Journal of Information Systems & Information Technology (JISIT) *Vol. 5 No. 2 2020*
ISSN: 2478-0677 *61-71*

Predictive Data Mining for Phishing Websites: A Rule Based Approach

Abdul Raheem Fathima Shafana[1], Abdul Raheem Fathima Shihnas Fanoon[2]

[1,2] Department of Information and Communication Technology,
Faculty of Technology,
South Eastern University of Sri Lanka

arfshafana@seu.ac.lk, fanoonarfs@gmail.com

Abstract. The rapid advancement in internet has paved way for several serious crimes, of which phishing occupies a very important place. Phishing is a form of cybercrime where an attacker mimicking a legitimate website or a person or an organization redirects the victims to steal confidential data through e-mail, malwares or some other social engineering platforms. Victims prominently suffer from financial loss and private data loss. The serious outbreak of phishing has paved way for many researches, though comprehensive and accurate solution has not been proposed so far for thwarting its impact. This paper aims to develop a resilient model to predict phishing scam by means of classification algorithms of data mining. Five algorithms were chosen for this purpose and a comparative study was undertaken for their performances, accuracy, error rate and efficiency. The rules generated from the algorithms showed up a relatively better performance than the existing phishing detection tools.

Keywords: Phishing, Data Mining, Classification, PART, Website Legitimacy

1 Introduction

Phishing is one of the most serious threat to internet users since it is one of the extremely common attack vectors that causes potential impacts and serious risks to the users. Phishing refers to the process of tricking or socially engineering the customers of organizations into disclose their confidential and sensitive information where it is used for despicable use [1]. The phishers make use of spam mails, bots, online business and online advertisements to drive phishing in a large scale and to impersonate as real firms and steal the confidential information tactfully.

Phishing is believed to be the oldest and easiest ways of stealing one's sensitive information, rather, its impact keeps on escalating exponentially. Presently, it is considered to be one of the most organized crimes of 21st century [2]. A report published by Forbes mentioned that approximately a sum of $500m has been lost in the US businesses alone by the phishing attack [3]. Nearly 90, 000 unique emails and more than 130,000 unique phishing websites have been reported by Anti-Phishing Working Group (APWG) in the year 2009[2]. 93% of the reported attacks are from financial and banking sector of the United States that accounts to monetary loss between $100m and $3bn [2].

The exponential growth of phishing tricks has superseded the existing protection mechanisms. Hence, it is more challenging globally to track the attackers easily. It has been

© *Faculty of Management and Commerce*
South Eastern University of Sri Lanka (SEUSL)
University Park, Oluvil 32360. Sri Lanka

Received: 19th Nov, 2020
Revised and Accepted: 22nd Dec, 2020

reported by the Anti- Phishing Working Group (APWG) that the phishing scams are increasing at a rate of 56% per month and has been forecasted that it would be the main risk over the internet [4]. As per the latest statistics from APWG [5] Phishing attacks that were recorded at the second quarter of 2019 showed a steady increase as that of second half of 2018. Therefore, developing a feasible protective mechanism is an urgent need of the cybersecurity to safeguard the online users from phishing attack.

There has been a series of researches globally for devising an efficient anti-phishing solution with special focus on detection and prevention. The proposed solutions had a long list and they can be majorly categorized into three classes namely phishing prevention solution, user training solution and phishing detection solution [6].

I. *Phishing Prevention*: Though this technique can serve as an extra protective layer, these mechanisms demand consistent and continuous update and support on both the website's side and on the user's side for the coordination. In addition, these solutions would lead to complex user interfaces, incur extra cost for the computation of each authentication, and would also require users to keep extra authentication devices, making its implementation and usage bit cumbersome.

II. *User Training Scheme*: This technique requires its users to gain insight of phishing and train them to protect themselves. However, this mechanism is not much preferred as it does not provide with fool-proof solution and it is hard to educate the novice users about the technical aspects of cybersecurity.

III. *Phishing Detection*: This technique has gained wide popularity over other mechanisms due to its competitive advantages provided for the novice users as well. As this either blocks or notifies the user of the authenticity of the website, this is considered to be the most efficient technique. This method expects minimal user training and does not require any changes to the existing authentication schemes used by a website.

