An Analysis of Various Social Engineering Attack in Social Network using Machine Learning Algorithm

Dalal Al-dablan1[1†], Amal Al-hamad[1†], Raghad Al-Bahlal[1†] and Maria Altaib Badawi [2].

College of science,
Department of Computer Science and Information in Majmaah university, Al Zulfi 15941, Saudi Arabia

Summary

"Social engineering explains how one can use the human mind for capturing useful information about organizations or individuals." With tremendous growth of internet, attack cases are increasing each day along with the modern attack method, and It targets emotional parts of human to gain access to controlled area or achieve sensitive information for various purposes. Since there is neither hardware or software available to protect an enterprise or individual against social engineering, it is essential that good practices be implemented .

The overall purpose of this research is to highlight the different social engineering attacks and how they can prevent in social network because Social engineering is one of the biggest problems in social network, a concern the privacy and security. And we have another problem on social networks, that it is difficult for users to judge if a friend request is trustworthy or not, and always users of online social networks tend to exhibit a higher degree of trust in friend requests and messages sent by other users, Social engineering lead to increase the Incorporate threats, fear and a sense of urgency in an attempt to manipulate the user into responding quickly. For this purpose, we will use J48 algorithm to implement a detection algorithm for social engineering attacks in URL links. Thereafter, this project is using a set of data then analysis it using the Weka tool, to defend against these attack

J48 approach is very simple and effective in decreasing the false alarm ratio and improving the intrusion detection accuracy.

The potential estimate of J48 algorithm is to help in an effective detection of probable attacks which could jeopardies the social engineering confidentiality, also this proposed technique could classify the data as either normal or abnormal.

This project presents the algorithm, that will help create a safer and more reliable computing environment around the world for our next generation.

Key words:

Social Engineering, Social engineering attacks, J48 algorithm, Uniform Resource Locator (URL).

1. Introduction

Social engineering has become an emerging threat in virtual communities and is an effective means to attack information systems. Social engineering, also known as human hacking, is the art of tricking employees and consumers into disclosing their credentials and then using them to gain access to networks or accounts. It is a hacker's tricky use of deception or manipulation of people's tendency to trust, be corporative, or simply follow their desire to explore and be curious. Sophisticated IT security systems cannot protect systems from hackers or defend against what seems to be authorized access. People are easily hacked, making them and their social media posts high-risk attack targets [1].

It is often easy to get computer users to infect their corporate network or mobiles by luring them to spoof websites and or tricking them into clicking on harmful links and or downloading and installing malicious applications and or backdoor's.

Hackers will use any information they can grab from your social networks and will use basic human nature against you. For example, it is much easier to fool someone into giving you their password than it is for you to try hacking their password (unless the password is really weak).

Security is all about knowing who and what to trust. It is important to know when and when not to take a person at their word and when the person you are communicating with is who they say they are. The same is true of online interactions and website usage: when do you trust that the website you are using is legitimate or is safe to provide your information? The weakest link in the security chain is the human who accepts a person or scenario at face value [2].

In these days, many emails come to us from known people the emails are getting smarter. They are getting much more difficult to discern. They are become more intelligent attacks and a harder methodology for people to recognize those attacks.

simulation experiment. And then, the practical effectiveness of the proposed method is experimentally confirmed by applying it to the actual road traffic noise data.

2. Theoretical Consideration

Social engineering attacks:

can be accomplished in any place which has human interaction involved as it has divergent or various forms. The five subsequent occurring of digital social engineering strikes are:

Manuscript received October 5, 2020
Manuscript revised October 20, 2020
https://doi.org/10.22937/IJCSNS.2020.20.10.7

1- Baiting: Baiting involves a faulty assurance material or curiosity. This scheme persuades the users in such a way that they confine them and steal all their vital data or impose a malware in their system. Physical Media is the savage form of Baiting which is used to diffuse the malicious malware in the system. The targeted person clicks the bait because of his/her curiosity and then places it in work or home computer evolving it in automatic installation of malware. Tempting and Attractive advertisements which guides to harmful sites or urges the users to download a malware- infected application are the online form of Baiting Scheme.

2- Phishing: These schemes are the emails and text messages whose main concern is to promote a sense of seriousness, necessity, strangeness or panic in the targeted person. It is famous social engineering strike. This scheme prompts them to disclose or release vital information by opening links to hostile websites or clicking attachments that accommodate malware. In Phishing technique the homogeneous messages are sent to all users.

3- Spear Phishing: This is the more focused version of the phishing scheme as in this the striker selects certain people or companies. Spear phishing technique needs more attempts on part of the striker and it may take a considerate time as to pull this scheme off. These schemes are done expertly therefore making them mostly undetectable. In this the striker customizes the messages established on features, job positions and contact possession of the targeted person as to make the attack less noticeable or observable.

