iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 4 **Issue:** VI **Month of publication:** June 2016

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Volume 4 Issue VI, June 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

A Survey Paper on Phishing Attacks with New Unsupervised Learning Model

Raina Bhushan Goswami[1], Sunil Sharma[2]
[1]*M. Tech Scholar Dr. C. V. Raman Universitry, Kargi Road Kota Bilaspur,*
[2]*Asst Professor Dr. C. V. Raman Universitry, Kargi Road Kota Bilaspur*

Abstract—Phishing attacks are very sensitive issue now days as we are in the world of connectivity that is internet. More the internet user more the attackers are in existence. Every email or any type of personal communication resources needs security. We have made a survey on this issue so that we can create some new concept to make our resources like email secure. We have studied phishing attack algorithm from which we have got so many good as well. As a result we found the unsupervised learning is appropriate for our upcoming research.
Keywords— Supervised Learning, Neural Network, Fuzzy Logic, Unsupervised learning.

I. INTRODUCTION

Now days the use of Internet is increasing rapidly to access information from the World Wide Web. Every organization like bank, insurance, industries have large volume of data. To secure such information, classification of information plays a very important role. Classification is one of the most important decision making techniques in many real world problems. Anti-phishing is one of the important areas to classify the phishing and normal e-mails[21]. Phishing is an Internet-based attack in which an attacker tricks a user into submitting his or her sensitive information to a fake website mimicking a legitimate site. This sensitive information ranges from usernames and passwords to bank account numbers and social security numbers. Phishing is a serious threat to the security of internet users' confidential information. Phishing is also a type of spam emails which redirect the users to fake websites and access the sensitive information from users. One of the major security issues associated with internet users these days is "phishing". Phishing is a fallacious action performed in order to acquire financial and personal information like usernames, passwords, credit card numbers, social security numbers, date of birth etc. It is an email spoofing in which a legitimate-looking email is sent to some target users. These emails appear to come from familiar and authentic websites. It usually includes exciting or bothersome statements and suspicious redirecting hyperlinks towards fake website spoofing Innocent internet users. A diagrammatic explanation of phishing process is given in fig. 1. The phisher installs phishing website and mass mailer to the victim server. The server unknowingly broadcast these phishing emails to the target users. User get forged by clicking hyperlinks embedded with the email.

Fig 1. Phishing Process

Volume 4 Issue VI, June 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

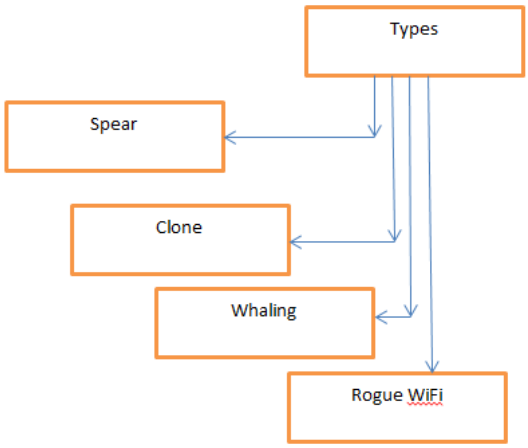


Fig 2. Types of Phishing Attacks

A. Spear phishing

It is one of the most successful techniques accounting 91% of attacks. It is accomplished by using personal information of the victim to earn trust thus increasing probability of success [20].

B. Clone phishing

A type of phishing in which a legitimate email is cloned completely replacing the attachment/link with the spurious version.

C. Whaling

It primarily targets high profile and senior executives. The content of email is often written as a legal subpoena, customer complaint, or executive issue. It involves some kind of falsified companywide concern [21].

D. Rogue WiFi (MitM)

Attackers compromise free Wifi access-points, and configure them to run man-inthe-middle (MitM) attacks [22].The Kaspersky Lab study „Financial Cyberthreats in 2014 reports that 28.8% of phishing attacks in 2014 were intended to steal financial data from users. While carrying out their scams, cybercriminals have shifted their focus from bank brands to payment systems and online shopping sites.

In the Payment Systems category, cybercriminals mostly targeted data belonging to users of Visa cards (31.02% of detections in the Payment Systems category), PayPal(30.03% of detections) and American Express (24.6%). Amazon remains the most ommonly-attacked brand in the Online Shopping category – 31.7% of attacks in this category used phishing pages mentioning Amazon.

