A MULTI-CLASSIFIER PREDICTION MODEL FOR PHISHING DETECTION

Sarju S[1], Riju Thomas[2], Emilin Shyni C[3]

[1, 2] PG Scholar, [3]Associate Professor, Department of Computer Science and Engineering, KCG College of Technology, Tamilnadu, India

Abstract

Phishing is the technique of stealing personal and sensitive information from an email client by sending emails that impersonates like the ones from some trustworthy organizations. Phishing mails are a specific type of spam mails; however the effects of them are much more terrible than alternate sorts. Mostly the phishing attackers aim the clients of the financial organizations, so its detection needs high priority. Lots of research activities are done to detect the phished emails, in the proposed methodology a multi-classifier prediction model is introduced for detecting phished emails. Our contention is that solitary classifier prediction might not be satisfactory to urge the clearest picture of the phishing email; multi-classifier prediction has accuracy 99.8% with an FP rate of 0.8%.

Keywords: *Phishing email detection, Machine learning techniques, Multi-classifier prediction model, Majority voting*

----------***----------

1. INTRODUCTION

Nowadays the emails turn into one of the generally utilized communication medium within the globe. Because of the fame of emails, the attackers utilized it to snatch the client data. Phishing messages are a particular kind of spam mail, which is used to take the individual and fiscal data from the email clients. Generally the attackers send an email that looks like legitimate messages from some reputed organizations, which lead clients to phishing sites. Phishing sites always have a user entry form, when he enters his data like user names, passwords and credit card details which in turn utilized by the attackers to do some deceitful exercises. As per the latest report issued by APWG [1] (Anti-Phishing Working Group), amount of phishing attack evolved and burgeoned in overabundance of 20% in 2013. Latest Trends Report of APWG quotes that the overall number of distinctive phishing internet sites rose to 143,353 throughout the July-September that is over the past quarter's 119,101.

Heaps of researches are carried out to detect the phished mails, in which the machine learning techniques are most prominent one due the higher precision of detection. The classification algorithms developed in the learning techniques are utilized for anticipating the class of the given mail, but each classification algorithms have their own particular blemishes. In the proposed technique we implemented a multi-classification method which fuses three most accurate classifiers for foreseeing the class label. A prediction model is constructed with the classifiers which incorporate Support Vector Machine (SVM), J48 and Instance Base 1, majority voting algorithm is used to settle on the last choice in regards to the class name of the given email. The dataset holds 5260 publically accessible email corpus for train the prediction model and 500 messages are utilized as the test mails.

2. BACKGROUND

Recently, a variety of anti-phishing techniques are introduced, out of which machine learning technique based approaches are most prominent one. Features are extracted from the html part and body part of the email is used by the machine learning algorithms to predict the possibility of the phishing.

Chandrasekaran et.al [2] proposed phishing email detection based on the structural properties of the phished mails. A total of 25 structural features are extracted and classification of the mails is done using the Support vector Machine algorithm. Data set only contains 200 mails from the public phishing mail corpus, so the results might not be accurate. Sarju et al.[3] utilized the structural properties to detect the spam emails and used Naïve Bayes, Adaboost and Random Forest are used to measure the accuracy. From the above works it identified that the structural properties can be used to discriminate the phished emails from ham mails.

A number of other novel features are also useful to identify phished emails. Bergholz et.al [4] analyzed different properties of the email to detect the phished ones. The content of the emails are evaluated for constructing the feature set and is used for phishing detection. Random Forest and SVM are used to measure the accuracy of the detection.

Abu-Nimeh et al.[5] compared machine learning techniques in phishing detection, they used six different classifiers. A total of 43 features are extracted from the dataset of 2889 phishing and ham emails. They found that Random Forest outperforms

all other classifiers used in their methodology. Miyamoto et al. [6] analyzed nine machine learning techniques for detection of phishing sites; they found that Random Forest and SVM outperforms all the other classifiers.

3. PROPOSED METHOD

In the proposed method structural features, content based features and element features of emails are analyzed and extracted for training the prediction model as shown in the Figure 1, Fi represents the feature set extracted from the mails and C1 and C2 the corresponding class labels.