This research aligns with the whitelist approach of the Phishing Detection Mechanisms where several features are gathered from the website in the real time and such features are used in classification of the website for its legitimacy. In the circumstance where fraudulent practices have become accustomed to the new technological opportunities in order to keep pace with, an anti-phishing mechanism promising a higher accuracy is of urgent need. Therefore, this paper utilizes the classification-based data mining techniques on a very large dataset obtained from a trustworthy source to solve the difficulty in detecting the phishing websites.

The rest of the paper is organized as follows: Section II outlines the related works in the field following the Methodology presented in Section III. Section IV provides the results obtained from the experiment. Finally, Section V concludes the paper with the conclusion and discussion.

2 Related Works

Web phishing involves the attempt of acquiring sensitive information for malicious reasons by impersonating the trustworthy websites on the internet. The sensitive information could be the password, username and credit card details. Researchers have carried on several researches to detect these websites.

The direct way of detecting a phishing website is the application of black list or white list. This is done by accessing the URL in a database and decide whether it is a phishing or a legitimate website [7]. Blacklist approach can be used in two ways to detect phishing websites. The first method is to include five heuristics to compute simple combinations of known phishing sites and then identify the new phishing URLs. The second method is an application

of an algorithm that could approximately match with the phishing websites and detect them [8]. Frequently used browsers Firefox [9] and Chrome [10] also applies their own or third-party black-white listing approach to detect phishing websites. But the drawback of this approach is that it is not real-time and hence will take more time and cost more to detect a website as a legitimate or phishing.

In the year 2006, Anthony Fu et al. proposed a methodology called Earth Mover's Distance (EMD) to detect the phishing websites [11]. This method is based on the visual similarity of the web pages. Though this approach has a higher accuracy, it requires a large amount of data as a priori knowledge. A supervised learning approach has been suggested by J. Ma et al. in order to classify URLs as legitimate or phishing [12]. Another approach has been suggested by Liu P and his fellow researchers for filtering spam emails. They used a text from a spam email as a keyword to perform a complex processing for the word and according to their study, they were able to obtain an accuracy rate of 92.8% [13].

An open source framework known as "Fresh_Phish" has been introduced by Hussein et al. in the year 2017, which creates machine learning data set and python is used for query purpose. The framework analyses on the time taken for training the detection model[14]. A combinational approach of algorithms has been introduced by Priyanka et al. in the year 2015. They have used a novel approach by combining the algorithms Adaline and Backpropion along with Support Vector Machine (SVM) [15].

Agrawal N et al., in the year 2016, proposes a content filtering technique to filter the spam emails using the header information on the incoming email. The main purpose of this method is to optimize the performance of the network and server [16]. Property selection has been used by Thomas J. et al. for spam filtering of emails. Different feature selection methods have been adopted here in comparison for classification and estimation of emails. Among the features, Weighted Information Mutual Feature has been identified to be the most effective approach [17].

3 Methodology

3.1 Attributes

The detecting of legitimacy of a website is a real-world classification problem in which the data mining approaches could be applied to extract the hidden patterns and the nontrivial knowledge in the data set. The data set chosen for this research has 10,000 instances and 48 attributes extracted from 5,000 legitimate websites and 5,000 web pages which were downloaded within the periods from January to May 2015 and from May to June 2017[18]. Attributes are the effective minimal set of phishing website features. Attributes take values -1 and 1 for phishing and legitimate websites respectively and few features have value 0 which denotes that they are suspicious. The attributes have been categorized into six categories as:

- Content based features (Table 1)
- Domain based features (Table 2)
- HTML based features (Table 3)
- Symbol related features (Table 4)
- Web page URL features (Table 5)
- Correlated features (Table 6)

Table 1. Content based features.