However, this is 29.41 percentage points less than in the previous year [2].The existing defense system (its designs and technology) against such malicious attacks needs to be greatly improved. Behdad et al. [1] pointed out that improving the defense system is not enough to stop fraudsters as some of them could still penetrate; the system should also be able to identify fraudulent activities and prevent them from occurring.To ensure cyber security and combat cybercrime, development and implementation of emphatic phishing

detection techniques is highly essential. Anti-phishing techniques based on machine learning methodology have already substantiated to be utterly effective due to

advances in data mining and learning algorithms.

Volume 4 Issue VI, June 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

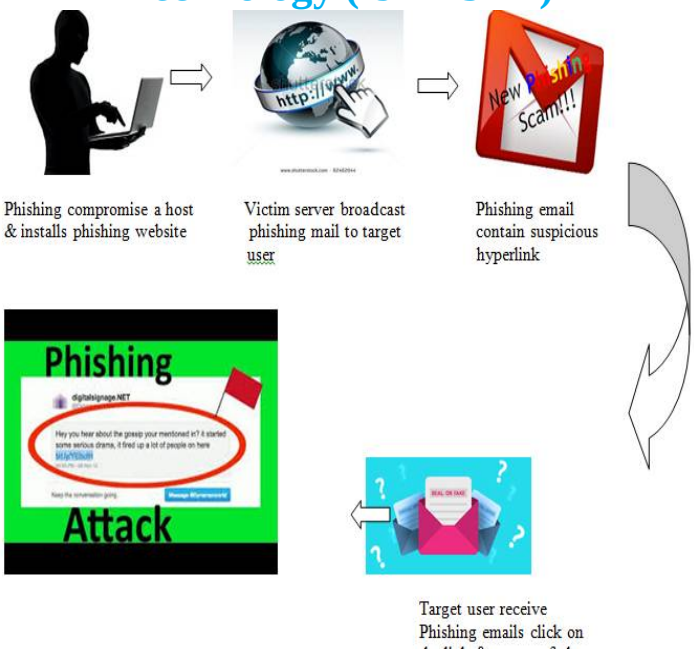


Fig 3: Phishing Scenario

II. EXISTING METHODOLOGY

Bayesian net is a statistical processing based on bayes decision theory and is a fundamental technique for pattern recognition and classification. It assumes that pattern possesses random characteristics and they are generated in a random way by some natural phenomena and process. It is a graphical model that encodes probabilistic relationships among variable of interest. The natural choice for dealing with random and uncertain pattern is to use statistical technique based on probabilistic characteristics of data. The Bayesian method is based on the assumption that the classification of patterns is expressed in probabilistic terms. It assumes that the statistical characteristics of random patterns are expressed as known probability values describing the random nature of pattern and their features. These probabilistic characteristics mostly concern a priori probability and conditional probability density of pattern of class [9].

CART (Classification and Regression Technique) is one of the popular methods of building decision tree in the machine learning community. It builds a binary decision tree by splitting the records to each node, according to a function of a single attribute. CART uses the Gini index for determining the best split. The initial split produces two nodes, each of which attempts to split in the same manner as the root node. Once again, all the input fields are examined to find the candidate splitters. If no split can be found that significantly decreases the diversity of a given node, labelled as leaf node. At the end of tree growing process, every record of the training set is assigned to some leaf of the full decision tree. Each leaf is assigned a class and an error rate. Error rate of a leaf node is the percentage of incorrect classification at that node [9].

CHAID (Chi-Squared Automation InteractionDetection) is a derivative of AID (Automatic Interaction Detection). It attempts to stop growing the tree before over fitting occurs. CHAID avoids the pruning phase. In the standard manner, the decision tree is constructed by partition the data set into two or more data subsets, based on the values of one of the non-class attributes. After the data set is partitioned according to the chosen attributes, each subset is considered for further partitioning using the same algorithm. Each subset is partitioned without regard to any other subset. The process is repeated for each subset until some stopping criteria is met. In CHAID, the number of subsets in a partition can range from two up to the number of distinct values of the splitting attribute [9].

D**.** *Artificial Neural Network (ANN)* is composed of a set of elementary computational units, called neurons, connected together through weighted connections. These units are organized in layers so that every neuron in a layer is exclusively connected to the neurons of the preceding layer and the subsequent layer. Every neuron, also called a node, represents an autonomous

computational unit and receives inputs as a series of signals that dictate its activation. Following activation, every neuron produces an output signal. All the input signals reach the neuron
simultaneously, so the neuron receives more than one input signal, but it produces only one output signal. Every input signal is associated with a connection weight. The weight determines the relative importance the input signal can have in producing the final impulse transmitted by the neuron. The connections can be exciting, inhibiting or null according to whether the corresponding weights are respectively positive, negative or null. A threshold value, called bias, is similar to an intercept in a regression model [10].The term neural network has moved round a large class of models and learning methods. The main idea is to extract linear combinations of the inputs and derived features from input and then model the target as a nonlinear function of these features. ANN is a large class of algorithms that has the capability of classification, regression and density estimation [11].

