Detecting E-Banking Phishing Website using C4.5 Algorithm

Haya Alhamad[1†], Tagreed Alzyadh [1†]and Maria Altaib Badawi [2†]

Majmaah University, Department of Computer Science & Inf

Summary

Technology in our current era has become very important, as it has facilitated many services for us and has become faster and easier than before, as we can complete many things simultaneously and as quickly as possible. One of the most important services affected by technology is the electronic banking services, where the services provided by banks to their customers can be dealt with at maximum speed and does not require much effort and time, and with the progress and development of these services there are problems that have emerged and among the most important of these problems facing electronic banking services that affect their security It is a phishing problem. The problem of phishing is one of the biggest and most important problems, and it means that there is an attack and sabotage on sensitive data by people trying to steal and destroy information by creating fake sites similar to the original sites of the bank aimed at manipulating customers and stealing important information and the most important of this information is the IBAN number of the account owner in The bank or password. In this project, we work to limit and reduce the attack by identifying legitimate and fake sites using one of the data mining algorithms, which is c4.5 algorithm As this algorithm works on classification, so it can classify fake and legitimate sites where used this algorithm in WEKA is a tool for analyzing data and categorizing fake websites and reducing phishing in electronic banking services. After applying the algorithm to a dataset containing 32 attributes, the accuracy rate was 98.11%, which is considered a good percentage for classification, while the instance rate was only incorrect 1.89%.

Key words:
Phishing, Data mining, E-banking, C4.5 algorithm, weka.

1. Introduction

Technology has revolutionized our world and everyday life. Technology has created amazing tools and resources, put useful information within our reach. Modern technology has paved the way for multifunction devices such as smart watches and smartphones. Computers are faster, more mobile, and more powerful than before. With all these revolutions, technology has made our lives easier, faster, better, better and more enjoyable[1], with the rapid development of technology, including electronic banks, many problems have emerged, including cyber-attacks They are activities primarily directed against computers or network resources, cybercrimes are offences that can only be committed using a computer, computer networks or other form of information communications technology (ICT). These acts include the spread of viruses or other malware, hacking and distributed denial of service (DDoS) attacks, although there may be a variety of secondary outcomes from the attacks. For example, data gathered by hacking into an email account may subsequently be used to commit a fraud [2]. Phishing, sometimes called brand spoofing, in-voles the use of e-mails that originate from businesses with which targeted victims have been, or are currently associated. In the past few years there has been an alarming trend both in the increase and complexity of phishing attacks. Some of the most common businesses and industries associated with phishing include banks, online businesses. Unsuspecting victims receive e-mails that appear to be from these entities, usually suggesting suspicious activity regarding the account and requesting personal information (e.g., personal identification numbers, credit card numbers, and social security numbers) [3]. In our paper, we discuss detect phishing sites using the C4.5 algorithm to select mechanism more flexible and efficient to reduce phishing in banking services to clients also to reduce lose confidence in the e-banking provided by the bank about attackers.

2. Literature review

In this part cover a similar and related work, detailed description of the scientific articles is presented as a literature review, in paper [4], They used the RIPPER data mining algorithm for classification. Moreover, after the email is evaluated and classified as a phishing email, the system proactively disposes of the phishing site or phishing page by sending a notification to the system administrator of the host server that it is hosting the phishing site that might lead to the site being removed. After categorizing phishing scams, the system retrieves the site, IP Detecting any Phishing website is really a complex and dynamic

Manuscript received November 5, 2020
Manuscript revised November 20, 2020
https://doi.org/10.22937/IJCSNS.2020.20.11.6

