Volume 6, Issue 6, June 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Anti Email Phishing Model Development with Unsupervised Learning Approach

Raina Bhushan Goswami, Sunil Sharma, Tanvi Sharma
CSE, Dr. C V R U Bilaspur, Chhattisgarh,
India

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 introduced a model on this issue so that we can create some new concept to make our resources, like email, secure. In this paper we studied prior work which was based on supervised learning so we come up with a plan using unsupervised learning method.

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 spoofingInnocent internet users. A diagrammatic explanation of phishing process is givenin 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 gets forged by clicking hyperlinks embedded with the email.

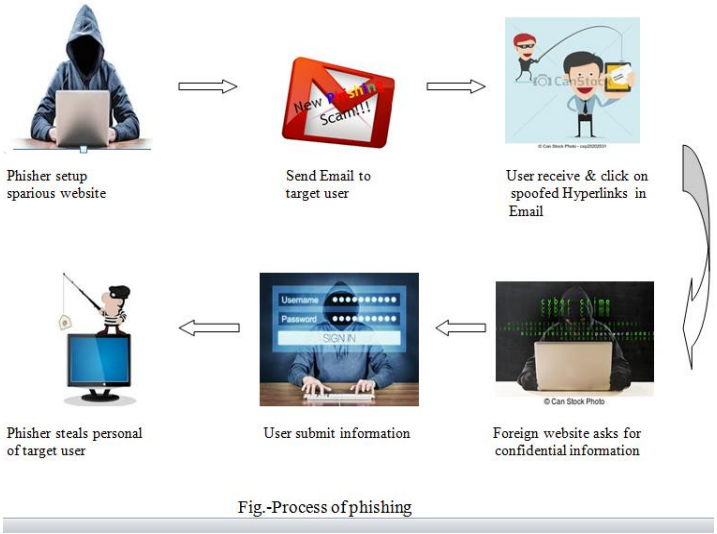


Fig 1. Phishing Process

1) **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].
2) **Clone phishing:** A type of phishing in which a legitimate email is cloned completely replacing the attachment/link with the spurious version.

© 2016, IJARCSSE All Rights Reserved

3) **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].

II. CHARACTERISTICS OF PHISHING EMAILS

A typical phishing email will have the following characteristics:

- 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.
- It sometimes contains **messages that sound attractive** rather than threatening e.g. promising the recipients a prize or a reward.
- It normally uses **forged sender's address** or spoofed identity of the organization, making the email appear as if it comes from the organization it claimed to be.
- 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.
- It usually contains **hyperlinks** that will take the recipient to a fraudulent website instead of the genuine links that are displayed.
- 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.

III. CHARACTERISTICS OF PHISHING WEBSITES

A typical phishing website will have the following characteristics:

- 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.
- It may contain **actual links** to web contents of the legitimate website such as contact us, privacy or disclaimer to trick the visitors.
- It may use a **similar domain name** or sub-domain name as that of the legitimate website.
- It may use **forms** to collect visitors' information where these forms are similar to that in the legitimate website.
- 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.
- 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.

IV. COMMON METHODS OF PHISHING ATTACKS

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

1. 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.
2. Use spyware, such as keyboard loggers, to capture information from the recipient's computer and sends the information back to the fraudsters.
3. 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 2: New Related To Phishing (Source :http://www.infosec.gov.hk/english/news/newsletters.html)