Support vector machine (SVM) design a hyperplane or set of hyperplanes in a high or infinite dimensional space, which can be used for classification, regression or other tasks. A SVM is a promising new method for classification of both linear and nonlinear data. Support Vector Machines are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships [12]. Support vector machine algorithms divide the n dimensional space representation of the data into two regions using a hyperplane. This hyperplane always maximizes the margin between the two regions or classes. The margin is defined by the longest distance between the examples of the two classes and is computed based on the distance between the closest instances of both classes to the margin, which are called supporting vectors [13]. The support vector machine is very popular as a high-performance classifier in several domains in classification. The basic idea is to construct a hyper plane as the decision surface such that the margin of separation between positive and negative examples is maximized. Here the error rate of a learning machine isconsidered to be bounded by the sum of the training error rate and a term depending on the Vapnik Chervonenk is (VC) 1 dimension. Given a labeled set of N training samples (Xi, Y). The expected loss of making decision is the minimum.

Decision tree induction is the learning of decision trees from class labeled training tuples. A decision tree is a flow chart like tree structure, where each internal node denote a test on an attribute, each branch represent an outcome of the test, and each leaf node hold a class label. The topmost node in a tree is the root node. Decision tree can handle high dimensional data. Their representation of acquired knowledge in tree from is intuitive and generally easy to assimilate to human. The learning and classification steps of decision tree induction are simple and fast. Decision tree algorithm is simple and fast. These tree classifiers have good accuracy. Decision tree induction algorithms have been used for classification in many application areas such as medicine, manufacturing, and production, Financial Analysis, astronomy, and molecular Biology. Decision tree are the basic of Several Commercial rule induction System. Decision tree are built, many of the branches may reflect noise or outliers in the training data. In this research work we will use various data mining based decision tree algorithm like CART, QUEST, CHAID, ID3, C5.0 etc to development of decision support system [14].

C5.0 is one of the more recent in a family of learning algorithms referred to as decision tree algorithms. This algorithm is an improvement of the C4.5 algorithm also developed by Quinlan. The improvements are merely in efficiency, the algorithm remains the same [16]. The algorithm is based on the concepts of entropy, the measure of disorder in the collection, and the information gain of each attribute. Information gain is a measure of the effectiveness of an attribute in reducing the amount of entropy in the collection. The C5.0 algorithm builds a decision tree for the data in question. This can be thought of as a sequence of if then rules that allow new instances to be classified. It begins by calculating the entropy of a collection (S) as shown in Equation 1. In this, c represents the number of classes in the system (2 in the phishing detection problem) and pi represents the proportion of instances that belong to class i. The next step is to calculate the information gain for each attribute. This is the expected reduction in entropy by partitioning the dataset on the given attribute. The information gain for attribute, A, in collection, S is shown in Equation 2 where E(S) is the entropy of the collection as a whole, Sv is the set of instances that have value v for attribute A.

QUEST uses a sequence of rules, based on significance tests, to evaluate the predictor variables at a node. For selection purposes, as little as a single test may need to be performed on each predictor at a node. Unlike C&RT, all splits are not examined,

and unlike C&RT and CHAID, category combinations are not tested when evaluating a predictor for selection. Splits are determined by running quadratic discriminate analysis using the selected predictor on groups formed by the target categories. This method again results in a speed improvement over exhaustive search (C&RT) to determine the optimal split [17, 18].

Hybrid: Two or more modes combined to form a new model is called an hybrid model. A hybrid model is a combination of two or more models to avoid the drawbacks of individual models and to achieve high accuracy. Bagging and boosting [14] are two techniques that use a combination of models. Each combines a series of k learned models (classifiers), M1, M2,…..Mk, with the aim of creating an improved composite model, M. Both bagging and boosting can be used for classification

III. CHARACTERISTICS OF PHISHING EMAILS

A typical phishing email will have the following characteristics:

A. It normally appears as an important notice, urgent update or alert with a deceptive subject line to entice the recipient to believe that the email has come from a trust source and then open it. The subject line may consist of numeric characters or other letters in order to bypass spamming filters.