problem involving many factors and criteria. Because of the ambiguities involved in phishing detection, fuzzy data mining techniques can be an effective tool in detecting phishing websites. we propose a method which combines fuzzy logic along with data mining algorithms for detecting phishing websites. Here, they define 3 different phishing types and 6 different criteria for detecting phishing websites with a layer structure. They have used RIPPER data mining algorithm for classification. Furthermore, after the email has been assessed and classified as a phishing email, the system proactively gets rid of the phishing site or phishing page by sending a notification to the system administrator of the host server that it is hosting a phishing site which may result in the removal of the site. Furthermore, after classifying the Phishing email, the system retrieves the location, IP address and contact information of the host server. address and contact information of the host server. Results in the research showed that the RIPPER algorithm correctly achieved 85.4% of phishing emails and 14.6% of incorrectly categorized phishing errors. Phishing page removal success rate is 81.81%. In paper [5], They developed an AC data mining method to discover correlations among features and produces them in simple yet effective rules, seems a potential solution that may effectively detect phishing websites with high accuracy. According to experimental studies, AC often extracts classifiers containing simple ''If-Then'' rules with a high degree of predictive accuracy. In this paper, they investigate the problem of website phishing using a developed AC method called Multi-label classifier based associative classification (MCAC) to seek its applicability to the phishing problem. They also want to identify features that distinguish phishing web- sites from legitimate ones. In addition, they survey intelligent approaches used to handle the phishing problem. Experimental results using real data collected from different sources show that AC particularly MCAC detects phishing websites with higher accuracy than other intelligent algorithms. Further, MCAC generates new hidden knowledge (rules) that other algorithms are unable to find and this has improved its classifiers predictive performance. The specific features of phishing sites. In addition, they scan smart methods used to tackle the problem of deception. They indicated that experimental results using real data collected from various sources that AC and MCAC discover phishing sites more accurately than other smart algorithms. Moreover, MCAC creates new hidden knowledge (rules) that other algorithms cannot find to improve the predictive performance of its classifiers. This search method has revealed new rules associated with more than one category giving the user a new type of useful information. These rules also enhanced the accuracy of ranking in discovering fake websites. Moreover, they were able to identify the important features of phishing sites using frequency analysis and the method of selecting the Chi-square feature. In paper [6], The concept the used is an end-host based anti-phishing algorithm, called the LinkGuard, they explain the basic algorithm of Link Guard Approach which can detect the phishing content, based on the characteristics of the phishing hyperlink. The LinkGuard algorithm: LinkGuard works by analyzing the differences between the visual link and the actual link. It also calculates the similarities of a URI with a known trusted site. Since LinkGuard is a rule-based heuristic algorithm, it may cause false positives (i.e., treat non-phishing site as phishing site) and false negatives (i.e., treat phishing site as nonphishing site) LinkGuard is based on the careful analysis of the characteristics of phishing hyperlinks. They have implemented Link Guard for Windows XP. their experiment showed that LinkGuard is lightweighted and can detect up to 96% unknown phishing attacks in real-time. In paper [7], they are applying fuzzy logics along with data mining algorithms. The first phase in preventing phishing problem is the detection of a phishing attack, they propose a technique to detect and prevent the phishing attacks on e-mail. It is an end user application that uses hyperlink and URL feature set to detect phishing attacks and makes the use of digital signature to prevent the attack. Several experiments have been conducted using different rule-based classification algorithms to extract new hidden knowledge that can help in detecting phishing websites. The results showed that we could improve the prediction accuracy relying only on nine features, those are: "Request URL, Age of Domain, HTTPS and SSL, Website Traffic, Long URL, Sub Domain and Multi Sub Domain, adding prefix or Suffix Separated by (-) to Domain, URL of Anchor and Using the IP Address". After conducting the experiments on the nine chosen features, the error-rate has decreased for all algorithms.

Table 1: Comparison of papers

Feature	Use of linguistic variable	Use IP address	Heuristic-based techniques	Generate new hidden knowledge	Determine the features that distinguish phishing sites
Detection of phishing sites using data mining techniques	✓	✓	✗	✗	✗
Phishing detecting based associative classification data mining	✗	✗	✗	✓	✗
Detecting of E-banking phishing website	✗	✓	✗	✗	✓
Modeling intelligent phishing detecting system for E-banking using fuzzy data mining	✗	✓	✗	✗	✗
Detecting E-Banking Phishing Website using c4.5 algorithm	✓	✓	✓	✗	✓

Table 1 presents some features extracted from the previous papers from our point of view and compared with our proposed paper.

