جامعة نايف العربية للعلوم الأمنية
Naif Arab University for Security Sciences

Naif Arab University for Security Sciences
Journal of Information Security & Cybercrimes Research
مجلة بحوث أمن المعلومات والجرائم السيبرانية
https://journals.nauss.edu.sa/index.php/JISCR

JISCR

Enhanced Classification Method for Phishing Emails Detection

Hussein Y. AbuMansour [*], Majed A. Alenizi
Forensic Sciences Department, College of Criminal Justice, Naif Arab University for Security Sciences, Riyadh, Saudi Arabia.

Received 01 Nov. 2020; Accepted 01 Dec. 2020; Available Online 20 Dec. 2020

CrossMark

Abstract

Emails are currently the main communication method worldwide as it proven in its efficiency. Phishing emails in the other hand is one of the major threats which results in significant losses, estimated at billions of dollars. Phishing emails is a more dynamic problem, a struggle between the phishers and defenders where the phishers have more flexibility in manipulating the emails features and evading the anti-phishing techniques. Many solutions have been proposed to mitigate the phishing emails impact on the targeted sectors, but none have achieved 100% detection and accuracy. As phishing techniques are evolving, the solutions need to be evolved and generalized in order to mitigate as much as possible. This article presents a new emergent classification model based on hybrid feature selection method that combines two common feature selection methods, Information Gain and Genetic Algorithm that keep only significant and high-quality features in the final classifier. The Proposed hybrid approach achieved 98.9% accuracy rate against phishing emails dataset comprising 8266 instances and results depict enhancement by almost 4%. Furthermore, the presented technique has contributed to reducing the search space by reducing the number of selected features.

I. Introduction

The Internet has become an essential component in many aspects in our life which necessities significant improvements in the development of the infrastructure aiming to facilitate making online procurements through proper dynamic interactions between different parties over the Internet in different architectures [1].

The internet and technology have become important parts of our life today and the number of internet users is increasing every moment and the communication between them that is becoming the extensively used are emails or websites. With the increasing use of the email communications, misused way have arisen that lead to phishing emails which intend to steal sensitive information such as credit cards and usernames which are numerous. Accordingly, phishing is a big challenge facing the communication era with rapidly increasing cost.

With more increasing usage, email traffic comes more increasing attacks of phishing emails, threatening, racial vilification, cyber bullying, terrorist activities, child pornography and sexual harassment, all of which they are common examples of abuses of emails [2]. Phishing attacks employ email messages and Websites that are designed in a professional manner like emails or websites from legitimate organizations. Usually the user is a customer for those organizations to persuade the targeted users into disclosing their personal or payment financial information [3].

Phishing becomes one of the biggest challenges and threats which increases year by year and is considered a

Keywords: *Cybersecurity, Classification, Phishing Emails, Detection, Feature Selection.*

Production and hosting by NAUSS

* Corresponding Author: Hussein Y. AbuMansour
Email: hmansour@nauss.edu.sa

doi: 10.26735/YGMY6142

criminal act that integrates social engineering and technical methods to steal confidential data of consumers such as usernames and passwords, credit cards, malicious files and other attackers' intents [17]. According to Anti Phishing Work Group (APWG), report that number of attacks are in thousands through the world. In online phishing techniques, the attackers will send as many emails as possible to deceive as many as possible [18]. Phishing is faced by two main solutions; detection and prevention either traditionally by using black and white list or in an automated manner that apply machine learning algorithms to effectively detect and prevent these emails.

Data Mining (DM) and Machine Learning (ML) are used in the phishing emails detection domains by applying many techniques to detect these phishing emails or websites, this is done through many steps starting by preprocessing steps, tokenization, stop removal words stemming, etc. Then, the feature extraction and selection deployed for the classification algorithms to start training and testing steps [4]. With the tremendous efforts in the phishing emails detection domain, the phishers continuously devise new techniques for phishing which necessitate the need for more efficient and effective techniques.