4- Vishing: (voice or VoIP phishing) is an electronic fraud tactic in which individuals are tricked into revealing critical financial or personal information to unauthorized entities. Vishing works like phishing but does not always occur over the Internet and is carried out using voice technology. A vishing attack can be conducted by voice email, VoIP (voice over IP), or landline or cellular telephone.

5- Pretexting: The scheme is initiated by a person pretending to need crucial information from a sufferer as to carry out an evaluative task. The striker obtains data through ingenious crafted lies. He/she identity is confirmed and through this they assemble the crucial data. The striker begins by developing as a co-worker, police, tax officials who have the authority to know things.

6- Scareware This scheme includes the victims who are flooded with flawed panic and counterfeit ultimatum. Scareware is also mentioned as deceitful software or fraud ware. It is diffused through spam emails which doles out fraudulent threats or create offers for users to buy harmful services. Users are mislead to believe that their system is damaged by the malware, persuading them to instate software that has no benefit to the person but the striker or it is a malicious malware itself [3].

In our data preparation, we have labeled the phishing URLs 1 and the benign URLs 0, in the training phase, we use one decision tree algorithm J48. This phase shows one train model, so we have evaluated one decision tree learning model in our dataset, in the testing phase we have used unknown URLs has tasted using the train model as phishing or benign.

Feature extraction:

URLs feature:

The 49 are from the literature [4], these features are shown in Taple1 Taple1.URLs Feature.

Sr.No	Feature name	Type
Features used in the literature		
1.	NumDots	numeric
2.	SubdomainLevel	numeric
3.	PathLevel	numeric
4.	UrlLength	numeric
5.	NumDash	numeric
6.	NumDashInHostname	numeric
7.	AtSymbol	numeric
8.	TildeSymbol	numeric
9.	NumUnderscore	numeric
10.	NumPercent	numeric
11.	NumQueryComponents	numeric
12.	NumAmpersand	numeric
13.	NumHash	numeric
14.	Num Chars	numeric
15.	No Https	numeric
16.	RandomString	numeric
17.	IpAddress	numeric
18.	DomainInSubdomains	numeric
19.	DomainInPaths	numeric
20.	HttpsInHostname	numeric
21.	HostnameLength	numeric
22.	PathLength	numeric
23.	QueryLength	numeric
24.	DoubleSlashInPath	numeric
25.	NumSensitiveWords	numeric
26.	EmbeddedBrandName	numeric
27.	PctExtHyperlinks	numeric
28.	PctExtResourceUrls	numeric

29.	ExtFavicon	numeric
30.	InsecureForms	numeric
31.	RelativeFormAction	numeric
32.	ExtFormAction	numeric
33.	AbnormalFormAction	numeric
34.	PctNullSelfRedirectHyperlinks	numeric
35.	FrequentDomainNameMismatch	numeric
36.	FakeLinkInStatusBar	numeric
37.	RightClickDisabled	numeric
38.	PopUpWindow	numeric
39.	SubmitInfoToEmail	numeric
40.	IframeOrFrame	numeric
41.	MissingTitle	numeric
42.	ImagesOnlyInForm	numeric
43.	SubdomainLevelRT	numeric
44.	UrlLengthRT	numeric
45.	PctExtResourceUrlsRT	numeric
46.	AbnormalExtFormActionR	numeric
47.	ExtMetaScriptLinkRT	numeric
48.	PctExtNullSelfRedirectHyperlinksRT	numeric
49.	{0,1}	nominal

Decision tree learning:
Decision Tree is the most widely used tool for decision making. To accomplish this one should draw a decision tree with different branches and leaves [5]. Linear Classifiers Is to use an object's characteristics to identify which class (or group) it belongs to [6].

Decision Tree algorithms in Machine Learning:
There are available data mining algorithms classification based on Bayesian classifiers, Artificial Neural Network and Decision tree. Decision tree method is the simplest and most widely used algorithm from data mining algorithms also it's easy for the user understanding and decision making. In decision tree algorithms there are different accuracy and cost effectiveness, it is very important to know which algorithm to use. "To accomplish this one should draw a decision tree with different branches and leaves. These branches and leaves should point to all various factors concerning a particular situation [5]. Decision tree it is one way to display the algorithm depends on situation and desired outcome, it is like a support toll.

J48 algorithm:
J48 is a Supervised Classification decision tree algorithm in ML, J48 is an extension of ID3. The additional features of J48 are accounting for missing values, decision trees pruning, continuous attribute value ranges, derivation of rules."[7].
"the J48 algorithm can divide the training data into many subsets which correspond to the various values of a chosen feature and this process is repeated for every subset till every subset is assigned to one class.