B. It sometimes contains messages that sound attractive rather than threatening e.g. promising the recipients a prize or a reward.

C. It normally uses forged sender's address or spoofed identity of the organisation, making the email appear as if it comes from the organization it claimed to be.

D. It usually copies contents such as texts, logos, images and styles used on legitimate website to make it look genuine. It uses similar wordings or tone as that of the legitimate website. Some emails may even have links to the actual web pages of the legitimate website to gain the recipient's confidence.

E. It usually contains hyperlinks that will take the recipient to a fraudulent website instead of the genuine links that are displayed.

F. It may contain a form for the recipient to fill in personal/financial information and let recipient submit it. This normally involves the execution of scripts to send the information to databases or temporary storage areas where the fraudsters can collect it later.

IV. CHARACTERISTICS OF PHISHING WEBSITES

A typical phishing website will have the following characteristics:

A. It uses genuine looking content such as images, texts, logos or even mirrors the legitimate website to entice visitors to enter their accounts or financial information.

B. It may contain actual links to web contents of the legitimate website such as contact us, privacy or disclaimer to trick the visitors.

C. It may use a similar domain name or sub-domain name as that of the legitimate website.

D. It may use forms to collect visitors' information where these forms are similar to that in the legitimate website.

E. It may in form of pop-up window that is opened in the foreground with the genuine web page in the background to mislead and confuse the visitor thinking that he/she is still visiting the legitimate website.

F. It may display the IP address or the fake address on the visitors' address bar assuming that visitors may not aware of that. Some fraudsters may perform URL spoofing by using scripts or HTML commands to construct fake address bar in place of the original address.

V. COMMON METHODS OF PHISHING ATTACKS

If the recipient believes that the email comes from a legitimate organisation, there are several common methods used by the fraudsters for phishing.

A. Install Trojan program or worms to the recipient's computer in form of email attachment to exploit loopholes and vulnerabilities or to take screenshots of the system, in order to obtain sensitive information from the recipient.

B. Use spyware, such as keyboard loggers, to capture information from the recipient's computer and sends the information back to the fraudsters.

C. Use deceit to gain recipient's confidence so that the recipient will visit the fraudulent website that appears as legitimate and provide sensitive information by completing a form on web page.

May 2016		
23-May-2016	Phishing email related to Standard Chartered Bank (Hong Kong) Limited	The Hong Kong Monetary Authority (HKMA) wishes to alert members of the public to a press release issued by Standard Chartered Bank (Hong Kong) Limited on phishing email, which has been reported to the HKMA.
23-May-2016	Fraudulent website related to Dah Sing Bank, Limited	The Hong Kong Monetary Authority (HKMA) wishes to alert members of the public to a press release issued by Dah Sing Bank, Limited on fraudulent website, which has been reported to the HKMA.
20-May-2016	Cyber Smart Advice: Advanced Security Settings for Intant Messaging Application (Chinese only)	Please refer to Chinese version
18-May-2016	Launch of Cybersecurity Fortification Initiative by HKMA at Cyber Security Summit 2016	To further enhance the cyber resilience of the banking sector in Hong Kong, the Hong Kong Monetary Authority (HKMA) announced today (May 18) the launch of a "Cybersecurity Fortification Initiative" (CFI) at the Cyber Security Summit 2016 (the summit), in which the HKMA also serves as the programme advisor for this prestigious event.
18-May-2016	Suspicious Internet banking mobile application related to Public Bank (Hong Kong)	The Hong Kong Monetary Authority (HKMA) wishes to alert members of the public to a press release issued by Public Bank (Hong Kong) Limited on suspicious Internet banking mobile application (Apps), which has been reported to the HKMA.
13-May-2016	Cyber Smart Advice: Defend against Bedep Malware (Chinese only)	Please refer to Chinese version
13-May-2016	GovCERT.HK - Security Alert (A16-05-04): Multiple Vulnerabilities in Adobe Flash Player	Security updates are released for Adobe Acrobat/Reader to address multiple vulnerabilities. It is reported that the vulnerability CVE-2016-4117 is being actively exploited.
11-May-2016	GovCERT.HK - Security Alert (A16-05-03): Multiple Vulnerabilities in Adobe Acrobat/Reader	Security updates are released for Adobe Acrobat/Reader to address multiple vulnerabilities.