Features of Websites	Description
PctExtHyperlinks	External hyperlinks percentage in the HTML
PctExtResourceUrls	External resource URLs percentage in the HTML
ExtFavicon	Installation of favicon different to the URL hostname
ExtFormAction	Existence of an external URL on the action form
PctNullSelfRedirectHyperlinks	Percentage of hyperlinks with empty value, auto redirecting value, abnormal values etc.
FakeLinkInStatusBar	Existence of MouseOver command in URL

Table 2. Domain based features

Features of Websites	Description
NumDashInHostName	Count of '-'character in Hostname
IpAddress	IP Address on the website URL
DomainInSubDomains	Use of TLD or cc TLD in the URL subdomain
DomainInPaths	Use of TLD or cc TLD in the URL link
HttpsInHostname	Disordering of Https in URL Hostname
EmbeddedBrandName	Existence of a brand name in the Domain
FrequentDomainNameMismatch	Matching of frequent domain name to the URL

Table 3. HTML based features

Features of Websites	Description
InsecureForms	Existence of URL contend without HTTPS protocol
RelativeFormAction	Existence of a relative URL in the action form
AbnormalFormAction	Existence of abnormal URL in the action form
RightClickDisabled	Existence of a JavaScript command to turn off right click
PopUpWindow	Existence of popup window command in JavaScript
SubmitInfoToEmail	Existence of 'mailto' source code in the HTML
IframeOrFrame	Usage of iframe or frame in HTML
MissingTitle	Leaving the title tag empty in HTML
ImagesOnlyInForm0	Existence of only images in the HTML form

Table 4. Symbol based features

Features of Websites	Description
NumDots	Count of '.' Character
NumDash	Count of '-' character
AtSymbol	Existence of '@' symbol
TildeSymbol	Existence of '~' symbol
NumUnderscore	Count of '_' character
NumPercent	Count of '%' character
NumAmpersand	Count of '&' character
NumHash	Count of '#' character
NumNumericChars	Count of Numeric character
DoubleSlashInPath	Existence of '//' character
NumSensitiveWords	Count of sensitive words (secure, account, login, etc.)

Table 5. Web page URL features

Features of Websites	Description
SubDomainLevel	Count of subdomain levels
PathLevel	URL depth
UrlLength	URL length
NumQueryComponents	Count of query components
NoHttps	Existence of HTTPS in URL
RandomString	Existence of random string in URL
HostNameLength	Length of hostname
PathLength	Length of the path
QueryLength	Length of the query

Table 6. Correlated features

Features of Websites	Description
SubdomainLevelRT	Sub-domain level correlated
UrlLengthRT	URL length correlated
PctExtResourceUrlsRT	External resources length correlated
AbnormalExtFormActionR	Form abnormal actions correlated
ExtMetaScriptLinkRT	Link of meta script correlated
PctExtNullSelfRedirectHyperlinksRT	Null self-redirect hyperlinks correlated

3.1 Classification Algorithms

A classifier, in a classification problem, predicts the output by learning the attributes which are the input for the data mining process. A website is predicted either as legitimate or phishing by learning the features and hidden patterns and correlations among those features.

Five different classification algorithms, OneR, PART, Decision Table, JRip and J48 have been implemented to predict the legitimacy and phishing nature of a website depending on their application of different strategies on the data sets.

3.1.1 OneR (One Rule):

OneR is a classification algorithm in which one rule will be generated for each predictor in the data and then a rule with the least total error rate will be selected as “One Rule”. OneR algorithm needs a frequency table for against the target for each predictor in order to create a rule. Results show that OneR produces rules only considerably less accurate than contemporary classification algorithms and produces results that could be easily interpreted by human [19].

OneR algorithm works as follows:

```
For each feature,
  For each feature, make a rule as follows:
      Count the occurrence of each target value
      Identify the mode class
      Assign the rule to the mode class and feature value
  Calculate the total error of the rules of each predictor
Choose the feature with the least total error.
```

3.1.2 PART Algorithm:

Project Adaptive Resonance Theory is known to be the PART algorithm. A separate-and-conquer technique is adopted to study the rules and build decision trees based on the divide-and-conquer method. A decision tree consists of planned set of rules in which a new data is compared to each rule and the feature will be assigned the class of the first matching rule. A partial C4.5 algorithm will be built in each iteration in PART algorithm and the best leaf is made into a rule [20].