The J48 algorithm uses an enhanced technique of tree pruning for decreasing the mis-classification error. Furthermore, this algorithm also used a greedy divide-and-conquer method for recursively inducing decision trees containing the database/dataset attributes for further classification. In any decision tree, classification is a major performance parameter. The classification error can be defined as the percentage of the misclassified cases.[8].
This J48 classifier algorithm can develop its decision tree depending on the information of the theoretical attribute values of the present training data. Also, in the case of a J48 algorithm, every feature or attribute separately estimates the gain value and the calculation process is continued till the prediction process is completed.

Basic Steps in the Algorithm:
1. In case the instances belong to the same class the tree represents a leaf, so the leaf is returned by labeling with the same class.
2. The potential information is calculated for every attribute, given by a test on the attribute. Then the gain in information is calculated that would result from a test on the attribute.
3. Then the best attribute is found on the basis of the present selection criterion and that attribute selected for branching.

Counting Gain:
This process uses the "Entropy" which is a measure of the data disorder. The Entropy of is calculated by And Gain is "[7].

$$Entropy(\vec{y}) = -\sum_{j=1}^{n} \frac{|y_i|}{|\vec{y}|} \log\left(\frac{|y_i|}{|\vec{y}|}\right)$$

$$Entropy(j|\vec{y}) = \frac{|y_j|}{|\vec{y}|} \log\left(\frac{|y_j|}{|\vec{y}|}\right)$$

$$Gain(\vec{y}, j) = Entropy(\vec{y} - Entropy(j|\vec{y}))\ |$$

Tool:
we have used Weka in our algorithm to gain the final decision on, weather the URL is benign or phishing.

3. Experimental Consideration:

Data source and dataset:

The dataset is from Mendeley Data contains 48 features extracted from 5000 phishing webpages and 5000 legitimate webpages. An improved feature extraction technique is. And this is the features that in the set.
Taple2. The dataset.

Task	Benign	Phishing	Total
Training	5000	5000	10000

Evaluation results:
we have evaluated the J48 decision tree algorithm classifier in our dataset.

Modify of features:
we find that when we are filtering our dataset with Supervised Attribute selection and evaluated by using info gain, also we delete one feature, the accuracy is increase.

Table 3. Feature deleted

Sr.No	Feature name	Type
5.	NumDash	numeric

We delete the number of dashes in the URLs because it is effect directly in the accuracy and the mistakes become less in our results.

Accuracy

Taple4. Accuracy.

Classifier	Accuracy without modifying	Accuracy with modifying	Change %
J48	98.84%	99.01%	0.17% ↑

4. Conclusion

In this paper, we have performed the detecting of URLs as benign or phishing. And how our modifying of the features effect on it. We have a prepared dataset 10000 URLs, among wish 4958 are benign and 4942 are phishing.
Also, we talk about social engineering and the effect of it in society.

Acknowledgment

First, the success of this project needs guidance and assistance. We are very lucky because we got this assistance during the completion of this project. we thank Allah for everything that made us capable of accomplishing this paper.

Many thanks to our project supervisor Dr. Maria Badawi, for sharing her expertise and helpful guidance, we are grateful to her. also, to our family who support and help us during this time, we are thankful to you.

References

[1] Conteh,N,Y. & Schmick,P,J . (2016). Cybersecurity: risks, vulnerabilities and countermeasures to prevent social engineering attacks. International Journal of Advanced Computer Research. http://dx.doi.org/

[2] (what is social engineering?) https://www.webroot.com/ca/en

[3] M. NazreenBanu et al, A Comprehensive Study of Phishing 1.Attacks‖ , IJCSIT, Vol. 4, 2013, pp.783-786.

[4] Phishing Datset,data.mendeley.com

[5] Navada, A, Ansari, A, N, Sonkamble,B (2011), Overview of Use of Decision Tree algorithms in Machine Learning, IEEE, 10.1109/ICSGRC.2011.5991826.

[6] Yuan,G, Ho,C, and Lin,C(2012),Recent Advances of Large Scale Linear Classification,IEEE,10.1109/JPROC.2012.2188013.

[7] Mazraeh, S., Modhej, A., Neysi,(2016), S.H.N.: Intrusion detection in computer networks using combination of machine learning tech- niques. Int. J. Comput. Sci. Netw. Secur. (IJCSNS) 16(8), 122 (2016).

[8] Medhat, K, Ramadan, R.A, Talkhan, I (2017), Security in mission critical communication systems: approach for intrusion detection. In: Multimedia Services and Applications in Mission Critical Communication Systems, pp.270–291. IGI Global (2017), 10.4018/978-1-5225-2113-6.ch012.