This Article aims investigate the feature selection methods and its impact in the classification efficiency and effectiveness. Particularly, the article attempts to answer the following research questions:

- Does dimensionality reduction impact the classification accuracy and performance?
- Does utilizing more than one feature selection method helping in selecting more significant features?

II. Related Work

This section provides an overview on common research efforts in phishing emails detection domain, those based on Data mining and Machine learning approaches.

Authors in [5] presented a simple methodology for phishing emails detection by utilizing Confidence Weighted Liner Classifiers (CWL). Its obtained results motivate to explore feature selection techniques and inclusion. Experimentation results showed competitive accuracy result of 99.77%, with FP rate of less than 1%. LIBLINEAR on the other hand gave the best accuracy of 99.58% with FPR less than 1% and the worst FNR of 2.3%.

Another hybrid-based feature selection approach was proposed in [6], it combines content-based and behavior-based features; it observes the sender behavior for identifying the phishing email using different classifiers (Bayesian network, AdaBoost, Decision Tree and Random Forest). They generated 3 datasets containing about 3000 email instances of phishing and ham emails that extracted the features using hybrid features selection (IG, GR & symmetrical uncertainty). Experimental results showed a promising 93% accuracy.

Authors in [7, 8] focused in their research on the enhancing the mitigation of bulk phishing emails. They examined the modified classification technique that proved to be effective in enhancing the classification accuracy of anti-phishing email filters efficiency. Their previously proposed technique was able to achieve 97% of classification accuracy by lexically analyzing their URL using 40 features. The used dataset comprises 4116 instances.

Later researchers in [9] proposed a new phishing detection model based on Artificial Neural Network. Particularly, they developed Feed Forward Neural Network model, trained by Back Propagation algorithm and was formed to classify websites as either legitimate or phishing. Experimentation results for the proposed model depicts high acceptance ability for noisy data, fault tolerance and high prediction accuracy rate.

In [10] there is another work in the domain that developed a mechanism for better classification of phishing emails using Random Forest Machine Learning algorithm. Their experimentation tested the dataset consisting of 2000 phishing and ham emails instances with a set of 15 prominent features which extracted and used by the machine learning algorithm and achieved the classification accuracy of 99.7% and low FN and FP rates. This algorithm is more efficient in terms of requiring fewer features to detect the phishing emails with more accuracy. On the other hand, because of the rapid change in phishing attack patterns the current phishing detection techniques need to be enhanced.

Authors in [11] proposed a method for detecting suspicious emails by using a Multilayer Neural Network Pruning approach which relies on a feedforward pruning algorithm that extracts only significant features. The Pruning strategy was used for Feature Extraction and select significant ones for identifying phishing emails. Particularly, 18 features have been considered when data set of 4000 emails instances is used. Experimentation results depict a minimized number of selected features is performed. The results in terms of FP and FN are satisfactory with good identification rate with short processing time with accuracy of 99.9%.

Another research in domain is [12] which presented a new system to detect the phishing emails by integrating supervised and unsupervised learning technique. Particularly, they compared the manual and automated feature selection groups of 47 features for the phishing emails structure on dataset of 4800 email instances through WEKA tool. Through this research, a comparative of algorithms (DT, Logistic Regression, CRT and SMO) are conducted. Experimental results showed that the best manually selected features achieved equal accuracy to automated ones of 98.25%. DT, J48 and SMO algorithms achieved the highest accuracy in both features selection methods and integration of multiple classifiers using three top algorithm SMO, DT and J48 by integrating unsupervised techniques with supervised ones before the testing gave more accuracy with 98.37% of all features.

