IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 20, Issue 2, Ver. IV (Mar. - Apr. 2018), PP 59-63
www.iosrjournals.org

An Associative Classification Approach Using Data Mining for Predicting Phishy Email

Nashwa khair[1], M. Daiyan[2], S. K. Tiwari[3],M. Aftab Alam[4]

Abstract*—Due to rapid development of internet peoples are more dependent on Internet. Email messages is one the communication medium over internet. Some messages are true and some are fake. In this we developed a method to predict Phishy email. Our experimental results shows that results are very effective over existing method.*

Keywords*—Data Mining, Machine Learning, Pattern Matching, Weka*

Date of Submission: 03-05-2018 Date of acceptance: 25-05-2018

I. Introduction

E-mail has become an effective way for the time being. People sends and receives many letters daily, and communicate with others, or share files and information. Phishing charges using email are the most common of electronic crime [1, 2].It is one of social engineering techniques used to get advantage of human unawareness. It allows abusive people to utilize the weaknesses in web security technologies which try to get confidential and private information, such as username, password, financial account credentials and detail of credit card, by veiling as a proper object in the email.

Dizzy Internet user may easily be deceived by this type of the scandal. Victims of phishing e-mail may lose the details of the account bank, passwords, number of credit card, or other private information to the phishing e-mail senders. In addition phishing is considered as spam; while it is being differs from spam [3]. Indeed spam almost seeks to sell a product or service, while a phishing message try to look like it is a form of legitimate organization. Straightforward, approaches that are handled the spam messages cannot be used to phishing messages [4, 3].

The reports of Anti Phishing Working Group (APWG) for the first quarter on 2012 [1] shows a terrible number of phishy emails attack, there were on average 28,481 unique reports of phishy emails (campaigns) received by APWG g from internet users. The email campaign is a unique email is sent out to multiple users mislead them accessing a specific phishy website. Undeniably, financial facilitiescreate to be the most-targeted industry subdivision in the first quarter of 2012. Furthermore, Payment amenitiesconcealed retail/amenities have the second-highest business sector for besiegedbouts (Activity, 2012).

FraudAction Research Labs divulge that, phishing attacks in the first six month of 2012 have been increased compared by the same period of 2011 (Kovacs, 2012).

II. Literature Review

Several ways are considered to pounce upon phishing. These rotate from communication-oriented methods like verification protocols over banning to content-based cleaning technique [5]. Currently the first two methods not muchimplemented. Blacklists, Whitlists or both of them are not that qualified when used in different filters approaches, while continuously a fresh phishy scam is engendered. In fact, these filters are dropped in scalable problems. Consequently content-based phishing filters are requisite and ample used. Soresearchers focus on machine learning and data mining techniques to carry this problem based on the email contents in the header and on the body of an email.

If these methods can completely eliminate the phishing ,hereit would be no need for other protection technologies .Though, current tools are unuseful to discovery phishing e-mail and phishing sites with a 100% of accuracy [6]. For example, in 2007, study showed that over 20% of phishing websites missed by the best anti-phishing toolbars and study in 2009 found that most anti-phishing tools did not start blocking phishing websites until several hours or days after the fraudulent e-mails were send to attract users to these websites [7].

1. Nashwa khair, M.tech Scholar, Mazooncollege,Oman
2. Dr. M. Daiyan, Assistant professor,CS&IT,daiyan.mohd@gmail.com
3. Dr. S.K. Tiwari, Associate Professor, MagadhUniversity,Bihar,India tiwari.dr.sanjay@gmail.com,
4. Dr. M. AftabAlam, Assistant Professor, Cs&IT, md.aftab05@gmail.com

Researchers have developed filters depended on different techniques to eliminate the phishy email problem based on traditional techniques such as network level protection and authentication protection as well as on modern techniques using machine learning and data mining approaches.

Black list is a network level protection that used as a filter to classify email as phishy or legitimate. This technique is examined the sender's address, IP address or DNS address by extracting these data from the email header with predefined list. If any one of these data has match what in the list it is rejected [8], then this email is classified as phishy so it is not received by their recipient [9]. Internet Server Providers (ISP) is the one who responsible for applying this procedure.

White List Filter This type of filters is connected with lists having static IP addresses for legitimate domains at network level protection. It is compared the email address with the IPs addresses that are found in the white list [10]. If the matching step handled a positive result then the email bypasses the filter and goes to the receiver's inbox. This white list is filled by emails received from legitimate companies or people who agree their addresses or IP addresses to be included in the list so this way the sender's identity always is known. This filter is categorized as legitimate emails classifier because it is based only on the legitimate address. However, every email is not considered in the white list is classified as a phishy email leading to suffer or incur loss significant emails. [11].