© 2016, IJARCSSE All Rights Reserved

V. PROPOSED METHODOLOGY

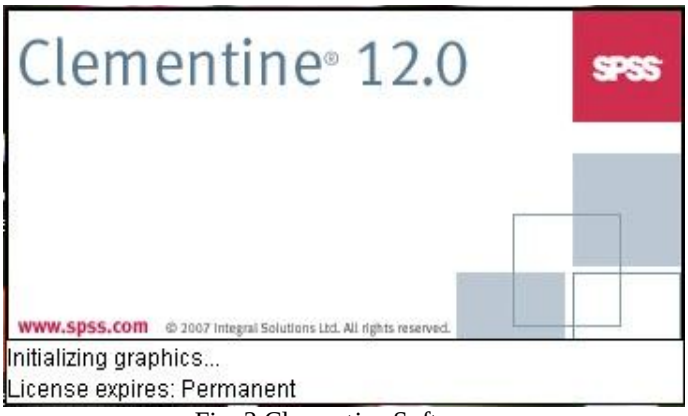
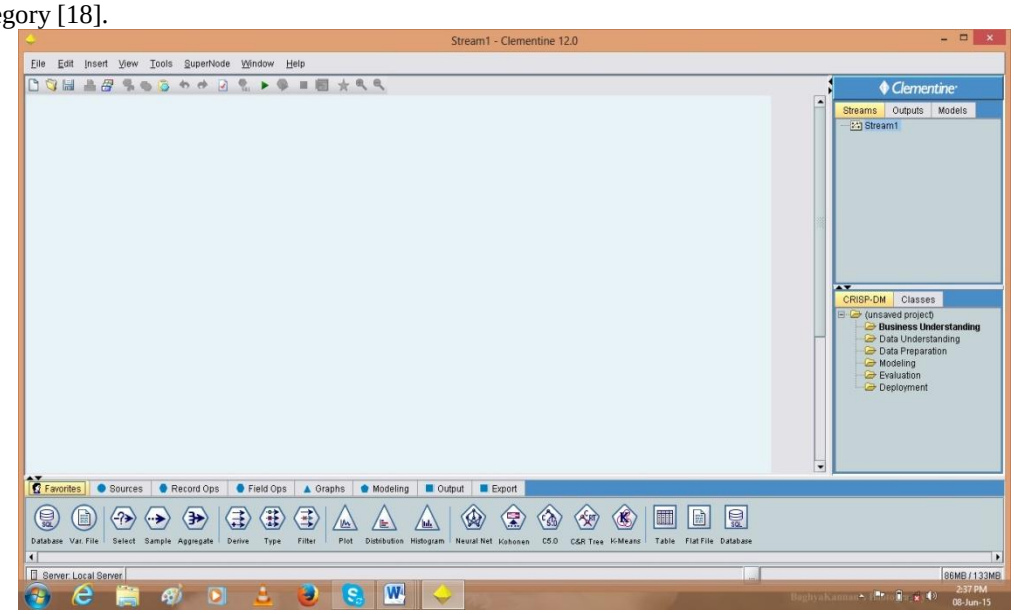


Fig. 3 Clementine Software

- Uncover key insights and use them to solve real business problems.
- Use IBM SPSS Modeler to solve your toughest challenges.
- Graphical interface makes modeling easy, saves time.
- Get faster results through automation.
- Use all of your data for maximum insight.
- Support for enterprise standards and technologies.

Clementine provides a number of so-called "supervised learning" and "unsupervised learning" techniques:

- **Supervised techniques model** an output variable based on one or more input variables. These models can be used to predict or forecast future cases where the outcome is unknown. Neural Networks, Rule Induction (decision trees), Linear Regression, and Logistic Regression, Bayes net, SVM, C5.0, CART, k-NN are some of these supervised techniques available in clementine 12.0
- **Unsupervised techniques** are used in situations where there is no field to predict but relationships in the data are explored to discover its overall structure. Kohonen networks, Two Step, and K-means belong to this category [18].



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, labeled 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 [40].

CART Algorithm

1. The basic idea is to choose a split at each node so that the data in each subset (child node) I "purer" than the data in the parent node. CART measures the impurity of the data in the nodes of a split with an impurity measure i(t).

© 2016, IJARCSSE All Rights Reserved

2. If a split s at node t sends a proportion pL of data to its left child node tL and a corresponding proportion pR of data to its right child node tR , the decrease in impurity of split s at node t is defined as
   $\Delta i(s,t) = i(t) – pLi(tL) – pRi(tR)$ = impurity in node t– weighted average of impurities in nodes tL and tR
3. A CART tree is grown, starting from its root node (i.e., the entire training data set) t=1,by searching for a split s* among the set of all possible candidates S which give the largest decrease in impurity

$$\Delta i(s^*,1) = \max_{s \in S} \Delta i(s,1)$$

Then node t=1 is split in two nodes t=2 and t=3 using split s*

4. The above split searching process is repeated for each child node.
5. The tree growing process is stopped when all the stopping criteria are met

Decision trees:

Decision tree induction is the learning of decision trees from class labelled 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 noiseor outliers in the training data [39].

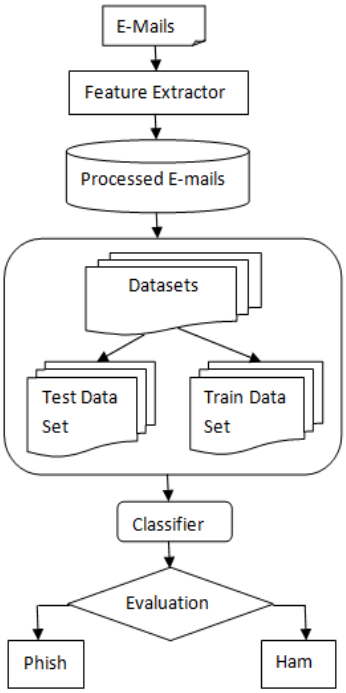


Fig4 Proposed architecture

Artificial Neural Network (ANN):

ANN is composed of a set of elementary computational units, called neurons, and 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. The weights are adaptive coefficients that, by analogy with the biological model, are modified in response to the various signals that travel on the network according to a suitable learning algorithm. A threshold value, called bias, is usually introduced. Bias is similar to an intercept in a regression model [1].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. Neural networks find applications in many different fields. Artificial Neural Network (ANN) is a large class of algorithms that has the capability of classification, regression and density estimation [41]. Main wok of our model is to combine any two methods mentioned above to get the accurate and better result.

