PHISHING E MAIL DETECTION IN E BANKING USING DATA MINING TECHNIQUES

AL.Hanees[1] and P.Thadshayini [2]

[1,2] Department of Mathematical Sciences, Faculty of Applied Sciences, South Eastern University of Sri Lanka.

alhanees@seu.ac.lk

Abstract

Phishing is a type of social engineering attack often used to steal user data. Especially in e-banking attackers sent an email that appears legitimate but is actually meant to lure a potential victim into providing some level of personal information for nefarious purpose, including login credentials and credit card numbers. In this paper we employ four data mining classification algorithms to detect the phishing emails in e-banking and then we use the weighted majority vote ensemble method to improve the detection of phishing emails and compare the performance of each. Experimental results shows that decision tree builds the best classifier.

***Keywords**:* Phishing emails, E-Banking, Ensemble, Weighted majority vote, data mining algorithm

Introduction

Phishing is a form of fraud and type of social engineering attack in which cyber criminals trick users into handing over sensitive information. Phishing attacks typically rely on social networking techniques applied to email or other electronic communication methods, including direct messages sent over social networks, SMS text messages. Phishers may use social engineering and other public sources of information, including social networks like Facebook, to gather sensible and believable background information about the user's personal. And also criminals gather the information about their co-workers and their information to make the users to pay their attention in that particular site. Attackers create a fake messages with collected believable information looks like a legitimate one.

Most of the phishing attacks occur through the email. Email phishing is a numbers game. An attacker sending out thousands of fake emails. They include about something like "if you win we will give gift vouchers" and "you got Rs.100000. To deliver the check fill the application". At least one out of thousand user will fall for their scam. There are some techniques attackers used to increase their success rate. For one, they will do whatever to design the phishing messages to look like an actual email from authorized organization. Using the same words as the authorized organizations usually use, typefaces, logos and signatures makes the messages appear legitimate. In addition attackers use the sentences when the users read that they will identify its urgent and pushing them to do it immediately. A spoofed messages often contains subtle mistakes that expose its true identity. But when the users receive the emails, 23% of users only consider about the site which the email to clarify whether the email is legitimate or not [1].

Banks offer convenient and flexible service to their customers to perform their financial transaction in order to increase the scope of the bank that is known as E-banking. Most of the phishing attacks occur in e-banking through the email. Attackers sent the fraudulent email that appears from the bank and the customers will process the contents included in the mail. It makes easy for attackers to collect the sensitive information of users.

According to above observation this paper presents work on detecting phishing emails in E-banking using data mining algorithms. We try to analyze the phishing emails from legitimate emails using four data mining algorithms and ensemble them using weighted majority vote method to improve the detection mechanism.

The rest of the paper is structured as follows: Section Ⅱ provides the previous approach for phishing emails; Section Ⅲ gives the details of data mining algorithms; Section Ⅳ gives the details of ensemble method; Section Ⅴ includes experimental result of phishing emails; Section Ⅵ concludes the work and direction for future work.

Related work

Phishing is a popular technique to thieve the user's information in the social network. There are many types of phishing attacks. Such as web Trojan, data theft, etc. Also to overcome these problem there are some techniques in [3]. Filtering the scam and spams using three learning methods and ensemble the result using majority vote method. The closest related method to our work is [2]. [5] develop an end user application to prevent from phishing attack and it act as an interface between the emails communications.

Data mining algorithm

A data mining algorithm is a set of examining and analytical algorithms. It help to create a model of data. To get a concrete model algorithm analyze the data that can finding a particular type of pattern from the data which we provide. The aim of this algorithm is an analysis of different more iterations. They can help in finding optimal and perfect new solution for mining model. These sets of parameters can be applied for the entire data set. They help in extracting the specific patterns and getting a detailed statistic of the data.

We are analyzing the accuracy of phishing emails and legitimate emails on the dataset. So it is a binary classification. In this paper we employ four data mining algorithms: Support vector machine (SVM), Decision tree, Naive Bayes probabilistic theory and K-nearest neighbor. Using these algorithms analyzing the dataset to find the accuracy of phishing emails. Then using the ensemble method to improve the detection mechanism.

i. Support vector machine (SVM)

Support Vector Machine (SVM) is a supervised machine learning algorithm. It can be used for both classification and regression. However, it is mostly used in classification problems. In this algorithm, every data item will be plot as a point. It will process in n-dimensional space (where n is number of features) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the accurate plane that differentiate the two classes very well.