3. Methodology

We will use data mining algorithm. Data mining is the process of discovering useful patterns and trends in large datasets [8], data mining also called knowledge discovery in databases (KDD) is a collection of exploration techniques based on advanced analytical methods and tools for handling a large amount of information. The techniques can find novel patterns that may assist an enterprise in understanding the business better and in forecasting techniques that have been developed over the last 50 years. Others are related to techniques that have been developed in statistics, sometimes called exploratory data analysis [9]. The C4.5 algorithm (known in weka j48) that we chose to solve the presented problem, C4.5 was proposed by J. Ross Quinlan which from the training set, forms a decision tree [10]. C4.5 is a suite of algorithms for classification problems in machine learning and data mining [11]. C4.5 is not one algorithm but rather a suite of algorithms—C4.5, C4.5-no-pruning, and C4.5-rules—with many features. We present the basic C4.5 algorithms first and the special features later. All tree induction methods begin with a root node that represents the entire, given dataset and recursively split the data into smaller subsets by testing for a given attribute at each node. The subtrees denote the partitions of the original dataset that satisfy specified attribute value tests. This process typically continues until the subsets are "pure," that is, all instances in the subset fall in the same class, at which time the tree growing is terminated [12].

3.1 Tools

Data mining involves analyzing a large set of data to reveal new forms and methods of database management and processing using an algorithm. One of the programs that deals with data mining is WEKA, that we will use, an open source software provides tools for data preprocessing, implementation of several Machine Learning algorithms, and visualization tools so that you can develop machine learning techniques and apply them to real-world data mining problems [13]. WEKA It is a set of tools and programs that help the user to collect, analyze and process data. Application of rules and algorithms. Weka provides accessible tools that allow opening and editing a set of data. It is also possible to change data contents, change features, and categorize data. Its modular, extensible architecture allows sophisticated data mining processes to be built up from the wide collection of base learning algorithms and tools provided. Extending the toolkit is easy thanks to a simple API, plugin mechanisms and facilities that automate the integration of new learning algorithms with WEKA's graphical user interfaces. workbench includes algorithms for regression, classification, clustering, association rule mining and attribute selection [14].

4. Experience and method

We deal with dataset contains 10000 phish and legitimate websites after reducing the number of instances from 11055, and its features [15].

4.1 Preprocessing

Preprocessing is the first step to dealing with a data set .The data pre-processing is the technique of converting data into a format that can be understood and dealt with, make sure the data is complete and ready for processing because often the data is incomplete or unformatted. In order to preprocess the dataset in Weka, the dataset must be converted into a format Weka understands, and Weka handles files in the ARFF (Attribute Relationship File Format) format.
The Excel file that contains tha datasetswill be converted to CSV format so we can then convert the file from CSV format to ARFF format. Convert the file to CSV directly through Excel after that we converted our file directly via Weka using ArffViewe (Tools -> ArffViewer -> then open our CSV file. -> Then File -> Then Save As and select Arff Data Files -> Save), this way our dataset was converted to ARFF format.

Preprocessing in research includes these steps:

1. Data cleaning: Data cleaning is a process of checking the quality of the data and that the dataset we are dealing with is free of errors, and some data may be unclean because it contains some problems of noise, outliers, missing values and duplicate data.
2. Data transformation: In this step, data is transferred from one type to another.
3. Data reduction: Is the use of methods that reduce original data in order to increase efficiency. There are several methods to reduce data, of these methods (feature selection method). This method determines the best attributes in the data set and gives the best result, and identifies which attributes are not good and eliminates them.

4.2 Apply Algorithm

After performing the data preprocessing in the WEKA program, the next step is to choose the algorithm and apply it. Since the C4.5(J48) algorithm is one of the classification algorithms, we select the algorithm from the classification list, as shown in figure 3.

After which the algorithm can be executed by select button "start" to start execution.

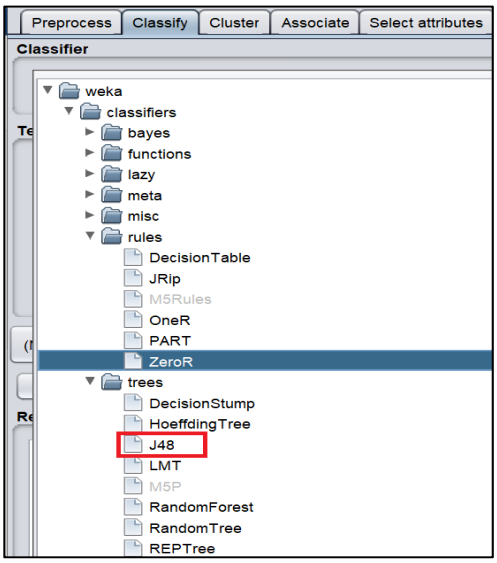

Figure 3: classification menu.