3.1.3 Decision Table:

Similar to Decision Trees, Decision Tables is also a classification model used for predictions in data mining approaches. A decision Table is a hierarchical table in which a new table will be formed by breaking down each higher-level entry in the Decision Table. The values of a pair of additional attributes is used to break the tables and form another new table. Also, a wrapper method is applied in order to identify the effective subset of features [21].

3.1.4 JRip Algorithm:

Repeated Incremental Pruning to Produce Error Reduction (RIPPER)", a propositional rule learner is implemented in JRIP algorithm and an ordered rule list is generated using a sequential covering algorithm. JRip algorithm has four stages, namely; Growing a Rule, Pruning, Optimization and Selection [22].

3.1.5 J48:

J48 is a classification algorithm that generates a Decision Tree by combining both C4.5 algorithm and an extension of ID3 algorithm. It uses divide-and-conquer approach to classify.

3.2 WEKA

The data mining tool, WEKA has been used for the classification process to detect the phishing websites. WEKA is a tool developed using Java language by the Machine Learning Group, University of Waikato, New Zealand with the vision to develop a state-of-the-art-software as to develop techniques on machine learning and apply them to real-world problems related to data mining. WEKA is an open source software with a collection of machine learning algorithms and data mining tasks [23].

3.3 Approach

Five different classification algorithms have been implemented to detect the legitimate and phishing websites. The data set was classified with ten-folds cross validation test using WEKA tool to produce rules on detection of phishing websites. The research focuses on the accuracy, error rate, the number of rules generated, and the time taken for the classification factors to study the performance of each algorithm.

4 Experimental Result Analysis

The repository contained 10000 instances of phishing website relations where they are classified with five data mining classifiers namely, Decision Table, J48, JRip, OneR and PART algorithms. The following table (Table 7) depicts the correct and incorrect classification of instances for each of the classifier concerned and the graphical representation is given in Fig. 1.

Table 7. Correct and Incorrect Classification Instances

Algorithms	Correctly Classified Instances	Incorrectly Classified Instances
Decision Table	9579	421
JRip	9730	270
OneR	9187	813
PART	9760	240
J48	9731	269

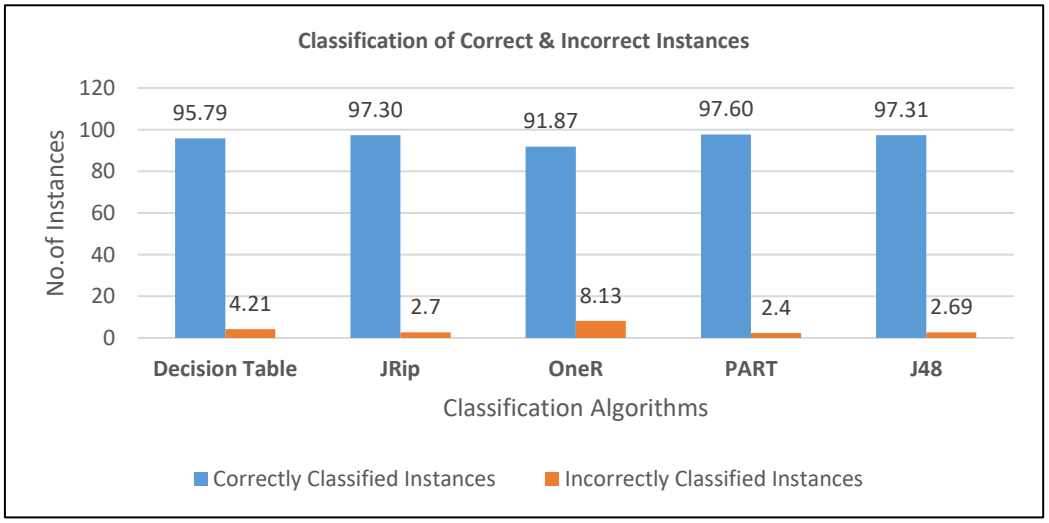

Fig. 1. Algorithms with its classification of correct and incorrect instances

The algorithms used for the study had a relatively higher prediction rate where all the algorithms had a prediction rate above 91%. Among the algorithms used for the study, by attaining a percentage of 97.60%, PART algorithm has obtained the maximum accuracy whereas OneR obtained the least prediction rate.