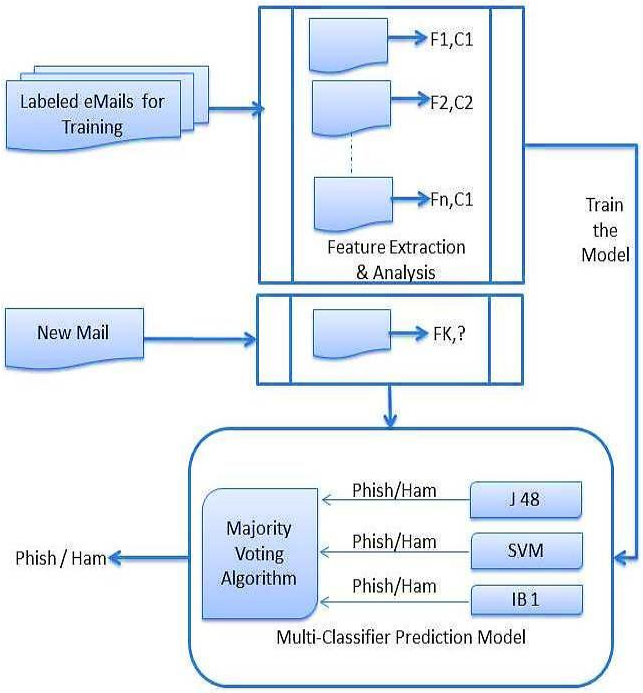


Fig 1 - Multi-Classifier Prediction system for Phishing email detection

3.1 Feature Extraction & Analysis

Feature Extraction and analysis phase, extracts the different features that plays important role in detecting the phished emails. In this work, we extracted the structural features, content based features and element features from the emails.

Emails are available in the Multipart Internet Mail Extension (MIME) format, an HTML parser is used to parse the mail and a HTML is tree constructed. Structural features are extracted from the HTML tree and are multi part count, non ASCII characters, non-text content and content labeling. The multipart feature divides the email into different parts. For example an email with an attachment has two parts, text content and attachment. An email in the MIME format contains a header which shows the character encoding used in that mail. This can be used to identify whether any non ASCII characters used in that mail. Table 1 shows the content label types used in this paper.

Table 1 - Email Content Types

File Extension	MIME Type	Description
.txt	text/plain	Plain text
.htm	text/html	Styled text in HTML format
.jpg	image/jpeg	Picture in JPEG format
.gif	image/gif	Picture in GIF format
.wav	audio/x-wave	Sound in WAVE format
.mp3	audio/mpeg	Music in MP3 format
.mpg	video/mpeg	Video in MPEG format
.zip	application/zip	Compressed file in PK-ZIP format

Element features includes the web technologies used in the email. In this work, we extracted element features of type Boolean which indicates whether HTML, JavaScript, VB Script, XHTML, and CSS used in the mail. Finally the content based features available in the email are extracted. Totally 42 features are extracted from the email corpus and give it as a training set to the prediction model.

3.2 Prediction Model

The extracted features from the feature extraction and analysis stage are used in the multi-classifier prediction model. The prediction model is built using the machine learning algorithms which includes J48, SVM and IB1, each one is capable of classifying the mails into phished or ham mails. The accuracy of the prediction can be improved by combining the classifiers. The final decision regarding the category of the mail is done through the majority voting algorithm. Different research works are done to combining classifiers [7-9]. Majority Voting is the mostly used way for combining classifiers, which count the votes for each class that are predicted by the classifiers and majority class is selected. The new confidence $f_i(x)$ for class i is calculated as

$$f_i(x) = \sum_j I(\max_j \left(p_{ji}(x)\right) = j) \quad (1)$$

in which I() is the identifier function: I(x) =1 if x is true else I(x) will be zero.

When a new mail comes the prediction model identifies its category based in the training test given.