4.3 Experiment Results

We used the C4.5 algorithm in the WEKA program to analyze and classify phish and legitimate sites according to the data set, and as a result, the accuracy of using the algorithm is shown by showing the classification accuracy. In Figure 4, the result of the algorithm's accuracy in classifying phishing sites is shown to 98.11%, since we have 10,000 cases, this means that the algorithm has classified 9811 sites correctly and 189 sites are classified incorrectly.

```
=== Summary ===

Correctly Classified Instances        9811               98.11 %
Incorrectly Classified Instances       189                1.89 %
Kappa statistic                          0.9617
Mean absolute error                      0.032
Root mean squared error                  0.1264
Relative absolute error                  6.4756 %
Root relative squared error             25.4473 %
Total Number of Instances            10000
```

Figure 4: The result of the classification accuracy

also, Figure 5 shows analysis the attributes in the decision tree. The attribute is selected at each stage by calculating "information acquisition". Information gain is the measure that is useful in building a decision tree, information gain ratio is the ratio of obtaining information gain with intrinsic information. To reduce the bias towards multi value attributes by taking the number and size of branches in a calculation when selecting attributes. This is useful as a consideration for logarithmic probabilities to measure the impact of this type of calculation in a dataset [16].

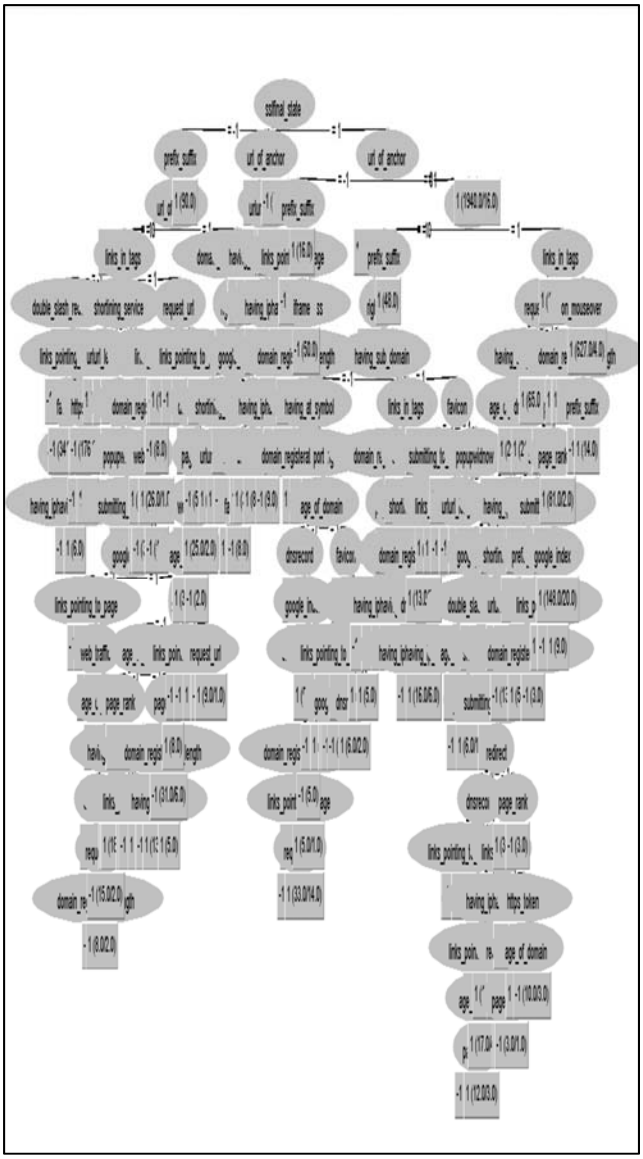

Figure 5: The constructed decision tree using WEKA

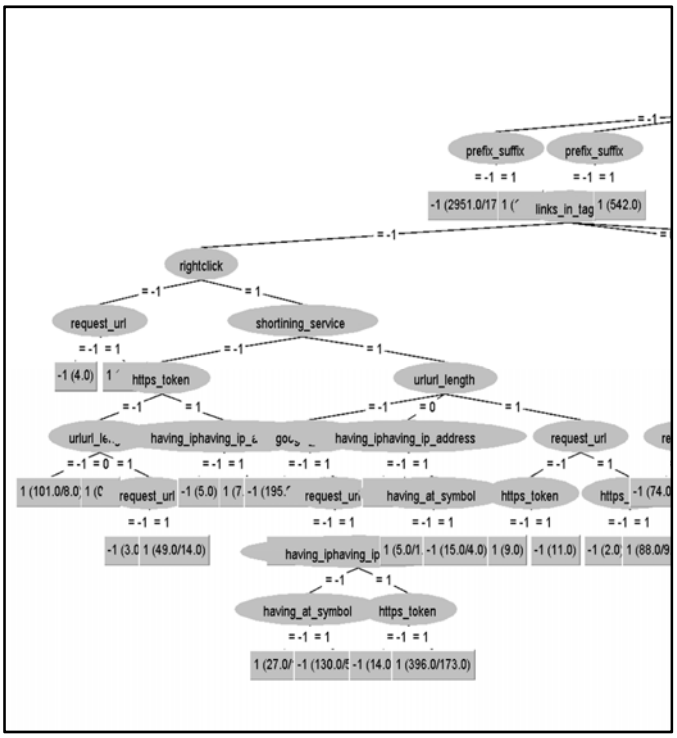

Figure 6: Part of decision tree