Error rate parameters were also used for the evaluation of the algorithms under study. For this, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Relative Absolute Error (RAE) and Root Relative Squared Error (RRSE) have been used. The error rate evaluation algorithms and their respective rate is as given in Table 8 below. It could be inferred from the table below that the PART algorithm exhibit the lowest error rate and the second lowest is by the J48 algorithm.

Table 8. Error rate evaluation of algorithms

Algorithms	MAE	RMSE	RAE	RRSE
Decision Table	0.0839	0.195	16.7876	39.007
JRip	0.0426	0.1585	8.5102	31.6925
OneR	0.0813	0.2851	16.26	57.0263
PART	0.0277	0.1517	5.5359	30.3361
J48	0.0361	0.1588	7.2138	31.7553

The time elapsed for the classification was also recorded as represented in Fig. 2. OneR algorithm records the least time where the Decision table takes the maximum time. As OneR algorithm prune the results with a single rule, it is the most effective one with one rule, as the name suggests. However, it could be noted that, JRip is considered to be effective beside OneR having the second least number of rules produced. Fig. 3 depicts the count of rules used by each algorithm.

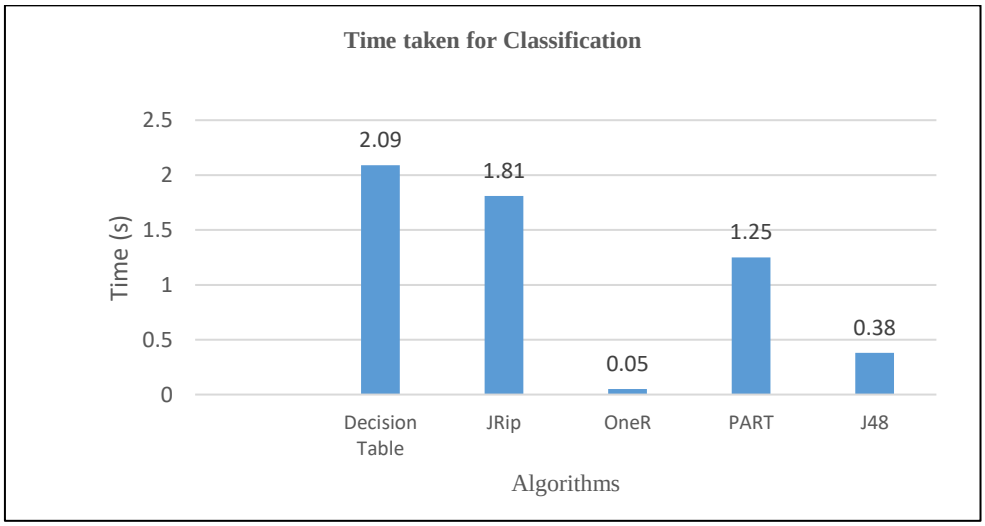

Fig. 2. Time elapsed for the classification

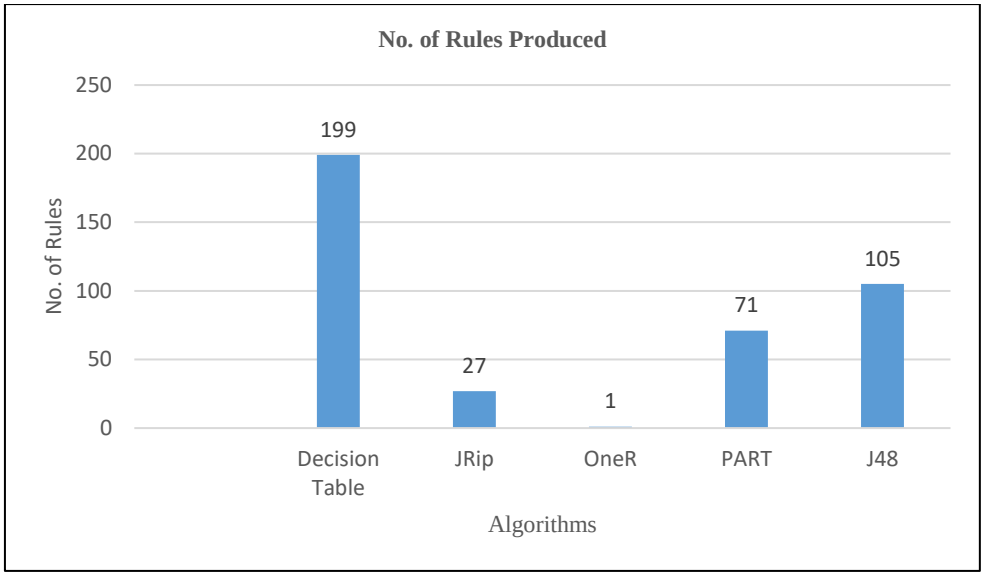

Fig. 3. Algorithms and the number of rules produced

The weighted average of the precision was also taken into consideration. Among all the five algorithms used for the study, PART achieves the highest precision with the weighted average precision of 97.60% which is graphically illustrated in Fig. 4.

