent to TP rate). Weighted average Recall of this data set is 0.959.

F-Measure: A combined measure for precision and recall calculated as 2 * Precision * Recall / (Precision + Recall). Weighted F-Measure value is 0.959.

4.1.4. RANDOM FOREST CLASSIFICATION

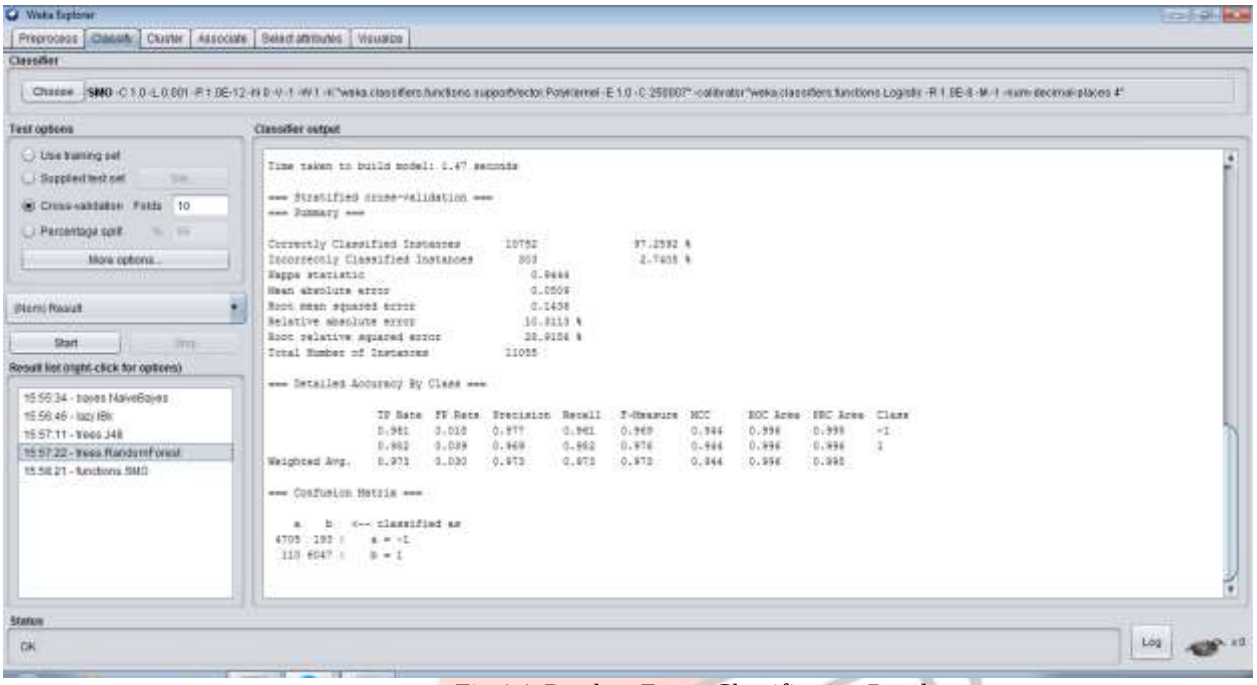


Fig.4.4. Random Forest Classification Result

97.2592% of web page phishing data instances are correctly classified and remaining 2.7408% of instances are incorrectly classified. The percentage of correctly classified instances is often called accuracy or sample accuracy. So this data set consists of 97.18% accurate instances.

TP Rate: rate of true positives (instances correctly classified as a given class). Weighted average TP Rate of this data set is 0.973.

FP Rate: rate of false positives (instances falsely classified as a given class). Weighted average FP Rate of this data set is 0.030.

Precision: proportion of instances that are truly of a class divided by the total instances classified as that class. Weighted average Precision value of this data set is 0.973.

Recall: proportion of instances classified as a given class divided by the actual total in that class (equivalent to TP rate). Weighted average Recall of this data set is 0.973.

F-Measure: A combined measure for precision and recall calculated as 2 * Precision * Recall / (Precision + Recall). Weighted F-Measure value is 0.973.