In Figure 6, the decision tree image is displayed after we reduced the number of features, to make it clearer.

5. Statistical Analyses and Evaluation

After trial and results, we came to the conclusion that C4.5 is one of the best data mining algorithms for classification, based on the accuracy of the result. In Figure 7, a comparison between the C4.5 algorithm and some of the classification algorithm in the Wicca software, the result indicates that the algorithm had an accuracy rate of 98.11%, which is the highest rate for the mentioned algorithms, while the invalid instance rate was only 1.89%.

We compared C4.5 algorithm with the following algorithms:

- REP Tree:

Basically Reduced Error Pruning Tree ("REPT") is fast decision tree learning and it builds a decision tree based on the information gain or reducing the variance. REP Tree is a fast decision tree learner which builds a decision/regression tree using information gain as the splitting criterion, and prunes it using reduced error pruning. It only sorts values for numeric attributes once. Missing values are dealt with using C4.5's method of using fractional instances [17].

- Naïve Bayes:

Naïve Bayes classifier is based on Bayes theorem. It has strong independence assumption. It is also known as independent feature model. It assumes the presence or absence of a particular feature of a class is unrelated to the presence or absence of any other feature in the given class [18].

- JRip:

JRip is an optimized version of IREP (Cohen, 1995). It was introduced by William W. Cohen. With the repeated incremental pruning JRip produce error reduction [19].

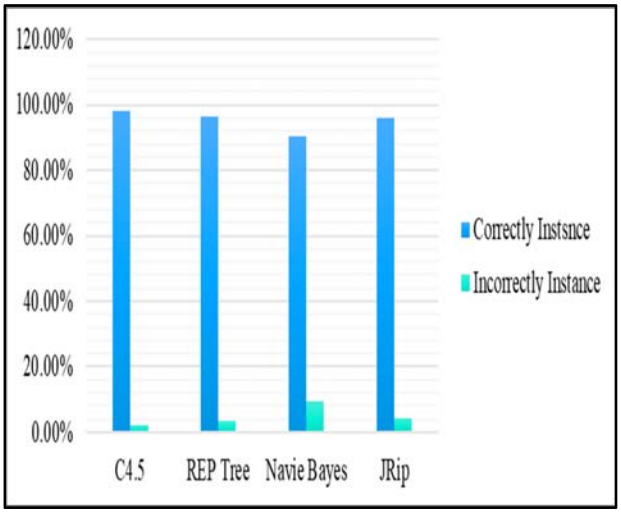

Figure 7: comparison between Classification algorithms.