Authors in [13] analyzed the emails structures and based on an improved recurrent convolutional neural networks (RCNN) model with multilevel vectors and attention mechanism. They proposed a new phishing email detection model named, THEMIS, which is used to model emails at the email header, the email body, the character level, and the word level simultaneously. They used an unbalanced dataset of phishing and legitimate emails. Their experimental results show that the overall accuracy of THEMIS reaches 99.848%, false positive rate (FPR) is 0.043%. High accuracy and low FPR ensure that the filter can identify phishing emails with high probability and filter out legitimate emails as little as possible.

As presented in the sample efforts in the field of phishing detection above, none have achieved 100% detection accuracy i.e. relatively high false alarm rates. As phishing techniques are evolving, the solutions need to evolve too as well as generalized in order to mitigate as much as possible. Hence, this article tries to use a hybrid feature selection method and apply it on different classifiers to get notable changes in the classification's effectiveness.

III. Proposed Phishing Detection

This section presents emergent common classification systems including K-Nearest Neighbors (KNN), Naïve Bayes and Support Vector Machine (SVM) and Decision Tree (J48) with a new proposed hybrid features selection method then adapting them for Phishing email detection problem.

A. Data Collection

The benchmark used for training comprises a real sample of existing emails which consists of two kind of email instances including Phish and Ham emails adopted from [4] and [5]. The dataset of 8266 email instances evenly divided into phishing and ham, 4133 instances each. The training dataset is then preprocessed in comma separated values (CSV) as well as Attribute-Relation File Format (ARFF) which are both suitable to be used in our tool. The dataset is already pre-processed according to typical standard. The Extracted features were 47 features as from the dataset as follows:11 features extracted from the email body, 11 features extracted from the email header, another 18 features extracted form URLs and finally 7 features extracted from JavaScript.

B. Preprocessing

A developed JAVA based program in conjunction with importing WEKA software package were used in implementing a new feature selection. The recently proposed hybrid feature selection by us [14] is merging two famous methods, Information Gain (IG) and Genetic Algorithm (GA). We used the IG for the feature selection method first to the most relevant subset features and taken these subset feature. The selected subset of features is then inputted with a customized Genetic Algorithm seeking further improvements to our initial selection, where in this step we ensure that only an informative high quality subset feature is selected and will be used for next step of the classification.

IG in Fig. 1 is used first to select most significant subset features to the class label, the selected subset is then inputted into a modified Genetic Algorithm (Algorithm 1) towards further improvements to initial subset of selected features by ensuring only informative and high quality subset feature remain for the final classification model.

GA is occurs after initial selection via IG method which is the first step that initializes individuals with genes and the gene length represents the features size which is set to a default value equal to 80. Once the initialization process is done, we can apply the Genetic Algorithm and its different process. The process is repeated until a predefined fitness value is met. At every iteration, a new

```
Finding best attributes using IG
D= dataset
Att= attribute
N= set of unique values
M= Regular-intervals
C1,2, ....p= class label, where C1, C2, Cp are same,
        different class label and child node
Da= decision node
Ba= best attribute
If( n ε c1) then
Split m
Else if (n ε c2)
Highest highest probability  c
C=c (highest probability)
Split c
Update split m
M n }
until termination condition is met}
dn= att+ highest normalized information gain
  recurse ba
Cp=ba
Repeat Until ba found
```

Fig. 1 Information Gain Pseudocode.

Algorithm 1: Genetic Algorithm *(n, χ, μ)*

```
// Initialise generation 0:
k := 0;
Pk := a population of n randomly individuals;
// Evaluate Pk:
Compute fitness(i) for each i ∈ Pk;
do
{          // Create generation k + 1:
// 1. Copy:
Select $(1 - \chi) \times n$ members of Pk and insert into Pk+1;
// 2. Crossover:
Select $\chi \times n$ members of Pk; pair them up; produce
        offspring; insert the offspring into Pk+1;
// 3. Mutate:
Select $\mu \times n$ members of Pk+1; invert a randomly-selected
        bit in each;
// Evaluate Pk+1:
Compute fitness(i) for each i ∈ Pk;
// Increment:
k := k + 1; }
while fitness of fittest individual in Pk isn't high enough;
return the fittest individual from Pk";
```