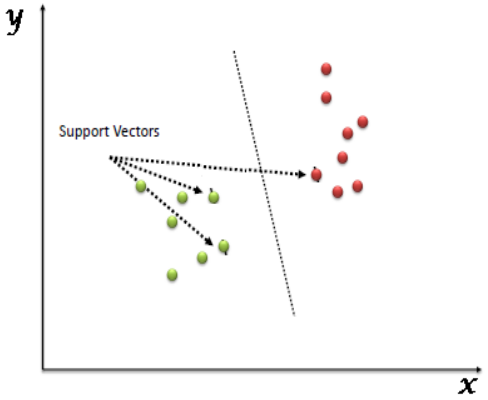

Figure 1

ii Decision tree

Decision tree can be used for both classification and regression. Algorithm breaks the dataset into smaller parts and again it breaks it into smaller parts and develop them in tree format. Finally the tree has leaf nodes and decision nodes. A decision node can have two or more branches. Leaf node cannot have branches it represents a classification or decision. The topmost decision node in a tree is called root node. We can use decision tree for both numerical data and classified data.

A decision tree is built top-down from a root node and divide the data into smaller parts that contain same and similar value instances. Entropy is used by the decision tree to calculate the similarity of the smaller parts of the dataset. If every parts of the dataset are similar to each other the entropy is zero and if the dataset is divided into equal smaller parts it has entropy of one.

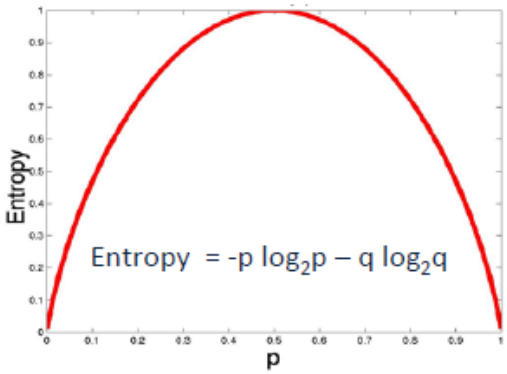

Figure 2

There is frequency table to calculate the entropy. Using that table we want to calculate two types of entropy to build a decision tree as follow:

1. Entropy using the frequency table of one attribute:
   $\boldsymbol{E(S)} = \sum_{i=1}^{c} -\mathrm{p}_i \log_2 p_i$ (01)

2. Entropy using the frequency table of two attributes:

   $$\boldsymbol{E(T,X)} = \sum_{c\in X} \boldsymbol{P(c)E(c)} \quad (02)$$

iii. Naive Bayes probabilistic theory

Naive Bayes is one of the probabilistic algorithm. It can be used for classification problems and also it is a classification technique based on Bayes' Theorem. It need some assumption based on the independence of the variables.

$$P(A/B) = \frac{P(B/A)\,P(A)}{P(B)} \quad (03)$$

$P(A)$ is the prior probability of A occurring independently.

$P(B)$ is the prior probability of B occurring independently.

$P(A/B)$ is the posterior probability that A occurs given B.

$P(B/A)$ is the likelihood probability of B occurring, given A.

iv. K-nearest neighbor

The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement and supervised machine learning algorithm. KNN can be used for both classification and regression problems.

Let m be the number of training data samples. Let p be an unknown point.

1. Store the training samples in an array of data point's arr []. This means each element of this array represents a tuple (x, y).
2. for i = 0 to m
   Calculate Euclidian distance d (arr[i], p).
3. Make set S of K smallest distances obtained. Each of these distances corresponds to an already classified data point.
4. Return the majority label among S.

III. Ensemble method

Ensemble methods is a machine learning technique. Ensemble model can combine the base model which are not same. It can improve the performance of the model. It can produce optimal and efficient model. There are many ways for ensemble [2]: Majority voting, Weighted majority voting, Naïve Bayes Combination and N dimensional Naïve Bayes sampling. In this paper we employ weighted majority voting for each classifiers.

Weighted majority voting

Performance is the use of voting and the weighted voting is an extension of simple voting [8]. We can compute a weighted majority vote by associating a weight wj with classifier Cj:

$$\hat{y} = \arg\max_i \sum_{j=1}^{m} w_i X_A(C_j(x) = i) \qquad (04)$$

Where X_A is the characteristic function,

$[C_j(x) = i \in A]$, and A is the set of unique class labels.