4. EXPERIMENTAL EVALUATION

The dataset holds 5260 publically accessible email corpus of phished and ham mails for train the prediction model and 500 messages are utilized as the test mails. The performance of the prediction model is analyzed using different measures like True Positive, False Positive, Accuracy and Receiver Operator Characteristics curve.

The information about actual and predicted classifications done by machine learning systems is represented in the form of a confusion matrix [10] as shown in the figure 2 and accuracy is measured based on entries.

		Predicted	
		Phish	Ham
Actual	Phish	**TN**	**FN**
	Ham	**FP**	**TP**

Fig 2 - Confusion Matrix

$$\text{Accuracy} = \frac{TP + TN}{TN + FN + FP + TP} \quad (2)$$

If the mail is ham and it is classified as ham, it is counted as a True Positive (TP); if it is classified as phish, it is counted as a False Negative (FN). If the mail is phished and it is classified as phish, it is counted as a True Negative (TN); if it is classified as ham, it is counted as a False Positive (FP).

The Figure 3 compares the FP rate obtained when the classifiers used independently and also combined using majority voting. It is shown that the multi-classification using majority voting outperforms individual classifier performance with an FP rate of 0.8%

.

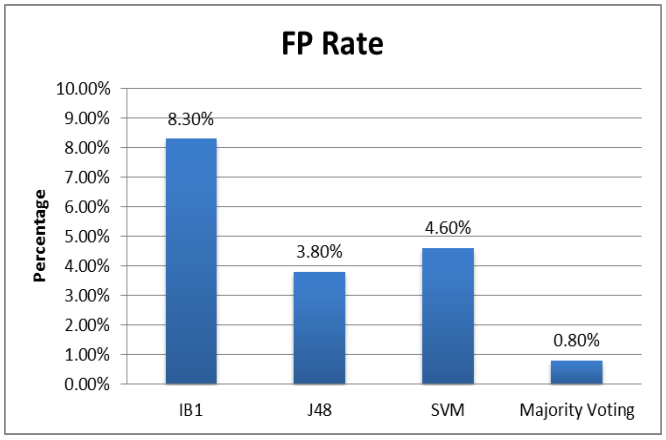


Fig 3 - Phishing Prediction FP Rate

The accuracy of the prediction model is evaluated using the equation 2. From the results, it is clearly understood that the accuracy of the multi-classification is higher than the classifiers applied individually. Multi-classification with majority voting gains an accuracy of 99.80% and is shown in the Figure 4.

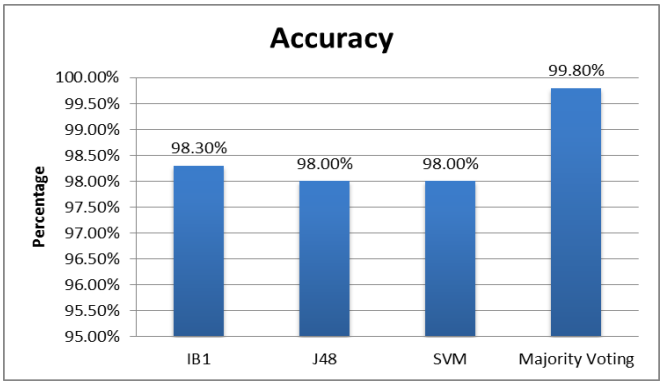
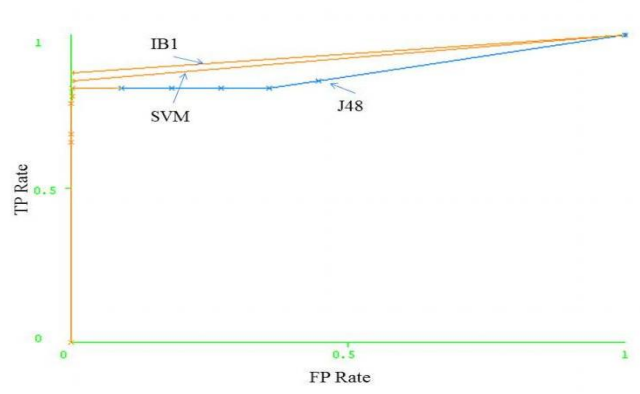
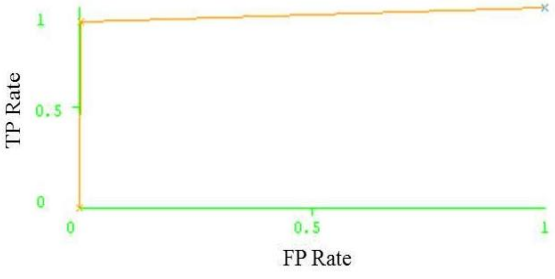