population is generated from the original parent population after evolving the population and kept repeating until an optimum solution is reached. Then, apply the fitness method "Refer to (1)" for obtaining the best individual where the possibility of survival in competition depends on its fitness value of individual. The fitness method was used to evaluate the individual value. Below is the fitness calculation method used which err is KNN err and nf represents cardinality of features extracted from prior step:

$$Fitness = err/nf.length*Math.exp(-1/nf.length) \quad (1)$$

The data set file imported in WEKA is a CSV file format with comma separated between the different attributes' values; WEKA then extracts feature and its datatype. The content of the file represented as: @ Relation < relation Name> which gives the brief description of relation. @attribute <attribute_name> <datatype>, @ Data appeared at the end of list to indicates the value declarations sections in the file. Data types includes numeric, nominal, string, and date. In the current research, majority of feature data type is numerical, where the last features (@attribute class {ham,phish}) is the target class identifier.

C. Classification

We have selected the foremost common classifier for the conducted empirical study, these are (k-Nearest Neighbors (KNN), Naïve Bayes and Support Vector Machine (SVM) and Decision Tree (J48). However, the pre-processing phase was altered in away the features used for training and testing phases are those extracted and selected using the proposed method. Below subsections discussing the steps of the feature extraction and selection in details as per the proposed method but on the case of phishing email detection problem where the number of features we dealt with is 47 features.

D. Experimental results and analysis

In this section, the considered classifiers K-Nearest Neighbors (KNN), Naïve Bayes and Support Vector Machine (SVM) and Decision Tree (J48) have been evaluated in two rounds as mentioned earlier. The bases of comparison are Precision, recall, relative accuracy (as per the equations 2, and 4 respectively) and number of selected features. Both rounds were evaluated against well-known benchmark dataset. Table I depicts the performance of the tested classifiers in terms of the three measurements i.e. precision, recall and accuracy rates in both cases, where IG alone is used in selecting features and with the hybrid feature selection method. It is obvious that the highest performance is achieved when the hybrid feature selection method is deployed.

TABLE I
PTESTED CLASSIFIERS' PERFORMANCE.

Algorithm	Precision		Recall		Accuracy		FP (false alarm)	
	IG	Proposed Method	IG	Proposed Method	IG	Proposed Method	IG	Proposed Method
KNN	0.986	0.993	0.86	0.993	96.55%	98.96%	0.025	0.007
Decision Tree-J48	0.952	0.989	0.952	0.989	95.17%	98.89%	0.053	0.011
NB	0.994	0.99	0.992	0.99	97.24%	98.97%	0.015	0.01
SVM	0.993	0.99	0.993	0.99	99.13%	99.00%	0.004	0.01
Avg	0.98125	0.9905	0.98075	0.9905	0.970675	0.98955	0.02425	0.0095

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

$$Precision = \frac{\mathrm{TP}}{TP+FP} \quad (3)$$

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

Where: TP True positive; FP False positive; TN True negative; FN False negative.

The won-tied-loss records of the considered classifiers with and without using the proposed feature selection method according to the average classification rate on the datasets are: (3,0,1) with deviation between the actual value and other measured ones as +1.395%, +2.65, +0.72%, -1.31 for KNN, Decision tree, NB and SVM respectively.

The average classification accuracy obtained is accuracy obtained after applying classification is 98.96 %, were correctly classified 8180 while 86 instances were incorrectly classified representing 1.04 %. On the other hand, Fig. 2 evaluation metrics when IG selection method is deployed alone against the ratio when the proposed hybrid feature selection method is deployed. Depicted results in Fig. 3 shows that the amount of false alarms has decreased when deploying the proposed selection method and this is affirming our assumption that when the selected features are limited on the foremost informative features , another line of inspection is conducted to ensure that only high quality features will remain in the classifier. This indeed enhances both, the performance and accuracy rates and reducing the amount of false alarms. The results indicate reduction of the false alarm by 0.0148 when the IG is used alone.