IV. Experimental Result

In this section we study the result. We describe first how we collect the dataset of phishing emails. We use the WEKA data mining tool to analyze the dataset. Then we analyze the dataset using four algorithms and finally ensemble those algorithms to find the best classifier for detection phishing emails.

i. Dataset

We used the phishing dataset from Phish tank dataset repository. Those dataset gathered by considering especially E-banking and E-commerce. It consist 1353 instances.

ii. Support vector machine (SVM) classifier

First the data set is divided into two classes as phishing emails and legitimate emails. To evaluate the data we employed support vector machine classifier with 10-fold cross validation mechanism. The result of the support vector machine is shows that 86.031% of instances are correctly classified. The result is shown in table 1.

iii. Decision tree classifier

Decision tree evaluate the instances on dataset according to binary mechanism. It will provide most accurate result. Because it will analyze every instances individually. We employed decision tree classifier with 10-fold

cross validation mechanism. The result is shows that 87.5831% of instances are correctly classified. The result is shown in table 1.

iv. Naïve Bayes classifier

Naïve Bayes classifier analyze the data using probabilities of phishing emails and legitimate email on total instances. We employed Naïve Bayes classifier with 10-fold cross validation mechanism. The result shows that 84.1094% of instances are correctly classified. The result is shown in table 1.

v. KNN classifier

KNN classifiers evaluate the data using the number of neighbor to get the accurate result. That means the accuracy depend on the number of neighbor. We employed KNN classifier with 10-fold cross validation mechanism. The result shows 88.3222% of instances are correctly classified. The result shows in table 1.

vi. Ensemble classification

We employed weighted majority voting ensemble method with iterations for each algorithms to improve the accuracy. Iterations makes to reduce the percentage of error in analyzing. The result of the algorithms with ensemble method is shown in table 2.

Table 1: using only data mining algorithm

Classifier	Phishing emails	Legitimate emails
SVM	1164	189
Decision tree	1185	168
Naïve Bayes	1138	215
KNN	1195	158

Table 2: using ensemble algorithm

Classifier	Phishing emails	Legitimate emails
SVM	1164	189
Decision tree	1217	136
Naïve Bayes	1138	215
KNN	1190	163

Conclusion and future work

In this paper we have presented an approach to detecting phishing emails in E-banking using data mining methods. We have used four well known data mining algorithms Support vector machine, Decision tree, Naïve Bayes and K-nearest neighbor.

We have employed weighted majority vote ensemble algorithm to improve the performance of detecting mechanism. The weighted majority vote algorithm uses the iterations to reduce the errors and then finalized the best accuracy in detecting for each algorithms. We have used the WEKA data mining tool for our practical analysis. The experimental result shows that the decision tree gives good accuracy with ensemble algorithm than other algorithms.

For the future work other types of ensemble methods could be employed to get more efficient accuracy in detecting phishing emails.

References

[1]. Liping, M., Bahadorrezda, O., Paul , W., & Simon , B. (2009). Detecting Phishing Emails Using Hybrid Features. 493-497.

[2]. Alhuseen , O., & Anwar, L. (2017). E-Banking Security: Internet Hacking, Phishing Attacks, Analysis and Prevention of Fraudulent Activities . *International Journal of Emerging Technology and Advanced Engineering* , 109-115.

[3]. Alireza , S., Mojtaba , V., & Behrouz , B. M. (2007). Learn To Detect Phishing Scams Using Learning and Ensemble Methods. 311-314.

[4]. Atiya , K., Farheen , M. M., Gauri , S., & Rasika , R. (2017). Detecting E Banking Phishing Websites Using Associative Classification. 261-264.

[5]. Daniel, B., & Aryeh, K. (n.d.). Consistencyofweightedmajorityvotes. 9.

[6]. James, L., Jason , L., & Anthony , B. (2019). Aprobabilisticclassifierensembleweightingschemebased oncross-validatedaccuracyestimates. *Springer*, 1674-1709.

[7]. Kashif , R., & Phaneendra P, D. (2015). IMPLEMENTATION OF METHODS FOR TRANSACTION IN SECURE ONLINE BANKING. 41-43.

[8]. Kuo, W. (2014). On Adjustment Functions for Weight-Adjusted Voting-Based Ensembles of Classifiers . *JOURNAL OF COMPUTERS*, 1547-1552.

[9]. Madhusudhanan, C., Ramkumar , C., & Shambhu , U. (2006). PHONEY: Mimicking User Response to Detect Phishing Attacks. *IEEE*.

[10]. Maher , A., Hossain , M., Keshav , D., & Fadi , T. (2009). Modelling Intelligent Phishing Detection System for e-Banking using Fuzzy Data Mining . 265-272.

[11]. Netra , P. (2007). Online frauds in Bank with phishing. *Journal of Internet Banking and Commerce* , 1-27.

[12]. Remya , K., & Ramya, J. (2014). Using Weighted Majority Voting Classifier Combination for Relation Classification in Biomedical Texts . *IEEE*, 1205-1209.

[13]. Sammar, M., Moustafa, Y., Nagwa, E., & Mohamed, S. (2018). Software bug prediction using weighted majority voting techniques. *ELSEVIER*, 2763-2774.

[14]. Suganya , V. (2016). A Review on Phishing Attacks and Various Anti Phishing Techniques. 20-23.

[15]. Ying , P., & Xuhua, D. (2006). Anomaly Based Web Phishing Page Detection.