Fig 4: New Related To Phishing (Source : http://www.infosec.gov.hk/english/news/newsletters.html)

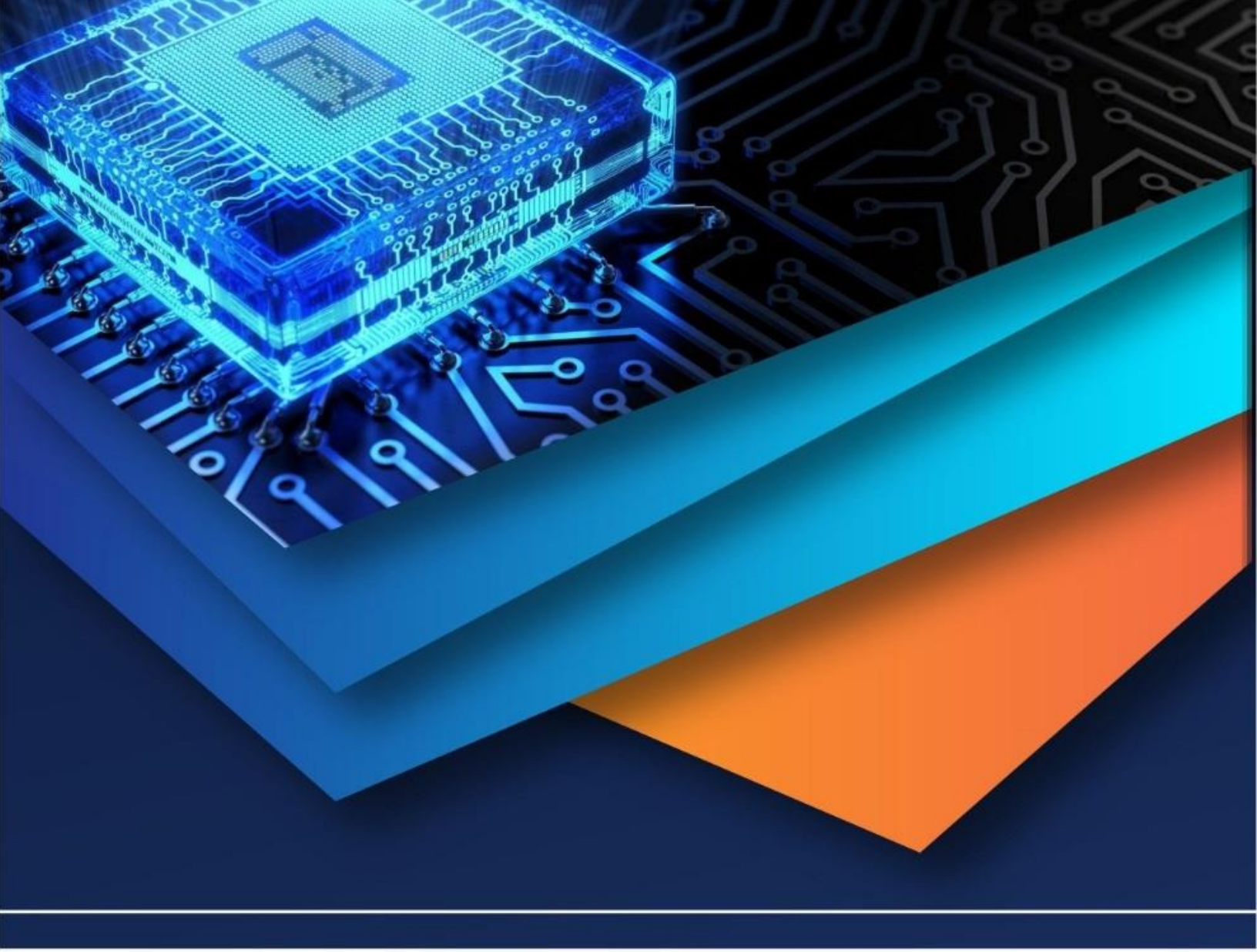
VI. CONCLUSION

In this paper we are trying to find out how unsupervised learning methodology where in algorithm, technique and attacks mechanism works. We have got a new combination that is any of the clementine tool with unsupervised learning technique.