Conclusion and Future work

Phishing has become one of the most important problems facing online banking services of our time. It is also a big problem facing everyone who uses the Internet today, and phishing sites are fake websites created by the attackers where the attacker tries to create fake websites that look like real sites and try to steal confidential customer data, then the information is used to impersonate the victims in order to empty their bank accounts, and manage Fraudulent auctions, money laundering, etc. In an attempt to reduce this problem in this paper entitled "Detecting E-Banking phishing website using the C4.5 algorithm", we

used the C5.4 algorithm for classification in the WEKA program, by analyzing data and classifying fake and legitimate sites to reduce the problem of phishing in banking services e. After processing the data and applying the algorithm, the accuracy rate reached by the algorithm 4 was 98.11%, while the algorithm's error rate was only 1.89%.
In future work, we aspire in to develop the algorithm to increase its accuracy in analyzing and identifying legitimate and phishing sites. We will also try to combine it with the Internet and search engines to try to prevent phishing sites from working automatically, to avoid its danger early.

Acknowledgment

We put a lot of effort into this research to make it an outstanding work and thus we gained good experience in writing a scientific paper. We would like to thank Majmaah University for its support and encouragement. The project supervisor, Dr. Maria Altaib Badawi, made great efforts with us on this research and answered questions at any time, and encouraged us to obtain the best results. We would also like to thank our family for our continuous support and assistance.

References

[1] (Technology in Our Life Today and How It Has Changed) https://www.aginginplace.org/technology-in-our-life-today-and-how-it-has-changed/

[2] McGuire,M & Dowling,S (2013). Cybercrime: A review of the evidence, Home Office, Cyber-crime-a-review-of-the-evidence-chapter-1-cyberdependent-crimes.pdf.

[3] Atkins.B, Huang.W, (2013), A Study of Social Engineering in Online Frauds, Open Journal of Social Sciences.

[4] Khade,A & Shinde,S,K,(2013), Detection of Phishing Sites Using Data Mining Techniques, International Journal of Engineering Research & Technology (IJERT).

[5] Abdel-Hamid,N . Ayesh,A, &Thabet.T(2014),Phishing detection based Associative Classification data mining , Expert Systems with Applications.

[6] Reddy,E,K . Rajamani & Saradhi.M.V ,DETECTION OF E-BANKING PHISHING WEBSITES, International Journal of Modern Engineering Research.

[7] Aburrous,M . Hossain,M,A . Dahal,K & Thabatah,F (2009) ,Modelling Intelligent Phishing Detection System for e-Banking using Fuzzy Data Mining , International Conference on CyberWorlds.

[8] UnnisaBegum.A, Hussain.M, Shaik.M. (August 2019). Data Mining Techniques for Big Data. International Journal of Advanced Research in Science, Engineering and Techology.

[9] GUPTA.G.K. (2015). Introduction to data mining with case studies. PHI Learning Private Limited. Delhi.

[10] S. Sikder, S. Metya, R. Goswami, (2019), Exception-Tolerant Decision Tree / Rule Based Classifiers, Ingénierie des Systèmes d Inf.

[11] Quinian. R. (1993). C4.5: programs for machine learning. Place of publication not identified: Morgan Kaufmann.

[12] Wu, X., & Kumar, V. (2009). The top ten algorithms in data mining. Boca Raton: Chapman & Hall-CRC.

[13] "Tutorials Point", (2019), https://www.tutorialspoint.com/weka/weka_tutorial.pdf

[14] Witten,l & Hall,M & Holmes,G & Frank,E .(2009) . The WEKA data mining software: An update. ReasearchGat, ACM SIGKDD Explorations Newsletter.

[15] Akash Kumar, "Phishing website dataset", https://www.kaggle.com/akashkr/phishing-website-dataset .

[16] Lestari.C, Alamsyah, (2020), Accuracy of C4.5 Algorithm Using Information Gain Ratio and Adaboost for Classification of Chronic Kidney Disease, Journal of Soft Computing Exploration.

[17] Srinivasan,B &.Mekala,P (2014) , Mining Social Networking Data for Classification Using Reptree , International Journal of Advance Research in computer science and management studies,

[18] Shinde,R , Arjun,S , Patil,P & Waghmare,J (2015), An Intelligent Heart Disease Prediction System Using K-Means Clustering and Naïve Bayes Algorithm , (IJCSIT) International Journal of Computer Science and Information Technologies, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.734.7958&rep=rep1&type=pdf

[19] Shahzad,W , Asad,S & Khan,M,A (2013) , Feature subset selection using association rule mining and JRip classifier ,International Journal of Physical Sciences , https://academicjournals.org/journal/IJPS/article-full-text-pdf/22AC4CB27262