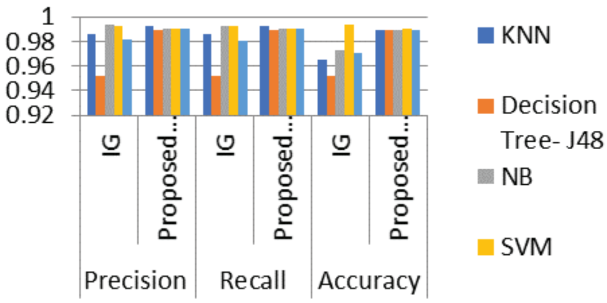

Fig. 2 Evaluation Metrics.

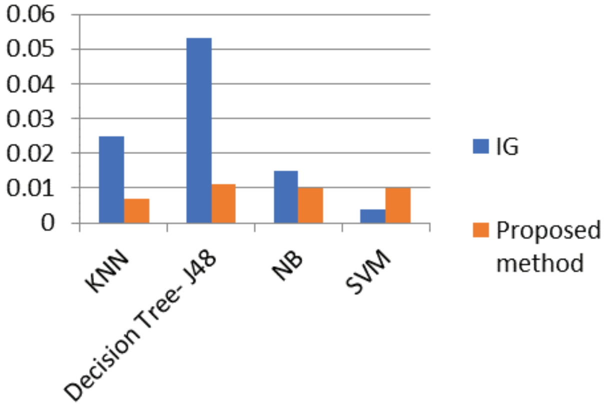

Fig. 3 Error ratio.

Another fact worthy to discuss is the number of attributes in the final classifier considering the 43 features extracted; according to the obtained results with regard to the number of selected features, we can see that when IG alone is used, the number of selected features is 12. On the other hand, only 10 features remain in the final classifiers when adopting the proposed method. No doubt, this reduced the classification time. Further, it has impacted the classification accuracy positively and achieved when using KNN, Decision tree, NB and SVM by 2.41%, 3.72%, 1.73%, -0.31% respectively.

IV. Conclusion

This article has examined the effectiveness of different classification system when applied to the phishing email detection problem. Particularly, we applied a newly proposed feature selection method to those classification systems and their effectives have been evaluated using well know phishing email dataset too. Several well-known classifiers including (KNN, NB, Decision tree and SVM). The bases of the comparison are the classification accuracy, precision, recall and the number of selected features. Experimentation results indicated that classification systems with the proposed feature selection method are becoming highly competitive when utilized in the phishing email problem. The obtained results indicate the superiority of the proposed feature selection method when compared to a single feature selection method, IG in our case. Using a good feature selection method in phishing email detection keeps only the high quality features which accordingly, enhance the accuracy rate as well as reducing the size of classifier which in turn increases the performance and that was proven in the proposed method. As a future piece of work, we'll expand our research effort into having an intelligent On-The-Fly phishing detection model where it works in real time.

References

[1] A. Yasin and A. Abuhasan, “An Intelligent Classification Model For Phishing Email Detection,” in *Int. J. Netw. Secur. Appl.,* vol. 8, no. 4, pp. 55-72, July 2016, doi: 10.5121/ijnsa.2016.8405.

[2] A. A. Akinyelu and A. O. Adewumi, “classification of phishing email using random forest machine learning technique,” in *J. Appl. Math.,* vol. 2014, 2014, Art no. 425731, doi: 10.1155/2014/425731.

[3] “Phishing Corpus”. http://monkey.org/~jose/wiki/doku.php?id=PhishingCorpus. 2006. Accessed July 2019.