Pattern Matching Filters are based on some concepts that classify the email as either phishy or legitimate at network level protection, by detect for particular words ,character sets, text strings in the content of email, sender and subject . But there is huge number of positives false in the result since many emails contain banned words or text strings [12].

Automated classifiers are classifiers and a server side filters based on machine learning and data mining approaches. Extracting different features from the email header and body is the current method followed to process the classifier depending on them allowing distinguish between emails if phishy or legitimate. Two most applied approaches types of classifiers, statistical such as Support Vector Machine (SVM). and Bayesian Second type is multi-layer classifiers such as Decision Trees (DT) and neural network.

III. Purposed model

The proposed detection model is shown in Figure 1. A collection of both, phishy and legitimate emails are used to obtain the significant features belonging to the phishy and legitimate emails permitting us to focus on them in the training step.

The proposed algorithm is trained on the dataset and using the specified thresholds to generate the knowledge (rules). Differently from other AC algorithms SCPE scan the dataset one time only by employing an intersection method based on Tid-list to enumerate the location, support and confidence of mining items inside the training dataset. The Tid-list offers a representation of the dataset having all necessary information related to each item (attribute value).

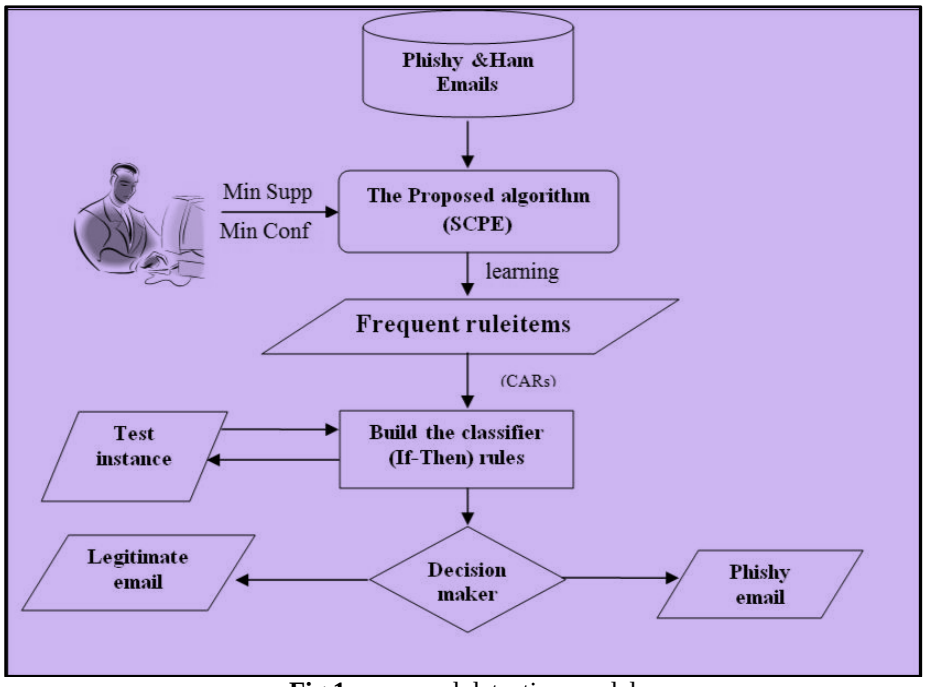

Fig 1- purposed detection model

Feature Assessment

In this assessment, the email features are used to classify the type of emails(phishy or legitimate). These features are divided into two types "nominal" and "continuous". Therefore, according to Figure 1, which shows the discrete step is done only for the continuous features to be nominal one since we are dealing with classification dataset. This has been performed using the multi-interval discretization technique [13] in WEKA software.

Firstly, the incessantcharacteristic is sorted in rising order with the class values related with the instance belonging to it. Then, breaking points is placed when the class ratevariations to compute the gen gain for each conceivable breaking point. The information gain represents the amount of information claimed to an attribute value with respect to its gain. Finally, the breaking point that minimizes the information gain over all possible breakingpoints is nominated and the algorithm is triggered again on the lower range of that attribute.