REFERENCES

[1] M. Behdad, L. Barone,M. ennamoun, and T. French, “Natureinspired techniques in the context of fraud detection,” IEEE Transactions on Systems, Man, and cybernetics C: Applications and Reviews, vol. 42, no. 6, pp. 1273–1290, 2012.
[2] http://www.kaspersky.com/about/news/virus/2015/Over-a-quarter-ofphishing-attacks-in-2014-targeted-users-financial-data.
[3] Isredza Rahmi A Hamid and Jemal Abawajy,” Phishing Email Feature Selection Approach”, International Joint Conference of IEEE TrustCom-11, pp. 916-921, 2011.
[4] F. Toolan, J. Carthy.: Phishing Detection using Classifier Ensemble. In E-Crime Researchers Summit,2009.
[5] Wilfried N. Gansterer David P\, et al., "E-Mail Classification for Phishing Defense,", Springer-Verlag, presented at the Proceedings of the 31th European conference on IR Research on Advances in Information Retrieval, Toulouse, France,PP. 449-460, 2009.
[6] M. Bazarganigilani, ”Phishing E-Mail Detection Using Ontology Concept and Nave Bayes Algorithm,” International Journal of Research and Reviews in Computer Science, vol. 2,no.2, 2011.
[7] del Castillo, M.Iglesias, Ángel Serrano, J., "An Integrated Approach to Filtering Phishing Emails Computer Aided Systems Theory –EUROCAST 2007." vol. 4739, R. Moreno Diaz, et al., Eds., ed: Springer Berlin / Heidelberg, pp. 321-328, 2007.
[8] N. Zhang and Y. Yuan, “Phishing detection using neural network,”http://cs229.stanford.edu/proj2012/ZhangYuanPhishingDetectionUsingNeuralNetwork.pdf.
[9] Arun K. Pujari, Data mining techniques, 4thedition, Universities Press (India) Private Limited,2001.
[10] Paolo Giudici, Silvia Figini, “Applied Data Mining for Business and Industry” ,John Wiley & Sons Ltd,. United Kingdom, 2009.
[11] Alessio Pascucci,“ Toward a PhD Thesis on Pattern Recognition” , 2006.
[12] V. N. Vapnik, “Statistical Learning Theory” , New York: John Wiley and Sons, 1998.
[13] V. Vapnik, "The Nature of Statistical Learning Theory” ,Springer; 2 edition , 1998.
[14] Han, J.,& Micheline, K., “Data mining: Concepts and Techniques”, Morgan Kaufmann ,Publisher, 2006.
[15] Quinlan, J. R. C4.5: Programs for Machine Learning, MorganKaufmann Publishers 1993.
[16] Tanvi Chauhan, Prof.Vineet Richhariya, Sunil Sharma, “Literature Report on Face Detection with Skin & Reorganization using Genetic Algorithm”, Tanvi Chauhan et al. / IJAIR Vol. 2 Issue 2 ISSN: 2278-7844
[17] Aanchal Chauhan , Zuber Farooqui, “AN INVENTIVE APPROACH FOR FACE DETECTION WITH SKIN SEGMENTATION AND MULTI-SCALE COLOR RESTORATION TECHNIQUE USING GENETIC ALGORITHM”, INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTERAPPLICATIONS AND ROBOTICS, Vol. 4 Issue 1, January 2016
[18] Tanvi Chauhan, Vineet Richhariya,” Real Time Face Detection with Skin and Feature Based Approach and Reorganization using Genetic Algorithm“, CIIT Digital Image Processing, Vol 5, No 1 (2013)
[19] D. Shanmuga Priya , B. Kavitha, R. Naveen Kumar and K. Banuroopa,”Improving BayesNet classifier using various feature reduction method for spam classification”, IJCST, Vol. 1 , Issue 2,2010.
[20] Stephenson, Debbie. "Spear Phishing: Who"s Getting Caught?".Firmex. Retrieved 27 July 2014.
[21] "What Is 'Whaling'? Is Whaling Like 'Spear Phishing'?". About Tech. Archived from the original on 2015-03-28. Retrieved March 28, 2015.
[22] "Black Hat DC 2009". May 15, 2011.

[23] Jose Nazario. Phishing corpus. http://monkey.org/~jose/wiki/doku.php?id=phishingcorpus.
[24] SpamAssassin. Public corpus. http: //spamassassin.apache.org/publiccorpus.
[25] Niharika Vaishnaw, S R Tandan, "Development of Anti-Phishing Model for Classification of Phishing E-mail" International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, Issue 6, June 2015
[26] Amit Dewangan, Sadaf Rahman, "Secured Wireless Content Transmission over Cloud with Intelligibility" International Journal of Engineering and Applied Sciences (IJEAS) ISSN: 2394-3661, Volume-2, Issue-5, May 2015.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY