Journal of Theoretical and Applied Information Technology
15th January 2021. Vol.99. No 1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A NEW ASSOCIATION CLASSIFICATION BASED METHOD FOR DETECTING PHISHING WEBSITES

FAISAL ABURUB [1], WAEL HADI [2]

[1]University of Petra, Department of Management Information Systems, Jordan

[2]University of Petra, Department of Information Security, Jordan

E-mail: [1]faburub@uop.edu.jo, [2]whadi@uop.edu.jo

ABSTRACT

Impacting businesses across the world, phishing remains today to be a serious problem: due to anonymous access to personal details, businesses and their consumers deal with the problems materialised out of fishing attacks, huge financial loss being one of these problems. Because of this, phishing needs to be identified and dealt with efficiently using intrusion detection techniques; such mechanisms are yet to be used. It is within this paper that, with the use of a newly-arisen method (Phishing Multi-Class, founded on the grounds of Association Rule) we will study the issue of predicting phishing websites. So as to weigh up the successful use of data mining algorithms using a publicly available dataset involving 10,068 incidents of legitimate and phishing websites, two experimental studies were conducted, in which the classifier model was built. In the first of these two studies, the capability of PMCAR (Phishing Multi-Class Association Rule) compared to three associative classification algorithms (CBA, MCAR, and FACA) was examined; additionally, five benchmark algorithms (SVM, LR, DT, and ANN) were assessed in the second experiment, so as to generalise the competence of utilising data mining for resolving the phishing websites detection issue. As a result of conducting these experiments, all data mining algorithms that were evaluated predict phishing websites with decent classification rate; and so we can conclude that, when looking to tackle the issue of predicting phishing websites, these can be successful methods.

Keywords: *Phishing Websites, Associative Classification, Phishing Multi-Class*

1. INTRODUCTION

Defined as "the art of imitating a truthful website for a company in order to steal critical financial information related to online users such as bank account numbers, credit card numbers, passwords, etc." [1], phishing is a persistent online security issue: in order to fool online users, the phished website frequently has similar content to the truthful website. Phishing costs credit card companies and banks in the USA billions of dollars per year alone, as demonstrated by recent studies (e.g. [2]).

Used as a practical alternative to traditional shopping, e-shopping has increased rapidly amongst internet-users in recent years; e-shopping worldwide sales increased by 20.1% to $1,500 trillion in 2014 [3]; thus, the central focus of e-shops is to attract a large number of customers, which has led to increased competition in product quality amongst e-shops. However, a hasty increase in e-shops has also led to a rapid increase in phishing websites. Many internet users fall victim to these phishing attacks, considering the primary objective of phishers it to convince users that they are using a trustworthy website; because of this, so users' personal data can be sheltered, it is paramount that a capable method of recognising phishing websites is founded.

In order to access personal information, phishing uses both social engineering and technical methods; phishing, which is considered to be one of the most frequent electronic crimes [4], is undergone by publishing a forged online site (e.g. a bank website), and requesting the user to enter their personal information (e.g., username, password, account number, credit card number); this damages the reputations of the targeted financial services, and is damaging to the targeted consumers. An increase of 18% compared to the third quarter of 2014, the amount of phishing records submitted to the organisation in the fourth quarter of 2014 was 197,252 [5].

In order to aid decision-makers in making profitable decisions, data mining is an area of study whose aim is to discover helpful information within vast databases [6,7]; it additionally involves a

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

number of tasks, such as: classification (assigning unseen samples to their predefined categories); association rules (discovering links between attributes, i.e. features, within a vast database); and association classification (AC; the new approach in machine learning and data mining, which intends to assign unseen samples based on association rules). Due to the fact that a number of scholars have indicated that it creates more accurate results than other classical data mining classification techniques, AC proves to be a capable approach [8–12].

The leading objectives within this paper are: to present a fresh AC classifier; and to compare this classifier with five recognised data mining algorithms and three advanced AC algorithms with reference to F1 assessment measures and classification accuracy on a publicly available phishing dataset (put forward by [13].

2. RELATED WORK

This section aims to review the large literature on approaches and methods of detecting phishing; commonly used techniques used for avoiding threat of phishing are categorised around the four categories detailed below:

1. **Type of approach utilised to diminish phishing**: four main approaches had been developed to diminish phishing websites—specifically: list-based approaches (including white and black lists) [14], heuristic approaches [15–17], content-based methods [18–20], and machine learning approaches [10,21], which have collected widespread popularity, due to the fact that they collected prior knowledge in order to envision the manifestations of phishing outbreaks in the future [22].

2. **Type of machine learning approach used**: Some of the following are commonly used: decision trees [23,24]; logistic regression [25–27]; support vector machine [28–30]; Naïve Bayes [31,32]; artificial neural networks [33–35]; and ensemble methods [36–38]. It is recognised that the majority of studies apply an integration of these approaches [35,39–42]. Including confidence-weighted algorithm, a number of studies across the years have fixated on suggesting specialised algorithms for phishing detection [2]: Bayesian adversary SVM classifier (BAAO-SVM) [43]; ball support vector machine method (BSVM) [44]; fuzzy classifier [45]; Transductive Support Vector Machine algorithm (TSVM) [46]; genetic algorithm [47]; passive aggressive algorithm [48]; k-Nearest Neighbour algorithm (k-NN) [49]; and Associative Classification algorithm [10,50].

3. **Types of variables used**: The variables which impact events of phishing attacks are recognised and modeled in a number of studies [4,51,52]. So as to approximate their impact on threat of phishing, some select features—such as Document Object Model (DOM), webpage URL structure, HTML objects, and IP address—are evaluated.

4. **Type of phishing attack**: A number of occurrences of phishing attacks are clear and have the potential to impact organisations greatly; further, it is worth noting that these data mining approaches have been created in the context of phishing variants prevention. In the same vein, the authors of [53] implemented association rule mining method for the purpose of identification of fraudulent phishing derived from instant messaging, predicting threats of phishing for text and audio messages. Similarly, so as to detect spam emails and threat of phishing, Pandey and Ravi (2013) [54] used text and data mining techniques. Further, the authors of [55] used data mining strategies so as to prevent and detect DNS attacks. A further list of researchers who have used corresponding techniques in order to identify phishing websites and emails by predicting attack patterns are: [56–58].

When it comes to predicting the events of phishing threat, these results clarify the essentiality behind machine-learning and data mining models; a number of methods have been shown from these studies (e.g. studying variables of the responsible predictors for phishing, as well as prediction and classification models for the purpose of determining the context of threats). However, it must also be addressed that there is a gap within these studies: considering the fact that the features evaluated for detecting phishing attacks are often minor in nature, this lessens the practicality of these predictive models [59,60]. By evaluating six relevant and basic features which affect phishing attacks, the study at hand attempts to bridge this

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

gap. The evaluated features are as follows: mouseover events; the usage of HTTPS token; the frequency of URL requests; popup windows; subdomain property; and redirection using double slashes. Another improvement that has been made within this study is the contrast of varying data mining classifiers with AC algorithms (e.g. CBA, MCAR and FACA). A new AC model has been created which utilises a number of rules for predicting instances so as to achieve superior capability from these AC algorithms. Considering previous literature fails to consider a multiple rule for detection of phishing detection websites, this step demonstrates the innovation of this approach. Further, the current work develops an intelligent system centered on the top predictor model for detection of phishing websites; notably, this tool can help network administrators, LAN administrators, and business owners to identify phishing websites centered on previously founded information extracted from the AC model. This is an advanced approach when you compare it to studies reliant on individual algorithms for phishing-attack detection using machine-learning strategies [28,34,41].

3. DATA MINING ALGORITHMS

It is necessary to select a suitable algorithm when predicting results for classification issues [2]; this section emphasizes the importance of five conventional learners, who were utilised in a number of studies in phishing attack prediction [2,24,25,29,34].

3.1. Logistic Regression (LR)

Being one of the statistical models utilised for binary response variable classification, logistic regression is very commonly used; this is due to its optimal performance when the link between data is linear [61], and its simplicity. This equation is written in terms of its logit function, as detailed below:

$$\beta^T(x) = \frac{\log P(x;\beta)}{1-P(x;\beta)} \quad (1)$$

In the equation detailed above, the regression parameter in a p_{X1} vector is represented by β, the 'p' predictors as a vector are represented by x ($x = (x_1, x_2 \ldots x_p)$), and the response attribute in binary form is represented by y [61]. However, this algorithm has some downfalls; it involves large-scale assumptions statistics-wise to be defined as compared to other classifiers; similarly, the algorithm additionally is unsuccessful at executing its service when a non-linear relationship is present amongst the data items. Further, it is known to produce unnecessary results for the data sets involving missing values.

Despite the limitations listed above, however, a number of studies have used logistic regression models so as to predict phishing threats [25,62,63]

3.2. Decision Trees

Listed after logistic regression for prediction and classification purposes, decision trees (DT) are some of the most widespread and commonly used machine-learning algorithms; founded by Quinlan [64], it is known to be very user-friendly and comprehensive, as well as being adaptable. C4.5 being a very commonly used tool for creating trustworthy tree structures [65,66], a binary tree is constructed through a 'divide and conquer' strategy so as to envision classification. Additionally, this specific algorithm is frequently used for predicting phishing and, after use in this context, has been known to produce trustworthy results [23,24,45,56].

3.3. Support Vector Machine

Founded on the grounds of enlarging the distance between class variables and hyperplanes [67,68], SVM was invented by Vapnik and, as a data mining and machine learning algorithm, it is used for resolving both linear and non-linear classification problems [69,70]. The middle of the separating margin is named as 'hyperplanes', and the boundaries of separation are named as support vectors. So as to strengthen the function of the algorithm, several kernel functions are defined—including string kernels, radial basis function (RBF) , and polynomial kernel [14]. A number of studies have evaluated SVM as a favourable strategy in predicting phishing attacks [28,29,71] indeed, when compared to other algorithms, SVM has functioned excelled [72].

3.4. Artificial Neural Networks

Founded on the grounds of functioning human brain cells, ANN is a supervised machine learning technique [73] made up of neurons (interconnected nodes), which transfer information across layers (normally three: output layer, one too many hidden layers, and input layer). The weight function is engineered amongst the differing layers so as to fine-tune the development of learning process built on a

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

pre-defined threshold value. Some being deep predictive networks, feedforward neural networks, deep belief networks, backpropagation networks, and convolution networks, there are many versions of neural networks present here [74,75]; a number of studies have emphasized the uses of neural networks when it comes to detecting threats of phishing [33,35,58,76].

3.5. Random Forest

Uniting outputs from a number of decision tree predictors produced at random from independent vector sampling [77,78], random forest is an ensemble learning method and is operated for classification and regression jobs; as well as successfully overseeing a vast amount of data attributes, the algorithm manages missing values [79]. It has been used within this study as a method to test phishing website presence, and has then been likened to other learners.

4. THE PROPOSED MODEL

A new classification method within data mining, AC mining has been analysed lengthily by a number of scholars in real world areas (such as email and websites phishing, medical diagnoses, bioinformatics, text categorisation, and more) spanning over the last decade [10,80]. This approach is very commonly used for two key reasons: how simple the rules are ('If-Then' rules), and the high predictive rate of resulting classifiers. However, despite these positive points, there is a disadvantage linked with the AC mining strategy—namely, the exponential rule-growth; however, this can be solved through suitable pruning in the classifier-construction stage.

So as to predict the label/type of unseen data (a.k.a. test data set), the ultimate objective through the AC mining algorithm is to construct a classification system, derived from a categorised historical data set (a.k.a. a training data set).

Some of which being MAC [81]; CPAR [82]; MCAR [8]; and FACA [10], there have been various commonly used AC algorithms within literature that have been constructed upon real-world domains; indeed, there have been a few attempts to confront the issue of detecting phishing websites through association rule-mining.

Generally speaking, there are two key stages which an AC algorithm must undergo, and recurrent rule-items (attribute values plus class attributes) are identified in the first stage. On the grounds of a threshold commonly known as 'minimum support' (which the end-user contributes), the AC algorithm discovers frequent rule-items (a rule-item which has a regularity within the training data set above the minimum support threshold). The AC algorithm weighs up their produce rule-items and confidence values that hold an acceptable amount of confidence into Class Association Rules (CARs) after all frequent rule-items are found.

After the complete rule-sets have been obtained, the algorithm then goes on to rank them in line with parameters (namely support and confidence); after that, it chooses the most prognostic rules as a classifier (classification system), derived from all the obtained pruning procedures. The final step is called 'the prediction step': the classifier is assessed on an independent data set as a way to gage its rate of prediction; thus, the output of this step is the prediction accuracy/error rate.

Within this paper, the suggested AC algorithm is titled 'Phishing Multi-Class Association Rule' (PMCAR); this algorithm, for one, ensures the end-user obtains a controllable amount of rules, in which they can maintain and understand further. Additionally, this algorithm uses a new class assignment strategy (which is founded on the grounds of group of prediction rules rather than a single rule, thus enhancing resultant accuracy in classifiers), ensuring only good quality rules are employed to forecast test cases—unlike that of the MCAR algorithm.

Below, PMCAR will be deliberated further, including rule pruning, rule discovery, and class assignment test cases.

4.1 Phishing Multi-Class Based on Association Rule

There are three key steps in which a Phishing Multi-Class based on Association Rule (PMCAR) will undergo: training, knowledge-base building, and predicting new cases (as demonstrated below in Figure 1). In the duration of the first phase, set input data is scanned in order to discover frequent items (named frequent one-items) in the form <AttributeValue, class> of size 1. After this step, the algorithm recurrently connects them to create frequent two-items, frequent three-items, and so forth. Notably, any item within the data set that bears an occurrence below the minimum support parameter, is omitted.

The PMCAR algorithm asserts their confidence values once all sizes of frequent items have been

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

noted, in these cases which have a bigger confidence score than the minimum confidence parameter will be a class association rule (CAR). In cases other than these, the item gets deleted; thus, the only items represented in the complete set of CARs are statistically representative and demonstrate high reliability. Now, selecting a part from the whole set of CARs to create a knowledge-base is the next step.

In the following subsections, further information regarding the PMCAR, involving prediction of test cases and knowledge-base construction, is supplied.

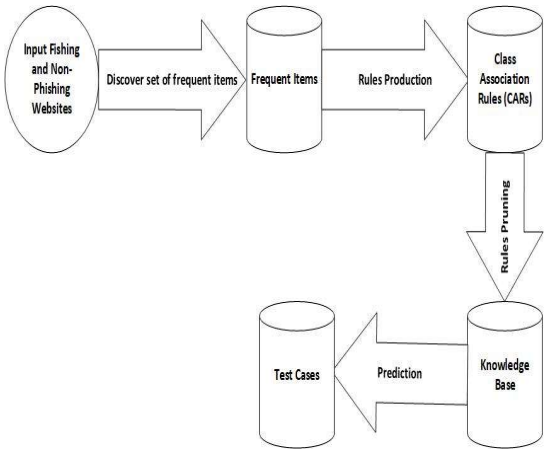

Figure 1. Phishing Multi-Class based on Association Rule (PMCAR) Model

4.1.1. Building the Knowledge Base

So as to construct the knowledge-base, before prune redundant guidelines, it is essential that the rules are fixed in order; this is to provide higher-quality rules higher priority in being selected as subset of the knowledge-base. Thus, the algorithm at hand fixes the rules in mind of the following directions:

1. The top rank is given to the rule with higher confidence score;
2. in the case of there being confidence scores consisting of two or more matching rules, then the rule with maximum support score gets a top rank;
3. in the case of there being confidence and support scores of two or more matching rules, the rule with a lesser amount of features in the body will be a top rank;
4. if all the previous standards are alike for two or more rules, the rule created first will be a top rank.

As follows, the construction of the knowledge base step in PMCAR is detailed:

When it comes to each sequentially-sorted/fixed rule (CAR), the PMCAR begins with the former and begins its application on the training dataset, whereby the knowledge-base receives the rule, assuming it includes one or more cases, irrespective of the training case rule class similarity. To word it differently, the top confidence rules are verified through the training instances; then, training instances that are comparable to the rule antecedent are signified for deletion, and the rule is stored in the knowledge base; additionally, when this is undergone, all training cases are deleted that are related to the knowledge base at hand. So that either no cases remain within the training data or so that all the rules have been verified, this procedure is recurring on the rest rules orderly; thus, this algorithm process promises that just confidence and high-quality rules are left for prediction.

The hope is that this creates added accuracy in prediction in training data sets—and not necessarily on novel unseen test cases.

4.1.2. Prediction Method

When it comes to the prediction of test data case, the strategy used is that of PMCAR: it splits all the rules that satisfy the test instance into collections—one for each class label—before counting the amount of rules per collection (or group), after which it labels the test instance of the group with the biggest count. If, however, we are looking at an instance whereby two or more groups with an alike count are present, the labelling of the class to the test instance is randomised. Whilst other modern AC algorithms (e.g. MCAR) use the top confidence rule only for forecasting test instances, the proposed method, on the other hand, creates the prediction on the grounds of a number of rules—evaluated by prior approaches based on prediction models (i.e. [83,84]. This acts as a benefit to our algorithm, as many support and high confidence rules added to the overall decision of assignment.

Lastly, the majority class in the training dataset (a.k.a. the default class) will be allocated to that case in cases where zero rules in the classifier are related to the test case.

The main difference between PMCAR and algorithms from prior work, it uses multiple rules in prediction stage.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

5. EXPERIMENTAL RESULTS

5.1 Description of Phishing Data Set

Performing a paramount role in the construction of predictive models of phishing classifiers, the attendance of high-quality data sets (with phishing classifiers) used within this study are obtained from UCI Machine Learning Repository (which stow data sets for public utilisation) [85]. A .csv file construction containing 30 web-based parameters linked with the event of phishing websites for 11,056 examples of data, this phishing data set comprises two non-phishing—where 4,899 instances were recorded—and phishing—where 6,157 instances were recorded—target classes. The class variable will be further adapted in the case of there being any inconsistencies found within the target class.

5.2. Data Pre-processing

So as to eradicate redundant tuples and outliers completely from the data, data processing is a paramount stage; as a matter of fact, it has been perceived that, within the phishing website data set—frequently referred to as 'imbalanced data set in classification problems'—, the target variable is not consistently dispersed amongst the negative and positive classes. Considering they result in biased predictions and misclassifications, it is harmful to use machine-learning algorithms on such data sets; thus, within literature, there are a number of strategies for dealing with those imbalanced data sets, cluster-based oversampling being the first we will discuss. This strategy, on the negative and positive events if the target variable, uses the k-means clustering algorithm so as to classify the data set clusters; so as to stem data groups that possess equal distribution of negative and positive classes, each cluster gained is, again, oversampled. This application of k-means clustering on the grounds of oversampling made identification of 988 data instances as inappropriate entities, resulting in 10,068 data instances with an equal proportion of phishing and non-phishing websites.

5.3. Evaluation on the Performance of the Proposed Algorithm

Two sets of experiment are to be used within this section as a way to investigate the applicability on detecting phishing websites using data mining classification algorithms. Within the first experiment, the algorithms FACA [10], CBA [86], and MCAR [8] were used to explore the PMCAR algorithm capability when it comes to predicting phishing websites. The above algorithms were chosen for study because they share comparable methodologies, which ensures our investigation is fair.

To begin with when weighing up the performance and reliability of PMCAR, we take the rules that have minimum confidence = 0.50 and that have minimum support = 0.05 and apply the F1 measures (which is a harmonic mean, or weighted average, of recall and precision, computed as in the equation below) and the standard classification accuracy (which is calculated by dividing the amount of correctly predicted groundwater locations by the total amount of groundwater locations within the testing data):

$$F1 = \frac{2 * precision * recall}{recall + precision} \quad (2)$$

In the above equation, the ratio of correct predictions divided by the total amount of the system's predictions is represented by 'precision', and the ratio of correct predictions divided by the total amount of predictions is represented by 'recall'.

Below, through an instance where a classification has been discovered to forecast phishing websites, let us demonstrate the performance measures; this particular sample involves 18 phishing websites—10 of which are labelled 'Yes', 8 'No'.

For the 8 phishing websites assigned 'No', 6 were predicted as 'No' and 2 'Yes' by the system; in the same vein, out of the 'Yes' websites, 7 were predicted 'Yes' and 3 'No'. The average scores for accuracy in estimation here was 0.737 for 'Yes' and 0.706 for 'No'.

Within the second experiment that was conducted, five commonly known benchmark classification algorithms (Random Forest, LR, ANN, DR and SVM) were explored within the corresponding data set so as to generalise the applicability of phishing website prediction in data mining algorithms.

Utilised to apply the algorithms evaluated in our experiments, the Waikato Environment for Knowledge Analysis (WEKA) tool [87] is commonly known as a data mining and machine-learning landmark system that has obtained large-scale acceptance when it comes to business circles and academia, and has also become a commonly used tool for data mining research [87].

In this case, in order to weigh up the considered algorithms within the experiments assessed, we

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

utilise a 10-fold cross-validation; notably, experiments are undergone on a 17 machine with 16GB main memory and a 3 GHz processor in a Windows 8 setting.

5.3.1. AC Algorithms Performance

Demonstrated in Table 1 is the classification accuracy (%) of the algorithms discussed; clearly, the PMCAR algorithm is more capable than that of the CBA, FACA and MCAR algorithms—whereby the CBA algorithm proves to be the least desirable in terms of predicting phishing websites, which is clear when we consider that PMCAR outperformed CBA by 5.9%; it additionally outperformed FACA and MCAR by around 2.7% and 4.6%.

Table 1 The classification accuracy (%) and F1 score of AC algorithms

AC Algorithms	Classification Accuracy	F1 score
PMCAR	83.8	0.843
MCAR	79.2	0.735
CBA	77.9	0.709
FACA	81.1	0.822

Also demonstrated in Table 1 are the F1 measures of the PMCAR, FACA, CMAR and CBA algorithms. Additionally, some noteworthy numbers to have been found here are PMCAR outperformed the algorithms by 13.4% for CBA, FACA and MCAR, respectively; 13.4% for CBA, 2.1% for FACA and 10.8% for MCAR. There is one paramount explanation behind why the PMCAR algorithm outperformed the above algorithms: it utilises a number of rules for predicting phishing websites.

A drawback for algorithms containing just one single rule is as follows: the highest confidence rule is sometimes unsuccessful, especially for those datasets that include an imbalanced distribution of phishing datasets (e.g. phishing datasets) [84]; thus, it seems that it is more successful to control a minimal amount of rules when it comes to prediction of phishing websites.

5.3.2. Benchmark Data Mining Algorithms Performance

We have evaluated five commonly used data mining classifiers (ANN, LR, DT, SVM, and Random Forest) so as to generalise the applicability of utilising data mining on identifying phishing websites; to do this, we utilised F1 measures and classification accuracy.

Table 2 The results of popular algorithms

Rule-based Algorithms	Classification Accuracy	F1
SVM	84.4	81.4
LR	83.3	86.5
DT	83.2	86.6
ANN	82.3	91.0
Random Forest	83.7	86.1

Regarding classification accuracy, after Table 2 has been successfully analysed, it was discovered that all the popular data mining algorithms were outperformed by the SVM classifier, the ANN algorithm proving to be the least capable; specifically, the SVM achieved the following amounts of higher classification accuracy over the algorithms: 2.1% over ANN, 1.1% over LR, 0.7% over Random Forest, and 1.2% over DT. Notably, the LR and DT algorithms both produced comparable results. When we evaluate the F1 measure, Random Forest obtained the best and ANN obtained the worst results, the latter proving to be 1.4% less reliable than the former; additionally, Random Forest obtained 6.1% over SVM, 1% over LR, and 0.90% over DT respectively.

The algorithm suggested within this study obtained lower classification accuracy over SVM by 0.6%; however, it proved more capable by 2.9% than SVM in terms of F1 measurement.

These same results additionally suggested that commonly used algorithms (e.g. ANN and SVM) achieve high classification accuracy; despite this, it is additionally worthy to note that they create complex and black box models, which are tough for the decision-maker to interpret and comprehend.

The rule-based model is necessary for decision-makers because of the following factors: 1) the classification process is similar to black box classifiers (e.g. ANN and SVM) and are well-organised; and 2) the decision-maker can comprehend, control, and read the produced rules with ease.

It seems reasonable to say that we can conclude, from all the above experiments, that all of the algorithms that have been evaluated create good enough F1 rates and classification accuracy, which

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

mirrors the features' importance for the phishing websites dataset. Further, it is safe to say that all algorithms have the potential to aid the problem of detecting phishing websites.

6. CONCLUSIONS

Phishing can be defined as: 'an attempt—frequently through fraudulent emails and/or websites—to thieve an individual's vulnerable information'. These, as we have established, are an especially relevant problem in this day and age, and successfully stop users from safely undergoing tasks via the internet.

Within this paper, the primary objective has been to construct a new, innovative AC algorithm named PMCAR, and to explore the effectiveness of this algorithm against three other well-known AC algorithms (FACA, CBA, and MCAR), as well as five other well-known data mining algorithms (Random Forest, LR, ANN, DT, and SVM), whereby F1 evaluation measures and classification accuracy towards the phishing dataset were taken into account. The results of these investigations suggest that, when it comes to the prediction of difficult problems concerning phishing websites, there is the potential for the application of computerised data mining strategies.

At some point in the near future, we would like to demonstrate the following as future works:

1. Data sets from live-time situations of recent origin concentrating on diverse parameters resulting in threats of phishing must be studied so as to consider the capability of the PMCAR model;
2. it is essential that a number of types of threats of phishing are identified, before concentrating on one type of security breach to comprehend the characteristics of phishing attacks;
3. as well as neural networks, deep learning strategies can be utilised as a way to trial the progression and capability in the predictive performance of the PMCAR model.

ACKNOWLEDGEMENT

This work was supported by University of Petra – Dean of Scientific Research [2018/2/2].

REFERENCES

[1] M. Ajlouni, W. Hadi, J. Alwidian, Detecting Phishing Websites Using Associative Classification, Eur. J. Bus. Manag. 5 (2013) 36–40.

[2] K. Nagaraj, B. Bhattacharjee, A. Sridhar, S. GS, Detection of phishing websites using a novel twofold ensemble model, J. Syst. Inf. Technol. 20 (2018) 321–357. https://doi.org/10.1108/JSIT-09-2017-0074.

[3] J. Ingham, J. Cadieux, A. Mekki Berrada, E-Shopping acceptance: A qualitative and meta-analytic review, Inf. Manag. 52 (2015) 44–60. https://doi.org/10.1016/j.im.2014.10.002.

[4] M. Khonji, Y. Iraqi, Lexical URL analysis for discriminating phishing and legitimate e-mail messages, Internet Technol. Secur. (2011) 11–14.

[5] G. Sonowal, K.S. Kuppusamy, PhiDMA – A phishing detection model with multi-filter approach, J. King Saud Univ. - Comput. Inf. Sci. 32 (2020) 99–112. https://doi.org/10.1016/j.jksuci.2017.07.005.

[6] N. El-Khalili, M. Alnashashibi, W. Hadi, A.A. Banna, G. Issa, Data Engineering for Affective Understanding Systems, Data. 4 (2019) 52. https://doi.org/10.3390/data4020052.

[7] W. Hadi, N. El-Khalili, M. AlNashashibi, G. Issa, A.A. AlBanna, Application of data mining algorithms for improving stress prediction of automobile drivers: A case study in Jordan, Comput. Biol. Med. 114 (2019) 103474. https://doi.org/10.1016/j.compbiomed.2019.103474.

[8] F. Thabtah, P. Cowling, Yonghong Peng, MCAR: multi-class classification based on association rule, in: 3rd ACS/IEEE Int. Conf. OnComputer Syst. Appl. 2005., IEEE, 2005: pp. 130–136. https://doi.org/10.1109/AICCSA.2005.1387030.

[9] N. Abdelhamid, A. Ayesh, W. Hadi, Multi-Label Rules Algorithm Based Associative Classification, Parallel Process. Lett. 24 (2014) 1450001. https://doi.org/10.1142/S0129626414500017.

[10] W. Hadi, F. Aburub, S. Alhawari, A New Fast Associative Classification Algorithm for Detecting Phishing Websites, Appl. Soft

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Comput. 48 (2016) 729–734. https://doi.org/10.1016/j.asoc.2016.08.005.

[11] W. Hadi, G. Issa, A. Ishtaiwi, ACPRISM: Associative classification based on PRISM algorithm, Inf. Sci. (Ny). 417 (2017) 287–300. https://doi.org/10.1016/j.ins.2017.07.025.

[12] W. Hadi, Q.A. Al-Radaideh, S. Alhawari, Integrating associative rule-based classification with Naïve Bayes for text classification, Appl. Soft Comput. 69 (2018) 344–356. https://doi.org/10.1016/j.asoc.2018.04.056.

[13] R. Mohammad, F. Thabtah, L. McCluskey, Phishing Websites Dataset, (2015). https://archive.ics.uci.edu/ml/datasets/Phishing+Websites (accessed November 1, 2015).

[14] Y. Cao, W. Han, Y. Le, Anti-phishing based on automated individual white-list, in: Proc. 4th ACM Work. Digit. Identity Manag. - DIM '08, ACM Press, New York, New York, USA, 2008: p. 51. https://doi.org/10.1145/1456424.1456434.

[15] L.A.T. Nguyen, B.L. To, H.K. Nguyen, M.H. Nguyen, Detecting phishing web sites: A heuristic URL-based approach, in: 2013 Int. Conf. Adv. Technol. Commun. (ATC 2013), IEEE, 2013: pp. 597–602. https://doi.org/10.1109/ATC.2013.6698185.

[16] R.S. Rao, S.T. Ali, PhishShield: A Desktop Application to Detect Phishing Webpages through Heuristic Approach, Procedia Comput. Sci. 54 (2015) 147–156. https://doi.org/10.1016/j.procs.2015.06.017.

[17] N. Khan, M. Kaur, R. Panchal, P.K. Rai, N. Rathod, Heuristic Based Approach for Fraud Detection using Machine Learning, Int. J. Innov. Res. Comput. Commun. Eng. 6 (2018) 908–911. https://doi.org/10.15680/IJIRCCE.2018.0602038.

[18] C. Ardi, J. Heidemann, Poster: Lightweight Content-based Phishing Detection, 2015. http://www.isi.edu/~johnh/PAPERS/Ardi15a.html.

[19] C. Ardi, J. Heidemann, AuntieTuna: Personalized Content-based Phishing Detection, in: Proc. 2016 Work. Usable Secur., Internet Society, Reston, VA, 2016. https://doi.org/10.14722/usec.2016.23012.

[20] A.K. Jain, B.B. Gupta, Phishing Detection: Analysis of Visual Similarity Based Approaches, Secur. Commun. Networks. 2017 (2017) 1–20. https://doi.org/10.1155/2017/5421046.

[21] N. Abdelhamid, A. Ayesh, F. Thabtah, Phishing detection based Associative Classification data mining, Expert Syst. Appl. 41 (2014) 5948–5959. https://doi.org/10.1016/j.eswa.2014.03.019.

[22] S. Wang, S. Khan, C. Xu, S. Nazir, A. Hafeez, Deep Learning-Based Efficient Model Development for Phishing Detection Using Random Forest and BLSTM Classifiers, Complexity. 2020 (2020) 1–7. https://doi.org/10.1155/2020/8694796.

[23] X. YANG, L. YAN, B. YANG, Y. LI, Phishing Website Detection Using C4.5 Decision Tree, DEStech Trans. Comput. Sci. Eng. (2017). https://doi.org/10.12783/dtcse/itme2017/7975.

[24] A.K. Shrivas, R. Suryawanshi, Decision Tree Classifier for Classification of Phishing Website with Info Gain Feature Selection, Int. J. Res. Appl. Sci. Eng. Technol. . 5 (2017) 780–783. www.ijraset.com.

[25] Y. Han, M. Yang, H. Qi, X. He, S. Li, The Improved Logistic Regression Models for Spam Filtering, in: 2009 Int. Conf. Asian Lang. Process., IEEE, 2009: pp. 314–317. https://doi.org/10.1109/IALP.2009.74.

[26] J. Ma, L.K. Saul, S. Savage, G.M. Voelker, Identifying suspicious URLs, in: Proc. 26th Annu. Int. Conf. Mach. Learn. - ICML '09, ACM Press, New York, New York, USA, 2009: pp. 1–8. https://doi.org/10.1145/1553374.1553462.

[27] K. Thomas, C. Grier, J. Ma, V. Paxson, D. Song, Design and Evaluation of a Real-Time URL Spam Filtering Service, in: 2011 IEEE Symp. Secur. Priv., IEEE, 2011: pp. 447–462. https://doi.org/10.1109/SP.2011.25.

[28] H. Huang, L. Qian, Y. Wang, A SVM-based Technique to Detect Phishing URLs, Inf. Technol. J. 11 (2012) 921–925. https://doi.org/10.3923/itj.2012.921.925.

[29] M. Moghimi, A.Y. Varjani, New rule-based phishing detection method, Expert Syst. Appl. 53 (2016) 231–242. https://doi.org/10.1016/j.eswa.2016.01.028.

[30] A. Abdulwakil, M.A. Aydin, D. Aksu, Detecting phishing websites using support vector machine algorithm, Pressacademia. 5 (2017) 139–142. https://doi.org/10.17261/Pressacademia.2017.582.

[31] R. Priya, An Ideal Approach for Detection of Phishing Attacks using Naïve Bayes

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Classifier, Int. J. Comput. Trends Technol. 40 (2016) 2016. http://www.ijcttjournal.org.

[32] N. Kumar, P. Chaudhary, Mobile Phishing Detection using Naive Bayesian Algorithm, Int. J. Comput. Sci. Netw. Secur. 17 (2017) 142. http://paper.ijcsns.org/07_book/201707/20170720.pdf.

[33] R.M. Mohammad, F. Thabtah, L. McCluskey, Predicting phishing websites based on self-structuring neural network, Neural Comput. Appl. 25 (2014) 443–458. https://doi.org/10.1007/s00521-013-1490-z.

[34] L.A.T. Nguyen, B.L. To, H.K. Nguyen, M.H. Nguyen, An efficient approach for phishing detection using single-layer neural network, in: 2014 Int. Conf. Adv. Technol. Commun. (ATC 2014), IEEE, 2014: pp. 435–440. https://doi.org/10.1109/ATC.2014.7043427.

[35] E.-S.M. El-Alfy, Detection of Phishing Websites Based on Probabilistic Neural Networks and K-Medoids Clustering, Comput. J. 60 (2017) 1745–1759. https://doi.org/10.1093/comjnl/bxx035.

[36] F. Toolan, J. Carthy, Phishing detection using classifier ensembles, in: 2009 ECrime Res. Summit, IEEE, 2009: pp. 1–9. https://doi.org/10.1109/ECRIME.2009.5342607.

[37] J.K. Virdi, A.K. Dewangan, An Ensemble Model for Identification of Phishing Website, Int. J. Res. Appl. Sci. Eng. Technol. 5 (2017) 1151–1153.

[38] H.S. Hota, A.K. Shrivas, R. Hota, An Ensemble Model for Detecting Phishing Attack with Proposed Remove-Replace Feature Selection Technique, Procedia Comput. Sci. 132 (2018) 900–907. https://doi.org/10.1016/j.procs.2018.05.103.

[39] S. Abu-Nimeh, D. Nappa, X. Wang, S. Nair, A comparison of machine learning techniques for phishing detection, in: Proc. Anti-Phishing Work. Groups 2nd Annu. ECrime Res. Summit - ECrime '07, ACM Press, New York, New York, USA, 2007: pp. 60–69. https://doi.org/10.1145/1299015.1299021.

[40] M. Aburrous, M.A. Hossain, K. Dahal, F. Thabtah, Predicting Phishing Websites Using Classification Mining Techniques with Experimental Case Studies, in: 2010 Seventh Int. Conf. Inf. Technol. New Gener., IEEE, 2010: pp. 176–181. https://doi.org/10.1109/ITNG.2010.117.

[41] N. Abdelhamid, F. Thabtah, H. Abdel-jaber, Phishing detection: A recent intelligent machine learning comparison based on models content and features, in: 2017 IEEE Int. Conf. Intell. Secur. Informatics, IEEE, 2017: pp. 72–77. https://doi.org/10.1109/ISI.2017.8004877.

[42] I. Qabajeh, F. Thabtah, F. Chiclana, A recent review of conventional vs. automated cybersecurity anti-phishing techniques, Comput. Sci. Rev. 29 (2018) 44–55. https://doi.org/10.1016/j.cosrev.2018.05.003.

[43] G. L'Huillier, R. Weber, N. Figueroa, Online phishing classification using adversarial data mining and signaling games, ACM SIGKDD Explor. Newsl. 11 (2010) 92. https://doi.org/10.1145/1809400.1809421.

[44] Y. Li, L. Yang, J. Ding, A minimum enclosing ball-based support vector machine approach for detection of phishing websites, Optik (Stuttg). 127 (2016) 345–351. https://doi.org/10.1016/j.ijleo.2015.10.078.

[45] M. Aburrous, M.A. Hossain, K. Dahal, F. Thabtah, Intelligent phishing detection system for e-banking using fuzzy data mining, Expert Syst. Appl. 37 (2010) 7913–7921. https://doi.org/10.1016/j.eswa.2010.04.044.

[46] Y. Li, R. Xiao, J. Feng, L. Zhao, A semi-supervised learning approach for detection of phishing webpages, Optik (Stuttg). 124 (2013) 6027–6033. https://doi.org/10.1016/j.ijleo.2013.04.078.

[47] V. Shreeram, M. Suban, P. Shanthi, K. Manjula, Anti-phishing detection of phishing attacks using genetic algorithm, in: 2010 Int. Conf. Commun. Control Comput. Technol., IEEE, 2010: pp. 447–450. https://doi.org/10.1109/ICCCCT.2010.5670593.

[48] W. Zhang, Y.-X. Ding, Y. Tang, B. Zhao, Malicious web page detection based on on-line learning algorithm, in: 2011 Int. Conf. Mach. Learn. Cybern., IEEE, 2011: pp. 1914–1919. https://doi.org/10.1109/ICMLC.2011.6016954.

[49] A. Altaher, Phishing Websites Classification using Hybrid SVM and KNN Approach, Int. J. Adv. Comput. Sci. Appl. 8 (2017). https://doi.org/10.14569/IJACSA.2017.080611.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[50] N. Abdelhamid, A. Ayesh, F. Thabtah, Phishing detection based Associative Classification data mining, Expert Syst. Appl. 41 (2014) 5948–5959. https://doi.org/10.1016/j.eswa.2014.03.019.

[51] W. Zhuang, Q. Jiang, T. Xiong, An Intelligent Anti-phishing Strategy Model for Phishing Website Detection, in: 2012 32nd Int. Conf. Distrib. Comput. Syst. Work., IEEE, 2012: pp. 51–56. https://doi.org/10.1109/ICDCSW.2012.66.

[52] M. Dadkhah, M. Jazi, V. Lyashenko, Prediction of phishing websites using classification algorithms based on weight of web pages characteristics, (2014). http://openarchive.nure.ua/handle/123456789/1934 (accessed March 31, 2016).

[53] M.M. Ali, L. Rajamani, Deceptive phishing detection system: From audio and text messages in Instant Messengers using Data Mining approach, in: Int. Conf. Pattern Recognition, Informatics Med. Eng., IEEE, 2012: pp. 458–465. https://doi.org/10.1109/ICPRIME.2012.6208390.

[54] M. Pandey, V. Ravi, Text and Data Mining to Detect Phishing Websites and Spam Emails, in: 2013: pp. 559–573. https://doi.org/10.1007/978-3-319-03756-1_50.

[55] H. Cui, J. Yang, Y. Liu, Z. Zheng, K. Wu, Data Mining-based DNS Log Analysis, Ann. Data Sci. 1 (2014) 311–323. https://doi.org/10.1007/s40745-014-0023-7.

[56] K.F. MBAH, A Phishing E-Mail Detection Approach Using Machine Learning, THE UNIVERSITY OF NEW BRUNSWICK, 2017.

[57] T. Peng, I. Harris, Y. Sawa, Detecting Phishing Attacks Using Natural Language Processing and Machine Learning, in: 2018 IEEE 12th Int. Conf. Semant. Comput., IEEE, 2018: pp. 300–301. https://doi.org/10.1109/ICSC.2018.00056.

[58] S. Smadi, N. Aslam, L. Zhang, Detection of online phishing email using dynamic evolving neural network based on reinforcement learning, Decis. Support Syst. 107 (2018) 88–102. https://doi.org/10.1016/j.dss.2018.01.001.

[59] A.C. Lorena, L.F.O. Jacintho, M.F. Siqueira, R. De Giovanni, L.G. Lohmann, A.C.P.L.F. de Carvalho, M. Yamamoto, Comparing machine learning classifiers in potential distribution modelling, Expert Syst. Appl. 38 (2011) 5268–5275. https://doi.org/10.1016/j.eswa.2010.10.031.

[60] Y. Bengio, A. Courville, P. Vincent, Representation Learning: A Review and New Perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2013) 1798–1828. https://doi.org/10.1109/TPAMI.2013.50.

[61] J. Feng, H. Xu, S. Mannor, S. Yan, Robust Logistic Regression and Classification, in: Proc. 27th Int. Conf. Neural Inf. Process. Syst. - Vol. 1, MIT Press, Cambridge, MA, USA, 2014: pp. 253–261. http://dl.acm.org/citation.cfm?id=2968826.2968855.

[62] M.N. Feroz, S. Mengel, Examination of data, rule generation and detection of phishing URLs using online logistic regression, in: 2014 IEEE Int. Conf. Big Data (Big Data), IEEE, 2014: pp. 241–250. https://doi.org/10.1109/BigData.2014.7004239.

[63] G. Sonowal, K.S. Kuppusamy, PhiDMA – A phishing detection model with multi-filter approach, J. King Saud Univ. - Comput. Inf. Sci. (2017). https://doi.org/10.1016/j.jksuci.2017.07.005.

[64] J.R. Quinlan, Induction of Decision Trees, Mach. Learn. 1 (1986) 81–106. https://doi.org/10.1023/A:1022643204877.

[65] B. HSSINA, A. MERBOUHA, H. EZZIKOURI, M. ERRITALI, A comparative study of decision tree ID3 and C4.5, Int. J. Adv. Comput. Sci. Appl. 4 (2014). https://doi.org/10.14569/SpecialIssue.2014.040203.

[66] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[67] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297. https://doi.org/10.1007/BF00994018.

[68] F. Aburub, W. Hadi, Predicting Groundwater Areas Using Data Mining Techniques : Groundwater in Jordan as Case Study, Int. J. Comput. Electr. Autom. Control Inf. Eng. 10 (2016) 1475–1478.

[69] R. Wang, Shu-Li, Hsu, Y.H. Lin, M.-L. Tseng, Evaluation of customer perceptions on airline service quality in uncertainty, Procedia - Soc. Behav. Sci. 25 (2011) 419–437. https://doi.org/10.1016/j.sbspro.2012.02.054.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[70] M.J. Zaki, C.-J. Hsiao, CHARM : An Efficient Algorithm for Closed Itemset Mining, in: Data Min. Knowl. Discov., 2001: pp. 457–473. https://doi.org/10.1.1.111.520.

[71] M. Moghimi, A.Y. Varjani, New rule-based phishing detection method, Expert Syst. Appl. (2016). https://doi.org/10.1016/j.eswa.2016.01.028.

[72] D.K. Tayal, A. Jain, S. Arora, S. Agarwal, T. Gupta, N. Tyagi, Crime detection and criminal identification in India using data mining techniques, AI Soc. 30 (2015) 117–127. https://doi.org/10.1007/s00146-014-0539-6.

[73] W.S. McCulloch, W. Pitts, A logical calculus of the ideas immanent in nervous activity, Bull. Math. Biophys. 5 (1943) 115–133. https://doi.org/10.1007/BF02478259.

[74] C. Shu, D.H. Burn, Artificial neural network ensembles and their application in pooled flood frequency analysis, Water Resour. Res. 40 (2004). https://doi.org/10.1029/2003WR002816.

[75] J.E. Villaverde, D. Godoy, A. Amandi, Learning styles' recognition in e-learning environments with feed-forward neural networks, J. Comput. Assist. Learn. 22 (2006) 197–206. https://doi.org/10.1111/j.1365-2729.2006.00169.x.

[76] E. Zhu, Y. Ju, Z. Chen, F. Liu, X. Fang, DTOF-ANN: An Artificial Neural Network phishing detection model based on Decision Tree and Optimal Features, Appl. Soft Comput. 95 (2020) 106505. https://doi.org/10.1016/j.asoc.2020.106505.

[77] L. Breiman, Random Forests, Mach. Learn. 45 (2001) 5–32. https://doi.org/10.1023/A:1010933404324.

[78] G. Biau, Analysis of a Random Forests Model, J. Mach. Learn. Res. 13 (2012) 1063–1095. http://dl.acm.org/citation.cfm?id=2503308.2343682.

[79] B. Gregorutti, B. Michel, P. Saint-Pierre, Correlation and variable importance in random forests, Stat. Comput. 27 (2017) 659–678. https://doi.org/10.1007/s11222-016-9646-1.

[80] W. Hadi, Q.A. Al-Radaideh, S. Alhawari, Integrating associative rule-based classification with Naïve Bayes for text classification, Appl. Soft Comput. J. 69 (2018). https://doi.org/10.1016/j.asoc.2018.04.056.

[81] N. Abdelhamid, A. Ayesh, F. Thabtah, S. Ahmadi, W. Hadi, MAC: A Multiclass Associative Classification Algorithm, J. Inf. Knowl. Manag. 11 (2012) 1250011. http://www.worldscientific.com/doi/abs/10.1142/S0219649212500116.

[82] X. Yin, J. Han, CPAR: Classification based on Predictive Association Rules, in: Proc. 2003 SIAM Int. Conf. Data Min., Society for Industrial and Applied Mathematics, Philadelphia, PA, 2003: pp. 331–335. https://doi.org/10.1137/1.9781611972733.40.

[83] W. Li, J. Han, J. Pei, CMAR: accurate and efficient classification based on multiple class-association rules, in: Proc. 2001 IEEE Int. Conf. Data Min., IEEE Comput. Soc, 2001: pp. 369–376. https://doi.org/10.1109/ICDM.2001.989541.

[84] F. Thabtah, W. Hadi, N. Abdelhamid, A. Issa, PREDICTION PHASE IN ASSOCIATIVE CLASSIFICATION MINING, Int. J. Softw. Eng. Knowl. Eng. 21 (2011) 855–876. https://doi.org/10.1142/S0218194011005463.

[85] M. Lichman, UCI Machine Learning Repository, 2013. https://archive.ics.uci.edu/ml/citation_policy.html (accessed May 16, 2016).

[86] B. Liu, W. Hsu, Y. Ma, B. Ma, Integrating Classification and Association Rule Mining, Knowl. Discov. Data Min. (1998) 80–86. https://doi.org/10.1.1.48.8380.

[87] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA data mining software, ACM SIGKDD Explor. Newsl. 11 (2009) 10. https://doi.org/10.1145/1656274.1656278.