4.1.5. SUPPORT VECTOR MACHINE CLASSIFICATION

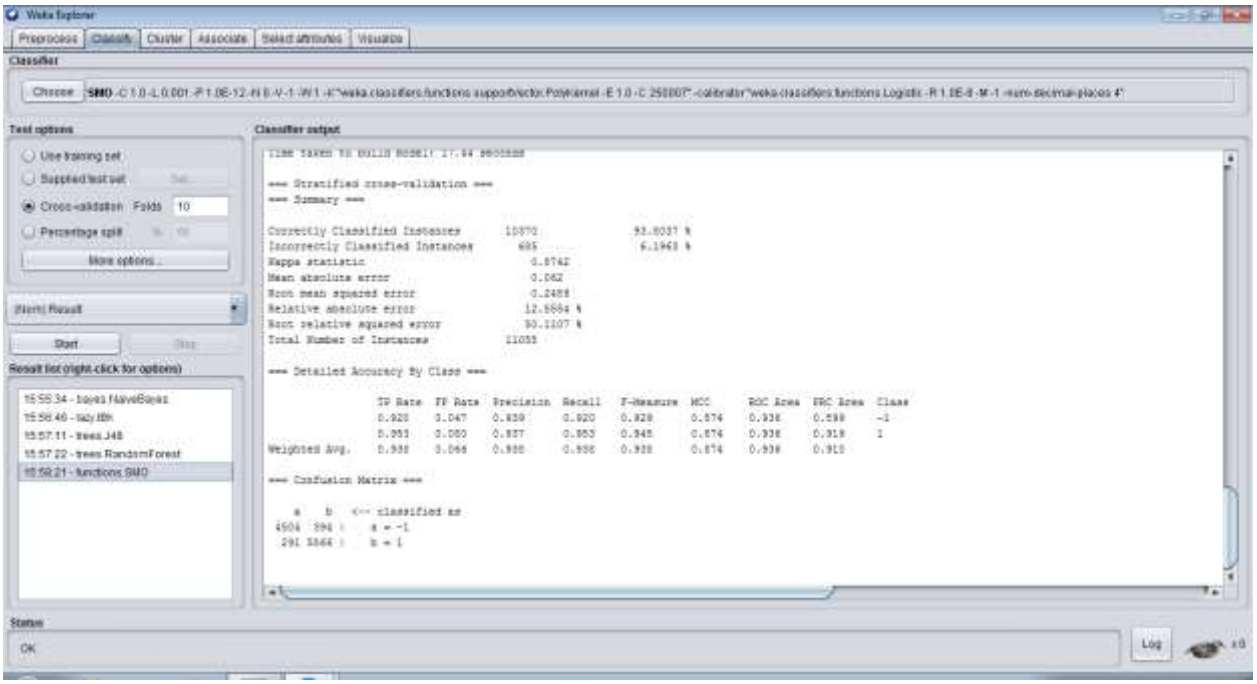


Fig.4.5 SVM Classification Result

93.8037% of web page phishing data instances are correctly classified and remaining 6.1963% of instances are incorrectly classified. The percentage of correctly classified instances is often called accuracy or sample accuracy. So this data set consists of 97.18% accurate instances.

TP Rate: rate of true positives (instances correctly classified as a given class). Weighted average TP Rate of this data set is 0.938.

FP Rate: rate of false positives (instances falsely classified as a given class). Weighted average FP Rate of this data set is 0.062.

Precision: proportion of instances that are truly of a class divided by the total instances classified as that class. Weighted average Precision value of this data set is 0.938.

Recall: proportion of instances classified as a given class divided by the actual total in that class (equivalent to TP rate). Weighted average Recall of this data set is 0.938.

F-Measure: A combined measure for precision and recall calculated as 2 * Precision * Recall / (Precision + Recall). Weighted F-Measure value is 0.938.

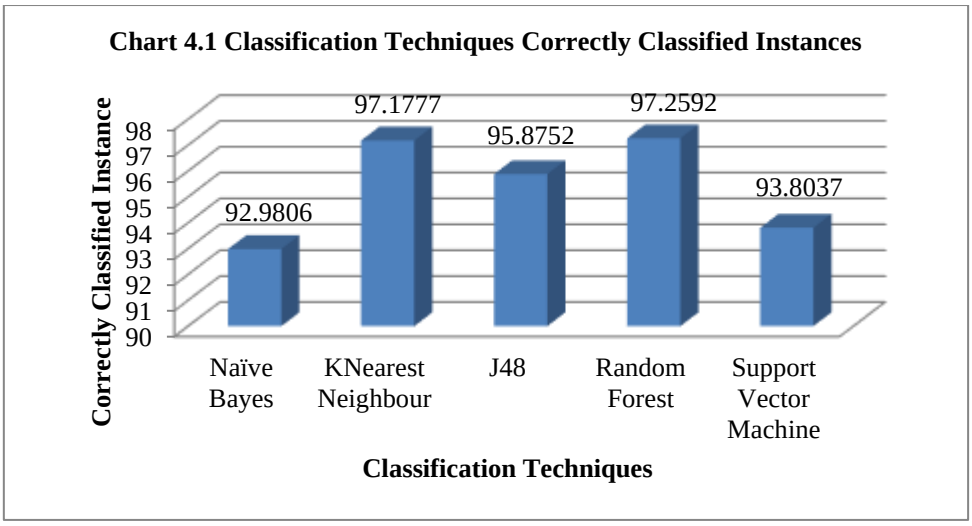
4.2. RESULTS AND DISCUSSIONS

4.2.1 COMPARISON OF CLASSIFICATION ALGORITHMS BASED ON CLASSIFIED INSTANCE

Classification	Correctly Classified Instances
Naïve Bayes	92.9806
KNearest Neighbour	97.1777
J48	95.8752
Random Forest	97.2592
Support Vector Machine	93.8037

Table 4.1. Correctly classified instances from various classifier models

The above table reveals that the classification of web page phishing instances as follows, Random Forest classification produced highest accuracy (97.26%) and closely followed by the KNN classification with 92.18% of accuracy, J48 classifier produced 95.88% accuracy, SVM produced 93.80% of accuracy and Naïve Bayes Classifier produced 92.984% accuracy.