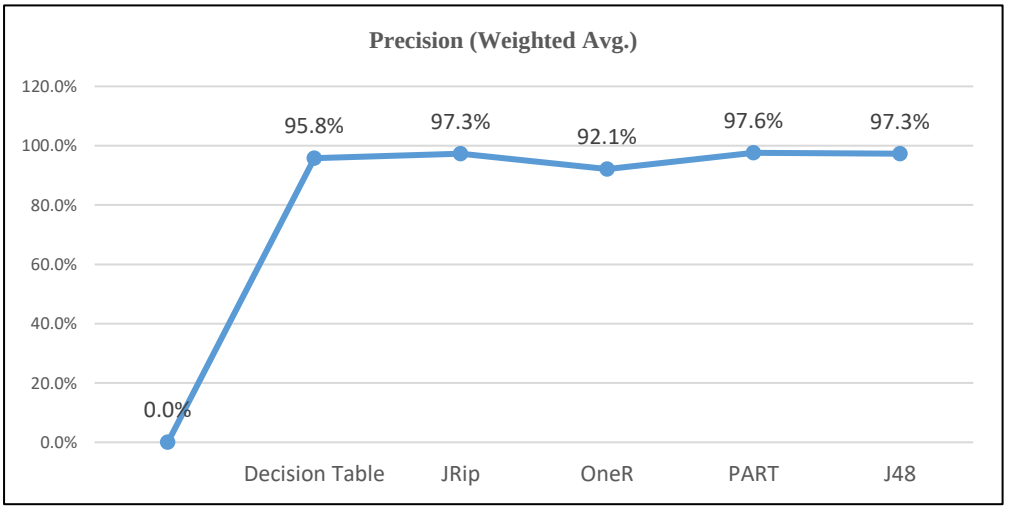

Fig. 4. Weighted Average Precision of Algorithms

4 Conclusion

The phishing websites have been predicted by using the competitive advantages of five algorithms namely Decision Table, J48, JRip, OneR and PART. The above algorithms have been widely used in classification-based data analytic domains. The algorithms allowed the research to unveil implicit knowledge from a large dataset consisting of 10000 instances to predict the legitimacy of websites. The classified outputs were compared for their efficiency and performances in terms of precision, error rate, accuracy, time duration and the number of rules produced.

It could be well noted from the experimental analysis that all the algorithms under this study had a higher prediction rate. In particular, PART algorithm can be considered as the most effective algorithm as it has the highest accuracy and precision of 97.60% while the error rate is also less as 0.0277.

The rules generated in the study confirmed that correlation exists between the website features and a model could be developed based on the rules generated. Hence, this particular model developed can be used for the prediction of phishing websites in order to ensure the internet users with the secure infrastructure. This model can be used as an extra shield against the phishing of confidential and sensitive data.

References

[1] Gunter Ollmann. (2016). The Phishing Guide Understanding; Preventing Phishing Attacks., IBM Internet Security Systems.