Fig 4 - Phishing Prediction Accuracy Comparison

Another performance measure used to evaluate our proposed system is ROC [11] curve. In an ROC graph X axis plotted with FP rate and TP on Y axis. Figure 5a and 5b shows the ROC measures for the prediction model, by analyzing the graphs it is identified that the multi-classification prediction model has improved results compared to the independent classifier prediction model.



(a)



(b)

Fig 5 – ROC Curves (a) when classifiers used independently ; (b) Multi- classification using Majority Voting

5. CONCLUSIONS

Our argument is that single classifier would not be adequate to urge the clearest image of the phishing email detection accuracy. The experimental results shown that proposed using multi-classifier prediction model outperforms the individual classifier based prediction models in many aspects. It preserves an accuracy of 99.8% with an FP rate of 0.8%. The ROC measure shows that the multi-classification with majority voting gives almost an ideal curve compared to independent classification algorithms.

In the future work, we are planning to incorporate the topic modeling features to the feature set generation stage, because it has the capability to overcome the novel techniques used by the attackers.

REFERENCES

[1] APWG, Anti phishing working group. *http://wwww.antiphishing.org* (2013). Accessed 31 September 2013

[2] Chandrasekaran M, Narayanan M and Upadhyaya S.: *Phishing email detection based on structural properties*. In: Proceedings of 9th Annual NYS Cyber Security Conference, pp. 2-8 (2006).

[3] Sarju S, Riju Thomas and Emilin Shyni C.: Spam *Email Detection using Structural Features*. International Journal of Computer Applications 89(3):pp.38-41 (2014).

[4] A Bergholz, JH Chang, F Reichartz, and S Strobel.: *Improved Phishing Detection using Model-Based Features*. In Proceedings of the Conference on Email and Anti-Spam (CEAS), Mountain View, CA,(2008).

[5] Saeed Abu-Nimeh , Dario Nappa , Xinlei Wang , Suku Nair.: *A comparison of machine learning techniques for phishing detection*, In. Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit, pp.60-69, (2007).

[6] Daisuke Miyamoto , Hiroaki Hazeyama and Youki Kadobayashi.: *An evaluation of machine learning-based methods for detection of phishing sites*, Proceedings of the 15th international conference on Advances in neuro-information processing, (2008).

[7] Geman D and Geman S.: *Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images*. IEEE Trans. on Pattern Analysis and Machine Intelligence, PAMI-6(6), pp.721–741, (1984).

[8] Jain AK, Zhong Y and Lakshmanan S.: *Object Matching Using Deformable Templates*. IEEE Trans. Pattern Analysis and Machine Intelligence 18(3), pp.267–278, (1996)

[9] Kumazawa I.: *Shape extraction by cellular Hough transform*, Technical report of IEICE, PRMU, pp.96–105, (1996)

[10] Provost F, Fawcett T and Kohavi R.: *The case against accuracy estimation for comparing induction algorithms*. In. Proceedings. ICML-98.pp. 445–453, (1998)

[11] T. Fawcett, *An Introduction to ROC Analysis*, Pattern Recognition Letters 27(8), pp. 861-874, (2006).

[12] WEKA. *http://www.cs.waikato.ac.nz/ml/weka/* (2013): Accessed 31 November 2013

[13] SpamAssassin, *http://spamassassin.apache.org* (2013) : Accessed 31 November 2013

[14] Phishingcorpus *http://monkey.org* (2013), Accessed 31 November 2013.

[15] Apache James. (2013) Mime4J Parser. *http://james.apache.org/mime4j*: Accessed 14 October 2013.