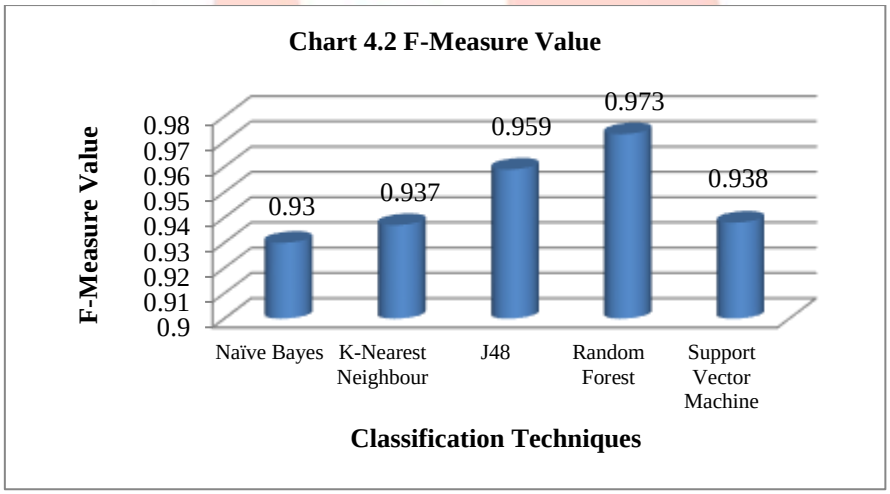


4.2.2. F-MEASURE VALUES FOR EACH CLASSIFICATION MODEL

Classification	F-Measure Value
Naïve Bayes	0.930
K-Nearest Neighbour	0.937
J48	0.959
Random Forest	0.973
Support Vector Machine	0.938

Table 4.2. F-Measure value of various classifier models

The above table reveals that the classification process in the web page phishing data set, Random forest classification holds high F-Measure value (0.973), J48 classifier's F-Measure value is 0.959, Support Vector Machine is 0.938, KNN's value is 0.937 and Naïve Bayes classifier's F-Measure value is 0.930.



4.2.3. COMPARISON MODELS

To calculate the performance of the various classification models, Correct Classified Rate (CCR), Recall Rate (RR), and F-measure to be used in document or data set classification criteria. The CCR is the rate of correct prediction, and Recall Rate is the ratio actually hit accurate predictions. And F-measure means the combinational mean of CCR and RR, and this is convenient expression method to compare models. The accuracy (AC) is the proportion of the total number of predictions that were correct. The recall or true positive rate (TP) is the proportion of positive cases that were correctly identified. The false positive rate (FP) is the proportion of negatives cases that were incorrectly classified as positive.

$$P = \frac{\text{No of correctly classified data}}{\text{No of correctly retrieved data}}$$

$$R = \frac{\text{No of correctly classified data}}{\text{Total No of Relevant data}}$$

$$\text{F- MEASURE} = \frac{2PR}{P+R}$$

Random Forest classification model has the best F-measure value (0.973), Correctly classified rate (97.25%) and best Recall rate (0.973) and this model showed the value compared to the other models such as Naïve Bayes, SVM, KNN and J48 decision tree in data classification of the web page phishing data set.

CONCLUSION AND FUTURE ENHANCEMENT

Data Classification is an important application area in web mining and web page phishing data sets why because classifying billions of phishing records manually is an expensive and time consuming task. Therefore, automatic classifier is constructed using pre classified sample phishing data set whose accuracy and time efficiency is much better than manual classification and prediction. Identifying efficient patterns also plays major role in text classification. Data mining classification techniques need to be designed to effectively manage large numbers of elements with varying frequencies. Almost all the known techniques for classification such as decision trees rules, Bayes methods and SVM classifiers have been used to the case of phishing data.

In this research work, web page's phishing data sets are used to analyse the various classification techniques and find out the efficient classifier. And we compared those data by applying the material to the conventional techniques of Bayesian statistical classification, J48 Decision tree, Random Forest, KNN and SVM to form a classification model.

The Random forest model shows better performance than KNN, SVM, J48 and Naïve Bayes classification models. Future works may also include hybrid classification models by combining some of the web mining techniques such as attribute selection and clustering.

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