[2] Vayansky, I. and Kumar, S. (2018). Phishing – challenges and solutions. Comput. Fraud Secur. 2018(1). 15–20

[3] Mathews, L. (2017). Phishing Scams Cost American Businesses Half a Billion Dollars A Year Retrieved on 8 February 2020 from https://www.forbes.com/sites/leemathews/2017/05/05/phishing-scams-cost-american-businesses-half-a-billion-dollars-a-year/#210c3d33fa1c.

[4] James, L. (2005). Phishing Exposed.

[5] APWG. (2019). Phishing Activity Trends Report 2 Quarter.

[6] Varshney, G., Misra, M. & Atrey, P.K. (2016). A survey and classification of web phishing detection schemes. Secur. Commun. Networks. 9(18). 6266–6284

[7] Yi, P., Guan, Y., Zou, F., Yao,Y., Wang,W., & Zhu,T. (2018). Web phishing detection using a deep learning framework. Wirel. Commun. Mob. Comput.2018

[8] Prakash, P., Kumar, M., Rao Kompella,R. & Gupta, M. (2010). PhishNet: Predictive blacklisting to detect phishing attacks. Proc. - IEEE INFOCOM

[9] Internet for people, not profit — Mozilla. Retrieved on 03 February 2020 from https://www.mozilla.org/en-US/.

[10] Google Chrome - The New Chrome & Most Secure Web Browser. Retrieved on 03 February 2020 from: https://www.google.com/chrome/

[11] Fu, A. Y., Liu,W., & Deng, X. (2006). Detecting phishing web pages with visual similarity assessment based on Earth Mover's Distance (EMD). IEEE Trans. Dependable Secur. Comput. 3(4). 301–311

[12] Ma, J., Saul, L. K. Savage, S. & Voelker, G. M. (2009). Beyond blacklists, In Proceedings of the 15th ACM SIGKDD International conference on Knowledge discovery and data mining - KDD '09, 1245.

[13] Liu, P., & Moh, T.S. (2016). Content based spam E-mail filtering. Proc. - 2016 Int. Conf. Collab. Technol. Syst. CTS-2016.218–224

[14] Shirazi, H., Haefner, K. &. Ray,I. (2017). Fresh-Phish: A framework for auto-detection of phishing websites. Proc. - 2017 IEEE Int. Conf. Inf. Reuse Integr. IRI 2017, 2017(Jan). 137–143

[15] Singh, P., Maravi,Y.P.S. & Sharma,S. (2015). Phishing websites detection through supervised learning networks. Proc. Int. Conf. Comput. Commun. Technol. ICCCT 2015. 61–65
[16] Agrawal, N. (2016). Detection and Spam Mail Filtering Approach. 99–104
[17] Thomas, J., Raj,N.S., & Vinod, P. (2014). Towards filtering spam mails using dimensionality reduction methods. Proc. 5th Int. Conf. Conflu. 2014 Next Gener. Inf. Technol. Summit, 163–168
[18] Phishing Dataset for Machine Learning: Feature Evaluation. (2018). 1. Mar.
[19] OneR. Retrieved on 8 February 2020 from https://www.saedsayad.com/oner.htm.
[20] Parsania, V. S., Jani, N. N. & Bhalodiya, N.H. (2014). Applying Naïve bayes , BayesNet , PART, JRip and OneR Algorithms on Hypothyroid Database for Comparative Analysis. Int. J. Darshan Inst. Eng. Res. Emerg. Technol. 3(1). 1–6
[21] Backer, B.G. (1998). Visualizing Decision Table Classifiers. Proceedings of the 1998 IEEE Symposium on Information Visualization, 1998. Retrieved on 31 January 2020 from https://dl.acm.org/doi/10.5555/647341.721218
[22] Gupta, A., Mohammad, A., Syed, A. & Halgamuge, M.N. (2016). A Comparative Study of Classification Algorithms using Data Mining: Crime and Accidents in Denver City the USA. 2016.
[23] Miertschin, S.L. WEKA: WEKA A Data Mining Tool.