Table 1.0 Algorithm SCPE learning algorithm

Input: Training data (T), minsupp and minconf thresholds
Output: A set of CARs

Preprocessing phase
Discretise continuous columns

The Algorithm

Scan T for the set R of frequent one attribute-value Do
For each pair of disjoint items V1, V2 in R

Intersect the sets of rowIds of V1, and V2 and store it in Ts If Ts size <itemsupp then

prune the new item else

begin
If (<V1 V2>, ci) passes the minsupp threshold
begin
if (<V1 V2>, ci) passes the minconf threshold begin
Generate a rule for <V1 V2> if it passes
R R <V1 V2>
end if

else discard the new item. end if

endif end

Table 1.1: Sample of feature frequency analysis function result for "SubjReply"

APPEARANCE	Class	Frequency
'1'	Phishy	44
'0'	Phishy	656
'1'	Legitimate (Ham)	89
'0'	Legitimate (Ham)	211

Rule Learning

Our algorithm is a private case of association rule that take into only the class label as a resultant of a rule to deduce a set of class association rules (CARs) from the training dataset which meet many user-constraints, minimum (confidence and support) thresholds. Our algorithm is a private case of association rule that take into only the class label as a resultant of a rule to deduce a set of class association rules (CARs) from the training dataset which meet many user-constraints, minimum (confidence and support) thresholds.

To show how we determine a frequent rule item, consider for instance item sets in Table 1.2< (Body Html, 1)> and < (EmailFunctionWords,1)>. The next two sets represent the sequences in which they occur, {1, 3, 4, 5, 6, 9, 10} and {1, 2, 4, 5, 8, 9, 10}. We can determine the support of the itemset< (Body Html, 1)>, <

(EmailFunctionWords, 1)> by intersecting the sequences sets for itemsets< (Body Html, 1)> and < (EmailFunctionWords, 1)>. The cardinality of the resulting set {1, 4, 5, 9, 10} represents thesupportforitemset< (Body Html, 1)>, < (EmailFunctionWords, 1)>, i.e. 5/10. If it passes the minimum supportthreshold, then we proceed by checking whether there is some class C such that < (Body Html, 1)>, < (EmailFunctionWords, 1)>C> passes the minimum supportthreshold, otherwise we prune it.

Table 1.2: Part of training data

Sequence	Body Html	EmailFunctionWords	Class
1	1	1	Phishy
2	0	1	Legitimate
3	1	0	Phishy
4	1	1	Phishy
5	1	1	Phishy

Rule Ranking

Ranking of generated rules reflects the strength of the classifier since in this step we use the selected rules to predict test instance in later phase. SCPE usually focuses on the superior rule in the final classifier. The superior rules are ones with large number of attribute not only the ones with high confidence values against the training dataset.

Our contribution in ranking is to deal with preferring superior rules that have maximum number of attributes. So our algorithm prefers specific long rules over general short rules in antecedent side.

So we follow the following procedure when two or more rules having same length: R1 > R2 if the confidence of R1 is greater than R2. When the confidence values of R1 and R2 are the same, but the support of R1 is greater than R2 the algorithm tend to favor R1. R1 precede rule R2 In the case of confidence and support values for both of them are the same, R1 is generated before R2 so R1 is preferred. As a result, there is minimal chance to choose any rules randomly which may led increasing classification error rate.

Rule Pruning

After rules ranking the pruning will start to choose only effective rules in the classifier. Now, the rules are ranked, then starting with superior rules if it covers at least one training instance it will be inputted into the classifier. The rule is pruned when it fails to classify at least a single instance. This way the algorithm is eliminating any rules that are redundant or contribute to incorrect classification. Our algorithm applies partial matching as new criteria if full matching of the candidate rule body and the training set is not met. This technique let the classifier contains less number of rules because a rule now has more training instance coverage.

Classifier Builder

SCPE deal with a rule can cover at least one training instance. After rules are generated, Table 1.1 presents the classifier builder algorithm used by SCPE:

Starting with the first rule ri, we fully match it with training rule if it classified a single rule at least, we simple delete its occurrences from training data (sequences) container, Tiand delete it from R' then put it into classifier Cl. Then For each other potential rule ri, we check if it partially covers at least one training instance into Tithen this rule will be inserted into the classifier. Finally, we choose a default class by subscribing the labels in Ti from C container which has all labels for the whole dataset or by take the majority class as a default class from the current Cl and add it to Cl as a default class if Ti is empty. Equation 1.1 represents the time complexity forthe classifier building phased which is used by SCPE.