[4] J. Mason, “The apache spamassassin public corpus. URL: http://spamassassin. apache. org/publiccorpus”. 2005. http://spamassassin.apache.org/, Accessed June 2019

[5] G. Mujtaba, L. Shuib, R. G. Raj, N. Majeed and M. A. Al-Garadi, "Email Classification Research Trends: Review and Open Issues," in *IEEE Access,* vol. 5, pp. 9044-9064, 2017, doi: 10.1109/ACCESS.2017.2702187.

[6] I. R. Hamid and J. Abawajy, “Hybrid Feature Selection for Phishing Email Detection,” in *Int. Conf. Algorithms Archit. Parallel Process.,* Berlin, 2011, pp. 266-275, doi: 10.1007/978-3-642-24669-2_26.

[7] M. A. Alenizi and H. Y. A. Mansour, "A New Intelligent Hybrid Feature Selection Method," *2018 IEEE/ACS 15th Int. Conf. Comput. Syst. Appl. (AICCSA)*, Aqaba, 2018, pp. 1-6, doi: 10.1109/AICCSA.2018.8612812.

[8] M. Khonji, Y. Iraqi and A. Jones, “Enhancing Phishing E-Mail Classifiers: A Lexical URL Analysis Approach,” in *Int. J. Inf. Secur. Res.(IJISR),* vol. 3, no. 1, Mar. 2013, pp. 236-245.

[9] M. Khonji, Y. Iraqi and A. Jones, "Phishing Detection: A Literature Survey," in *IEEE Commun. Surv. Tutor.,* vol. 15, no. 4, pp. 2091-2121, Fourth Quarter 2013, doi: 10.1109/SURV.2013.032213.00009.

[10] P. S. Bogawar and K. K. Bhoyar, “Email Mining: A Review,” in *Int. J. Comput. Sci.,* vol. 9, no. 1, pp. 429-434, Jan. 2012.

[11] R. B. Basnet and A. H. Sung, “Classifying Phishing Emails Using Confidence-Weighted Linear Classifiers,” *2010 Int. Conf. Inf. Secur. Artif. Intell. (ISAI 2010),* China, Dec. 2010, pp. 109-112.

[12] R. M. Mohammad and H. Y. AbuMansour, “An intelligent model for trustworthiness evaluation in semantic web applications,” *2017 8th Int. Conf. Inf. Commun. Syst. (ICICS),* Irbid, 2017, pp. 362-367, doi: 10.1109/IACS.2017.7921999.

[13] R. Mohammad, T.L. McCluskey and F. A. Thabtah, “Predicting Phishing Websites using Neural Network trained with Black-Propagation,” *Proc. 2013 World Congr. Comput. Sci., Comput. Eng., Appl. Comput., WORLDCOMP,* Las Vegas, NA, USA, 2013, pp. 682-686.

[14] S. A. Al-Saaidah, “Detecting Phishing Emails Using Machine Learning Techniques,” M.S. Thesis, Dept. Comput. Sci., Middle East Univ., Jorden, Jan. 2017.

[15] T. Kathirvalavakumar, K. Kavitha and R. Palaniappan, “Efficient Harmful Email Identification Using Neural Network,” in *Br. J. Math. & Comput. Sci.,* vol. 7, no. 1, pp. 58-67, 2015, doi: 10.9734/BJMCS/2015/15279.

[16] Y. Fang, C. Zhang, C. Huang, L. Liu and Y. Yang, "Phishing Email Detection Using Improved RCNN Model With Multilevel Vectors and Attention Mechanism," in *IEEE Access,* vol. 7, pp. 56329-56340, 2019, doi: 10.1109/ACCESS.2019.2913705.

[17] G. Aaron and R. Manning, “R. APWG phishing activity trends report 2015”. 2015.

[18] A. Almomani, B. B. Gupta, T.-C. Wan, A. Altaher and S. Manickam, “ Phishing Dynamic Evolving Neural Fuzzy Framework for Online Detection “Zero-day” Phishing Email,” in *Indian J. Sci. Technol.*, vol. 6, no. 1, pp. 3960-3964, 2013.