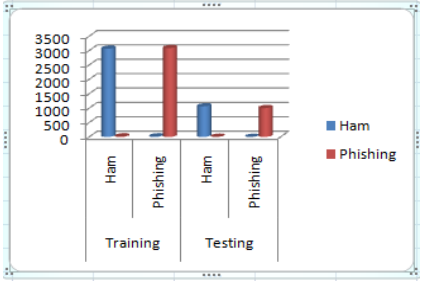
© 2016, IJARCSSE All Rights Reserved

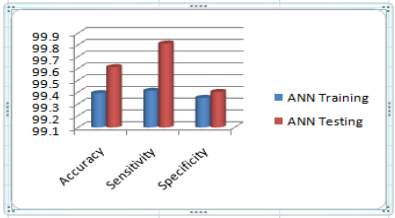
VI. RESULT ANALYSIS

Techniques	Accuracy of models	
	Training	Testing
CART	98.29	98.84
ANN	**99.39**	**99.61**

Vs Predicte	Training		Testing	
	Ham	Phishing	Ham	Phishing
Ham	3066	18	1064	2
Phishing	20	3088	6	1002

Performance Measures	ANN	
	Training	Testing
Accuracy	99.39	99.61
Sensitivity	99.41	99.81
Specificity	99.35	99.4

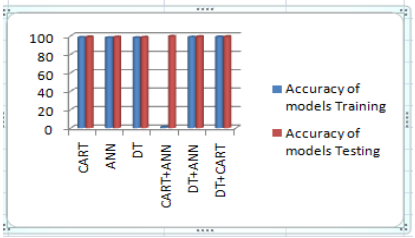


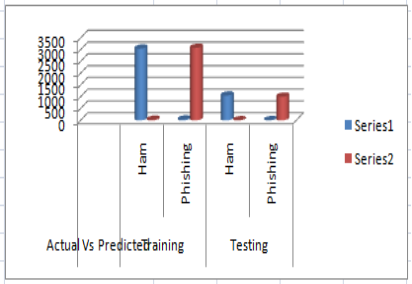


Techniques	Accuracy of models	
	Training	Testing
CART	98.29	98.84
ANN	98	98.65
DT	98	98.41
CART+ANN	**99%**	**99.32**
DT+ANN	**98.66**	**99.04**
DT+CART	**98.79**	**99.08**

Actual Vs Predicted	Training		Testing	
	Ham	Phishing	Ham	Phishing
	3049	35	1057	9
	27	3081	5	1003

Performance Measures	Random Combo	
	Training	Testing
Accuracy	99	99.32
Sensitivity	98.86	99.15
Specificity	99.13	99.5





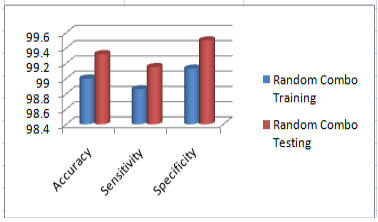


Fig 5: Result analysis is shown in different figure above

VII. CONCLUSION

In this paper we are using unsupervised learning methodology where in algorithm, technique and attacks mechanism works for better accuracy. We have got a new combination that is any of the clementine tool with unsupervised learning technique. Email Phishing is tough for hackers with this techniques for far extent as this technology mixes the combination randomly hackers need to get the combination first then apply their algorithm to hack our email which is very difficult to get. We are trying here to make our emails data secure.

REFERENCE

[1] NiharikaVaishnaw, S R Tandan "Development of Anti-Phishing Model forClassification of Phishing E-mail"., International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, Issue 6, June 2015

[3] IsredzaRahmi A Hamid and JemalAbawajy," 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, "Phishingdetection 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] AlessioPascucci," 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.

© 2016, IJARCSSE All Rights Reserved

[16] TanviChauhan, Prof.VineetRichhariya, Sunil Sharma, "Literature Report on Face Detection with Skin & Reorganization using Genetic Algorithm", TanviChauhan et al. / IJAIR Vol. 2 Issue 2 ISSN: 2278-7844

[17] AanchalChauhan ,ZuberFarooqui, "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] TanviChauhan, VineetRichhariya," 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. ShanmugaPriya , 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] JoseNazario. Phishing corpus. http://monkey.org/~jose/wiki/doku.php?id=phishingcorpus.

[24] SpamAssassin. Public corpus. http: //spamassassin.apache.org/publiccorpus.

[25] NiharikaVaishnaw, 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, SadafRahman, "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.

© 2016, IJARCSSE All Rights Reserved