Prediction of Test Data

In data mining, predictions is the process that forecast the class label for unseen instance which is ultimate goal of classification. To illustrate the idea, let R be the set of generated rules and Ts be the set of test data instance to classify a test instance, the proposed algorithm used a simple way, which states that the first rule in the set of already ranked rules that contained in the test instance classifies it. If there isn't rule do completely

matching the test instance, SCPE uses a new process to find all rules that match part of the test instance and calculates the average confidence of rules that have the same class then applies the class of the high average of confidence. So the prediction here is based on mathematical formula (1.0) below where conf is the rule confidence and ri is the number of rules belonging to the same class.

$\sum ri\ (conf)/ri$ .....(1.0)

This way we ensure the given class is dependent on a number of rules applicable to the test instance. Also more than one rule is playing a role to classify test instances which is surly better of using one rule as MCAR and CBA. In cases where no instructionequals the test instance, the evasion class is allocated to the check instance.

IV. Results & Discussion

To classify a test instance, the proposed algorithm used a simple way, which states that the first rule in the set of already ranked rules that contained in the test instance classifies it. If there isn't rule do completely matching the test instance, SCPE uses a new process to find all rules that match.

The proposed model is an AC approach that operates two steps (rule ranking, rule pruning). We have used a new ranking that prefers superior rules and then enhanced rule pruning by using partial match between the rule's body and the training instance. Moreover, the prediction phase is enhanced by using group of rules average confidence for predicting the test data instances. SCPE offers also vertical data format which let the scans over the training dataset to be only one time during learning rules differently from most AC algorithms. Then a comparison is done with different classification algorithms.

For an experimental work we collect data from various sources. Weka used tools is used for result analysis.

V. Conclusion

We have proposed a new algorithm using AC approach to handle the phishy emails problem. We have gathered phishy and legitimate emails to train our model and test the applicability of AC in this kind of problem. This problem has many effects in our life causing different kind of losing. There are many exists solutions based on many sciences like statistical and probability, these solution did not have a 100% accuracy. However, we are achieved higher accuracy rate between different algorithms, Naïve Bayes, J48 and Prism which reflect that the AC approach is effective to deal with such problem. As a future work, the usage of genetic algorithm could offer extracting more features

References

[1]. Activity, P., & Report, T. (2012). Phishing Activity Trends Report 1 Quarter

[2]. Iraqi, Y &Khonji, M.,. (2011).Lexical URLs analysis for separate legitimate and phishing e-mail messages. Secured and Internet Technology , 11–14.

[3]. Webb, S Irani, D.,.,Pu, C. &Giffin, J., (2008). Evolutional Study of Phishing.University of Computing Georgia Institute of Technology, Georgia..

[4]. SonicWall. (2008). Applied Bayesian Spam Classification to Phishing E-Mail. 2008 SonicWall.

[5]. Paaß, G., &Bergholz, A. (2009).AntiPhish - Machine Learning for Phishing Detection., Project Exhibition at ECML/PKDD 2009.7-8

[6]. Sheng, S., Kumaraguru, P,.,Cranor, L. F., Hong, J &. Acquisti, A (2010).Learning Johnny not to down for phish.ACM Transactions online

[7]. Parmar, A. (2013). securing against spear-phishing. Computer Security & Fraud January 2013.

[8]. Sheng, S., Kumaraguru, P,.,Cranor, L. F., Hong, J &. Acquisti, A (2010).Learning Johnny not to down for phish.ACM Transactions online.

[9]. Paaß, G., &Bergholz, A. (2009).AntiPhish - Machine Learning for Phishing Detection., Project Exhibition at ECML/PKDD 2009.7-8.

[10]. Han, W., Le, Y & Cao, Y.,. (2008). Anti-phishing tools depends on automated individual white-list. Proceedings of the 5th ACM workshops on Digital identity management DIM '08, 51., New York, United State America : ACM Press.

[11]. Reichartz, F., Strobel, S &Paaß, G.,. (2008).Developed Phishing Detection by Model-Based Features.Fraunhofer IAIS SchloßBirlinghoven 53754 St. Augustin, Germany.

[12]. Shalendra, C. (2005). PhishingSpam, and E-mail Hacking. University of California.

[13]. K. I. andFayyad, U., (1993). Multi- clincherdiscritisation of continues-valuedattributes for classification learning, Californiainstitutetechnology, USA 1022-1027.

Nashwa khair "An Associative Classification Approach Using Data Mining for Predicting Phishy Email." IOSR Journal of Computer Engineering (IOSR-JCE) 20.2 (2018): 59-63